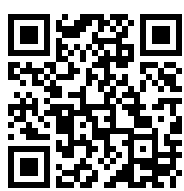


[Formal] Approaches to [Slavic] Linguistics

The Philadelphia Meeting

1999

Michigan Slavic Publications

Formal Approaches to Slavic Linguistics 8, 1999

Michigan Slavic Publications is a non-profit organization associated with the Department of Slavic Languages and Literatures of the University of Michigan. Its goal is to publish titles which substantially aid the study and teaching of Slavic and East European languages and cultures. The present volume, based on a colloquium held at the University of Pennsylvania, Philadelphia in May 1999, continues a series of conference proceedings devoted to formal approaches to Slavic linguistics.

Michigan Slavic Materials, vol. 45

Annual Workshop on Formal Approaches to Slavic Linguistics

The Philadelphia Meeting 1999

edited by
Tracy Holloway King
and
Irina A. Sekerina

Michigan Slavic Publications
Ann Arbor 2000

Library of Congress Cataloging-in-Publication Data

Workshop on Formal Approaches to Slavic Linguistics (8th : 1999 : Philadelphia, Pa.)
Annual workshop on formal approaches to Slavic linguistics : the Philadelphia meeting, 1999 / edited by Tracy Holloway King and Irina A. Sekerina.
p. cm. -- (Michigan Slavic materials ; v. 45)
Includes bibliographical references.
ISBN 0-930042-84-0 (alk. paper)
1. Slavic languages--Congresses. 2. Computational linguistics--Congresses. I. King, Tracy Holloway, 1966- II. Sekerina, I. A. (Irina A.), 1961- III. Title. IV. Michigan Slavic materials ; no. 45.
PG13 .M46 no. 45 ,
[PG 11]
491.'8--dc21

99-089439

Michigan Slavic Publications
Department of Slavic Languages and Literatures
3040 Modern Languages Bldg.
University of Michigan
Ann Arbor, MI 48109-1275

e-mail: michsp@umich.edu

The present volume contains a selection of the papers presented at the Eighth Meeting of Formal Approaches to Slavic Linguistics which took place May 21-23, 1999, at the Institute for Research in Cognitive Science at the University of Pennsylvania. We wish to thank the Institute for generously funding the conference.

This meeting of FASL reflected two newly emerging fields in Slavic studies: computational linguistics and adult psycholinguistics. This is inspired by the explosion of new developments in cognitive science which are transforming traditional linguistics. Each of these areas had a special session devoted to it, and five of the six papers presented are included in this volume: The session on computational Slavic linguistics comprised papers by P. Chew, N. Leko, M. McShane et al., and S. Sheremetyeva and S. Nirenburg; the session on Slavic adult psycholinguistics comprised papers by J. Nicol and R. Wilson, I. Sekerina, and G. Zybatow and G. Mehlhorn. It is our hope that this initiative, which emerged from FASL 8, will continue to be supported and grow through forthcoming FASL meetings.

We are very pleased to note that the prestige of FASL has grown over time. The number of abstracts submitted has steadily increased and the reviewing has become correspondingly more stringent. This year also witnessed a burgeoning of representation from many more countries including Canada, England, Germany, Norway, Poland, Russia, as well as the United States.

There were twenty-seven papers presented at the conference, three of which were contributed by invited speakers: G. Corbett, A. Kroch and C. Heycock, and D. Zec. All participants were invited to submit their papers for publication in this volume. The papers here have been rigorously edited for both content and style, with special assistance on some of the manuscripts from C. Bethin (SUNY at Stony Brook), E. Borg (Reading University, England), K. Dziwirek (University of Washington), J. Fry (Stanford University), and N. Kim (IRCS, University of Pennsylvania). We are grateful for their help in maintaining the high editorial standards of the FASL volumes.

Twenty-one of twenty-seven papers presented appear in this volume. The remaining six papers which were not submitted for publication are: G. Alexandrova "Where Derivational Space and Time Meet, Clitics Have a Say"; M. Babyonyshev "'Missing' Verb Classes in Russian"; N. Friedberg "Russian metrics and Stochastic Constraints: Determining Metrical Complexity"; A. Kroch and C. Heycock "A Cross-Linguistic perspective on Word-Order Variation in Copular Sentences"; I. Sekerina "On-line Processing of Russian Scrambling Constructions: Evidence from Eye Movements During Listening"; and D. Willis "Verb-Raising in Slavic Conditionals."

Tracy Holloway King
Xerox PARC

Irina A. Sekerina
Rutgers University Center for Cognitive Science

Contents

Infinitival Existential Sentences in Russian: A Case of Syntactic Suppletion

Leonard H. Babby
Princeton University

1. Introduction

There is general agreement that the morphosyntactic relations of sentence pairs like (1) are highly complex and "marked." I shall argue that this impression is due in part to the fact that, while the affirmative sentence in (1a) has the same structure as it did in Old Russian (OR), its negated counterpart in (1b) has developed a structure that is radically different from (1a). (2) is the past tense of (1); (3)-(4) are additional examples.

(1) a. Nam est' gde spat'.
us.DAT there-is where to-sleep
'There is somewhere for us to sleep.'
b. Nam negde spat'.
us.DAT nowhere to-sleep
'There is nowhere for us to sleep.'

(2) a. Nam býlo gde spat'.
us.DAT was.N.SG where to-sleep
'There was somewhere for us to sleep.'
b. Nam negde bylo spat'.
us.DAT nowhere was.N.SG to-sleep
'There was nowhere for us to sleep.'

(3) a. Est' komu o nix pisat'.
there-is who.DAT about them to-write
memuary.
memoirs.ACC
'There is someone to write memoirs about them.'
b. Nekomu o nix pisat' memuary.
noone.DAT about them to-write memoirs.ACC
'There is noone to write memoirs about them.'

(4) a. Nam est' čto čitat'.
us.DAT there-is what.ACC to-write
'There is something for us to read.'
b. Nam nečego čitat'.
us.DAT nothing.GEN to-read
'There is nothing for us to read.'

I shall argue that while the semantic relations between the affirmative and negated sentences in (1) are compositional, the morphosyntactic relations are suppletive. Sentences like (1) are of considerable theoretical interest because their derivation involves the interaction of several disparate constructions, i.e.: existential sentences, impersonals, infinitive clauses, free relatives, the genitive of negation, *wh*-movement and extraction from nonfinite CP, copula-introduction, and the formation of postsyntactic words.

In the process of determining what is regular and predictable in the morphosyntactic relations between (1a) and (1b), I will attempt to pinpoint the property or properties that make these relations appear to be anomalous. Our focus will be the status of [*né*-K] words (*negde* 'nowhere,' *nekogo* 'noone,' *nečego* 'nothing' etc.): Are they lexically-stored negative pronouns, parallel to *nigde* 'nowhere,' *nikogo* 'noone' etc. (cf. *Nikomu ne sygrat' ètu rol'* 'No one can play this role' ~ *Nekomu (*ne) sygrat' ètu rol'* 'There is no one to play this role'), or are they derived predicates that result from a post-lexical operation that merges *ne* and the K word (= *wh*-word in English) into a single word in the course of the sentence's derivation? Both positions have been proposed in the recent literature (see Apresjan and Iomdin 1989, Yoon 1989).

2. The Structure of the Affirmative IES

The function of affirmative infinitive existential sentences (hereafter IES) like (1a) is to assert that the necessary conditions exist for the realization of the action denoted by the infinitive (which determines the K word's case, theta role, and grammatical function). Thus a form of the existential verb *est'* 'there-is' and the

infinitive CP are the two obligatory components of the IES construction, i.e.: [IP est' [CP Ki [IP PRO [VP Vinfin...ti...]]]].

2.1. The "Understood" Dative Subject of the Infinitive

The infinitive clause *gde spat'* 'where to-sleep' in (1) is an argument of *est'* 'there is.' I assume that (1a-b) are impersonal rather than having a null expletive, but this assumption is not crucial in what follows (see Babby 1989). Since it has been established that infinitive clauses in Russian all have a dative subject, overt or null (see Babby 1998), the null hypothesis is that the optional dative NP in IESs, which is always construed as the infinitive's subject, is the preposed dative subject of the infinitive clause, i.e., the structure of (1a) can be schematically represented as (5). If the subject of the infinitive clause is null (PRO), as in (5b), it receives an arbitrary-reference interpretation; if it is overt, it is normally fronted, occupying the Spec position in the sentence's highest functional projection. The same thing happens in other impersonal infinitive constructions: *Nami bylo [IP ti ne provesti ee]* 'We couldn't fool her'; *Vami ne imeet smysla [IP ti zanimat'sja muzykoj]* 'It doesn't make sense for you to study music' (see Lavine, Harves, and Billings (to appear) for discussion of XP fronting in Russian).

(5) a. Nami [IP [VP [V est'] [CP gdej [IP ti spat' tj]]]]
b. [IP [VP [V est'] [CP gdej [IP PROi spat' tj]]]]

It has been proposed that the dative *nam* 'us' in (1a) is an argument of the matrix existential verb *est'* and obligatorily controls the PRO subject of the infinitive clause (see Rappaport 1986). The reason for this claim appears to be that extraction of the subject of the infinitive clause would incur a subjacency violation. There are, however, at least three reasons for assuming that the dative in (1) is in fact the subject of the infinitive. First, it is clear from sentences like (6)-(7) that pronouns other than the dative pronouns systematically prepose out of the infinitive clause, and we do not

want to claim, for example, that accusative *menja* in (6) is also an argument of *bylo* that controls a null direct object of *vstrečat'*:

(6) Menja$_i$ nekomu bylo [CP vstrečat' t$_i$].
me.ACC noone.DAT was.N.SG to-meet
'There was no one to meet me.'

(7) Ne dumaju, čtoby teper' nas$_i$ bylo [CP za čto upreknut' t$_i$].
'I don't think that there is anything to rebuke us for now.'

The unnaturalness of the word order in sentences like *Est' gde nam spat'* 'There-is somewhere for-us to-sleep,' which are closer to the "underlying" structure than (1a), is due to a violation of the EPP; see Lavine, Harves, and Billings (to appear) for discussion. This word order is, however, entirely natural in questions, where *est'* raises to the Spec position and the infinitive's dative subject *nam* remains *in situ*: *Est' gde nam spat'?* 'Is-there somewhere for-us to-sleep?' (cf. *Gde nam spat?'* 'Where should we sleep?', which is not an IES).

Second, when the K-word is itself the overt dative subject of the infinitive, the putative (non-K) dative argument of *est'* cannot occur (see (3a) and (6)), which is not what is predicted under the hypothesis that the dative in IESs is an argument of of *est'*; (3a) is in fact precisely what we would expect under the hypothesis proposed here that the dative is the subject of the infinitive, not the finite matrix verb.

Third, it will be argued below that while stressed *ést'* (*býlo*, *búdet* 'there-is (was, will be)' in Affirm-IESs is an existential verb, which can take arguments (i.e., assign theta roles), the finite matrix verb in the corresponding Neg-IES is the (unstressed) copula. Since the copula, like all auxiliary verbs in Russian, does not assign theta roles (Babby, to appear), the dative in Neg-IESs cannot be an argument of the matrix verb. Since we obviously do not want to treat the dative in Affirm- and Neg-IESs differently, this argument, along with the other two, constitutes a sufficiently strong case for treating the optional dative in IESs as the preposed subject of the infinitive clause.

2.2. The CP Infinitive Clause

It has been widely assumed that the CP infinitive clause in IESs is a *free relative* clause. However, Izvorski 1998 argues that it has the structure of an indirect question: the K words used in IESs can function as interrogative as well as relative pronouns. We return to this problem in Section 6, after discussing Neg-IESs. We assume for now that the infinitive clause in IESs is just what it appears to be: a bare CP (with a K-word in Spec positon) that serves as an argument of the matrix predicate. Infinitive clauses are never "independent," i.e., they must be embedded in a higher structure (Babby 1998). I shall argue that, while the bare infinitive CP structure in (8) is common to IESs, free relatives, and indirect questions, we cannot reduce IESs to the other two: all three differ in terms of what CP is embedded in.

(8) $[_{CP}\ K_i\ [_{IP}\ PRO_{dat}\ [_{VP}\ V_{infin}\ ...t_i...]]]]$

The use of putative "free relative" clauses in constructions other than the IES construction is highly restricted in standard Russian: *najti(s')* 'to-find' is the most common matrix verb used in free relatives, e.g.: *Najdetsja [komu sčitat' den'gi]* 'Someone.DAT to-count the-money will-be-found'; *V xolodil'nike najdete [čto / čego poest']* 'You'll find something.ACC/GEN to-eat in-the-frige' (see Rappaport 1986; Zemskaja 1973 for discussion of the status of free relatives in standard and colloquial Russian). But the fact that the dative subject of the infinitive in IESs, as well as other constituents of the infinitive clause (cf. (6)-(7)), can be systematically extracted casts doubt on the hypothesis that we are dealing with free relatives in the case of Russian IES: true free relatives are headless relative clauses and CP is thus embedded in a higher NP (see Hirschbuhler and Rivero 1983). Now, if there were an NP dominating the CP in IESs, we would not expect the dative subject or any other constituent of the infinitive clause to be extracted since this would constitute an unacceptable subjacency (complex-NP) violation: *Nam$_i$ est'* $[_{NP}...[_{CP}...[_{IP}\ t_i\ [_{VP}\ ...]]]]$.

3. Negated Infinitival Existential Sentences

It is clear from a comparison of (1)-(4) that the b)-sentences are not the simple syntactic negation of the a)-sentences, i.e., the structure of the negated sentence is not identical to that of the affirmative plus the introduction of the canonical negation marker *ne* (cf. "simple negation" in *Ona ljubit tancevat'* 'She loves to dance' ~ *Ona ne ljubit tancevat'* 'She doesn't love to dance'). If we assume that *ne* in (1b) is the canonical negation marker used in simple negation, then we must treat the following five distinctive morpho-syntactic properties of the Neg-IES as anomalies.

3.1. Position of *Ne* in Neg-IES

Ne in (1b) is syntactically related to the K-word of the infinitive clause rather than to the matrix verb. In Russian, *ne* is canonically placed immediately before the verb in cases of *sentence negation*, marking the left periphery of the scope of negation:..*ne* [VP V..; so IESs do not appear to invlove simple sentence negation. We shall see below that the position of *ne* in Neg-IESs cannot be accounted for by claiming that (1b) involves *constituent* negation. Sentence vs. constituent negation is illustrated by the following pair: *On ne kupil gazet* 'He didn't buy newspapers.GEN' ~ *On kupil ne gazety/*gazet* 'It wasn't newspapers.ACC(*GEN) that he bought'. If (1b) involves sentence negation, not constituent negation, as suggested by its semantics, we expect *ne* to be associated with the matrix verb *est'*, i.e., the result of negating (1a) should thus be (9), not (1b): *net* < *ne est'*. While sentences like (9) do in fact occur in colloquial Russian, the existence of the Neg-IES construction in (1b) appears to block their use in standard Russian. I argue below that the Neg-IES does in fact involve sentence negation and that its formal resemblance to constituent negation is entirely fortuitous.

(9) a. Nam (u nas) net [gde spat'].
us.DAT (at us) there-is-not where to-sleep
'There is (we have) nowhere to sleep.'

b. Nam (u nas) ne bylo [gde spat'].
us.DAT (at us) neg was.N where to-sleep
'There was (we had) nowhere to sleep.'

3.2. Obligatory Absence of *Est'* in Neg-IESs

The existential verb *est'* in (1a) is obligatorily absent in its negated counterpart (1b). Given that *ne* is associated with the K-word, we might expect the negated counterpart of (1a) to be (10) rather than (1b). Thus what appears to be anomalous about (1b) is that the matrix verb *est'* 'there is' in (1a) has a null realization in (1b).

(10) *Nam est' negde spat'.
us.DAT there-is nowhere to-sleep

3.3. The Position of *Bylo (Budet)* and the K-word

The third anomaly deals with the relative position of the past tense *bylo* 'was.N' and the K-word; we shall be considering only the neutral, unscrambled word order. Compare the relative position of *bylo* and *gde* in (2a-b) above: *bylo* must precede *gde* in the affirmative but it must follow *negde* in the negated sentence.

3.4. The Stress of *Bylo (Budet)*

In Affrim-IESs, *est'/bylo/budet* 'there-is/was/will-be' are stressed, whereas in Neg-IESs *bylo* and *budet* are de-stressed and behave like clitics (see Kubík 1971:53-9), just as in sentences whose main predicate is a modal: *Ja dolžen byl ostat'sja* 'I had to-stay' ~ **Ja býl dolžen ostat'sja /*Ja dolžen býl ostat'sja* (cf. (2)).

3.5. Systematic Gaps In The Case Paradigm Of K-words

The fifth apparent anomaly deals with gaps in the case paradigm of the nominal K-words, *kto* 'who' and *čto* 'what.' When they are the accusative direct object in the Affirm-IES, they must be in the

genitive case rather than the accusative in the corresponding Neg-IES; see (4).

The nominative forms of *kto* and *čto* do not occur in either affirmative or negated IESs. This nominative case gap is a natural consequence of the fact that the K-word receives its case in the infinitive clause and, as noted above, infinitives in Russian assign dative to their subject, not nominative. We shall see below that the nonoccurrence of bare accusative K-words in Neg-IESs has an entirely unrelated explanation.

Summary: Given the structure of Affirm-IESs proposed in (5) and the canonical position of *ne* in sentences with sentence negation, there is no obvious explanation for why negation of Affirm-IESs should entail precisely the set of five "anomalous" properties described above. In a truly explanatory analysis, all five properties should fall out naturally, i.e., none of them should have to be stipulated as a unique property of the IES construction. I shall argue that the properties of Neg-IESs cannot be explained in terms of the principles of Russian negation as long as *ne* in *negde* 'nowhere,' *nekogo* 'noone,' etc. is construed as the canonical negation marker.

4. Diachronic Analysis: IES in Old Russian

The explanation I shall propose for the five properties identified above is based on the following diachronic observation (Potebnja 1958, Lomtev 1956): The *ne* in *negde* in (1b) is not the negation marker *ne* (не); it is clear from Old Russian that *ne* in *negde*, *nekuda* 'nowhere,' *nečego* 'nothing,' etc. derives from the finite negated existential verb *ně* (нѣ) 'there-is-not' (/ě/ subsequently merged with /e/, obscuring their different origins). Consider the OR Neg-IES in (11a) (OR *kamo* = *kuda* 'whence' in MR); (11b) is its structure; (12) is an example in the past tense:

(11) a. Uže nam ně kamo sja děti.
already us.DAT there-is-not where clitic to-go
'There is no longer anywhere for us to go.'

b. Nam_i [IP [VP [ně]V [CP $kamo_j$ sja_k [IP t_i děti t_k t_j]]]]

(12) a. Ne bystь imъ kuda pereĕxati. (OR)
NEG was them.DAT where to-cross-to
'There was nowhere for them to cross over.'
b. Im nekuda bylo pereexat'. (MR translation of (12a))
'There was nowhere for them to cross over.'

The OR present-tense negated existential verb *nĕ* in (11a) was formed earlier from the canonical negative marker *ne* + *e(st')* (cf. *ne* + *byst'* in (12a)) and was thus one of five competing finite negated existential verbs in OR: *nĕ*, *nětъ*, *nĕstь*, *nětu*, *nětutь*. Only *net* remains in MR as the *finite* negated form of existential *est'* (*netu* has been reanalyzed as its emphatic form). (13) is an example of the use of *nĕ* in OR not involving IES.

(13) Nĕ togo, iže by moglъ
there-is-not that-one.GEN who might able
na onu stranu doiti.
on other side to-go
'There is no one who is able to go to the other side.'

Note that OR (11) is identical to the structure of the MR Affirm-IES (1a) proposed in (5): the only significant difference is that in place of the affirmative existential verb *est'* we find the OR negated existential verb *nĕ*; compare:

(5a) Nam$_i$ [IP [VP [V est'] [CP gde$_j$ [IP t$_i$ spat' t$_j$]]]]
(11) Nam$_i$ [IP [VP [V nĕ] [CP kamo$_j$ sja$_k$ [IP t$_i$ děti t$_k$ t$_j$]]]]

OR *nĕ* was lost everywhere except in IESs, where it was immediately adjacent to the K-word in the Spec of its CP complement, i.e., *nĕ-K* in *...nĕ [CP K C'...* was reanalyzed as a single word [*nĕ*-K]. The details of this reanalysis and the morphosyntactic status of [*né*-K] in MR are discussed in Section 6. For now we need note only that *nĕ* is no longer an independent, finite verb in MR. It has become a dependent predicate word that is "bound" to the K-word it is adjacent to. We shall refer to OR *nĕ* in MR as *né-* so as not to confuse it with the canonical negation marker *ne*.

Thus *né-* is a dependent negated existential predicate in MR. OR *ně* also survives in MR in the nonfinite modal predicate word *nel'zja* (+ infinitive clause) 'it is impossible to' (*nel'zja* < OR *ně lьzě* (GEN) 'there is no use'). I argue below that the [*né*-K] words are themselves predicates (i.e., *negde* is the matrix predicate of (1b)), which explains why they freely conjoin with other nonfinite predicate words; for example: *Pridetsja uxodit'. No ved' nekuda i nevozmožno* 'I have to leave, but there is nowhere to go and it is impossible to go' (cf. (27)-(28); Apresjan and Iomdin 1989:43). Russian has other types of K-predicates, e.g.: *Sprosili, kakov bol'noj* 'They-asked how the-patient (was feeling).'

5. The Explanation of Sections 3.1.-3.5.

The reanalysis of OR *ně* as MR *né-* enables us to show that the five properties described in Sections 3.1-3.5 are not anomalies.

The obligatory positioning of *ne* before K in Neg-IESs is neither deviant sentence negation nor constituent negation (cf. 3.1.). The existential verb *ně* in OR was immediately adjacent to its CP complement (cf. (11)) and, therefore, the position of *né-* in MR Neg-IESs before the K word is exactly what we expect: CP is the complement of *né-* and K is in Spec of CP and thus its left-most constituent. In MR Affirm-IESs, the neutral position of *est'* (*býlo*) is also immediately before the K-word (cf. (5a)), only here *est'* and K remain independent words, as in OR. Thus the constituent negation hypothesis is wrong because it is based on a fortuitous superficial similarity between *né-K* and the position of *ne* in constituent negation (*On pošel ne tuda, kuda nado* 'He didn't go where he should have' (cf. **nétuda*)).

We now have a natural explanation for the ungrammaticality of (10) (**Nam est' negde spat'*), i.e., for why the existential verb *est'* must be absent in MR Neg-IESs (cf. 3.2.). According to the analysis of *né-* from OR *ně*, the existential verb is contained in the MR words *negde, nekogo*, etc. (*né-* < OR *ně* < *ne+e(stь)*) and thus cannot occur for a second time in the matrix clause as *est'*.

This brings us to a crucial point in our analysis. We see that Affirm- and Neg-IESs have radically different syntactic structures

in MR, which is due to the reanalysis of the OR finite existential verb *ně* as a nonfinite, bound predicate that obligatorily merges with the immediately adjacent K-word. This means that *né-* in MR still functions as an existential predicate, despite the fact that it is patently no longer an independent, finite verb. This reanalysis had the following consequence: While *ést', býlo, búdet* in MR Affirm-IESs are existential forms of the verb, *bylo* and *budet* in Neg-IESs are *not* forms of the existential verb, which is, as we saw above, encapsulated in [*né*-K]. Unstressed *bylo* and *budet* are forms of the copula, which is null in the present tense (see 3.4., Apresjan and Iomdin 1989). We see this use of the copula with other nonfinite predicate words, e.g.: *Nam nel'zja bylo spat'* 'We couldn't sleep'. In other words, the copula is automatically introduced in a derivation to express tense whenever a finite predicate is made nonfinite (cf. passive participles and infinitival impersonals like *Teper' ego bylo ne ostanovit'* 'It-was now impossible to-stop him.'

We saw in Sections 3.3 and 3.4 that, while *ést'/býlo/búdet* is stressed and precedes the K word in Affirm-IESs, *bylo* and *budet* in the corresponding Neg-IES are de-stressed (encliticized) and normally follow the [*né*-K] word (*né-* is itself stressed because it derives from the negated existential verb, which is stressed in this construction); cf. (14)-(15).

(14) a. Nam býlo gde spat'.
us.DAT there-was where to-sleep
'There was somewhere for us to sleep.'
b. *Nam gde býlo spat'.

(15) a. Nam negde bylo spat'.
us.DAT nowhere was to-sleep
'There was nowhere for us to sleep.'
b. *Nam býlo negde spat'. (cf. (10))
c. ?Nam bylo negde spat'.

The fact that the existential verb is stressed and precedes the K-word in Affirm-IESs follows from the structure given in (5). The de-stressed, enclitic status of *bylo* and *budet* in Neg-IESs is precisely what we expect given our hypothesis that they are forms of

the copula: It is normally the case in MR that the copula cliticizes onto matrix nonfinite modal predicates (see (16)). Thus *né-* is behaving here like the other nonfinite predicate words with respect to the positioning of the copula (examples from Isačenko 1965:284).

(16) a. Ja dólžen byl vernut'sja.
I.NOM supposed was to-return
'I was supposed to return.'

b. Emu móžno bylo kurit'.
him.DAT possible.N.SG was.N.SG to-smoke
'He could smoke.'

The question naturally arises here why the unscrambled order in Neg-IESs is different from that of the corresponding affirmative, i.e., why isn't (15c) the basic order, with de-stressed *bylo* cliticized onto *nam* (it is after all closer to the "underlying" structure)? While the order in (15c) is not ungrammatical, it is clear that *Nam negde bylo spat'* ((1b)) is the neutral order. In other words, why does [*né*-K] appear to raise to a position in the matrix sentence higher than the copula in (2b)? The answer to this question is that [*né*-K] behaves like the other nonfinite predicate words, occupying a position to the left of the copula (see (16), Isačenko 1965). But the reason that the other modal nonfinite predicates precede the copula goes well beyond the scope of this paper because it would require extensive discussion of the basic position of the modal projection (MP) in Russian.

We now come to the fifth of the putative "anomalies" (cf. 3.5). The accusative form of *kto* and *čto* in Affirm-IESs is always replaced by the genitive in the corresponding Neg-IES; see (4). The null hypothesis is that this is simply an instance of the genitive of negation, which is consistent with our claim that *né-* is a negated existential predicate in MR: The NP argument of a negated existential verb in Russian is always genitive, never nominative or accusative (cf. *net* 'there isn't'+ genitive).

There is one systematic exception to the use of genitive for accusative in Neg-IESs: if the K-word is governed by a preposition requiring the accusative case, the accusative is normally retained:

(17) a. Nam est' vo čto zavernut' rybu.
us.DAT there-is in what.ACC to-wrap fish.ACC
'There is something for us to wrap the fish in.'
b. Nam ne vo čto zavernut' rybu.
us.DAT there-is-not in what.ACC to-wrap fish
'There is nothing for us to wrap the fish in.'

Sentences like (17b) appear at first glance to provide independent evidence for the hypothesis that the genitive instead of the accusative in Neg-IESs is an instance of the genitive of negation: PP, the maximal projection of prepositions, normally forms a barrier to the genitive of negation; cf. (18):

(18) a. Ona smotrit [PP na sestru].
she.NOM looks at sister.ACC
'She is looking at her sister.'
b. Ona ne smotrit na sestru (*sestry).
she.NOM not look at sister.ACC/*GEN
'She is not looking at her sister.'

However, there is evidence suggesting that this explanation of the genitive in Neg-IESs may be an oversimplification. First, Affirm-IESs sentences like (19) show that the genitive in Neg-IESs may in fact be an instance of the quantitative (partitive) genitive; see Apresjan and Iomdin 1989:47.

(19) Každomu est' [čego skryvat'].
each.DAT there-is what.GEN to-hide
'Everyone has something to hide.'

(20) Tam est' čego est'? Est' čego.
there there-is what.GEN to-eat there-is what.GEN
'Is there anything to eat there? There is.'

Second, in sentences like the following, there is a CP boundary separating the negated matrix verb from the genitive K-word in the complement clause: case assignment in general, and the genitive of

negation in particular, does not as a rule cross clause boundaries ((21) is from MR, (22) is Old Church Slavonic, (23) is OR).

(21) Net li u tebja [čego poest']? [MR]
there-is-not Q at you what.GEN to-eat
'Don't you have anything to eat?'

(22) Ne imamь [čьso položiti prědъ nimь]. [OCS]
not have what.GEN to-put before him
'I do not have anything to place before him.'

(23) Ne imutъ [česo ěsti]. [OR]
not have what.GEN to-eat
'They do not have anything to eat.'

Finally, since PP is a barrier to the genitive of negation in standard Russian (see (18)), sentences like the following constitute particularly compelling evidence against the hypothesis that the genitive in Neg-IESs is simply the genitive of negation (the preposition *čerez* takes an accusative NP complement). (24a) is from D. Truskinovskaja; (24a) is from I. Sekerina (cf. (17b)).

(24) a. Emu i kuvyrkat'sja ne
him.DAT and stumble there-is-not
čerez čego bylo.
over what.GEN was
'There was really nothing for him to stumble over.'

b. Nam ne v čego zavernut' rybu.
us.DAT there-is-not in what.GEN to-wrap fish
'There is nothing for us to wrap the fish in.'

As far as I have been able to determine, examples like (24) have never been discussed in the literature. The data are quite robust: the native speakers I have consulted all find these sentences to be natural and colloquial.

6. The Structure of Neg-IESs in Modern Russian

It emerges clearly from the analysis proposed in the first five sections that the IES construction's central problem is the morphosyntactic status of the [*né*-K] predicate in MR: While we know that *né-* (< *nĕ*) and K were independent words in OR and that they combined as the result of the reanalysis triggered by the loss of *nĕ* as an independent finite verb, it is not immediately clear how to treat [*né*-K] in MR. Should it be treated as a single lexical item, e.g., a negative pronoun or predicate, as often proposed in the literature, or as two separate words, each with its own lexical entry in the mental lexicon, that merge in the course of the sentence's (synchronic) derivation to form a single syntactic unit? If [*né*-K] results from combining two lexical items, what kind of a derived word is it: a *morphological* word (e.g., a compound or the result of incorporating K into *né-*), a *syntactic* word (cf. Chvany's 1996 treatment of the relation between suppletive *net* in the negated existential paradigm *ne bylo, net, ne budet* 'there wasn't, isn't, will not be,' or a postsyntactic *phonological* word (e.g., contractions like *shouldn't*)? See DiSciullo and Williams 1987 for discussion. Below I will eliminate some of these alternatives and outline what I take to be the simplest analysis that accounts best for the facts.

One of the problems with treating [*né*-K] as a single, nonderived lexical item (a morphological word) in MR is that it does not have the syntactic distribution of a word. Actually, [*né*-K] simultaneously satisfies the distribution and selectional properties of two different words, which strongly suggests that it is the product of a late syntactic rule or, far more likely, a postsyntactic merger of two lexical items: *né-* and K each satisfies its own syntactic distribution in its own clause before combining into a single unit (see Christensen 1986 for a similar treatment of Norwegian *igen* < *né einn-gi* 'not anyone'); Apresjan and Iomdin 1989 argue for essentially this analysis.

I am thus proposing that the derivation of [*né*-K] is parallel to the derivation of complex words formed from the bound quantifier *pol-* 'half' in MR: The syntax of NPs containing *pol-* clearly demonstrates that it initially occupies the same head-external position

in the NP as the numbers 2-4 before it merges with the immediately adjacent head noun N (to which it assigns genitive case) to form a phonological word, i.e.: [$_{N}$*pol-* N$_{gen}$]. The genitive plural of the adjective *celyx* 'entire' in *On vypil [$_{NP}$acc celyx polstakana vodki]* 'He drank an entire.GEN.PL half glass.GEN of vodka' shows that *pol-* must occupy the head-external, QP position early in the sentence's derivation, i.e. [$_{NP}$ [$_{AP}$ *celyx*] [$_{QP}$*pol-*] [$_{N}$*stakana*] *vodki*]: *pol-* here c-commands both *celyx* and *stakana* and assigns them both genitive case (see Babby 1987 for details). But the fact that *polstakana* can be the object of the preposition *s* 'approximately' demonstrates that *pol-* and the head noun *stakana* must have subsequently merged to form a phonological word [$_{N}$*polstakana*] since *s* can combine only with nouns (with N, not N' or NP): *On vypil s [$_{N}$ polstakana] vodki* 'He drank about a half glass of vodka,' but **On vypil s polnyj stakan (tri stakana) vodki* 'He drank almost a full glass (three glasses) of vodka.'

While it might seem feasible to treat indeclinable [*né*-K] words like *negde* 'nowhere' as synchronically nonderived lexical items, this analysis runs into serious difficulties when extended to the declined [*né*-K] words. But the most convincing argument against the single-word analysis comes from sentences like (17), (24), and (25): the K-word here is governed by a preposition, which is determined inside CP as a selectional property of the infinitive, i.e.: [*né-* [P K]$_{pp}$]. (The convention is to spell all three components separately, i.e., *ne s kem*, not *neskem*). Assuming that the optimal solution to the problem of the status of [*né*-K] words is the one that treats all Neg-IESs in a unified way, sentences like (17), (24), and (25) virtually eliminate the underived analysis: Assuming the Lexicalist Hypothesis, a morphological word cannot contain maximal (syntactic) projections (XP), i.e., *[*né*-PP]$_{X}$, where X defines a morphological word (see DiSciullo and Williams 1987; Lieber 1992 and Hale and Keyser 1993 propose a different view of the autonomy of morphology and syntax). Note in (26b) that *ne s kem*, like *negde*, functions as a single unit with respect to final word order (see Section 3.3, where it is suggested that the [*né*-K] predicate word raises to a position higher than the copula projection). (26) is the past tense of (25).

(25) a. Nam ést' s kem posovetovat'sja.
us.DAT there-is with whom.INST to-consult
'There is someone for us to consult with.'
b. Nam né s kem posovetovat'sja.
us.DAT there-is-not with whom to-consult
'There is noone for us to consult with.'

(26) a. Nam býlo s kem posovetovat'sja.
'There was someone for us to consult with.'
b. Nam né s kem bylo posovetovat'sja.
'There was noone for us to consult with.'

We shall thus assume that [*né*-(P)K] is composed of two discrete expressions, *né-* and (P)K, which merge into a single derived syntactic unit (word) in the course of the sentence's derivation. *Né-* is a predicate word (Pr(edicate)) that selects as its agrument a "K-clause" which is closer in structure to an indirect question than to an infinitival free relative, but cannot be reduced to either. Unlike the other nonfinite predicate words, *né-* is a dependent predicate.

The initial, pre-movement structure of an Affirm-IES and its negated counterpart are schematically represented in (27)-(28): *est'* and *né-* are both predicates, as in OR, but, in MR, *est'* is a finite form of the verb, while *né-* is a nonfinite bound predicate word. According to this analysis, the preposition does not "intervene" between the *né-* and *kem* (**né-kto* > *ne s kem*): *né-* composes with [PP *s* [NP *kem*]]); see Harves 1998. ("Pr" stands for nonfinite predicate word.)

(27) a. (Nam) ést' s kem posovetovat'sja. (= (25a))
b. [IP [est']$_V$ [CP s kem$_i$ [C' [IP (nam) posovetovat'sja t$_i$]]]]

(28) a. (Nam) né s kem posovetovat'sja. (= (25b))
b. [IP [cop]$_V$ [PrP [né-]$_{Pr}$ [CP s kem$_i$ [IP (nam) posovetovat'sja t$_i$]]]]

Given this analysis of [*né*-(P)K], the crucial issue in the derivation of IESs can be narrowed down to the following question: How

late in the sentence's derivation do *né-* and (P)K merge to form a word? In other words, what kind of a "word" do *né-* and (P)K form? While the simplest answer, as suggested above, would be that [*né-*(P)K] is a "phonological" word, i.e. *né-* and K merge to form a unit after the syntactic derivation is completed, there is at least one phenomenon that appears to suggest that they may in fact merge at an earlier point in the derivation: If the genitive case of the direct object of the infinitive in Neg-IESs like (4) were in fact assigned by *né-*, this would constitute evidence that *né-* and (P)K would have to merge in the syntax per se, assuming that genitive case assignment in the scope of negation is a syntactic operation and that the scope of negation does not extend beyond the negative operator's clause (i.e., *né-* cannot assign genitive case to *kogo/čego* across CP). But we saw above that there is some evidence that the genitive of the K-word in IESs is not simply the genitive of negation (see (19)-(24)), which means that it does not constitute a particularly strong argument against the postsyntactic merger of *né-* and (P)K. The genitive of negation argument against postsyntactic merger seems less significant in the Minimalist approach, where only fully formed words merge and case is "checked" rather than assigned in the course of a syntactic derivation (see Brown 1999).

Since there is no compelling evidence to the contrary, I conclude that *né-* and (P)K merge into a single expression postsyntactically, which has the desirable effect of greatly simplifying the derivation and syntactic representation of IESs: If *né-* and (P)K merged as the result of a syntactic operation, we would have to claim, for example, that (P)K raises out of its Spec position in CP and incorporates into [*né-*]$_{Pr}$, forming a [*né-*(P)K] predicate word whose (P)K component binds a K-trace in Spec of CP, which, however, violates the Lexicalist Hypothesis (according to which, internal components of words cannot have syntactic relations outside the domain of the word).

A few words need to be said about *lexicalization*, which figures prominently in discussions of whether or not [*né-*(P)K] is derived or lexical. Consider the sentences in (29) (what is said about *nečego* holds for *nekogda* 'there is no time to').

(29) a. Mne nečego stydit'sja.
me.DAT nothing.GEN to-be-ashamed-of
'There is nothing for me to be ashamed of.'
b. Mne nečego stydit'sja svoix roditelej.
me.DAT no-point to-be-ashamed my parents.GEN
'There is no point in my being ashamed of my parents.'

(29a) is a canonical Neg-IES: *nečego* is genitive, which is the lexical case assigned by *stydit'sja* to its object (cf. *Mne est' čego stydit'sja* 'There is something.GEN for me to be ashamed of'). But in (29b) *stydit'sja* assigns its genitive case to *roditelej*. *Nečego*, which is historically genitive, has become an independent indeclinable modal predicate word in MR meaning roughly 'there is no point to;' it is no longer the object of the infinitive here. Thus, while *nečego* is a nonfinite predicate word in both (29a) and (29b), its lexical status is entirely different in the two sentences: In (29a) *nečego* is a derived by the merger of *né-* with the K component, which is assigned genitive by the infinitive: [*né-* [$_K$ *čego*]]; but in (29b), *nečego* is a single indeclinable (caseless) lexical item in the mental lexicon that takes an infinitival clause complement (not a K-clause) and has an idiomatic meaning. Thus *nečego* in (29b) has the same lexical status as the predicate word *nel'zja* 'it is impossible.'

Summary: I have argued in this paper that affirmative IESs and their negated counterparts have radically different morphosyntactic structures. This suppletion can be traced back to the loss in OR of the negated existential verb *ně* in all positions except when followed by the K-word of its CP infinitive complement in IESs, in which case it was reanalyzed as a bound predicate *né-* and merged with K to form a phonological word. This derivation accounts for the five "anomalous" properties of Neg-IESs presented in Sections 3.1-3.5 and is a particularly dramatic confirmation of Pinker's observation that "reanalysis is an inexhaustible source of new complexity" (Pinker 1994:244).

References

Apresjan, J.D. and L.L. Iomdin. 1989. "Konstrukcija Tipa *Negde Spat'*: Sintaksis, Semantika, Leksikografija." *Semiotika i Informatika* 29, 34-92.

Babby, L.H. 1987. "Case, Prequantifiers, and Discontinuous Agreement in Russian." *NLLT* 5, 91-138.

Babby, L.H. 1989. "Subjectlessness, External Subcategorization, and the Projection Principle." *Zbornik Matica Srpske za Filologiju i Lingvistiku* XXXII/2 (Novi Sad), 7-40.

Babby, L.H. 1998. "Subject Control as Direct Predication: Evidence from Russian." Z. Bošković, S. Franks, and W. Snyder (eds.) *Formal Approaches to Slavic Linguistics*, 17-37. Ann Arbor: Michigan Slavic Publications.

Babby, L.H. To appear. "Voice and Diathesis in Slavic." George Fowler (ed.) Volume of Papers from the Comparative Morphosyntax Conference at Indiana University, June 5-7, 1998. Indiana: Slavica Publishers.

Brown, S. 1999. *The Syntax of Negation in Russian: A Minimalist Approach*. Stanford Monographs in Linguistics. CSLI Publications.

DiSciullo, A.M. and E. Williams. 1987. *On the Definition of Word*. Cambridge, Mass.: MIT Press.

Chvany, C. 1996. "Syntactically Derived Words in a Lexicalist Theory." O.Yokoyama and E. Klenin (eds.) *Selected Essays of Catherine V. Chvany*, 43-54. Indiana: Slavica Publishers.

Christensen, K. 1986. "Norwegian *Igen*: A Case of Postsyntactic Lexicalization." O. Dahl and A. Holmberg (eds.) *Scandinavian Syntax*, 21- 35. University of Stockholm.

Harves, S. 1996. "The Syntax of Negated Prepositional Phrases in Slavic." Z. Bošković, S. Franks, and W. Snyder (eds.) *Formal Approaches to Slavic Linguistics*, 166-186. Ann Arbor: Michigan Slavic Publications.

Hale, K. and S.J. Keyser. 1993. "On Argument Structure and the Lexical Expression of Syntactic Relations." K. Hale and S.J. Keyser (eds.) *The View from Building 20*, 53-109. Cambridge, Mass.: MIT Press.

Hirschbuhler, P. and M. Rivero. 1983. "Remarks on Free Relatives and Matching Phenomena." *Linguistic Inquiry* 14, 505-520.

Isačenko, A.V. 1965. *Grammatičeskij Stroj Russkogo Jazyka v Sopostavlenii s Slovackim (vol. I)*. Bratislava.

Izvorski, R. 1998. "Non-Indicative *Wh*-Complements of Possessive and Existential Predicates." P.Tamanji and K. Kusomoto (eds.) *Proceedings of NELS 28*, 159-173.

Kubík, M. 1982. *Russkij Syntaksis v Sopostavlenii s Češskim.* Praha: SPN.

Lavine, J, S. Harves, and L. Billings. To appear. "Feature Checking, the EPP, and Diatheses." George Fowler (ed.)Volume of Papers from the Comparative Morphosyntax Conference at Indiana University, June 5-7, 1998. Indiana: Slavica Publishers.

Lieber, R. 1992. *Deconstructing Morphology: Word Formation in Syntactic Theory*. The University of Chicago Press.

Lomtev, L.P. 1956. *Očerki po Istoričeskomu Sintaksisu Russkogo Jazyka*. Moskva: Izd.MGU.

Pinker, S. 1994. *The Language Instinct*. N.Y.: M. Morrow & Co.

Potebnja, A.A. 1958 (=1888). *Iz Zapisok po Russkoj Grammatike*. Moskva: AN SSSR.

Rappaport, G. 1986. "On a Persistent Problem of Russian Syntax: Sentences of the Type *Mne Negde Spat'*." *Russian Linguistics* 10, 1-31.

Yoon, E. 1989. *A Syntactic Study of Russian Sentences of the Type "Mne nečego delat'."* Doctoral Dissertation, University of Washington.

Zemskaja, E.A. 1973. *Russkaja Razgovornaja Reč'*. Moskva: Nauka.

Slavic Departtment, E. Pyne Building
Princeton University
Princeton, N.J. 08544-5264
babbylh@princeton.edu

The *Wh*/Clitic-Connection in Slavic: Theoretical Implications*

Cedric Boeckx and Sandra Stjepanović
University of Connecticut

1. Outline

This paper addresses several, we believe, intricately related issues in Slavic syntax. Rudin (1988) claims that Slavic languages fall into two classes when it comes to multiple *wh*-fronting: the Bulgarian (BG) type, where all *wh*-phrases form a 'group,' and the Serbo-Croatian (SC) type, where *wh*-phrases can be split. The difference is illustrated in (1).

(1) a. ?*Koj, spored tebe, kakvo e kazal? [BG]
who according to you what is said
'Who, according to you, said what?'

b. Ko, po tebi, šta pije. [SC]
who according to you what drinks
'Who, according to you, drinks what?'

What, to the best of our knowledge, no one has noticed is that clitics abstractly pattern the same way as *wh*-elements. Although this may not be visible at the surface, we will show in Section 3

* This is a substantially revised version of the paper we presented at FASL 8. As is obvious from the text, we are most indebted to Željko Bošković for raising the issues discussed here. Deepest thanks also go to Penka Stateva for invaluable comments, and providing crucial data. Thanks, finally, to David Pesetsky, Norbert Hornstein, Arthur Stepanov, Saša Vukić, and the audience at FASL 8 (in particular, Tony Kroch and Martha McGinnis) for their interest. We would also like to thank the audience of the Perspectives on Clitics and Affixes Workshop at the University of Illinois, Urbana-Champaign (July 1999), where this particular version of our *Wh*/Clitic-connection account was presented. Special thanks to Arhonto Terzi, Catherine Rudin, and Juan Uriagereka for invaluable comments on that occasion.

that by various tests, clitics form a cluster in Bulgarian,[1] but not in Serbo-Croatian.

(2) a. Te sa mu go predstavili. [BG]
they are him-dat him-acc introduced
'They introduced him to him.'

b. Oni su mu ga predstavili. [SC]
they are him-dat him-acc introduced
'They introduced him to him.'

This parallelism between phenomenologically distinct elements like clitics ("topics") and *wh*-phrases ("foci"), which we refer to here as the *Wh/Clitic-Connection*, will be at the core of our discussion.

Before proceeding, we must mention that the inseparablity of *wh*-phrases in Bulgarian hasn't gone unchallenged. Rudin (1994), Mišeska-Tomić (1996), Kim (1998), Stateva (1998), Bošković (forthcoming) have provided some compelling evidence against Rudin's original treatment. That is, the above-mentioned authors observe that *wh*-phrases can be separated by the arguably second-position clitic *li*.

(3) Koj li kakvo na kogo e dal? [BG]
who interr.cl. what to whom is given
'Who gave what to whom, I wonder?'

Lambova 1999 furthermore shows that even parentheticals and adverbs of various kinds can intervene:

(4) a. Koj sigurno kakvo e kupil? [BG]
who probably what aux.PRES.3.SG. bought

[1] By *cluster* we mean syntactic cluster. As discussed in Bošković forthcoming, Serbo-Croatian clitics must also form a group at PF. But this can be achieved without having the clitics form a group in the syntax.

'Who has probably bought what?'

b. Koj pravilno kakvo e rešil?
who correctly what aux.pres.3p.sg. solved
'Who has correctly solved what?'

c. Koj prŭv kogo e udaril?
who first whom is hit
'Who hit whom first?'

d. Koj, kazvaš, kakvo šte donese?
who say-pres.2p.sg. what will bring
'Who, you are saying, will bring what?'

However, there seems to be considerable disagreement among native speakers with respect to the data in (4). Following the majority of speakers, we will disregard the data in (4) as idiolectal, reflecting some special affinity with the Serbo-Croatian pattern.

The data in (3) cannot be so easily dismissed, and as such threatens the connection we are trying to establish since separability of the type observed in (3) is unavailable for clitics. Here the connection we just highlighted seems to breakdown.

The Bulgarian/Serbo-Croatian divide nonetheless remains (e.g., *wh*-phrases in Bulgarian all move successive cyclically, not obligatorily as in Serbo-Croatian (5)), which we believe indicates that the connection is a genuine one. But clearly any adequate theory must account for the connection breakdown in (3).

(5) a. Koj$_1$ kogo$_2$ misliš, če t_1 e celunal t_2 [BG]
who whom think-you that has kissed
'who do you think kissed who?'

b. *Koj$_1$ misliš, če kogo t_1 e celunal

c. Ko koga misliš da je poljubio [SC]
who whom think-you that has kissed

d. Ko misliš da je koga poljubio

Both the *Wh*/Clitic-connection, and the environments where it seems to break down will be the focus of this paper. In Section 2

we outline our assumptions concerning multiple *wh*-fronting. In Section 3 we do the same for clitic placement. In Section 4 we provide an account for the *wh*/clitic-connection. Section 5 is a summary.

Our purpose is to uncover the theoretical implications of such a connection. In particular, we will argue that the *Wh*/Clitic-connection favors the treatment of clitics defended in Uriagereka (1995a,b), where clitics are said to occupy a specific functional projection associated with such notions as theme and rheme. We will also argue that the *Wh*/Clitic-connection favors Chomsky's (1995) interpretation of Movement in terms of Attract/Suicidal Greed, and, de facto, Bošković's treatment of multiple (*wh*) fronting in terms of Attract-All (see also Brown 1999).

2. Multiple *Wh*-Fronting: Some Refinements

As far as Multiple *Wh* Fronting (henceforth, MWF) is concerned, we follow Bošković (1996, 1997a, 1998a, 1999, to appear a,b) (in turn based on Stjepanović 1995) in taking MWF to be an epiphenomenon, decomposable into more familiar *wh*-movement for one *wh*-phrase, and independently motivated focus movement for the remaining *wh*-items.

Despite his previous arguments to the contrary (see Bošković 1996, 1998a), Bošković (1998b, to appear a,b) provides some evidence that all types of fronting might be analyzable in terms of Attract (property/ies of the target(s)), and not necessarily by an Attract/Move combination.[2] Bošković proposes that *wh*-movement is subject to the Attract Closest principle, which in the realm of *wh*-movement captures Chomsky's (1973) Superiority Condition (see Bošković 1998a on that point). By contrast, focus movement has an Attract-All property (an extension of Chomsky's 1995 multiple spec hypothesis), which essentially amounts to saying that

[2] Brown (1999) extends Bošković's treatment of multiple *wh* fronting as Attract All-F to multiple neg-fronting ("Neg(ative)-concord").

the attracted elements can move in any random order, provided they all move. What we might call 'the Selective/Unselective Attraction hypothesis' provides an elegant account of the superiority/antisuperiority effects in Bulgarian first noted in Bošković (1997a). He observes that although Bulgarian requires the highest *wh*-phrase to move first, no such requirement is imposed on subsequent *wh*-frontings. As shown in (6a), when only two *wh*-phrases are used, superiority holds; *kogo* must be higher than *kak*, but when three *wh*-phrases are used, the very same *wh*-phrases, now in the second and third positions, are freely ordered ((6c-d)).

(6) a. Kogo kak e celunal Ivan? [BG]
whom how is kissed Ivan
'Whom did Ivan kiss how?'
b. *Kak kogo e celunal Ivan?'
c. Koj kogo kak e celunal?
who whom how is kissed
'Who kissed whom how?'
d. Koj kak kogo e celunal?

As Bošković (1998a, to appear a,b) shows, the facts in (6) follow if we assume that the first instance of *wh*-fronting is a case of *wh*-movement, subject to superiority/Attract Closest, whereas the subsequent frontings are cases of focus movement. To account for the impenetrability of the *wh*-sequence in this language, Bošković proposes that Bulgarian uses the same target for *wh*-, and focus movements, viz. CP.

This account works nicely for Bulgarian. For Serbo-Croatian, Superiority is 'latent,' detectable in various contexts including long-distance *wh*-movement, correlatives/embedded clauses, root clauses with an overt C^0 (given in (7)), while it is absent in root clauses with a null C^0 ((8)). Bošković capitalizes on Bošković (1998b, to appear c), and argues that CP is absent when there are no superiority effects, as in (8).

(7) *Long-distance questions*:

a. Ko si koga tvrdio da je istukao?
who are whom claim.2.SG that is beaten
'Who do you claim beat who?'

b. ?*Koga si ko tvrdio da je istukao?
whom are who claim.2SG that is beaten

Embedded contexts:[3]

c. Ko koga voli, taj o njemu i govori.
who whom loves, that-one about him even talks
'Everyone talks about the person they love.'

d. ?*Koga ko voli, taj o njemu/o njemu taj i govori.
Whom who loves, that-one about him even talks

e. Ima ko šta da ti proda.
has who what that you sells
'There is someone who can sell you something.'

f. *Ima šta ko da ti proda.
has what who that you sells

Root questions with overt C:

g. Ko li šta kupuje.
who C what buys
'Who on earth buys what?

h. *Šta li ko kupuje?
what C who buys

(8) *Short distance matrix questions with a null C*:

a. Ko koga voli?

---

[3] Bošković (1996, 1997a, 1998a) avoids giving indirect questions as examples of embedded questions because such questions involve an interfering factor. As Bošković notes, indirect questions formally do not differ at all from matrix questions in Serbo-Croatian. As a result, there is always a danger that they might be analyzed as matrix questions, with the superficial matrix clause treated as an adsentential. Instead, Bošković gives examples of correlative and existential constructions which, as shown by Izvorski (1996, 1998), also contain embedded questions. Bošković (1997a) does show that when this interfering factor in indirect questions is controlled for, true indirect questions in SC also exhibit Superiority effects.

who whom loves
'Who loves whom?'

b. Koga ko voli?
whom who loves
'Who loves whom?"

In cases like (8), the target of movement is some focus projection F. The latter has an attract-all property, which allows *wh*-phrases to move in random order. To account for the separability of *wh*-phrases in the language, more than one projection must be involved in the case of Serbo-Croatian (since moving to the same projection accounts for the Bulgarian pattern). Bošković argues that two AGR-projections can assume the role 'F' (see also Stjepanović 1995).

It is fair to say that Bošković's analysis in terms of selective/unselective attraction is elegant and consistent with the most recent tenets of the Minimalist Program (Chomsky 1995; 1998). However, it seems that Bošković's account might face some problem. By assuming that all *wh*-phrases target the same position in Bulgarian, he fails to predict the separability noted in (3). One might be tempted to project a distinct Focus projection immediately below CP, as argued for by Kim (1998), and Stateva (1998), but if we were to do so, we would lose the superiority facts.[4] Assuming that all *wh*-phrases must be focused, hence must at the very least pass through SpecFocP, nothing forces the highest *wh*-phrase to move first to SpecFocP. Therefore nothing ensures that the *wh*-phrase attracted by CP will be the highest one.

This wrong prediction might force us to take another tack, but we believe otherwise. The separability of *wh*-phrases is the crucial argument put forth by those who argue in favor of a projection distinct from, and located below CP to host some *wh*-forms (Kim 1998; Stateva 1998, among others). However, we saw that the

[4] Thanks to David Pesetsky (p.c.) for discussion of this issue.

separability data in (4) were not reliable.[5] What remains of the separability argument is the *li*-phenomenon in (3). Here, we would like to follow Bošković's suggestion (class notes 1998; Bošković forthcoming) that *li* is a focus affix in Bulgarian (unlike in Serbo-Croatian, where it is a complementizer), and as such does not mark the head of a phrase, but attaches to a word. In other words, due to the nature of *li* as a PF-affix, the *li*-placement data in (3) are not a conclusive test for the separability of *wh*-phrases in Bulgarian. The ultimate placement of *li* is a PF-phenomenon, not a structural one.

With the recalcitrant data in (4) analyzed away in terms of a PF-filter, we can assume Bošković's analysis of multiple *wh*-fronting in terms of Attract-All. In addition, and quite crucially, we need not assume any *Wh*/Clitic-connection breakdown.

Before turning to clitics, let us say a few words about the situation in Serbo-Croatian.

The issue to be addressed as far as Serbo-Croatian is concerned is the separability of *wh*-phrases. The solution put forward by Bošković is that more than one projection must be involved for focus in Serbo-Croatian. We will follow him in that regard.

Bošković and Stjepanović have argued in various works that at least two projections are needed to license *wh*-phrases and account for their separability.[6] Bošković (1998a) claims that the simplest solution is to argue that those two projections are AGRs and AGRo, and equate AGRP with FocP. Here, a parallel could be drawn with many studies which take FocP to be identical to IP (see

[5] If it were to turn out that the data in (4) are shared by a significant number of speakers (which we doubt from the judgments we got from Bulgarian speakers from various areas), we might try to accommodate them in terms of a recursive CP mechanism (see Watanabe 1993, and references therein), which would only be available in case of adverbs (and other adverb/adjunct-like elements) are part of the numeration, assuming with Cinque (1999) that adverbs need to be license in specific Adverbial Projections.)

[6] In work in progress, we examine the idea that the number of AGRPs/FocPs is not an accident, and they correspond rather closely to Chomsky's (1998) notion of phase (we reanalyze AGRs and AGRo in Serbo-Croatian as the edges of the CP and VP phases, respectively).

the papers in Kiss 1995). Assuming a Split-IP hypothesis with several AGRPs (Chomsky 1991), it follows that FocP will also be split into as many AGR projections as are needed. We therefore see the separability of *wh*-phrases in Serbo-Croatian as a consequence of the split nature of INFL (more precisely, the presence of several AGR projections, each hosting a *wh*-phrase).[7]

We see the difference between Serbo-Croatian (multiple FocPs) and Bulgarian (one CP/FocP) as the result of a macroparameter distinguishing the two languages. The macroparameter we have in mind is Discourse Configurationality. As is well-known, Serbo-Croatian, but not Bulgarian, has Scrambling, a hallmark of discourse-configurationality. The multiplicity of FocPs in Serbo-Croatian is another way of exhibiting this discourse-configurational character. Having established how *wh*-phrases are licensed in Bulgarian and Serbo-Croatian, we can now turn to clitics, where we will show a situation similar to the one found for *wh*-phrases is found.

3. Clitic Placement

One difference between Serbo-Croatian and Bulgarian clitics is that in Serbo-Croatian clitics do not form a cluster in overt syntax, while in Bulgarian, they do. As shown by the ellipsis tests by Stjepanović (1998) in (9a), the VP fronting tests by Wilder and Ćavar (1997), in (9b), and the possibility of parenthetical insertion between the clitics by Bošković forthcoming, in (9c), Serbo-Croatian clitics can be split in overt syntax.

---

[7] The Split Agr hypothesis does not prevent all *wh*-phrases from forming a cluster in overt syntax, as in (i):

(i) Šta kome Marija daje?
what whom Marija gives
'What is Marija giving to whom?'

In this case we assume that the topmost AGRP in Serbo-Croatian acts as an Attract-All Projection, i.e., in this case, the topmost AGRP should be seen as equivalent to Richards's (1997) Agr∀P.

(9) a. Mi *smo mu ih* predstavili,a i vi
we are him.dat them.acc introduced and also you
ste ~~*mu ih*.predstavili~~ (takodje)
are him.dat them.acc, introduced too
'We introduced them to him, and you did, too.'

b. Dali ga Mariji su Ivan i Stipe
given it.acc Marija.dat are Ivan and Stipe
'Give it to Marija, Ivan and Stipe did.'

c. ?Oni su, kao što sam vam rekla, predstavili se Petru.
they are, as am you told introduced self Petar
'They, as I told you, introduced themselves to Petar'

Bulgarian, however, fails all of these tests, as shown by Bošković (forthcoming), which means that clitics in Bulgarian cannot be split in overt syntax:

(10) a. ?* Te sa ja celunali, i nie sŭšto sme.
they are her kissed and we too are
'They kissed her, and we did too.'

b. *Celunala go Maria e.
kissed him Maria is
'Maria kissed him.'

c. * Te sa, kakto ti kazax, predstavili go na Peter.
they are, as you I-told introduced him to Peter
'They, as I told you, introduced him to Peter.'

As far as Serbo-Croatian auxiliary clitics are concerned, Stjepanović (1997) and Bošković (forthcoming) argue that they occur in AgrS in overt syntax, based on the following facts:

(11) Oni su navodno istukli Petra.
they are allegedly beaten Petar
'They allegedly beat Peter.'

In (11), the auxiliary clitic *su* 'are' precedes the sentential adverb *navodno* 'allegedly'. As shown by Watanabe (1993) and Bošković (1995, 1997b), sentential adverbs are adjoined to TP. This means that the clitic auxiliary in (11) precedes TP. Stjepanović (1997) and Bošković (1995, 1997b) take the auxiliary to be in AgrS, where it moves in overt syntax.

As far as pronominal clitics are concerned, based on the ellipsis facts in (12), Stjepanović concludes that they are base-generated as arguments of the verb within its VP, and move overtly out of VP to AgrIO and AgrDO.

(12) Oni su mu ga predstavili, a vi mu niste.
they are him.DAT him.ACC introduced and you him.DAT aren't
'They introduced him to him, but you didn't.'

Bošković (forthcoming) also comes to a similar conclusion with respect to pronominal clitics based on the following data with adverb placement:

(13) Oni su joj pravilno odgovorili.
they are her correctly answered
'*They did the right thing in answering her.'
'They gave Milena a correct answer.'

In (13), the pronominal clitic *joj* precedes the adverb *pravilno* 'correctly'. This adverb is normally ambiguous between a sentential and manner reading. Note, however, that in (13), where a pronominal clitic precedes it, the adverb cannot have a sentential reading. It only has a manner reading. Now, as mentioned above, on the sentential reading, the adverb is adjoined to TP. Manner adverbs, however, are standardly assumed to be adjoined to VP. Based on this, Bošković concludes that pronominal clitics in Serbo-Croatian occur in a functional projection lower than TP, but higher than VP, i.e. in AgrOPs.

So, Serbo-Croatian clitics are located in AgrPs in overt syntax. However, we have seen that *wh*-phrases in Serbo-Croatian also occupy AgrPs in overt syntax, and with respect to *wh*-phrases, we have equated AgrPs with FocPs in Serbo-Croatian. However, the fact that clitics also occupy AgrPs, at first sight argues against equating FocP with AgrP. Clitics being topics, i.e., the opposite of focused phrases, would not be expected to occupy the same position as focussed phrases, in this case, *wh*-phrases. However, in Section 4, we will show that the problem is only apparent.

As far as Bulgarian clitics are concerned, we have seen in (9) that they must form a cluster in overt syntax, i.e. they cannot be split. In Section 4, we will show that there is a similar parallelism between Bulgarian clitics and *wh*-phrases with respect to the position they occupy in overt syntax, as there is in Serbo-Croatian.

4. Accounting for the *Wh*/Clitic-Connection

Having spelled-out our assumptions about clitics and *wh*-phrases, and their respective licensing mechanisms, we are now ready to provide an explanation for the *wh* /clitic-connection.

Given what has been said in the previous sections, the connection should already be obvious at this point. Clitics in Serbo-Croatian are licensed in AGRPs. So are *wh*-phrases. It might seem odd to say that clitics (old-information carrying elements),[8] and *wh*-phrases (new-information carrying elements) are licensed by the same projection. But the oddity disappears once we say that AGRP in Serbo-Croatian (and by hypothesis, in all discourse-configurational languages (see Kidwai 1999)) is a Discourse-related phrase, hosting both foci and topics irrespective of their conflicting nature.[9]

---

[8] We abstract away here from clitics in clitic-doubling constructions in Bulgarian, which, as Catherine Rudin reminded us (personal communication), might not always convey old information. The point in the text remains unaffected.

[9] Note that sharing licensing positions might shed light on why there seems to be a contiguity/adjacency relation between *wh*-phrases and clitics in the language (as noted in

Interestingly, the idea that clitics are licensed in a Discourse-related phrase has already been put forward by Uriagereka (1995a,b). Uriagereka argues that clitics in Western Romance are hosted by a specific functional projection called "FP." The term 'FP' is rather unfortunate, because FP is often used in the literature to identify 'some' projection which seems to be needed, but for which we do not have any name yet. Uriagereka's FP is very different:[10] It is a specific functional projection, related to concepts like theme and rheme (discourse properties in general). It is important to bear this fact in mind when we relate our proposal to Uriagereka's.

To the extent that our analysis of the *Wh*/clitic-connection is correct, it lends credence to Uriagereka's theory of clitic-licensing, as well as to Bošković's theory of multiple *wh* fronting as Attract All.

Our analysis in terms of Discourse configurationality has a further non-trivial theoretical implication. As already said above, it is expected that clitics and *wh*-phrases pattern together if one concentrates on their intrinsic properties (old and new information, respectively). The *Wh*/Clitic-connection would therefore remain a mystery in those theories that assume that elements move to satisfy their own needs (the original notion of Greed defended in Chomsky 1993). If elements were moving to satisfy their own needs, we would expect clitics and *wh*-phrases to target different landing sites;[11] and so, no *Wh*/clitic-connection (or, at least, no connection of the type we have argued for here, that is, in terms of common landing site, which seems to us to have received considerable support in the previous sections).

With the notion Attract/Suicidal Greed, it becomes possible to establish a landing-site-based connection between clitics and *wh*-

---

various papers in Halpern and Zwicky 1996).

[10]The term 'FP' was introduced in Uriagereka (1988), with the idea of 'Further' (Projection), at the time when Pollock was advocating the existence of AGRsP.

[11]Or, at least, we would not necessarily expect them to move to a common projection.

phrases. All we have to do is assume the existence of a Discourse-related phrase capable of hosting elements with conflicting needs.

What this amounts to is that our analysis seems to lend support to the Attract-based theory of movement. To recap, clitics and *wh*-phrases are licensed in the various Discourse-related projections available in Serbo-Croatian.

In Bulgarian, clitics occupy IP, forming a cluster. *Wh*-phrases occupy some projection related to focus. For reasons of uniformity and symmetry, we would like to analyze Bulgarian IP as having some Discourse-related feature, and equate IP and FocP (cf. Zubizarreta's 1998 notion of 'syncretic category'). That IP is focus-related in Bulgarian has been independently argued for by Rudin (1986), and for Romanian (a language that patterns like Bulgarian in all relevant respects, as first proposed by Rudin 1988) by Montapanyane (1998). So unlike Serbo-Croatian IP, Bulgarian IP would be "unsplit," another reflex of Bulgarian not being (strongly) Discourse-related (strong Discourse-relatedness would entail multiplicity of Discourse-related projections). The Split/Unsplit-IP parameter has been argued on independent grounds (having to do not with discourse configurationality, but morphology) by Bobaljik and Thráinsson (1998); so we take it to be an option of Universal Grammar. Much like Bobaljik and Thráinsson argues that their Split-Unsplit Infl parameter cuts across morphology, we would like to argue that the Split-Unsplit-Infl cuts across discourse-configurationality.

As already noted above, the number of Discourse-related phrases differentiating Serbo-Croatian and Bulgarian is the manifestation of a macroparameter. We consider it an advantage of our analysis that we have been able to reduce the *Wh*/Clitic-connection to a larger manifestation, in accordance with the Principles-and-Parameters approach to cross-linguistic variation.

5. Conclusion

This paper has tried to provide an explanation for the puzzling

symmetry of behavior of clitics and *wh*-words in Bulgarian and Serbo-Croatian. Armed with adequate licensing mechanisms for both clitics and *wh*-phrases, we accounted for the *wh*/clitic connection in terms of their occupying the same Discourse-related projections, which are split in Serbo-Croatian, but not in Bulgarian, hence their separability in the former, but not in the latter. We suggested that the discourse-configurational character of Serbo-Croatian might be the determining factor for the multiplicity of Discourse-related positions.[12]

It goes without saying that we have not tried to deny all the (too obvious) differences between clitics and *wh*-phrases by establishing a *Wh*/Clitic-connection. What we have tried to do is relate those elements with conflicting demands in terms of licensing, which allowed us to reach interesting theoretical conclusions like the licensing of clitics in FP (Uriagereka 1995a,b), the superiority of the Suicidal Greed/Attract conception of movement, and the correctness of the macroparametric view of cross-linguistic variation. The idea behind the *Wh*/Clitic-connection is that there are highly abstract properties uniting disparate elements, differences among which (locality, say) result from independent factors (e.g., XP- vs. X^0-movement).

References

Bobaljik, J., and H. Thráinsson. 1998. Two Heads aren't always better than one. *Syntax* 1, 37-71.

Bošković, Ž. 1995. Principles of Economy in non-finite complementation. PhD dissertation, University of Connecticut, Storrs.

Bošković, Ž. 1996. "Fronting *wh*-phrases in Serbo-Croatian." In Franks, S. and M. Lindseth (eds.), *Formal Approaches to Slavic Linguistics 1996. The Indiana Meeting*, , 86-107. Ann

[12] For extensive discussion of the discourse-configurational character of Serbo-Croatian, see Stjepanović (1999).

Arbor, Mich.: Slavic Publications.
Bošković, Ž. 1997a. Superiority effects with multiple *wh*-fronting in Serbo-Croatian. *Lingua* 102, 1-20.
Bošković, Ž. 1997b. *The syntax of non-finite complementation: an economy approach*. Cambridge, Mass.: MIT Press.
Bošković, Ž. 1998a. "Multiple WH-fronting and economy of derivation." In Curtis, E., J. Lyle, and G. Webster (eds.), *Proceedings of the 16th West Coast Conference on Formal Linguistics*, 49-63. Stanford, CA: CSLI.
Bošković, Ž. 1998b. "LF-movement and the minimalist program." In Tamanji, P. and K. Kusumoto, *Proceedings of NELS 28*, 43-57. GLSA, University of Massachusetts, Amherst.
Bošković, Ž. 1998c. On the interpretation of multiple questions. Ms., University of Connecticut, Storrs. [Available at *http://mitpress.mit.edu/celebration*]
Bošković, Ž. 1999. What is special about multiple *wh*-fronting. Ms., University of Connecticut, Storrs.
Bošković, Ž. To appear a. *Wh*-movement and *wh*-phrases in Slavic. Paper presented at the Comparative Slavic Morphosyntax Workshop, Spencer, Ind. To appear in Proceedings.
Bošković, Ž. To appear b. "On multiple feature checking: Multiple *wh*-fronting and multiple head-movement." In Epstein, S. D., and N. Hornstein (eds.), *Working minimalism*. Cambridge, Mass.: MIT Press.
Bošković, Ž. To appear c. "Sometimes in situ, sometimes in SpecCP." In Martin, R., D. Michaels, and J. Uriagereka (eds.), *Step by step: Essays on minimalism in honor of Howard Lasnik*. Cambridge, Mass.: MIT Press.
Bošković, Ž. Forthcoming. The syntax-phonology interface: A view from the syntax. Ms., University of Connecticut, Storrs.
Brown, S. 1999. Negative concord and Attract-All. Ms., Harvard University.
Chomsky, N. 1973. "Conditions on transformations." In Anderson, S., and P. Kiparsky (eds.), *A festshrift for Morris Halle*, 232-

286. New-York: Holt, Rinehart, and Winston.

Chomsky, N. 1991. "Some notes on economy of derivation and representation." In Freidin, R. (ed.), *Principles and parameters in comparative grammar*, 417-454. Cambridge, Mass.: MIT Press. [Reprinted in *The minimalist program*, N. Chomsky, 129-166. Cambridge, Mass.: MIT Press. 1995.]

Chomsky, N. 1995. "Categories and transformations." In *The minimalist program*, N. Chomsky, 219-394, Cambridge, Mass.: MIT Press.

Chomsky, N. 1998. "Minimalist inquiries: The framework." MIT Occasional Papers in Linguistics #15. Distributed by MITWPL, MIT, Cambridge, Mass. [To appear in Martin, R., D. Michaels, and J. Uriagereka (eds.), *Step by step: Essays on minimalism in honor of Howard Lasnik*. Cambridge, Mass.: MIT Press.

Cinque, G. 1999. *Adverbs and functional heads: a cross-linguistic perspective*. Oxford: Oxford University Press.

Halpern, A., and A. Zwicky, eds. 1996. *Approaching second*. Stanford, CA: CLSI.

Horvath, J. 1986. *Focus in the theory of grammar and the syntax of Hungarian*. Dordrecht: Foris.

Izvorski, R. 1996. The syntax and semantics of correlative proforms. In Kusumoto, K. (ed.), *Proceedings of NELS 26*, 133-147. GLSA, University of Masssachusetts, Amherst.

Izvorski, R. 1998. Non-indicative wh-complements of possessive and existential constructions. In Tamanji, P., and K. Kusumoto, 159-173, *Proceedings of NELS 28*. GLSA, University of Masssachusetts, Amherst.

Kidwai, A. 1999. The topic interpretation in Universal Grammar. Ms., Jawaharlal Nehru University.

Kim, J.-S. 1998. Superiority effects in multiple *wh*-fronting. Paper presented at the Comparative Slavic Morphosyntax Workshop, Spencer, Ind.

Kiss, K. E, ed. 1995. *Discourse-configurational languages*. Oxford: Oxford University Press.

Lambova, M. 1999. On Multiple *Wh*-fronting languages in Slavic: New evidence from Bulgarian. Ms., University of Connecticut, Storrs.

Mišeska-Tomić, O. 1996. The Balkan Slavic clausal clitics. *Natural Languages and Linguistic Theory* 14, 811-872.

Montapanyane, V. 1998. Focus, checking theory, and fronting strategies in Romanian. *Studia Linguistica* 52, 227-243.

Richards, N. 1997. What moves where when in which language. PhD. Dissertation, MIT.

Rizzi, L. 1997. The fine structure of the left periphery. In Haegeman, L. (ed.), *Elements of grammar*, 281-337. Dordrecht: Kluwer.

Rudin, C. 1986. *Aspects of Bulgarian syntax: Complementizers and* wh-*constructions*. Columbus: Slavica.

Rudin, C. 1988. On multiple questions and multiple *wh*-fronting. *Natural Language and Linguistic Theory* 6, 445-501.

Rudin, C. 1994. On focus position and focus marking in Bulgarian questions. In Davison, A. et al. (eds.), *Papers from the 4th Annual Meeting of Formal Linguistics of MidAmerica*, 252-265.

Stateva, P. 1998. On focus features and movement in Bulgarian. Paper presented at ConSOLE 7, Bergen.

Stjepanović, S. 1995. Short distance movement of *wh*-phrases in Serbo-Croatian matrix clauses. Ms., University of Connecticut, Storrs. [Paper presented at the Comparative Slavic Morphosyntax Workshop, Spencer, Ind., 1998.]

Stjepanović, S. 1997. VP ellipsis in a verb raising language: Implications for verbal morphology. Paper presented at Escol 97. [To appear in *Proceedings of Escol 97*, Cornell University.]

Stjepanović, S. 1998. On the placement of Serbo-Croatian clitics: evidence from VP-ellipsis. *Linguistic Inquiry* 29: 527-537.

Stjepanović, S. 1999. What do second position cliticization, scrambling and multiple *wh*-fronting have in common?, PhD dissertation, University of Connecticut, Storrs.

Uriagereka, J. 1988. On government. PhD. dissertation, University

of Connecticut.
Uriagereka, J. 1995a. Aspects of the syntax of clitic placement in Western Romance. *Linguistic Inquiry* 26, 79-123.
Uriagereka, J. 1995b. An F-position in Western Romance. In Kiss, K. E. (ed.), *Discourse-configurational languages*, 153-175. Oxford: Oxford University Press.
Watanabe, A. 1993. AGR-based Case theory and its interaction with the A-bar system. PhD. Dissertation, MIT.
Wilder, C., and D. Ćavar. 1997. Verb movement, cliticization, and coordination. Ms., ZAS Berlin, and University of Hamburg.
Zubizarreta, M.-L. 1998. *Focus, prosody, and word order*. Cambridge, Mass.: MIT Press.

Department of Linguistics
University of Connecticut, U-1145
341 Mansfield Rd
Storrs, CT 06269-1145 (USA)
ceb99001@sp.uconn.edu (Boeckx)
sas94006@sp.uconn.edu (Stjepanović)

Paths and Boundaries in Croatian Prefixed Motion Verbs: Establishing Componential Relations between Form and Meaning*

Maya Brala
University of Cambridge, UK

1. Introduction

In 1957, Chomsky observed that we

> find many important correlations ... between syntactic structure and meaning; or, to put it differently, find that the grammatical devices are used quite systematically. These correlations could form part of the subject matter for a more general theory of language concerned with syntax and semantics and their points of connection.
>
> (Chomsky, 1957:108)

The idea that elements of meaning could be systematically related to elements of linguistic form, and that, furthermore, systematic relations observed crosslinguistically could lead to a set of universal and, perhaps, primitive linguistic elements underlying and 'guiding' language, was not new. It had been around since Greek times, but it was not until Chomsky's influential interest in syntactic universals, that the idea of linguistic universals became widely recognised and thoroughly investigated. Chomsky's studies in formal syntactic properties of language were soon to be followed by an interest in semantic universals. Research into semantic universals was largely underpinned by studies such as Berlin and Kay's (1969) work on colour terminology, the Piagetian revolution, and the extensive Standford Project on Language Universals.

More recently, the interest in semantic universals has revolved around the work by Beth Levin (1993), Anna Wierzbicka (1996),

* I would like to thank Dr. Keith Brown, Dr. Henriette Hendriks and two anonymous reviewers for helpful comments on earlier drafts of this paper.

Ray Jackendoff (1996), and Leonard Talmy (1983, 1985). As in this paper I shall be exploring ways in which motion is lexicalised in Croatian, and as motion events are the primary concern of Talmy's work, let us start by briefly reviewing some of his main findings.

2. Talmy's Conflation Patterns

In his analysis of systematic relations in language between meaning and surface expression of language, Talmy (1975, 1985, 1988) has analyzed and systematized semantic-to-surface associations in a large number of languages, singling out, on the one hand, typological patterns that are language-specific, and isolating from these patterns, on the other hand, the components or semantic categories recurring crosslinguistically.

The components Talmy sees as constituents of a basic motion event include: (a) *Figure* (object moving or located with respect to another object), (b) *Ground* (the reference object), (c) *Motion* (the presence per se of motion in the event), (d) *Path* (the course followed by the Figure with respect to the Ground), and (d) other (e.g., *Manner* and *Cause*). The ways in which these components are lexicalised or combined in words vary from language to language. It is important to note here that the main issue underlying Talmy's approach is the quest for semantic components and typical combinatorial patterns, or, put in a broader context, the question of how language organises conceptual content.

On the basis of the most frequently used combinatorial patterns, Talmy concludes that languages can be categorized into several groups. On the one hand, there are the languages, including Chinese, and apparently all branches of Indo-Europen except (post-Latin) Romance, that conflate Motion with Manner or Cause, and lexicalise the other components separately. An illustrative example is given by the sentence: 'John ran across the bridge' where in fact, the verb conflates Motion and Manner.

Semitic, Polynesian, and Romance languages do, on the other hand, typically conflate Motion and Path in the verb root, expressing Manner/Cause separately. A good example showing this is given by the Spanish sentence:

(1) La botella entro` a la cueva (flotando).
the bottle moved-in to the cave (floating)
Lit. 'The bottle floated into the cave'.

We see that Spanish tends to conflate Motion and Path in the verb *enter*, while lexicalising Manner separately[1].

It needs to be stressed out at this point that what has been said so far is not a rule but rather a regularity. In fact, English *can* in certain verbs that genuinely incorporate Path, such as *rise*, *descend*, *circle*, or *cross*, realise the Spanish-type pattern, just as Spanish or Italian have lexical means for expressing the English-type pattern. However, what Talmy is considering is the *frequency of use*, rather than the (im)possibility of realising certain patterns.

The remainder of this article is divided into three main parts. In Section 3, Talmy's findings relative to motion verbs are probed on the Croatian language, the main problems are set out, and theoretical tools for their possible solution are provided. Then, in Section 4, which is the focal section of the paper, selectional properties of Croatian prefixed motion verbs are examined. The aim of this section is to establish links between surface form and underlying meaning. Finally, in Section 5 problems and notions introduced in previous sections are brought together, and further evidence supporting Talmy's conflation patterns is provided.

3. Motion Events in Croatian

According to Talmy, the conflation pattern for Croatian should be the same as that of English, i.e.,, the one combining Motion and Manner in the verb, and expressing Path separately. The English sentence 'John ran across the bridge' translates into Croatian as:

---

[1] In contrast, and as predicted by the pattern, English tends to conflate Motion and Manner in the verb *to float*.

(2) John je pretrcao most.
John be.AUX over-run.PAST bridge ACC
'John ran across the bridge.'

The verb *pretrcati* is a prefixed form of the verb *trcati* 'to run' to which the prefix *pre-* is added. Three things are worth observing at this point: the first one is that the prefix *pre-* has a cognate preposition *preko* which means 'across', 'over', which is frequently used for expressing Path. The second one is that prefixation is the most productive process in Croatian word formation; in fact, over half of all Croatian verbs are of the form prefix+root verb (cf. Babic 1986:477). Interestingly, rather than be seen as conflating only Motion and Manner, prefixed verbs of motion of this type are frequently said to conflate Path[2] as well. Thirdly, the verb *pretrcao* in (2) is a perfective verb, which raises the question of Aspect.

Let us try and gain deeper insight into these issues by noting that the sentence 'John ran across the bridge' is a Past event, and that it is bounded[3], i.e., bound to a result (in this case, crossing the bridge). If our sentence were in the Present Tense, *John runs/is running*, we could have rendered it in three different ways. In (2'a, b), we only have the progressive reading, *John is running across the bridge*. Example (2'c) can also have the progressive reading of (2b).

(2') a. John trci preko mosta.
John run.PRES.3SG over bridge.ACC
b. John pretrcava most.
John *over-run.PRES.3SG bridge ACC
c. John pretrcava preko mosta.
John *over-run.PRES.3SG over bridge.GEN
'John runs/is running across the bridge.'

We observe immediately that in (2a), the pattern, posited by Talmy as typical for the Indo-European languages (except the Romance

[2] Path is here, supposedly, expressed by the prefix *pre-*.
[3] Bounded events are also referred to as conclusive or telic events.

family) holds. In fact, the verb *trci* 'runs' conflates Motion and Manner, and the Path is lexicalised by the preposition *preko* 'across'. However, in examples (2b) and (2c), we encounter the verb *pretrcava*, which is the Present form of the Infinitive *pretrcavati* 'to be running across'. *Pretrcavati* is formed by adding to the root verb *trcati* 'to run' a prefix *pre-*[4], and an infix *-va-*, which lexicalises progressiveness and iterativity. Furthermore, if we compare (2b) and (2c), we note that while in (2b) the prefixed verb is followed by a direct object, in (2c), it is followed by a PP, and that, even more interestingly, in (2c) there is repetition of form between prefix and preposition. The (possible) differences in meaning between (2b) and (2c) are quite opaque, much like the rule governing selectional properties of prefixed verbs.

Apart from opening up the possibility that there might be components, other than Manner and Motion, that are lexicalised by Croatian prefixed verbs of motion, the observations spelled out above thus raise the issue of prefixal semantics. Furthermore, if we accept the idea that prefixed verbs of motion incorporate the Path, Talmy's prediction for Croatian, which sees it as expressing Path separately, would not hold. So, how could we try systematising the conflation patterns of Croatian motion verbs and relate them to Talmy's framework, once we include prefixed verbs of motion as well? We will not be able to do so unless we take a closer look at two further theoretical notions, the first one being Aspect.

3.1. The Notion of Aspect

Defining Slavic Aspect by means of the English metalanguage of syntax is an extremely difficult task. The pitfalls are numerous, the most treacherous being the fact that Aspect in its original diathesis, i.e., perfect vs. imperfect[5], is used today in English in collocation with Tense. In contrast, Slavic Aspect is a verbal category distinct from Tense.

---

[4] As we have seen, there is also the cognate preposition *preko*, which means 'over', 'across'; in (2'c) we see it expressing Path.

[5] Respectively, complete vs. non-complete

In Croatian grammar books Aspect is defined as follows:

> Verbs that express an action in process are called verbs of imperfective aspect... Verbs that express full completion of an action are called verbs of perfective aspect.
>
> (Baric et al. 1995:225; my transl.).

It is, hence, the *meaning* of the verb that tells us if it is a perfective or an imperfective verb. However, the above definition of Aspect is insufficient, and, in our case, potentially misleading. Let us see why.

To begin with, we might want to note that a definition of Aspect, while being necessarily based on semantic criteria, is not exhausted by the completion vs. non-completion diathesis. As Klein (1995:675-677) points out, this distinction, apart from being too metaphorical, fails on at least three substantial grounds. First of all, there are many common usages of Imperfective Aspect in which the situation is clearly completed.[6] Secondly, it is unclear whether 'completion' is meant in the target sense (does the action have a goal? Has the goal been reached?), or in a simpler, temporal sense (is the action simply over?). Finally, the notion of 'completion' only makes sense if, i.e., when, related to a time T. Failing to include the relation between the action and a point in time T, leaves the entire characterisation of Aspect 'hanging in the air' (ibid.:676).

Klein's insightful proposal for coping with the issue of Aspect is based on temporal grounds. Klein (1994: *passim.*) introduces three frames of reference: the *'Time of the Utterance'* to refer to the time at which the utterance is made (TU), the *'Time of the Situation'* to refer to the time at which the situation expressed by the predicate obtains (TSit), and, finally, the *'Topic Time'* to refer to that subinterval of TSit for which the assertion is made (TT). Aspect is seen as a relation between TT and TSit.

Since prefixation of an imperfective verb *always* transforms it into a perfective verb (cf. Baric et al. 1995:379; Babic 1986:478)

---

[6] In Croatian, the use of the Present of an Imperfective to refer to completion of an action in the future, of a generalisation, or the 'fact constatation' use of the Imperfective (cf. Klein 1995:693).

we could, at this point, be easily led to conclude that the meaning component 'completion' is the first element to be associated with Croatian prefixes,[7] and add that prefixes lexicalise Aspect.

Going back to Talmy's framework, we note that prefixes - falling within the category of verb-constituents termed 'satellites' - have, indeed, in Talmy's analysis been identified as expressing, among other components, also Aspect. However, in the remaining part of this paper I shall try and argue that a) prefixes do not lexicalise Path, and b) prefixes do not lexicalise Aspect. In order to do so, a further notion is needed: that of *Aktionsart*.

3.2. The Notion of Aktionsart

Aktionsart (German for 'form of action') is a view of predicates in terms of their inherent temporal properties. The most common classification of verbs on an Aktionsart basis is the original one proposed by Vendler (1957 [1967]), in which distinction is made between: states, achievements, accomplishments and activities. Klein's (1994) proposal sees Aktionsart 'rewritten' and yielding three possibilities: a) *∅-state contents*, b) *1-state contents*, and c) *2-state contents*.

∅-state contents are seen as atemporal, they do not have boundaries, and obtain without temporal limits. A good example of a ∅-state content is provided by the verb *to be* in *Two plus two is four*. 1-state contents are expressed by verbs for which it can be assumed that, if there is a time T at which they are true for an argument, then there also must be a 'contrasting time T' at which they are not true. Good examples are provided by verbs such as *to sleep*, *to be hungry*, and *to work*. Finally, and most interestingly for the purposes of our analysis, there are the 2-state contents. These are expressed by verbs whose content relates to a time span TT within which there is a *change of state*, i.e., within TT a situation first obtains, and then, within the same TT, it does not obtain. 2-state verbs include e.g., *to arrive*, *to wake up*, and *to die*.

---

[7] Thus Path being not the *only* semantic component that a prefix 'adds' to the root verb expressing motion.

Now, since prefixation of a verb automatically transforms it into a perfective, i.e., since the 'default' semantic value of the prefixed infinitive is perfective, it is easy to see why the semantic role of a prefix would automatically be interpreted as Aspect-marking. However, we need to note here that prefixed perfectives can, through infixation, be imperfectivized.

Another important fact which ought to be noted at this point is that prefixation of a verb automatically brings about a 2-state content. Furthermore, we can have a prefixed verb which is not perfective, but we cannot have a prefixed verb which is not a 2-state content. Given this latter, crucial remark, and recalling Klein's (1995) view of Aspect as a primarily temporal relation, I argue that *prefixes are to be interpreted as Aktionsart markers, rather than Aspect markers*. This is made even more clear when we observe that the distinction between 1-state and 2-state contents is not boundedness (in that both describe situations that involve boundaries), but rather if the lexical content involves a phase and its opposite. The claim here then is that the semantic role of the prefix is to introduce of a 'target state' within TT.

Now, having introduced the issue of prefixal semantics, we can move on to examine the seemingly 'unpredictable' selectional properties of the Croatian prefixed verbs of motion.

4. Analysis of the Syntactic Behaviour of the Croatian Prefixed Motion Verbs

In the first part of this paper we have viewed some general issues related to (Croatian) prefixed motion verbs. Let us now turn to a more opaque, and, in my view, quite interesting problem, namely the fact that there seems to be no rule in reference to their subcategorization. A thorough investigation of the Croatian prefixed verbs of motion reveals that (a) some of these prefixed verbs obligatorily select a Prepositional Phrase (PP), (b) others cannot be followed by a preposition but always select a direct or indirect object (usually in the Genitive, Dative or Accusative case); and (c) a third group can optionally take either one or the other.

This fact has defied any satisfactory description, analysis, classification, or explanation; no grammar book offers any type of account, and the representation in dictionaries is completely arbitrary and highly unsystematic. As such, this issue represents a large problem for foreign learners, but also a field in which to seek potential evidence underpinning hypotheses regarding the semantico-syntactic interface, i.e., gain further insight into lexicalisaton patterns of motion verbs.

A closer look at these three groups of verbs is hence needed; the task is to search for any semantic features that motion verbs within each of the three categories might share, with the aim of finding an explanation for their shared syntactic behaviour. The best starting point for such an investigation seems to be the category of prefixed verbs exhibiting the optionality pattern (cf. sentences (2b) and (2c) above). In fact, if, by examining the optionality category, we manage to pinpoint a semantic difference between the prefixed verb+direct object construction and the prefixed verb+PP construction, we could then move on from there and try to extend our findings to the two categories which obligatorily select one of the two possible constructions.

In order to investigate native speakers' intuitions with respect to semantic nuances which (may) differentiate sentences where optionality is allowed, a questionnaire[8] was set up and administered to 30 fourth-year university students (13 male and 17 female), all Croatian native speakers. They were asked to give their acceptability judgements on 10 pairs of sentences containing prefixed verbs exhibiting the optionality pattern. The respondents were also asked to indicate whenever they perceived the two sentences forming a pair were somewhat different in meaning, and

[8] The respondents were presented with pairs of sentences (cf. Section 4.1.) in which the optionality pattern is allowed, and asked to judge them as '—', '+', or '/'. In the latter case, the respondents were also asked to make the perceived semantic differences explicit.

— 'grammatically unacceptable' (for one sentence of the pair)

\+ 'both grammatically acceptable and meaning the same'

/ 'both sentences forming the pair are grammatically acceptable, but there is a difference between sentences a) and b)'.

to describe the two meanings. A summary of the findings follows in Section 4.1.

4.1. Prefixed Verbs of Motion which Allow Optionality (Prefixed Verb + PP *or* Prefixed Verb + Direct Object)

Let us begin by considering the following pairs of sentences:

(3) a. Avion je preletio polje.
plane be.AUX *overflew field.ACC
b. Avion je preletio preko polja.
plane be.AUX *overflew over field.GEN
'The plane flew over the field.'

(4) a. Preskocio je ogradu.
*overjumped be.AUX fence.ACC
b. Preskocio je preko ograde.
*overjumped be.AUX over fence.GEN
'He jumped over the fence.'

(5) a. Avion je preletio planinski vrh.
airplane be.AUX *overflew mountain top.ACC
b. Avion je preletio preko planinskog vrha.
airplane be.AUX *overflew over mountain top.GEN
'The airplane flew over the mountain top.'

(6) a. Obletjeti kucu.
*roundfly house.ACC
b. Obletjeti oko kuce
*roundfly around house.GEN
'Fly around the house.'

(7) a. Carl Lewis je pretrcao stazu.
Carl Lewis be.AUX *crossran track.ACC
b. Carl Lewis je pretrcao preko staze.
Carl Lewis be.AUX *crossran across track.GEN
'Carl Lewis ran (across) the track.'

The pairs of sentences (3)-(4) are all taken from the questionnaire that has been used for assessing native speaker's perception of (possible) semantic differences between sentences a) and b) in each of the pairs. The first interesting result that came up is that, except for sentence pair (6), most speakers (97%) found the sentence under a) (the one in which the verb is followed by a direct object) to be the 'more natural' one. Several of them expanded on this by saying that it had more of a 'completion sense' or something alike. Sentences under b) were generally (86%) rated as 'acceptable' or 'grammatically correct', but 'somewhat redundant'.

Furthermore, the general (83%) remark for sentences under a) was that whenever the prefixed motion verb is followed by a direct object, the objective is 'more easily achieved'. Conversely, sentences under b) were seen as being more focused on the physical aspects of the Ground. One respondent, very interestingly, pointed out that in sentence pair (4,) the fence is higher, and more difficult to *overjump in b) than in a), whereas another respondent noted that in sentence pair (5), the plane flies much higher in a), whereas in b) it almost 'scratches' the mountain top.

For sentences under a), there also seems to be a shared tendency among the respondents (53%) to continue talking about motion 'beyond what is expressed in the sentence'. This can be easily explained for the sentence pair (3), where in (3a) the field is seen more as just 'one of the things which were *overflown while the plane was flying', whereas in (3b), there seems to be some implication where the flying of the plane could (not necessarily, but could) be confined to the field only (where the plane flies from one side of the field to the other, and then lands).

Finally, in native speakers' responses, directionality seems to be another important element at play. All of the respondents (100%) pointed out that in sentence pair (7), the sentence (7a) *Carl Lewis **crossran the track* is to be interpreted as in 'Carl Lewis is running in a race, and has run the whole track (100 meters, or 400 meters etc.)', whereas in (7b), *Carl Lewis **crossran across the track*, the athlete is seen as running from one side of the track to the opposite side. However, I would like to suggest that this fact is not so much

linked to 'directionality proper' as it is to whether the meaning of the sentence focuses on the verb (i.e., on the action or motion) itself, or rather on physical, or rather 'regional' elements of the Ground. In fact, if one is running in a race, the only sensible way of envisaging the 'completion of the task' is that of considering the 'length' (rather than the 'width') of the track, and put 'motion' in focus. Whence the absolute preference for the verb+direct object construction by the respondents. On the other hand, the case with verb+PP constructions seems to dictate a perceptual shift of focus from 'motion' to 'Path region or part/boundary'. In that case the physical (part) of Path (expressed by the preposition) seems to gain semantic prominence over the 'completion of the task' element, and then it appears more sensible to think of the 'shorter' axis (cf. Landau and Jackendoff 1993: 227).

In summary, optionality between PP and direct object with prefixed verbs of motion seems to occur with verbs, i.e., Grounds that can be viewed both from a 'physical' as well as from the 'completion' point of view. A PP following a verb of motion bears some relation to path-boundedness. A sentence in which a prefixed motion verb is followed by a direct object focuses more on the fact that the action has been completed. In fact, the main component of meaning being expressed is given by the fact that a target state has been reached, while the physical boundaries of G, or rather the boundaries to which motion is confined, are left out of focus.

4.2. Prefixed Verbs of Motion which Obligatorily Select a PP

Let us begin by considering a few examples of sentences that always require a PP, and where, furthermore, the preposition retains the same form as the prefix.

(8) *U*letjeti *u* gnijezdo.
*infly in nest.ACC
'fly into the nest'

(9) *Is*trcati *iz* kuce.
*outfly from house.GEN
'fly out of/from the house'

(10) *Na*trcati *na* prepreku .
*onrun onto obstacle.ACC
'bump into/against an obstacle'

In sentences of this type the prefixed motion verb is to be followed by a preposition, whose meaning also seems to be ***the same*** as that of the prefix.

However, upon closer analysis, we might want to ask whether the prefix and the preposition really have the same meaning, or is it just the form that is the same, whereas the semantic content might be different. In fact, departing from the conclusion we reached in Sections 3 and 4.1 above, we might want to state that the role of the prefix is to determines a change of state, i.e., to introduce the target state to be brought about by the motion expressed by the verb. In such cases, if any (part of) Path, such as the source or the goal of motion, is to be overtly expressed, it needs to be *specified separately* (the semantic role of the preposition).

All the prefixed verbs of motion, which require a PP appear to share a common semantic feature: they all involve a close, tight relation between the act of motion and some element of physical Path. It could be said that these verbs refer to motion which is less about 'motion bringing about a new. i.e., target state' than it is about 'physical elements of the Ground being covered by motion'.

As an interesting case, which can be seen in line with what has been said so far, is the sentence pair (6) (repeated here for convenience) from the 'optionality pattern questionnaire'.

(6) a. Obletjeti kucu.
*roundfly house.ACC
b. Obletjeti oko kuce
*roundfly around house.GEN
'Fly around the house'

In fact, while both are acceptable, the majority of the respondents (63%) rated (6b) as being 'better sounding', 'more accurate', or 'more natural' than (6a) (while, as seen in Section 4.1, on most other optionality pattern sentence pairs, the preferred construction was that of the sentence in which the verb is followed by the direct object). The reason I see for this preference for the verb+PP option with *obletjeti* '*roundfly', is that this verb intrinsically incorporates a very well defined Path (i.e., around the house), and does, as such, sound better in a syntagm where the prefixed verb is lexicalised by the preposition.

Concluding, we note that prefixed motion verbs, whose semantics is such as to involve explicit expression of (part of) physical Path involved in motion, require a PP.

4.3. Prefixed Verbs of Motion which Obligatorily Select a Direct Object

Finally, let us consider the following sentences:

(11) a. John je prestigao protivnika.
John be.AUX *overcome opponent.ACC
b. **John je prestigao preko protivnika.*
John be.AUX *overcome over opponent.GEN
'John has overcome the opponent.'

(12) a. Prebroditi krizu
*Oversail crisis.ACC
b. *Prebroditi preko krize
*Oversail over crisis.GEN
'Overcome the crisis'

(13) a. Premostiti slijedeca 2 tjedna.
*overbridge next two weeks.ACC
b. Premostiti preko slijedeca 2 tjedna.
*overbridge over next two weeks.GEN
'Bridge the next two weeks'

As (11-13b) intend to show, in all three cases, the prefixed option is not an option, since it is ungrammatical. The meaning of none of the three sentences above (under a) seems to focus on the 'physical' nature of the action, i.e., motion expressed, but rather on the 'achievement', i.e., 'accomplishment' component of the action. This would appear to be in line with our findings so far. A natural implication following from the conclusion that absence of physical path leads to direct object selection by the verb is that all metaphorical usages of motion verbs would disallow the PP. However, there are a number of real life examples that prove this hypothesis wrong. In fact, in Croatian, 'to fall for a joke' is translated as *nasjesti na salu,* where *nasjesti* is a prefixed verb formed by prefixing the verb *sjesti* 'to sit down' and the prefix *na,* meaning 'on'. The construction **nasjesti salu,* with direct object for 'joke', is ungrammatical. *Nasjesti,* as '*onsit onto something', obligatorily selects a PP both in the physical and in the metaphorical senses. At first glance, this fact might be seen as representing a flaw in the interpretative framework suggested in this paper, but a possible explanation for this phenomenon is suggested below.

Summarising, we state that the category of prefixed motion verbs which disallow a PP and require a direct object to follow the verb include all those instances in which there is no reference to the physical elements linked to motion such as Path, Directionality, or Ground's physical[9] boundaries, either because (a) this 'physical' element is totally irrelevant for the meaning of the sentence, or (b) there is no 'physical' element at all, in that the sense of the sentence is metaphorical, i.e., the G has no physical properties at all.

Within the latter category, however, we are to distinguish between (b1) metaphorical usages of prefixed motion verbs which come into metaphorical language from the optionality pattern group and which in the metaphorical sense allow only the verb+direct object construction, and (b2) metaphorical usages of prefixed

[9] Where 'physical' needs to be added in order to allow for the explication of measure phrases **John crossran 100 meters*, where **John crossran across 100 meters* would be ungrammatical. In fact, 'a 100 meters' is just a 'distance to be *crossrun', and as such has nothing to do with the physical aspects of the Path, and it has all to do with the 'Aktionsart' elements associated with the verb.

motion verbs which come into metaphorical language from the prefixed verb+obligatory PP, and which, as such, retain the PP in the expression of metaphorical sense as well.

5. Conclusions

In viewing 'where' and 'motion' as distinct primitives, Wiezbicka (1996) offers us an argument for positing the activity expressed by a sentence containing a (Croatian) prefixed motion verb as possibly focusing on *either* the 'act of motion' itself (the primitive 'motion'), *or* the 'Path', or rather 'Part of Path' (the primitive 'where'). We could also further expand on what we posited as the 'focusing on act of motion' and say that since the prefix comes from spatial language,[10] prefixation automatically brings about a region (which nonprefixed motion verbs do not do *explicitly*). With prefixation, such a region necessarily involves Klein's (1994) 2-state situations or contents with a source state and a target state, or rather, it automatically opens up a scenario which involves a state and its opposite. The Aktionsart element of 'crossing the boundary' follows from there. We could then summarise our findings by stating that prefixation is about lexicalising Aktionsart. This would also appear to be a valid answer to the question of why prefixed verbs followed by a direct object have more of the 'completion' sense to them. Here, Wierzbicka's (1996) before/after distinction in terms of primes seems to be very much in line with Klein's (1994) 'change of state', i.e., his 2-state contents.

Furthermore, if we take prefixation to be an Aktionsart marker, any 'physical' elements related to this change of state, such as (part) of Path, or Path boundedness, have to be taken up by a preposition. Hence, overt expression of such elements requires the prefixed verb to be followed by a PP. This could also be seen as an explanation as to why native speakers whose acceptability judgement and semantic ratings were tested on the optionality pattern see the verb+PP as 'redundant' when they are talking about 'completion', while preferring the verb+PP construction with

[10] In fact, all prefixes have a cognate spatial preposition of an identical surface form.

prefixes which are very 'physical', rating in such cases the verb+direct object option as 'bare', and 'lacking something'.

Still further support for the above hypotheses is more syntax-based. Van Valin and Lapolla's (1997) analysis of prepositional phrases termed 'adpositional phrases' (ibid.:52-53,159-162) focuses on a distinction between non-predicative prepositions, such as *to* in *Bob gave the book to Mary*, which do not license their object, and predicative prepositions, such as *in* in the sentence *Bob read in the library*, which do function as predicates and license their object. Departing from this observation, we note that in our prefixed verb+direct object case, the object is the argument of the verb, hence 'suffering' the action of the verb, i.e., the motion. With the prefixed verb+PP constructions, the NP is the argument of the preposition, hence 'suffering' the semantic selectional properties of the preposition. It is as if in the latter case the 'predication' over the object is that of the preposition, whereas in the former case it is the prefixed verb that predicates over the direct object.

In conclusion, let us state that with prefixation of motion verbs, the main semantic load of the prefix is the expression of a change of location, i.e., state, with the 'crossing the boundary'[11] as the main semantic feature. However, whenever prefixation yields a motion which semantically includes (part) of clearly bounded (physical) Path, then the Path brought about by this change of location has to be expressed by a PP. This is the case with the prefixed verb+PP construction. Other prefixed verbs are concentrated exclusively on the 'change of location', or rather 'change of state', where 'change of state' is the only semantic component which is semantically, and pragmatically salient. Such verbs, consequently, disallow a Path to be expressed, and obligatorily appear in the Prefixed motion verb+direct object construction. Finally, a third group of prefixed motion verbs, by alternating between the addition and non-addition of the PP, can optionally focus either on the change of state or on the Path or part thereof.

By analysing ways in which various semantic elements interact, thus determining systematic relations between semantics on the one

[11] Into a target state, with both source and target state lexicalised within the TT.

hand, and surface, syntactic behaviour of the Croatian language in its lexicalisation of motion events, on the other, this paper is aimed at addressing the following three issues: (a) the contribution of different linguistic tools, among which are lexical items and grammatical markers, to the expression of motion; (b) the ways in which (inherent) temporal information and spatial information interact; and (c) by showing that the Croatian prefixed verb does not conflate Path, provide further evidence supporting the validity of Talmy's (op.cit.) predictions relative to typical conflation patterns for expressing motion.

The idea underlying the whole approach is the quest for semantic universals, and the subsequent question of semantic primitives. Further crosslinguistic exploration into issues of this nature could, in my view, offer interesting insights into aspects of language specificity vs. universality, more widely issues related to language acquisition and processing, and, finally and most speculatively, issues regarding the cognitive architecture of the human language faculty.

From the viewpoint of this article, and within the general question of how language organises conceptual content, particular relevance is attributed to the proposal by Talmy (1983, 1985) and Slobin (1985) according to which there exists a difference between the kinds of meaning expressed by open-class and closed-class forms: the former are essentially unbounded, while the latter are constrained .

Of particular interest for further research in the direction of language universals is Talmy's (in press) view of the language system as consisting of two sub-systems: the expressional and the conceptual. The closed class portion of the expressional system is seen as reflecting the fundamental structuring properties of the underlying cognitive system. If there is a sound possibility that the closed class items of language can be seen as an explicit pointer to its structural properties, investigations such as the one presented in this article become helpful in answering questions concerning the cognitive architecture of the human language faculty, and possibly beyond. After all, the brain seems to be the direction in which linguistic science is necessarily headed.

References

Babic, S. 1986. *Tvorba Rijeci u Hrvatskom Knjizevnom Jeziku (Croatian Word Formation)*. Zagreb: Globus.

Baric, E. et al. 1995. *Hrvatska Gramatika (A Grammar of Croatian Language)*. Zagreb: Skolska Knjiga.

Berlin, B & Kay, P. 1969. *Basic Color Terms: Their Universality and Evolution*. Berkeley, California: University of California Press.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.

Jackendoff, R. 1996. "The Architecture of the Linguistic Spatial Interface". In Bloom, P. et al. (eds.) *Language and Space*. Cambridge, MA.: MIT Press.

Klein, W. 1994. *Time in Language*. London: Routledge.

Klein, W. 1995. "A Time-Relational Analysis of Russian Aspect". *Language* 71, 669-695.

Landau, B. and Jackendoff, R. 1993. "*What* and *Where* in Spatial Language and Spatial Cognition". *Behavioral and Brain Sciences* 16, 217-265.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Quirk, R. et al. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Slobin, D. 1985. "Crosslinguistic Evidence for the Language Making Capacity". In Slobin, D. (ed.) *The Crosslinguistic Study of Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Talmy, L. 1975. "Semantics and Syntax of Motion". In Kimball, J. (ed.) *Syntax and Semantics*. New York: Academic Press.

Talmy, L. 1983. "How Language Structures Space". In Pick, H. and L. Acredolo. (eds.) *Spatial Orientation: Theory, Research, and Application*. New York: Plenum.

Talmy, L. 1985. "Lexicalization Patterns: Semantic Structure in Lexical Forms". In Shopen, T. (ed.) *Language Typology and*

Syntactic Description, Vol. III: Grammatical Categories and the Lexicon. Cambridge: Cambridge University Press.

Talmy, L. 1988. "The Relation of Grammar to Cognition". In Rudzka-Ostyn, B. (ed.) *Topics in Cognitive Linguistics*. Amsterdam: John Benjamins.

Talmy, L. (in press). *Toward a Cognitive Semantics*. Vol. I and II. Cambridge: MIT Press.

Van Valin, R. and Lapolla, R. 1997. *Syntax: Structure, Meaning and Form*. Cambridge: Cambridge University Press.

Vendler, Z. 1957 [1967]. *Linguistics in Philosophy*. Ithaca: Cornell University Press.

Wierzbicka, A. 1996. *Semantics, Primes and Universals*. Oxford: Oxford University Press.

University of Cambridge
Research Centre for English and Applied Linguistics
Keynes House, Trumpington Street,
Cambridge CB2 1QA,
United Kingdom
mb250@hermes.cam.ac.uk

Full Form Auxiliaries in Serbian/Croatian/Bosnian

Andrew Caink
University of Wolverhampton, U.K.

1. Introduction

In Wilder & Ćavar (1994:3.2) and Mišeska Tomić (1996), it is proposed that the full form auxiliaries in Serbian/Croatian/Bosnian (henceforth SCB) in (1a,b) result from the syntactic adjunction of the clitic form in (1c) to the head of a NegP/AssertionP. This functional head contains the morphemes *je-* and *ne-* respectively.

(1) a. Ja jesam student.
I be.1.SG student
'I am a student'.
b. Ja nisam student.
I be.1.SG.[NEG] student
'I am not a student'
c. Ja sam student.
I be.1.SG student
'I'm a student'.

This paper argues against such an analysis on the grounds of morphology, semantics and syntactic distribution and supports the proposal that the full form auxiliaries are formed in the lexicon (Bošković 1995, King 1996b). This analysis lends support to the argument in Caink (1999) that the presence of a semantic feature in the lexical entry of a full form auxiliary triggers syntactic insertion whereas the so-called clitic auxiliary undergoes Phonological Lexicalization. In Section 2, I review a number of well-established differences in the distribution of the full and clitic auxiliaries. In Section 3, I list a number of shortcomings of the syntactic adjunction account before outlining in Section 4 the minimalist theory of lexicalization adopted, and how this can account for the data under consideration. Importantly, Section 5 suggests that the

same account is also superior with respect to the full and clitic form auxiliary distinction found in SCB future constructions.

2. The Full Form/Clitic Distinction

The paradigm for the present tense full and clitic auxiliaries[1] is listed in (2).

(2) **Table 1.** *Jesam* 'Am' And *Nisam* 'Not Am'

	full form	negative full form	clitic
1.SG	jesam	nisam	sam
2.SG	jesi	nisi	si
3.SG	jest(e)	nije	je
1.PL	jesmo	nismo	smo
2.PL	jeste	niste	ste
3.PL	jesu	nisu	su

Note that the clitic auxiliary forms are to some extent suppletive, whereas the full forms exhibit a stem with person/number inflection. Observe also that, despite a fair degree of similarity between the inflectional endings on the full form auxiliaries, the 3rd person singular declarative ending *–st(e)* is distinct from the clitic *si*.

The clitic auxiliary appears as part of a 'clitic cluster' in (3a). The full form auxiliaries are distinct from the clitic cluster, as shown in (3b) where the clitic cluster is represented by the single pronominal clitic *mu* 'him'.

[1] I refer to the clitic morphemes in (2) as 'clitic auxiliaries' throughout this article in order to avoid confusion with the literature. However, Caink (1998, 1999) argues that so-called clitic auxiliaries in SCB are the Alternative Realization of I^0 features, hence able to appear on the 'highest head' in the extended projection (see Emonds 1999 for a slightly different characterization of Alternative Realization).

(3) Stefan tvrdi...
S. claim.3.SG
a. ... da [mu ga je] Petar poklonio.
that 3.SG.DAT 3.SG.ACC be.3.SG.P. given
'Stefan claims that Peter has given it to him as a present'.
b. ... da [mu] Ivan i Marija jesu/nisu pisali.
that 3.SG.DAT I. and M. be.3.PL./[NEG] written
'He claims that Ivan and Maria *did*/didn't write to him'.

In (3a), the 3rd person singular form *je* 'is' appears in the final position in the bracketed cluster; all other forms appear in first position.

Secondly, the clitic auxiliary forms are phonologically idiosyncratic, being enclitic and hence requiring a host to their left. Lack of such a host leads to the ungrammaticality in (4a). In contrast, the full forms can bear stress and can appear in the first position in (4b,c), just as any 'lexical' verb.

(4) a. *Sam mu ga dala.
be.1. SG 3.SG.DAT 3.SG.ACC given
b. Jesam li mu ga dala?
be.1.SG Q 3.SG.DAT 3.SG.ACC given
'Did I give it to him?'.
c. Nisam mu ga dala.
be.1.SG.[NEG] 3.SG.DAT 3.SG.ACC given
'I didn't give it to him'. (Mišeska Tomić 1996:842)

Thirdly, the full form and clitic auxiliaries appear in different positions. Whilst the full form auxiliaries are generally assumed to appear in I^0 (e.g. following the subject *Ivan i Marija* in (3b)), it has been argued that the clitic form appears on the highest head in any given extended projection of the verb (Caink 1997, 1998, 1999; Franks 1998). In (1c), for example, there is no reason to assume that anything more than an IP has been generated, and thus the clitic auxiliary appears on I^0, the highest head available. In (5a) below, however, the clitic appears on C^0, following the complementizer *da* 'that'.

(5) Nedžad tvrdi da...
N. claims that
a. ... su Ivan i Marija čitali knjigu.
be.3.PL I. and M. read book
'...Ivan and Maria were reading the book'.
b. *...Ivan i Marija su čitali knjigu.
c. *...jesu/nisu Ivan i Marija čitali knjigu.

Examples (3b,c) indicate that the positions of the full and clitic forms are not interchangeable.

Fourthly, a number of authors have observed - but so far failed to explain - why the full form auxiliaries are able to license a trace of movement in (6 and 7a,b), but the clitic form cannot in (6 and 7c).

(6) a. [Pio vina]$_i$ jesam t_i
drunk wine be.1.SG
'I have drunk wine'.
b. [Pio vina]$_i$ nisam t_i
drunk wine be.1.SG.[NEG]
'I haven't drunk wine'
c. *[Pio vina]$_i$ sam t_i
drunk wine be.1.SG
'I have drunk wine'. (Mišeska Tomić 1996:857)

(7) a. [Jako dosadna] jest, ali...
very boring be.3.SG but
'She is very boring, but...'.
b. [Jako dosadna] nije, ali...
very boring be.3.SG.[+NEG] but
'She is not very boring, but...'.
c. *[Jako dosadna] mi je njegova knjiga
very boring 1.SG.DAT be.3.SG his book
'His book is very boring to me'. (Browne 1975:118)

In (6), a VP has been topicalized, and in (7), an AP has been topicalized. In both cases, the full forms are able to license the trace but a clitic auxiliary cannot (assuming some form of head licensing of traces is required, hence a conjunctive ECP as in Aoun et al. (1987), Rizzi (1990)).

Finally, note the difference in semantic interpretation between the full and clitic forms throughout the above data. The negative form clearly carries negation and the declarative full form carries assertion, similar to the use of auxiliary *do* in English *I do like minimalism, honest.*

We summarise the above differences in (8).

(8) **Table 2.** Differences between Full and Clitic Forms

	full form	clitic form
A stem and regular inflectional paradigm	✓	✖
Syntactic position is I^0	✓	✖
Can appear in sentence-initial position	✓	✖
Bears stress/emphasis	✓	✖
Licenses a movement trace	✓	✖

3. The Syntactic Adjunction Account

Wilder and Ćavar (1994:3.2) and Mišeska Tomić (1996) suggest that the full form auxiliary stems *je-* and *ni-* are generated in a functional head Assertion/Neg^0, essentially equivalent to Laka's ΣP (Laka 1990). In keeping with a dominant trend in the 'Government and Binding' theory of the late eighties, these authors propose that the full forms result from the cliticization of the clitic auxiliary to the head of Assertion/NegP.

Hence, the formation of these auxiliaries mirrors the way in which a lexical verb is negated, through head adjunction to Neg^0 as in, e.g., King (1996b). For example, the negative morpheme *ne-* 'not' is cliticized to a lexical verb like *čitam* 'I read' to create sentence negation: *ne-čitam* 'I don't read'.

Wilder and Ćavar point out that this accounts for the complementary distribution of the *je-* and *ni-* morphemes, since they are generated in the same position: the declarative auxiliary *jesam* 'I *am*' never adjoins to Neg^0 to create the form ***nijesam*** (Wilder and Ćavar 1994:23).

There are a number of shortcomings to this account to which we now turn.

Firstly, it is noteworthy that the word order of the clitic cluster is strictly that of (9a), illustrated in (9b,c).

(9) a. Strict Clitic Cluster Word Order:
clitic auxiliary – dative – accusative – reflexive – 3.SG. *je* 'is'

b. Ja [sam mu ga] dala.
I be-1. SG 3.SG.DAT 3.SG.ACC given
'I gave it to him'.
c. Da li [mi ga je] dao?
that Q. 1.SG.DAT 3.SG.ACC be-3.SG given
'Did he give it to me?'.

All forms of the clitic auxiliary appear in the first position, except for the 3rd person singular *je* 'is', which must appear in the final position. However, this otherwise strict word order is apparently violated just in the case where Assertion/NegP is generated:

(10) a. Ja [mu se] nisam predstavio.
I 3.SG.DAT REFLbe.1.SG.[+NEG] introduced
'I have not introduced myself to him'. (Rivero 1991:336)
b. Nije mi ga dao.
be.3.SG.[+NEG] 1.SG.DAT 3.SG.ACC given
'He didn't give it to me'. (Mišeska Tomić 1996:844)

In (10a), for example, the supposed clitic auxiliary *sam* 'am' appears cliticized to Neg^0 and follows the pronominal clitics *mu se* 'to him, self', contra to the word order in (9b). In (10b), *je* 'is' appears in front of the pronominal clitics, unlike in (9c). Whatever

one's account of the clitic cluster, a question remains as to why the order is so strict in all but these cases.

Indeed, considering specific accounts within a Principles and Parameters framework, we may go further and question why the supposed clitic auxiliary may appear in a wholly different position in the clause once it has adjoined to Assertion/Neg^0. This is highly problematic for analyses that stipulate that the clitic auxiliary must 'check a feature' in C^0 in examples such as (3a) and (5a) (e.g. Ćavar and Wilder 1994, Progovac 1996). Whenever Assertion/NegP is generated, this feature must be checked covertly instead, raising further questions about the nature of weak and strong features which seem unlikely to provide revealing answers.

As we have seen, an intuitively pleasing aspect of the syntactic adjunction account is that the formation of the full form auxiliaries mirrors that of the negated lexical verbs. However, the morpheme *ne* on a lexical verb (e.g. *ne-ćitam* 'I don't read') is clearly distinct from the negative stem *ni-* in the negative auxiliary (e.g. *ni-sam* 'I am not'). Interestingly, if we consider other cases of lexical negation in SCB, we find that many, if not all, also utilize the negative morpheme *ni-* rather than *ne-* (e.g. *niko* 'nobody', *ništa* 'nothing', *nikad* 'never'). If the full form auxiliaries are indeed the result of lexical rules rather than syntactic adjunction, this suggests that lexical negation is typified by the morpheme *ni-* whilst syntactic negation is formed with the morpheme *ne-*[2]. The alternative is to suggest that the head of NegP generates both *ni-* and *ne-* in negative clauses, depending on whether a lexical verb or the clitic auxiliary is to be generated lower in the tree[3].

Next, the syntactic adjunction account makes incorrect predictions. First, a minor point, the 3rd person singular declarative form is predicted to be the unattested **jeje*. More seriously, we

---

[2] Though see Section 5 and the negative future full form auxiliary *neću* 'I won't'.

[3] Of course, one might argue that powerful rules in a separate post-syntactic morphological component are able to switch the morpheme as required (along the lines of, say, Halle & Marantz 1993), but this would lose the intuitive attraction of the syntactic adjunction account.

have seen that Wilder and Ćavar cite the ungrammaticality of **nijesam* in favour of their account, yet precisely this form is attested in Standard Montenegrin, and some dialects of Serbian/Croatian. The full paradigm consists of the following:

(11) **Table 3.** Standard Montenegrin Nijesam 'Am Not'

	negative full form
1.SG	nijesam
2.SG	nijesi
3.SG	nije
1.PL	nijesmo
2.PL	nijeste
3.PL	nijesu

This data presents us with two theoretical options: (i) we stipulate a substantial difference in phrase structure between Standard Montenegrin and SCB, despite the fact that they enjoy largely identical syntax, or (ii) we assume two distinct lexical entries for 'not be' in the lexicons of Montenegrin and SCB. The first option is highly stipulative. Given that we are proposing that the full form auxiliaries result from lexical rules on independent evidence, the second option is evidently the more promising alternative, but one which undermines the syntactic adjunction account.

4. A Semi-Post-Lexicalist Approach

In this section, I outline a theory of lexicalization compatible with Bare Phrase Structure (Chomsky 1995: chapter 4) that is developed out of Emonds' theory of syntactic and phonological lexicalization (Emonds 1985, 1994, 1999). As suggested in Caink (1999), the crucial distinction between the full form and clitic auxiliaries stems from the content of their lexical entries; this in turn leads to distinct lexicalization processes which underlie the distributional differences noted in Section 2. We relate the distribution of the SCB auxiliaries with the full form and clitic forms in Standard English and the presence/absence of the copula in Black English Vernacular.

I assume that a lexical item consists of a pairing of phonological features π and purely semantic features λ and formal syntactic features FF: [[π λ] FF]. In the terms of Chomsky (1965), FF are the features involved in a syntactic operation, whereas λ are 'purely semantic' and do not contribute to syntactic operations. A λ feature is however required at the LF interface.

Assume next a form of Global economy, such that it is cheaper for the operation Select to take only the syntactic features of a lexical item in the numeration for merge in the computational system. The operation is therefore better termed as Select F (on a par with Affect F). When a lexical item in the numeration lacks any purely semantic content, such as the clitic auxiliary *je* 'is', then the feature matrix consists of only FF and phonological features, as in (12):

(12) [[π ∅] FF].

In such a case, it is cheaper for Select to take only the FF for merge in the computational system. The phonological features π are introduced at PF.

Open class items contain a full feature matrix [[π λ] FF]. In such a case, lexical feature decomposition is barred; Select is forced to pied-pipe the full feature matrix for merge in the computational system.

In both cases, the FF of lexical items are merged in the syntax and construct the syntactic tree as in Chomsky's Bare Phrase Structure account. However, a terminal node that contains an open class item differs from one that contains a semantically null item by the fact that it dominates both phonological and purely semantic features. This system therefore rejects the post-lexicalism argued for in Zwart (1996), and argues for a 'semi-post-lexicalism', whereby the phonological features of semantically null items are subject to Phonological Lexicalization[4]. The result is

[4] This distinction between the presence or absence of phonological features in the syntax is formally equivalent to Emonds' distinction between lexical items that are inserted at D-structure or at PF respectively.

that the phonological features of the clitic auxiliary are not introduced into the derivation until PF.

Consider now the full form auxiliaries. The declarative full form carries more semantic weight than the clitic auxiliary, captured in the syntactic adjunction account by the notion of an 'AssertionP'. This extra semantic weight must be captured in the lexical entry in some way; I maintain it constitutes a 'purely semantic' feature λ. Let us call it [assertion]. The lexical entry for the stem of the declarative form will therefore include at least the information in (13).

(13) je-, [+V,-N], [ASSERTION]

In the system outlined above, the lexical item in (13) will be inserted 'into the syntax': Select will pied-pipe the full feature matrix, hence the terminal node will dominate phonological features that are visible at PF.

Equally, the negative full form carries the extra semantic weight of 'negation', and will have a lexical entry along the lines of (14):

(14) ni-, [+V,-N], [NEGATION]

Here again, I assume that the presence of a full feature matrix forces Select to pied-pipe all the features. Hence the terminal node in question dominates phonological features which are visible at PF.

Contrary to Chomsky's minimalism, I assume that some form of trace licensing is still required, albeit at the interface levels only. I assume some form of head licensing of a movement trace is required at the level of PF as in, say, Aoun *et al.* (1987): a head properly head governs a movement trace if it is 'visible' at PF. The kind of visibility that concerns us here is defined in (15).

(15) *A terminal node is visible for head government at PF if it dominates phonological features π.*

If we assume that Phonological Lexicalization occurs *after* the level at which a trace is licensed, it becomes evident why there is a distinction between the full form and clitic auxiliaries in terms of trace licensing. In (6a,c) repeated here as (16a,b), the full form auxiliary in (16a) is able to license a movement trace, whereas the clitic form whose phonological features are introduced at a later level of PF cannot license the trace[5].

(16) a. [Pio vina]$_i$ jesam t_i
drunk wine be.1.SG
'I have drunk wine'.
b. *[Pio vina]$_i$ sam t_i
drunk wine be.1.SG
'I have drunk wine'. (Mišeska Tomić 1996:857)

As noted in Caink (1999), precisely the same distinction occurs in Standard English. The full form *is* can license a trace in a way that the clitic form *–s* cannot:

(17) a. You think he is/'s where today?
b. Where$_i$ do you think he is/*'s t_i today? (Zagona 1982)
c. John is/'s a teacher.
d. I wonder what$_i$ John is/*'s t_i now.

A similar distinction is found in Black English Vernacular between the null form and the full form. Although the copula may be null, a movement trace must be licensed by the full form.

(18) a. She ∅/*is the first one started us off.
b. You ∅/*are out the game.
c. I don't care what$_i$ you are/*∅ t_i
d. Do you see where that person is/*∅ t_i (Labov 1972)

[5] Note that this system assumes that syntactic information is retained en route to PF, unlike in Chomsky (1995) where it is assumed that only the phonological features are 'stripped away'.

In our terms, the whole feature matrix of the full form has been introduced into the computational component and is present at the relevant level for PF head licensing. The null copula appears to lack any phonological features at this level as well. Whether no phonological features are introduced, or they are introduced only to be subsequently deleted, is a separate issue.

5. And the Future Forms?

Finally, SCB exhibits a similar full form/clitic distinction in the future tense. In this section, we will briefly view the data and consider how successful the respective accounts are in dealing with these forms. Preliminary examples are given in (19) and the full paradigm in the table in (20).

(19) a. Hoću mu ga dati.
will.1.SG 3.SG.DAT 3.SG.ACC give.INF
'I will give it to him'.
b. Neću mu ga dati.
will.1.SG.[NEG] 3.SG.DAT 3.SG.ACC give.INF
'I will not give it to him'.
c. Da- ću mu ga.
give.INF will.1.SG 3.SG.DAT 3.SG.ACC
'I will give it to him'.

(20) **Table 3.** *Hoću* 'Will' and *Neću* 'Not Will'

	full form	negative full form	clitic
1.SG	hoću	neću	ću
2.SG	hoćeš	nećeš	ćeš
3.SG	hoće	neće	će
1.PL	hoćemo	nećemo	ćemo
2.PL	hoćete	nećete	ćete
3.PL	hoćê	nećê	ćê

In (19a,b), the full forms appear in sentence-initial position, indicating they are not enclitic, unlike the clitic form in (19c)

which appears cliticized to the reduced infinitive form of *dati* 'give'.

Just as the full forms in Section 2, the full future forms appear in different positions to the clitic form:

(21) On tvrdi ...
he claims
a. da [mu] Ivan i Marija hoćê /nećê pisati.
that 3.SG.DAT I. and M. will.3.PL/won't write.INF
'He claims that I.and M. will/won't write to him'.
b. da ćê mu Ivan i Marija pisati.
that will.3.PL 3.SG.DAT I. and M. write.INF
'He claims that I.and M. will write to him'.

In (21a), the full forms appear in I^0, following the subject *Ivan i Marija*. In (21b), the clitic form *ćê* 'will' appears with the clitic cluster in C^0.

The data in (22) suggests that the full forms are able to license a trace, just as their counterparts in Section 2.

(22) a. [Piti vina]$_i$ neću t_i
drink wine will.1.SG.[NEG]
'I won't drink wine'.
b. [Biti jako dosadna]$_i$ neće t_i , ali...
be.INF very boring will.1.SG.[NEG] but
'She will not be very boring, but...'.

Although the authors associated above with the syntactic adjunction account do not explicitly discuss the future forms in these terms, it is worth considering how such an account might deal with these forms. Presumably, the clitic form *-ću* 'will' adjoins to a head position that contains either *ho-* or *ne-*. Presumably this must also be the head of Assertion/NegP. Hence this head position generates either *je-* or *ho-* if it is declarative, along with *ne-* or *ni-* if specified as negative, depending on the nature of the clitic auxiliary generated further down.

A major problem with this account is that the full form has more semantic possibilities than the clitic form. Whereas the clitic form *-ću* 'I will' carries only future modality, the full future auxiliary, whether negated or not, carries either future modality *or* the sense of 'will, desire'. This is illustrated with the negative forms in (23):

(23) a. Neću biti kod kuće.
will.1.SG.[NEG] be at home.GEN
'I won't be at home'.
b. Neću kući.
will.1.SG.[NEG] home.DAT
'I don't want to go home'.

Hence, *neću* 'I will not' is more than simply a syntactic combination of Neg^0 and the clitic future *-ću* 'I will'. In the semi-postlexicalist account, the full and clitic forms have distinct lexical entries, independently motivated by the trace licensing facts.

Contra to our analysis is the fact that the negative morpheme is *ne-* in (23). This contradicts the earlier argument that *ni-* and *ne-* are distinct morphemes, indicative of lexical negation and syntactic negation respectively.

6. Summary and Conclusion

In this paper, I have reviewed the distinctions between the full and clitic forms of the SCB auxiliaries and, concentrating on the full forms, have argued that they result from lexical rules, rather than GB style adjunction in the syntax. Their lexical entries contain the features [assertion] or [negation] which carry enough semantic weight to trigger 'syntactic insertion'. In contrast, the clitic auxiliary carries no semantics and is phonologically lexicalized in the way discussed in detail in Caink (1999).

Presence in the syntax allows the full form auxiliaries in both SCB and English to license movement traces, assuming some form of PF head licensing in Aoun *et al.* (1987). In contrast, the clitic

forms in SCB and English are assumed to be introduced after the level of PF at which head licensing occurs.

I outlined a reformulation of this lexicalization theory for a system with Bare Phrase Structure, thus enabling us to combine the insights of Emonds' (1985, 1994) theory of lexicalization with the minimalist aspirations of Chomsky (1995). 'Syntactic insertion' means that Select F has to pied-pipe the phonological and semantic features of a lexical item along with the formal syntactic features for merge in the computational system. 'Phonological Lexicalization' indicates that lexical feature decomposition occurs: Select takes only the formal syntactic features of a lexical item FF on grounds of economy, and the remaining features are introduced at the relevant interface. In this case, phonological features are introduced at PF, following the level at which PF head licensing applies.

References

Aoun, J., Hornstein, N., Lightfoot, D. & A. Weinberg. 1987. "Two Types of Locality". *Linguistic Inquiry* 18,537-577.

Bošković, Ž. 1995. "Participle Movement and Second Position Cliticization In Serbo-Croatian". *Lingua* 96, 245-266.

Browne, W. 1975. "Serbo-Croatian for English Speaking Learners", in Filipović (ed.) *Kontrastiva Analisa Engleskog i Hrvatskog ili Srpskog Jezika*, University of Zagreb.

Caink, A. 1997. "Extended Projections in South Slavic", Paper presented at *Formal Approaches to South Slavic Linguistics II*, Sofia.

Caink, A. 1998. *The Lexical Interface: Closed Class Items in South Slavic and English*, Doctoral Dissertation, University of Durham.

Caink, A. 1999. "Serbian/Croatian/Bosnian Clitics at the Lexical Interface", in H. Coates, K. Dziwirek & C. Vakareliyska (eds.) *Formal Approaches to Slavic Linguistics 7*, 81-100. Ann Arbor: Michigan Slavic Publications.

Ćavar, D. & C. Wilder. 1994. "Word Order Variation, Verb Movement, and Economy Principles". *Studia Linguistica* 48, 46-86.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*, Cambridge, Mass.: MIT Press.

Chomsky, N. 1995. *The Minimalist Program*, Cambridge, Mass.: MIT Press.

Collins, C. 1997. *Local Economy*, Cambridge, Mass.: MIT Press.

Emonds, J. 1985. *A Unified Theory of Syntactic Categories*, Dordrecht: Foris.

Emonds, J. 1994. "Two Principles of Economy", in G. Cinque et al. (eds.) *Paths Towards Universal Grammar; Studies in Honor of Richard S. Kayne*, 155-172, Georgetown: Georgetown University Press.

Emonds, J. 1999. "How Clitics License Null Phrases; a Theory of the Lexical Interface", in H. van Riemsdijk (ed.) *Empirical Approaches to Language Typology: clitics in the languages of Europe*, Berlin: Mouton de Gruyter.

Franks, S. 1998. "Clitics in Slavic", Position paper, Comparative Slavic Morphosyntax workshop, Spencer, Indiana. (downloadable at http://www.indiana.edu/~slavconf/linguistics/index.html)

Halle, M. & A. Marantz. 1993. "Distributed Morphology and the Pieces of Inflection" in K. Hale & S. Keyser (eds.) *The View from Building 20*, 111-176, Cambridge, Mass.: MIT press.

King, T. H. 1996a. "Slavic Clitics, Long Head Movement, and Prosodic Inversion". *Journal of Slavic Linguistics* 4/1, 274-311.

King, T. H. 1996b. "Structuring Negation in Slavic", in J. Toman (ed.) *Formal Approaches to Slavic Linguistics vol. 3*, 145-170, Ann Arbor: Michigan Slavic Publications.

Labov, W. 1972. *Language in the Inner City: Studies in the Black English Vernacular*, Philadephia: University of Pennsylvania Press.

Laka, I. 1990. *Negation in syntax: on the nature of functional categories and projections*, Doctoral Dissertation, MIT.

Mišeska Tomić, O.M. 1996. "The Balkan Slavic Clausal Clitics". *Natural Language & Linguistic Theory* 14/4, 811-872.

Progovac, L. 1996. "Clitics in Serbian/Croatian: Comp as the Second Position", in A. Halpern & A. Zwicky (eds.) *Approaching Second Position: Second Position Clitics and Related Phenomena*, 411-428, Stanford: CSLI Publications.

Rivero, M.-L. 1991. "Long Head Movement and Negation: Serbo-Croat vs. Slovak And Czech". *The Linguistic Review* 8, 319-351.

Rizzi, L. 1990 *Relativized Minimality*, Cambridge, Mass.: MIT press.

Wilder, C. & D. Ćavar. 1994. "Long Head Movement? Verb Movement and Cliticisation in Croatian". *Lingua* 93, 1-58.

Zwart, C. J.-W. 1996. "Morphosyntax of Verb Movement: a Minimalist Approach to the Syntax of Dutch", Dordrecht: Kluwer Academic Publishers.

Zagona, K. (1982) *Government and Proper Government of Projections*, Doctoral Dissertation, University of Washington.

Linguistics Division
School of Humanities, Languages & Social Studies
University of Wolverhampton
Stafford Street
Wolverhampton WV1 1SB
United Kingdom
a.d.caink@wlv.ac.uk

The Unergative/Unaccusative Split and the Derivation of Resultative Adjectives in Polish*

Bożena Cetnarowska
University of Silesia, Sosnowiec (Poland)

1. Introduction

It has often been suggested in the linguistic literature (Perlmutter 1978; Burzio 1986; Hoekstra 1984, among others) that there are two classes of intransitive verbs: unergatives and unaccusatives. Subjects of unaccusative verbs pattern syntactically together with objects of transitive verbs, while subjects of unergative verbs share many characteristics of subjects of transitive verbs. Perlmutter (1978) views the intransitivity split as universally relevant. Consequently, I attempt to show below that it characterizes intransitive predicates in Polish. The class of intransitive predicates considered here includes "intransitives proper", such as *upaść* 'to fall', as well as verbs occurring with the reflexive clitic *się*, e.g., *potknąć się* 'to slip'.[1]

---

* An earlier version of this paper was presented at Indiana University in Bloomington. I would like to thank the anonymous reviewers for this volume and the audiences at Indiana University and at UPenn (FASL-8) for their comments. I am particularly grateful to Masha Babyonyshev, George Fowler, Steven Franks, Kyle Johnson, Eva Juarros, Jim Lavine, Anita Nowak, Tom Roeper, Bob Rothstein and Misha Yadroff for helpful discussion. Special thanks are due to Bob Rothstein for reading the final version of the manuscript. All errors are my own. The research reported here was supported by a grant from the Fulbright Program.

[1] Verbs such as *potknąć się* 'to slip', *śmiać się* 'to laugh' or *spóźnić się* 'to come late' can be regarded as intransitive since they denote one-place predicates (i.e., predicates involving one argument). The verbs mentioned in the preceding sentence belong to the class of *reflexiva tantum* which require the presence of the reflexive clitic. The clitic *się* occurs also in reflexive sentences (i), in reciprocal sentences (ii), in impersonal sentences (iii) and in middle sentences (iv).

(i) *Dzieci pomalowały się szminką mamy.*
"The children painted themselves with their mother's lipstick."

(ii) *Zobaczymy się wkrótce.*
"We will see each other soon."

(iii) *Tu się ciężko pracuje.*
"One works hard here."

Following Perlmutter (1978), Zaenen (1993) and Levin and Rappaport Hovav (1995), among others, I assert that the unergative/unaccusative distinction has a semantic basis but its consequences can be observed in the syntax (i.e., it is a semantically determined syntactic phenomenon). The surface subject of an unaccusative predicate has the semantic role of Theme/Patient and it typically undergoes a change of state or a change of location. Subjects of unergative verbs are Agents.

The standard way of capturing the distinction between the two classes of intransitive verbs, proposed by Burzio (1986) within the framework of Government and Binding, is to analyze 'surface' subjects of unaccusative verbs as originating in the position of objects (internal arguments) at D-structure. Since an unaccusative verb is unable to assign the accusative case, its D-structure object must move to the surface subject position. Subjects of unergative verbs, in contrast, are base-generated as D-structure subjects (as external arguments). I assume, following, among others, Hale and Keyser 1993 and Babyonyshev 1996, that unaccusative verbs lack an agentive (vP) projection above VP.

Levin and Rappaport Hovav (1995) draw a distinction between deep and surface unaccusativity. They point out that diagnostics of surface unaccusativity apply only if the surface subjects of unaccusatives remain in the postverbal position (for instance, locative inversion and *there*-insertion in English or *ne*-cliticization in Italian). The tests for surface unaccusativity tend to identify only a subclass of unaccusative verbs, namely, verbs of existence and verbs of appearance. The diagnostics of deep unaccusativity (such as auxiliary selection in Dutch and the resultative construction in English) depend on the semantic properties of predicates.[2]

Various tests have been proposed in the literature to diagnose unaccusativity in Russian (Pesetsky 1982; Babyonyshev 1996;

(iv) *Ta bluzka prasuje się dobrze.*
"This blouse irons well."
See Wilczewska (1966) for more discussion of reflexive sentences.

[2] Alexiadou and Anagnostopoulou (1998) demonstrate for Greek that some diagnostic criteria for unaccusativity do not coincide since they are sensitive to distinct properties of predicates.

Schoorlemmer 1995). These tests include the genitive of negation, unmarked word order, quantification by prefixed *pere-/po-* and *na-* verbs, conjunction agreement and the use of adverbial *po*-phrases. They appear to be primarily tests for surface unaccusativity since they involve quantifier scope or discourse functions. They tend to select verbs of appearance and existence and to disfavour change-of-state verbs. The diagnostic power of the genitive of negation in Polish is much more restricted than in Russian. This construction is obligatory with the verb *być* 'to be' and impossible with other intransitive verbs. In the case of the unmarked word order test, on the other hand, it is difficult to control the influence of discourse factors and the animacy of subject NPs on native-speaker judgments.

In this paper I propose to modify two of the diagnostics of deep unaccusativity employed in Germanic languages to make them suitable for diagnosing the intransitivity split in Polish. Firstly, I show that the existence of resultative adjectives terminating in *-ły*, e.g., *wychudły* 'thinned' or *zmarły* 'dead', is indicative of the unaccusative status of the related verb in Polish. Secondly, I consider the applicability of the impersonal construction for identifying Polish unaccusative verbs. The predictions of both purported unaccusativity diagnostics are then shown to coincide. Finally, it is explained why those tests for unaccusativity are one way implicational (e.g., why the non-occurrence of a related *-ły* adjective should not be taken as evidence for the unergative status of a given verb).

2. Resultative *-ły* Adjective Test

As has been observed for Dutch, German and English, the past participles of unaccusative verbs can be used prenominally as adjectival modifiers (Hoekstra 1984; Mulder 1992). The past participles in (1) are formed from unaccusative verbs in Dutch whereas those in (2) come from unergative verbs. (3) shows passive participles in Dutch acting as premodifiers.

(1) a. de gevallen man
the fallen man

(2) a. *de gelachen man
the laughed man

b. de gestorven man
the died man
c. het gezonken schip
the sunken ship

b. *de gewerkte man
the worked man
c. *de gewandelde man
the strolled man

(3) a. de geslanden hond
the beaten dog
b. het gezongen song
the sung song

The generalization manifested in (1-3) is that prenominal participles can modify internal arguments of verbs (i.e., objects of transitive verbs and surface subjects of unaccusatives).

Polish lacks past participles as independently occurring inflectional forms, hence the unaccusativity diagnostic exemplified in (1-3) for Dutch is not directly applicable in Polish. It is instructive, however, to consider the derivational pattern of resultative adjectives terminating in *-ły*:

(4) a. *przybyły* 'arrived'
b. *upadły* 'fallen'
c. *posiwiały* 'grey, grizzled'
d. *zmarły* 'dead'
e. *zgniły* 'putrid, rotten'
f. *zbiegły* 'escaped'
g. *zdziczały* 'savage'
h. *zwiędły* 'faded'

The adjectives in (4) denote the state resulting from the event denoted by the related verb. The semantic classes of verbs which form *-ły* adjectives most productively are given in (5), following the terminology employed in Levin and Rappaport Hovav (1995):

(5) a. Verbs denoting change of state (especially entity-specific and internally caused change of state):
oślepnąć 'to become blind' *opuchnąć* 'to become swollen'
zachrypnąć 'to grow hoarse' *zemdleć* 'to faint'
b. Verbs of inherently directed motion:
rozpierzchnąć się 'to scatter' *przybyć* 'to arrive'
upaść 'to fall' *zbiec* 'to escape'
c. Verbs of disappearance:

polec 'to die (in battle)' *przepaść* 'to disappear'
umrzeć 'to die' *wygasnąć* 'to expire'
d. Verbs of appearance and occurrence:
wyniknąć 'to ensue' *zaistnieć* 'to come into being'
e. Verbs of assuming position:
obwisnąć 'to hang down' *zalec* 'to lie down somewhere'

The verb classes listed in (5) take non-agentive subjects, denote a change of state or position and are often found cross-linguistically to pattern like unaccusatives (as observed in Levin and Rappaport Hovav 1995).[3] I will show in the next section that a very similar set of verbs are not allowed in the impersonal *-no/-to* construction, which is sensitive to the unaccusative/unergative status of verbs.

3. Impersonal *-no/-to* Construction

It has been postulated for Dutch and German (Perlmutter 1978), that unergative verbs, in contrast to unaccusative verbs, allow impersonal passivization. This is illustrated for Dutch in (6):

(6) a. Unergative:
Er werd gelachen (door Jan).
there was laughed (by Jan)
b. Unaccusative:
*Er werd gevallen (door Jan).
there was fallen (by Jan)

[3] An anonymous reviewer remarks that Rosen (1984) has demonstrated the fallacy of using semantic criteria and cross-linguistic analogy in identifying unaccusative verbs. While Polish verbs recognized here as unaccusatives are grouped into semantic classes for ease of exposition, it is not claimed here that all verbs exhibiting an appropriate semantic interpretation will be necessarily unaccusative. The status of each verb needs to be tested by means of the unaccusativity diagnostics discussed above.

Furthermore, impersonal passives can be formed with transitive verbs but not with passive verbs. As argued in Perlmutter (1978), these facts allow a syntactic explanation. Promotion of internal arguments (to surface subject position) excludes the possibility of impersonal passivization.

The impersonal *-no/-to* construction in Polish cannot be regarded as the impersonal passive synchronically, though it was historically the nominal neuter form of the passive participle (Siewierska 1988). Although the verb occurs in the morphologically passive form, it does not allow the modification by the agentive adjunct phrase *przez kogoś* 'by someone' (see Dziwirek 1994). However, the *-no/-to* construction seems to target similar semantico-syntactic classes of intransitive/reflexive verbs as the impersonal passives in Dutch/German. Apart from a transitive predicate in (7a), it allows an intransitive predicate with an agentive subject in (7b) (Rozwadowska 1992; Śpiewak and Szymańska 1997).[4] Intransitive predicates with non-agentive surface subjects in (7d, e, f) and passive predicates (as in 7g) are unacceptable in this construction.[5]

(7) a. Transitive:
Zbudowano szpital (*przez żołnierzy)
no-built.PF hospital.ACC (*by soldiers)
'They built a hospital (*by soldiers).'
b. Unergative:
Zatańczono (*przez Jana).
no-danced.PF (by Jan)
'They danced'
c. Unergative:

[4] PF stands for 'perfective', IMPF for 'imperfective' and M for 'masculine' in (7). See Babby (1998), Puzynina (1993) and Wolińska (1978) for more discussion of impersonal constructions in Russian and Polish.

[5] The *-no/-to* construction in Polish differs in this respect from the superficially similar nonagreeing passive-participial constructions in North Russian (*-no/-to*) and Lithuanian (*-ma/-ta*). They are analyzed in Lavine (1999) as involving ergative quirky-case subjects and nominative objects.

Zadzwoniono do nas wieczorem.
no-phoned.PF to us evening.INSTR
'They phoned us in the evening.'

d. Unaccusative:
*Wyrośnięto w atmosferze terroru.
no-grew-up.PF in atmosphere.LOC terror.GEN
'They grew up in an atmosphere of terror.'

e. Unaccusative:
*Wychudnięto w ciągu zeszłego lata.
no-thinned.PF in course last summer
'They grew thin last summer.'

f. Unaccusative:
*?Upadnięto na kolana przed cesarzem.
no-fell.PF on knees.ACC before emperor.INSTR
'They fell on their knees in front of the emperor.'

g. Passive:
*Byto poniżanymi.
no-was.IMPF humiliated.M.PL.INSTR
'They were humiliated.'

Let us emphasize that the requirement of the non-passive nature of the surface subject is not a characteristic of another impersonal construction attested in Polish, namely, the reflexive speaker/addressee inclusive construction. As shown in (8a-d), the equivalents of (7d-g) are fully felicitous in the impersonal reflexive.[6]

(8) a. Wyrosło się w atmosferze terroru.
grew.3.N.SG.PF. r.cl in atmosphere.LOC terror.GEN
'One/you grew up in an atmosphere of terror.'

b. Wychudło się w ciągu zeszłego lata.
grew-thin.3.N.SG.PF r.cl. in course last summer
'One/you grew thin in the course of the last month.'

c. Jeśli nie upadło się na kolana przed
if not fell.3.N.SG.PF r.cl. on knees.ACC before

[6] The abbreviation 'r.cl.' stands for 'reflexive clitic' and 'N' for 'neuter'.

cesarzem, trzeba ponieść karę.
emperor.INSTR necessary suffer punishment.ACC
'If one/you didn't fall to their knees in front of the emperor, one/you will have to be punished.'

d. Było się poniżanym.
was.3.N.SG r.cl. humiliated.M.SG.INSTR
'One/you was/were humiliated.'

Consequently, I will assume that the non-reflexive impersonal construction in Polish has the same potential as the impersonal passive in Dutch or German to distinguish unergative verbs from unaccusative ones.[7]

The list of intransitive verb classes in Polish which are not felicitous in the impersonal *-no/-to* construction is given in (9). Native speakers of Polish decidedly reject *byto*, *żyto* or *umarto* (related to *być* 'to be', *żyć* 'to live' and *umrzeć* 'to die', respectively) as ill-formed. There is some variation between acceptability judgments offered by native speakers for sentences containing the *-no/-to* forms related to other verb classes listed in (9), for instance to nonagentive verbs of inherently directed motion (in 9b). Forms such as *przybyto* *'no*-arrived' may be viewed either as unacceptable or as marginally acceptable.[8]

---

[7] Babby (1998) suggests that impersonal sentences in Slavic have no external theta role, i.e., they are derived unaccusatives (their formation involves the suppression of the verb's initial external theta-role). This would account for the impossibility of using underlying unaccusatives (such as *być* 'to be') or passive verbs (i.e., other derived unaccusatives) in the *-no/-to* impersonal construction in Polish. Śpiewak and Szymańska (1997) propose that the position of the defocused agent participant in the *-no/-to* sentences is occupied by a null element. The whole matter deserves more attention but, for reasons of space, is beyond the scope of the present paper.

[8] One of the reviewers for this volume regards *-no/-to* forms of nonagentive change-of-state verbs in (9b) as legitimate. The forms *osiwiano* *'no*-turned grey' and *zmoknięto* *'no*-became wet' are included in the conjugational paradigms of the corresponding verbs *osiwieć* 'to turn grey' and *zmoknąć* 'to soak' in Doroszewski (1980) but they are qualified there as "not used". Puzynina (1993) treats *skamieniano* *'no*-turned to stone' or *zmiękniętо* *'no*-became soft' as impossible forms.

(9) a. Verbs denoting change of state (unless interpretable as denoting volitional acts):
oślepnąć 'to become blind' *opuchnąć* 'to become swollen'
zachrypnąć 'to grow hoarse' *zemdleć* 'to faint'
b. Verbs of inherently directed motion:
rozpierzchnąć się 'to scatter' *przybyć* 'to arrive'
upaść 'to fall' *zbiec* 'to escape'
c. Verbs of disappearance:
polec 'to die (in battle)' *przepaść* 'to disappear'
umrzeć 'to die' *zniknąć* 'to disappear'.
d. Verbs of appearance and occurrence:
wyniknąć 'to ensue' *zaistnieć* 'to come into being'
e. Verbs of assuming position:
obwisnąć 'to hang down' *zalec* 'to lie down somewhere'
f. Verbs of maintaining position:
leżeć 'to lie' *stać* 'to stand'
g. Verbs of existence:
być 'to be' *istnieć* 'to exist'

There is a striking overlap between the classes of verbs listed in (5) and (9), which provides support for the claim that both constructions are sensitive to the unaccusative/unergative status of verbs.

There is, however, some disparity between the results of the purported unaccusativity diagnostics proposed here for Polish, which needs to be accounted for. I will show that these differences are due to additional morphological and semantic restrictions on the tests in question.

4. Restrictions on the Resultative Adjective Test and *-no/-to* Construction Test

4.1. Resultative *-ły* Adjective Formation

Resultative adjectives can be derived from telic verbs only. The only exceptions are the largely lexicalized forms given in (10) below, which are related to stative verbs.

(10) a. *były* 'former' (from *być* 'to exist')
b. *bywały (w świecie)* 'experienced, knowledgeable' (cf. *bywać* 'to frequent')
c. *rosły* 'tall' (cf. *rosnąć* 'to grow')
d. *stały* 'constant' (cf. *stać* 'to stand')
e. *trwały* 'persistent, durable' (cf. *trwać* 'to persist')

While it is frequently claimed that unaccusative verbs are necessarily telic (van der Putten 1997), there is evidence that some atelic verbs, such as verbs of existence and verbs of maintaining position, exhibit unaccusative properties. The evidence comes from word order phenomena in Russian and auxiliary selection in Italian (Levin and Rappaport 1995; Babyonyshev 1996). The fact that the adjectives in (10) are related to such atelic classes of verbs further strengthens the hypothesis of the link between verb unaccusativity and the occurrence of *-ły* adjectives.

There exist morphological restrictions on verbs deriving resultative *-ły* adjectives. Telic and non-reflexive verbs do not normally form *-ły* adjectives if they are nonprefixed, hence the unacceptability of the nonprefixed form **pękły* 'burst' compared to the prefixed adjective *rozpękły* 'burst'.[9] Moreover, obligatorily reflexive verbs in (11a-c) and verbs participating in the transitive/inchoative alternation in (11d, e) form resultative adjectives by means of the *-n/-t* suffix.

(11) a. *spóźniony* 'late' (from *spóźnić się* 'to come late')
b. *uśmiechnięty* 'smiling' (from *uśmiechnąć się* 'to smile')
c. *zamyślony* 'lost in thought' (from *zamyślić się* 'to fall into thinking')
d. *złamany* 'broken' (from *złamać (się)* 'to break something/somebody, to break oneself')
e. *zmęczony* 'tired' (from *zmęczyć (się)* 'to tire somebody; to get tired')

[9] The form *pękły* is listed in Bajerowa (1992:191) as archaic.

As the *-n/-t* suffix appears also in passive forms, it is difficult to distinguish between a passive participle in (12a) and a non-passive resultative adjective in (12b) without considering their context and the semantic interpretation of the related verb. This is why I focus here on *-ły* adjectives.

(12) a. Passive participle:
złamane przez dzieci drzewo
broken by children tree.N.NOM
b. Non-passive resultative adjective:
zamyślona studentka
lost-in-thought student.F.NOM

Due to the high productivity of the *-n/-t* suffix and the diminishing productivity of the suffix *-l*, one can occasionally find resultative *-n/-t* adjectives competing with *-l* forms (in 13) or ousting them completely (as in 14).

(13) a. *nasiąknięty* or *nasiąkły* 'that has absorbed (water)'
b. *namoknięty* or *namokły* 'saturated (with water)'
c. *popękany* or *popękały* (rare) 'cracked'
d. *wyschnięty* or *wyschły* 'dried'
e. *spuchnięty* or *spuchły* (rare) 'swollen'
f. *zziębnięty* or *zziębły* (rare) 'freezing, chilled, that feels cold'
g. *umarty* (dial.) or *umarły* 'dead'

(14) a. *padnięty* (coll.) 'dead tired' (cf. *paść* 'to fall' and *padły* 'dead')
b. *wypoczęty* 'rested' (cf. *wypocząć* 'to rest')
c. *zwariowany* 'mad' (cf. *zwariować* 'to go mad')

As noted in Bajerowa (1992:190ff.), the number of neologisms terminating in *-ły* diminished in the second half of the nineteenth century. Moreover, certain existing *-ły* formations had by then become archaic, e.g. *wybuchły* (arch.) 'that has erupted', *wyszły* (arch.) 'that has left' or *weszły* (arch.) 'that has come inside'.

In conclusion, while the nonoccurrence of *-ły* resultative adjective is not a proof of the unergative status of an intransitive or reflexive verb, the existence of a related *-ły* adjective seems to be a good indication of the verb's unaccusativity.

4.2. The Impersonal -no/-to Construction

One of the disadvantages of the impersonal construction as a diagnostic for verb unaccusativity is that it cannot distinguish an unaccusative verb from an unergative one if they both take inanimate/non-human subjects. This is shown in (15). Although the impersonal *-no/-to* construction is generally well-formed with unergative verbs, as in (7b, c) in Section 3, the sentence in (15a) is ill-formed if the unergative verb *toczyć się* 'to roll' denotes the movement of inanimate objects.

(15) a. Unergative:
*Toczono się po stole.
no-rolled.IMPF r.cl. on table.LOC
'They rolled on the table.' (referring to billiards balls)
b. Unaccusative:
*Roztapiano się na słońcu.
no-melted.IMPF r.cl. on sun.LOC
'They melted in the sun.' (about icecream)

The use of the iterative/habitual interpretation makes some hypothetically unaccusative verbs permissible in the impersonal *-no/-to* construction.[10] One can contrast, in this respect, the felicity of the imperfective verb forms in (16a, b) with the infelicity of the corresponding perfective verbs in (16c, d).

[10] It is plausible to suggest the existence of two types of *-no/-to* constructions: the non-iterative one (which can serve as an unaccusativity diagnostic) and the iterative one (which admits a wider range of verbs). I do not intend to imply that imperfective verbs such as *umierać* 'to die, impf' and *padać* 'to fall, impf' are unergative (and hence possible in the *-no/-to* form) in contrast to the related perfective verbs *umrzeć* 'to die, pf' and *paść* 'to fall, pf'.

(16) a. Unaccusative:
Podczas wojny umierano z głodu.
during war.GEN *no*-died.IMPF from hunger.GEN
'People would die from hunger during the war.'

b. Unaccusative:
W tamtych czasach padano na kolana przed
in those days *no*-fell.IMPF on knees.ACC before
cesarzem.
emperor.INSTR
'In those days people would fall on their knees in front of the emperor.'

c. Unaccusative:
*Umarto z głodu.
no-died.PF from hunger.GEN
'They died of hunger'.

d. Unergative:
*?Upadnięto na kolana przed cesarzem.
no-fell.PF on knees.ACC before emperor.INSTR
'They fell on their knees in front of the emperor.'

Consequently, in order to get sharper acceptability judgments, it is necessary to restrict the test to non-iterative contexts.

Finally, the felicity of a verb in the *-no/-to* construction increases when it implies greater control of the subject over the event.

(17) a. *Znieruchomiano z przerażenia.
no-became.PF-motionless from fright
'They became still because of fright.'

b. ?Znieruchomiano, a potem znów rozpoczęto
no-became.PF-motionless and then again *no*-began.PF
taniec.
dance.
'They became motionless and then again began the dance.'

The data in (17) can be interpreted as exemplifying the variation between the unaccusative use of the verb in (17a) and the unergative use in (17b).[11]

To sum up, it was shown in Section 4.2. that the acceptability of sentences with the impersonal *-no/-to* construction is determined not only by the unergative or unaccusative status of the verb. It is sensitive to additional factors, such as the availability of the iterative/habitual interpretation and the requirement that the event is predicated of a human participant.[12]

4.3. Comparison of the Results of Both Diagnostic Tests

Below I assess the (degree of) correlation between the results of the two unaccusativity tests discussed above, namely the resultative *-ły* adjective formation and *-no/-to* construction. The initial prediction is that verbs which derive resultative *-ły* adjectives are unable to occur in the impersonal *-no/-to* construction and vice versa. However, since there exist additional factors, apart from the unaccusative/unergative status of the verb, to which these diagnostic tests are sensitive, the correlation between the results of both tests is blurred.

A strong correlation between the predictions of my unaccusativity diagnostics can be observed in the case of non-alternating change-of-state verbs (group A), in other words intransitive verbs which do not allow the transitive (causative) usage, e.g., *zachrypnąć* 'to grow hoarse', *zemdleć* 'to faint'. Such verbs derive adjectives terminating in *-ły* in a very productive manner and are decidedly unacceptable in the impersonal *-no/-to* sentences.

---

[11] Variable behaviour of verbs with respect to unaccusative diagnostics is discussed at length in Chapter 5 of Levin and Rappaport Hovav (1995) and in Dixon (1994). Levin and Rappaport Hovav point out, for instance, that the *roll* verbs show unergative behaviour when used agentively. When used nonagentively (with the external cause left unexpressed), the *roll* verbs are unaccusative.

[12] Puzynina (1993) and Rothstein (1970) observe that most verbs with the thematic suffixes *-e-* and *-nq-* have no related *-no/-to* forms, e.g., *ginąć* 'to die' and *siwieć* 'to turn grey'. However, it appears plausible to restate such a morphological restriction in terms of syntactico-semantic constraints discussed in the present paper.

Weaker correlation of the results of those tests has been noted in the case of verb groups C, E, F and H in Table 1. Group C consists of verbs denoting inherently directed motion (*upaść* 'to fall', *zbiec* 'to escape'). Verbs of appearance and occurrence, such as *powstać/zaistnieć* 'to come into being' and *wyniknąć* 'to ensue', are included in group E. Group F contains verbs of disappearance, such as *wygasnąć* 'to expire' and *zniknąć* 'to disappear'. Verbs of assuming position, e.g., *przylec* 'to come to adjoin' and *zawisnąć* 'to hang down', are referred to in Table 1 as group H. The derivation of resultative adjectives from those groups of verbs involves cases of unpredictable lexical gaps, such as *znikły* (arch.) 'that has disappeared' and *wyszły* (arch.) 'that has left', or of forms with non-compositional semantics (*upadły* 'fallen; immoral' or *zawisły* 'dependent'). Moreover, the *-no/-to* construction with verbs belonging to those groups can be marginally acceptable for some speakers, as is the case with the verb *upaść* 'to fall' in (16d) in Section 4.2. above.

No correlation of the results of the diagnostics for unaccusativity proposed here can be observed for verb classes B, D and G in Table 1 since those verbs fail the resultative *-ły* adjective test. Verbs participating in the transitive-inchoative alternation (group B), e.g., *złamać się* 'to break (oneself); to get broken', have related *-ny/-ty* adjectives which allow both passive and non-passive interpretation. Such verbs are cross-linguistically found to be core examples of unaccusatives (e.g., in Levin and Rappaport 1995; Hale and Keyser 1993). The prediction concerning the unaccusativity of verbs of existence (class D) and verbs of maintaining position (class G) can be made again on the basis of cross-linguistic evidence, though here presumably resort could be taken as well to diagnostics of surface unaccusativity mentioned in Section 1 (e.g., distributive *po-* phrases, unmarked word order or the genitive of negation).

Table 1. Correlation of the Results of the Two Unaccusativity Diagnostics for Polish Verbs

Verb type	Do *-ły* adj. exist?	Is *-no/ -to* felicitous?	Correlation between the tests.[13] Verb status
A. nonalternating change of state verbs: *zachrypnąć* 'to grow hoarse'	+	-	++; telic, unacc.
B. alternating change-of-state verbs: *złamać (się)* 'to break (oneself)'	N/A[14]	-	-; telic, presum. unacc.
C. inherently directed motion verbs: *upaść* 'to fall'	?+[15]	?[16]	+; telic, unacc.
D. verbs of existence: *być* 'to be'	-	-	-: atelic, presum. unacc.
E. verbs of appearance and occurrence: *zaistnieć* ' to come into being'	?+	?	+; telic, unacc.
F. verbs of disappearance: *zniknąć* 'to disappear'	?+	?	+; telic, unacc.
G. verbs of maintaining position: *leżeć* 'to lie'.	-[17]	?	-; atelic, presum. unacc.
H. verbs of assuming position: *zawisnąć* 'to hang down'	?+	?	+; presum. telic and unacc.

[13] The symbol '++' is used to mark strong correlation, '-' signals no correlation and '+' weak correlation between the tests employed.

[14] Resultative adjectives are formed from such verbs with the suffix *-n/-t* .

[15] The symbol '?+' in this column shows that *-ły* adjectives from some of those verbs are lexicalized or not attested.

[16] The symbol '?' in this column indicates that for some speakers the *-no/-to* construction is marginally acceptable (with human participants).

[17] There exist some lexicalized *-ły* adjectives related to such verbs.

5. Conclusion

In this paper I argued that the unaccusative/unergative distinction is relevant for intransitive/reflexive verbs in Polish. I proposed that the existence of resultative adjectives terminating in *-ły* is a convenient deep unaccusativity diagnostic in Polish. There is a correlation between the occurrence of resultative adjectives terminating in *-ły* and the nonoccurrence of related verbs in the impersonal *-no/-to* construction, which can be interpreted as showing that both tests are sensitive to the unergative/unaccusative status of verbs in Polish. Differences between the predictions of both unaccusativity diagnostics are due to additional morphological and semantic restrictions imposed on the formation of resultative *-ły* adjectives and on the impersonal *-no/-to* construction.

References

Alexiadou, A. and E. Anagnostopoulou. 1998. "Unaccusativity Mismatches in Greek." Ms., ZAS, Berlin.

Babby, L. 1998. "Voice and Diathesis in Slavic." Position Paper for Comparative Slavic Morphosyntax workshop. Spencer, Indiana, June 1998. Available on-line at: <http://www.indiana.edu/~slavconf/linguistics/index.html>

Babyonyshev, M. 1996. *Structural Connections in Syntax and Processing: Studies in Russian and Japanese.* Ph.D. dissertation, MIT, Cambridge, Mass.

Bajerowa, I. 1992. "Polski język ogólny XIX wieku. Stan i ewolucja. Tom II: Fleksja". Katowice: Uniwersytet Śląski.

Bartnicka, B. 1970. "Adiektywizacja imiesłowów w języku polskim". Warszawa: PWN.

Burzio, L. 1986. "Italian Syntax. A Government-Binding Approach". Dordrecht: Reidel.

Dixon, R.M. 1994. "Ergativity". Cambridge: Cambridge University Press.

Doroszewski, W. 1980. "Słownik poprawnej polszczyzny". Warszawa: PWN.

Dziwirek, K. 1994. "Polish Subjects". New York and London: Garland Publishing Inc.

Grzegorczykowa, R. 1979/1982. "Zarys słowotwórstwa polskiego. Słowotwórstwo opisowe". Warszawa: PWN.

Hale, K. and S. J. Keyser. 1993. "On Argument Structure and the Lexical Expression of Syntactic Relations". In Hale, K. and S.J. Keyser (eds.) *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, 53-109. Cambridge, Mass.: MIT Press.

Hoekstra, T. 1984. "Transitivity". Dordrecht: Foris.

Lavine, J. E. 1999. "Subject Properties and Ergativity in North Russian and Lithuanian". In Dziwirek, K., H. Coats and C.M. Vakareliyska (eds.) *Annual Workshop on Formal Approaches to Slavic Linguistics. The Seattle Meeting 1998,* 307-328. Ann Arbor: Michigan Slavic Publications.

Levin, B. and M. Rappaport Hovav. 1995. "Unaccusativity. At the Syntax-Lexical Semantics Interface". Cambridge, Mass.: MIT Press.

Mulder, R. 1992. *The Aspectual Nature of Syntactic Complementation*. Ph.D. dissertation, HIL, Leiden.

Perlmutter, D. 1978. "Impersonal Passives and the Unaccusative Hypothesis". In Jaeger, J.J. (ed.) *Proceedings of the Fourth Annual Meeting of the Berkeley Linguistic Society*, 159-189. University of California at Berkeley.

Pesetsky, D. 1982. *Paths and Categories*. Ph.D. dissertation, MIT, Cambridge , Mass.

Puzynina, J. 1993. "Die Sogenannten 'Unbestimmt-Persönlichen' Formen in der Polnischen Sprache". In Hentschel, G. and R. Laskowski (eds.) *Studies in Polish Morphology and Syntax*, 31-62. München: Verlag Otto Sagner.

Rosen, C. 1984. "The Interface between Semantic Roles and Initial Grammatical Relations". In Perlmutter, D.M. and C. Rosen (eds.) *Studies in Relational Grammar 2*, 38-77. Chicago, Ill.: University of Chicago Press.

Rothstein, R. 1970. "The Polish Verb". Ms, University of Massachusetts, Amherst.

Rozwadowska, B. 1992. "Thematic Constraints on Selected Constructions in English and Polish". Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.
Schoorlemmer, M. 1995. *Participial Passive and Aspect in Russian.* Ph.D. dissertation, OTS, Utrecht.
Siewierska, A. 1988. "The Passive in Slavic". In Shibatani, M. (ed.) *Typological Studies in Language 16: Passive and Voice,* 243-289. Amsterdam: John Benjamins.
Śpiewak, G. and I. Szymańska. 1997. "Polish Impersonal Constructions: an Exercise in Head Formation and Argument Licensing". In Junghanns, U. and G. Zybatow (eds.) *Formale Slavistik*, 145-154. Frankfurt am Main: Vervuert Verlag.
van der Putten, F. 1997. *Matter and Mind in Morphology. Syntactic and Lexical Deverbal Morphology in Dutch.* Ph.D. dissertation, HIL, Leiden.
Wilczewska, K. 1966. "Czasowniki zwrotne we współczesnej polszczyźnie". Toruń: Uniwersytet Mikołaja Kopernika.
Wolińska, O. 1978. "Konstrukcje bezmianownikowe we współczesnej polszczyźnie." Katowice: Uniwersytet Śląski.
Zaenen, A. 1993. "Unaccusativity in Dutch: Integrating Syntax and Lexical Semantics". In Pustejovsky, J. (ed.) *Semantics and the Lexicon*, 129-161. Dordrecht: Kluwer Academic Publishers.

University of Silesia
Department of English
ul. Żytnia 10
41-205 Sosnowiec
Poland
cetnarow@uranos.cto.us.edu.pl

A Comparative Computational Evaluation of Three Theories of Russian Stress

Peter Chew
Oxford University

1. Introduction

Halle (1997), Melvold (1989) and Zaliznjak (1985) all propose derivational theories of stress assignment: all three theories apply to modern Russian, and Halle (1997) claims that the principles he advances also hold for other Indo-European languages. The aim of this paper is to show computationally that Halle's and Melvold's theories are preferable to Zaliznjak's on the criteria of simplicity, scope, and, to a lesser extent, coverage; and on the criterion of scope alone, Melvold's theory is optimal. I also show that derivational theories such as Melvold's can easily be reformulated as declarative phrase-structure grammars.

All three of these theories are 'morpheme-based', which means that they assume that each morpheme of Russian has its own accentual specification, and stress assignment results from applying a set of rules referring to accentual specifications of a word's constituent morphemes. The view that stress in Russian is at least partially conditioned by morphology is widely accepted; in one way or another this view also underlies Garde (1980), Lagerberg (1992) and Brown, Corbett, Fraser, Hippisley and Timberlake (1996). On the other hand, the view that individual morphemes must be accentually categorized is less widely accepted. In general, however, a theory which is based upon smaller linguistic units will be more economical than one which is based upon larger units, because the lexicon required will be smaller — so, for example, a theory of stress which categorizes morphemes accentually will be preferable to one which categorizes stems or words (both of which

may be polymorphemic). There are no theories which claim to assign stress in Russian on the basis of linguistic units smaller than morphemes (e.g., phonemes).

This article is organized as follows. I begin in Section 2 with a detailed look at the three theories. Section 3 is an examination of the factors which influence the constrainedness of the theories, and presents empirical data allowing the theories to be directly compared. Section 4 then presents experimental evidence on the respective coverage of the theories. The evidence suggests that of the three theories, Melvold's is to be preferred. Section 5 describes how Melvold's theory could be re-implemented in the framework of context-free grammar, and finally Section 6 outlines the theoretical implications of the results presented in this paper.

2. A Sketch of the Three Theories

2.1. Accentual Categories

Halle (1997:278) states that morphemes can be either 'accented' or 'unaccented'. The accent of an 'accented' morpheme is either borne by one of the morpheme's constituent vowel segments, or it follows all the morpheme's constituent segments, in which case the morpheme is 'post-accented'. This means that 'post-accented' is a subcategory of 'accented'. As far as the categorization of Russian morphemes is concerned, Melvold's analysis is exactly the same as Halle's, except that she introduces the additional lexical distinction of 'dominant' versus 'recessive' which only applies to suffixes and is independent of the accented/unaccented distinction. Zaliznjak's analysis, like Melvold's, incorporates the same dominant/recessive distinction (except this time the distinction is not restricted to suffixes), but a new accentual distinction in Zaliznjak's theory is between morphemes marked ↓ (accented), → (post-accented) and

← (pre-accented). Again, the two distinctions are independent of one another.

2.2. Stress Assignment Rules

In both Zaliznjak's and Melvold's theory, stress is determined by the rightmost dominant morpheme if there is one. If there are no dominant morphemes, then stress is determined by the leftmost accented morpheme according to Melvold, or by the leftmost morpheme marked ← or ↓ according to Zaliznjak. Failing this, stress falls on the leftmost syllable of the word in Melvold's theory, or is determined by the rightmost morpheme (which must be marked →) in the case of Zaliznjak's theory.

Apart from these (perhaps minor) differences between Zaliznjak's and Melvold's theories, there are two other important differences. First, in Melvold's theory, stress assignment is cyclic: this means that stress is first assigned to simple stems, then re-assigned within more complex stems. Secondly, Zaliznjak alone distinguishes between the two concepts of stress being 'attached' to certain morphemes, and stress being 'determined' by particular morphemes. Specifically, according to Zaliznjak, if the morpheme which 'determines' stress is classified →, then stress is 'attached' to the following morpheme; if the morpheme which 'determines' stress is classified ←, then stress is 'attached' to the preceding morpheme. (Although Zaliznjak does not say so, we may assume that if the morpheme which 'determines' stress is classified ↓, then stress is 'attached' to that morpheme.) Further, if the morpheme to which stress is 'attached' is classified ↓, then stress falls on the vowel which bears the accent; if the morpheme to which stress is 'attached' is classified ←, then stress falls on the syllable immediately to the left of the morpheme; if the morpheme to which stress

is 'attached' is classified →, the stress falls on the syllable immediately to the right of the morpheme.

The stress assignment rules in Halle's theory are the simplest of the three theories: stress is determined by the leftmost accented morpheme if there is one, or falls on the initial syllable of the word otherwise; there is no cyclicity, and stress is only 'determined', not 'attached'.

2.3. Examples.

Table 1 gives some examples of word-derivations in accordance with Halle's rules:

Table 1. Derivations in Accordance with Halle (1997)

Word	górodu	koroljú
Meaning	'town.DAT.SG'	'king.NOM.SG'
Underlying form[1]	/gorod+u/	/korolʲˈ+u/
Surface form	/gˈorod+u/	/korolʲ+ˈu/

Table 2, by contrast, gives a derivation in accordance with Melvold's rules. As Melvold (1989:48) notes, the cyclicity of the rules is the device which guarantees that the stress of derived words (words which include a derivational suffix) will, in most cases, be on the same syllable throughout the paradigm (see also Zaliznjak 1985:30, Lagerberg 1992:11).

[1] Accentuation is represented by the symbol ˈ ; where this symbol immediately precedes a segment, that segment is accent-bearing; where it is placed in morpheme-final position, the morpheme is post-accented.

Table 2. Derivations in Accordance with Melvold (1989)

Word	universitétu	universitétami
Meaning	'university.DAT.SG'	'university.INST.PL'
Underlying form[2]	/unʲivʲˈɛrsʲ+*itʲ*ɛt+u/	/unʲivʲˈɛrsʲ+*itʲ*ɛt+ˈamʲi/
Surface form	/unʲivʲɛrsʲitʲˈɛt+u/	/unʲivʲɛrsʲitʲˈɛt+amʲi/

For Zaliznjak's theory we shall consider just one example which highlights the importance of the difference between stress being 'determined' by a morpheme, and stress being 'attached' to a morpheme. Zaliznjak claims that the explanation for the stress in *čertjóžik* 'small sketch' is that the accentuation of the word's constituent morphemes is as shown in (1):

(1) čertj ož ik
→ →D ←

Because */ož/* is the rightmost dominant morpheme, it 'determines' the stress placement. But because it is also post-accented, stress is 'attached' to the following morpheme. According to Halle and Melvold, */ik/* would now be stressed regardless of its accentuation. But in Zaliznjak's theory, stress surfaces on the syllable to the *left* of */ik/*, since */ik/* is pre-accented. The consequence of this difference is that Zaliznjak's theory is considerably less constrained than either of the other two theories, since it effectively has more ways of explaining a given surface stress. I shall return to this point in the next section.

3. Consistency and Constrainedness of the Three Theories

The often unspoken assumption of derivational grammar theories is that underlying forms are known *a priori*. No doubt this is

[2] Symbols are as in Table 1; additionally, dominant suffixes are italicized.

because the process of compiling by hand a comprehensive list of underlying forms is, for a realistically sized lexicon, impracticable. Yet if a list of underlying representations does not accompany the rules, the grammar can never be properly tested: the grammar is under-determined.

All three theories have this shortcoming: stress assignment rules are discussed in some detail, but the underlying forms of comparatively few morphemes are specified. Thus, these theories *cannot* be tested on a large corpus as they stand: the question of how well they fit the data on a large scale therefore remains open.

Since I was interested in obtaining an empirical answer to the question of the respective coverage of different theories of Russian stress assignment, I had somehow to obtain, for each theory and dataset to be tested, the underlying form for each morpheme represented in the given dataset — in other words, to build up a lexicon of fully-specified underlying forms which best fits both the theory and the data. In general, the problem is that the surface stress of a given word may be explained by a given theory in many different ways: for example, without knowledge of the accentuations of */čertj/*, */ož/*, and */ik/*, the stress of *čertjožik* could be explained by */ož/* being dominant ↓, or by */ik/* being dominant ←.

By implementing the rules of each grammar computationally and applying them 'backwards', I found, for each theory, how many different accentuations words in a given dataset could have such that the correct surface stresses were derived. For example, it turned out that the stress of *čertjóžik* could be explained in 324 ways in Zaliznjak's theory, and in only 63 ways in Melvold's theory. This gave some measure of the extent to which Zaliznjak's theory is less constrained: the difference is largely attributable to the fact that his theory incorporates the more complex mechanism identified in the previous section.

An interesting feature of morpheme-based theories of Russian stress is that a list of words will constrain the underlying forms of morphemes to a greater extent than a single word. For example, in Halle's theory the surface stress of *rúku* can be explained, essentially, in two ways: either the root */ruk/* is accented, or both */ruk/* and */u/* are unaccented. The stress of *ruká* can also be explained in two ways: either */ruk/* is post-accented, or */ruk/* is unaccented and */a/* is accented. If we take these examples in conjunction with one another, however, logic dictates that */ruk/* must be unaccented, as this is the only accentuation which is possible in both cases. In other words, the alternative explanations can be ruled out.

A computational algorithm (henceforth referred to as the 'elimination algorithm' or EA) was employed to apply logic in this way based on the individual constraints of each of the three theories. Clearly, the more words are considered, the more can be deduced about the underlying forms of morphemes in the lexicon.[3] It was hoped that by considering a large enough corpus of words in this way, it would be possible to arrive at a single underlying form for each morpheme, even though initially the size of the solution space appeared to be very large.

Unfortunately, this was not the case for any of the three theories. In the case of Halle's and Melvold's theories, eliminating inconsistent underlying forms in this way leads to the absurd conclusion that some words have *no* possible underlying forms. This means that the constraints of the theories are unsatisfiable. For example, consider the following words:

---

[3] Applying this kind of logic on a large scale is impracticable unless done computationally, although I shall not go into the precise details of the algorithm here.

(2) a. sociál'nij 'social' /sots+ij+ˈalʲ+n+ij/[4]

b. socialistíčeskij 'socialist' /sots+ij+alʲ+istʲ+ˈitʃʲ+ɛsk+ij/

For both of these words the three initial morphemes are the same. Based on (2a) one must conclude, using Halle's theory, that one of the following propositions is true:

(3) a. /alʲ/ is accented

b. /ij/ is post-accented

However, none of these possibilities could be consistent with (2b). Clearly, the existence of inconsistencies such as this shows that the rules of the theory are not sufficient to account for 100% of the data, no matter what accentual specification is chosen for the lexicon. Yet there would be problem words even for a theory with a coverage of 99.9%, and one would surely not wish to reject such a theory. This point is made in Chomsky and Halle (1968:ix):

> We see no reason to give up rules of great generality because they are not of even greater generality, to sacrifice generality where it can be attained. It seems hardly necessary to stress that if we are faced with the choice between a grammar G_1 that contains a general rule along with certain special rules governing exceptions and a grammar G_2 that gives up the general rule and lists everything as an exception, then we will prefer G_1.

Unfortunately, it is not possible to isolate a list of exceptions to Halle's (or any other) theory using the EA. The reason for this is that in reality it is *combinations* of words (such as *sociál'nij*, *socialistíčeskij*), rather than *individual* words, which can cause problems for a theory. In other words, we do not know from the above whether it is *social'nij* or *socialističeskij* which is the exception to the theory (it could also be that both words are except-

[4] Morphological tokenization follows Oliverius (1976).

ions). For this reason, logical inference is of little use in identifying a comprehensive list of exceptions to a given derivational theory: all we have been able to do is prove that there must be exceptions to that theory. However, in Section 4 I shall present other experimental evidence which does answer the question 'what is the exact coverage of Halle's and Melvold's theories, and what are the exceptions?'.

Applying the EA to Melvold's theory yields the result that this theory, too, must have exceptions. This can be seen from the chain of inferences in Table 3.

Table 3. Inferences Showing Melvold's Theory Has Exceptions

Word	Morphological tokenization, following Oliverius (1976)	Inferences
načálo 'beginning'	/na+tʃˈa+l+o/	/l/ is not dominant[5]
právilo 'rule'	/prˈavʲ+i+l+o/	If /i/ is dominant, it is also unaccented
rodítel' 'parent'	/rodʲ+ˈi+tʲɛlʲ+/	/i/ is not dominant unaccented, so /rodʲ/ is post-accented
ródina 'homeland'	/rˈodʲ+in+a/	/in/ is dominant unaccented
glubiná 'depth'	/glubʲ+in+ˈa/	No possible analysis given previous inferences

[5] If /l/ were dominant post-accented, the stress would be *načaló*; if it were dominant unaccented the stress would be *náčalo*; and it cannot be dominant accented as it does not contain a vowel.

When the EA is applied to Zaliznjak's theory, it turns out that there are no instances of this kind. This represents a lack of evidence that Zaliznjak's theory has exceptions rather than positive evidence that the theory is exceptionless. However, a different problem emerges: this becomes clear if we consider how many morphemes in the lexicon have how many candidate accentuations before and after the EA is run. The figures are shown in Table 4.

Table 4. Data Relating to Constrainedness of Zaliznjak's Theory

No. of candidate accentuations	No. of morphemes before EA is run	No. of morphemes after EA is run
x	n_x	n_x
0	0	0
1	0	15
2	0	62
3	0	44
4	72	119
5	0	90
6	463	237
7	0	0
8	54	24
9	0	0
10	3	1
Total	592	592

If we define the 'lexical solution space' (LSS) for a given theory as the number of distinct ways in which the lexicon[6] may be accentually specified within the constraints of a given theory, then for a lexicon where the number of morphemes with x underlying forms is n_x, the size S of the overall LSS will be given by (4):

[6] In this case, the 'lexicon' consists of 592 morphemes.

$$(4) \qquad S = 0^{n_0} \times 1^{n_1} \times \ldots \times x^{n_x}$$

Thus, for Zaliznjak's theory the value of S after the EA is run will be given by the expression in (5):

$$(5) \qquad 1^{15} \times 2^{62} \times 3^{44} \times 4^{119} \times 5^{90} \times 6^{237} \times 8^{24} \times 10^{1} \approx 2 \times 10^{390}$$

For comparison, the value of S before the EA is run (calculated from Table 4 in the same way) is approximately 3×10^{455}. In other words, the EA reduces the size of the LSS by a factor of approximately 10^{65} — a substantial reduction, but not enough to prevent the size of the final LSS from being substantially more than the number of atoms in the universe, let alone the number of bytes of memory space available in the universe. Clearly, this is a far cry from the hoped-for scenario where the number of possible underlying forms per morpheme is reduced to one. One is led to conclude that the proposition that Zaliznjak's theory accounts for Russian stress without exceptions is computationally unverifiable.

4. Tests to Ascertain the Coverage of the Three Theories

In the previous section, we considered the results of testing the theories for consistency and, in the case of Zaliznjak's theory, constrainedness. We now turn to a different type of computational test, which aims to compare the possible coverage of the three theories, that is, the percentage of cases in which stress is correctly assigned on the basis of the rules and a fixed lexicon of accentually-specified morphemes.

4.1. Method of Ensuring that the Lexicon is Uniquely Specified

It should be reiterated that a test of this type cannot function unless the size of the LSS is 1 (in other words, each morpheme in the lexicon has exactly one underlying form). It was shown in the

previous section that logical inference fails to reduce the LSS size to 1, so to obtain a lexicon with a reasonable accentual specification for the test, I had to use a procedure which is partly linguistic and partly statistical. Briefly, the procedure was as follows.

For each morpheme, all but one candidate underlying form were eliminated: in each case, the remaining one was the one which might most reasonably be considered to be the 'default', in accordance with Kiparsky's (1973) 'alternation condition'. I was then in a position to test the given theory, including the accompanying lexicon, to determine its coverage (which may well be suboptimal). The next step was to inspect the list of exceptions to the theory to determine whether there were morphemes which occurred repeatedly in incorrectly stressed words. Where it was suspected that the incorrect stress of these words was attributable to an incorrect accentual specification of the morpheme common to them, the common morpheme's accentual specification was overridden. Each such modification of the lexicon meant that some words in the corpus which were previously incorrectly stressed now became correctly stressed — although in some cases some other words which were previously correctly stressed became incorrectly stressed. However, modifications of this type were only accepted where the number of words in the former category exceeded the number in the latter (in other words, where the modification resulted in an increase in coverage). A point was eventually reached where no further increases in coverage were possible: at this point, the coverage of the theory was recorded. In each case, the figure recorded was probably less than optimal, but was significantly better than the initial coverage based on the 'default' accentuations.

This method is of no great interest in itself, since it is based on a continuous reanalysis of previous failures. This notwithstanding, the accentual specification obtained for the lexicon should stand or

fall on its own linguistic merits, one of which is presumably the extent to which words are assigned correct stresses when the lexicon is used in conjunction with a particular theory of stress assignment.

4.2. The Algorithm Used to Test Coverage

With a uniquely specified lexicon in place (determined individually for each of the three theories in accordance with the procedure in Section 4.1), it is now possible to test the theories' respective coverages.

For each theory, an appropriate corpus of words was chosen as the sample for the test. What was 'appropriate' was determined by reference to the scope of the theory or theories being tested: thus, for the comparison of Zaliznjak's theory and Melvold's theory, only nominative singular derived nouns were used as the sample. This is because Zaliznjak states that his theory applies only to derived non-inflected words, and the relevant part of Melvold's theory applies to nouns but not other parts of speech.

A stress checking algorithm (SCA) was then run. This algorithm involved the straightforward subprocedures in (6):

(6) The Stress Checking Algorithm (SCA)

a. For each morphologically-tokenized stressed word, remove the stress and look up the underlying accentuations of the constituent morphemes in the lexicon.
b. Assign stress to the word (how this is done depends on which theory is being tested).
c. Compare the stress computed in (b) with the actual stress from (a). Keep count of the number of words correctly stressed and the number of words incorrectly stressed.

The SCA was run six times, each time using either a different sample or a different theory. The samples were chosen from a source computer file consisting of words and morphological tokenizations from Oliverius (1976), which lists 2,493 frequently occurring headwords of Russian. This file was augmented by matching headwords from Oliverius (1976) to nominative singular nouns from Brown, Corbett, Fraser, Hippisley and Timberlake's (1996) theorem dump (835 nouns were common to both sources) in order to extrapolate the morphological tokenization of 8,798 inflected forms. As a result, the total size of the source file was 11,291 words.

4.3. Analysis of Results

Table 5 shows the quantitative results obtained by running the SCA as described in the previous section. First and foremost, these results show that Russian word-stress can in the overwhelming majority of cases be determined by looking at their morphological composition. In particular, the fact that Halle's theory accounts correctly for almost 98% of non-derived nouns (see row 1 of Table 5) disproves Zaliznjak's (1985) claims that 'as far as non-derived

words are concerned, the only way of [predicting stress] is to list [the stress-pattern of every paradigm]' (p. 22), and that 'the stress-pattern and placement of stress within the stem for derived words differs from that of non-derived words in that it can be determined from a word's morphological composition' (p. 29) [my translations — PC].

Further, comparison of rows 1 and 2 shows that Halle's theory does a little worse at assigning stress to derived nouns than to non-derived nouns. However, incorporating cyclicity and the dominant-recessive distinction into the theory (as Melvold does) remedies this problem, bringing the theory's coverage back up to almost 96% of all nouns (row 3 of Table 5). Melvold's theory, as outlined in 1.2.2, appears to fit parts of speech other than nouns less well; this can be attributed largely to the fact that the theory does not account adequately for verbs, since many of the 771 exceptions are verbs.[7]

[7] The difficulty can be explained as follows. The verb-forming suffixes /i/ and /a/ (as identified by Oliverius 1976) are frequently found in nouns and are almost always stressed (e.g., *pred+sed+á+tel'+* 'president', *uč+í+l+išč+e* 'college'). This suggests that these suffixes are dominant accented. However, numerous infinitives containing these suffixes are stressed on a syllable before the suffix (e.g., *plák+a+t'* 'to cry', *u+dár+i+t'* 'to hit'), a fact which is inconsistent with the hypothesis that the suffixes are dominant accented. If Melvold's theory could be modified to account for cases such as these, the overall error rate for nouns, adjectives and verbs would be much more satisfactory.

Table 5. Results of Tests of Coverage

	Theory	Sample description	Correct stresses	Incorrect stresses	Sample size
1	Halle	Non-derived[8] nouns	3,163	71	3,234
			97.8%	2.2%	100.0%
2		All nouns	8,416	1,105	9,521
			88.4%	11.6%	100.0%
3	Melvold	All nouns	9,126	395	9,521
			95.9%	4.1%	100.0%
4		Nouns, verbs and adjectives	9,914	771	10,685
			92.8%	7.2%	100.0%
5	Zaliznjak	Nom sg derived nouns	482	42	524
			92.0%	8.0%	100.0%
6	Melvold	Nom sg derived nouns	498	26	524
			95.0%	5.0%	100.0%

Finally, rows 5 and 6 of Table 5 provide an interesting comparison of the ability of Melvold's and Zaliznjak's theories to account for exactly the same dataset. The coverage of Melvold's theory is slightly higher, and Melvold's theory is simpler than that of Zaliznjak. Furthermore, the scope of Melvold's theory is greater, because it accounts for non-derived *and* derived nouns in all six grammatical cases, whereas Zaliznjak's is restricted to derived nouns in the nominative singular only. In other words, Melvold's theory not only accounts for a greater percentage of cases than Zaliznjak's on a like-for-like basis, but it does so at less computational cost and it also accounts for words which are completely outside the scope of Zaliznjak's theory.

[8] A 'non-derived' noun is defined as any noun which consists morphologically only of a root and an inflectional ending.

5. Melvold's Theory as a Context-Free Grammar

With strong evidence that, of the three theories, Melvold's is to be preferred, I shall show in this section how the theory can be reformulated as a context-free phrase-structure grammar.

Context-free grammars are one of a number of types of phrase-structure grammar classified by Chomsky (1959) into what is now known as the Chomsky Hierarchy. More powerful grammars within this hierarchy are capable of describing a wider range of languages, but this can be a disadvantage: after all, Occam's razor would presumably lead one to use the most restrictive grammar possible. Additionally, grammars which are no more powerful than context-free have computational advantages: first, they are Turing-decidable, which means they can always accept or reject strings in a finite amount of time; secondly, the ordering of context-free rules will not affect the way in which they function or their end result; and finally, context-free rules can be used both 'forwards' and 'backwards' (for generation or parsing), so a context-free grammar can be said to be an equally valid model both of linguistic perception and production.

We saw in Section 2 that in Melvold's theory word-stress is ultimately determined by one, and only one, morpheme — for example, the rightmost dominant morpheme. This idea can be very straightforwardly incorporated into phrase-structure morphology. As an example, consider the word *spec+ij+ál+'n+ost'* 'speciality'. Instead of being represented as a string of morphemes, this word could also be represented as a tree, assuming we have a context-free grammar of Russian morphology. For example, given the grammar in (7), the representation of *spec+ij+ál+'n+ost'* would be as in Figure 1:

(7)

noun	→	noun_stem, infl_ending
noun_stem	→	adj_stem, noun_suffix
noun_stem	→	verb_stem, noun_suffix
adj_stem	→	adj_stem, adj_suffix
adj_stem	→	noun_stem, adj_suffix

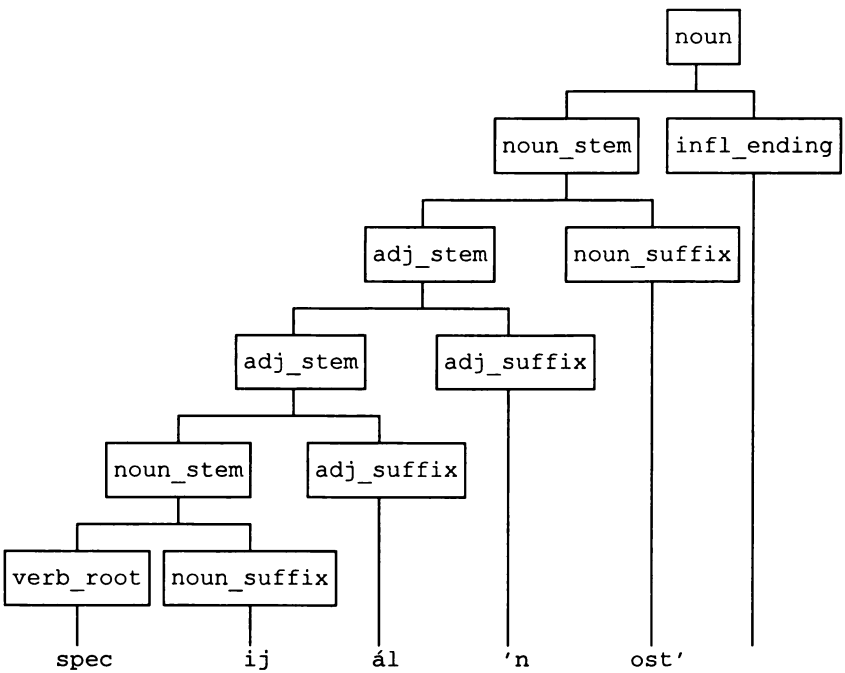


Figure 1. Parse Tree for spec+ij+ál+'n+ost' 'speciality'

Now suppose that the stress of this word is explained by the fact that */al/* is a dominant accented morpheme. Since 'dominant accented' is a value which is lexically specified, this value must be carried by the grammar's terminal symbols (that is, the leaves of the tree). Now suppose we assign some binary feature to each node in the tree, both terminal and non-terminal nodes: we shall call this

feature [±stress_det(ermining)]. By stipulating that the start symbol of the morphology grammar is [+stress_det] and enriching the morphology grammar in (7) appropriately, it is possible to ensure that the value of [+stress_det] is inherited from the start symbol by exactly one of the terminal symbols. Further, by referring to the lexical accentuation of the terminal symbols, we can ensure that the terminal symbol which inherits [+stress_det] is always the rightmost dominant morpheme. The feature [+stress_det] is distinct from the features [±accented] and [±dominant]; the latter two are lexically determined, and it is possible that a word may have more or less than one [±accented] and/or [±dominant] morpheme. The feature [±stress_det] is also distinct from [±stressed]: for example, where a post-accented morpheme determines word-stress, the morpheme which includes the [+stressed] vowel or syllable will not be the same as the morpheme which is [+stress_det]. The four features [±stress_det], [±stressed], [±accented] and [±dominant] are, however, all related by their distribution within the context-free rules.

Accordingly, to encapsulate Melvold's rule which assigns stress to the rightmost dominant morpheme in a word, the second rule from (7) is expanded into the rule schemata in (8) and (9) (each of which abbreviates two context-free rules).

(8)	noun_stem [αstress_det]	→	adj_stem, [αstress_det]	noun_suffix [–stress_det] [–dominant]
(9)	noun_stem [αstress_det]	→	adj_stem, [–stress_det]	noun_suffix [αstress_det] [+dominant]

For example, for *soc+ij+al+íz+m* 'socialism', which includes two dominant accented morphemes, */al/* and */iz/*, rules similar to those

in (8) and (9) would yield the structure in Figure 2 (italics denote those symbols which have the feature [+stress_det]). Note that the form of (9) ensures that no stem, regardless of where it is embedded in the overall tree structure, can ever be [+stress_det] if the suffix attached to it is dominant.

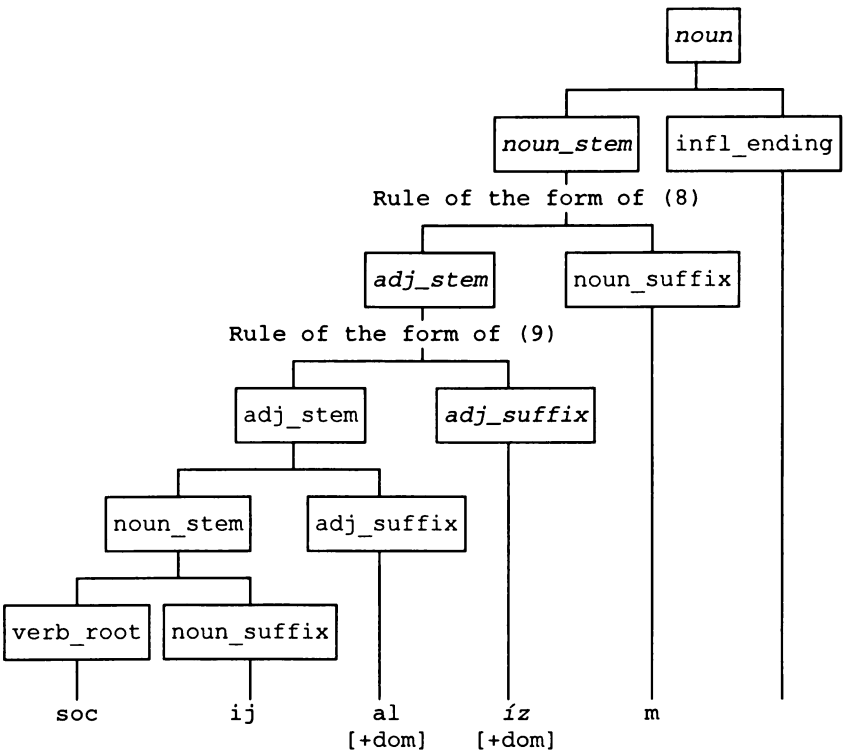


Figure 2. Parse Tree for soc+ij+al+íz+m 'socialism'

For non-derived nouns (where by definition there can be no dominant suffixes), the rules in (10)-(12) would be used to ensure that the leftmost accented morpheme (or the leftmost morpheme, if no morphemes are accented) inherits [+stress_det].

(10)	noun	→	noun_root [+stress_det] [+accented]	infl_ending [–stress_det]
(11)	noun	→	noun_root [–stress_det] [–accented]	infl_ending [+stress_det] [+accented]
(12)	noun	→	noun_root [+stress_det] [–accented]	infl_ending [–stress_head] [–accented]

Finally, Figure 3 exemplifies the use of rules like (10)-(12); it shows the structure of *stol+ám*, in which the root */stol/* is post-accented and the inflectional ending */am/* is accented. Crucially, stress surfaces on */am/* not because */am/* is accented, but because */stol/* is the leftmost (post-)accented morpheme; it therefore 'determines' the stress (in other words, the feature [+stress_det]).

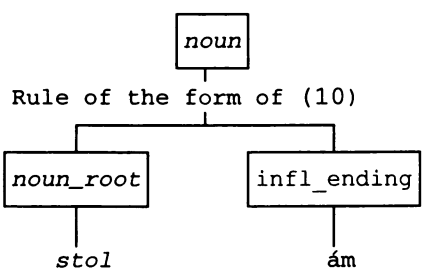


Figure 3. Parse Tree for stol+ám 'tables' (dat pl)

6. Summary

To conclude, the tests carried out suggest that of the three theories considered, Melvold's is to be preferred on the grounds that it has

a high coverage rate (95.8% of a dataset of 9,521 nouns), a broad scope (it is not restricted to non-derived nouns as Halle's theory is, or to nominative singular derived nouns as Zaliznjak's is), and is relatively simple (it is no more powerful than context-free, and involves fewer variables than Zaliznjak's theory).

The fact that Melvold's theory can successfully be implemented as a context-free grammar has theoretical importance: since context-free (but not necessarily context-sensitive) grammars are Turing-decidable, Melvold's theory is shown to be psychologically plausible, at least inasmuch as the ability to recognize and reject strings is arguably one of the linguistic faculties possessed by humans. Furthermore, context-free grammars also have a clear declarative interpretation: parse trees can be seen as linguistic structures, no part of which 'changes' during the course of a derivation. The fact that Melvold's grammar can be interpreted as a declarative grammar refutes Brown, Corbett, Fraser, Hippisley and Timberlake's (1996:59) implied claims that a rule-based grammar cannot be declarative, and that it is somehow harder for derivational grammars to combine prosody, phonology and morphology.

A less positive conclusion about Melvold's theory (which also applies to the other two theories) is that it omits a fully specified lexicon of morphemes. Of course, this information can be supplied, and a way in which this can be done was given in 4.1. Also omitted is a well-defined grammar of morphology, which is necessary if the theory is to be implemented as a practical means of assigning stress. It should be emphasized that both the accentual specification of the lexicon and the morphological grammar are as much a part of the theory as the stress assignment rules themselves, since the stress assignment rules are unverifiable otherwise.

Despite these omissions — and they are omissions, not faults — Melvold's theory appears to be a promising means of ac-

counting with a high level of accuracy for stress assignment in Russian.

References

Brown, D., G. Corbett, N. Fraser, A. Hippisley and A. Timberlake. 1996. "Russian Noun Stress and Network Morphology." *Linguistics* 34, 53-107.

Chomsky, N., and M. Halle. 1968. *The Sound Pattern of English.* New York: Harper & Row.

Garde, P. 1980. *Grammaire russe: I. Phonologie. Morphologie*, Paris: Institut d'Études Slaves.

Halle, M. 1997. "On Stress and Accent in Indo-European." *Language* 73, 275-313.

Kiparsky, P. 1973. "Phonological Representations." In O. Fujimura (ed.) *Three Dimensions of Linguistic Theory*, 1-136. Tokyo: TEC.

Lagerberg, R. 1992. *The Relationship between Stress and Affixation in Russian.* Ph. D. dissertation. School of Slavonic and East European Studies, University of London.

Melvold, J. 1989. *Structure and Stress in the Phonology of Russian.* Ph.D. dissertation. MIT, Cambridge, MA.

Oliverius, Z. 1976. *Morfemy Russkogo Jazyka: Častotnyj Slovar'.* Prague: Univerzita Karlova.

Zaliznjak, A. 1977. *Grammatičeskij Slovar' Russkogo Jazyka: Slovoizmenenie.* Moscow: Russkij jazyk.

Zaliznjak, A. 1985. *Ot Pravoslavjanskoj Akcentuacii k Russkoj.* Moscow: Nauka.

University of Oxford Phonetics Laboratory
41 Wellington Square
Oxford, OX1 2JF, U.K.
pchew@jesus.oxford.ac.uk

Formal Approaches to Slavic Linguistics[1]

Greville G. Corbett
University of Surrey, UK

1. Introduction

This paper is intended to celebrate the FASL series of conferences, by reflecting on the fine choice of title. I imagine we do *Linguistics* because we find language fascinating and believe that, as technical means of communication become ever more available, issues of the use of those channels, and specifically communication through language, will in turn grow in importance. For the value of the *Slavic* contribution to linguistics one need only think of Jakobson and Trubetzkoy. However, those who set up the series might have been tempted to focus it on Russian. While Russian has a dominant position, given its status as a world language and hence its role in educational institutions, it was so much better to have *Slavic Linguistics* as the subject. The linguistic interest of Slovene and Sorbian for instance is just as great as that of Russian. So to the more substantial issue, that of *Formal Approaches*. There is a variety of formal approaches which may be of benefit for Slavic linguistics. Some have been discussed at previous conferences, and the current collection is refreshingly diverse in this respect. I will outline three different approaches. In each I will report on joint work with several colleagues and highlight the importance of Slavic data for wider typological concerns. The three approaches involve morphology, lexical semantics and corpus linguistics. Since I wish to illustrate breadth I shall not be able to cover each approach in depth, rather I shall give illustrations with pointers to fuller accounts.

---

[1]The support of the ESRC UK (grant R000237845) is gratefully acknowledged. I also wish to thank my co-researchers. This draft was improved following suggestions by Andrew Hippisley, lively discussion at FASL 8, and helpful comments by the editors.

2. Network Morphology

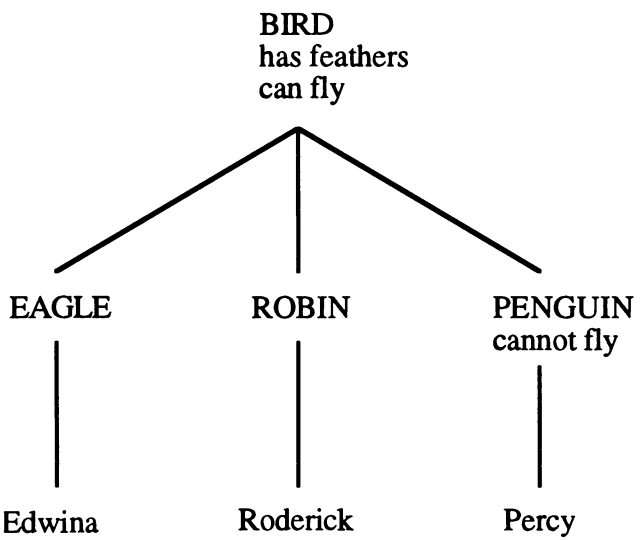
The first formal approach is Network Morphology. I will introduce some basic ideas and then consider two illustrative examples, gender assignment and syncretism. The research to be discussed also involves Norman Fraser and Dunstan Brown, and the account draws on previous publications. There is space here for a sketch; details are given in Corbett & Fraser (1993, 1997), Brown & Hippisley (1994), Fraser & Corbett (1995, 1997), Brown, Corbett, Fraser, Hippisley & Timberlake (1996), Hippisley (1997), Brown (1998).

Network Morphology is an approach to morphology which distributes information across a network in which generalizations can be optimally expressed. Generalizations become available in specific cases by the operation of default inheritance. Network Morphology theories are expressed in a formal representation language called DATR developed by Roger Evans and Gerald Gazdar. DATR is a particularly useful formalism for developing explicit accounts of complex linguistic data because it is formally well-defined (Evans & Gazdar 1989a) and it allows for the construction of largely declarative accounts which rely on a limited set of basic operations, of which default inheritance is one (Evans & Gazdar 1989b); it has been a major source of inspiration in the development of Network Morphology. Helpful introductions to DATR for linguists can be found in Evans & Gazdar (1996) and Gazdar (forthcoming). Added to the formal rigour and rich expressiveness of the DATR language is a third valuable feature: computer interpreters (and compilers) exist which are capable of taking a linguistic theory expressed in DATR as input and automatically generating as output all the forms which the theory allows. Working in this way means taking seriously some of the basic ideas of generative grammar. The computer thus has a valuable checking role: we are interested in *theoretical* linguistics, being concerned with observation, description and explanation rather than computational issues like algorithmic efficiency. We wish to separate the questions of linguistic theory (Network Morphology) from the formalism we use (DATR). This separation focuses

attention on the substance of theories rather than on their notation (Shieber 1987).

We should first consider the concept of default inheritance, which we approach using the taxonomic hierarchy in Figure 1. The lines in the taxonomy indicate instantiation rather than sub-classification. So an eagle is a bird, as is a robin and a penguin; Edwina is an eagle, Roderick is a robin, and Percy is a penguin.

Figure 1. A Simple Instantiation Hierarchy



Given an instantiation hierarchy of this kind, default inheritance allows all attributes of a given node in the hierarchy (such as BIRD) to be inherited by a node which instantiates it (such as EAGLE). This is the case except where the lower node already has a value for some attribute and thus overrides the default (that is, inheritable) value for that attribute. In Figure 1 a BIRD has feathers and can fly. These facts are inherited by EAGLE and ROBIN and, indirectly, by Edwina and Roderick. The attribute of having feathers is also inherited by PENGUIN and, through it, by Percy. However, specific information that PENGUINs cannot fly blocks inheritance

of the more general information about BIRDs. Although Percy is a BIRD, he cannot fly.

Default inheritance allows generalizations to be expressed once at a high level, and then automatically to apply to everything which inherits from there. Regularities, subregularities and exceptions can be represented easily and economically. This approach has the added advantage of marking exceptions as such, as in the case of PENGUIN in Figure 1. If Percy were a penguin who could fly, this extreme exceptionality would be visible because an exceptional fact (PERCY can fly) would override an exceptional fact (PENGUINs cannot fly), overriding a default (BIRDs can fly). Further information on default inheritance can be found in Gazdar (1987) and Daelemans, de Smedt & Gazdar (1992).

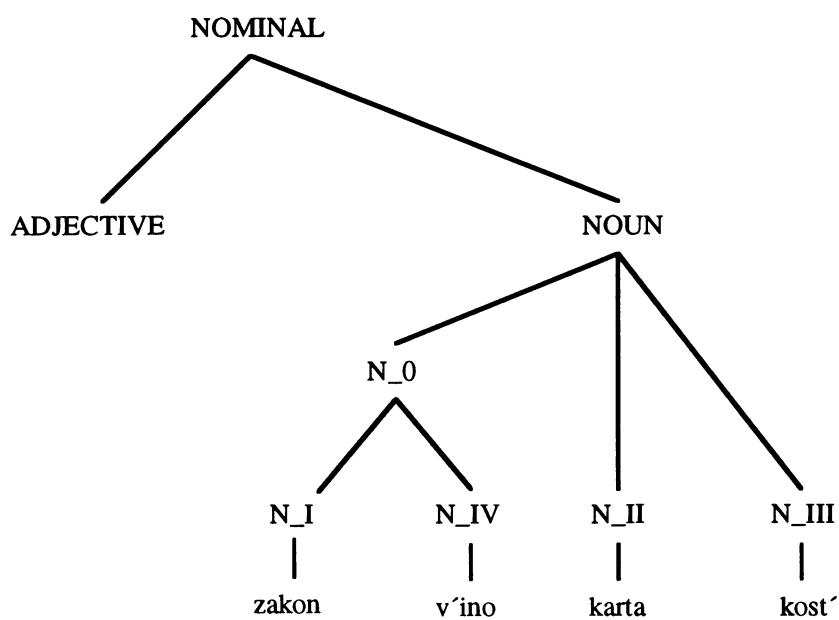
The information in Figure 1 could be expressed in DATR as follows:

(1)
```
BIRD:
    <has_feathers> == yes
    <can_fly> == yes.

EAGLE:
    <> == BIRD.

ROBIN:
    <> == BIRD.

PENGUIN:
    <> == BIRD
    <can_fly> == no.

Edwina:
    <> == EAGLE.

Roderick:
    <> == ROBIN.

Percy:
    <> == PENGUIN.
```

The labels preceding colons are 'nodes'; the angle bracket expressions to the left of the '==' symbol are 'paths'; the words to the right of non-empty paths are 'values'. Thus, the value of the <can_fly> path at the PENGUIN node is 'no'. The form '<>' is the special case in which a path is empty (hence maximally underspecified). This allows, for example, the node 'EAGLE' to inherit all values available at the node BIRD. Since we are dealing with default inheritance, PENGUIN inherits all values from BIRD, except the one which is overridden.

Of course, we are interested in the use of this kind of formalism for expressing linguistic generalizations. Figure 2 shows a simplified inheritance network for the morphology of Russian nominals.[2]

Figure 2. An Inheritance Structure for Russian Nominals



[2] Corbett & Fraser (1993): later papers use a network of hierarchies rather than a single one.

N_I to N_IV are nodes where information which distinguishes declensional classes is stored. Notice the posited node N_0 from which N_I and N_IV inherit; it allows us to capture the advantages of analyses of Russian which postulate three noun declensional classes and of those which distinguish four. The following (incomplete) fragment is taken from our earlier analysis (Russian forms in this section are given in phonemic transcription):

(2)
```
NOUN:
    <mor loc sg> == "<stem>" _e
    <mor nom pl> == "<stem>" _i.

N_III:
    <> == NOUN
    <mor loc sg> == "<mor dat sg>".

Kost´:
    <> == N_III
    <stem> == kost´.
```

The first fact at NOUN should be read as saying that the locative singular consists of the stem followed by an *-e* ending. A path enclosed in double quotes in a DATR sentence is used to retrieve the specified value for that path at the node from which the query originates. If we wanted to find the nominative plural of *Kost´*, we would inherit the sentence `<mor nom pl> == "<stem>" _i.` We would have to find out what the `<stem>` of `Kost´` is. Since the answer is *kost´*, the nominative plural is *kost´i*. The quoted path means that we take the stem of `Kost´` (not of NOUN, which has no stem). If, however, we wanted to know the locative singular of `Kost´`, we would never inherit the definition of locative singular at NOUN because it is overridden at N_III, from which *Kost´* inherits. The definition of locative singular at N_III establishes an asymmetric identity between the locative singular form of an N_III noun and its dative singular.

It may be desirable to inherit most information from one source, but to have access to some information stored elsewhere. Consider the following fragment:[3]

(3)
```
N_II:
     <> == NOUN
     <mor gen sg> == "<stem>" _i.

N_III:
     <> == NOUN
     <mor gen sg> == N_II.
```

This says that N_III may inherit its schema for forming the genitive singular from N_II, even though N_III (like N_II) inherits primarily from NOUN. This may be expressed more explicitly as follows:

(4)
```
N_III:
     <mor gen sg> == N_II:<mor gen sg>.
```

This was a brief introduction to default inheritance and to the DATR formalism. We now consider two linguistic problems and show how the formal approach sketched combines with a typological approach.

2.1. Gender Assignment

Gender systems have agreement as their defining characteristic. Nouns of a gender language can be grouped analytically according to agreement evidence. We then ask how the native speaker, who produces the agreement evidence, 'knows' the gender of the different nouns. Assignment to a particular gender is always

---

[3] Only facts relevant to the discussion are shown at nodes N_II and N_III in (3). In our full analysis of Russian nominal morphology, each of these nodes records a much richer set of facts. The facts for N_III are quite different from those for N_II, so an analysis of the form 'N_III: <> == N_II' would fail.

possible for the vast majority of nouns, from information required independently in the lexical entry (Corbett 1991:7-69). The particular type of information which may be used gives us a typology of assignment systems. We find *semantic systems* (where only semantic information is required) and *semantic + formal systems* (where semantic information is supplemented by morphological and/or phonological information). Purely formal systems (where gender would be predicted by formal means but where the different agreement classes of nouns would have no semantic significance) are not found.

Languages with semantic assignment and those with semantic rules supplemented by phonological rules are relatively unproblematic. The most difficult are the formal-morphological systems (see, for instance, Aronoff 1994:73-74), and this is precisely what is proposed for Russian, and Slavic more generally. These systems have often been analysed differently; instead of gender being predictable (and therefore not needing to be specified in the lexical entry), some treat gender as specified, and from it attempt to predict the morphological class of nouns. When the number of genders and the number of declensional classes are the same or nearly so, it is not immediately obvious which analysis is to be preferred. I propose that Russian has a gender assignment system in which morphological information supplements semantics (recall that in this section Russian examples are transcribed):

(5) *Semantic Assignment Rules*
(5a) Sex-differentiable nouns denoting males (humans and higher animals) are masculine: *sin* 'son', *d´ad´a* 'uncle', *lev* 'lion';
(5b) Sex-differentiable nouns denoting females are feminine: *doč´* 'daughter', *t´ot´a* 'aunt', *l´v´ica* 'lioness'.

Nouns which are sex-differentiable are those where the sex matters to humans (as for humans and domesticated animals) and where the difference is striking (as in the case of lions). There are few exceptions to these rules but many nouns are not covered by them. Unlike Godoberi, Russian does not treat all noun in the semantic

residue in the same fashion. They are subject to further rules, including the following:

(6) *Morphological Assignment Rules*
(6a) nouns of declensional class I (*zakon* 'law' type) are masculine;
(6b) nouns of declensional classes II (*karta* 'map') and III (*kost'* 'bone') are feminine;
(6c) others are neuter.

Given the dispute as to whether this is the right analysis, there are two traditional types of argument available here. First, and most important, there are language-specific arguments. It can be shown that predicting gender on the basis of declensional class is simpler and involves fewer exceptions than the attempt to predict declensional class on the basis of gender. These arguments are treated at length in Corbett (1982), and will not be repeated here. Second, there is the typological argument: since there are many languages where gender is straightforwardly predictable, it is simpler to claim that it is predictable in all languages, with typological variation being restricted to the type of information used for prediction.

By giving a Network Morphology analysis, using DATR as a tool, we have access to third type of argument: since DATR comes with a compiler, we can demonstrate that our analysis (which is an explicit account of the interactions of semantics, gender and declensional class) does indeed yield the correct results. Given our analysis and the lexical entries for a range of Russian nouns, a computer can be used to verify that our analysis makes the right predictions as to gender. Such an analysis is presented in Fraser & Corbett (1995). The aim of this section is not to justify that analysis. Rather we want to emphasize that the analysis, that of a theoretical linguist working within the Network Morphology framework, can be shown to work using computational methods. Other analyses of gender in Russian are not backed by similar demonstrations of accuracy. Thus formal tools like DATR can elucidate cases which are crucial for typological purposes.

2.2. Syncretism

In our first example, the impetus came from typology for us to provide a formal and hence testable account of critical data. Now we move to an example of the reverse, an attempt at formal description which leads to a typology. The varying patterns of neutralization in Slavic are intuitively of different types, and the differences in the natural ways of handling them in DATR support this view. Some correspond to the notion of 'syncretism', where a single inflected form corresponds to more than one morphosyntactic description (Spencer 1991:45), or informally where the morphology 'fails' the syntax. Work within Network Morphology has led us to a typology of these neutralizations, first according to their domain and second according to their nature.

The Domain of Syncretism. Consider the singular paradigm of *kost´* 'bone':

(7)

NOM(inative)	kost´
ACC(usative)	kost´
GEN(itive)	**kost´i**
DAT(ive)	**kost´i**
INST(rumental)	kost´ju
LOC(ative)	kost´i

Among other things, we want to say that genitive and dative singular are identical. We could reflect this in the lexical entry for *kost´*:

(8)
```
Kost´:
    <> == N_III
    <gloss> == bone
    <mor dat sg> == <mor gen sg>
    ...
```

However, if we express the identity in this way, it will apply only to the lexical entry for this one word. Such a situation, a syncretism involving a single lexical item, forms the first part of our typology.

However, we are not aware of any instances in Russian. In fact, this syncretism holds for all members of class N_III, the nouns like *kost´*. We may position this at a higher point in the inheritance hierarchy, at the node for nouns of declensional type III, from which *kost´* inherits:

```
(9)  N_III:
          ...
          <mor dat sg> == "<mor gen sg>"
```

In both instances the identity is handled under a single node — there is no need for multiple inheritance. The significance of the quotes, as discussed earlier, is that the dative singular is whatever "your own" genitive singular is: in other words dative singular will take the value of genitive singular at the original query node.

The statement of identity as given could be pushed ever higher up the inheritance tree (DATR encourages us to state generalizations higher and higher), and this gives us the first parameter of our typology:

(10) The Domain of Syncretism

- a single word
- a single inflectional class
- a subset of the inflectional classes of a word class
  (In Russian, dative and locative singular are identical in two inflectional classes.)
- across all inflectional classes in the word class
  (Russian adjectives have genitive plural identical to locative plural.)
- across more than one word class
  (In Slovene, nominative and accusative dual are identical for all nouns and adjectives, Priestly 1993:399.)
- across all potentially relevant word classes
  (In Slovene, dative and instrumental dual are identical for anything which can mark them, Priestly 1993:399.)

The Nature of Syncretism. Turning to the nature of syncretism, let us consider the data in Table 1, giving some of the forms of Russian *student* 'student' and *zakon* 'law' (in transcription):

Table 1. The Morphological Effect of Animacy

	Singular		Plural	
NOM	student	zakon	studenti	zakoni
ACC	studen**ta**	zakon	studen**tov**	zakon**i**
GEN	studen**ta**	zakon**a**	studen**tov**	zakonov

Let us start with the first column of forms. There is syncretism of accusative and genitive singular (conditioned by whether the noun is animate or inanimate). As we shall see below, this can be captured by a DATR statement including the following:[4]

```
(11)
     <acc sg animate masc> == "<mor gen sg>"
```

Basically this is saying that the accusative is the same as the genitive. If we looked at the other paradigms, we would see that they share the same *pattern* of identity, even though the particular inflections differ. It would not be sufficient to state the identity of forms separately for each paradigm; that would imply that the cases involved could equally well differ from paradigm to paradigm. This regularity can be captured in the DATR account by a statement high up the inheritance tree:

```
(12)
NOMINAL:
     <acc> == "<mor nom>"
     <acc pl animate> == "<mor gen pl>"
     <acc sg animate masc> == "<mor gen sg>"
```

[4] The gender value is available as described in section 2.1.

```
<mor acc $number>
      == < acc $number "<syn animacy>"
                 "<syn gender>" >
```

Let us go back to our DATR statement (11), which is embedded in (12). Note that this is not a symmetrical relationship. The genitive form is "right", the accusative is a copy of it. This can be seen by comparing with the second column: any noun of this type will have the genitive singular in *-a*, the accusative matches this genitive if the noun is animate, and the nominative if not. The question of directionality is a live issue. Rules which specify that one morphological form will be realized identically to another are often called 'rules of referral', following Zwicky (1985:372). They may be seen as comparable to Perlmutter & Orešnik's 'prediction rules' (1973). It is precisely because of their directionality that Aronoff (1994:83) criticizes the use of rules of referral, in certain analyses. It is thus worth demonstrating that there are instances of syncretism that are definitely not symmetrical. The Russian example appears well-founded. However, Slavic provides an even clearer case in Slovene (see Corbett & Fraser 1997 for the data).

Since the possible types of such neutralizations cannot be constrained within the formalism, we should look for constraints to impose externally. This is a part of the general enterprise of Network Morphology. The essential point, however, is that the use of the formal language DATR, which forces us to clarify distinctions often left vague, has led us towards the formulation of a typology.

2.3. Network Morphology: Conclusion

In our first example, gender assignment, we saw how typological work led us to use formal methods to clarify the analysis of the crucial language type. In the second example, syncretism, we saw how the different types of expression in the formal language led us towards a typology. From these cases we conclude that computational linguistics and typology are not opposite poles of linguistics but rather they are complementary approaches.

3. Lexical Semantics: the Case of Colour Terms

Our second area is lexical semantics; within this, colour terms form a topic of unique interest, which produces a steady stream of publications by linguists (interested in the implications for linguistic relativity), anthropologists, psychologists, psychophysicists and philosophers (see Hardin & Maffi 1997 for representatives of these different interest groups). For linguists, the Berlin & Kay hypothesis (1969) enjoys a special place, and this confers a unique aura on Russian, as the best-studied problem case.

3.1. Typological Constraints and the Russian System

While the work of Berlin & Kay on basic colour terms continues to provoke interest and research, doubts have remained about their criteria for identifying basic colour terms (as in Cromer 1991:138–140). And so there have been various attempts to find more objective measures, going on alongside extensive empirical work on ever more languages (see Davies, Sosenskaja & Corbett (forthcoming) for a recent example). As originally formulated by Berlin & Kay (1969:5), the hierarchy consists of the following positions:

Figure 3. The Berlin and Kay Hierarchy

white black	<	red	<	green yellow	<	blue	<	brown	<	purple pink orange grey

The hierarchy constrains the possible inventories of colour terms since the presence of any given term implies the existence of all those to the left (thus a language with a basic term for YELLOW will have basic terms for WHITE, BLACK and RED). It makes diachronic predictions in that languages must move from one state allowed by the hierarchy to another. (Thus a language with basic

terms for WHITE, BLACK, RED and YELLOW would next gain a basic term for GREEN, followed by a basic term for BLUE.) There have been various revisions, in Kay (1975:257-262), Kay & McDaniel (1978:638-640) and Kay, Berlin & Merrifield (1991) and throughout the situation of Russian remains of great interest. It has two basic terms for BLUE, a possibility noted by Berlin & Kay (1969:36, 99) and later in Kay & McDaniel (1978:640-641). Our work has confirmed Russian's unique status with twelve basic terms (instead of the normal maximum of eleven): *belyj* 'white', *čërnyj* 'black', *krasnyj* 'red', *zelënyj* 'green', *žëltyj* 'yellow', *sinij* 'dark blue', *goluboj* 'light blue', *koričnevyj* 'brown', *fioletovyj* 'purple', *rozovyj* 'pink', *oranževyj* 'orange', *seryj* 'grey'.[5]

While several of these terms are straightforward, others require comment. Our research suggests very strongly that both terms for blue are indeed basic (see, for example, Corbett & Morgan 1988;[6] Morgan & Corbett 1989, Davies & Corbett 1994.[7] Our list varies in two respects from that provided by Slobin for Berlin & Kay (1969:98-99): first we believe the basic term for ORANGE is *oranževyj,* and second that for PURPLE is *fioletovyj* (see the list experiment in Morgan & Corbett 1989 and the instrumental data on the referent of *fioletovyj* in Moss 1989).

---

[5] Examples are transliterated in this section.

[6] Unknown to Corbett and Morgan (1988), Vamling (1986) claimed that Russian has two basic terms for blue, on the basis of frequency in texts. She quoted the list proposed by Kulick and Vamling (1984) which corresponds exactly to the twelve given above, having been established independently. She noted, however, (Vamling 1986: 226) that *fioletovyj* 'purple' and *oranževyj* 'orange' 'seem to have a less certain status as basic colour terms'.

[7] For instrumental data on referents of the two terms see Morgan & Moss (1988/89) and for data on children's acquisition of the terms see Davies, Corbett, McGurk & MacDermid (1998). Differences between the terms are treated from the perspective of translation by Alimpieva (1982a) and from a diachronic perspective by Alimpieva (1982b). Examples from early texts are given in Baxilina (1975: 174-207).

3.2. A Formal Approach to Diachrony

It is naturally of interest to consider how such a situation can develop. Let us put that question together with the notion of default inheritance. Taking the historical view, we might claim that by default nothing will happen. Any change may be seen as an override. This provides a means of investigating diachrony in a formal way. We are investigating this idea in joint work involving Ian Davies and Andrew Hippisley, with assistance from Gerald Gazdar, taking the information on the basic colour terms of all the Slavic languages as our starting point (these data are available in Comrie & Corbett 1993). We are attempting to build a computable model of the colour term systems of Slavic, working backwards from current inventories to the earliest times. If we can demonstrate techniques which prove valid where the earlier data are available, they could be employed for other families where the earlier data are lacking. In common with the previous example, this formal approach interfaces with typology, and it also extends into historical linguistics. Initial results are given in Hippisley & Gazdar (1999).

When looking at Slavic, we are looking at a family for which extensive data are available and using a novel method based on the notion of defaults, a method that is computable, so allowing us to check that our claims are valid. We aim to arrive at an account of the colour term systems of the twelve contemporary members of the Slavic language family, of their common ancestor Proto-Slavic, and of the developments which have led to the present situation. We hope to offer a formal, computable approach to diachrony, and a detailed account of the colour system of Slavic, which deserves intensive study given the uniquely exceptional nature of Russian.

There are two main hypotheses we wish to test. First that default inheritance provides feasible reconstructions of ancestor languages, including the common ancestor of the family. Not only will our default inheritance model capture in an elegant way diachronic change in a language family, it will relate these changes in such a way as to reconstruct unchronicled stages in the history of the language family. In the same way, it will be used to give an account of the common ancestor of the language family, in our case

Proto-Slavic. And second that the default inheritance model will reflect the evolutionary dimension of the Berlin and Kay colour term hierarchy. The Berlin and Kay hierarchy in Figure 3 was arrived at by studies on different languages of the world. The claim made is that the universality of the hierarchy is due to the fact that each point on the hierarchy reflects an evolutionary stage in language development. Thus: 'The logical, partial ordering of [the hierarchy]...corresponds, according to our hypothesis, to a temporal-evolutionary ordering...' (Berlin & Kay 1969:4). Note that Proto-Slavic has been claimed to be a Stage V language, that is, having the basic terms as far as BLUE on the hierarchy (Priestly 1981-83:247). This is in itself a hypothesis worth testing. But more important is whether or not our model will yield results consistent with the Berlin and Kay hypothesis according to the most recent modifications. Russian is already problematic for Berlin and Kay, and further inconsistencies with the hierarchy have been found in recent experimental work (Davies & Corbett 1998).

3.3. Lexical Semantics: Conclusion

We hope to demonstrate that a diachronic account can be adequately expressed in DATR, which will allow a degree of rigour and testability not normally available in historical linguistics. This should shed light on an area of special interest in lexical semantics, namely the colour term systems of Slavic, contemporary and historical, and contribute to the typological enterprise initiated by Berlin & Kay.

4. Corpus Linguistics

Our first two formal approaches have both been of the symbolic type. We now turn to one of the stochastic type. Everyone knows that there is a connection between irregularity and frequency (see, for example, Greenberg 1966, Bybee 1995). But there is the question of whether the frequency envisaged is based on the lexeme and all its forms, or just on the irregular form(s). To investigate further we have examined nouns in the Uppsala corpus, a one million word Russian corpus. There are various analytical choices

which had to be made, which are justified at length in the paper on which this section is based (Corbett, Hippisley, Brown & Marriott, forthcoming), and which here we shall take as given. Thus, based on distributional criteria we assume a paradigm of twelve cells, while recognising that no noun has twelve distinct forms; the statistical method too will be accepted without argument here.

The general claim that there is a relationship between frequency and irregularity is something with which almost any linguist would agree. However, that relationship is too vague to be testable. Once we start clarifying the claim, we find an interesting range of possibilities. For instance, we looked initially for a straightforward linear correlation between regularity and frequency; however, the data suggested that it was more appropriate to search for a more complex relationship. Let us start with irregularity and consider its *extent*. Within a given lexeme it might be that every form could be irregular independently; or else it might be that forms come in groups which are regular or irregular together. A second question concerns the *degree* of irregularity. Russian *č´elovek ~ l´ud´-i* 'person ~ people' form an irregular relation, but so do *mést-o ~ mest-á* 'place ~ places'. Intuitively, the first type of irregularity is more severe than the second. If we believe there is a relationship between frequency and irregularity, then we might claim that it will be sensitive to degrees of irregularity. To test this claim we set up a scale of irregularity, devised of course without reference to frequency (section 4.3).

Frequency then can be viewed in two ways. Given a noun whose plural is irregular, with what precisely do we expect to find a relationship? It is easiest to see the alternatives if we consider a corpus and look at the tests we might apply. We might compare lexemes one with another or we could compare regular and irregular forms within lexemes. For the first approach, we could count up how many times each lexeme occurs in the plural. Since we are counting only plurals (without respect to other forms, i.e. the singular) we call this the *absolute frequency* of a lexeme's plural. We can then compare the absolute frequency of plural of different lexemes, to see if there is a relationship between irregular plurals and their absolute frequency. There is, however, a quite different

way to look at the plural (and indeed at any cell or combination of cells in a paradigm). That is to compare it, within the lexeme, with the other available forms. For a given lexeme, we could count how many times it occurs in the plural as compared with the number of times in the singular. This is the *relative frequency* of the plural. We can then compare the relative frequency of the plural in lexemes where it is irregular with that in lexemes where it is regular, as we consider further in the next section.

4.1. Terms and Hypotheses

We now set out a number of hypotheses to test the relationship between irregularity and frequency. We will look for a particular kind of anomaly in the corpus. An anomaly in the plurals of the corpus can be of two distinct types. The first is in terms of an anomalous count of plurals for a lexeme compared to the amount one would expect for a typical lexeme of the corpus; this is *absolute plural anomaly*. What is being compared is an absolute number of plurals for a lexeme with the distribution of the absolute number of plurals in the corpus.

The second type of anomaly is a relative one. Here it is the proportion of instances of the lexeme that are plural which is examined. The distribution of plural proportions can be calculated for the lexemes of the corpus, and if the given lexeme's proportion of plurals is extreme compared to this distribution, we would have identified a *relative plural anomaly*.

We also wish to allow for the possibility of the anomaly being due to a single cell of the paradigm. If one specific cell has an extreme proportion compared to the distribution of the proportion of that cell throughout the corpus, then we have an instance of *cell anomaly*. The anomaly is that a given lexeme has a significantly higher (or lower) than average proportion of word forms for a given cell. We define cell anomaly in relative terms only, because formulating it in absolute terms might lead us to observe plural (or singular) anomaly in disguise. (The cell might be above or below the average simply as a consequence of its singular or plural being above or below the average.)

We will investigate three hypotheses:

(13) *Hypothesis 1a*
There is a relation between absolute plural anomaly and irregularity

If Hypothesis 1a is confirmed, we will have shown that there is a relation between irregularity and frequency. In order to state the relationship more precisely, we would need to go a little further. If we observed absolute plural anomaly in certain groups of lexemes this might still be because the lexeme as a whole was anomalously frequent. We need a test which will tell us whether the frequency relationship is with the general lexeme, or whether it is specifically with the lexeme's irregular forms. Recall our original question: is frequency related to the lexeme as a whole, or to its irregular word forms? This is provided by Hypothesis 1b.

(14) *Hypothesis 1b*
There is a relation between relative plural anomaly and irregularity

We will also need to test whether there is a stronger relationship with irregularity when we combine plural anomaly (either absolute or relative) with the more specific cell anomaly. In other words, if a lexeme's plural forms occurred more frequently than average, and a particular cell in the plural was proportionally more frequent than average, are we right in expecting the noun in question to be even more irregular? This is provided by Hypothesis 2, which allows us to look for a stronger (and more fine grained) relationship with irregularity.

(15) *Hypothesis 2*
Given Hypothesis 1a or Hypothesis 1b is true, there is a stronger relationship between irregularity and the combination of plural anomaly and cell anomaly

A particular case and number may occur more frequently than average either due to the lexeme occurring frequently or to the fact that the cell occurs unusually out of proportion to all word forms in the corpus (absolute frequency of the cell).

4.2. The Data

We test the hypotheses on the nouns in a corpus. Russian is a good choice for this type of investigation. First noun paradigms have sufficient cells for us to tease apart the irregularity of the lexeme in a sub-paradigm, and that of one of its word-forms. Second, irregularity in Russian is highly varied, ranging from full suppletion to shift in stress. We use the Uppsala corpus, which is a set of Russian sub-corpora of various genres, containing in total about one million words. It is considered the best Russian corpus available, in terms of scope and design. For information on the corpus, see Lönngren (1993). The dataset which we created is in the form of a Microsoft Excel document.[8] Since we were interested in estimating proportions in different categories, there would be large standard errors in our estimates where observed numbers in each category are small. Large sampling errors would complicate detailed cluster analysis. For this reason we recorded only those lexemes which occur at least five times. Our dataset contains around 5440 lexemes, accounting for around 243 000 word forms from the entire one million word corpus.

4.3. The Irregularity Scale

We specifically wish to tease apart the irregularity of a lexeme and that of one of its inflectional forms. We expect a regular noun to have a single (unchanging) stem, as part of that, a fixed stress, and a consistent set of endings. We treat each irregularity type as a step away from regularity. Suppletion is the most severe type of irregularity but even this does not define an end point, since a noun

---

[8] The basic dataset is available on the world wide web, and can be found at: *http://surrey.ac.uk/LIS/SMG,* along with a readme file.

with suppletive stems and irregular inflections is more irregular than a noun with suppletive stems but regular inflections. We are investigating *structural irregularity*, i.e. irregularity determined by comparing forms according to a set of principles. Since we wish to investigate the relationship with frequency, we must exclude any frequency consideration when determining regularity. For determining paradigms, we start from the distributional criterion, that is, we determine how many distinctions are justified by the syntax (Comrie 1986, 1991). We accept the traditional view of six cases and two numbers, hence twelve cells in all. We propose the following scale, which is justified in Corbett, Hippisley, Brown & Marriott (forthcoming).

(16) Irregularity Scale

suppletion >
pluralia tantum >
stem augments >
segmental stem irregularity >
stress stem irregularity >
segmental inflectional irregularity >
stress inflectional irregularity >
full regularity

4.4. Discussion of Results

Our results proved to be extremely interesting. We found relations between frequency and irregularity and a certain degree of correspondence with the Irregularity Scale. We also found evidence for a split between prosodic and non-prosodic morphology.

Absolute Plural Anomaly. The first of our hypotheses, Hypothesis 1a, was confirmed. There is a relation between absolute plural anomaly and irregularity. Below we give eight groups of nouns from the corpus divided up according to our Irregularity Scale; we made a further distinction between two stress patterns which divide the singular and plural. These patterns are, according to the classification in Zaliznjak (1977): pattern C (stem stress

throughout singular, ending stress throughout plural); pattern D (ending stress throughout singular, stem stress throughout plural). The eight groups are given in Table 2.

Table 2. Absolute Plural Anomaly in Eight Groups of Nouns

Group	Type of irregularity	Stress Pattern	Median plural count	Observed number of types	p-value[9]
1	end stress pl	C	9	64	< 0.001
2	end stress sg	D	5	80	< 0.05
3	stem stress alternation	n/a	22	2	0.25
4	stem alternation	n/a	96	3	< 0.001
5	stem aug in pl	n/a	10	24	< 0.001
6	stem aug in sg	n/a	15	10	< 0.05
7	stem aug in both	n/a	14	14	< 0.05
8	suppletion	n/a	935.5	3	< 0.001

For each of the groups in Table 2 the median value for plural occurrences was significantly higher, as the p-values show, than for the corpus as a whole, with the single exception of Group 3. If we list the groups according to the median value, we get the following: Group 2, Group 1, Group 5, Group 7, Group 6, Group 3, Group 4, Group 8. The data do not support irrefutably the ordering given here as the differences in some cases are insignificant. We cannot reject an ordering of the groups according to the indexing we gave them in Section 4.3. In fact the data here could still be consistent with the principled ordering of the Irregularity Scale, which is an interesting result. Groups 3 and 4 have small sample sizes and their place in the ordering may well be anomalous.

What is shown definitely is that both singular augments and plural augments are related to absolute plural anomaly. While we might argue that singular augments mark the unexpected number

[9] The p-value represents the probability that a median value more extreme than that observed could have occurred purely by chance. A value < 0.05 is reasonable evidence that there is a relationship between anomaly and irregularity. A value < 0.01 is strong evidence that there is a relationship.

with plural anomaly, this cannot be the case with plural augments, which mark what is the expected number. In other words, it appears that having an augment throughout a particular number (irrespective of whether it is singular or plural) is related to a lexeme having a high plural anomaly. We might have expected an augment in the plural to be associated with higher occurrence of singulars than the average for the corpus. The opposite is the case. In sum, there is a relationship between frequency and irregularity in absolute terms. We must now test our Hypothesis 1b in order to see if this is true in relative terms.

Relative Plural Anomaly. The groups 1-8 were tested for the next of our hypotheses. Evidence for Hypothesis 1b turned out to be not as strong as that for Hypothesis 1a, and involved groups of a specific type. We found evidence for Hypothesis 1b for two groups, and arguably for a third. The stronger evidence is for group 6 (where there is a stem augment in the singular), and group 5 (where there is a stem augment in the plural), and the weaker evidence is for group 4 (where there is a stem alternation). In each case the irregularity is segmental rather than prosodic. The results are given in Table 3.

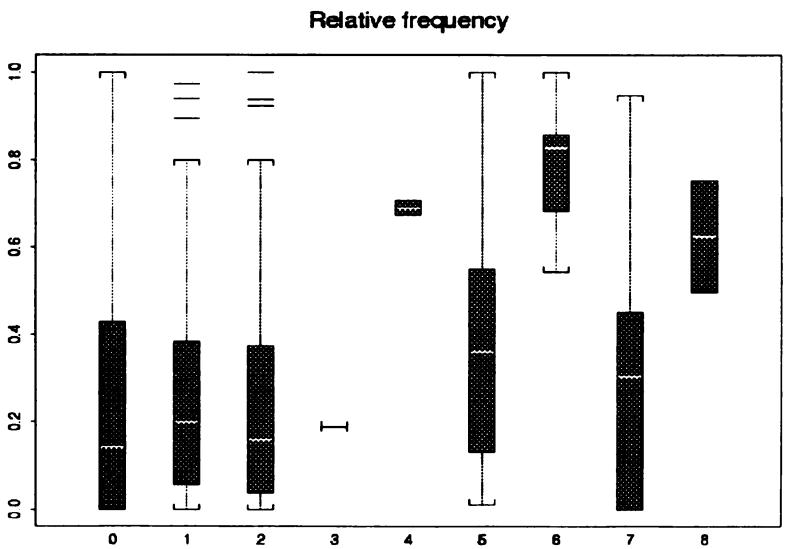
Table 3. Relative Plural Anomaly

Group	Type of irregularity	Median plural proportion	p-value
1	end stress pl	0.2	0.1
2	end stress sg	0.15	0.54
3	stem stress alternation	0.18	0.54
4	stem alternation	0.68	0.06
5	stem aug in pl	0.36	0.03
6	stem aug in sg	0.82	< 0.001
7	stem aug in both	0.32	0.4
8	suppletion	0.62	0.16

Thus we find some evidence that the frequency of occurrence of the irregular forms, and not just frequency of occurrence of the lexeme as a whole, does relate to irregularity of the forms in question. However, frequency of occurrence of forms which are irregular

only in terms of stress does not appear to relate to irregularity. In the box plot in Figure 4 the prosodic groups (Groups 1, 2 and 3) have much lower medians than the others.[10] The median is represented by the white line in the middle of the box; the box itself represents a range of proportions covering the middle 50% of the lexemes in the category; the whiskers cover the remaining 50%, except outliers which are indicated separately with horizontal bars (Daley, Hand, Jones, Lunn & McConway 1995).

Figure 4. Irregularity Type and Plural Anomaly



Key: *y* axis = proportion of plurals, *x* axis = irregularity type: 0 = regular, 1= stress C, 2 = stress D, 3 = stem stress alternation , 4 = stem segment alternation, 5 = stem augment in plural, 6 = stem augment in singular, 7 = different stem augment in singular and plural, 8 = suppletion

[10] Recall that for Hypothesis 1a this does not exclude a relationship between absolute frequency and irregularity for these prosodic irregularities.

It is an extremely interesting and significant result to find that relative frequency of occurrence in the plural appears to be important where non-prosodic irregularity is concerned, but not where prosodic irregularity is concerned. Thus degree of irregularity matters.

Cell Anomaly. Delving deeper into the paradigm, we looked to see if frequency of occurrence of individual case and number cells could be related to their irregularity. We looked at the absolute frequency of occurrences for all cells of given lexemes with one individual irregular cell.[11] This is in order to address Hypothesis 2, which is looking for a stronger relationship based on cell irregularity and cell anomaly. Since we are looking for an effect not caused by Hypothesis 1, we must concentrate on cells which do not have a significantly high lexeme frequency. Having investigated this (the nominative plural proved the best candidate) we found little evidence for Hypothesis 2.

4.5. Frequency and Irregularity: Conclusions

Our Hypothesis 1a, that there is a relation between absolute plural anomaly and irregularity, is strongly confirmed. More specifically, nouns which have an irregularity involving a split between singular and plural will tend to be nouns which occur frequently in the plural. There is a less dramatic but still significant effect when only stress is involved. There are some indications of a relation between the degree of irregularity, and the degree of plural anomaly, with cases of suppletion being an extreme case.

Hypothesis 1b, that there is a relation between relative plural anomaly and irregularity was less strongly confirmed. Here we are concerned with the plural forms of a lexeme as a proportion of all its occurrences. Where we did observe an effect, where the plural was used in proportion to the singular significantly more frequently than

---

[11] This includes lexemes for which the cell in question is the only irregularity, as well as lexemes for which the cell irregularity is accompanied by a singular-plural irregularity defined independently of that cell irregularity.

found generally through the corpus, the irregularity was always a segmental one. Furthermore whether the irregularity concerns the singular or the plural, we still find a high relative plural frequency .

When we moved down to examine single cells (Hypothesis 2), we found no evidence that irregularity is related to a high relative frequency of a specific cell in the paradigm, once the effects discussed under Hypothesis 1a and 1b are factored out. This is an interesting result, since it implies a structuring of lexical items. It suggests that an individual irregular cell does not stand out from its subparadigm (singular or plural) in terms of frequency.

There is a relation between frequency and irregularity but this claim is so general as to be relatively uninteresting. Once we clarify the claim, using a formal approach, we see that the relation is more intricate and interesting than we imagined. We find the strongest relation in the "middle ground" where we consider lexemes by splitting them into singular and plural sub-paradigms.

5. General Conclusion

Following the impetus of the FASL series, we have looked at a variety of *Formal Approaches to Slavic Linguistics*. Each one highlights an area of special interest within Slavic and each interfaces with different branches of linguistics, notably typology.

References

Alimpieva, R. V. 1982a. "Slova *sinij*, *goluboj* v proizvedenijax S. Esenina i ix èkvivalenty v pol´skix perevodax". *Materialy po russko-slavjanskomu jazykoznaniju,* 8-14. Voronež.

Alimpieva, R. V. 1982b. "Stanovlenie leksiko-semantičeskix grupp cvetovyx prilagatel´nyx v russkom jazyke pervoj poloviny XIX v.". *Voprosy semantiki: Issledovanija po istoričeskoj semantike,* 49-60. Kaliningrad.

Aronoff, M. 1994. *Morphology By Itself: Stems and Inflectional Classes* (=Linguistic Inquiry Monograph 22). Cambridge, MA.: MIT Press.

Baxilina, N. 1975. *Istorija cvetooboznačenij v russkom jazyke.* Moskva: Nauka.

Berlin, B. & Kay, P. 1969. *Basic Colour Terms: Their Universality and Evolution.* Berkley: University of California Press. [Revised edition 1991]

Brown, D. 1998. From the General to the Exceptional: a Network Morphology account of Russian nominal inflection. PhD Dissertation, University of Surrey.

Brown, D. P., Corbett, G. G., Fraser, N. M., Hippisley, A. & Timberlake, A. 1996. "Russian Noun Stress and Network Morphology". *Linguistics* 34, 53-107.

Brown, D. & Hippisley, A. 1994. "Conflict in Russian Genitive Plural Assignment: a Solution Represented in DATR". *Journal of Slavic Linguistics* 2, 30-48.

Bybee, J. 1995. "Regular Morphology and the Lexicon". *Language and Cognitive Processes* 10, 425-455.

Comrie, B. 1986. "On Delimiting Cases". In R. D. Brecht & J. S. Levine (eds.) *Case in Slavic*, 86-106. Columbus, Ohio: Slavica.

Comrie, B. 1991. "Form and Function in Identifying Cases". In F. Plank (ed.) *Paradigms: The Economy of Inflection* (=Empirical Approaches to Language Typology 9), 41-55. Berlin: Mouton de Gruyter.

Comrie, B. & Corbett, G. G. (eds.) 1993. *The Slavonic Languages.* London: Routledge.

Corbett, G. G. 1982. "Gender in Russian: an Account of Gender Specification and its Relationship to Declension". *Russian Linguistics* 6, 197-232.

Corbett, G. G. 1991. *Gender.* Cambridge: Cambridge University Press.

Corbett, G. G. & Davies, I. R. L. 1997. "Establishing Basic Colour Terms: Measures and Techniques". In C. L. Hardin & L. Maffi (eds.) *Color Categories in Thought and Language,* 197-223. Cambridge: Cambridge University Press.

Corbett, G. G. & Fraser, N. M. 1993. "Network Morphology: A DATR Account of Russian Nominal Inflection". *Journal of Linguistics* 29, 113-42.

Corbett, G. G. & Fraser, N. M. 1997. "Vyčislitel´naja lingvistika i tipologija". *Vestnik MGU: Serija 9: Filologija* no. 2, 122-140.

Corbett, G. G., Hippisley, A., Brown, D. & Marriott, P. forthcoming. "Frequency and Regularity Revisited: Knowing what to Count". Paper presented at the Symposium "Frequency Effects and Emergent Grammar", Carnegie-Mellon University, 28-30 May 1999, and to appear in the Proceedings, to be published by John Benjamins.

Corbett, G. G. & Morgan, G. 1988. "Colour Terms in Russian: Reflections of Typological Constraints in a Single Language". *Journal of Linguistics* 24, 31-64.

Cromer, R. F. 1991. *Language and Thought in Normal and Handicapped Children.* Oxford: Blackwell.

Daelemans, W., De Smedt, K., & Gazdar, G. 1992. "Inheritance in Natural Language Processing". *Computational Linguistics* 18, 205-218.

Daley, F., Hand, D., Jones, C., Lunn, D. & McConway, K.1995. *Elements of Statistics.* London: Addison-Wesley.

Davies, I. R. L. & Corbett, G. G. 1994. "The Basic Colour Terms of Russian". *Linguistics* 32, 65-89.

Davies, I. R. L. & Corbett, G. G. 1998. "A Cross-cultural Study of Colour-groupings: Tests of the Perceptual-physiology Account of Colour Universals". *Ethos* 26, 338-360.

Davies, I. R. L., Corbett, G. G., McGurk, H. & MacDermid, C. 1998. "A Developmental Study of the Acquisition of Russian Colour Terms". *Journal of Child Language* 25, 395-417.

Davies, I. R. L., Sosenskaja, T. & Corbett, G. G. Forthcoming. "First Account of the Basic Colour Terms of a Daghestanian Language: the Case of Tsakhur". To appear in: *Journal of Linguistic Typology.*

Evans, R. & Gazdar, G. 1989a. "Inference in DATR". *Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics,* 66-71. Manchester, England.

Evans, R. & Gazdar, G. 1989b. "The Semantics of DATR". In A. G. Cohn (ed.) *Proceedings of the Seventh Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour,* 79-87. London: Pitman/Morgan Kaufmann.

Evans, R. & Gazdar, G. 1996. "DATR: a Language for Lexical Knowledge Representation". *Computational Linguistics* 22, 167-216.

Fraser, N. M. & Corbett, G. G. 1995. "Gender, Animacy, and Declensional Class Assignment: a Unified Account for Russian". In G. Booij & J. van Marle (eds.) *Yearbook of Morphology 1994*, 123-50. Dordrecht: Kluwer.

Fraser, N. M. & Corbett, G. G. 1997. "Defaults in Arapesh". *Lingua* 103, 25-57.

Gazdar, G. 1987. "Linguistic Applications of Default Inheritance Mechanisms". In P. Whitelock, M. McGee Wood, H. L. Somers, R. L. Johnson & P. Bennett (eds.) *Linguistic Theory and Computer Applications,* 37-68. London: Academic Press.

Gazdar, G. 1990. "An Introduction to DATR". In R. Evans & G. Gazdar (eds.) *The DATR Papers.* Cognitive Science Research Paper CSRP 139, 1-14. School of Cognitive and Computing Sciences, University of Sussex.

Gazdar, G. forthcoming. "Ceteris Paribus". To appear in J. A. W. Kamp & C. Rohrer (eds.) *Aspects of Computational Linguistics.* Berlin: Springer.

Greenberg, J. 1966. *Language Universals, with Special Reference to Feature Hierarchies.* The Hague: Mouton.

Hardin, C. L. & Maffi, L. (eds.) 1997. *Color Categories in Thought and Language.* Cambridge: Cambridge University Press.

Hippisley, A. 1997. Declarative Derivation. Ph.D. Dissertation, University of Surrey.

Hippisley, A. & Gazdar, G. 1999. "Inheritance Hierarchies and Historical Reconstruction: towards a History of Slavonic Colour Terms". Paper read at the 35th Regional Meeting of the Chicago Linguistic Society, April 1999.

Huntley, D. 1993. "Old Church Slavonic". In B. Comrie & G. G. Corbett (eds.), *The Slavonic Languages*, 125-187. London: Routledge.

Kay, P. 1975. "Synchronic Variability and Diachronic Change in Basic Color Terms". *Language in Society* 4, 257-70.

Kay, P. & McDaniel, C. 1978. "The Linguistic Significance of the Meanings of Basic Colour Terms". *Language* 54, 3, 610-646.

Kay, P., Berlin, B. & Merrifield, W. 1991. "Biocultural Implications of Systems of Color Naming". *Journal of Linguistic Anthropology* 1, 12-25.

Kulick, Don & Vamling, Karina 1984. "Ryska". In T. Pettersson (ed.) *Färgterminologi: Seminarieuppsatser i Allmän Språkvetenskap* (=Praktisk Lingvistik 9), 79-109. Lund. [Quoted from Vamling (1986).]

Lönngren, L. 1993. *Častotnyj slovar´ sovremennogo russkogo jazyka*. (=Acta Universitatis Upsaliensis, Studia Slavica Upsaliensis 33). Uppsala.

Morgan, G. & Corbett, G. G. 1989. "Russian Colour Term Salience". *Russian Linguistics* 13, 125-141.

Morgan, G. & Moss, A. E. St. G. 1988/89. "The two Blues of Russian: the Referents of *sinij* and *goluboj*". *Die Farbe* 35/36, 353-357.

Moss, A. E. 1989. "Basic Colour Terms: Problems and Hypotheses". *Lingua* 78, 313–320.

Perlmutter, D. M. & Orešnik, J. 1973. "Language-particular Rules and Explanation in Syntax". In S. R. Anderson & P. Kiparsky (eds.), *A Festschrift for Morris Halle*, 419-59. New York: Holt Rinehart.

Priestly, T. M. S. 1981-83. "On Basic Color Terms in Early Slavic and Ukrainian". *The Annals of the Ukrainian Academy of Arts and Sciences in the United States,* vol. XV [appeared 1987], nos 39-40, 243-251.

Priestly, T. M. S. 1993. "Slovene". In B. Comrie & G. G. Corbett.(eds.), *The Slavonic Languages*, 388-451. London: Routledge.

Shieber, S. M. 1987. "Separating Linguistic Analyses from Linguistic Theories". In P. Whitelock, M. McGee Wood, H. L. Somers, R. L. Johnson & P. Bennett (eds.) *Linguistic Theory and Computer Applications,* 1-36. London: Academic Press.

Spencer, A. 1991. *Morphological theory.* Oxford: Blackwell.

Vamling, Karina 1986. "A Note on Russian blues". *Slavica Lundensia* 10, 225-33.

Zaliznjak, A. A. 1977. *Grammatičeskij slovar' russkogo jazyka: slovoizmenenie.* Moscow: Russkij jazyk

Zwicky, A. 1985. "How to Describe Inflection". In M. Niepokuj, M. Van Clay, V. Nikiforidou & D. Feder (eds.), *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society,* 372–386. Berkeley, California: B. L. S., University of California.

SLIS, University of Surrey,
Guildford, Surrey, GU2 5XH, GB
g.corbett@surrey.ac.uk

Length Neutralization and Vowel Lengthening in Orlec Čakavian

Katherine M. Crosswhite
University of Rochester

1. Introduction

In this paper, I examine a system of vowel lengthenings and shortenings found in the Orlec dialect of Čakavian Croatian (Houtzagers 1982, 1985). This dialect has phonemic vowel length, but this contrast is neutralized in two different contexts: in unaccented syllables, underlying length contrasts are neutralized in favor of the short vowels, while in *certain* accented syllables, they are neutralized in favor of the long vowels. Of the two recent theoretical approaches to positional neutralizations within Optimality Theory—licensing (Steriade 1994a,b, Zoll 1996, 1998) and positional faithfulness (Beckman 1998)—only licensing can adequately account for both the neutralization-via-lengthening and the neutralization-via-shortening phenomena. Thus, the Orlec dialect supports Zoll's (1998) refutation of the claim made by Beckman (1998) that positional faithfulness can subsume licensing as a formal mechanism.

1.1. Background on The Orlec Dialect

The dialect examined in this paper is spoken on the Island of Cres, specifically in the town of Orlec. The data on Orlec Čakavian presented here is based on the published descriptions of Houtzagers (1982, 1985), who did extensive fieldwork in the town of Orlec, as well as in other villages of the island.[1] Typologically, the Orlec dialect belongs to the Northwestern branch of the Čakavian group of the Serbo-Croatian dialect continuum. Linguistic characteristics of this dialect include:

---

[1] Houtzagers notes that the Orlec variant is fairly representative of the dialect of the island of Cres in general, but most of his fieldwork was done in Orlec.

(1) a. No leftward accent shift (cf. Inkelas and Zec 1988 on accent shift in Štokavian dialects)
b. Pitch Accents: long rising, long falling, short falling
c. Pitch-accent placement is non-predictable
d. Maintains phonemic vowel length, but this is neutralized in certain contexts
e. Lexical borrowings from Italian

1.2. Background on Positional Neutralizations in OT

A positional neutralization is a situation where two underlyingly contrastive elements [A] and [B] are neutralized in certain phonological environments ("weak" positions); in other phonological environments, [A] and [B] remain contrastive ("strong" positions). One question addressed by recent research on positional neutralizations in OT deals with predicting the positions where [A] and [B] will merge, and where they will remain distinct. Two observations have been made:

(2) Positions immune to a neutralization may be associated with certain psychological effects (cf. Beckman 1998 for review). For example, word- or root-initial segments are often immune to neutralization; the same segments play an important role in Marslen-Wilson's (1984, 1987) cohort model of spoken word recognition.

(3) Positions singled out for neutralizations are usually phonetically impoverished while positions immune to neutralizations are phonetically more robust (cf. Steriade 1994a, b). For example, stressed vowels are often immune to neutralizations; stressed vowels are also generally louder and longer than their unstressed counterparts.

These two observations lead to a chicken-and-the-egg paradox: Are strong positions phonetically robust because they are psychologically important, or are strong positions psychologically

important because they are phonetically robust? Both alternatives make sense: If certain points in a word (such as the stressed syllable) are important for spoken word recognition, it would be important to make sure the segments filling those positions are clearly articulated, thus making recognition easier. However, it is also true that if certain positions are phonetically more robust for articulatory or physiological reasons, it would be reasonable for speakers and listeners to pay more attention to these positions, since they will be more informative than other positions.

Both of these alternatives have given rise to formal machinery intended to account for positional neutralizations. These two approaches are referred to as positional faithfulness (Beckman 1998) and licensing (Steriade 1994a, b; Zoll 1998). In positional faithfulness, special faithfulness constraints specific to "strong positions" are allowed: If a given position is important for lexical access, a phonological alternation affecting that position might hinder lexical access, and therefore such alternations should be *blocked*. For example, stressed syllables might be subject to faithfulness requirements not affecting unstressed syllables. Such a requirement might be enforced in an Optimality-Theoretic grammar using a constraint such as the following, which refers to tongue root advancement ([ATR]):

(4) **Ident-σ́-[ATR]**: A stressed syllable must be faithful to underlying specifications for the feature [ATR].

If this constraint is ranked highly enough in the grammar, phenomena neutralizing [ATR] distinctions will be able to play themselves out in unstressed positions, but any attempt to do so in stressed syllables will be blocked.

In licensing, special markedness constraints specific to "weak positions" are allowed. The rationale for this approach is that phonetically complex or unusual structures should not occur in phonetically impoverished contexts where they will by necessity be implemented poorly. For example, the following constraint can be thought of as a context-sensitive markedness constraint affecting [-ATR] vowels:

(5) **Lic[-ATR]/stress:** Unstressed [-ATR] vowels may not occur in the output.

In other words, this constraint says that unstressed syllables are too phonetically impoverished to house the phonetically marked category [-ATR]. Given the correct ranking for this constraint, we might expect to see underlyingly [-ATR] vowels become [+ATR] in unstressed syllables. This would cause the neutralization of underlying [ATR] distinctions in unstressed positions, which is the same result predicted by the positional faithfulness constraint discussed above. In this case, however, limitation to unstressed vowels is a result of the fact that neutralization is motivated by the Lic[-ATR]/stress constraint only in that context. No blocking effect is necessary in this case.

As pointed out by Beckman (1998), many types of positional neutralizations can be accounted for using either type of constraint (vowel reduction in unstressed positions or consonant devoicing in coda position, for example). However, Zoll (1998) convincingly argues that although both positional faithfulness and licensing account for similar types of cases, some positional neutralizations nonetheless require a licensing approach. The reason that many positional neutralizations can be accounted for in either formal approach seems to stem for the fact that most positional neutralizations are reductive in nature: That is, positional neutralizations usually involve the removal of phonetic material from a weak position, which causes the simplification of marked phonetic structures. This is not the only way to cause neutralizations however; neutralizations can also be augmentative in nature. In such as case, two underlyingly contrastive elements would become neutralized through the addition of phonetic material. For example, neutralization of /V/ and /V:/ via lengthening or neutralization of oral and nasal vowels via

nasalization would be examples of augmentative neutralization.[2] The Orlec dialect is an interesting case for positional neutralization, since it employs both reductive vowel-length neutralization (shortening in unaccented syllables) as well as augmentative vowel-length neutralization (lengthening in certain accented syllables). The overarching generalization seen in the Orlec dialect is that long vowels cannot occur in unaccented syllables, but they can occur in accented syllables. Importantly, long vowels that are derived via lengthening are allowed to occur in accented syllables, but not in unaccented ones. As we shall we, this fact can be accounted for easily in a licensing-based analysis, but not in a positional faithfulness analysis.

2. A Licensing-Based Approach to Orlec

In the Orlec dialect, vowels can be underlyingly long or short. The phonemic status of vowel length in this dialect is demonstrated by the following forms.[3]

Table 1: Minimal And Near-Minimal Pairs for Vowel Length in Orlec

Words w/ Long Vowels		Words w/ Short Vowels	
ú:sta	*'mouth'*	úski	*'narrow'*
tí:rat	*'pull'*	tírat	*'urge on'*
tá:t	*'thief'*	ták	*'wooden chopping block'*
sé:st	*'sit down'*	ʃést	*'type of cooper's tool'*
blí:ʒu	*'near'*	sísa	*'nipple'*
muhí:ć	*'weed name'*	miʃíć	*'muscle'*
zvoní:k	*'bell-tower'*	zvonít	*'ring, toll'*

[2] Zoll points out that in addition to augmentative neutralizations, conspiratorial neutralizations (i.e., neutralizations that are carried out through various phenomena that all achieve the same outcome) also require licensing rather than positional faithfulness.

[3] For purposes of this paper, only the location of accent will be shown, with an acute accent mark. Different pitch-accents will not be noted.

However, this distinction is neutralized in unaccented syllables; unaccented vowels are realized as short. Since stress in Orlec is mobile, stress shifts and concomitant changes in vowel length are common. The example forms listed below demonstrate this alternation:

Table 2: Shortening Under Accent Shift in Orlec

Accented Long Vowel		Shortened Vowel (Unaccented)	
ʒú:j	*'blister'*	ʒují	*'blisters'*
nó:s	*'nose, spout'*	nosú	*'nose, spout' (loc.)*
tsvé:st	*'to bloom'*	tsvetú:ć	*'to bloom' (gerund)*
trú:t	*'duty' (nom.)*	trudá	*'duty' (gen.)*
bó:s	*'barefooted'*	bosá	*'barefooted' (fem.)*
frí:ʃ	*'stripe'*	friʒíć	*'stripe' (dim.)*
sné:h	*'snow' (nom.)*	snehú	*'snow' (loc.)*

To account for this alternation, I posit the following licensing constraint:

(6) **Lic-μμ/σ́:** Unaccented vowels may not be bimoraic.

This licensing constraint disallows long vowels in unaccented syllables; unaccented syllables must contain short vowels. The rationale behind this constraint is that long vowels are phonetically long, while unaccented syllables are phonetically short. Therefore, including a long vowel in an unaccented syllable will necessarily mean that the long vowel is implemented with less duration than it normally would be: the durational impoverishment of the unaccented syllable results in poor execution of the category V:. By ranking this constraint above **Max-μ** (defined below in (7)), the correct neutralizations are predicted.

(7) **Max-μ:** Do not delete underlying moras.

Tableau 1: Shortening Under Affix Shift

/sne:h/ 'snow' (nom.)	**Lic-μμ/σ́**	**Max-μ**
☞ sné:h		
snéh		*!

/sne:h + ú/ 'snow' (loc.)	**Lic-μμ/σ́**	**Max-μ**
☞ snehú		*
sne:hú	*!	

As shown in the tableau for *sné:h* 'snow' (nom.), long vowels in accented syllables are not affected by **Lic-μμ/σ́**. Therefore, an output candidate with vowel shortening is ruled out due to violation of **Max-μ** (cf. **snéh*). In the tableau for *snehú*, 'snow' (loc.), the output candidate with shortening wins. The alternative candidate without shortening fatally violates **Lic-μμ/σ́** (cf. **sne:hú*).

However, length in the Orlec dialect can also be neutralized via lengthening. In particular, underlying short vowels will become long in accented syllables that are closed by a sonorant consonant. Examples are provided in Table 3:

Table 3: Presonorant Lengthening in Orlec

Lengthening Does Not Apply		Lengthening Does Apply	
govoríla	*'spoke' (fem.)*	govorí:l	*'spoke' (masc.)*
kantát	*'to sing'*	kantá:l	*'sang' (masc.)*
ʒúɲa	*'July' (gen.)*	ʒú:ɲ	*'July' (nom.)*
terpét	*'to suffer'*	terpé:l	*'suffered' (masc.)*

Notice that lengthening only occurs if the accented vowel is followed by a *coda* sonorant. It does not occur if the accented vowel is followed by a heterosyllabic sonorant, as in *govoríla* 'she spoke' (cf. **govorí:la*). In addition, presonorant lengthening is not observed in unaccented syllables (cf. *kantá:l* 'he sang' vs. **ka:ntá:l*). Recall, however, that vowel length is phonemic in the language: a form like **govorí:la* is not impossible to pronounce in the language (cf. *tí:rat* 'pull' and *zé:nica* 'eyelash'); it is simply an

incorrect pronunciation for the form /govoríla/, which has an underlying short /i/.

The phenomenon of presonorant lengthening is not unique to the Orlec dialect. This is an important fact, since the other attestations of presonorant lengthening reveal that this is not a process that is limited in occurrence to accented syllables. Presonorant lengthening has been observed, for example, in Štokavian dialects (Zec 1994). In this variant of the process, presonorant lengthening is not limited to accented syllables. If we assume that sonorant codas are obligatorily moraic in these languages (cf. Zec 1995), then we might expect the vowels of such syllables to undergo lengthening (if they are not already long) in order to maintain their status as the most moraicly-prominent element of their syllable.[4] The following two constraints derive this situation:

(8) **Sonorant Coda:** Sonorant codas are moraic.

(9) **Moraic-Sequencing:** A non-nuclear element of some syllable must be less prominent at the moraic level than the nuclear element of the same syllable. (Abbreviated: **μ-Seq**)

Thus, if a coda consonant is monomoraic, the nucleus of that syllable must be bimoraic, which may cause lengthening of an underlyingly short vowel. The moraic sequencing constraint used here is quite similar to a constraint discussed in Crosswhite (1999) for European Portuguese. In that case, unstressed vowels undergo extreme shortening and reduction, sometimes resulting in devoicing and/or deletion of the unstressed vowel. These vowels are analyzed as structurally nonmoraic. Vowels found in syllables closed by a sonorant are not subject to this process. Thus, both the Štokavian and European Portuguese examples demonstrate that,

[4] As pointed out by Christina Bethin, this does not necessarily mean that Presonorant Lengthening creates trimoraic syllables (a marked structure). It could be, for example, that the vowel obtains an extra mora by sharing the mora of the sonorant coda consonant. Thus, the vowel can lengthen without mora insertion.

cross-linguistically, presonorant lengthening is not a phenomenon limited to stressed positions. Yet, in Orlec, presonorant lengthening does not occur in unstressed syllables. To account for this fact, it is necessary to merely rank both the **μ-Seq** constraint and a moraic faithfulness constraint (**Ident-μ**) below **Lic-μμ/σ́**. This is demonstrated in the tableaux shown below. I will assume in the tableaux that consonants are not moraic underlyingly, although this is not a crucial assumption:

(10) **Ident-μ:** A segment in an output candidate must be associated with the same number of moras as is its correspondent in the input.

Tableau 2: Presonorant Lengthening

/govorí + l/ | | | μ μ μ	SONOR-ANT CODA	LIC-μμ/σ́	μ-SEQ	IDENT-μ
☞ govorí:l | | /\ | μ μ μ μ				* (i) *(l)
govoríl | | | μ μ μμ			*!	*(l)
govoríl | | | μ μ μ	*!			

/tombat/ | | μ μ	SONORANT CODA	LIC-μμ/σ	μ-SEQ	IDENT-μ
☞ tombát | | | μ μ μ			*	*(m)
to:mbát |\ | | μ μ μ		*!		*(o) *(m)
tombát | | μ μ	*!			

As demonstrated in the first tableau, an accented vowel that is followed by a sonorant coda does undergo the lengthening required by **μ-Seq**. That is, presonorant lengthening is not blocked in accented syllables. Presonorant lengthening is blocked in unaccented syllables, however, as demonstrated in the second tableau. There, the unaccented vowel of *tombát* 'to drop' is found in a syllable with a sonorant coda, but it does not undergo lengthening. This is due to the constraint **Lic-μμ/σ**; presonorant lengthening in this case would create a long vowel in a position where it is not licensed. Note that Moraic Sequencing must outrank the constraint **Ident-μ**; the reverse ranking would predict that vowels are always faithful to their underlying moraic status.

Thus, both types of length neutralization seen in the Orlec dialect are accounted for: the ranking **Lic-μμ/σ** » **Max-μ** generates shortening of unaccented long vowels, while the ranking **μ-Seq** » **Ident-μ** generates lengthening of presonorant accented vowels. The ranking of **μ-Seq** below **Lic-μμ/σ** accounts for the fact that presonorant lengthening is limited to accented syllables. In the next section, I will show that this result cannot be replicated using positional faithfulness constraints.

3. Alternative Analysis: Positional Faithfulness

An alternative account for the Orlec vowel shortening facts is also possible in a framework utilizing positional faithfulness constraints instead of licensing constraints. In this section, I will present such an alternative analysis, and demonstrate how it cannot account for the vowel lengthening facts. Possible reformulations that avoid this problem while maintaining a positional faithfulness analysis will also be considered, but ultimately rejected.

In a positional faithfulness analysis of the Orlec dialect, shortening of unaccented long vowels is motivated by a context-free markedness constraint that punishes occurrences of long vowels in output forms:

(11) ***V:** No long vowel may occur in an output form.

This type of markedness constraint is referred to as "context-free" since it punishes all occurrences of long vowels equally; no reference is made to particular contexts where long vowels are worse (such as unaccented syllables) or better (such as accented syllables). They are all equally bad. The effects of this constraint are limited by a positional faithfulness constraint, such as the one shown below:

(12) **Max-μ/σ́:** Do not delete underlying moras in accented syllables.

By ranking **Max-μ/σ́** above ***V:**, long vowels will be avoided only in unaccented syllables. This is demonstrated in the tableaux below:

Tableau 3: Unaccented Shortening, Positional Faithfulness

/sne:h/ 'snow' (nom.)	**Max-μ/σ́:**	***V:**
☞ sné:h		*
snéh	*!	

/sne:h + ú/ 'snow' (loc.)	**Max-μ/σ́:**	***V:**
☞ snehú		
sne:hú		*!

In the first tableau, the constraint ***V:** militates towards eliminating the underlying long vowel, but this is over-ridden by the higher-ranked constraint **Max-μ/σ́**. In other words, loss of a mora from an accented vowel is a more serious violation than the occurrence of a marked long vowel. In the second tableau, no such blocking occurs, since the long vowel in this form is in an unaccented syllable, and therefore outside the domain of **Max-μ/σ́**.

However, when we add the **μ-Seq** constraint, this analysis predicts the incorrect outcome. This is demonstrated in the following tableaux, demonstrating one possible ranking of the constraints involved:

Tableau 4: Presonorant Lengthening Attempt 1; **μ-Seq** » ***V:**

/govori + l/ 'he spoke'	**Max-μ/σ́**	**μ-SEQ**	***V:**
☞ govorí:l			*
govoríl		*!	

/tombat/ 'to drop'	**Max-μ/σ́**	**μ-SEQ**	***V:**
☞ to:mbát			*
tombát		*!	

As shown above, the ranking **μ-Seq** » ***V:** correctly accounts for the fact that presonorant lengthening is allowed in accented syllables (first tableau), but the same ranking predicts the incorrect result for unaccented syllables (second tableau). However, reversing the ranking of these two constraints does not solve this problem, as demonstrated below:

Tableau 5: Presonorant Lengthening Attempt 2; ***V: » μ-Seq**

/govori + l/ 'he spoke'	**Max-μ/σ**	***V:**	**μ-SEQ**
☜ govoríl			*
govorí:l		*!	

/tombat/ 'to drop'	**Max-μ/σ:**	***V:**	**μ-SEQ**
☞ tombát			*
to:mbát		*!	

With the reversed ranking ***V: » μ-Seq**, the correct results are now generated in unaccented syllables (second tableau), but the incorrect result is predicted for accented syllables (first tableau). Note that under both ranking, none of the output candidates violates the **Max-μ/σ** constraint: This constraint can only be used to rule out inappropriate shortening in accented syllables, not inappropriate *lengthenings* in *unaccented* syllables. Thus, the positional faithfulness approach is only able to account for one type of length neutralization: unaccented shortening. It is unable to account for presonorant lengthening in accented syllables.
One possible repair strategy for this problem would be to use a second positional faithfulness constraint to bar lengthening in unaccented syllables:

(13) **Dep-μ/σ̆**: Do not insert moras into unaccented syllables.

By adding this constraint to the grammar, and ranking it above **μ-Seq**, the correct results will generated: **μ-Seq** will be able to cause lengthening in accented syllables, but this effect will be blocked by **Dep-μ/σ̆** in unaccented syllables. However, note that the **Dep-μ/σ̆** constraint refers specifically to a *weak* position. This is precisely the type of positional faithfulness constraint that is not supposed to occur. That is, unaccented syllables do not play a privileged role in lexical access tasks, and should not be afforded special protection via positional faithfulness.

Another possible repair strategy is to use a reformulated version of the **μ-Seq** constraint. Such a reformulation might be:

(14) **μ-Seq/σ́:** The nucleus *of an accented syllable* is more prominent at the moraic level than non-nuclear elements of the same syllable.

This version of moraic sequencing is accent-sensitive: it only enforces moraic sequencing in accented syllables. This revised constraint will also allow the positional faithfulness analysis to generate the correct results with respect to presonorant lengthening. However, the cost of this repair strategy is also rather high: the revised moraic sequencing constraint is itself formally a licensing constraint. Thus, this analysis requires the use of both a licensing constraint (**μ-Seq/σ́**) and a positional faithfulness constraint (**Max-μ/σ́**). To see how this is the case, consider the following:

The general version of the moraic sequencing constraint (**μ-Seq**) says that if the coda is moraic, then a monomoraic syllable nucleus is "too insubstantial" for that syllable. Such a syllable nucleus must be augmented via lengthening. The accent-specific version of moraic sequencing (**μ-Seq/σ́**) says that same kind of insubstantial syllable nucleus that precedes a sonorant coda is acceptable in a phonetically impoverished unaccented syllable, but not in a phonetically robust accented syllable. In other words, a phonetically impoverished syllable nucleus is allowable only in phonetically impoverished syllables — a licensing effect.

4. Conclusion

In the preceding sections, I presented two different Optimality-Theoretic analyses that account for the predictable shortening of unaccented vowels in the Orlec dialect of Čakavian: one based on licensing, and one based on positional faithfulness. The results of comparing these analyses show that in order to account for both vowel shortening and vowel lengthening in Orlec, you must adopt one of the following strategies:

(15) *Strategies for Analyzing Orlec Vowel Length*
   a. Use a licensing constraint
   b. Use two positional faithfulness constraints, including one that protects weak positions (contrary to the positional faithfulness hypothesis)
   c. Use both positional faithfulness, and licensing constraints

In the broader scheme of things, the Orlec example tells us that licensing constraints are a necessary formal mechanism. This mirrors the results discussed by Zoll (1998), where certain positional neutralizations in the languages Hamer and Guugu Yimidhirr require analysis via licensing, rather than via positional faithfulness. Examples such as these guarantee that licensing constraints must be authentic members of the set of universal phonological constraints used in human languages. It may be the case that there are other cases of positional neutralizations that cannot be accounted for by using licensing, but can only be accounted for by positional faithfulness. If this is the case, then we will know that we need to maintain *both* licensing and positional faithfulness as part of phonological theory.

References

Alderete, J. 1995. *Faithfulness to Prosodic Heads*. Ms., University of Massachusetts, Amherst.

Beckman, J. 1998. *Positional Faithfulness*. Ph.D. Dissertation, University of Massachussetts, Amherst.

Crosswhite, K. 1999. Vowel Reduction in Optimality Theory." Doctoral Dissteration, UCLA.

Houtzagers, H.P. 1982. "Accentuation in a Few Dialects of the Island of Cres." *Studies in Slavic and General Linguistics* 2, 117-129.

Houtzagers, H.P. 1985. *The Čakavian Dialect of Orlec on the Island of Cres*. Rodopi: Amsterdam.

Inkelas, S., and D. Zec. 1988. "Serbo-Croatian Pitch Accent: The Interaction of Tone, Stress, and Intonation." *Language* 64.2, 227-248.

Marslen-Wilson, W.D. 1984. "Function and Process in Spoken Word-Recognition". In Bouma, H. and D.G. Boushuis (eds.) *Attention and Performance X: Control of Language Processes.* Hillsdale, NF: Lawrence Erlbaum.

Marslen-Wilson, W.D. 1987. "Functional Parallelism in Spoken Word-Recognition." In Frauenfelder, U. and L.K. Tyler (eds.) *Spoken Word Recognition.* Cambridge, MA: MIT Press.

McCarthy, J. and A. Prince. 1993. *Prosodic morphology I: Constraint interaction and satisfaction.* Technical Report 3, Rutgers University Center for Cognitive Science.

Prince, A. and P. Smolensky. 1993. *Optimality Theory: constraint interaction in generative grammar.* TR-2, Rutgers University Cognitive Science Center.

Steriade, D. 1994a. *Licensing by Cue.* Ms., UCLA.

Steriade, D. 1994b. *Positional Neutralization and the Expression of Contrast.* Ms, UCLA.

Zec, D. 1994. "Footed Tones and Tonal Feet: Rhythmic Constituency in a Pitch Accent Language." *Working Papers of the Cornell Phonetics Laboratory*, vol. 9.

Zec, D. 1995. "Sonority Constraints on Syllable Structure," *Phonology* 12.1, 85-129.

Zoll, C. 1996. *Parsing Below the Segment in a Constraint-Based Framework.* Ph.D. dissertation, UC Berkeley.

Zoll, C. 1998. *Positional Asymmetries and Licensing.* Ms., MIT

University of Rochester
Center for the Sciences of Language
Lattimore Hall 503
Rochester, NY 14627
crosswhi@ling.rochester.edu

Nominative Object Constructions in Old Russian and Finnish

Edit Jakab
Princeton University

1. Introduction

In this paper I will explain the Nominative Case assignment to direct objects and bare NP adverbials in Finnish and Old Russian making use of the movement and feature checking theory of the Minimalist framework (Chomsky 1995; Ferguson 1996; Martin 1996; Bošković 1997). Although this phenomenon exists in several languages[1], I chose to focus on these two because it may have been a syntactic borrowing from Finnish and most Baltic Finnic languages[2] into the Northern dialects of Old Russian through geographic proximity and linguistic contact (Timberlake 1974). This seems to be supported by the semantic similarity between the Finnish modal sentences and imperatives (which also have a modal meaning) and the Old Russian sentences: they all have a modality meaning, expressing that something is necessary, mandatory, appropriate or worthwhile for someone to do, as illustrated in (1a), (1b), (11), (12) and (13). These Nominative object constructions always contain a modal predicate; this may be either overt, in which case it is an invariable third person singular form, as shown in (11), or non-overt, as can be seen in (12). A necessary condition in all languages is that the subject NP of the clause must be oblique, and the verb must be nonfinite. This includes Finnish but not Slavic or Baltic imperatives.[3] Since the phenomenon of Nominative objects is tightly connected to the

[1] Nominative direct objects can be found, for example, in Japanese (Fujita 1993; Koizumi 1994; Tada 1992; Takahashi 1995). However, they occur in syntactic environments that are different from the Finnish examples; their discussion would go far beyond the scope of this paper.

[2] With the exception of Livonian.

[3] This was first argued by Timberlake (1974:172-179).

existence of modal predicates, and consequently, modal semantics, this issue must be investigated in a discussion about infinitives and modality.

2. What Constructions Contain Nominative Objects?

In Finnish, Nominative is assigned to direct objects in four types of constructions (1) (Taraldsen 1985); from these the first two directly bare on the issue of modality while the last two have indirect relevance to the topic: constructions with third singular modal predicates as seen in (1a), imperatives illustrated in (1b), passives shown in (1c) and existentials/unaccusatives in (1d).
In this paper I will only discuss the first two constructions.

(1) a. Sinu-n täyty-y [t luke-a tämä kirja].
you.GEN must.3SG read.INF this.NOM book.NOM
'You must read this book.'
a'. Minu-n kannatta-a [t osta-a auto].
I.GEN worth.3SG buy.INF car.NOM
'It is worthwhile for me to buy a car.'
b. Lue (sinä) tämä kirja!
Read.IMPV.2SG (you) this.NOM book.NOM
'Read this book.'
c. Lue-taan kirja.
read.PASS book.NOM
'The book is being read.'
d. Tuli vaikea-t aja-t.
came.3SG hard.NOM.PL time.NOM.PL
'There came hard times.'

Otherwise direct objects are marked with the Genitive or the Partitive.

In Finnish direct object singular count nouns receive (a) the Genitive[4] (*-n*), (b) the Partitive *(-a/-ä, -ta/tä)* or the Nominative

[4] The Genitive morpheme is used to express the Accusative since the original *–t* desinence has been kept only in certain pronominal forms.

(-Ø) Case depending on the context. The Finnish Accusative morpheme (*-t*) was preserved only in personal and iterrogative pronouns.

(2) a. (Minä) lu-i-n tämä-n kirja-n.
I.NOM read.PAST.1SG this.GEN book.GEN
'I have read this book [the entire book].'

b. (Minä) lu-i-n tä-tä kirja-a.
I.NOM read.PAST.1SG this.PART book.PART
'I was reading this book [not the entire book].'

That the NPs marked Nominative are indeed objects, and not subjects at any derivational level is shown by Timberlake with cross-linguistic evidence:

a. The verb preceding the Nominative object is never plural[5];
b. The conditions for Genitives and Partitives replacing Accusative DOs are the same as for them replacing Nominative objects;
c. Depending on the language, direct object pronouns etc. may take Genitive form (Russian) or may have the original Accusative form (5) (Finnish[6]). These object pronouns do

[5] See example (1d); the verb is in the third person singular although the Nominative object is a plural NP. This verb form may be either a default form, or it may agree with a null expletive.

[6] This is the original Accusative ending, *-t* which was preserved only in personal and interrogative pronouns; these are the only animate categories in Finnish. It is quite universal that pronouns always exhibit more inflection than other categories (e.g., English *him*, *her* or *whom*, etc.)

i. Sinu-n täyty-y kutsu-a minu-t/*minä.
you.GEN must.3SG invite.INF I.ACC/*NOM
'You have to invite me.'

ii. Kutsu • minu-t/*minä!
invite.IMPV.2SG I.ACC/*NOM
'Invite me.'

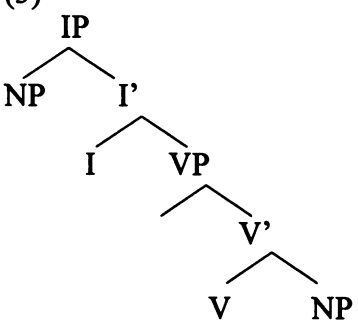
not become Nominatives, and so indicate that no general "subjectivization" of objects occurs;

d. Subjects usually precede the verb, but Nominative objects follow it.

3. Conceptual Background

In Government and Binding Theory (Webelhuth 1995) it is assumed that Nominative is assigned to the subject NP of a finite clause under government by I^0, and Accusative is assigned to the direct object NP under government by V^0 (3). This Case assignment strategy, however, is not able to account for the Nominative object construction; a V^0 cannot assign Nominative Case to its object.

(3)



In the Minimalist Program (Chomsky 1995) the Projection Principle and conventional X-bar theory have been abandoned because of their lack of conceptual necessity. Under Minimalist Syntax, the structural relations of a given element are defined in terms of the relations that the element has with other elements in the structure. It is assumed that formal features such as Case or categorial features are syntactic primitives and that they may enter into checking relations. Thus, it is natural to hypothesize that grammatical functions are related to checking relations. For example, if an argument has a feature-checking relation with Infl (or T, as I will refer to it), then that argument assumes the

grammatical function subject; hence to tell what kind of grammatical function an argument bears in this theory, it is sufficient to distinguish what kind of checking relation it enters into with what kind of functional head.

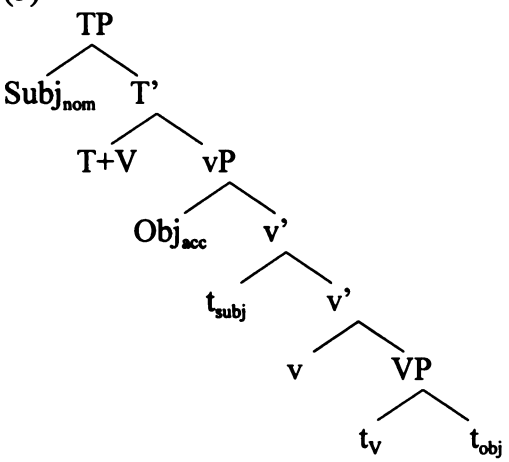
In his earlier Minimalist theory (1993) Chomsky attributes abstract Case assignment not to government, but to Spec-head agreement in AgrPs (4).

(4) $[_{AgrSP}$ Subj $[_{AgrS'}$ Agr_s $[_{TP}$ T $[_{AgrOP}$ Obj $[_{AgrO'}$ Agr_o+V $[_{VP}$ t_{subj} $[_{V'}$ t_V $t_{obj}]]]]]]]$

Thus abstract Nominative (both morphological Nominative and quirky Case) is checked in Spec-Agr_S against a T head, while Accusative (morphological Accusative and quirky Case) is checked in Spec-Agr_O configuration against a V head. Thus, the subject NP must move from its base VP-internal position to Spec-Agr_S to check its Case, and the object must raise to Spec-Agr_OP to check its features against those of V, which has already head-moved to $Agr_{O'}$.

In this paper I am assuming the typical Minimalist structure (with no Agr projections) for an ordinary active transitive sentence such as *I read the book*; its representation can be seen in (5). The nominal feature (D-feature) of the subject is checked in Spec-TP, whereas the nominal feature of the object is checked in the Spec of the higher head of the two-layered VP-shell, that is, in Spec-vP, if these checkings take place before Spell-Out (in overt syntax). The same mechanism applies to transitive sentences of this kind in Finnish and Old Russian as well, i.e., the canonical Genitive (which corresponds to the regular Accusative Case in languages that did not lose this ending) Case-feature is checked off by v before Spell-Out (see (5)) the same way.

(5)



4. Previous Theories

4.1. The Case Tier Hypothesis

One of the several theories attempting to explain the phenomenon of Nominative Case assignment is the Case Tier Hypothesis (Zaenen et al. 1985; Yip et al. 1987) according to which the highest available grammatical function (GF) is assigned Nominative, and the next highest is assigned Accusative on the basis of the notion that grammatical Cases form a Case tier. Hence, when the subject NP is assigned quirky Case (Dative in Icelandic) (6), the next highest available NP (the DO NP) will be assigned Nominative.

(6) Henni hefur alltaf Þótt Ólafur leiðinlegur.
she.DAT has always thought Olaf.NOM boring.NOM
'She always thought Olaf to be boring.'

The Case Tier Hypothesis and standard Case Theory do not allow bare NP adverbials[7] (adverbial adjuncts) to be treated the same way as argument NPs under the same syntactic conditions. Icelandic and German adverbial adjuncts have inherent Case marking which is invariant, i.e., it does not change under passivization or negation. However, since there are languages, such as Finnish and Russian,[8] in which adverbial NPs of temporal and spatial extension can have structural Case and in which they behave (change) the same way as argument NPs in the same environment, the Case Tier Hypothesis must be modified.

4.2. Timberlake's Impersonal Theory

Another theory is Timberlake's Impersonal Theory (1974) which establishes the Nominative object construction's systematic cross-linguistic status and its syntactic conditioning factors: whenever the V is "systematically impersonal," the direct object will be in the Nominative. He establishes Nominative as the "default Case" for the direct object, which occurs in environments that systematically lack a grammatical subject, i.e., the Nominative object represents the failure to specify the object as Accusative.

In government-binding terms Nominative objects are a failure to assign morphological Accusative, even though Nominative objects have "abstract"[9] Accusative Case. The rules for the phonological realization of abstract Case are spell-outs of the set of abstract Case categories on an NP.

---

[7] In several languages adverbials are PPs, and by analogy even bare NP adverbials can be viewed as receiving Case from a null P (Emonds 1989).

[8] Russian bare NP time adverbials can receive Accusative, and Genitive under negation just as direct objects can. We will see this in detail in Section 5.4.

[9] In GB terms abstract Case is the property of an NP which identifies its governor:
Dative/Oblique = NP_P
Genitive Case = NP_D or NP_N (D = the specifier of the NP)
Accusative Case = NP_V
Nominative Case = NP_I (I = the finiteness category)

4.3. Mitchell's New Functional Projections For The Subject

Mitchell (1991) suggests a new functional projection, the PredP (Predicative Phrase), in the Spec of which the subject is base-generated in order to explain the Case "assignment" in the passive (1c) and unaccusative constructions (1d). According to a rule that she proposes, i.e., "Assign Nominative Case to Spec-Pred," the subject in this position will be assigned Nominative, and the object, which is base-generated in Spec-VP, will be assigned the canonical Genitive the usual way. If, however, there is no subject in the sentence (as in passive constructions[10] such as (1c)), the direct object can raise to Spec-PredP where it will be assigned Nominative. Her explanation for unaccusative sentences is similar: the single argument (the direct object) is base-generated in Spec-VP and moves to Spec-PredP where it is assigned Nominative (1d).

Before rejecting this analysis, I am going to present her account of the remaining two constructions (1a) and (1b). To explain the structure of these sentences, she proposes another new functional projection (OblP=Obligation Phrase[11]) in the Spec of which their subject will be base-generated. This is necessary because Spec-PredP must be made empty (subject trace-free) in order for the object to be able to move there and get Nominative Case. This way, since the subject (oblique or Nominative) is base-generated in Spec-OblP, the direct object has leeway to raise to Spec-PredP to be assigned Nominative.

Even though her analysis seems to be able to explain the Case assignment in these constructions with "Nominative object", there are several conceptual and empirical problems with it. First of all, the notion of Case assignment is not valid any more in the Minimalist framework; the arguments check their Case features

---

[10] She argues that since in Finnish passive constructions the agent can never be expressed, i.e., a *by*-phrase is never allowed, Spec-PredP is empty (nothing is base-generated there).

[11] She posits an OblP on the basis of the semantics of the subject in these constructions in that it expresses some type of obligation. Although it is not entirely accurate to state that the subject determines the meaning of the construction, its semantics is clearly related to the predicate.

against those of functional categories rather than being assigned Case.[12] Second, it is not advantageous, and against Minimalist goals, to have two different positions for the base-generation of the subject NP: Spec-PredP and Spec-OblP.[13] It is easier to base-generate the subject in Spec-vP following Chomsky (1995), and from there it can raise to the appropriate functional categories to check its features.

The empirical problem with her analysis is that it fails to account for the behavior of third person imperatives (9); this will be detailed in the end. It also fails to account for the lack of agreement in unaccusative constructions such as (1d).

5. The Feature Checking Mechanism

I will show that these theories cannot account for all the Finnish data since in Finnish the direct object can be assigned Nominative even when there is another overt Nominative in the clause (1b). Moreover, when the Finnish infinitive bears a possessive suffix, thereby making the verb form personal, the object will still be in the Nominative contradicting the impersonal theory (8). Following Ura (1996), I assume that

a. The EPP-feature [D-feature] of T [=Infl] is strong in Finnish and Old Russian, hence must be checked and deleted before Spell-Out;[14]
b. The experiencer subject is inherently assigned Genitive in Finnish and Dative in Old Russian by the complex Mod + V_{inf} head;
c. T's ϕ-features are strong;

---

[12] There is no longer a one-to-one relation between a functional category and an argument; e.g., the object is also able to check T's Nominative Case feature if it bears the appropriate features.

[13] It can be asked how the difference between unaccusative and unergative verbs is accounted for if there is only one subject position. The surface subject is the direct object for unaccusatives but it is the subject position for unergatives.

[14] Strong features that remain unchecked at PF cause the derivation to crash.

d. T's Nominative Case-feature is weak, hence it can be checked at LF;
e. The matrix verb in modal constructions, which is a modal, has no ϕ-features; it is a functional head Mod rather than a full lexical verb[15] (hence it cannot induce agreement);
f. The imperative verb's D-feature and ϕ-features are strong.[16]

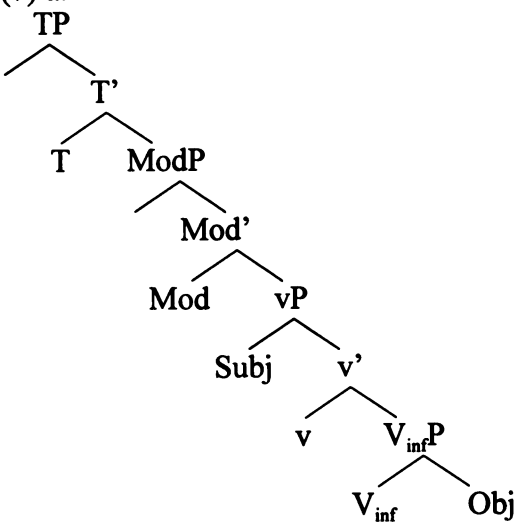
The base-generated tree for Nominative object constructions is represented in (7a). I propose that first (7b) the subject moves to Spec-ModP to check its oblique Case with the complex Mod + V_{inf} head (the infinitive has already adjoined to Mod). Then, because of the strong EPP and ϕ-features of T, the subject NP is further attracted to Spec-TP to check these features before Spell-Out. However, it cannot check T's Nominative feature because it inherently has an oblique Case-feature (Genitive in Finnish (1a) and Dative in OR (11) and (12)). Since the Case-feature of T is not strong, it can be checked by another NP at LF (7c); hence T's weak Nominative feature will be checked by the Nominative Case-feature of the direct object NP which then moves to the outer Spec of TP at LF to check its Case-feature against the weak Case-feature of T[17] As I already mentioned, I am assuming multiple Specs (Ura 1996).

[15] S. Wurmbrand (1998:260) draws the same conclusion about modals on the basis of German modal constructions.

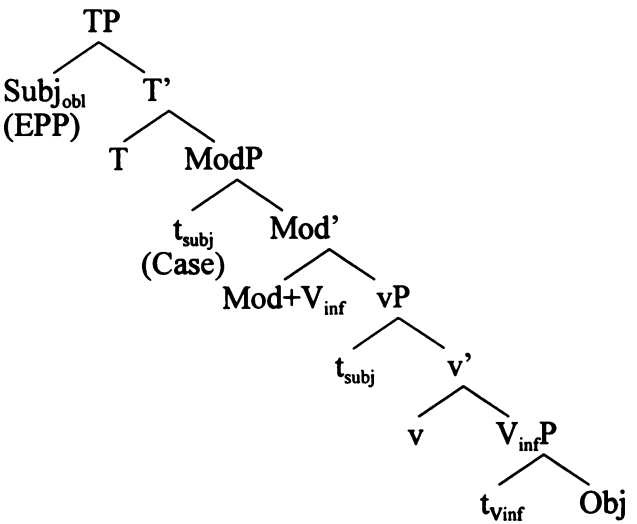
[16] I base this assumption on an attested nominal suffix -**k* in the imperative verb form which is visible only in the plural. That it exists in the singular as well is evidenced by the obligatory glottal stop that follows the second singular imperative form.

[17] I assume that only features, not categories, move at LF, and features at LF always move to a head that attracts them.

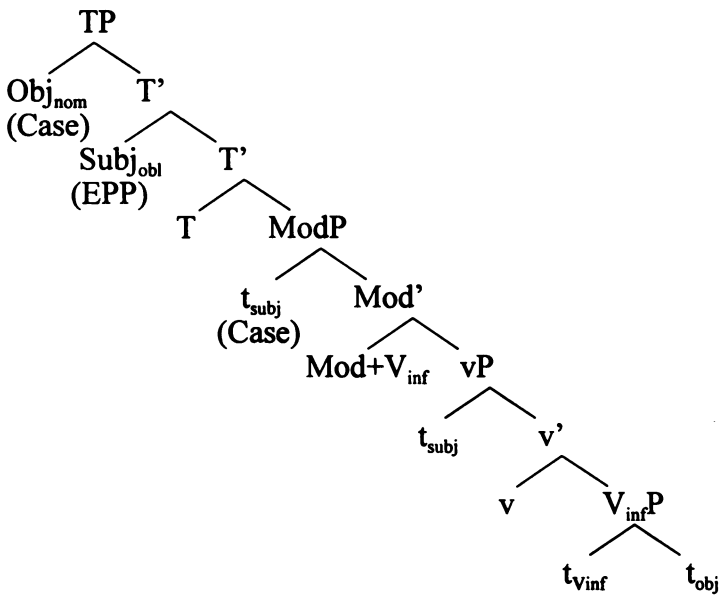
(7) a.



b.



c.



In imperatives, as was shown in (1b) above, the Nominative direct object checks its Case against T's Nominative Case-feature at LF as well; however, the imperative verb itself has a D-feature which checks off T's EPP and ϕ-features. We will see this later in detail.

5.1. Finnish Modal Constructions

In the modal constructions (1a), the oblique (Genitive) subject, which is base-generated in Spec-vP (7a), raises to check off T's strong EPP (D) feature in Spec-TP, and it checks its quirky Case-feature against Mod which has already inherently assigned Genitive to it together with the infinitive head.[18] The modal

[18] We know that it is the subject, and not the object, that checks T's D-feature because the word order shows that the object did not, and could not, move overtly.

predicate in these constructions has no ϕ-features (though it has tense), therefore it is not able to induce agreement. Since T's Nominative Case-feature also needs to be checked, the direct object NP moves to the outer Spec of T at LF to check its Nominative Case-feature against that of T. This is possible because T's Nominative Case-feature is weak. I assume Ura's theory of multiple Specs and multiple feature-checking in accounting for Nominative objects which are one instance of the phenomenon of grammatical function splitting.

According to Timberlake's theory (1984) the object in (8) should be in Accusative since the possessive suffix on the infinitive makes the infinitive "personal"; however this prediction is not borne out. The object is in the Nominative, which shows that the personal/nonpersonal distinction cannot always account for Nominative objects. In my proposal the fact that the object is Nominative follows from the rules of Minimalist theory: it must be Nominative in order to be able to check T's weak Nominative Case-feature.

(8) a. Sinu-n täyty-y matkusta-a Suome-en
you.GEN must.3SG travel.INF Finland.ILLAT
[t osta-a-kse-si auto].[19]
buy.INF.transl.POSS.2SG car.NOM
'You have to travel to Finland to buy a car.'

b. Matkusta Suome-en
travel.IMPV.2SG Finland.ILLAT
osta-a-kse-si auto.
buy.INF.TRANSL.POSS.2SG car.NOM
'Travel to Finland to buy a car.'

[19] This construction can be tied to the issue of the Case licensing domain: how large can it be? How is it possible for the object to move out from the embedded clause (the purpose clause) in order to get Case from the matrix modal? Perhaps it is not a full clause but rather some kind of a nominal form. It is also plausible that the infinitive is transparent and it somehow inherits the Nominative Case assigning property from the modal, and thus can be licensed locally.

5.2. Finnish Imperatives

In imperatives (1b) repeated as (9) the Nominative direct object *tämä kirja* 'this book' checks its Case against T's Nominative Case feature at LF as it does in the previous construction; however, here the imperative verb itself has a D-feature which checks off T's EPP and ϕ-features.[20]

(9) a. Lue (sinä) tämä kirja!
read.IMPV.2SG (you.NOM.SG) this.NOM book.NOM
'Read this book [sg].'
b. Luke-kaa (te) tämä kirja!
Read.IMPV.2PL (you.NOM.PL) this.NOM book.NOM
'Read this book [pl].'

I am suggesting that the imperative verb has a D-feature on the basis of a nominal morpheme **-k*, which was historically present also in the second person singular form of the imperative. This is evidenced by the glottal stop that is pronounced after this second singular verb form, and also by the fact that the second person plural still has an overt *–k* morpheme.[21]

This feature of the imperative verb also explains why the other Nominative NP *sinä* 'you' is optional and it never occurs in the canonical sentence-initial subject position.[22] Since the imperative verb has a D-feature, which can check off T's EPP and ϕ-features, it automatically determines the person of its subject in both the main and the subordinate clauses but does not itself have a Case-

[20] This is in concord with Timberlake's observation (1974:171) according to which the category of imperative determines the person of the subject participant, and not vice versa, but it contradicts his nonpersonal theory since the imperative shows person, yet it takes a Nominative direct object, and not an Accusative one.

[21] It is also possible that the imperative element (its D-feature) assigns quirky Nominative to its object.

[22] Another possibility is that, similar to English, the imperative V moves to C, and the overt or understood NP *sinä* 'you' checks the EPP feature.

feature; hence it cannot check T's Nominative Case, and so the object NP must move at LF to check it off.

However, the objects of imperative verbs with third person subjects are in the canonical Genitive Case (10).

(10) Lue-koon hän kirja-n.
read.IMPV.3SG he book.GEN
'May he/she read the book.'

This can be explained not only by their different semantics, i.e., that the third person forms are optative forms expressing wish or permission, but also theoretically: Finnish verbs in the third person lack a D-feature, hence they need a third person pronoun; this is reflected by the pro-drop rules in Finnish according to which only first and second person pronouns can be dropped, whereas third person pronouns are obligatorily present in a sentence.

5.3. Russian Modal Constructions

The Russian examples (11) and (12) can be explained in the same way; the oblique (Dative) subject checks off T's EPP-feature, and T's Nominative Case-feature is checked against the Case-feature of the Nominative object NP at LF. The Russian predicate, like the Finnish one, has no ϕ-features, therefore it is not able to induce agreement.

(11) Ino dostoit'' [mužu žena svoja nakazyvati].[23]
for fit.3SG man.DAT wife.NOM his.NOM punish.INF
'For it is fitting for a man to punish his wife.'

(12) Tym' znati svoja služba.
they.DAT know.INF own.NOM duty.NOM
'It is for them to know their own duty.'

[23] This modal sentence has an epistemic reading; the EPP is checked by a *pro* or an expletive in the matrix, whereas the mechanism proposed for Nominative object constructions applies to the embedded clause.

When both an overt subject NP and an overt modal predicate are missing (13),[24] the object overtly moves to the subject position to be able to check T's EPP as well as Case-features. There is, however, no agreement since the infinitive has no ϕ-features.

(13) Zemlja paxat'.
earth.NOM plow.INF
'It is necessary to plow the land.'

In example (14), however, the direct object is in the Accusative, and the reason for this is that *voevoda* 'general' belongs to the animate category in Russian since it is a masculine animate N.[25]

(14) I mne poslat' svoego voevodu s' tvoim' voevodoju.
I.DAT send.INF own.ACC general.ACC with your.INST general.INST
'And it is for me to send my general with your general.'

5.4. Bare NP Adverbials

Bare NP adverbials exhibit the same alternations between Nominative and Accusative[26] as direct objects in the same environments (Babby 1991; Franks 1995): (15a) has the same structure as (1a). (15b) and (15c) show that when the NP adverbial

[24] This is an example from a modern North Russian dialect.

[25] It is noteworthy that only animate categories (personal pronouns in Finnish and animate masculine nouns in OR) preserve the canonical Case ending for a direct object. It is plausible that Nominative animate nouns and pronouns have some strong feature that must be deleted in the checking relation with EPP, and since the oblique subject already checks the EPP, no other element is necessary to perform this, and therefore the direct object must be in the Accusative.

[26] In Finnish the original Accusative marker *–t* has been preserved only in certain pronominal forms (*minut* 'me'); otherwise the Accusative is morphologically the same as the Genitive (*-n*) (see fn. 3).

occurs with an aspectually neutral verb, only the direct object NP can be assigned Nominative, whereas the adverbial receives an Accusative desinence since only one NP is needed to check off T's Case-feature.

(15) a. Minu-n täyty-y [t luke-a koko päivä/*päivän].
I.GEN must.3SG read.INF whole day.NOM/*GEN
'I have to read all day.'
b. Muista matka koko vuode-n!
remember.IMPV.2SG trip.NOM whole year.GEN
'Remember the trip for an entire year.'
c. Ranska juliste-ttiin kolmanne-n kerra-n/*kolmas
France.NOM declare.PASS third.GEN time.GEN/*third
kerta tasavalla-ksi.
time.NOM republic.TRANSL
'France was declared a republic for the third time.'

Modern Russian bare NP time adverbials can receive Accusative (16a), and Genitive under negation just as direct objects can, but they do not become Nominative as Finnish adverbials do under the same syntactic conditions. Although Old Russian did not exhibit the Nominative object phenomenon, it is important to note what happens to Russian adverbial NPs because they change their Case under certain circumstances as Finnish adverbial NPs do in the scope of an Operator. In Russian, the adverbial can be assigned Genitive only in the scope of the negative particle *ni*, i.e. only if *ni* immediately precedes the adverbial (16b). In the absence of the *ni* particle a Genitive object and a Genitive bare NP adverbial cannot cooccur (16c); the latter has to be in the Accusative. This follows the logic of the Case Tier Hypothesis in that the most marked Case (here the Genitive of negation) is assigned to the highest available grammatical function just as the most marked Case (the Nominative) in Finnish is assigned to the highest grammatical function (15a), and the unmarked (Accusative) is assigned to the next highest grammatical function (the adverbial) just as the unmarked Finnish objective Case (=the

Genitive) is assigned to the next highest grammatical function (15b) or (15c).

(16) a. Ivan čital ètu knjigu vsju nedel-ju.
I. read this.ACC book.ACC all.ACC week.ACC
'Ivan has been reading this book all week.'
b. Ivan ne čital ètoj knigi[27] ni minuty.
I. NEG read this.GEN book.GEN NEG minute.GEN
'Ivan didn't read this book for (even) a minute.'
c. Ivan ne čital ètoj knigi[28] vsju nedelju/*vsej nedeli.
I. NEG read this.GEN book.GEN all.ACC week.ACC/*all.GEN week.GEN
'Ivan hasn't read this book all week.'

In Finnish, the choice of the Partitive (which can be viewed as the equivalent of the Russian Genitive) (Itkonen 1979; Karlsson 1987) or the Genitive (which is the equivalent of the Russian Accusative) works in a similar way. Finnish direct object NPs and bare NP adverbials receive the Partitive Case in the scope of an Operator (negation, indefiniteness or numerals other than 'one'). Similarly to Russian, there can be only one NP in the Partitive in a sentence, and the adverbial will receive the regular Genitive Case (17a) and (17c). The only difference is that in certain Finnish constructions, such as (17b) both the direct object NP and the time adverbial NP can be in the Partitive since the Operator (here negation) is strong enough to have scope over both elements. The Case Tier Hypothesis fails to explain this instance.

(17) a. Ivan on lukenut tätä kirjaa koko viiko-n.
I. has read this.PART book.PART whole week.GEN
'Ivan has been reading the book all week.'

[27] The Accusative *ètu knigu* is also possible.
[28] See fn. 27.

b. Ivan ei lukenut tätä kirjaa I.
I. NEG read this.PART book.PART
minuuttia(kaan).
minute.PART (even)
'Ivan didn't read this book for (even) a minute.'
c. Ivan ei ole lukenut tätä kirjaa
I. NEG be read this.PART book.PART
koko viiko-n.
whole week.GEN
'Ivan hasn't read this book all week.'

6. Conclusion

In this paper I gave a unified account for the usage and occurrence of Nominative objects in Finnish and Old Russian modal constructions (including Finnish imperatives) in the Minimalist framework by showing that the object NP indeed has to bear the Nominative Case to be able to check off T's weak Nominative case feature at LF. I also examined bare NP adverbials and showed that they pattern similarly to argument NPs in the same environments.

References

Babby, L.H. 1991. "Noncanonical Configurational Case Assignment Strategies". In Toribio A. and W. Harbert (eds.) *Cornell Working Papers in Linguistics* 9, 1-55. NY: Ithaca.

Bošković, Ž. 1997. *The Syntax of Nonfinite Complementation: An Economy Approach*. Cambridge, Mass.: The MIT Press.

Chomsky, N. 1993. "A Minimalist Program for Linguistic Theory." In Hale K. and S. J. Keyser (eds.) *The View from Building 20: Essays in Linguistics in honor of Sylvain Bromberger*, 1-52. Cambridge, Mass.: MIT Press.

Chomsky, N. 1995. *The Minimalist Program.* Cambridge, Mass.: MIT Press.

Emonds, J. 1989. *Timberlake's Nominative Objects in Finnish and Old Russian.* Paper Presented at the Annual Conference of the Linguistic Society of America.

Ferguson, K.S. 1996. "Shortest Move and Object Case Checking". In Abraham W., S. Epstein, H. Thráinsson and C. Zwart (eds.) *Minimal Ideas: Syntactic Studies in the Minimalist Framework*, 97-111. Amsterdam: John Benjamins.

Franks, S. 1995. *Parameters of Slavic Morphosyntax.* Oxford: Oxford University Press.

Fujita, K. 1993. "Object Movement and Binding at LF." *Linguistic Inquiry* 24, 381-388.

Harbert, W. and A. J. Toribio. 1991. "Nominative Objects". *Cornell Working Papers in Linguistics* 9, 127-192. NY: Ithaca.

Itkonen, T. 1979. "Subject and Object Marking in Finnish: An Inverted Ergative System and an "Ideal" Ergative Sub-System". In Plank T. (ed.) *Ergativity: Towards a Theory of Grammatical Relations.* NY: Academic Press.

Karlsson, F. 1987. *Finnish Grammar.* Juva: WSOY.

Koizumi, M. 1994. "Nominative Objects: The Role of TP in Japanese". *Proceedings of FAJL 1: MIT Working Papers in Linguistics* 24, 211-230.

Martin, R. 1996. *A Minimalist Theory of PRO and Control.* Ph.D. dissertation, University of Connecticut.

Mitchell, E. 1991. "Case and the Finnish Object". *Cornell Working Papers in Linguistics* 9, 193-228.

Mitchell, E. 1994. "When Agr_o Is Fused To Agr_s: What Morphology Can Tell Us About The Functional Categories". In Harley H. and C. Phillips (eds.) *MIT Working Papers in Linguistics* 22, 111-131.

Tada, H. 1992. "Nominative Objects in Japanese". *Journal of Japanese Linguistics* 14, 91-108.

Takahashi, D. 1995. "On Antecedent-Contained Deletion". *University of Connecticut Working Papers in Linguistics* 5, 65-80.

Taraldsen, T. 1985. "On the Distribution of Nominative Objects in Finnish". *Proceedings of the 12th Scandinavian Conference on Linguistics.* Joensuu: University of Joensuu.

Timberlake, A. 1974. "The Nominative Object in Slavic, Baltic, and West Finnic". *Slavistische Beiträge* 82. München: Verlag Otto Sagner.

Ura, H. 1996. *Multiple Feature-Checking: A Theory of Grammatical Function Splitting*. Ph.D. Dissertation, MIT, Cambridge, Mass.

Yip, M., J. Maling and R. Jackendoff. 1987. "Case in Tiers." *Language* 63, 217-250.

Webelhuth, G. 1995. *Government and Binding Theory and the Minimalist Program*. Blackwell Publishers.

Wurmbrand, S. 1998. *Infinitives*. Ph.D. Dissertation, MIT, Cambridge, Mass.

Zaenen, A., J. Maling and H. Thráinsson. 1985. "Case and Grammatical Functions: The Icelandic Passive". *Natural Language and Linguistic Theory* 3, 441-483.

Dept. of Slavic Languages
and Literatures
028 East Pyne
Princeton University
Princeton, NJ 08544
enjakab@princeton.edu

Infinitives and Constraints on Pronominals in Russian

Konstantin I. Kazenin
Moscow State University and University of Tübingen

1. Introduction

This paper deals with some aspects of distribution of pronominals in Russian. Though usually thought to follow Principle B of the Binding Theory in its standard formulation, Russian pronominals in certain configurations can unexpectedly corefer with the subject of their clause. This is observed in infinitive clauses (except subject-controlled ones) and in simplex sentences with dative subjects. I argue that, despite the apparent diversity of the two constructions, the "non-standard" use of pronominals in them can receive a unified account based in a crucial way on Babby's (1996) analysis of Russian infinitive complements. After suggesting an analysis for the "local" coreference of pronominals, in the final section I show that some facts related to it are not accounted for by the proposed analysis and therefore still need an explanation.

2. The Binding Theory

Principles A, B, and C of the Binding Theory as formulated by Chomsky (1981:188) postulate non-overlapping domains for the three classes of nominal expressions, as shown in (1):

(1) Principle A: An anaphor is bound in its binding domain.
Principle B: A pronominal is free in its binding domain.
Principle C: An R-expression is free.

The binding domain (or governing category) initially received a uniform definition putatively valid for every language (Chomsky, 1981:188). Roughly, it was defined as a minimal clause or NP containing the element in question. Under this concept of binding domain, the contrast in grammaticality between (2) and (3) in

Russian is captured. Indeed, in (3) the subject is not bound as the corefering NP does not c-command it; hence the ungrammaticality of (3). By contrast, in (2) the object anaphor obligatorily corefers with the c-commanding subject, thus yielding a grammatical sentence:[1]

(2) Ivan [uvidel sebja v zerkale].
Ivan saw self.ACC in mirror
'Ivan saw himself in the mirror.'

(3) *Sam [uvidel Ivana v zerkale].
self saw Ivan.ACC in mirror

Similarly, Principle B captures the key distributional properties of Russian pronominals. The grammaticality of (4) follows from Principle B as it allows free pronominals in any configuration:[2]

(4) [$_{IP}$ On$_{i/*j}$ [$_{VP}$ uvidel v zerkale Ivana$_j$]].
He saw in mirror Ivan.ACC
'He$_{i/*j}$ saw Ivan$_j$ in the mirror.'

By contrast, (5a) with the given indexing is ungrammatical because the pronominal is bound by the subject in its binding domain (clause). Co-reference of a pronominal with a non-subject NP in its clause, as in (5b), by contrast, is allowed:[3]

---

[1] I assume that in Russian, the subject is generated under the Spec of VP and further moves into the Spec of IP where it is assigned Nominative case; in this I follow the general approach of Koopman and Sportiche (1988), adopted for Russian by Franks (1995: 226ff.)

[2] Note that the coreference between the accusative NP *Ivana* and the pronominal is precluded by Principle C, which does not allow an R-expression to be bound.

[3] I do not discuss here the question of whether the grammaticality of (5c) is due to a special configuration of sentences with three-place verbs in Russian, where the object does not c-command and bind the pronominal, or whether (5c) implies that Principle B adequate for Russian prohibits local binding of pronominals only by subjects but not by any other NPs.

(5) a. [$_{IP}$Ivan$_i$ [$_{VP}$ uvidel ego$_{j/*i}$ v zerkale]].
Ivan saw he.ACC in mirror
'Ivan$_i$ saw him$_{j/*i}$ in the mirror.'
b. Ivan$_i$ pokazal Petru$_j$ ego$_{j,*i}$otraženie v zerkale.
Ivan showed Peter.ACC his image.ACC in mirror
'Ivan showed Peter his image in the mirror.'

The distribution of nominal expressions in Russian constructions with finite complements is also captured by the Binding Theory as formulated above. Thus, in (6) the anaphor can only be bound by the embedded subject, since its coreference with the matrix subject would make the anaphor free in its binding domain (the minimal clause). By contrast, the pronominal in (6) can get any interpretation except being coreferent with the embedded subject, because this would yield binding the pronominal within its minimal clause, in violation of Principle B; if the pronominal is coindexed with the matrix subject or with some element outside the sentence (discourse anaphora), Principle B is not violated:

(6) Ivan$_i$ poprosil, [čtoby gosti$_j$ nalili sebe$_{j/*i}$/emu$_{i/*j}$ čaju].
Ivan asked COMP guests pour themselves/him tea.DAT
'Ivan$_i$ asked the guests$_j$ to pour him$_i$/themselves$_j$ some tea.'

Certain complications for the Binding Theory, however, arise in Russian constructions with infinitive complements. As noted by Rappaport 1986, an anaphor used within an infinitive complement can corefer either with the embedded or with the matrix subject:

(7) Ivan$_i$ poprosil$_j$ gostej [nalit' sebe$_{i/j}$ čaju].
Ivan asked guests pour.INF self tea.DAT
'Ivan asked the guests to pour some tea for him.'
or 'Ivan asked the guests to pour some tea for themselves.'

The coreferentiality of *sebe* with the matrix subject is not predicted by the Binding Theory introduced above, because the anaphor remains free in its minimal clause. Based on such examples,

Rappaport 1986 has argued that the binding domain for Russian anaphors must be the minimal finite clause containing them. Under this assumption, both of the binding possibilities for the anaphor in (7) are predicted, as the minimal finite clause coincides with the matrix clause in (7).

Consider now pronominals in infinitive clauses. An object pronominal within an infinitive clause cannot be bound by the unexpressed embedded subject (which is PRO, according to the standard generative analysis which I will abandon below)), but can be bound by the matrix subject:[4]

(8) Ivan$_i$ poprosil soseda$_j$ [PRO$_j$ nalit' emu$_{i/*j}$ čaju].
Ivan asked neighbour pour.INF him tea.DAT
'Ivan asked the neighbour to pour him some tea.'

The impossibility of the pronominal coreferring with the PRO (and, via it, with the matrix object) is predicted by Principle B as long as the minimal clause counts as the binding domain for the pronominal. By contrast, the possibility of *emu* coreferring with the matrix subject is not predicted by Principle B under the definition of binding domain which Rappaport has justified for Russian. Indeed, the pronominal appears to be bound in its minimal finite clause.

The comparison between (7) and (8) reveals an overlap of distribution of anaphors and pronominals. The way to account for (7) and (8) in a non-contradictory way which I adopt in this paper is to acknowledge different binding domains for Russian anaphors and pronominals. For anaphors, it is the minimal finite clause, whereas for pronominals it is *any* minimal clause. This allows either an anaphor or a pronominal in an infinitive clause to corefer with the matrix subject. The potential difference between binding domains for anaphors and pronominals has been assumed in quite a few studies on the Binding Theory (see Reinhart and Reuland,

[4] Here and below, the coindexing does not indicate the possibility for a pronominal to have no antecedent in the sentence including it; this possibility always exists, as predicted by Principle B.

1993 and references there.) So the facts observed in (7)-(8) actually do not pose a new challenge for the Binding Theory.

A new challenge, however, comes from the behaviour of possessives in infinitive clauses. This will be the problem on which I concentrate below. That the possessive anaphor in (9) can equally corefer with the embedded and with the matrix subject clearly is predicted by the Binding Theory given Rappaport's amendments; however, it is not predicted that the possessive pronominal has the same binding possibilities as the anaphor does:

(9) Otec$_i$ poprosil Ivana$_j$ [zabrat' ego$_{i/j}$/svoi$_{i/j}$ veščі
father asked Ivan take.INF his things
so staroj kvartity].
from old flat
'Father$_i$ asked Ivan$_j$ to take his$_{i/j}$ things away from the old flat.'

Indeed, under the standard analysis of infinitive complements, the subject position of the infinitive is occupied by PRO, which under the given reading will bind the pronominal:

(10) Otec poprosil Ivana$_j$ [PRO$_j$ zabrat' ego$_j$ veščі
father asked Ivan take.INF his things
so staroj kvartity].
from old flat

Thus we get a configuration where a pronominal is bound in its minimal clause; however, it does not yield ungrammaticality.

Two ways to accommodate (10) within the current grammatical theory a priori look equally plausible: (1) the Binding Theory has to be reformulated in a way to accommodate (10); or (2) the analysis of infinitive complements with PRO in the subject position has to be dismissed; as long as there is no subject position within the infinitive clause in (10), the pronominal is free in its minimal clause, being coindexed only with the matrix object, in agreement with Principle B.

The former possibility looks fairly uncertain, however. In (7)-(8), the issues problematic for the Binding Theory were resolved

by redefining the binding domain for the respective group of expressions in Russian. In order to account for (10) in a similar fashion, we will need to adopt the idea that the binding domain for the pronominal in (10) is less than its minimal clause. For example, that could be the NP containing the pronominal. In this case, (10) would not contradict Principle B. However, this approach would not allow us to explain why in (5b) the same possessive pronoun cannot corefer with the local subject. Treating the governing categories for pronominals as the minimal NP in one construction and the minimal clause in another construction would be a mere stipulation, imposing undue complexity on the Binding Theory.

For this reason, I will explore the possibility of an account for (10) elaborating on certain peculiarities of Russian infinitive constructions. The analysis proposed below will require some amendment of the Binding Theory, but of a completely different spirit than redefinition of governing categories. The analysis will allow us to explain some restrictions on the binding pattern in (10). However, as Section 8 will show, it will still leave a number of questions unanswered.

3. Possessive Pronominals And Control

There is a restriction on the use of pronominals which is imposed by matrix verbs. The coreference of a pronominal with the understood subject of the infinitive[5] is regularly possible when that subject is controlled by the matrix (accusative or dative) object. Thus, on a par with *prosit'* 'to ask' in (9), the same construal of the pronominal is possible with the following matrix predicates with object control: *predložit'* 'to offer', *zastavit'* 'to cause', *zapretit'* 'to forbid', *naučit'* 'to teach, train', *pomoč'* 'to help'. By contrast, when the understood subject of the infinitive is controlled by the matrix subject, the former can never corefer with the possessive pronominal inside the infinitive clause:

[5] Here and below I prefer the term 'understood subject' rather than PRO. The reason for this is that the analysis of Russian infinitives I adopt below actually eliminates the category of PRO in them.

(11) Ivan$_i$ rešil [otremontirovat' ego$_{j/*i}$ mašinu].
Ivan decided fix.INF his car
'Ivan$_i$ decided to fix his$_{j/*i}$ car.'

(12) Ivan$_i$ obeščal otcu$_j$ [otremontirovat'
Ivan promised father.DAT fix.INF
ego$_{j/*i}$ mašinu].
his car.ACC
'Ivan$_i$ promised his father$_j$ to fix his$_{j/*i}$ car.'

Looking for an account of the pronominal coreference in (9), we have to make sure that our analysis also predicts the ban of the coreference in examples like (11)-(12). The analysis I am going to suggest will capitilize on the structural difference between subject-controlled and object-controlled infinitival clauses argued for by some scholars in Russian syntax. I will turn to this in the following section.

4. Russian Infinitive Clauses: IPs vs. VPs

The structural difference between subject-controlled and object-controlled infinitives in Russian was first suggested in Comrie 1974. The most recent version of this analysis, elaborated within the current generative framework, was proposed by Babby (1996).[6] The central idea of this analysis is that subject-controlled infinitive clauses are bare VPs, whereas other infinitive clauses are IPs. This means that subject-controlled infinitive clauses do not have a position for the subject, but other infinitive clauses do.

The crucial evidence for this analysis comes from the behaviour of the floating quantifiers *sam* 'himself, on his own', *vse* 'all' and

[6] Among the most relevant publications between Comrie's and Babby's work are Neidle (1988), Franks and Hornstein (1992), Laurençot (1997).

odin 'alone'.[7] A well known property of these quantifiers is that they agree in case with the NP they refer to:

(13) On sam postroil ètot garaž.
he.NOM self.NOM built this.ACC garage.ACC
'He build this garage himself.'

(14) Mne samomu bylo grustno.
I.DAT self.DAT was sorry
'I felt sorry myself.'

These quantifiers (which I will collectively refer to as SAM, following Babby) are possible in infinitive complements, where they can refer to the understood subject. The crucial question is about the trigger for their case agreement in infinitive constructions. Consider first subject-controlled infinitives. Here SAM obligatorily takes the nominative case:

(15) Ivan xotel [sam pogovorit' s bratom].
Ivan.NOM wanted self.NOM talk.INF with brother.INST
'Ivan wanted to talk to his brother himself.'

Given the standard account of PRO as caseless, the nominative case of *sam* in (15) can only come from the matrix subject. This is confirmed by the impossibility of the nominative SAM with dative matrix verbs — here SAM also must be in the dative case:

(16) Ivanu xotelos' [samomu pogovorit' s bratom].
Ivan.DAT wanted self.DAT talk.INF with brother.INST
'Ivan wanted to talk to his brother himself.'

The contrast between (15) and (16) can be straightforwardly accounted for on the assumption that SAM in these sentences agrees in case with the matrix subject.

---

[7] Following Babby, I will treat them as VP-adjuncts, although nothing hinges on this solution in the analysis I will argue for below.

Consider now SAM in object-controlled infinitive clauses. Here, rather unexpectedly, SAM takes the dative case by default:[8]

(17) Otec poprosil Ivana [ne xodit' tuda odnomu].
fatherasked Ivan.ACC NEG go.INF there alone.DAT
'Father asked Ivan not to go there alone.'

The use of the dative case in (17), unlike the nominative in (15) and the dative in (16), cannot be triggered by the matrix NP coreferent with the understood subject of the infinitive: in (17) that NP is in the accusative, which does not preclude the dative marking of SAM.

It is exactly this difference between case agreement of SAM under subject control and without it that led scholars in Russian syntax starting with Comrie (1974) to the conclusion that the presence or absence of this control trigger different structures for

---

[8] It was noted in Babby (1996) that in Colloquial Russian, case agreement with the controller (which, unlike in (15)-(16), is not a subject) is also possible instead of the 'default' dative marking; so, (16) has the grammatical variant (17'):

(17') Otec poprosil Ivana [ne xodit' tuda odnogo].
father.NOM asked Ivan.ACC NEG go.INF there alone.ACC
'Father asked Ivan not to go there alone.'

In Kazenin (1999) I argue that the preference for either the 'default' dative or case agreement depends upon the semantics of the verb. In general, matrix verbs with a high degree of 'semantic transitivity' in the sense of Hopper and Thompson (1980) prefer case agreement (i), whereas verbs with a low degree of semantic transitivity prefer the 'default' dative (ii)-(iii):

(i) Ja zastavil Ivana pojti tuda odnogo/ ??odnomu.
I.NOM forced Ivan.ACC go.INF there alone.ACC/alone.DAT
'I forced Ivan to go there alone.'

(ii) Ja nadoumil Ivana pojti tuda ??odnogo /odnomu.
I.NOM advised Ivan.ACC go.INF there alone.ACC/alone.DAT
'I advised Ivan to go there alone.'

(iii) Ja naučil Ivana ??samogo/samomu rešat' takie zadači.
I.NOM trained Ivan.ACC self.ACC/self.DAT tackle.INF such problems
'I trained Ivan to tackle such problems on his own.'

Russian infinitive clauses. Instead of complicating the case agreement rules for SAM, it was suggested that SAM always agrees in case with its most proximate subject. At the same time, it was assumed that under subject control, an infinitive clause is a bare VP lacking the position for the subject (Comrie explained this structural phenomenon as an instance of 'cohesion' between the matrix and embedded clauses under subject coreference). Given this, the most proximate subject of SAM in (15)-(16) is the matrix subject, with which SAM thus agrees in case. In the absence of subject control, however, the infinitive is viewed as having its own empty subject, which it assigns the dative case. This is what happens in (17).

It is important to note that dative case assignment to the empty subject of an infinitive complement is not an ad hoc stipulation: as shown, among others, by Comrie (1974) and Babby (1996), infinitives can be used with overt dative subjects elsewhere in Russian syntax, specifically in bare infinitival constructions as in (18) (Babby's (12)) and as sentential subjects in (19) (Babby's (8)):

(18) Vam samoj ne spravit'sja.
you.DAT self.DAT.F NEG manage.INF
'You won't be able to manage yourself.'

(19) [Tebe ujti na pensiju] označalo by
you.DAT go.INF on pension.DAT mean MOD
kapitulirovat' pered vragom.
capitulate.INF before enemy.INST
'For you to retire would mean capitulating to the enemy.'

Aiming at a generative analysis of Russian infinitive clauses, Babby notes that the account of case agreement of SAM as outlined above actually leads to eliminating the notion of PRO for Russian infinitives. In the absence of subject control, the empty subject of the infinitive is assigned the dative case by the infinitive; since PRO cannot have Case, this means that the subject of infinitive in (17) is *pro*, i.e., an empty pronominal requiring Case (see also Laurençot 1997 for discussing of possible consequences

of the dative case assigned to SAM for the theory of PRO). By contrast, under subject coreference illustrated in (15), the infinitive clause is a bare VP and as such lacks the subject position. Babby suggests that the subject theta-role of the verb is saturated via a 'direct predication' relation between the VP and the matrix verb, in the spirit of Williams 1994 (I omit the appropriate technical details as irrelevant here). The proposed analysis of subject-controlled and non-subject-controlled infinitive clauses is schematically represented in (15') and (17') respectively:

(15') Ivan xotel [$_{VP}$ sam pogovorit' s bratom].
Ivan.NOM wanted self.NOM talk.INF with brother
'Ivan wanted to talk to his brother himself.'

(17') Otec poprosil Ivana$_i$ [*pro*$_{DAT i}$ ne xodit'
father.NOM asked Ivan.ACC NEG go.INF
tuda odnomu].
there alone.DAT
'Father asked Ivan not to go there alone.'

5. The Binding Domain of Russian Pronominals

Now I would like to argue that the analysis of Russian infinitives outlined above allows us to account for the pronominal binding phenomena we are considering. For this, it is only necessary to reformulate the definition of the binding domain for Russian pronominals (i.e., the domain in which they cannot be bound), which, as already pointed out in the introduction, is standardly assumed to be the minimal clause (IP). I suggest, instead of this, the following definition of the local domain (below I will argue that this definition is independently motivated for Russian):

(20) The local domain for Russian pronominals is their minimal clause if that clause contains subject agreement.

This definition implies that if a pronominal occurs in a minimal clause that lacks subject agreement, i.e. has no nominative subject, then it can be bound inside that clause. Obviously, under this definition Principle B does not rule out (9). Under Babby's analysis, (9) has the structure in (21):

(21) Otec$_i$ poprosil Ivana$_j$ [$_{IP}$*pro*$_{DATj}$ zabrat' ego$_{i/j}$
father.NOM asked Ivan.ACC take.INF his
veščI so staroj kvartity].
thingACC from old flat.GEN
'Father$_i$ asked Ivan$_j$ to take his$_{i,j}$ things away from the old flat.'

Here the minimal clause (IP) containing the pronominal *ego* is the infinitive clause. However, it does not contain subject agreement; therefore co-indexing of the pronominal with the embedded subject is not prohibited by Principle B according to the given definition of the binding domain. Nor is its co-indexing with the matrix subject prohibited, since the former is outside the minimal clause of the pronominal.

Let us now consider the case of coreference between the matrix subject and the understood subject of the infinitive in (11)-(12). Under Babby's analysis, (11) gets the structure in (22):

(22) Ivan$_i$ rešil [$_{VP}$ otremontirovat' ego$_{j/*i}$ mašinu].
Ivan.NOM decided fix.INF his car.ACC
'Ivan$_i$ decided to fix his$_{j/*i}$ car.'

Here the minimal clause containing the pronominal is identical with the matrix clause (the embedded VP cannot count as a clause because it does not have subject position). Now, since the matrix clause contains the subject c-commanding the pronominal, co-indexing between that subject and the pronominal is ruled out by

Principle B in combination with the proposed definition of the binding domain of the pronominal.[9]

Now I would like to show that the proposed definition of the local domain of Russian pronominals receives independent motivation. It is well known that Russian has a number of predicates assigning quirky (most often dative) case to their subjects (see Kondrashova, 1994; Schoorlemmer, 1994). The predicates often assumed to belong to this set include, among others, *stydno* 'be ashamed', *veselo* 'be merry', *xolodno* 'feel cold', *žalko* 'feel pity':

(23) Ivanu bylo stydno/ veselo/ xolodno.
Ivan.DAT was ashamed/ merry/ cold
'Ivan was ashamed/merry/cold.'

(24) Ivanu bylo žalko mal'čika.
Ivan.DAT was pity boy.ACC
'Ivan felt pity for the boy.'

Care should be taken to make sure that the dative NP subcategorised by each of these predicates indeed possesses the subject properties. Of course, with all these predicates the dative NPs lack one subject property: they do not trigger subject agreement. However, at least with some of these predicates, the dative NP possesses two other properties restricted to subjects in Russian. First, it can bind anaphors and gerunds (Kondrashova,

[9] The proposed definition also accounts for the possibility of the pronominal to corefer with the matrix subject in (21), but not in (22). The possibility of coreference with the matrix subject in (21) is predicted since the coreferent subject is outside the binding domain of the pronominal. By contrast, in (22) the matrix clause counts as the minimal clause for the pronominal, and thus Principle B rules out the co-indexing under consideration. (Of course, taken alone, this contrast between (21) and (22) could be accounted for on the standard assumption that the local domain for Russian pronominals is the minimal clause containing them. However, we have already seen that this view of local domain cannot account for the coreference in (10)).

1994). This is demonstrably the case with *veselo* 'be merry' and *strašno* 'be scared' (Franks, 1995:253):[10]

(25) [Slušaja ètot rasskaz,] emu$_i$ bylo
listening this story he.DAT was
veselo/ grustno/strašno.
merry/ sad/ scary
'While listening to this story, he was merry/sad/scared.'

(26) Ivanu$_i$ bylo stydno pered soboj$_i$.
Ivan.DAT was ashamed before self.INST
'Ivan was ashamed of himself.'

Discussing examples like (25)-(26), Franks (1995:253ff) notes that the evidence they provide for subjecthood of the dative NPs is "actually far from definitive". First, Franks correctly observes that under certain conditions Russian anaphors can be anteceded by a non-subject. Although this can take place only under very special conditions, (26) as a piece of evidence for the subjecthood of datives definitely is weakened by this observation. Second, Franks claims that control of subject reference into gerunds also is not restricted to matrix subjects, but is available for some other matrix NPs as well. However, I find the examples adduced by Franks (p. 255) for non-subject control in gerunds ungrammatical, and do not see any other support for the possibility of such control. Given this, (25) in my opinion gives good evidence for the subjecthood of dative NPs combined with the predicates under consideration. Below I treat these NPs as subjects.

Second, the observation crucial for the proposed analysis of pronominals is that the datives combined with the adjectival predicates can bind a pronominal on a par with an anaphor:

[10] Admittedly, the possibility of controlling gerunds is rather marginal for the adjective predicates; in Section 8.2, however, I will consider some dative predicates which allow it freely.

(27) Ivanu$_i$ bylo nelovko za svoju$_i$/ego$_i$ ošibku.
Ivan.DAT was ashamed of own/his mistake.ACC
'Ivan was ashamed about his mistake.'

(28) Ivanu$_i$ bylo teplo v ego$_i$/?svojom$_i$ zimnem
Ivan.DAT was warm in his/ own winter
pal'to.
coat.PREP
'Ivan$_i$ felt warm in his$_i$ winter coat.'[11]

This variability might seem puzzling as soon we treat the dative NPs of these predicates as subjects. However, it is everything but unexpected under the definition of the local domain for Russian pronominals proposed in (20). Note that in the minimal clause containing the pronominal, which is here the independent sentence, there is no subject agreement. Therefore, binding of the pronominal within this clause does not violate Principle B.

To summarize, we have seen that both non-overt subjects of non-subject-controlled infinitival clauses and overt dative subjects can antecede pronominals in their own clause; for overt nominative subjects this is not the case. This result is reflected in Table 1, where the coreference possibilities which the standard Binding Theory cannot account for are shaded. The suggested modification of the definition of the binding domain of pronominals has allowed me to account for these "offending" binding facts.[12]

---

[11] I am not considering here conditions determining the choice of an anaphor or a pronominal in such constructions. These conditions are pragmatic in nature, and probably rather complex. What is relevant for my purposes is that the grammar does not preclude both the anaphor and the pronominal.

[12] It should be noted that Babby's explanation of the "default" dative case of SAM in the absence of subject control is not the only possible one. Franks and Hornstein (1992) (see also Franks 1995:238-249) propose for another explanation of the case marking of SAM. Under their analysis, PRO is always present in Russian infinitival clauses. They argue for the existence of a mechanism of case transmission, which is blocked exactly in those case when SAM cannot agree in case with the matrix subject and is assigned dative. Despite certain conceptual advantages of Franks and Hornstein's analysis (e.g. under their analysis it is expected that the dative SAM is possible also when the infinitive cannot be

Table 1. Coreference of Possessive Anaphors and Pronominals with Local Subjects in Russian

	Possessive Pronominals	Possessive Anaphors
Matrix predicates with NOM subjects	no	yes
Matrix sentences with DAT subjects	**yes**	yes
Subject-controlled infinitives	no	yes
Non-subject-controlled infinitives	**yes**	yes

Before I close this section, let me note that the facts we have observed once more show that, first, binding domains of anaphors and pronominals can overlap, and, second, they are subject to parametric variation across lamguages (the proposed definition of the binding domain of pronominals was shown to be correct only for Russian and may be unapplicable to any other language).

6. The Impossibility of a "Clause Union" Analysis

Before we proceed further, let me show an incorrect analysis which looks like an attractive alternative to the one I have just argued for. It could be tempting to treat sentences like (9) as instances of clause union in the spirit of Relational Grammar (Gibson and Raposo, 1986); under such an analysis, the embedded clause node

combined with an overt subject, which is a serious problem which Babby only sketchily discussed in his paper), it cannot explain why the same coreference possibilities are observed in non-subject-controlled infinitives and in matrix sentences with dative subjects; under Babby's analysis, as we have just seen, this does not come as a mere coincidence, as in both cases it is a dative subject that antecedes a possessive pronominal in its own clause.

in (9) would be eliminated, its predicate being amalgamated with the matrix predicate, and its dependents becoming matrix constituents. Given that Russian pronominals only have to be subject-free in their binding domain, coreference of the pronominal directly with the matrix object as its clause-mate would be predicted. However, this analysis will not predict the other binding possibility for the pronominal in (9), specifically, its coreference with the matrix subject, which will be in the same clause with the pronominal under such analysis. I conclude, therefore, that the clause union analysis is not valid for the data we have observed.

7. Dative Controllers

A difficulty for the analysis proposed in Section 5 comes from sentences where the understood subject of an infinitive is controlled by a dative matrix subject. The definition of the local domain of pronominals proposed in Section 5 predicts that in this case, in the same way as in monoclausal structures, a possessive pronominal can be co-indexed with the dative subject. However, this is not the case. To see this, consider (29). The possessive pronominal in the dative-controlled complement of (29a) looks considerably worse than in the monoclausal (29b):

(29) a. Ivanu$_i$ bylo studno [razgovarivat' s ego$_{j/??i}$
Ivan.DAT was ashamed talk.INF with his
synom].
son.INST
'Ivan$_i$ was ashamed to talk to his$_{j/??i}$ son.'

b. Ivanu$_i$ bylo stydno za ego$_i$ syna.
Ivan.DAT was ashamed of his son.ACC
'Ivan$_i$ was ashamed of his$_i$ son.'

In Section 5 I have already provided some evidence in favor of treating the dative NPs subcategorised by these adjective predicates as subjects. Given that the dative NP in (29) is the subject, we are invited to further reconsider the definition of the local domain for Russian pronominals proposed in Section 5, in order to account for

the impossibility of the possessive pronominals in (29a) to corefer with the matrix dative subject.

First of all let me note that, in light of the analysis of Russian infinitive complements introduced in Section 5, the infinitive complement of (29a) can receive two different interpretations: they can be treated either as bare VPs or as 'full' clauses (IPs) including the subject position occupied by an empty pronoun. The reason for this ambiguity is that in sentences like (29a), unlike in sentences with matrix Nominative subject, case agreement of SAM seems to say nothing about the structure. For example, the dative case of the floating quantifier in (30) can equally arise via case agreement with the empty pronominal dative subject of the infinitive (if any), or with the matrix dative subject:

(30) Ivanu$_i$ ne bylo veselo [(*pro*$_{DATi}$) odnomu$_i$
Ivan.DAT NEG was merry alone.DAT
idti po gorodu].
go.INF around town.DAT
'Ivan didn't feel merry going around the town alone.'

Since I am not aware of any other test revealing whether an infinitive clause has the subject position or not, we are free either to view or not to posit *pro* in the structure of (30) (as well as (29a)). I choose the latter option, since positing a subject *pro* of infinitives in these sentences would lead to the same structure for the infinitive clauses in them as we have as we have assigned to the infinitive clause in (21). In (21), however, the possessive pronominal can be bound by the dative subject of an infinitive. Thus, if in (29a) the infinitive clause had a subject position, we would get two kinds of infinitive complements with exactly the same structure, but different binding possibilities for pronominals. We are well advised to avoid this complication and to treat the infinitive complements in (29a) and in (30) as subjectless.

On this assumption, we have to redefine once again the local domain for Russian pronominals, as proposed in (31), in order to account for the binding facts observed in (29a).

(31) The local domain for Russian pronominals is their minimal clause if that clause contains subject agreement, or the clause in which the subject theta-role of the minimal clause is saturated.

Let me give some independent evidence for the necessity of this reformulation. Consider two more Russian verbs subcategorising for a dative NP and a subordinate clause: *xotet'sja* 'want' and *sledovat'* 'necessary, must'. Subjects of these verbs control zero subjects of gerunds especially freely, compared with other dative subjects:

(32) [Soxranjaja dannyj analiz infinitivnyx
retaining given analysis.ACC infinitive
konstrukcij], nam sleduet prinjat'
constructions.GEN we.DAT should adopt.INF
nekotorye izmenenija teorii svjazyvanija.
some change.ACC theory.GEN binding.GEN
'Retaining the given analysis of infinitive constructions, we have to adopt certain changes of the Binding Theory.'

Now if a gerund clause dependent upon one of these verbs contains a possessive pronominal, the latter cannot be co-indexed with the matrix subject, this co-indexation being available exclusively for anaphors:

(34) [Pogovoriv so svoim$_{i/*j}$/ ego$_{j/*i}$ synom,]
having.talked with own.INST/his son.INST
Ivanu$_i$ sledovalo nemedlenno uexat'.
Ivan-DAT should.PAST immediately leave.INF
'Having talked to his son, Ivan should have left immediately.'

As shown by Babby (1996), gerund clauses always are bare VPs saturating their subject theta-role in the sentence they are adjoined to (this is evident from case agreement of the SAM quantifiers in gerund clauses, which is very similar to the one observed in

subject-controlled infinitive clauses). The impossibility of the binding of *ego* in (34) by the matrix subject thus is predicted by Principle B under the definition of the local domain proposed in (31), but not by the definition considered earlier.

8. Further Problems

8.1. Possessives

A problem for which no solution can be proposed within the suggested analysis is why binding by the understood subject of an infinitive or by a matrix dative subject is possible only for possessive, but not any other pronominals, cf.:

(35) Otec$_i$ poprosil Ivana$_j$ [otnosit'sja k nemu$_{i/*j}$
father.NOM asked Ivan.ACC treat.INF to he.DAT
bolee uvažitel'no].
more respectfully
'Father$_i$ asked Ivan$_j$ to treat him$_{i/*j}$ with more respect.'

(36) Ivanu$_i$ stalo grustno za nego$_{j/*i}$.
Ivan.DAT became sad for he.ACC
'Ivan$_i$ became sad for him$_{j/*i}$.'

The exclusive status of the possessive pronominals with respect to the binding phenomena we are studying can hardly be explained away under the view proposed in Reinhart and Reuland (1993), who suggest a kind of 'out-of-theoretic' status for possessive pronouns, treating them as elements not accountable by the principles of any valid Binding Theory. We have seen that Russian possessive pronominals generally demonstrate the standard Principle B effects (cf. (5)). The sole systematic violation of Principle B under the standard definition of the local domain can lead to a redefining of the domain (what I actually have suggested), but can hardly be a reason to claim that the possessive pronominals are not subject to the Principle B at all. The question, however,

remains why the local domain of the possessive pronominals differs from that of other pronominals.[13]

8.2. Purpose Clauses

Russian purpose clauses containing the complementizer *čtoby* and an infinitive normally entail coreference of the understood subject with the matrix subject (for a detailed analysis of Russian purpose clauses, see Junghanns, 1994)[14]:

(37) Ivan priexal, [čtoby uvidet'sja s drugom].
Ivan.NOM came COMP see.INF with friend.INST
'Ivan came to see his friend.'

Despite the subject coreference, the infinitive clause in (38) is an IP with an empty pronominal in the subject position, rather than a bare VP. This is evidenced by the dative subject of SAM in it, cf.:

---

[13] The fact that the asymmetry between possessive and non-possessive pronominals holds both for non-subject-controlled infinitive clauses and simplex clauses with the dative subject supports the uniform structural analysis of the two kinds of clauses.

[14] With a group of verbs with causative semantics, *čtoby* -infinitives marginally allow object control as well:

(i) Komandir$_i$ ostavil Ivana$_j$ v derevne, čtoby
commander.NOM left Ivan.ACC in village.PREP COMP
pro$_j$ uxaživat' za ranenymi.
look.INF after injured.PL.INST
'The commander$_i$ left Ivan$_j$ in the village in order (for him$_j$) to look after the injured.'

Note also that in the other type of Russian purpose clauses, which also include an infinitive but lack a complementizer, object control is perfect, and binding possibilities of pronominals confirm to our analysis. For example, in (ii) the pronominal co-refers with the understood subject of the infinitive clause:

(ii) Komandir$_i$ ostavil Ivana$_j$ v derevne samomu
commander.NOM left Ivan.ACC in village.PREP self.DAT
uxaživat' za ego$_{i/j}$ ranenym drugom.
look.INF after his injured.INST friend.INST
'The commander$_i$ left Ivan$_j$ in the village in order (for him$_j$) to look after his$_{i/j}$ injured friend.'

(38) Ivan priexal, [čtoby pro_{DAT} samomu/ *sam
Ivan.NOM came COMP self.DAT/*self.NOM
uvidet'sja s drugom].
see.INF with friend.INST
'Ivan came to see his friend.'

This seems to be the only construction in Russian where the infinitive clause has a subject position despite the coreference between the matrix and embedded subjects. Babby (1996) explains this via barrierhood of the embedded CP node for theta-role saturation, which would be necessary if the embedded clause was a bare VP.

Crucially, the possessive pronominal inside the *čtoby*-clause, contrary to expectation, cannot be bound by its empty subject:

(39) $Ivan_i$ priexal, [čtoby pro_{DATi} uvidet'sja
Ivan.NOM came COMP see.INF
s $ego_{j/*i}$ drugom].
with his friend.INST
'$Ivan_i$ came to see $his_{j/*i}$ friend himself.'

At present, I cannot suggest a solution to this problem. It can only be noted that the exceptional status of the purpose clauses with respect to pronominal binding agrees with their exclusive status with respect to coreference in general, since, as already mentioned, these are the only full IPs in Russian which require coreference between their subject and the matrix subject.

8.3. Arbitrary *pro*

One more difficulty comes from infinitive clauses occupying the subject position. They are definitely IPs, as the dative case of SAM indicates:

(40) [Odnomu spravit'sja s takoj zadačej]
alone.DAT cope.INF with such.INST task.INST
neprosto.
difficult
'It is difficult to cope with such a task alone.'

The problem is that when the subject position of the infinitive in this construction is empty,[15] a pronominal cannot corefer with the subject, contrary to the prediction of our analysis:

(41) [pro_i idti na konflikt so svoim$_i$/ego$_{j/*i}$ načal'nikom]
go on conflict with own.INST/his boss.INST
— vrjad li mudroe rešenie.
hardly wise solution
'To get into conflict with the boss is hardly a wise move.'

It is important to note, however, that if the empty subject is replaced by an overt NP, that NP can corefer with a pronominal:

(42) [Ivanu$_i$ ujti s ego$_i$ novoj raboty]
Ivan.DAT leave.INF from his new job.GEN
označalo by poterjat' v žizni vsjakuju perspektivu.
would mean loose.INF in life.PREP any future.ACC
'For Ivan to leave his new job would mean to lose any future in this life.'

Given (42), we can conclude that in general infinitive clauses in the subject position fit the prediction of Principle B under the given definition of the local domain for Russian pronominals. The ungrammaticality of (41) needs a separate explanation, of course, which probably should be related to some properties of the arbitrary interpretation of empty pronominals.

[15] Most often in this case the subject is understood as generic, therefore traditionally this type of infinitive subject is termed 'arbitrary PRO'; under the analysis of infinitive constructions adopted here, this rather is 'arbitrary *pro*'.

9. Conclusions

I have shown that the coreference possibilities of possessive pronominals in Russian are not accounted for on the whole by Principle B as long as the standard definition of the local domain of pronominals is assumed, taking it always to be the minimal clause. I argued for another definition of the local domain for pronominals in Russian, which also helped to account for some peculiar binding effects outside the infinitive constructions. At the same time, it turned out that the proposed redefinition of the local domain is valid only for possessive, but not for other kinds of pronominals. Although this definitely is an undesired complication of the proposed analysis, I argued against replacing this analysis by an apparently simpler one (positing 'clause union'), but making some wrong predictions.

References

Babby, L. 1996. "Subject Control as Direct Predication: Evidence from Russian." Paper presented at the 6th Annual Workshop on Formal Approaches to Slavic Linguistics.

Comrie, B. 1971. *Aspects of Sentence-Complementation in Russian*. Ph.D. Dissertation, Cambridge University.

Comrie, B. 1974. "The Second Dative: A Transformational Approach." In Brecht, R. and C. Chvany (eds.) *Slavic Transformational Syntax*. Ann Arbor: Michigan Slavic Publications, 123-150.

Franks, S. 1995. *Parameters of Slavic Morphosyntax*. Oxford University Press.

Franks, S. and N. Hornstein. 1992. "Secondary Predication and Null Case." In Larson, R., S. Iatridou, U. Lahiri, and J. Higginbotham (eds.) *Control and Grammar,* 1-50. Dordrecht: Kluwer.

Gibson, J. and E. Raposo. "Clause Union, the Stratal Uniqueness Law, and the Chômeur Relation." *Natural Language and Linguistic Theory* 4, 95-331.

Hopper, P. and S. Thompson. 1980. "Transitivity in Grammar and Discourse. " *Language.*

Junghanns, U. 1994. *Syntaksische und Semantische Eigenschaften Russischer Finaler Infinitiveinbettungen.*" (Slavistische Beitrage 315). München: Verlag Otto Sagner.

Kazenin, K. 1999. Sintaksicheskie Nuli v Konstrukcijax s Infinitivami: Real'nost' Jazyka ili Artefakt Teorii? In *Trudy Mež-dunarodnogo Seminara "Dialog-99".* Moscow.

Kondrashova, N. 1994. "Agreement and Dative Subjects in Russian." In Avrutin, S., S. Franks, and L. Progovac (eds.) *Proceedings of the Second Annual Workshop on Formal Approaches to Slavic Linguistics. The MIT Meeting, 1993,* 255-285. Ann Arbor: Michigan Slavic Publications.

Laurençot, E. 1997. "On Secondary Predication and Null Case." In Lindseth, M. and S. Franks (eds.) *Proceedings of the Fifth Annual Workshop on Formal Approaches to Slavic Linguistics. The Indiana Meeting 1996,* 191-206. Ann Arbor: Michigan Slavic Publications,.

Neidle, C. 1988. *The Role of Case in Russian Syntax.* Dordrecht: Kluwer.

Rappaport, G . 1986. "On Anaphor Binding in Russian." *Natural Language and Linguistic Theory* 4, 97-120

Reinhart, T. and E. Reuland. 1993. "Reflexivity." *Linguistic Inquiry* 24.

Schoorlemmer, M. 1994. "Dative Subjects in Russian." In Toman, J. (ed.) *Functional Categories in Slavic Syntax. Proceedings of the First Annual Workshop on Formal Approaches to Slavic Linguistics: The Ann Arbor Meeting 1993,* 129-172, Ann Arbor: Michigan Slavic Publications,.

Williams, E. 1994. *Thematic Structure in Syntax.* Cambridge, Mass.: MIT Press.

Department of Theoretical and Applied Linguistics
Philological Faculty, MGU
Vorobjovy Gory
Moscow Russia
kazenin@philol.msu.ru

Lexical Analysis of Polish Multifunctional Reflexive Marker

Anna Kupść
Polish Academy of Sciences and Université Paris 7

1 Introduction

The aim of this paper is to argue for and propose a lexical analysis of (all types of) the Polish reflexive marker (RM) *się* 'self'.[1]

As in many Slavic and Romance languages, Polish RM is multifunctional. The anaphoric RM (ARM), unlike other types of RM, is usually treated as a counterpart of the anaphoric complement *siebie*. We provide evidence that ARM is not fully equivalent to *siebie*, either, and we propose a lexical analysis of all verbs that occur with RM. Our analysis is couched within the theoretical framework of HPSG (Head-driven Phrase Structure Grammar, cf. Pollard and Sag (1994)). We use lexical rules and lexical specification to account for various constructions with RM as well as these verb forms which obligatorily occur with RM, e.g., inherent reflexive verbs. We also show how this analysis interacts with an HPSG-encoded binding theory.

[1] We wish to thank Anne Abeillé, Małgorzata Marciniak, Adam Przepiórkowski, Marek Świdziński and editors of this volume for very useful comments and suggestions. We benefited also from the comments of audiences at the 1999 Spring Linguistic Colloquium at the University of North Carolina and the FASL-8 Workshop. Special thanks to Tracy Holloway King and Irina Sekerina for typesetting the paper in Word. The research presented in this paper was partly supported by the CRIT-2 grant.

2. Properties of RM

2.1. Multifunctional RM

In its most prominent function, RM *się* is often considered a weak counterpart of the anaphoric (both reflexive and reciprocal) complement *siebie* 'oneself', cf. (1).

(1) a. Janek$_i$ zobaczył się$_i$/siebie$_i$ w lustrze.
John saw self.RECP.ACC in mirror
John saw himself in the mirror.'

b. reciprocal
Dzieci$_i$ się$_i$/siebie$_i$ straszã
children Self.RECP.ACC frighten
opowieściami o duchach.
frighten about ghosts
'Children frighten each other with stories about ghosts.'

Polish anaphoric pronouns, e.g., ARM and *siebie*, have only the case value assigned but are unspecified for number, gender and person. Unlike in French or English, there is no morphological agreement between an anaphoric pronoun and its nominal antecedent.

In addition to (1), Polish RM has other functions. It can be an impersonal, a middle (or inchoative), or an inherent RM, (2a-c).[2]

(2) a. Impersonal
Czyta się/*siebie tę książkę
reads self.IMP this book
z przyjemnością.
with pleasure
'One reads this book with pleasure.'

---

[2] See e.g., Wilczewska (1966), Saloni (1976), Kański 1984, Kubiński (1987) for various classifications of types of *się*.

b. Middle
Zupa się /*siebie szybko gotuje.
soup self.MIDDLE quickly cooks
'Soup cooks quickly.'

c. Inherent
Jan boi się/*siebie Piotra.
John fears self.INH Peter
'John is afraid of Peter.'

In none of these functions RM can be replaced by the anaphoric complement *siebie*, unlike in (1). Verbs which occur in (2) are usually stored with *się* in the lexicon, unlike transitive verbs in (1). In subsequent sections we show, however, that the correspondence between *siebie* and ARM is not always maintained and we argue that the presence of ARM is lexically conditioned as well.

2.2. RM as a Weak Form

Cardinaletti and Starke (1994) provide an apparently universal classification of pronominals. Although the overall classification is not unproblematic, some properties which distinguish ARM from the complement *siebie* may follow from the fact that ARM is a weak pronoun whereas *siebie* is a strong form.

Weak forms, in contrast to strong forms cannot be used in isolation:

(3) Kogo widzisz na tym zdj•ciu?
who See-you on this picture
Siebie./*Się.
self.STRONG/self.WEAK
'Whom do you see in this picture? Myself.'

A weak pronoun, unlike a strong one, cannot occur as a conjunct in a coordinated structure:

(4) Jan umył siebie/*się i dzieci.
John washed self.STR/self.WEAK and children
'John washed himself and the children.'

Weak forms can never be modified:

(5) Jan całego siebie/*się umył.
John whole self.STR/self.WEAK washed
'John washed all of himself.'

Note that in (3)-(5) the presence of RM is not excluded in general; (1a) and (6) are fully grammatical.

(6) Jan się umył.
John self washed
'John washed himself.'

Positions of weak forms are more restricted than those of strong ones. If postverbal, RM must be adjacent to a verb, unlike *siebie*, cf. (7) (a postverbal RM can be separated from the verb only by another weak form).

(7) Jan umył wczoraj siebie/*się
John washed yesterday self.STR/self.WEAK
'John washed (himself) yesterday.'

RM cannot be used as a complement of a preposition, cf. (8).[3]

(8) Położył kurtkę obok siebie/*się
put-he jacket next-to self.STR/self.WEAK
'He put the jacket next to himself.'

[3] The only counterexample in the contemporary Polish seems to be a fixed expression *samo przez się* 'all by itself'. In Old Polish, however, weak forms after prepositions were fully correct, e.g., Klemensiewicz et al. (1955).

Note that the ungrammaticality of (8) does not follow from the fact that prepositions can never host RM. If a preposition has no complement, e.g., *dookoła* 'around' in (9), it may host RM.

(9) Szybko dookoła się$_{inh}$ rozejrzała.
quickly around self$_{inh}$ looked-she
'She quickly looked around.'

Also the sentence initial position of RM is excluded, cf. (10a).[4] This restriction, however, seems to follow from prosodic rather than syntactic constraints since (10b) is fully grammatical.

(10) a. Siebie/*się wczoraj nie umył.
self.STR/self.WEAK yesterday not washed-he
'He didn't wash himself yesterday.'

b. Piotr powiedział, że siebie/*się
Peter said COMP self.STR/self.WEAK
wczoraj nie umył.
yesterday not washed-he
'Peter said that he hadn't washed (himself) the other day.'

According to Cardinaletti and Starke (1994), strong forms, unlike deficient (weak) forms, may have only human reference. This distinction, however, is not respected, e.g., by Polish *siebie* and ARM. In (11), both pronouns may refer to dogs.

(11) Gdy psy$_i$ zobaczyły się$_i$/siebie$_i$
when dogs saw self.STR/self.WEAK
na wzajem zaczęły szczekać.
each other started-they bark.INF
'When dogs saw each other in the street, they started to bark.'

Since ARM and *siebie* successfully pass most of the tests to distinguish weak from strong forms proposed in Cardinaletti and

[4] In informal speech, sentence-initial positions of RM are more acceptable.

Starke (1994), we assume here that the presented contrasts follow from this distinction.

2.3. Restrictions on Anaphoric RM

In this section we show that various verbs may have only the strong form as a complement while ARM is excluded.

As (12) illustrates, certain transitive verbs may occur only with the anaphoric complement *siebie* whereas ARM is blocked.[5]

(12) a. Jan siebie/?*się
John oneself.REFL/RECP/self?*REFL/*RECP
lubi/ rozumie/ kocha.
likes understands loves
'John likes/understands/loves himself.'

b. Oni siebie/się
they oneself.REFL/RECP/self??REFL/RECP
lubiã/ rozumiejã/ kochajã.
like understand love
'They like/ understand/ love each other/ themselves.'

Such contrasts are easily explained if verbs are lexically (or semantically) specified whether they may occur with RM. The reflexive interpretation of RM is highly degraded in both (12) but a reciprocal interpretation is possible. Since the reciprocal pronoun requires a collective (plural) antecedent, it is available only in (12b).

In Polish, the presence of RM (in inherently reflexive verbs) need not exclude another nominal argument, e.g., the verb *bać się* 'be afraid' in (2c) has a genitive argument as well. If used with such verbs, the anaphoric complement cannot be substituted by ARM, cf. (13).

(13) Jan się siebie/*się boi.
John self.INH oneself.REFL/self.REFL.GEN fears
'John is afraid of himself.'

[5] This fact has been brought to our attention by Adam Przepiórkowski.

Impersonal constructions with *się*, e.g., (2a), are very productive in Polish and can be formed from various predicates, cf. e.g., Dziwirek (1994) for an overview. If the impersonal construction is formed from a verb which may have the anaphoric complement, the strong form *siebie* is possible but ARM cannot be used, cf. (14).

(14) Janowi łatwo się siebie/*się.
John$_{dat}$ easily Self.IMP oneself/self.REFL
myje w ciepłej wodzie.
washes in warm water
'John finds it easy to wash himself in warm water.'

Note that the ungrammaticality of (13) and (14) if ARM is used cannot be due to the fact that the presence of RM blocks occurrence of another anaphoric argument. The sentences with *siebie* are grammatical. Also if only one RM is present, the sentences are fully acceptable:

(15) a. Jan się boi.
John self.INH fears
'John is afraid.'
b. Janowi łatwo się myje w
John$_{dat}$ easy self$_{refl/imp}$ washes in
ciepłej wodzie.
warm water
'John finds it easy to wash (himself) in warm water.'

The contrasts in (13)-(14) and the grammaticality of (15) indicate that a verb can license at most one RM whatever its function is.[6]

Polish RM exhibits so-called *haplology*. Haplology presents an apparently paradoxical situation: one occurrence of a morpheme fulfills several functions of this morpheme. As illustrated in (16a),

[6] A similar observation is also made, e.g., in Rozwadowska (1992) or Rappaport (1997).

a single *się* serves both as an impersonal and an inherent RM (*przechadzać się* 'stroll' is an inherently reflexive verb). Several RMs cannot be used, (16b).[7]

(16) a. Przyjemnie się nam przechadza
nicely Self.INH/IMP us.DAT strolls
po parku.
in park
'We enjoy strolling in the park.'
b. * Przyjemnie się się nam przechadza po parku.

As observed above, Polish verbs may license at most one RM. Therefore, one might attribute the obligatory single occurrence of RM (13)-(15)to haplology. In contrast to (16a), however, sentences with a single RM in (15), do not have a haplology interpretation. (15a) has an 'object-drop' meaning, i.e., 'John is afraid', and does not have the intended haplology interpretation 'John is afraid of himself'. The sentence in (15b) is ambiguous. It can mean either 'John finds it easy to wash himself' (a haplology interpretation) or 'John finds it easy to wash something', washing of an unspecified object (an 'object-drop' meaning). Therefore haplology does not explain the ungrammaticality of (13)-(14) if several RMs co-occur.

Cross-linguistically, haplology is often morphologically conditioned and explained by the fact that affixes (inflectional or phrasal) cannot be doubled, e.g., Stemberger (1981), Zwicky (1987), Miller (1992). Nevertheless, a morphological explanation of haplology is not possible in the case of Polish RM. As extensively argued in Kupść (1999), Polish RM is a syntactic item (a postlexical clitic in the terminology of Miller (1992)). Unlike affixes, RM does not form a prosodic lexical unit with the host, does not have a fixed position and can be elided on a par with syntactic complements.

In Section 4, we provide a syntactic analysis which correctly accounts for data discussed so far, including haplology.

---

[7] For the discussion of haplology of Polish RM in other environments see Fowler (1993), Rappaport (1997) and Kupść (1999).

3. RM and Binding

In Polish, as in many other Slavic languages, binding of anaphoric pronouns is subject-oriented, see (17) (=(1)).[8]

(17) a. Janek$_i$ zobaczył się$_i$/siebie$_i$ w lustrze.
John saw Self.REFL.ACC in mirror
'John saw himself in the mirror.'

b. Reciprocal
Dzieci$_i$ się$_i$/siebie$_i$ straszã
children self$_{recp,acc}$ frighten
opowieściami o duchach.
frighten about ghosts
'Children frighten each other with stories about ghosts.'

If RM has other functions, see (2), it seems to be only a syntactic marker rather than an anaphoric pronoun. Although it is not clear to us whether these RMs should be subject to binding principles, anaphoric co-arguments of RM in inherent reflexives and middles are bound by the subject, cf.(18a) and (18b), respectively.

(18) a. Jan$_i$ bał się siebie$_i$
John feared self.INH oneself
'John was afraid of himself.'

b. Gałãź$_i$ złamała się pod
branch broke self$_{midd}$ under
ciężarem swoich$_i$ owoców.
weight self's fruits
'The branch broke under the weight of its fruits.'

In impersonal constructions, cf. (19), binding of anaphoric complements is less straightforward.

---

[8] Unlike reflexive pronouns, reciprocal pronouns may be bound by a non-subject NP as well, e.g., Marciniak (1999). We ignore this fact here.

(19) a. Trudno rozmawia się z innymi$_{*i}$
difficult talks self.IMP with others
o sobie$_i$
about self
'It is difficult to talk to other people about oneself.'

b. Trudno się mi$_i$ rozmawia z
difficult self.IMP me.DAT talks with
innymi$_{*i}$ o sobie$_i$
others about self
'It is difficult for me to talk to other people about myself.'

Unlike in (17) or (18), the anaphoric complement in (19a) refers to some unspecified (human or animate) object. In (19b), the binder is overtly expressed but it is a dative rather than a nominative phrase.

The specification of the syntactic subject (or a binder) in impersonal constructions has been a long-standing issue of Polish grammar, cf. e.g., Wilczewska (1966), Saloni (1976), Kański (1984), Kubiński (1987). Saloni (1976) argues that in such constructions there is a nominative subject, namely *się*. Hence in (19a), RM can be an overt binder for the anaphoric *sobie*. In (19b), however, the complement is coindexed with the dative phrase *mi* rather than with RM. Note that the main difference between both sentences is the presence of NP.DAT in (19b) and lack thereof in (19a). Due to these similarities, RM in both constructions should have the same status. If RM is the syntactic subject in both constructions, then the obligatory coindexation with NP.DAT in (19b) could be obtained only if RM and NP.DAT were coindexed. Observe that if RM is the subject, it (locally) commands NP.DAT. Thus (pronominal) NP.DAT in (19b) would be (locally) bound ((locally) commanded by and coindexed with *się*), in violation of Principle B (or Principle C if NP.DAT is an R-expression).

The account we propose is quite different. We assume that impersonal constructions are subjectless, i.e., there is no overt subject in any of (19). Neither RM nor the dative NP in (19) are syntactic (surface) subjects. We assume that in impersonal

constructions the binder need not be overtly realized but it is always present at a different (semantic) level. Once realized, however, it surfaces as NP.DAT.

A similar distinction between covert and overt subjects has been used in Marciniak (1999) to account for binding of arguments of participial impersonals:

(20) Rozmawiano o Tomku$_{*i}$ ze sobą$_i$
talked.PART.IMP about Tom about self
'They were talking to each other about Tom..

In (20), no overt subject can be present but the anaphoric pronoun must be bound by an implicit argument (a syntactically unrealised subject). Coindexation with any other argument is ungrammatical.

4. HPSG Account

The analysis presented in this section is mostly based on that of Kupść (1999), extended to account for binding facts discussed in section 3.

4.1. General Assumptions

This lexical entry comprises all types of RMs distinguished in
As stated in section 2.3, Polish RM is a syntactic item. A schematic lexical entry for *się* is given in (21).

(21)
$$\left[\begin{array}{l} \textit{word} \\ \text{PHON } \langle \text{się} \rangle \\ \text{SS} \left[\begin{array}{l} \textit{clitic} \\ \text{LOC} \left[\begin{array}{l} \textit{local} \\ \text{CAT}|\text{HEAD } \textit{noun} \\ \text{CONT} \left[\begin{array}{l} \textit{ana} \\ \text{INDEX } \textit{index} \\ \text{RESTRICTION } \textit{set\_psoa} \end{array}\right] \end{array}\right] \end{array}\right] \end{array}\right]$$

This lexical entry comprises all types of RMs distinguished in Section 2.1. We assume that there is only one lexical entry for *się*, e.g., contrary to Saloni (1976). Various functions of RM, exemplified in (1) and (2), will be established by (semantics of) a verb. Referential properties of nominal objects (hence RM as well) are encoded as the value of the INDEX attribute, i.e., *index*, which can be split into *referential* and *non-referential* subsorts. The RESTRICTION (RESTR) attribute allows one to restrict content of nominals, e.g., require NPs to denote only animate objects.

We follow Sag (1997), Miller and Sag (1997), Abeillé et al. (1998) and split *synsem* objects into subsorts (*synsem*s represent syntactico-semantic properties of linguistic objects). In order to distinguish RM from other NPs, we introduce subsorts *clitic* and *non-clitic*, correspondingly.[9]

We assume that RM is subcategorized for by a verb on a par with other complements.[10] In HPSG, the subcategorization frame is encoded via VALENCE (VAL) attributes. A schematic lexical entry for the verb *boi się* 'fears' is given in (22).

(22)
$$\left[\begin{array}{l} \textit{word} \\ \text{PHON } \langle \text{boi} \rangle \\ \text{SS|LOC|CAT} \left[\begin{array}{l} \text{HEAD } \textit{verb} \\ \text{VAL} \left[\begin{array}{l} \text{SUBJ } \boxed{1}\langle \text{NP}[\textit{str}] \rangle \\ \text{COMPS } \boxed{2}\langle \textit{ana-cl}, \text{NP}[\textit{lgen}] \rangle \end{array}\right] \\ \text{ARG-ST } \boxed{1} \oplus \boxed{2} \end{array}\right] \end{array}\right]$$

The VAL attributes contain information about arguments of the (lexical or phrasal) head. Values of these attributes are lists of *synsem*s. The ARG-ST (ARGUMENT-STRUCTURE) attribute is specified only for *word*s and is the locus of HPSG binding theory, cf section 4.6. ARG-ST and VAL are related. In most cases, ARG-ST

[9] The Polish *się* is a syntactic clitic, cf. Kupść (1999). Thus, *clitic* is a subsort of the *canonical synsem* (on a part with *non-clitic*), unlike French morphological clitics in Miller and Sag (1997).

[10] Unlike Kupść (1999), we do not use a separate attribute for.

is just concatenation of VAL|SUBJ and VAL|COMPS lists (in this order) but this relationship need not always be preserved, cf. Manning and Sag (1999). In a more elaborated account, once the precise relationship between ARG-ST and VAL is established, it is sufficient to specify only the ARG-ST value in the lexicon as VAL can be deduced. Since we do not formalise the principle which relates them, we explicitly specify both VAL and ARG-ST in the paper.

We use *ana-cl* as the abbreviation for the SYNSEM (SS) value given in the lexical entry for *się* in (21). Hence, RM is a (subcategorized for) nominal object of the sort *clitic* with (referential or not) anaphoric properties ([CONT *ana*]).

We assume that RM is present on ARG-ST, cf. (22). Therefore if RM bears a referential index, it will obey binding principles, cf. Section 4.6. Also if case assignment is encoded on ARG-ST, as proposed in Przepiórkowski (1999), the (structural) case value of RM will be resolved by the case principle.[11]

As discussed in Section 2.3, a verb can license at most one RM. The following constraint ensures that ARG-ST contains at most one *ana-cl*:[12]

$$(23)\quad \left[\begin{array}{l} \textit{word} \\ \text{SS|LOC|CAT|ARG-ST}\ \boxed{1}\textit{ne\_list}(\textit{ana-cl})\ \bigcirc\textit{list}(\textit{non-ana-cl}) \end{array}\right] \rightarrow \boxed{1}\langle \textit{ana-cl} \rangle$$

This constraint is implicational and does not force verbs to occur with *się*. It says that if *ana-cl* is present on the ARG-ST list, this is the only *ana-cl* element of this list. We use *non-ana-cl* as the abbreviation for a *synsem* object distinct from *ana-cl*.

Since RM is a syntactic item, traditional HPSG syntactic principles, i.e., the Valence Principle and Immediate Dominance

---

[11] We follow Przepiórkowski (1996) and adopt lexical vs. structural case dichotomy for Polish.

[12] The '0' indicates the 'shuffle' relation introduced in Reape (1992). This relation operates on lists and 'shuffles' their elements in such a way that the relative order of elements of original lists is preserved.

Schemata, ensure its realization. No additional mechanisms are necessary.

Having introduced these general assumptions of our analysis, we may proceed to the account of data discussed in previous sections.

4.2. Inherent Reflexives

Inherent reflexive verbs do not have a non-reflexive counterpart and RM is specified directly in the lexical entry. The lexical entry for *boi się* 'fears' given in (22) is repeated as (24).

Due to the constraint in (23) and the lexical entry in (24), multiple RMs in (13) are correctly ruled out. The verb has two elements on the COMPS list: *się* and a genitive NP. The realization of the nominal genitive argument as another *się* is correctly blocked by (23). On the other hand, nothing prevents the co-occurrence of several anaphoric arguments, i.e., (13) is licensed as desired. There is no principle that forces haplology in general. Hence, an 'object-drop' reading in (15a) is predicted: only RM is realised while the complement is missing.

(24)
$$\left[\begin{array}{l} \textit{word} \\ \text{PHON } \langle \text{boi} \rangle \\ \text{SS|LOC|CAT} \left[\begin{array}{l} \text{HEAD } \textit{verb} \\ \text{VAL} \left[\begin{array}{l} \text{SUBJ } \boxed{1}\langle \text{NP}[\textit{str}] \rangle \\ \text{COMPS } \boxed{2}\langle \textit{ana-cl}, \text{NP}[\textit{lgen}] \rangle \end{array}\right] \\ \text{ARG-ST } \boxed{1} \oplus \boxed{2} \end{array}\right] \end{array}\right]$$

4.3. Verbs with ARM

Verbs such as *zobaczył* 'saw-he' or *straszã* 'frighten-they' in (1) can occur both with ARM or *siebie*. We assume that their COMPS value is left unspecified, cf. (25), i.e., the complement NP can be of any *synsem* sort. Hence it can be realized either as *ana-cl*

or any other accusative NP. This correctly predicts that in (1) both forms, *się* and *siebie*, are possible.

(25)
$$\left[\begin{array}{l} \textit{word} \\ \text{PHON } \langle \text{zobaczył} \rangle \\ \text{SS|LOC|CAT} \left[\begin{array}{l} \text{HEAD verb} \\ \text{VAL} \left[\begin{array}{l} \text{SUBJ } \boxed{1}\langle \text{NP}[\textit{str}] \rangle \\ \text{COMPS } \boxed{2}\langle \text{NP}[\textit{str}] \rangle \end{array}\right] \\ \text{ARG-ST } \boxed{1} \oplus \boxed{2} \end{array}\right] \end{array}\right]$$

For verbs such as in (12), we stipulate here that their only NP complement cannot be the *reflexive* clitic, i.e., *reciprocal* or *non-clitic:refl* . This rules out ungrammatical forms with the reflexive RM. Certainly, a more principled (semantic) explanation should be proposed since similar restrictions hold in other languages, e.g., Serbian/Croatian, according to Wayles Browne (p.c.). We leave this issue for future study.

4.4. Impersonal Verbs

We assume that impersonal verbs with *się* are obtained via a lexical rule[13] given schematically in (26).[14]

[13] See Monachesi (1995) who uses lexical rules to account for the Italian reflexive marker *si*.

[14] NP: [1] (boxed 1) abbreviates an NP with the [SS|LOC|CONTENT [1]] path value.

(26)
$$\left[\begin{array}{l} \textit{word} \\ \text{SS|LOC|CAT} \left[\begin{array}{l} \textit{category} \\ \text{HEAD } \textit{verb} \\ \text{VAL} \left[\begin{array}{l} \textit{val} \\ \text{SUBJ } \boxed{6}\langle \text{NP}[\textit{str}]:\boxed{1}\rangle \\ \text{COMPS } \boxed{4}\text{list}(\textit{ana-cl})\oplus\boxed{2}\text{list}(\textit{non-ana-cl}) \end{array}\right] \\ \text{ARG-ST } \boxed{6}\oplus\boxed{4}\oplus\boxed{2} \end{array}\right] \end{array}\right]$$

$$\mapsto \left[\text{SS|LOC|CAT} \left[\begin{array}{l} \text{HEAD|VFORM } \textit{fin}_{\textit{3rd\_neut}} \\ \text{VAL} \left[\begin{array}{l} \text{SUBJ } \langle\rangle \\ \text{COMPS } (\boxed{3}\langle \text{NP}[\textit{ldat}]:\boxed{1}\rangle\oplus)\boxed{5}\langle \textit{ana-cl}\rangle\oplus\boxed{2} \end{array}\right] \\ \text{ARG-ST } \boxed{3}\oplus\boxed{5}\oplus\boxed{2} \end{array}\right]\right]$$

Although impersonal constructions are quite productive in Polish, they are not allowed with all verbs. First, impersonals are possible only with verbs which require a human (or animate) subject. This can be obtained by restricting the value of input's VAL|SUBJ to animate NPs (via the CONT|RESTR value). Second, certain psych-verbs, e.g., *bać się* 'to be afraid' do not have impersonal forms with *się* although they do require an animate subject. Thus (26) has to be further restricted. Precise semantic constraints are not clear to us at the moment and we concentrate on syntactic issues mainly.

(26) provides only this information which distinguishes input and output verbs. The ARG-ST lists of the input and output verbs can be distinct if the input verb is not reflexive, i.e., if there is no *ana-cl* on the ARG-ST list. Note that the COMPS value of the input contains a list of *ana-cl*, i.e., [4] may be either empty or non-empty. Distinct VAL values of input and output verbs correctly block recursive application of the rule. The output verb must occur in third neuter singular form. This verb form is predicted by the subject-predicate agreement principle proposed in Czuba and Przepiórkowski (1995).

The lexical rule in (26) licenses only one RM in the output verb. Irrespectively of the input specification of the verb only one RM results. If [4] is not empty, (26) correctly rules out (14) and (16b) and predicts that only one RM appears in (15b) and (16a). The optionality of the dative argument in impersonal constructions, cf. (2a) vs. (18a) and (16a), is indicated by the brackets around [3].

Note, however, that NP[dat] is always present on the ARG-ST. This fact will be crucial for binding in section 4.6.

Our analysis accounts also for the ambiguity of (15b). We assume that *myje* 'washes' has the lexical entry analogous to (25), i.e., the complement can be either realized as *ana-cl* or any other NP. If in the input of (26) a non-reflexive verb is used, an 'object-drop' reading results. Only the impersonal RM is realised while the complement of the transitive *myje* 'washes' remains unrealized, i.e., [2] is non-empty. However, if the input verb occurs with ARM, haplology results. The input's NP complement is realized as *ana-cl* and [2] is the empty list. Hence in the output of (26), RM in [5] serves both as an impersonal RM and ARM.

4.5. Middles

We propose a (schematic) lexical rule in (27) to relate transitive verbs and their 'middle' counterparts.

As in the case of impersonals, there are semantic restrictions on formation of middle constructions. For example, (27) should mainly apply to verbs which have a non-animate complement. Thus, the CONT value of input's complement should be restricted. Middle constructions in Polish are not very productive and much stronger semantic restrictions should apply. At present, we assume that middle forms are derived by the lexical rule (rather than enumerated in the lexicon) and we discuss briefly syntactic aspects of (27).

(27)
$$\left[\begin{array}{l} \textit{word} \\ \text{SS|LOC|CAT} \left[\begin{array}{l} \textit{category} \\ \text{HEAD } \textit{verb} \\ \text{VAL} \left[\begin{array}{l} \textit{val} \\ \text{SUBJ } \boxed{2}\langle \text{NP}[\textit{str}] \rangle \\ \text{COMPS } \boxed{3}\langle \text{NP}[\textit{str}]{:}\boxed{1} \rangle \end{array}\right] \\ \text{ARG-ST } \boxed{2} \oplus \boxed{3} \end{array}\right] \end{array}\right]$$
$$\mapsto \left[\text{SS|LOC|CAT} \left[\begin{array}{l} \text{VAL} \left[\begin{array}{l} \text{SUBJ } \boxed{4}\langle \text{NP}[\textit{str}]{:}\boxed{1} \rangle \\ \text{COMPS } \boxed{5}\langle \textit{ana-cl} \rangle \end{array}\right] \\ \text{ARG-ST } \boxed{4} \oplus \boxed{5} \end{array}\right]\right]$$

The effect of (27) is to change the complement of the (transitive) input verb into the output's subject and add RM. We do not impose any restrictions on the COMPS value of the input verb in spite of the fact that middles cannot be formed from verbs that occur with RM. The derivation of middle forms from such verbs, however, will be blocked. In Polish, anaphoric NPs are not subjects. The SUBJ value of the output is non-empty and NP's CONTENT value, i.e., [1], is token-identical with that of the complement of the input. Thus, the output's subject has the same semantic properties as the complement of the input. If the latter is anaphoric, the general principle of Polish blocks its realization as the subject. This also blocks recursive application of (27) (the output's VAL|COMPS is *ana-cl*).

4.6. Binding Theory

As mentioned in Section 4.1, HPSG binding theory is encoded on the ARG-ST list, cf. Abeillé et al. (1998), Manning and Sag (1999). In HPSG, binding theory is defined in terms of the relative obliqueness of grammatical functions rather than in purely configurational terms. Pollard and Sag (1994) define o-command (obliqueness-command), (28), and o-binding (obliqueness-

binding), (29), relations which serve for the definition of binding principles, (3).[15]

(28) Let X and Y be *synsem* objects, X referential. Then X *locally o-commands* Y just in case X is less oblique than Y (X precedes Y on the ARG-ST list).

(29) X *locally o-binds* Y just in case X and Y are coindexed and X locally o-commands Y.

(30) PRINCIPLE A
A locally o-commanded anaphor must be locally o-bound.

Such a formulation of anaphoric binding accounts for these cases when 1) there is a local binder, i.e., a preceding element on ARG-ST exists, and 2) the local binder is referential. Otherwise, anaphoric elements present on ARG-ST are exempt (30) but can be bound on other non-syntactic grounds.

In Polish, (reflexive) anaphora are subject-oriented[16] and the definition of the binder must be more specific. We quote below relevant definitions of the binding theory for Polish presented in Marciniak (1999).

(31) Let X and Y be *synsem* objects, where X is referential. Then X *locally s-commands* Y if X and Y are elements of the same ARG-ST list and X is the first element of this ARG-ST list; ...

(32) X *locally s-binds* Y just in case X and Y are coindexed and X locally s-commands Y.

(33) PRINCIPLE A
A reflexive anaphor must be locally s-bound.

---

[15] We present only these definitions of Pollard and Sag (1994) which are relevant for our present discussion.
[16] See Fn. 8.

As mentioned in Section 4.1, the ARG-ST value is usually the concatenation of values of VAL attributes in the canonical oblique-based order, i.e., the subject (VAL|SUBJ) plus complements (VAL|COMPS), e.g., (24) and (25). For example, the ARG-ST list for *boi się* 'fears' in (18a) is given as follows:

(34) ARG-ST < NP_i[*str*]:*npro*, *ana-cl*, NP_i[*lgen*]:*ana* >

The definitions in (31)-(33) correctly predict that *siebie* refers to *Jan* 'John'. The (referential) complement *siebie*, indicated as NP_i[*lgen*]:*ana*, and the subject *Jan*, i.e., NP_i[*str*]:*npro*, are elements of the same ARG-ST list. According to (31), the latter locally s-commands the former. Since these NPs are coindexed, *Jan* locally s-binds *siebie*, cf. (32), and the Principle A in (33) is satisfied as desired. The binding of anaphora in (17) (and (18b)) is obtained analogously.[17]

This theory accounts also for binding in impersonal constructions. The ARG-ST value for *rozmawia* 'talks' in (19) is given in (35).

(35) ARG-ST <NP_i[*ldat*]:*npro*, *ana-cl*, PP_{*i}[z]:*npro*, PP_i[o]:*ana*>

The dative phrase is the *first* element on ARG-ST. As mentioned above, NP[dat] need not be overtly realised. For example, in (2a) and (19a), there would be no corresponding NP[dat] on the VAL|COMPS lists. Since NP[dat] is always present on ARG-ST of impersonal verbs, the anaphoric complements, e.g., PP *o sobie* 'about oneself' in (19) are locally s-bound. In (19a), NP[dat] denotes an unspecified (animate) object while in (19b) this object is overtly expressed as *mi*. Due to definitions in Marciniak (1999), NP[dat] locally s-commands *sobie*. Since these NPs are coindexed,

[17] The anaphoric pronoun in (18b) is not a direct argument of the verb. If we apply remaining (not quoted here) clauses of *the local s-command* definition in Marciniak (1999), binding is correctly obtained.

NP[dat] locally s-binds the anaphora, (32), and (33) is satisfied as desired.

Note that if RM were the first element of ARG-ST in (35), the correct binding in (19b) would not be predicted. If RM was referential, then *się* (rather than NP[dat]) would be the local s-binder for the anaphora. One can suggest that NP[dat] might become a local s-binder indirectly, i.e., if it were coindexed with (referential) RM. Note, however, that in this case (non-anaphoric) NP[dat] becomes (locally) s-bound, contrary to Principle B and C. On the other hand, if the first element of ARG-ST was a non-referential RM, binding of PP *o sobie* with *mi* would be exempt from our binding theory. Therefore, s-binding with NP[dat] might be predicted on different grounds. If (33) does not hold, however, nothing prevents illegal coindexation of the anaphora with PP *z innymi*.

5. Conclusions

In this paper we presented a lexical analysis of the Polish reflexive marker *się*. Such an analysis allows us to account for these cases when the correspondence between the anaphoric complement *siebie* and ARM is not maintained. Our approach accounts also for so-called haplology of RM. Although Polish verbs may license at most one RM, haplology does not always result. Our analysis correctly predicts this fact since haplology is accounted for by a lexical rule which derives impersonals. Due to the lexical specification of the ARG-ST value for impersonals, we capture problematic bindings in impersonal constructions.

Although the presented analysis correctly deals with the data, it is mostly syntactic. For example, we have to stipulate the subcategorization frame for verbs that do not allow ARM whereas a more principled semantic explanation should be proposed. Also lexical rules we sketched reflect mostly syntactic properties of discussed constructions. Semantic restrictions have not been incorporated. We hope to resolve these issues in future study.

References

Abeillé, A., D. Godard, and I. A. Sag. (1998). "Two Kinds of Composition in French Complex Predicates." In Hinrichs, E., A. Kathol, and T. Nakazawa (eds). *Complex Predicates in Nonderivational Syntax*, volume 30 of *Syntax and Semantics,* 1-41. Academic Press, New York.

Borsley, R.D. and A. Przepiórkowski. (eds.) (1999). *Slavic in Head-Driven Phrase Structure Grammar.* CSLI Puclications, Stanford.

Cardinaletti, A. and M. Starke. (1994). "The Typology of Structural Deficiency on the Three Grammatical Classes." *University of Venice Working Papers in Linguistics* 4(2), 41-109.

Czuba, K. and A. Przepiórkowski. (1995). *Agreement and Case Assignment in Polish: An Attempt at a Unified Account.* Technical Report 783, Institute of Computer Science, Polish Academy of Sciences.

Dziwirek, K. (1994). *Polish Subjects.* Garland, New York.

Fowler, G. (1993). "A syntactic Account of Derivational *–sja* in Russian." In Maguire, R. and A. Timberlake (eds.) *American Contributions to the Eleventh International Congress of Slavists*, 270-284. Slavica Publications, Columbus, OH.

Kański, Z. (1984). *Arbitrary Reference and Reflexivity: A Generative Study of the Polish Pronoun się and its English Equivalents.* Ph. D. Dissertation, University of Silesia, Katowice.

Klemensiewicz, Z., T. Lehr-Spławiński, and S. Urbańczyk. (1955). *Gramatyka Historyczna Języka Polskiego.* Państwowe Wydanwnictwo Naukowe.

Kubiński, W. (1987). *Reflexivization in English and Polish: An Arc Pair Grammar Analysis*, volume 178. *Linguistische Arbeiten.* Niemeyer, Tuebingen.

Kupść, A. (1999). Haplology of the Polish Reflexive Marker. In Borsley, R. D. and A. Przepiórkowski (eds.), 91-124.

Manning, C. and I. A. Sag. (1999)."Dissociations Between Argument Structure and Grammatical Relations." In Webelhuth, G. et al. (eds.), 63-78.

Marciniak, M. (1999). "Toward a Binding Theory for Polish." In Borsley, R. D. and A. Przepiórkowski (eds.), 125-147.

Miller, P. (1992). *Clitics and Constituents in Phrase Structure Grammar.* Garland, New York.

Miller, P. and I. A. Sag. (1997). "French Clitic Movement Without Clitics or Movement." *Natural Language and Linguistic Theory* 15, 573-639.

Monachesi, P. (1995). *A Grammar of Italian Clitics.* Ph.D. Dissertation, Tilburg University, Tilburg. ITK Dissertations Series 1995-3 and TILDIL Dissertation Series 1995-3.

Pollard, C. and I. A. Sag. (1994). *Head-Driven Phrase Structure Grammar.* Chicago University Press, Chicago.

Przepiórkowski, A. (1996). "Case Assignment in Polish: Towards an HPSG analysis." In Grover, C. and E. Vallduví (eds.) *Studies in HPSG*, volume 12 of *Edinburgh Working Papers in Cognitive Science, 191-228.* Center for Cognitive Science, University of Edinburgh.

Przepiórkowski, A. (1999). "On Case Assignment and 'Adjuncts as Complements'". In Webelhuth, G. et al. (eds.), 231-245.

Rappaport, G. (1997). "Clitics as Features: A Non-Semiotic Approach." To appear in *American Contributions to the International Congress of Slavists.* Version of 7th March 1997.

Reape, M. (1992). *A Formal Theory of Word Order: A Case Study in West Germanic.* Ph.D. Dissertation, University of Edinburgh.

Rozwadowska, B. (1992). *Thematic Constrains on Selected Constructions in Polish and English.* Wydawnictwa Uniwersytetu Wrocławskiego, Wrocław.

Sag, I. A. (1997). "English Relative Clause Constructions." *Journal of Linguistics* 33, 431-483.

Saloni, Z. (1976). *Cechy Składniowe Polskiego Czasownika.* Ossolineum, Wrocław.

Stemberger, J. (1981). "Morphological Haplology." *Language* 57, 791-817.

Webelhuth, G., J.-P. Koenig, J.-P., and A. Kathol (eds.) (1999). *Lexical and Constructional Aspects of Linguistic Explanation.* CSLI Publications, Stanfod.

Wilczewska, K. (1966). "Czasowniki Zwrotne We Współczesnej Polszczyźnie." *Towarzystwo Naukowe w Toruniu, Prace Wydziału Filolgiczno-Filozoficznego.* XVII(3).

Zwicky, A. (1987). "Suppressing the Zs." *Journal of Linguistics* 23, 133-148.

Institute of Computer Science
Polish Academy of Sciences
21 Ordona
01 237 Warzsawa
Poland
aniak@ipipan.waw.pl

The Typology of Multiple Wh-Fronting in Slavic Revisited*

Mariana Lambova
University of Connecticut

In this paper I look at Multiple Wh-Fronting (MWF), a topic that has been extensively discussed in the literature. Early studies are those of Wachowitz (1974) and Toman (1981). More recently, the work of Rudin (1988), Koizumi (1994), Richards (1997) and especially Bošković (1996; 1997c; 1998) has brought to light significant insights on the subject.

Since Rudin (1988) it has been standardly assumed that MWF languages fall into two types. In some, like Bulgarian (BG), preposed wh-phrases are all in SpecCP while in others, like Serbian/Croatian (SC), only the first one is located there and the rest are adjoined to IP. Rudin provides several types of evidence for her claim but I will discuss only her constituency data. Specifically, I will show that in a sequence of fronted wh-phrases BG, like SC, allows intervening lexical material after the first wh-phrase. I will conclude, contra Rudin, that preposed wh-phrases do not form a constituent in this language. The facts considered will be shown to support the Economy/Focus movement account of MWF (Bošković, 1997c) and provide further empirical evidence for the possibility of pronouncing lower copies of movement (Franks, 1998; Bošković, forthcoming).

The paper is organized as follows. First, I review Rudin's constituency evidence and present new data on the penetrability of preposed wh-phrases in BG. Next, I observe that the split is not allowed in the presence of a topic. To account for this, I will propose a slight modification of Bošković's (1997c) account: the wh-

* For encouragement and helpful discussion, I am indebted to Howard Lasnik, David Michaels, and especially to Željko Bošković. Thanks are also due to my informants: Ilka Lambova, Danka Ilieva, Anna Toncheva, Ivan Karamanov, Keti Polendakova, Ivan Koprinkov, Nadežda Slavova, Neda Dimitrova-Gilina, Liljana Vrublevska. As is usual, the blame for the conclusions drawn is solely mine.

cluster, in the sense of Rudin (1988), is in PolP, a projection immediately below CP which licenses both focus and topic, and only one wh-phrase is in SpecCP. Following Kim (1997) I assume a prior formation of the wh-cluster which subsequently moves as a complex to SpecPolP. Finally, I show that the presence of a topic forces pronunciation of a lower copy of the first wh-phrase.

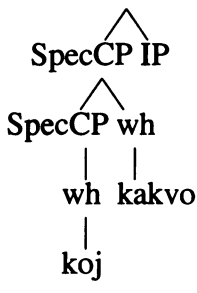
MWF is a characteristic of Slavic, and involves overt movement of all wh-phrases to the front of the sentence. The following examples with wh-phrases "in situ" are not acceptable as true questions in either BG or SC:

(1) a. *Koj kupuva kakvo? [BG]
who buys what
b. Koj kakvo kupuva?

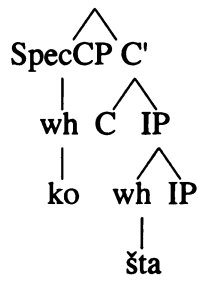
(2) a. *Ko kupuje šta? [SC]
who buys what
b. Ko šta kupuje?

However, according to Rudin, this similarity is only superficial. She claims that the two languages represent two different types of MWF. In BG, all fronted wh-phrases are in SpecCP. In SC, only the first wh-phrase is located there. Rudin (1988:479ff) proposes the following structures for multiple questions: rightward adjunction to SpecCP in BG and leftward adjunction to IP in SC:

(3) a. [BG]

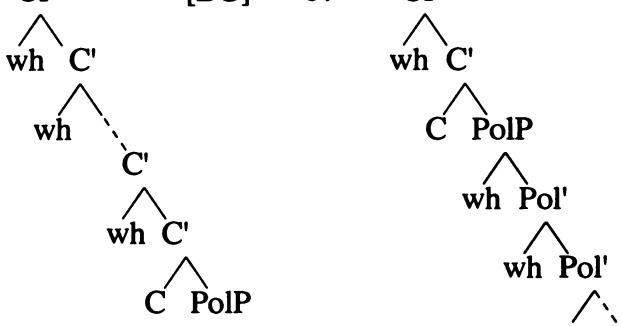


b. [SC]



Koizumi (1994) offers a reanalysis based on the layered specifiers hypothesis. The licenser of multiple specifiers is a different head in the two languages: C in BG and Pol in SC. Accordingly, the structures are:

(4) a. [BG] b. [SC]



Since nothing in this paper depends crucially on the choice between adjunction and multiple specifiers I will adopt Rudin's analysis for ease of exposition.

The reason for proposing that all BG wh-phrases are in SpecCP is that, according to Rudin, fronted wh-phrases in this language form a constituent that cannot be split:

(5) $[_{SpecCP}$ WH WH WH] ... $[_{IP}$ ...] [BG]

In support of her claim Rudin shows that no lexical material can intervene between BG preposed wh-phrases. Thus the auxiliary clitic can (and must) follow the first wh-phrase in SC but not in BG:

(6) a. *Koj e kakvo kupil? [BG]
who AUX.PRES.3P.SG what bought
b. Koj kakvo e kupil?
'Who bought what?'

(7) a. Ko je šta kupio? [SC]
who AUX.PRES.3P.SG what bought
b. *Ko šta je kupio?

This is allegedly true of parentheticals (8), as well as adverbs (9):

(8) a. *Koj, spored tebe, kakvo pie? [BG]
who according-to you what drinks
b. Ko, po tebi, šta pije? [SC]
'Who, according to you, is drinking what?'

(9) a. *Koj prâv kogo e udaril?[1] [BG]
who first whom AUX.PRES.3P.SG hit
b. Ko je prvi koga udario? [SC]
'Who hit whom first?'

Bošković (1998: 2) notes in passing that the distinction between BG and SC with respect to penetrability is not clear cut as some BG speakers allow the split. Note that the clitics evidence is irrelevant since, in contrast to SC clitics, BG clitics are verbal, and therefore must be adjoined to the verb.[2,3] Significantly, the data on parentheticals and adverbs are incorrect. In what follows I will be concerned exclusively with BG and will put SC aside.

---

[1] Rudin's (1988:467) original example [42b] is an indirect question. I changed it to have a minimal pair. (It is actually good, like [9a]; see below.):

(i) *Zavisi ot tova koj prâv kogo e udaril?
depends on this who first whom AUX.PRES.3P.SG hit

[2] SC clitics are second position clitics. Given the standard assumption that second position clitics can follow either the first word or the first constituent, the crucial example for SC is (7b): it shows that SC fronted wh-phrases do not form a constituent. For a recent discussion of SC clitics, see Franks (1998) and Bošković (in press a; forthcoming).

[3] As noted in the literature (Mišeska-Tomić, 1996:867, fn. 51; Bošković, 1999:5) the BG question particle *li* , arguably a second position clitic, can split fronted wh-phrases. Rudin (1994:260) judges (i) as marginal:

(i) Koj li kakvo na kogo e dal?
who Q what to whom AUX.PRES.3P.SG given
'Who gave whom what, I wonder?'

The facts are slightly more complicated in that *li* can occur further down in the clause, as in (ii), a fact I will not discuss here as it does not affect my argument:

Consider first the possibility for various particles to break the wh-cluster. The following sentences are fully acceptable:

(10) a. Koj čunkim kakvo mi e dal?
who dub.part. what me.CL AUX.3P.SG given
'Who, for God's sake, has given me what for free?'

b. Koj ma /be kakvo iska ot tebe?
who voc.part.F/M (substand.) what wants from you
'Hey, who wants what from you?'

c. Koj pâk kakvo ti dava na tebe?
who emph.part. what you.CL gives to you
'Who gives you (of all people) what?'

Next, all speakers I consulted allow without exception parentheticals, contra Rudin's (1988:468; ex. [46a]) judgment:

(11) Koj, spored tebe, kakvo pie? [=(8a)]
who according-to you what drinks.
'Who, according to you, is drinking what?'

(12) a. Koj, misliš, kakvo šte kaže?
who think.2P.SG what will say
'Who, you think, will say what?'

b. Koj, sâs sigurnost, kakvo koga šte donese?
who for sure what when will bring
'Who, for sure, will bring what when?'

The same applies to the example with an intervening adverb, repeated below, which is perfect:

(13) Koj prâv kogo e udaril? [=(9a)]
who first whom AUX.PRES.3P.SG hit
'Who hit whom first?'

(ii) Koj (li) kakvo (li) koga (li) šte ni servira?
who Q what Q when Q will us.CL surprise-with
'Who will surprise us with what when, I wonder?'

There is no disagreement regarding example (13)'s grammaticality status, or that of the original in Fn.1 above. Furthermore, various other adverbs are allowed in this position.[4] Notice also that ambiguous adverbs (15b) preserve both readings:

(14) Koj včera kakvo e zagubil?
who yesterday what AUX.PRES.3P.SG lost
'Who lost what yesterday?'

(15) a. sentential adverbs
Koj sigurno kakvo e kupil?
who probably what AUX.PRES.3P.SG bought
'Who has probably bought what?'
b. sentential and manner readings
Koj umelo kakvo e prikril?
who cleverly what AUX.PRES.3P.SG hidden
'Who has cleverly hidden what?'
c. manner adverbs
Koj grozno kakvo e bojadisal?
who ugly what AUX.PRES.3P.SG painted
'Who has painted what uglily?'

I have just shown that the sequence of fronted wh-phrases can be split. Now there is a possibility for a topic to precede the wh-phrases. Surprisingly, when this happens the wh-sequence cannot be broken. The pattern, schematically illustrated below, is rather robust in that no lexical material can intervene, and judgments involving intervening material are very clear cut:

(16) a. √ wh_1 ... wh_2 (wh_3)
b. √ Topic ... wh_1 wh_2 (wh_3)
* Topic wh_1 ... wh_2 (wh_3)

[4] A couple of my informants found intervening manner adverbs somewhat marginal at first. (Two fellow linguists reject adverbs altogether but at least one of them seems to do it inconsistently: for instance, allowing *prâv* 'first' but not others. They both allow particles and parentheticals, though.)

Topics are common in both yes-no and wh-questions (17a,b). These are clause internal, not left-dislocated structures. The relevant data are given in (18)-(21):[5]

(17) a. Ivan na tebe li dade knigata?
Ivan (top) to you Q give.3P.SG book-the
'Ivan, did he give you the book?'

b. Decata koj kâde šte vodi?
kids-the (top) who where will take
'The kids, who will take them where?'

(18) a. Na mene koj kakvo mi e dal?
to me (top) who what me.CL AUX.3P.SG given

b. Na mene, čunkim, koj kakvo mi e dal?
dub.part.

c. *Na mene koj, čunkim, kakvo mi e dal?
'To me, for god's sake, who's given me what?'

(19) a. Šefât kakvo koga iska?
boss-the (top) what when wants

b. Šefât, kazvaš, kakvo koga iska?
say.2P.SG

c. *Šefât kakvo, kazvaš, koga iska?
'The boss, you're saying, what does he want when?'

(20) a. Stipendija koj za kâde e polučil?
scholarship (top) who for where AUX.3P.SG gotten

b. Stipendija, s položitelnost, koj za kâde e polučil?
with certainty

c. *Stipendija koj, s položitelnost, za kâde e polučil?
'A scholarship, for sure, who's gotten it for where?'

(21) a. Vazata koj kak e sčupil?
vase-the (top) who how AUX.PRES.3P.SG broken
'The vase, who has broken it how?'

[5] Examples with three wh-phrases are given in (33) and (43).

b. Vazata, verojatno, koj kak e sčupil?
vase-the (top) probably who how AUX.3P.SG broken
c. *Vazata koj, verojatno, kak e sčupil?
probably
'The vase, who's probably broken it how?'

Based on the data in (10)-(15) I conclude, contra Rudin (1988), that fronted wh-phrases in BG do not form a constituent. Therefore, postulating two types of MWF languages is at best dubious. Still BG appears to be singled out among the Slavic languages. The question now is how to explain the contrast between (10)-(15) and (18)-(21), when a topic is present.

Rudin's (1986:32) structure for topicalization offers no answer. Clearly, if the topicalized phrase is above CP, as Rudin suggests, the penetrability of the fronted wh-phrases should not be affected:

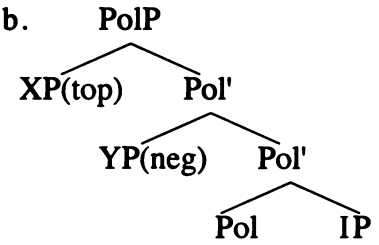
(22) $[_{TopP}$ XP $[_{SpecCP}$ WH WH WH] ... $[_{IP}$ ...]

I propose that in BG topics are below CP. Furthermore, I show that the observed pattern provides evidence for Bošković's (1997c) Economy/Focus movement account of MWF, under which the first wh-phrase in BG undergoes wh-movement, and the rest undergo focus movement. Bošković (1996; 1999), however, suggests that in BG the interrogative complementizer C licenses both movements. As a first approximation, the data I have provided necessitate a minor modification. Following Kim (1998) I assume that C licenses wh-movement while Pol licenses focus movement. Pol is the head of PolP in the sense of Koizumi (1994), and is equivalent to Foc of Kim (1998).

I follow Koizumi (1994:259ff) in that C takes PolP as its complement. He claims that Pol has two strong NP-features, namely +Top and +Neg, as evidenced by the co-occurrence of a topicalized phrase and a preposed negative/affective constituent in English:

(23) PolP (Polarity phrase)
a. He said that [beans] [never in his life] had he liked.
topic negative preposing SAI

b.

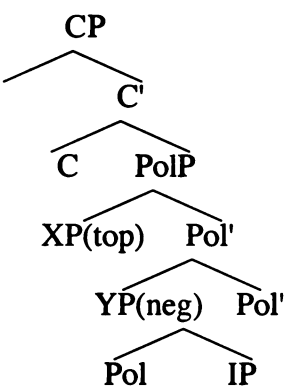


The relevance of this structure to MWF in BG becomes obvious on the suggestion that in examples (18)-(21) topics are in XP and fronted wh-phrases are in YP. I take Koizumi's +Neg to be a more general +Foc feature.

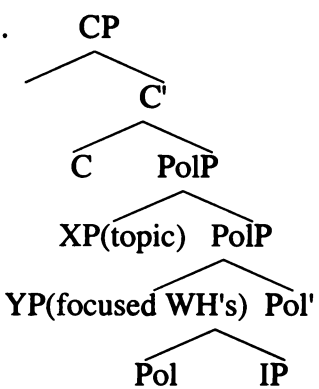
As Bošković (1997c) maintains, MWF in BG involves two distinct processes: wh-movement and focus-movement. Assuming Attract Closest (Chomsky, 1995), it should suffice for only one wh-phrase to go to SpecCP in order to check the apparently strong +wh feature of C. The rest of the wh-phrases move for an independent reason, claimed to be focus checking, as proposed by Stjepanović (1995/1998) for SC. She argues that SC wh-phrases undergo focus movement because they are inherently focused following Rochemont's (1986) and Horvath's (1986) work on other languages.

I suggest that Pol is a discourse-oriented head which licenses topicalization and focus movement, both of which are discourse related (for some relevant discussion, see Boeckx and Stjepanović, 1999). Koizumi (1994) argues that PolP has multiple specifiers:

(24) a.



b.



Assuming Chomsky's (1993) definition of Checking Domain,[6] the layered specifiers structure with XP in the higher SpecPolP (24a) is equivalent to XP being adjoined to PolP (24b). The simplest possible analysis will have focused wh-phrases in SpecPolP (=YP) and the topicalized phrase adjoined to PolP.[7] I first discuss constructions without a topic, and then return to constructions involving a topic.

Following Kim (1997) I claim that wh-phrases form a cluster by adjoining to each other. I assume that both Pol (=Foc, for Kim) and the wh-phrases have a strong focus feature. In a construction like (25a) *kakvo* 'what' first adjoins to *koj* 'who', which happens to be in SpecIP, to check its own strong focus feature, and then the complex *koj-kakvo*, whose focus feature is interpretable, moves to SpecPolP to check the strong focus feature of Pol:

(25) a. Koj kakvo e kupil?
who what AUX.PRES.3P.SG bought

b. $[_{CP}\ [_{SpecPolP}\ [\ koj\ ]_k\ [_{IP}\ t_k \ldots t_j\ ]]]$

koj kakvo$_j$

Now C has a strong +wh feature, and *koj* 'who' excorporates[8] to check it. Kim (1998) suggests that in the wh-cluster *koj* 'who' is higher than *kakvo* 'what' (i.e., one segment of *koj* 'who' is higher than *kakvo* 'what'), and therefore Superiority forces *koj* 'who' to move to SpecCP.

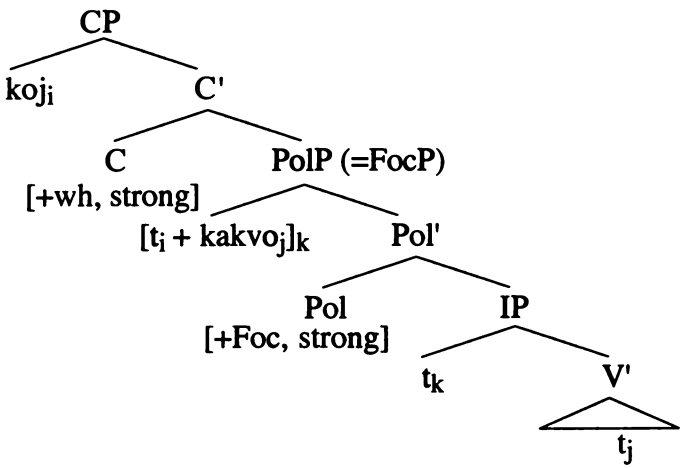
---

[6] According to this definition SpecXP, the SpecXP adjoined position, and the XP adjoined position are in the checking domain of X.

[7] Alternatively, I can assume multiple specifiers but will have to stipulate that +Foc is checked first.

[8] Excorporation is forced by Economy of Derivation in accordance with Watanabe's (1993) theory (see Bošković, 1997d; also, Kim, 1998).

(26)



As first discussed by Bošković (1997c), BG exhibits Superiority effects selectively. The structurally highest wh-phrase surfaces first but after that the order of the remaining wh-phrases is free:

(27) a. Kogo kakvo e pital?
whom what AUX.PRES.3P.SG asked
b. *Kakvo kogo e pital?
'Whom did you ask what?'

(28) a. Koj kogo kakvo e pital?
who whom what AUX.PRES.3P.SG asked
b. Koj kakvo kogo e pital?
'Who asked whom what?'

Let me show how these facts can be captured. I continue to assume wh-clustering, modifying Kim's (1997) analysis by proposing that the cluster is accomplished by rightward adjunction. For Kim, the direction of adjunction is free. Bošković (1997c; 1998; in press b) shows that when more than one element move to the same position, the order of movement is free. Thus regardless of whether x and y move to z in a x, y or y, x order, the same number of nodes are crossed:

(29) [z [x y]]

This applies to the wh-cluster analysis in the following way. In (28) *kogo* 'whom' and *kakvo* 'what' move to *koj* 'who' to check their own focus feature.[9] They right-adjoin to *koj* 'who'. The order of adjunction to *koj* 'who' does not matter for Economy of Derivation since *kogo* 'whom' and *kakvo* 'what' cross the same number of nodes regardless of the order of movement:

(30) a. [koj kogo] kakvo]
[IP koj [t_i t_k]
koj kogo$_i$
koj kakvo$_k$

b. [koj kakvo] kogo]
[IP koj [t_i t_k]
koj kakvo$_k$
koj kogo$_i$

The complex wh-cluster then moves to SpecPolP to check the strong focus feature of Pol. I assume that excorporation cannot take place here since all wh-phrases in BG are required to be located in the focus position (see Bošković, 1997c; 1998; in press b), which in my analysis is SpecPolP. The highest wh-phrase, namely *koj* 'who', then excorporates to move to SpecCP to check the strong +wh feature of C in the most economical way:[10]

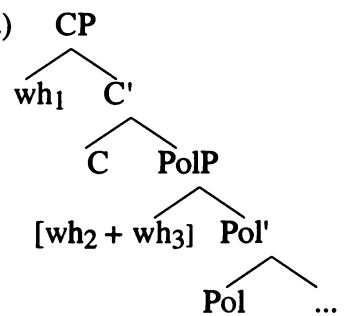
---

[9] Here I depart from Kim (1998) who would allow the second and the third wh-phrase to adjoin to each other. I leave open how this possibility should be ruled out.

[10] Economy of Derivation forces movement to carry as little material as possible. Moving *koj* 'who' alone is therefore more economical than moving the whole wh-cluster. See Kim for an alternative explanation of why excorporation takes place with wh-movement but not with focus movement.

(31) [SpecCP koj [C [SpecPolP t_{koj} [Pol ...]]]]
+wh, strong +Foc, strong
koj kogo$_i$
koj kakvo$_k$

The proposed analysis makes a testable prediction. Given the above discussion, in a sequence of three wh-phrases an adjunct is expected to be able to intervene between the first and the second wh-phrase but not between the second and the third one since the latter two form a constituent:

(32)


This prediction is borne out, as I show below:

(33) a. Koj, navjarno, kâde koga šte porâča tortata?
who perhaps where when will order cake-the
b. *Koj kâde, navjarno, koga šte porâča tortata?
who where perhaps when will order cake-the
c. *Koj koga, navjarno, kâde šte porâča tortata?
who when perhaps where will order cake-the
'Who will perhaps have the cake made where and when?'

Returning to the observed pattern that the split is not allowed in the presence of a topic, I will argue that this, too, is consistent with the analysis presented so far. I propose that the effect is due to the possibility of pronouncing lower copies of movement. Franks (1998) and Bošković (forthcoming) claim that what normally gets

pronounced is heads of non-trivial chains, but a lower member of a chain can be pronounced if this is necessary to avoid a PF violation.

I suggest that when a topic is present, intonation forces the pronunciation of a lower copy of the wh-phrase in SpecCP. As a result the wh-phrase below the topic is pronounced. Sometimes a topic is described as a preposed phrase characteristically set off by "comma intonation". As I will show below the BG facts make this intuition more precise.

The structures for (34a,b) are as given in (35) and (36), respectively:

(34) a. Koj kâde e porâčal tortata?
who where AUX.PRES.3P.SG ordered cake-the
'Who had the cake made where?'

b. Tortata koj kâde e porâčal?
cake-the (top) who where AUX.3P.SG ordered
'The cake, who had it made where?'

(35) [$_{SpecCP}$ koj [C [$_{SpecPolP}$ [~~koj~~] kâde]] [$_{Pol}$ e porâčal [$_{IP}$ [~~koj~~] ~~kâde~~]] ~~e porâčal~~ [$_{VP}$ ~~koj~~ ~~porâčal~~ ~~kâde~~ tortata]]]]]]

(36) [$_{SpecCP}$ ~~koj~~ [C [$_{PolP}$ tortata [$_{SpecPolP}$ [koj] kâde]] [$_{Pol}$ e porâčal [$_{IP}$ [~~koj~~] ~~kâde~~]] ~~e porâčal~~ [$_{VP}$ ~~koj~~ ~~porâčal~~ ~~kâde~~ ~~tortata~~]]]]]]]

The copy of *koj* 'who' in SpecCP (i.e. the head of the chain) is pronounced in (35) while in (36) the copy of *koj* 'who' in SpecPolP is pronounced.

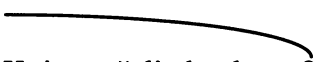
The responsible PF violation is intonation clash. At this point I need to make a small digression on BG intonation facts. As Penchev (1978) notes there are two major intonational contours in BG: neutral and marked. The former involves a medium to high fall, the latter involves a medium or high fall followed by a rise-fall. The neutral contour is equally available to statements and questions:

(37) Statements and questions:

a. statement: medium fall, neutral

Ivan spečeli konkursa.
Ivan won competition-the
'Ivan has won the competition.'

b. question: high fall, neutral



Koj spečeli konkursa?
who won competition-the
'Who has won the competition?'

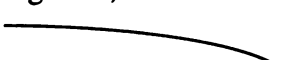
Penchev observes that a wh-word (or word followed by *li*) always attracts the high tone. Thus when a wh-word is sentence initial, the initial high tone falls on it and the contour is neutral (38a,b). When it is inside the sentence, the rise of a rise-fall is realized on it and the contour is marked (38c):

(38) Questions:

a. high fall, neutral

Ivan li spečeli konkursa?
Ivan Q won competition-the
'Did Ivan win the competition?'

b. high fall, neutral



Koj spečeli konkursa? [=(37b)]
who won competition-the
'Who won the competition?'

c. high fall followed by a rise-fall, marked

Znaeš li koj spečeli konkursa?
know.2P.SG Q who won competition
'Do you know who won the competition?'

Topics require the marked contour. The topic is delimited by an initial fall, and the rise-fall marks off the comment. Specifically, in (39) the rise is realized on *Ivan* :

(39) Topics: medium fall followed by a rise-fall, marked

Konkursât Ivan spečeli.
competition-the Ivan won
'(As for) The competition, Ivan has won it.'

A clause internal topic within a question also requires the marked contour. Again, an initial fall marks off the topic. The question word following the topic attracts the high tone, i.e., the rise of the rise-fall is realized on it:

(40) Question with a topic: medium fall followed by a rise-fall, marked

Konkursât koj spečeli?
competition-the who won
'The competition, who has won it?'

No other intonational contour is possible in this case. In particular, a wh-phrase at the beginning of a sentence requires a high gradually falling tone, i.e., the neutral contour of high fall. In this case, as well as in statements, the fall is actually realized at the end of the

sentence. The topic itself requires a medium fall. Moreover, the fall should be realized on it rather than at the end of the sentence. If the head of the chain (i.e., the copy of the wh-phrase in SpecCP) were to be pronounced on a neutral contour, the fall would fail to mark off the topic since it is not at the end of the sentence. On the other hand, on the marked contour the topic will be delimited by an initial fall, the rise marking off the lower copy of the wh-phrase in SpecPolP. Recall that question words normally require a high tone, even when inside the sentence. Thus in (40), as well as in (34b), the marked contour accommodates both the question word and the topic but is consistent with pronouncing the second copy of *koj* 'who'.

If this reasoning is correct, the penetrability pattern, observed earlier and repeated below, should follow immediately:

(41) a. √ wh_1 ... wh_2 (wh_3) [=(16)]
b. √ Topic ... wh_1 wh_2 (wh_3)
* Topic wh_1 ... wh_2 (wh_3)

(42) a. Koj, navjarno, kâde šte porâča tortata?
who perhaps where will order cake-the
'Who will perhaps have the cake made where?'
b. Tortata, navjarno, koj kâde šte porâča?
cake-the (top) perhaps who where will order
'The cake, who will perhaps have it made where?'
c. *Tortata koj, navjarno, kâde šte porâča?
cake-the (top) who perhaps where will order

The split (41a) is possible since the first wh-phrase is in SpecCP and the rest are in SpecPolP (see (10)-(15) and (42a)). In the presence of a topic (41b) fronted wh-phrases cannot be broken (see (18)-(21) and (42b,c)). I have argued that this is so because what gets pronounced is the wh-cluster in SpecPolP which forms a constituent.

The prediction that an adjunct can never intervene between a second and a third wh-phrase (32)-(33) should not, and indeed is not, affected by the presence of a topic:

(43) a. *Tortata koj kâde, navjarno, koga šte porâča?
cake-the (top) who where perhaps when will order
b. *Tortata koj koga, navjarno, kâde šte porâča?
cake-the (top) who when perhaps where will order
'The cake, who will perhaps have it made where when?'

Summarizing, I have presented new data on MWF in BG concerning penetrability of fronted wh-phrases. I have shown that first, a wh-cluster can be broken in the absence of a topic but not in the presence of one, and second, the split can only occur between the first and the rest of the wh-phrases, never further down in the wh-cluster. I have proposed an analysis that accounts for these facts. The analysis provides evidence for Bošković's (1997c; 1998) proposal that in BG the first wh-phrase undergoes wh-movement and the rest undergo focus movement. I have modified his account by locating the phrases undergoing different movements in different projections.

More generally, the new data on MWF in BG and the proposed analysis provide further empirical support for the possibility of pronouncing lower copies of non-trivial chains motivated by PF conditions as argued by Franks (1998) and Bošković (forthcoming). This, in turn, provides evidence for the copy theory of movement (Chomsky, 1995). Finally, the above discussion supports the conclusion of Bošković (1999) that MWF is an epiphenomenon and should be eliminated from the crosslinguistic typology.

References

Boeckx, C. and S. Stjepanović. 1999. The Wh/Clitic Connection in Slavic: Theoretical Implications. Paper presented at FASL 8. In King, T.H. and I. Sekerina (eds.) *Annual Workshop on Formal Approaches to Slavic Linguistics, The Philadelphia Meeting,* [this volume]. Ann Arbor, MI: Michigan Slavic Publications.

Bošković, Ž. 1996. "Fronting Wh-Phrases in Serbo-Croatian." In Lindseth, M. and S. Franks (eds.) *Annual Workshop on Formal Approaches to Slavic Linguistics, The Indiana Meeting at*

Crawfordsville, 86-107. Ann Arbor, MI: Michigan Slavic Publications.

Bošković, Ž. 1997a. "On Certain Violations of the Superiority Condition, AgrO, and Economy of Derivation." *Journal of Linguistics*, 33: 227-254.

Bošković, Ž. 1997b. "Superiority Effects with Multiple Wh-Fronting in Serbo-Croatian." *Lingua* 102, 1-20.

Bošković, Ž. 1997c. "Multiple Wh-Fronting and Economy of Derivation." In Curtis, E., J. Lyle, and G. Webster (eds.) *The Proceedings of the West Coast Conference on Formal Linguistics* 16, 49-65. Stanford, CA: CSLI Publications.

Bošković, Ž. 1997d. *The Syntax of Nonfinite Complementation: An Economy approach* . Cambridge, MA: MIT Press.

Bošković, Ž. 1998.*Wh-Phrases and Wh-Movement in Slavic*. Position paper, *Workshop on Comparative Slavic Morphosyntax: The State of the Art*, 5-7 June 1998, Indiana University. http://www.indiana.edu/~slavconf/linguistics/download.html.

Bošković, Ž. 1999. *What is Special About Multiple Wh-Fronting?* Ms. University of Connecticut.

Bošković, Ž. (in press a). "Second Position Clitics: Syntax and/or Phonology." In Beukema, F. and M. den Dikken (eds.) *Clitics in Europe*. Dordrecht: Foris.

Bošković, Ž. (in press b). "On Multiple Feature Checking: Multiple Wh-Fronting and Multiple Head-Movement." In Hornstein, N. and S. Epstein (eds.) *Working Minimalism*. Cambridge, MA: MIT Press.

Bošković, Ž. (forthcoming). *On the Syntax-Phonology Interface: A View from the Syntax*. Ms. University of Connecticut.

Chomsky, N. 1993. "A Minimalist Program for Linguistic Inquiry." In Hale, K. and J. Kayser (eds.) *The View from Building 20: Essays in Honor of Sylvian Bromberger*. Cambridge, MA: MIT Press.

Chomsky, N. 1995. *The Minimalist Program* . Cambridge, MA: MIT Press.

Franks, S. 1998. *Clitics in Slavic.* Position paper, *Workshop on Comparative Slavic Morphosyntax: The State of the Art*, 5-7 June 1998, Indiana University. http://www.indiana.edu/~slavconf/linguistics/download.html.

Horvath, J. 1986. *Focus in the Theory of Grammar and the Syntax of Hungarian* . Dordrecht: Foris Publications.

Kim, J.-S. 1997. *Syntactic Focus Movement and Ellipsis: A Minimalist Approach* . Ph.D. Dissertation: University of Connecticut.

Kim, J.-S. 1998. Superiority Effects in Multiple Wh- Fronting. Paper presented at the *Workshop on Comparative Slavic Morphosyntax: The State of the Art*, 5-7 June 1998, Indiana University.

Koizumi, M. 1994. "Layered Specifiers. " In Gonzàlez, M. (ed.) *The Proceedings of North East Linguistic Society* 24 (vol. I), 255-269. Amherst, MA: GLSA.

Mišeska-Tomić, O. 1996. "The Balkan Slavic Clausal Clitics." *Natural Language and Linguistic Theory* 14, 811-872.

Penchev, J. 1978. "Osnovni Intonacionni Konturi v Bûlgarskoto Izrechenie." *Bûlgarski Ezik* 28, 293-302.

Richards, N. 1997. "Shortest Moves to (Anti-)Superiority." In Curtis, E., J. Lyle, and G. Webster (eds.) *The Proceedings of the West Coast Conference on Formal Linguistics* 16, 335-349. Stanford, CA: CSLI Publications.

Rochemont, M. 1986. *Focus in Generative Grammar* . Philadelphia: John Benjamins.

Rudin, C. 1986. *Aspects of Bulgarian Syntax: Complementizers and Wh-constructions* . Columbus, OH: Slavica.

Rudin, C. 1988. "On Multiple Questions and Multiple Wh-Fronting." *Natural Language and Linguistic Theory* 6, 445-501.

Rudin, C. 1994. "On focus position and focus marking in Bulgarian questions." In Davidson, A. et al. (eds.) *Papers from the 4th Annual Meeting of the Formal Linguistics Society of Mid-America*, 252-265. Iowa City, IO: University of Iowa.

Stjepanović, S. 1995/1998. "Short-Distance Movement of Wh-Phrases in Serbo-Croatian Matrix Clauses." Paper presented at *Workshop on Comparative Slavic Morphosyntax: The State of the Art*, 5-7 June 1998, Indiana University.

Toman, J. 1981. "Aspects of Multiple Wh-Movement in Polish and Czech." In May, R. and J. Koster (eds.) *Levels of Syntactic Representation* , 293-302. Dordrecht: Foris.

Watanabe, A. 1993. *Agr-Based Case Theory and Its Interaction with the A-bar System* . Ph.D. Dissertation: MIT.

Wachowitcz, K. 1974. "Against the Universality of a Single Wh-Question Movement." *Foundations of Language* 11, 155-166.

Department of Linguistics, U-1145
University of Connecticut
341 Mansfield Road
Storrs-Mansfield, CT 06269
mdl97002@uconnvm.uconn.edu

Syntactic vs. Semantic Agreement in the Oslo Corpus of Bosnian Texts

Nedžad Leko
University of Oslo

1. Introduction

Agreement in Slavic languages poses a serious problem for linguists seeking to establish a set of syntactic rules that would generate appropriate agreement forms in appropriate positions. It seems that the alternative forms, one strictly syntactic, another more semantic, allowed in certain agreement positions imply that this phenomenon is not governed strictly by syntactic rules.

Two hierarchies, the Agreement and Predicate Hierarchies, proposed by Corbett (1979) seem to be the most promising in dealing with these agreement alternations. The validity of these hierarchies is extensively tested on a wide range of Slavic languages (Corbett 1983). I will concentrate on agreement in Bosnian and show that the hierarchies formulated by Corbett are indeed effective. The Oslo Corpus of Bosnian Texts, now available on the Web at: *http://www.tekstlab.uio.no/Bosnian/Corpus.html*, is used to support Corbett's proposal. It contains about 1.5 million words from Bosnian sources from the nineties.

I will follow the terminology adopted by Corbett (1983:5) and distinguish two types of agreeing constituents: *a controller* and *a target*. The former determines agreement and the latter agrees with the former. Adjectives, verbs and pronouns are agreement targets in Bosnian, and nouns are agreement controllers. The Agreement Hierarchy is formulated with respect to the agreement possibilities of four different types of targets, viz. attributive modifiers, predicates, relative pronouns, and personal pronouns, whereas the

Predicate Hierarchy makes hierarchical divisions of syntactic positions within the predicate.

Certain controller types permit alternative agreement forms (syntactic vs. semantic) in some target positions. Alternative agreement forms occur most frequently in a personal pronoun position. On the other hand, the likelihood of semantic agreement in attributive position is very low. Of the remaining two positions, semantic agreement is more probable with the relative pronoun than predicative position.

Therefore, it is possible to establish a hierarchy of four agreement positions according to the likelihood of semantic agreement with controllers that permit alternative agreement forms:

(1) attributive > predicate > relative pronoun >personal pronoun
----------->
semantic agreement likelihood increase

Hierarchical divisions of agreement positions within the predicate are established using an honorific pronoun *vi* 'you-PL' as a controller. It is plural in form but has a singular referent. On the basis of the likelihood of semantic agreement depending on the type of predicate, the following hierarchy is established:

(2) finite verb > active participle > predicate adjective > noun
----------->
semantic agreement likelihood increase

All the examples to be reported in subsequent sections will be discussed with respect to the hierarchies given in (1) and (2).

Various factors relating to the controller may influence agreement. Animate controllers and controllers preceding their targets favour semantically justified agreement forms. Other factors that may influence agreement forms of targets are syntactic distance from the controller and apposition.

This article is organized as follows. We begin in Section 2 with the discussion of different positions of agreement on the

Agreement Hierarchy. Section 3 deals with exceptional controllers in Bosnian corpus that permit alternative agreement forms in various agreement positions. Subsection 3.1 illustrates the application of agreement hierarchies proposed by Corbett to Bosnian data. Subsections 3.2 and 3.3 present corpus data that support the proposed hierarchies. Finally, subsection 3.4 is a summary of the results obtained by analyzing corpus data.

2. The Agreement Hierarchy

The four positions of agreement are illustrated in (3):

(3) **Jadni** *čovjek* **koji** plače je **izgubio** roditelje. **On** tuguje.
poor.MASC man who.MASC cries lost.MASC parents he.MASC m
'The poor man who is crying has lost his parents. He mourns.'

The first position is prenominal and we find an attributive modifier there (*jadni*). The second is a relative pronoun position (*koji*), the third is a predicative position (*izgubio*), and in final position we find a personal pronoun (*on*). The head noun in (3) (*čovjek*) is clearly masculine and it does not permit alternative agreement forms in any of these positions. However, certain controller types permit alternative agreement forms in some of these positions.

For example, there are morphologically neuter nouns which denote a female person, so that alternative neuter or feminine agreement is possible with such controllers. If a controller contains the numerals *dva* 'two', *tri* 'three', *četiri* 'four', *oba* 'both', the target may exhibit either dual (syntactic) or plural (semantic) agreement. Alternative agreement forms are also found when the controller position is occupied by conjoined noun phrases with both conjuncts being plural, but of different genders. Finally, alternative gender agreement is possible with a limited class of nouns which are masculine in the singular, but in the plural may also behave as feminine nouns.

The Agreement Hierarchy makes very strong predictions. It claims, for example, that if the predicate permits semantic agreement, then this type of agreement will be possible in the other

two positions to the right. Similarly, the possibility of semantic agreement in the two adjacent positions is the same or greater in the position to the right on the scale than in that to the left. It should be pointed out that the hierarchy applies at the corpus level. It may also operate as a sentence-level constraint, but not always. In particular, there are examples of a single sentence with semantic agreement of the verb and syntactic agreement of the relative pronoun. If the Agreement Hierarchy applies exclusively at the level of a single sentence, this should not be allowed.

3. Exceptional Controllers in Bosnian

Most controllers are like the one in (3): their morphological properties do not clash with their meaning. This means that the target must agree with the morphological gender, number and person of the controller. However, there are controllers which are exceptional in the sense that their morphological properties and their meaning do not correspond, e.g. a neuter noun may denote a female person.

We find this type of gender conflict with *djevojče* 'teenage girl', a neuter noun which clearly denotes the female sex, and thus allows for both neuter and feminine agreement in some target positions. Gender agreement is also involved with controllers like *gazda* 'landlord', *kalfa* 'journeyman' *aga* 'aga', *halvadžija* 'halva maker', etc., all of which denote a profession or a social status. They are masculine nouns in the singular, but may also behave as feminine in the plural, so that both masculine and feminine agreement occur with plural nouns of this type as controllers. Agreement variation with such controllers is found in three agreement positions: attributive, predicative and relative pronoun.

Number agreement is involved in the case of controllers containing the numerals *dva*, *tri*, *četiri*, *oba*. In such cases, we find alternative number agreement, dual or plural, in two agreement positions: the predicate and relative pronoun. A similar type of agreement alternation in the same two positions is found when the controller position is occupied by conjoined noun phrases of different genders. This time the variation in gender agreement is

manifested. The target in predicative and relative positions may agree with the nearest conjunct or with both conjuncts.

Space limitations prevent us from discussing all types of exceptional controllers. Instead, we will discuss only alternative agreement with *djevojče*, and controllers containing the numerals *dva*, *tri*, *četiri*, *oba*.

3.1. Application of Agreement Hierarchies in Bosnian

The four positions of agreement were illustrated with the example in (3). There would be no reason to postulate any hierarchy of target types if all controllers had morphological properties which were not in conflict with their meaning. But the example in (4) shows that certain controller types permit alternative agreement forms.

(4) **To** *djevojče* **koje** plače je **došlo** juče,
that.NEUT girl.NEUT who.NEUT cries AUX came.NEUT yesterday
ali sam **ga**/ **je** već zavolio.
but AUX it.NEUT her.FEM already like
'That girl who is crying arrived yesterday but I already like her'

In one of the four positions two agreement forms are possible, one being syntactic (*ga*), the other semantic (*je*). Morphological properties of the controller determine the usual form of syntactic agreement, and no considerations of meaning are taken into account. That is a type of agreement we find in three positions in (4), with neuter forms *to* 'that', *koje* 'who' and *došlo* 'came'.

Since the noun (*djevojče*) clearly denotes female sex, alternative agreement, which takes into account only meaning, occurs in one position: the personal pronoun position, that is, the rightmost position on the hierarchy. But other types of controllers allow for agreement alternation in more than one position.

We find such examples with controllers containing the numerals *dva* 'two', *tri* 'three', *četiri* 'four', and *oba* 'both' with a masculine noun. Such controllers allow for agreement alternations in two positions, the predicate and relative pronoun, as in (5):

(5) **Ova** *četiri* **zasebna** *položaja*
these four discrete.DUAL[1] position.DUAL.MASC
koja/ **koji** postoje u rečenici su
which.DUAL which.MASC.PL exist in sentence are
popunjena/ **popunjeni** imeničkim modifikatorima.
filled.DUAL filled.MASC.PL nominal modifiers.
Oni/ ***ona** se nalaze u hijerarhijskom odnosu.
they.MASC.PL DUAL are in hierarchical relation
'These four discrete positions which are found in a sentence are filled with noun modifiers. They are in a hierarchical relation.'

The first position on the Agreement Hierarchy, the attributive position, always shows syntactic dual agreement, so that the attributive modifier cannot take an alternative masculine plural form *zasebnih*; only the dual form *zasebna* is allowed.

Alternative forms of agreement are found in the predicative and relative pronoun positions, as shown in (5): the dual (*koja*; *popunjena*) representing syntactic agreement, and the masculine plural forms (*koji*; *popunjeni*) representing semantic agreement. Finally, in the last position only semantic agreement is possible, so that the personal pronoun obligatorily occurs in the masculine plural form: *oni* in (5). Next we will discuss examples of agreement with *djevojče* found in the Oslo Corpus of Bosnian Texts (OCBT).

3.2. *Djevojče*

The OCBT contains 11 occurrences of *djevoč e*, and they trigger agreement in 17 target positions: 7 attributes, 6 predicates, 3

[1] Bosnian used to have three numbers in nominal declensions (singular, plural and dual). In later developments dual was lost as a separate category, but the special form of the noun after *dva* 'two', *tri* 'three', *četiri* 'four', and *oba* 'both' is a remnant of the old dual. It corresponds in form to the genitive singular.

personal pronouns, and 1 relative pronoun. The distribution of neuter and feminine agreement forms is shown in Table 1.

Table 1. Agreement Patterns for the Word *Djevojče*

	Attribute	Predicate	Relative Pronoun	Personal Pronoun
Neuter	7	3	1	0
examples	(6c-d)	(6b, 6d)	(6b)	
Feminine	0	3	0	3
examples		(6c-d)		(6a, 6c)
Total	7	6	1	3

The data from the OCBT show that the choice of agreement forms with the noun *djevoj e* as an agreement controller is possible not only in a personal pronoun position, as in (4) above, but also in predicate positions:

(6) a. ...da pripravi *djevojče*. Njezina je želja bila da
to prepare girl.NEUT her AUX wish was that
je uda za Nadi-bega.
her.FEM.ACC marry for Nadi-bey
'to prepare the girl. It was her wish to marry her to N.-bey.'
(Source: B/BA/M/96)

b. ...*djevojče* **koje** je **sjedilo** na starinskoj klupi...
girl.NEUT who.NEUT AUX sat.NEUT on ancient bench
'...the girl who was sitting on an ancient bench...'
(Source: B/BA/M/96)

c. *Ona*, **to** **malo**, gotovo **nedoraslo** *djevojče*
she.FEM that.NEUT little.NEUT almost immature.NEUT girl.NE
znala je...
knew.FEM AUX
'She, that little, almost immature girl, knew...'
(Source: B/MA/HN/96)

d. ...**jadno** *djevojče* [je] **doznalo** šta se desilo
poor.NEUT girl.NEUT AUX found out.NEUT what happened
sa njezinom prijateljicom i **došla** da joj olakša sramotu.
with her friend.FEM and came.FEM to her alleviate shame

'...the poor girl found out what happened to her friend and she came in order to alleviate her shame.'
(Source: B/MA/HN/96)

We find a split distribution of neuter and feminine forms in the predicate: three neuter and three feminine active participle forms.[2] All attributive modifiers are neuter, whereas all personal pronouns are feminine. There is only one relative pronoun and it is neuter.

In (6a), there is only one agreement position, a personal pronoun position, and we find a feminine pronoun (*je*) there, an instance of semantic agreement, and that is an alternative predicted by the Agreement Hierarchy. In (6b) two agreement positions are filled, the relative pronoun and predicate position, and we find neuter forms in both positions (*koje* and *sjedilo*), an instance of syntactic agreement.

The feminine forms of the predicate in (6c,d) apparently contradict the Agreement Hierarchy. Instead of the expected neuter participle form *znalo*, we find a feminine form *znala* in (6c). Notice, however, that there is comma intonation after the personal pronoun *ona*, which means that the following phrase containing the noun *djevojče* is an appositive phrase, so that the predicate actually agrees with the pronoun *ona*, rather than with the noun *djevojče*. In (6d) we find two participle forms in the predicate, one neuter (*doznalo*), the other feminine (*došla*). The second participle is separated from its conjunct by a rather long string of words, so that the relative distance from the controller is at work here.

These examples show that various factors may influence agreement. Corbett (1983) emphasized and discussed at length animacy and precedence as the most prominent of the controller factors. Animate controllers and controllers preceding their targets favour semantically justified agreement forms. The examples in (6c-d) show that syntactic distance and apposition may also influence agreement, favouring semantic agreement. These factors

[2] There are no examples of predicative adjectives in the OCBT with *djevojče* as a controller of agreement. Only active participle forms are found and therefore the Predicate Hierarchy is irrelevant here.

will come into play again in discussing examples involving the numerals *dva*, *tri*, *četiri*, and *oba* in Section 3.3.

The Agreement Hierarchy claims that of the two positions allowing for choice of agreement forms, the relative pronoun will show semantic agreement more readily (or at least with equal frequency) than the verb in the predicate. Let us check this claim by examining agreement with the controllers containing the numerals *dva*, *tri*, *četiri*, and *oba* found in the OCBT, since such controllers allow for agreement alternation both in the predicate and relative pronoun.

3.3. Numerals *Oba, Četiri, Tri, Dva*

These numerals are represented with 2536 occurrences. Their distribution is shown in Table 2.

Table 2. Distribution of Numerals

Oba	154
Četiri	415
Tri	857
Dva	1110
Total	2536

The following subsections will present agreement data concerning each of these numerals, starting with *dva*.

Dva. The distribution of the numeral *dva* is represented in Table 3.

Table 3. Distribution of *Dva*

Dva + Masculine Noun	941
Dva + Neuter Noun	143
Dva Alone	26
Total Occurrences	1110

Only examples with masculine nouns (941 examples) are relevant for our purposes. Neuter nouns with *dva* allow for only plural

forms in all agreement positions, because there is no distinct neuter dual form. The data in Table 4 present the distribution of the agreement forms.

Table 4. Agreement with *Dva* in Agreement Hierarchy Positions

	Attribute	Predicate	Relative Pronoun	Personal Pronoun
Dual	332	86	17	0
examples	(9c-d)	(11a)	(9a, 9c, 11b)	
Plural	1	91	17	14
examples	(7)	(11b, 13a-c)	(9b, 9d)	(8)
Singular	0	4	0	0
examples		(10, 12)		
Total	333	181	34 (43)	14

In 332 examples of attributive modifiers we find dual forms of modifiers ending in *–a*, an instance of syntactic agreement. There is one example with the plural form of the attributive modifier, as in (7), with the plural form *nekih*:

(7) ...**nekih** *dva kilometra* niže od krajnje zapadne granice...
some.PL two kilometers lower from furthest western border
'...some 2 kilometers further from the furthest western border...'
(Source: E/KD/DS/95)

The OCBT contains 28 occurrences of the plural genitive form *nekih* followed by a quantity expression. In most of these examples (22 examples) we find numerals higher than five, but there are six examples in which numerals *dva*, *tri*, and *četiri* occur, as in (7). *Nekih* may be regarded as a frozen form which doesn't show agreement at all in these expressions used as a measure of distance or time.

There are 14 examples involving agreement in the personal pronoun position. Only examples of semantic agreement are found, as in (8), with the plural form *oni*:

(8) Angažirao mu je *dva advokata* čak iz Beograda.
hired him AUX two lawyers.DUAL even from Belgrad
Oni su uz malu pomoć...
they.PL AUX with little help
'He hired two lawyers from Belgrade for him. They, with little...' (Source: B/KN/IP/95)

There are 43 examples involving relative pronouns. Nine of these examples are not relevant because their relative pronouns appear in oblique forms which do not have distinct dual forms.

(9) a. Samo *dva čovjeka* **koja** su boravila u
only two men.DUAL who.DUAL AUX stayed in
blizini mosta mogla su...
vicinity bridge could.DUAL AUX
'Only two men who stayed in the vicinity of the bridge could...'
(Source: D/MA/PK/94)

b. ...*dva čovjeka* u crnim odijelima sa konjskim glavama
two men.DUAL in black suits with horses' heads
koji su nam donijeli stolice.
who.PL AUX us brought chairs
'...the two men in black suits with horses' heads who brought us the chairs.'
(Source: D/HI/VA/94)

c. ...više kao susret *dva neuvježbana hrvč a*
more as meeting (of) two untrained wrestlers-DUAL
koja nastoje da ne povrijede...
who.DUAL try that not injure
'...more as a meeting of two untrained wrestlers who try not to injure...'
(Source: E/LA/UZ/96)

d. Nešto kao susret *dva uvježbana hrvača*
something as encounter two trained wrestlers-DUAL
koji misle na trećeg.
who.PL think about third one
'Something like an encounter between two trained wrestlers who think about the third one.'
(Source: E/LA/UZ/96)

The examples (9a,b) contain the same quantified phrase (*dva čovjeka*), but whereas in (9a) the relative pronoun referring to this phrase is in its dual form (*koja*), in (9b) we find the plural form (*koji*). The plural form in (9b) may be explained as being due to the syntactic distance from the head, since two prepositional phrases separate the noun head from the relative pronoun.

However, the contrast in (9c,d) cannot be explained in this way. In these examples we again find the same quantified phrase (*dva (ne)uvježbana hrvača*) immediately followed by a relative clause but introduced by different forms of the relative pronoun – the dual form *koja* in (9c) and the plural form *koji* in (9d). Notice incidentally that both examples come from the same source, so that the different pronouns used cannot be due to the idiolectal differences of the authors. Obviously, this author simply doesn't have any preferences for dual or plural forms, and the use of dual or plural forms cannot be attributed to any of the factors that we are discussing here.

Finally, we will discuss the application of the Predicate Hierarchy in the examples with *dva*. There are 181 examples involving agreement in predicative position, as shown in Table 5. Only examples with participles and predicative adjectives are

relevant for our purposes since they may appear either in dual or plural form.[3]

Table 5. Agreement with *Dva* in Predicate Hierarchy Positions

Predicate:	Present	Aorist	Future	Nominal	Participle	Adjective
Plural examples	59	10	1	1	12 (11b, 13b)	8 (13a-c)
Dual examples	0	0	0	0	58 (11a)	28
Singular examples	0	1 (10)	0	0	3 (12)	0
Total	59	11	1	1	73	36

Aorist verbs regularly appear in the plural form. But there is one example with an aorist verb in which we find a singular, instead of the expected plural form, as in (10):

(10) **Prođe** mjesec. **Prođe** dva.
passed.SG month passed.SG two (months)
'A month passed. Two months passed.'
(Source: B/MA/HN/96)

We find here the singular aorist form *prođe* instead of the expected plural form *prođoše*, as in: *Prođoše dva (mjeseca)*. This shows that *mjesec/ dva (mjeseca)* in these sentences is not a subject. The underlying subject in these sentences is an unexpressed expletive pronoun corresponding to English *it*. Therefore we find the singular form of the verb in agreement with the unexpressed singular pronoun. So (10) doesn't contradict our claim that aorist verbs always occur in their plural form with subjects containing the numeral *dva*. *Dva* (*mjeseca*) in (10) simply is not a subject, and therefore the verb doesn't agree with it.

In (11) the verb *stajati* 'to stand' occurs in its dual and plural participial forms:

[3] Predicates with present or future tense verbs, and nominal elements in copula constructions, are not relevant because they can only appear in the plural.

(11) a. Pred vratima samoga Enderona **stajala** su *dva oficira*..
in front of door itself Enderon stood.DUAL two officers
'Two officers were in front of Enderon...'
(Source: B/BA/M/96)

b. Sa strana su u počasnoj straži smrtno ozbiljni **stajali**
on sides AUX in honour guard deadly serious stood.PL
po *dva studenta* Sarajevskog univerziteta,...,
DISTR two students.DUAL (of) Sarajevo University
koja su pokušavala da...
who.DUAL AUX tried that
'Two students of Sarajevo University stood deadly serious as honour guards on each side, and they tried...'
(Source: B/AM/GO/97)

The verbs are in front of the subject in both examples, so that this factor must be eliminated as a possible cause of the different forms of the verb. We may suggest that the plural form of the verb in (11b) may be due to the fact that we find the distributive construction in this example with the distributive preposition *po*, roughly corresponding to English 'each'.

There are three examples with active participles which occur neither in their dual or plural forms, but unexpectedly in the singular neuter form, as in (12):

(12) *Dva vojna džipa i više automobila*, paralelno s
two military jeeps.DUAL and several cars.PL parallel with
tramvajem je **gorjelo**...
tram.SG AUX burned.SG.NEUT
'The two military jeeps and several cars, parallel with the tram, were burning...'
(Source: E/PM/SR/93)

The subject contains co-ordinated quantified phrases, and the neuter singular participial form is in agreement with the second conjunct (*više automobila* 'several cars'). The quantifier *više*

'several, more' is the head of the phrase and the participle actually agrees with this head and therefore we find a singular neuter form (*gorjelo*).

There are 36 examples with adjectival elements in the predicate. They manifest their plural form only in eight of these examples. The one containing a plural predicative adjective *ozbiljni* 'serious' is given in (11b). Some more examples of plural adjectival elements are given in (13):

(13) a. *Dva sarajevska društvena sloja,* takozvana elita
two Sarajevan social layers.DUAL so-called elite
i sirotinja, iako oštro međusobno **razgraničeni,**
and poor although sharply mutually separated.PL
'The two Sarajevan social layers, the so-called elite and the poor, although sharply mutually separated,...'
(Source: B/AM/GO/97)

b. *Dva brata* od dvije matere šutili su **odvojeni.**
two brothers of two mothers were silent separated.PL
'The two brothers from different mothers were silent, separated from each other.'
(Source: B/IN/U/96)

c. U blizini ovoga kraja, prirodno **povezani** s njim,
in vicinity (of) this region naturally linked.PL with it
nalaze se *dva grada* **koja**...
there are.PL two towns which.DUAL
'There are two towns in the vicinity of this region, naturally connected with it, which...'
(Source: E/LI/LI/94)

In (13a,b) the adjectival phrase follows the quantified phrase, whereas in (13c) the quantified phrase is preceded by the adjectival phrase. In (13a) the plural form of the adjectival phrase in the predicate *razgraničeni* may be explained as due to syntactic distance, since the subject and the predicate are separated by intervening material. In (13b) we find the plural form of the verb (*šutili su*), and thus the following adjectival element (*odvojeni*) is also plural. The choice of the plural adjectival form in (13c) may

be influenced by the reversed order of the subject and the predicate. Notice that the relative pronoun in (13c) has dual form.

Space limitations prevent us from discussing examples of agreement with controllers containing the numerals *oba, tri*, and *četiri*. However, we will present the tables showing the distribution of these numerals in the OCBT and the distribution of the agreement forms.

Table 6. Distribution of *Oba*

Oba + Masculine Noun	138
Oba + Neuter Noun	16
Total Occurrences	154

Table 7. Agreement with *Oba* in Agreement Hierarchy Positions

	Attribute	Predicate	Relative Pronoun	Personal Pronoun
Dual	20	13		0
Plural	0	5		2
Total	20	18 (37)	2	2

Table 8. Agreement with *Oba* in Predicate Hierarchy Positions

Predicate:	Present	Future	Nominal	Participle	Adjective
Plural	15	3	1	3	2
Dual	0	0	0	7	6
Total	15	3	1	10	8

Table 9. Distribution of *Četiri*

Četiri + Masculine noun	176
Četiri + Feminine noun	179
Četiri + Neuter noun	38
Četiri Alone	22
Total Occurrences	415

Table 10. Agreement with *Četiri* in Agreement Hierarchy Position

	Attribute	Predicate	Relative Pronoun	Personal Pronoun
Dual	31	21	3	0
Plural	3	6	4	2
Total	34	27 (36)	7	2

Table 11. Agreement with *Četiri* in Predicate Hierarchy Positions

Predicate:	Present	Future	Nominal	Participle	Adjective
Plural	8	1	0	4	2
Dual	0	0	0	17	4
Total	8	1	0	21	6

Table 12. Distribution of *Tri*

Tri + Masculine Noun	422
Tri + Feminine Noun	324
Tri + Neuter Noun	59
Tri Alone	52
Total Occurrences	857

Table 13. Agreement with *Tri* in Agreement Hierarchy Positions

	Attribute	Predicate	Relative Pronoun	Personal Pronoun
Dual	119	29	3	0
Plural	1	8	8	0
Total	120	37 (61)	11 (14)	0

Table 14. Agreement with *Tri* in Predicate Hierarchy Positions

Predicate:	Present	Aorist	Past Perfect	Participle	Adjective
Plural	18	6	0	5	3
Dual	0	0	1	19	9
Total	18	6	1	24	12

3.4. Summary

In conclusion, we will present tables including combined data with all numerals discussed in Section 3.3.

Table 15. Distribution of *Oba, Dva, Tri, Četiri*

Oba/ Dva/ Tri/ Četiri + Masculine Noun	1677
Tri/ četiri + Feminine Noun	503
Oba/ Dva/ Tri/ Četiri + Neuter Noun	256
Dva/ Tri/ Četiri Alone	100
Total Occurrences of *Oba/ Dva/ Tri/ Četiri*	2536

Table 16. Agreement Forms in Agreement Hierarchy Positions

	Attribute	Predicate	Relative Pronoun	Personal Pronoun
Dual	502 (99%)	149 (56,5%)	23 (44%)	0
Plural	5 (1%)	110 (42%)	29 (56%)	18 (100%)
Singular	0	4 (1,5%)	0	0
Total	507 (100%)	263 (100%)	52 (100%)	18 (100%)

Table 17. Agreement Forms in Predicate Hierarchy Positions

Predicate:	Present	Aorist	Future	Noun	Participle	Adjective
Dual	0	0	0	0	101 (79%)	47 (76%)
Plural	100	16	5	2	24 (19%)	15 (24%)
Singular	0	1	0	0	3 (2%)	0
Total	100	17	5	2	128 (100%)	62 (100%)

The figures in Table 16 show that The Oslo Corpus of Bosnian Texts supports the hierarchy proposed by Corbett. The likelihood of semantic agreement in attributive position is very low. The Corpus contains 502 examples of dual agreement forms in attributive position, and only 5 plural forms, which is less than 1%. On the opposite side of the scale, in the personal pronoun position, only plural forms are found (18), an example of semantic agreement. Of the remaining two positions, semantic agreement is more probable in the position of the relative pronoun than in

predicative position. We find more examples of semantic than syntactic agreement in the relative pronoun position: 29 plural forms (56%) vs. 23 dual forms (44%). On the other hand, there are more examples of syntactic than semantic agreement in the predicative position: 149 dual (56,5%) vs. 110 plural forms (42%).

Table 17 shows that the Predicate Hierarchy is also supported by the data from the OCBT. Since finite verbs and nominal elements in the predicate do not distinguish distinct dual forms these two positions are not relevant. Of the remaining two positions, semantic agreement is more probable in the adjectival position than in the active participle. In both positions we find more examples of syntactic than semantic agreement. However, percentages show that semantic agreement is more represented in the adjectival position (24%) than in the position of the active participle (19%), in accordance with the Predicate Hierarchy.

References

Corbett, G. 1979. "The Agreement Hierarchy". *Journal of Linguistics* 15, 203-224.

Corbett, G. 1983. *Hierarchies, Targets and Controllers. Agreement Patterns in Slavic*. London: Croom Helm.

Corbett, G. 1998. "Agreement in Slavic". Position Paper for the Workshop *Comparative Slavic Morphosyntax*. Indiana University. January 1998. Available on the Web at: *http://www.indiana.edu/~slavconf/linguistics/index.html*

Sources (from: *http://www.tekstlab.uio.no/Bosnian/Korpus2.html*)

B/AM/GO/97: Ajanovi , M. 1997. "Goeteborg".
B/BA/M/96: Bjelevec, A. H. 1996. "Minka". Preporod.
B/IN/U/96: Ibrišimovi , N. 1996. "Ugursuz". Preporod.
B/KN/IP/95: Kurspahi , N. 1995. "Iščezavanje plavih jahača".
B/MA/HN/96: Muradbegovi , A. 1996. "Haremske novele". Prep.
B/OS/B/96: Omeragi , S. 1996. "Bunda". Lettre international.
D/HI/VA/94: Horozovi , I. 1994. "Vauvan". Bosanska riječ.
D/MA/PK/94: Musi , A. 1994. "Poljana kod Šeste vodenice".

E/KD/DS/95: Karahasan, Dž. 1995. "Dnevnik selidbe". Durieux.
E/LA/UZ/96: Lazarevska, A. 1996. "U znaku ruže". Bosanska k.
E/LI/LI/94: Lovrenović, I. 1994. "Labirint i pamćenje".
E/PM/SR/93: Prstojević, M. 1993. "Sarajevo - ranjeni grad". Ideja.

Department of Linguistics
Boks 1102 Blindern
0317 Oslo
nedzad.leko@ilf.uio.no

Slavic as Testing Grounds for a Linguistic Knowledge Elicitation System

Marjorie McShane, Stephen Helmreich,
Sergei Nirenburg, Victor Raskin
New Mexico State University

1. Introduction

Among many other tasks, NLP (natural language processing) systems must be able to determine which words to list in the dictionary and how to determine citation forms in texts—a non-trivial matter for languages that have extensive inflectional and/or derivational morphology, spelling mutations, etc. This issue is central for Boas (Nirenburg and Raskin 1998), the linguistic knowledge elicitation component for the Expedition project of the Computing Research Laboratory at New Mexico State University. (See *http://crl.nmsu.edu/expedition* for an overview.) The goal of Expedition is to develop the capability for fast deployment of a machine translation system between any so-called "low-density" language (one lacking significant machine-tractable resources) and English. Boas must guide non-expert human informants through questions about the morphology, syntax, lexical stock, syntax, and ecology (letters, symbols, punctuation, etc.) of their language. Here we focus on those components of Boas associated with morphology.

Since Boas must accommodate any low-density language, and since linguistic materials are scarce or unavailable for many such languages, the system cannot be 'primed' to cover individual languages, leading to questions of coverage and efficient testing. With respect to coverage, a broad survey of languages reveals that even significantly diverse morphological phenomena fall into groups whose elements can be handled similarly by Boas. With respect to testing, since the major Slavic languages contain morphological phenomena representing each of the groups we have delineated, we hypothesize that if Boas provides sufficient

declarative knowledge for the processing engines in trials on Slavic languages (to be carried out by personnel at the lab), it should have similar success with the low-density languages for which it is being designed.

The rules of the Boas game are as follows: one language informant, who need not be a linguist, and one programmer, who need not be versed in NLP, will work for six months guided by the materials resident in the system. They may start from scratch or may incorporate existent on-line resources if the programmer can make them compatible with the relevant components of Boas. A strong informant-programmer team can incorporate extensive language-specific plug-ins, while a less experienced team can limit themselves to tasks explicitly set by Boas. At the end of six months, a moderate-quality, broad-coverage translation system should be in place.

This paper focuses on the scope of morphological phenomena presented by natural language, the distribution of these phenomena among the components of Boas, and the role that Slavic languages play in building and testing the system. The paper is organized as follows. Section 2 discusses the morphologically relevant components of Boas and the Slavic phenomena that will test each one; Section 3 discusses why most derivational morphology is not handled productively in Boas; and Section 4 concludes the paper.

2. Morphology in Boas

Morphological phenomena will be gathered in various components of Boas: inflectional paradigms, derivational affixation, and the closed- and open-class lexicons, as shown in Figure 1 (SL stands for 'source language'). The sections below describe each component, with emphasis placed on the method of elicitation employed, the types of cross-linguistic phenomena targeted, and the Slavic examples that will serve to test the system.

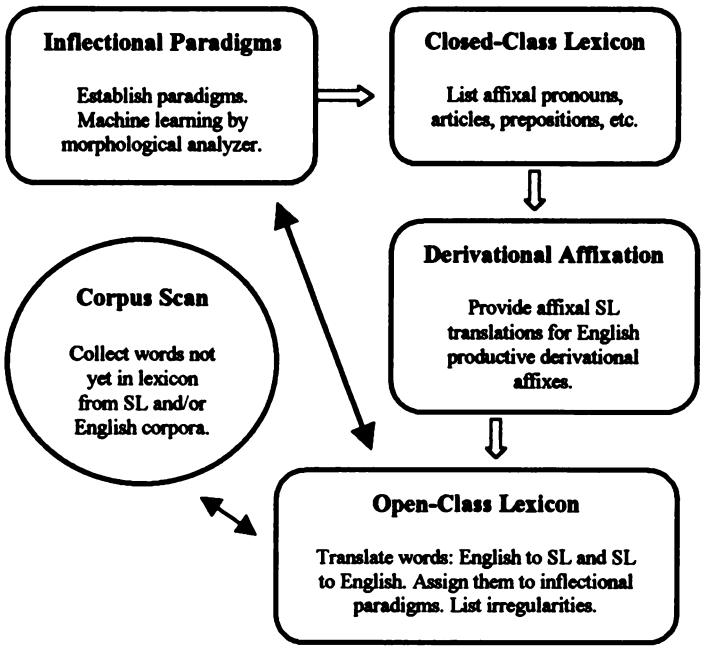


Figure 1. The Stages of Morphological Acquisition in Boas.

2.1. Inflectional Paradigms

The establishment of paradigms is intended to speed up open-class lexical acquisition: rather than type in all the inflectional forms for each of the 60,000 English word senses to be translated into the source language (SL), the informant can type in just the citation form and then assign the word to one of the paradigms established earlier.

Dividing inflecting words into paradigms is anything but a trivial task for linguists, not to mention the non-expert users of Boas, since paradigms represent the epitome of regularity but natural language is often far from regular. Therefore, in addition to supplying extensive pedagogical support, Boas walks the informant through the process of establishing paradigms, first determining for which combinations of parameter values the given

part of speech (PoS) inflects,[1] then providing a paradigm template that associates each licit (as indicated by the informant) combination of parameter values with a text box, as the miniature paradigm template in Figure 2 shows.

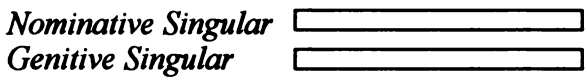


Figure 2. Templates for Inflectional Paradigms.

Then comes the hard part: the informant must select words that represent all the regular patterns of inflection in SL, interpreting 'regular' in a computationally valid sense (in contrast to the extended notion of paradigm found in many grammars). Obviously, no non-expert is expected to know intuitively what a computationally valid paradigm is, so Boas will help. First, the informant posits one citation form for what he thinks is each major inflectional pattern—a sort of first approximation of 'the truth'.[2] Let us assume he selects four patterns, instantiated by *word1*, *word2*, *word3*, *word4*. Each of these citation forms is generated in each of the textboxes of the paradigm template, and the informant edits the citation forms as necessary to reflect the given combination of parameter values (e.g., in Russian, the citation form *stol* would be edited to *stola* in the box labeled Genitive Singular).

After the informant establishes these paradigms, the morphological analyzer generates a preliminary set of rules for them (Oflazer and Nirenburg 1999). Then the informant provides

[1] Extensive lists of parameters and values are presented in the form of tables containing brief definitions and check boxes. Further pedagogical support can be accessed via links to a comprehensive glossary of linguistic terms, each page of which contains links to related topics such that the glossary can be employed as a free-form tutorial for novice language informants. At all stages, the informant can supplement the supplied list of parameters and values.

[2] Although the informant is free to choose what form to consider the 'citation form', it should match the citation form found in any on-line dictionaries that might be employed.

three additional citation forms for each paradigm to test these rules: if the analyzer generates all correct inflectional forms, the user can be relatively confident that the rules for the given paradigm are correct and comprehensive, and the paradigm is set. If, however, the analyzer generates some incorrect forms, the informant manually corrects the errors and tells the analyzer to relearn the rules for the paradigm. This learning loop proceeds as long as necessary until the analyzer produces correct forms for a representative sample of words belonging to the paradigm.

This learning loop is necessary not only to create a robust morphological analyzer, but also to teach the informant the strengths and limitations of computational methods. For example, the informant will be told, and may test for himself, that mutations of a similar sort can be handled within a single paradigm as long as the inflectional endings are the same and an example of each mutation is provided explicitly. For example, the analyzer has proven capable of handling Russian mutations of the type *s/Ö* and *z/Š* within a single verbal paradigm. Deviations from the major paradigms will be listed explicitly in the open-class lexicon.

It goes without saying that the rich and complex inflectional morphology of Slavic languages provides rigorous testing grounds for Boas's morphological analyzer. More importantly, however, Slavic languages contain all the *categories* of paradigm-related complications that we have found in surveys of other languages. Consider the following sampling of phenomena:[3]

Stem-internal alternations will be handled by positing multiple stems in the lexicon for each word assigned to the given paradigm. In Slavic, listing multiple stems will account for Russian fleeting vowels (*otec* 'father.NOM.SG' ~ *otca* 'father.GEN.SG') and Belorussian graphotactic vowel reduction (*stol* 'table.NOM.SG.' ~ *stala* 'table.GEN.SG.'). Outside of Slavic, listing multiple stems

[3] For reasons of space, examples are presented sparingly; "etc." should be assumed in all instances. In addition, sources are not presented after each language example, as all the language examples were either taken from the grammars listed in the bibliography (page numbers are cited there) or gathered from in-house native speakers.

will account for suppletion in Comanche and Blackfoot verbal paradigms: in Comanche, intransitive verbs are suppletive for singular versus plural subjects, while transitive verbs are suppletive for singular versus plural objects (113); in Blackfoot, intransitive verbs have different stems for animate and inanimate subjects (*siksinámma* 'it.ANIMATE is black' / *siksinátts̆iwa* 'it.INANIMATE is black') (38).[4] To reduce overgeneration, the user will be asked to indicate, when possible, which stem is associated with which parameter value or combination of parameter values (e.g., stem 1: present tense, stem 2: past tense).

Boundary alternations can be incorporated into paradigms as long as they are mandatory; for example, one cannot include in a single paradigm words that do and do not have *s* → *Ö* shift in the Present 1st Singular. The morphological analyzer has been positively tested for boundary alternations in Polish verbs (*wożę* 'drive.1.SG.PRES' ~ *wozisz* 'drive.2.SG.PRES') and is expected to work equally well for Finnish consonant gradation (*kauppa* 'shop' ~ *kaupat* 'shops') and Blackfoot vowel shortening (*kakkóówa* 'pigeon' ~ *kakkóiksi* 'pigeons') (9).

Multiple realizations of an inflectional form can be incorporated into the paradigm, when regular. In Slavic, the future tense has regular duplicate forms in Polish (*będę robił* ~ *będę robić* 'will work.1.SG') and Ukrainian (*robitimu* ~ *budu robiti* 'will work.1.SG'). In Blackfoot, many verbs have two or three acceptable past tense forms, which are formed by rules of different paradigms: e.g., *Nitókska'si/Nitsíikska'si* 'I ran' (36). If duplicate forms are idiosyncratic or apply to a limited number of lexical items, they can be added to the paradigm at the stage of open-class acquisition. This applies to the so-called 2nd locative in Russian (*lese* ~ *lesu* 'forest') and to variant realizations of absolutive nouns in Nahuatl (*tochin* ~ *tochtli* 'rabbit') (17).

---

[4] Page numbers for examples that are drawn from grammars are presented in 'bare' form, since there is generally only one grammar per language listed in the references.

Slight Irregular Modifications of Paradigms will be handled by overriding one or more inflectional forms during open-class lexical acquisition. The user will assign the word to the paradigm that has the closest fit, then click on the 'show me forms' button to see what inflectional forms the morphological analyzer generates. The forms are generated in an editable field where the user can make all necessary corrections. In Slavic, this "manual override" process can be used for Russian masculine plurals in stressed *–á* (*adres* 'address' ~ *adresá* 'addresses') and for Polish masculine dative singulars in *–u* (*brat* 'brother.NOM' ~ *bratu* 'brother.DAT'). Outside of Slavic, it should be particularly useful for languages that have highly complicated, largely unpredictable inflectional patterns. For example, in describing nominal inflection in Irish, Ó'Siadhail (1989:159) notes,

> It is very difficult to predict how the plural of any given noun is formed; nevertheless the phonetic environment and the function of the plural play a certain part in determining the formation.

Thus, rather than force the user to create some ideal system of paradigm delineation, which might necessitate positing hundreds of paradigms, Boas permits him to posit far fewer paradigms and tweak them as necessary during open-class lexical acquisition.

Multiple Paradigm Templates can exist for a given part of speech. For example, *pluralia tantum* nouns have no singular forms (e.g., Russian *časy* 'watch'), and certain classes of verbs lack a given mood or aspect in some languages (e.g., intransitive verbs commonly have no passive). Multiple templates can also exist for reasons idiosyncratic to a given language: for example, Nahuatl permits plurals only for animate nouns, so inanimate paradigms will be half the size of animate ones (16). Similarly, Blackfoot has no singular-plural distinction for non-particular nouns (11). The user can create as many different templates as necessary when establishing paradigms in the morphological component of Boas.

The only paradigm-related process for which there is no obvious test material in Slavic is inflectional reduplication, which is used, for example, to create different verbal aspects in Ponapean (tense is conveyed pragmatically; assume past tense in this example): *kang* '(I)-ate' ~ *kangkang* '(I)-was-eating.DURATIVE' (74). Reduplication is also used to form plurals in Nahuatl according to the following rule: reduplicate the first syllable and add the suffix *-tin: teuctli* 'lord' ~ *teteuctin* 'lords' (17).[5]

Since the morphological analyzer in Boas is a finite state machine, it interprets only strings of characters, not patterns. Therefore, reduplication must be handled by scripts unconnected to the morphological analyzer. When the informant is establishing inflectional paradigms, he will indicate what, if any, forms are generated via reduplication. These forms will be excluded from the learning loop of the morphological analyzer. Then, in a separate task, the informant will select from a list the appropriate pattern of reduplication and provide a handful of examples. Based on this information, a script will be automatically generated, which the informant and programmer can modify, as necessary. If the patterns of reduplication are too complex to be captured in explicit rules, the reduplicative forms will have to be listed individually for every lexical item.

2.2. Closed-Class Lexicon

In some languages, closed-class items regularly attach to stems and must be stripped off to reveal the citation form. The affixal realization of closed-class items will be captured during closed-class lexical elicitation, in which English closed-class senses can be translated into SL as a *word*, *phrase*, *affix*, or *feature* (e.g., case). Below are some Slavic and non-Slavic examples.

(1) BULGARIAN articles: *more* ~ *moreto* 'sea ~ the sea'

[5] Derivational reduplication will not be handled productively, for reasons explained in Section 3.

RUSSIAN reflexive/reciprocal affix: *myt'* ~ ***myt'sja*** 'wash ~ wash oneself'
PERSIAN possessive pronouns: ***kt\|b*** ~ ***kt\|bt*** ' book ~ your book'
ARABIC prepositions: *byt* ~ ***bbyt*** 'house ~ in a house'
Cree possessive pronouns: *astotin* ~ ***nitastotin*** 'cap ~ my cap' (44)
NAHUATL possessive pronouns: *michin* ~ ***nomich***[6] 'fish ~ my fish' (26)
PONAPEAN demonstratives: *wahr* ~ *wahret* 'canoe ~ this canoe' (86)
MALAY interrogative particle: ***-kah*** attaches to the word questioned (123)
COMANCHE reflexive/reciprocal affix: ***na-*** attaches to the verb (103)
MALAY prepositions: *rumah sakit* ~ ***di-rumah sakit*** 'hospital ~ in the hospital' (79)

When applicable, allomorphs of attached affixes and/or their inflectional forms will be listed explicitly in the closed-class lexicon. In Bulgarian, e.g., allomorphs will be listed for the masculine singular definite article: *-ъt, -jat, -a, -ja*; in Cree, they will be listed for the possessive pronoun 'my': *ni-, nit-, n-* (44-46); and in Comanche, they will be listed for the locative postpositions *on, under, in, at, from,* and *beside*, each of which has between two and five allomorphs (73).

For each affixal realization of a closed-class sense, the user will indicate what part(s) of speech it can attach to for purposes of disambiguation. For example, in Russian and Polish, instrumental *with* is reflected on noun phrases by the feature 'Instrumental case', while in Ponapean it is reflected on verbs by the suffix *–ki*. Compare the translations of *I will write with this pen.*

(2) I pahn ntingki pehnet. [Ponapean: 224]
I will write-**with** pen-this

[6] The absolutive suffix is removed when the possessive prefix is added.

Ja	budu	pisat'	ètoj	ru̯koj.	[Russian]
I	will	write	this.INSTR	pen.INSTR	

2.3. Derivational Affixation

For reasons explained in Section 3.1, Boas will not attempt to generate rules for all derivational processes in SL. There is, however, one subset of derivational forms that will be handled productively: forms created by affixes that have a direct English counterpart. Productive affixes in English have been divided into several dozen semantic classes, like 'simple negation' (*un-, in-, im-, non-*), 'very' (*super-, extra-*), 'against' (*anti-*). The informant will supply corresponding affixes in SL, if they exist. Although this elicitation might yield no results for some languages, for others it will prove fruitful: for example, it will catch negated verbs in Czech, which are formed by the prefix *ne-* (*mluvím* ~ ***nemluvím*** '(I) speak ~ (I) don't speak') and negated adjectives in Ponapean, which are formed by the prefix *sa-* (*peik* ~ ***sapeik*** 'obedient ~ disobedient'). Likewise, it will cover Blackfoot words modified by affixes meaning 'very' (*iik-*) and 'extraordinarily' (*sska'-*) (92). The English generator will be responsible for mapping, 'not + proper' to *improper* and 'not + obedient' to *disobedient* (avoiding invalid formulations like ****inproper*, ****inobedient*** and ****non-proper*, ****non-obedient***).

This elicitation also gathers affixes whose main function is to change the part of speech with little or no accompanying semantic shift (e.g., noun → adjective in *success* → ***successful***). This subtype of derivational morphology was singled out because it permits (relatively) direct transfer from SL to English, something impossible for many other derivational processes.

2.4. Open-Class Lexicon

The open-class lexicon in Boas, as everywhere, is the seat of things unpredictable: translations of words, their paradigm membership (if not fully predictable by the spelling of the citation form),

irregular inflectional forms, allomorphs, etc. As is well known, the bigger the lexicon, the better the machine translation system. The challenge for Boas lies in maximizing the effectiveness of limited lexicon-building resources—namely, having just one language informant devoting less than six months to the task. Due to these time constraints, it is unlikely that the informant will be able to translate all of the 60,000 English word senses resident in Boas, as well as all the words in the SL corpus not covered by these word senses. Therefore, the informant is encouraged to organize open-class lexical acquisition in the way most efficient for his language and most in keeping with his goals (e.g., coverage of articles dealing with medicine or nuclear proliferation). Below we consider the most morphologically salient aspects of open-class lexical acquisition, again focusing on classes of phenomena and the Slavic examples that will serve as testing material.

Non-Inflectional Spelling Variants. Some languages have non-inflectional spelling variants of lexical items, like Russian *znaŗen'e/znaŗenie* 'meaning' and *predstavlen'e/predstavlenie* 'presentation'. Mokilese, for example, has so-called cluster metathesis, by which clusters composed of a labial and a velar stop can occur in either order: *apkas* ~ *akpas* 'now'. Since these alternations are rather idiosyncratic, they are best listed in the open-class lexicon as allomorphs.

More problematic are widespread alternations, like Irish lenition and eclipsis, which are phonologically driven processes that modify word-initial consonants based on the preceding lexical item. Table 1 presents a sample of such alternations:

Table 1. Lenition and Eclipsis in Irish

basic consonant	lenited consonant	eclipsed consonant
c	ch	gc
b	bh	mb
g	gh	ng

Lenition can occur, for example, after the preposition *ar* 'on': ***bad*** 'boat' → *ar* ***bhad*** 'on (the) boat'. Eclipsis can occur after the

positive interrogative particle *an*: ***bris*** 'break' → ***An mbriseann se...?*** 'Does he break...?'

The more primitive but fool-proof way to handle such alternations would be to list the variant spellings as allomorphs in each relevant lexical entry, but this method would carry high time costs for the informant. Alternatively, the linguist/programmer team could write lexicon-wide rules for all such alternations, but this would require a type of unguided rule writing potentially out of the reach of less experienced linguist/programmer teams (much depends on their respective knowledge of NLP). We are currently working to develop rule templates to assist in this process, and will test them on the Ukrainian word-initial alternations *u-/v-* and *i-/j-* as in: *uɾitel/vɾitel* 'teacher' and *idu/jdu* '(I) go'.[7]

Rule writing carries its own complications, as evidenced by the Ukrainian phenomena mentioned above. For example, imposing lexicon-wide rules can lead to overgeneration: in Ukrainian, place names like *Ural* 'Urals' and foreign words like *uran* 'uranium' do not have a *v-* variant (27-8). In most instances, such overgeneration is irrelevant (and simply adds a bit of dead weight to the lexicon), since the translation system will be primarily interpreting, not generating, SL.[8] However, in some instances overgeneration will lead to ambiguity. For example, Ukrainian *uklad* means 'regime' while *vklad* means 'contribution'. A lexicon-wide rule that puts *u-* and *v-* in free variation word initially will cause each instance of *uklad* and *vklad* to be incorrectly tagged with two meanings. But considering the amount of lexical and other ambiguity that all translation systems face, these additional sources of ambiguity are relatively insignificant.

---

[7] These letters also alternate as freestanding prepositions, but as prepositions they will simply be listed as allomorphs in the closed-class lexicon.

[8] 'Primarily' interpreting because the morphological analyzer will be used to generate forms in the morphological learning loop and in the open-class lexicon during paradigm selection/modification.

3. Derivational Morphology

Words formed by derivational morphological processes present significant problems to MT systems even if the source language is known and can be prepared for individually. These problems are compounded when source language is unknown. The sections below detail the problems inherent in derivational morphology and Boas' rather unconventional approach to dealing with this aspect of the grammar.

3.1. Derivational Morphology: The Problem

Source language words formed by productive derivational processes (like the German *Donaudampfschiffahrtskapitän* 'Donau steam ship driver captain') will, in large part, not be captured by Boas' English-driven open-class lexical acquisition since such words are equivalent to multi-word English phrases. This presents a considerable problem for languages with widespread compounding (German, Swedish) and/or reduplication (Tagalog, Ponapean). One obvious way to handle derivational word formation would be to prepare the system to analyze such forms based on knowledge elicited from the informant. While creating a series of questions would be trivial ('How many roots can typically be joined in a compound?' 'What, if any, letters can be added between compounding roots?' 'Do compounding forms use different roots than non-compounding forms?'), processing the results presents significant complications. The problem lies in the fact that derivational forms are often semantically ambiguous and/or non-compositional. Thus, even correct formal analysis of derived words would often be of little help in SL-to-English lexical transfer.

Consider, for example, the Swedish surface form *frukosten*, which can have the following five parses (from Karlsson *et al*, 1995:28).

(3) a. *frukost + en* 'the breakfast'
b. *frukost_en* 'breakfast juniper'
c. *fru_kost_en* 'wife nutrition juniper'
d. *fru_kost+en* 'the wife nutrition'
e. *fru_ko_sten* 'wife cow stone'

Such compounding ambiguities abound in Swedish. Dura (1998) suggests that the best way to deal with them is to list the most common compounds explicitly in the lexicon, then use these ready-made chunks as set units for further analysis of compounding forms.

Another problem inherent in compounding is the opaque semantics of many compounds. For example, a Comanche grammar calls the word for 'Mexican restaurant' a compound composed of the elements 'fat-white-man-*possessive*-eat-house.' Even if Boas could decompose the components of such a compound, it would never generate a correct English equivalent.

Derivational reduplication poses even more fatal problems, both in formal and in semantic terms. For example, in Tagalog the meaning 'a vendor of the product indicated by the base' is created as follows (103): [prefix *mag*] + [first two letters of the base, reduplicated] + [base].

(4) *magbubulaklak* 'flower vendor' (*bulaklak* 'flower')
magkakandila 'candle vendor' (*kandila* 'candle')

Clearly, such word formation processes can only be captured by language-specific rules that are (i) difficult, if at all possible, to elicit in a generalized way, (ii) limited to certain semantic classes of lexical items, and (iii) not always strictly compositional in meaning.

Further complications arise from the 'theme and variations' nature of reduplication. For example, Turkish color terms can be intensified by reduplication that includes various consonant additions/mutations: *siyah* ~ *simsiyah* 'black ~ very black', *mor* ~ *mosmor* 'purple ~ very purple'. Ponapean shows similar formal

variations, as evidenced by the following reduplicative forms (leaving the meanings aside): *pa ~ pahpa, it ~ itiht, alu ~ alialu.*

3.2. Derivational Morphology: Boas' Answer

Because of the complications associated with derivational word formation, Boas will treat it lexically, assisted by the on-line corpus.[9] More specifically, the informant will begin open-class lexical acquisition by providing SL equivalents for some minimum number of high-frequency English words. Open-class acquisition can then proceed in a number of ways, as deemed best by the informant. He can: (i) continue to translate English word senses resident in Boas; (ii) scan the SL corpus and translate the most frequent SL words into English; or (iii) scan an English corpus devoted to some special topic and translate the most frequent words therein. These methods of lexical acquisition can be carried out in loops, as dictated by the informant (e.g., 200 English words, then 300 words from the SL corpus, then another 200 words from the SL corpus...).

For languages with extensive derivational morphology, the SL corpus scan should be used broadly, since English-driven lexical acquisition will miss many common words (cf. 'flower vendor' and 'very black' above). For languages with less extensive derivational morphology, English-driven lexical acquisition should provide relatively good coverage.

When the corpus scan is employed, it will generate a list of unknown words in order of frequency for the informant to potentially translate. 'Potentially' is an important notion, as a given corpus might include a large number of rare words.

While Slavic languages do not show extensive compounding or reduplication, they have other derivational word formation processes on which the corpus-scan method of lexical supplementation can be tested. For example, English-driven lexical

[9] The primary goal of Boas is to translate on-line resources into English; thus we assume the existence of on-line texts that can be compiled into a corpus by the programmer.

acquisition will not capture quasi-productive prefixation (like Russian *dopet'* 'sing to the end' and *pereutomit'sja* 'get overtired') or diminutive/endearing forms (like Russian *košerka* < *koška* 'cat'). These forms, like compounds and reduplicative forms, often do not have entirely compositional semantics.

4. Conclusions

It is virtually impossible to build a knowledge elicitation system that specifically caters to every linguistic eventuality encountered in every natural language, since it is virtually impossible even to list all such eventualities. Therefore, when facing the task of creating a knowledge elicitation system with maximal coverage, strategy plays a crucial role. Chance also plays some role. In the case of Boas, chance dictated that the linguist developers had more knowledge of Slavic languages than, say, of African languages, making testing of the former language group more realistic than testing of the latter. However, strategy proves no less important: by dividing language phenomena into typologically valid classes whose members can be handled similarly, we can test a given elicitation process on Slavic languages with relative confidence that equally good results will be achieved in the more 'exotic' languages for which testing lies beyond our reach.

References

Charney, J.O. 1993. *A Grammar of Comanche*. Lincoln: University of Nebraska Press.

deBray, R.G.A. 1980. *Guide to the East Slavonic Languages*. Columbus, Ohio: Slavica Publishers.

Dura, E. 1998. *Parsing Words*. Göteborg, Sweden: Göteborg University.

Frantz, D.G. 1991. *Blackfoot Grammar*. Toronto: University of Toronto Press.

Harrison, S.P. 1976. *Mokilese Reference Grammar*. Honolulu: The University Press of Hawaii.

Heim, M. 1982. *Contemporary Czech*. Columbus, Ohio: Slavica Publishers, Inc.

Hubenova, M. et al. 1968. *A Course in Modern Bulgarian*. Columbus, Ohio: Slavica Publishers, Inc.

Karlsson, F. 1995. "Designing a Parser for Unrestricted Text." In Karlsson, F., A. Voutilainen, J. Heikkilä, A. Anttila (eds.) *Constraint Grammar*. New York: Mouton de Gruyter.

Lewis, M.B. 1954. *Teach Yourself Malay*. London: English Universities Press, Ltd.

Medushevsky, A. and R. Zyatkovska. 1963. *Ukrainian Grammar*. Kiev: Radyanska shkola.

Nirenburg, S. and V. Raskin. 1998. "Universal Grammar and Lexis for Quick Ramp-Up of MT Systems," in *COLING-ACL '98* (36th Annual Meeting of the Association for Computational Linguistics), vol. II, 975-979.

Oflazer, K. and S. Nirenburg. 1999. "Practical Bootstrapping of Morphological Analyzers." In *Proceedings of the Workshop on Computational Natural Language Learning at EACC '99*, Bergen, Norway.

Ó'Sé, D. and J. Sheils. 1993. *Irish*. Lincolnwood, Illinois: NTC Publishing Group.

Ó'Siadhail, M. 1989. *Modern Irish*. Cambridge: Cambridge University Press.

Ó'Siadhail, M. 1995. *Learning Irish*. New Haven: Yale University Press.

Rehg, K.L. 1981. *Ponapean Reference Grammar*. Honolulu: The University Press of Hawaii.

Schachter, P. 1972. *Tagalog Reference Grammar*. Berkeley: University of California Press.

Sullivan, T.D. 1988. *Compendium of Nahuatl Grammar*. Translated from the Spanish by T.D. Sullivan and N. Stiles. Salt Lake City: University of Utah Press.

CRL, Box 30001, 3CRL
New Mexico State University
Las Cruces, NM 88003
marge@crl.nmsu.edu

Optimal Stress Patterns In Russian Nouns*

Irina Milnes and Caroline Wiltshire
University of Florida, Gainesville

1. Introduction

Stress in Russian is traditionally described as free and mobile, and both Russian and Western linguists have made numerous attempts to reduce the seemingly unlimited number of stress paradigms. A recent approach in Optimality Theory (Prince and Smolensky 1993) analyzes Russian within a theory of Root Controlled Accent systems (Alderete 1999). Alderete's account handles data with fixed root stress throughout the paradigm by ranking a constraint on faithfulness to underlying root stress above constraints such as faithfulness to underlying suffix stress, an analysis we agree with here. However, in order to analyze paradigms in which stress is fixed on inflectional suffixes, he posits default right stress for Russian. Furthermore, in order to handle mobile stress patterns, he introduces an entirely new family of constraints on Transderivational Anti-Faithfulness (TFA).

We propose here an account of suffixal and mobile stress which avoids introducing this problematic new family of TFA constraints. As a case study, we analyze stress patterns in Russian second declension feminine nouns, and show that our account also handles Russian long-form adjectives, without recourse to TFA. While we believe that our account can be extended to other grammatical classes in Russian, due to limitations of space, we have not attempted to do more than suggest such extensions here.

The paper is organized as follows. Section 2 provides the Russian noun data. Section 3 briefly reviews surface vs. underlying approaches, in order to show how we make use of both, and

* The authors' names are in alphabetical order, since our contributions were equal. We would like to thank Bill Sullivan for insight into the problem and for the data provided, and Irina Sekerina and Christina Bethin for their careful reviews which greatly improved this paper. All remaining errors are our own.

illustrates the parts of Alderete's analysis that we will adopt here. Section 4 develops our account of Russian noun stress in the second declension feminine, and extends the analysis to adjective stress. Section 5 compares our approach to Alderete's and summarizes the research and possible extensions.

2. Noun Data

The first set of data consists of nouns (fem., second decl.) grouped (a-e) based on stress patterns in the paradigm (Sullivan 1969).[1]

(1) Russian Feminine Second Declension Nouns[2]

	a.	b.	c.	d.	e.
SG NOM	[kajmá]	[žába]	[žená]	[zimá]	[ruká]
ACC	[kajmú]	[žábu]	[ženú]	[zímu]	[rúku]
GEN	[kajmɨ́]	[žábɨ]	[ženɨ́]	[zimɨ́]	[rukí]
DAT	[kajmé]	[žábe]	[žené]	[zimé]	[ruké]
INST	[kajmój]	[žáboj]	[ženój]	[zimój]	[rukój]
PL NOM	[kajmɨ́]	[žábɨ]	[žónɨ]	[zímɨ]	[rúki]
ACC	[kajmɨ́]	[žáb]	[žón]	[zímɨ]	[rúki]
GEN	[kajóm]	[žáb]	[žón]	[zím]	[rúk]
DAT	[kajmám]	[žábam]	[žónam]	[zímam]	[rukám]
LOC	[kajmáx]	[žábax]	[žónax]	[zímax]	[rukáx]
INST	[kajmámi]	[žábami]	[žónami]	[zímami]	[rukámi]
	'border,edge'	'toad'	'wife'	'winter'	'arm,hand'

Groups (a) and (b) showed fixed stress, on the stem and suffix respectively. Groups (c-e) show mobile stress, on the stem or suffix in different parts of the paradigm. As revealed by forms with similar syllabic and segmental structures yet different stress, such as [ženú]

[1] There is dialectal, social, and situational variation for some of the forms here; accounting for such variation is beyond the scope of our research.

[2] [x] = voiceless velar fricative, [ž] = post-alveolar voiced fricative, [ɨ] = close back unrounded vowel. [ó] in the (1c) plurals, orthographically represented as ё, is treated here as a reflex of /é/ following Melvold (1990). Palatalization is not represented on consonants before front vowels, where it is predictable.

vs. [zímu], the placement of stress is morpho-phonological rather than purely phonological, as discussed in Section 4.2.

3. Approaches to Stress Patterning

Traditionally, Russian nouns have been analyzed from one of two perspectives, based on either surface or underlying forms. Section 2.1 compares these two approaches, while Section 2.2 suggests a combined surface-underlying analysis of Russian stress, the approach we use. Section 2.3 describes the machinery from Alderete (1999) which forms the basis of our account in OT.

3.1. Surface vs. Underlying

Surface representation approaches attempt to give a detailed description of the whole inventory of Russian nouns, grouping them according to how native speakers stress forms of the nouns in a paradigm. A typical surface representation approach is presented in Avanesov (1956). Differentiating between fixed and mobile stresses in Russian, Avanesov suggests at least six types of mobile stress. Even though his classification is quite detailed, he does not cover every Russian noun, e.g., case (1d) is not included. To describe the whole inventory of Russian nouns based on surface representations requires a lengthy list and results in a picture of stress in Russian as abundantly variable and idiosyncratic.

In underlying representation approaches, deep structure stress and phonological rules which manipulate it in the lexicon are used to account for the surface variety in Russian. A typical underlying representation analysis is Coats (1976). Assuming that stress is an inherent property of morphemes, Coats, after Halle (1973), introduces four types of Russian polysyllabic nouns and a set of rules to convert them into surface forms. His rules, while intended to reflect the "natural tendency for lexical entries to become simpler" (1976:8), in fact result in a complex system of stress assignment, such as rules which first remove and then reapply stress in his Type 3 nouns. Although the input is simplified, the required sequencing of the rules results in a complicated route from input to

output. Compared with surface representation analyses, underlying representation analyses considerably reduce the number of stress patterns by assigning inherent stress in the lexicon. However, the rules which turn underlying stresses into surface stresses remain language-specific and idiosyncratic.

3.2. Combined Surface/Underlying Representation Approach

The combined approach uses both underlying and surface representations, reflecting the morpho-phonological nature of stress in Russian. Examples of the combined approach are Sullivan (1969), Kiparsky and Halle (1977), and Melvold (1990). For the data in (1), Sullivan assigns morphologically-based stress in the following underlying specifications, which we use in our analysis. Roots are stressed, post-stressed, or unstressed; affixes are stressed or unstressed. Note that some roots have allomorphs with different stress patterns, depending on case and number.

(2) Underlying Stress specifications for (1)
- a. stressed roots: /žáb/ /kajm′/ /žen′/ (/žén/ if PL.)
- b. unstressed roots: /zim/ (in SG.) /ruk/
- c. stressed suffixes: /á/ /í/ (GEN.SG.) /é/ /ój/ /áx/ /ám/ /ámi/
- d. unstressed suffixes: /i/ /ɨ/ (NOM.PL.) /u/

In forms like /kajm′/ and /žen′/, the stress is not associated with a particular vowel, although its presence in the underlying specification means that such roots are also in the stressed class. As this stress can be realized on a suffix rather than a root (see Section 4.2), it has also been called floating. The above stress assignments results in four main patterns in underlying forms:

(3) Four Russian Noun Patterns

a) no stress:	/ruk + u/	/ruk + i/ (NOM. PL.)
b) one stress, stem:	/žáb + u/	/kajm′ + u/
c) one stress, suffix:	/zim + á/	/ruk + é/
d) two stresses:	/kajm′ + á/	/žáb + á/ /zím + ám/

Stress assignment on the surface gives the following patterns (the stressed morphemes are given in capitals):

(4) Russian Nouns Stress patterns (ours, not Sullivan 1969)

a.	/stem/ + /affix/ =	[STEM.affix]	[rúku]
b.	/STEM/ + /affix/ =	[STEM.affix]	[žábu]
c.	/stem/ + /FLOAT/ + /affix/ =	[stem.AFFIX]	[kajmú]
d.	/stem/ + /AFFIX/ =	[stem.AFFIX]	[zimá]
e.	/STEM/ + /AFFIX/ =	[STEM.affix]	[žába]
f.	/stem/ + /FLOAT/+/AFFIX/=	[stem.AFFIX]	[kajmú]

With phonemic stress indicated in the input, the output results from the simple generalization that the leftmost wins, and the left edge is stressed if there are no stressed morphemes in the input.

The combined approach is able to account for some 'marginal' stress patterns (e.g., GEN.SG. [rukí] and NOM.PL. [rúki], or NOM.SG. [zimá] and LOC.PL. [zímax]), which finally fit the Russian stress paradigm. This is due to considering underlying, as well as surface, properties of Russian nouns. We further simplify the analysis and connect it to cross-linguistic tendencies by applying OT methodology to the data in (1).

3.3. Alderete's Approach

Optimality Theory provides powerful tools for matching a certain phenomenon, like stress, with its motivations, i.e., the constraints that outputs aim to satisfy. Given an input, possible outputs are evaluated to determine which best satisfies the ranked set of constraints. A phenomenon results from a variety of constraints, universal in their nature, but languages differ in how these constraints are ranked. The optimal candidate may violate a particular constraint in order to better satisfy a more highly ranked constraint in the language-specific ranking.[3] Drawing on both phonological (surface) and morpho-phonological (correspondence)

[3] For readers unfamiliar with OT, see Prince and Smolensky (1993) or Archangeli and Langendoen (1997).

constraints, OT makes it possible to simplify the analysis of Russian noun stress. Alderete (1999) provides an account of Russian as a case of Root Controlled Accent, and we adopt some of his analysis here. The differences in our analysis, in order to handle the data in (1) without resorting to TFA constraints, are developed in Section 4.

Alderete (1999) discusses languages in which the input specification of the root as stressed or unstressed is of crucial importance to the output.[4] The correspondence constraints MAX-PROM$_{ROOT}$ and MAXPROM relate input forms to output, and are morpho-phonological in that they tend to preserve phonemic stress, stating that input (S_1) stress is realized as output (S_2) stress.

(5) MAXPROM$_{ROOT}$: Every stress in a root in S_1 has a correspondent in S_2.

(6) MAXPROM: Every stress in S_1 has a correspondent in S_2.

The specific constraint on roots in (5) gives a special role to root stress relative to any other stress in input, following McCarthy and Prince (1995) on morphologically dispersed faithfulness, Beckman (1997) on positional faithfulness, and Alderete (1999) on root controlled systems. Both constraints (5) and (6) are similarly evaluated, in that they can only be violated if an input stress is present. That is, a root or affix without input stress cannot violate either constraint. Roots and affixes with input stress that is realized as output stress satisfy the constraints; an affix with input stress realized without output stress violates (6); and a root with input stress realized without output stress violates both (5) and (6). Thus stems and affixes may surface with or without stress, and whether these output forms satisfy constraints (5) and (6) depends on their input specifications.

---

[4] In his account, surface stress is evaluated relative to input stress by correspondence constraints (McCarthy and Prince 1995), which are specified for Root faithfulness vs. Affix faithfulness. In our account, following Beckman (1997), we use Root faithfulness vs. faithfulness in general, but the difference is of no consequence to the cases discussed here.

The cross-linguistic character of MAXPROM$_{ROOT}$ is readily observable and supported by analogous examples from other languages. Smith (1998) uses a similar constraint for Tuyuca and Japanese. Furthermore, as Alderete argues, this sort of constraint fits into the observations of other properties controlled by roots rather than affixes, such as vowel harmony in Beckman (1997).

A second type of constraint we adopt from Alderete's analysis is the phonological constraint HEADED. In the prosodic hierarchy, prosodic words (PWs) dominate at least one foot, and feet are the domain of stress. As a result, every PW has stress.

(7) HEADED: Every prosodic word has exactly one stress.

In fact, the PW should have exactly one main stress, meaning that the constraint can be violated in two distinct ways: a word without any stress violates once, and words with more than one stress violate with each additional stress. This double evaluation implies that HEADED can be split into two separate constraints, but we use a single constraint here since both are undominated and unviolated in Russian.

HEADED applies to the Russian data on both the theoretical and native speaker levels. Similar primary stress patterns obeying HEADED have been analyzed by, e.g., Halle and Vergnaud (1987), Hayes (1995), Kager (1995), and Alderete (1999) for numerous languages. Typologically different languages support the cross-linguistic character of its motivation.

4. Our Development of an OT Account

Section 4 is organized as follows. Section 4.1 specifies the remaining constraints we use here, and Section 4.2 develops their ranking, showing how we account for the noun data. Section 4.3 uses a subset of the same ranking to account for stress in Russian long-form adjectives.

4.1. Constraint Specification

In addition to MAXPROM$_{ROOT}$, MAXPROM, and HEADED, we use two further constraints: EDGE(L) and POSTSTEMSTRESS. First, EDGE(L) prefers stress at the left-edge of the word, so that this is the default location for Russian:

(8) EDGE(L): a peak of prominence lies at the left edge of PW.

EDGE(L) is unviolated if the left edge has stress. If stress occurs elsewhere, a violation is assessed for each syllable that appears between the stress and the left edge of the word. The constraint originates from EDGEMOST (Prince and Smolensky 1993), a kind of alignment constraint. Earlier observations of the phenomenon date back to Trubetzkoy, who noted word-initial stress in a large number of languages (1971:277); more recent observers include Halle and Vergnaud (1987), Hayes (1995), and Halle and Idsardi (1995). The cross-linguistic character of the constraint is easily supported by analogous examples from many languages.

For Russian, stress at the left edge of a word usually means stressing the stem, so that it is difficult to separate the effects of EDGE(L) from MAXPROM$_{ROOT}$. However, stress is preferred on the left not only when marked in input (stressed root, unstressed suffix), but also in the cases in which neither stem nor affix are stressed, so that neither MAXPROM constraint can play a role.

The final constraint appears to be more language specific:

(9) POSTROOTSTRESS: the left edge of a stress aligns with the right edge of a root

This constraint, also a member of the alignment family (McCarthy and Prince 1993), specifies that stress immediately follows the root. It will be crucial for roots with a floating stress combined with an unstressed affix (c.f., /kajm′ + u/ = [kajmú]). The constraint used here is similar to the constraint used by Alderete (1999); however, we diverge in the choice of underlying representations (see Section 5) and the interpretation of this constraint. In addition to Russian, a

similar constraint ALIGN-STEM-L was supported by Kager (1994) based on data from Diyari and Dyirbal. The cross-linguistic character of this constraint is thus supported by analogous examples from other languages.

4.2. Constraint Ranking

We now consider the role each constraint plays in Russian noun patterns by looking at the ranking relationships between them.

HEADED and EDGE(L). Together these determine stress in no-accent root + suffix combinations. Although we do not rank the two yet, we show in Tableau 1 that these phonological constraints alone assign default stress in unmarked cases. Other constraints (particularly MAXPROM) are unviolated and irrelevant to the pattern in question.[5]

Tableau 1. Analysis of /zim/ + /u/ (Accusative Singular)

/zim/ + /u/	HEADED	EDGE(L)
→ zímu		
zimú		*!
zimu	*!	
zímú	*!	

Only the output with stress on the left satisfies both HEADED and EDGE(L). Candidates with right-edge stress, two stresses, or no stress violate one of these constraints and are therefore suboptimal. Thus HEADED and EDGE(L) alone correctly determine the output.

MAXPROM. This constraint plays a decisive role in the case of unstressed roots plus stressed affixes, and therefore must outrank

[5] The presence of an output stress that was not present in input is also a violation of a constraint against the insertion of stress, DEP-IO(stress), which must therefore rank lower than HEADED. As DEP-IO(stress) does not determine the output in any of these forms, it is not discussed further here.

EDGE(L). Winning forms surface with stress on the affix, since MAXPROM$_{ROOT}$ cannot be violated in cases in which the root is unstressed in the input:

Tableau 2. Analysis of /zim/ + /í/ (Genitive Singular)

/zim/ + /í/	HEADED	MAXPROM	EDGE(L)
→ zimí			*
zími		*!	
zimi	*!	*!	

The unstressed root which surfaces violates EDGE(L), but the violation is tolerated due to higher ranking MAXPROM.

MAXPROM$_{ROOT}$. The general faithfulness to underlying prominence, captured by MAXPROM, is outranked however by the specific faithfulness to prominence marked on a root syllable, as shown by two-stress noun patterns. Being equally ranked with MAXPROM$_{ROOT}$, HEADED assures that the winning form has exactly one stress, although this means violation of MAXPROM in cases of two stresses in input. Crucially, therefore, HEADED and MAXPROM$_{ROOT}$ both outrank MAXPROM.

Tableau 3. Analysis of /žáb/ + /á/ (Nominative Singular)

/žáb/ + /á/	HEADED	MAXPROM$_{ROOT}$	MAXPROM
→ žába			*
žabá		*!	*
žábá	*!		

MAXPROM violations are tolerated, so that a single stress surfaces and suffix stress is lost. EDGE(L) cannot be used to place stress on the root here, as we saw previously that MAXPROM outranks EDGE(L); hence MAXPROM$_{ROOT}$ plays the crucial role in cases of competing stresses. Ranking MAXPROM$_{ROOT}$ highly does not affect our previous analyses, since the constraint is vacuously satisfied in the forms in Tableaux 1 and 2.

POSTROOTSTRESS. Also outranking EDGE(L) is the constraint that prefers stress after the root, as shown in the cases of roots with a floating stress. The floating stress in roots like /kajm′/ surfaces with stress on the suffix, due to POSTROOTSTRESS. MAXPROM is satisfied because the input stress is present in the output. HEADED is satisfied by the presence of a single stress, and we consider $\text{MAXPROM}_{\text{ROOT}}$ to be satisfied by the realization of root stress in the output, despite the location of stress on the suffix, since location is not specified in floating stress roots. The winner violates only EDGE(L).

Tableau 4. Analysis of /kajm′/ + /u/ (Accusative Singular)

/kajm′/+/u/	MAXPROM	POSTROOTSTRESS	EDGE(L)
→kajmú			*
kájmu		*!	
kajmu	*!		

If a suffix that is stressed in input is added to a root with floating stress, the placement of output stress on the suffix involves either MAXPROM or POSTROOTSTRESS, both of which outrank EDGE(L).

Tableau 5. Analysis of /kajm′/ + / á/ (Nominative Singular)

/ kajm′/+/á/	MAXPROM	POSTROOTSTRESS	EDGE(L)
→ kajmá			*
kájma		*!	
kajma	*!		

EDGE(L). Finally, we have seen that the constraint preferring stress on the left is outranked by MAXPROM and POSTROOTSTRESS. As these are in turn dominated by $\text{MAXPROM}_{\text{ROOT}}$ and HEADED, EDGE(L) thus ranks lower than all other constraints. Winning forms can violate EDGE(L) except for unstressed affixes added to stressed or unstressed roots. Only in such cases does EDGE(L) assert itself, determining that these combinations are stressed on the left.

Constraint Ranking Conclusions. Based on the analysis of the feminine second declension nouns in (1), we come to the following conclusions. The location of stress results from a compromise between the intrinsic prominence specified for some roots and the preferred location for output prominence as governed by a set of constraints. OT provides the factors governing optimal stress patterns, which are ranked for Russian to capture how the compromise is reached. In the cases from (1), the optimal stress location is determined by the ranking:

(10) { HEADED, $\text{MAXPROM}_{\text{ROOT}}$ } >>
{ MAXPROM, POSTROOTSTRESS } >> EDGE(L)

The above pattern is both simple and general enough to describe one category of Russian nouns. The OT analysis also proves that stress assignment in Russian is a morpho-phonological rather than purely phonological phenomenon. Genitive singular and nominative plural forms, identical in their segmental representation, acquire different stress patterns due to the differently accented allomorphs they belong to: /zim/ in the singular (Tableau 2) and /zím/ in the plural (below).

Tableau 6. Analysis of /zím/ + /i/ (Nominative Plural)

/zím/ + /i/	HEADED	$\text{MAXPROM}_{\text{ROOT}}$	MAXPROM	EDGE(L)
→ zími			*	*
zimí		*!		
zimi	*!			

While listing allomorphs adds a form of complexity to the grammar, it is unavoidable due to the morpho-phonological nature of stress patterns in Russian nouns. Further evidence of the nature of the system is that neither correspondence nor surface based constraints alone can provide a complete analysis. The highest ranking HEADED (output-based) and $\text{MAXPROM}_{\text{ROOT}}$ (input-output correspondence) constraints are principal in stress assignment. The purely phonological constraint EDGE(L) is regularly violated in the surface

representation, as stress is forced to non-initial positions by other constraints.

Similarly, the morphologically based constraints MAXPROM$_{ROOT}$ and MAXPROM taken alone do not give the full picture either, especially in cases in which both morphemes are stressed or unstressed. There must be an interface between the underlying (morphological) and surface (phonological) constraints for the stress assignment to proceed optimally. Such an interface is easily provided within an OT correspondence account, and has proven useful for a wide range of phenomena which are influenced by both morphology and phonology, such as reduplication (McCarthy and Prince 1995) and syllabification (McCarthy and Prince 1993).

Russian noun paradigms contain language specific peculiarities due to the fact that a single morpheme may have two differently stressed allomorphs. The recognition of the morpho-phonological character of stress assignment in Russian allows us to separate out the characteristics which are truly phonological and systematic, from arbitrary facts that must be memorized, such as phonemic stress distinctions.[6]

4.3. Supporting the Ranking: Adjectives

The ranking from the nouns is supported by data from Russian long-form adjectives. We use adjectives in the masculine, switching gender to preserve parallel syllable structure. The paradigms below are grouped (1-2) in (11) below based on stress patterns both within and across the paradigms; note that stress is fixed either on the root, as in (1a-b) or on the suffix (2a-b). Each group is subdivided based on the so-called 'hard' (a) and 'soft (b) endings (/ɨ/ as opposed to /i/).

[6] The arbitrariness in the allomorphy may also be reducible, using some kind of classification with default inheritance, as in Brown et al. (1996).

(11) Masculine adjectives[7]

	1a.	1b.	2a.	2b.
SG NOM	[bélɨj]	[gíbkij]	[živój]	[bol'šój]
ACC	[bélɨj]	[gíbkij]	[živógo]	[bol'šój]
GEN	[bélogo]	[gíbkogo]	[živógo]	[bol'šógo]
DAT	[bélomu]	[gíbkomu]	[živómu]	[bol'šómu]
LOC	[bélom]	[gíbkom]	[živóm]	[bol'šóm]
INST	[bélɨm]	[gíbkim]	[živím]	[bol'ším]
PL NOM	[bélɨe]	[gíbkie]	[živíe]	[bol'šíe]
ACC	[bélɨe]	[gíbkie]	[živíe]	[bol'šíe]
GEN	[bélɨx]	[gíbkix]	[živíx]	[bol'šíx]
DAT	[bélɨm]	[gíbkim]	[živím]	[bol'ším]
LOC	[bélɨx]	[gíbkix]	[živíx]	[bol'šíx]
INST	[bélɨmi]	[gíbkimi]	[živímі]	[bol'šímі]
	'white'	'flexible'	'alive'	'big'

Adjectives are assigned the following underlying stresses:

(12) Underlying Stress Specifications
a. stressed roots: /bél/ /gíb/
b. post-stressed roots: /živ'/ /bol'š'/
c. unstressed suffixes: /ɨj/≈/ij/ /ɨm/≈/im/ /ɨe/≈/ie/
/ɨx/≈/ix/ /ɨmi/≈ /imi/ /oj/ /ogo/
/omu/ /om/

Because the suffixes are all unstressed, in contrast to the noun suffixes which may be stressed or unstressed, the resulting patterns are simpler. The morpho-phonological input of the above stress assignments takes the form of two main stress patterns:

(13) Two Russian Adjectives patterns
a. one stress, root: /bél + ɨj/ = [bélɨj]
b. one stress, floating: /bol'š' + oj/ = [bol'šój]

---

[7] The accusative has the [-ogo] variant if the noun is animate; we have illustrated this with the adjective 'alive' for which this variant would be most natural.

The ranking for adjectives is a subset of the ranked constraints from (10) for nouns. Tableaux 7-8 provide illustrative examples.

(14) MAXPROM$_{ROOT}$ >> POSTROOTSTRESS >> EDGE(L)

Tableau 7. Analysis of /bél + ɨj/ (Nominative Singular)

/bél + ɨj/	MAXPROM$_{ROOT}$	POSTROOTSTRESS	EDGE(L)
→[bélɨj]			
[belɨj]	*!		*
[belɨj]	*!		

The dominant MAXPROM$_{ROOT}$ chooses the winner for root stressed adjectives (groups 1a-b). In affix stressed adjectives (groups 2a-b), POSTROOTSTRESS asserts itself.

Tableau 8. Analysis of /bol'š'/ + /oj/ (Nominative Singular)

/bol'š'/+/oj/	MAXPROM$_{ROOT}$	POSTROOTSTRESS	EDGE(L)
→[bol'šój]			*
[ból'šoj]		*!	
[bol'šoj]	*!		

The constraint ranking for adjectives is thus compatible with that of the nouns. As they have limited stress patterns relative to nouns, long-form adjectives show predictable and unmarked stress assignment.

5. Comparison and Conclusion

5.1. Comparison with Alderete (1999)

Although we cannot do justice to Alderete's account in this limited space, we do want to point out the major differences in our analysis. Following traditional accounts, we have proposed three possible input forms for roots: stressed, unstressed, and floating,

and two forms for suffixes: stressed and unstressed. In accounting for mobile stress forms, we have relied on allomorphy in combination with a consistent phonological system.

Alderete proposes to analyze our post-stressed roots, such as /kajm′/, as unstressed and sets default stress in Russian as being on the right. He treats our unstressed roots as lexically marked for stress on the left. To handle mobile stress, he proposes a new kind of constraint, called Transderivational Anti-Faithfulness (TFA), that requires an output form to be distinct in some way from other output forms in the paradigm. These constraints are triggered by specific affixes, such as NOM. PL. /-a/, and also must specify which output form the affixed form is to be compared to for distinctness.

(15)¬OO-MAX-PROM: In a pair of words in an Output-Output correspondence relation, at least one pair of correspondent segments is non-identical for prominence.

The insight is that distinct stress in different parts of the paradigm provides for easy recognition of which suffix has been added. However, the use of TFA increases the complexity of the grammar, as it requires additional constraints, specified not only for specific affixes but also for which forms are compared. Such constraints increase the power of OT grammars, blur distinctions between phonology and morphology, and raise questions of learnability.

By the account we have proposed, we showed that a successful account of Russian need not rely on such constraints, and, if true, this weakens Alderete's case for them. As only certain roots are mobile stressed, we believe that marking the roots for allomorphy captures the Russian system better than marking the affixes for anti-faithfulness.

5.2. Conclusions

We conclude that OT can help to simplify an analysis of Russian stress in nouns and adjectives. In addition to the cases discussed here, we believe that an extended account could be developed to

handle a larger portion of the data. For example, to handle polysyllabic nouns, in which placement of stress on the first or second syllable is distinctive (*kómnat-a* 'room' vs. *tetrád'* 'exercise book'), an additional constraint would prevent movement of stress within a root. We leave such further cases to future research.

To handle the admittedly limited cases discussed here, we have proposed an analysis based on only three types of input forms, five cross-linguistically motivated constraints, and a single constraint ranking. Our account does rely on allomorphy, which we believe must be a part of Russian grammatical competence. Insofar as Russian stress is also controlled by the phonological system, however, using OT tools shows to what extent the Russian system incorporates common cross-linguistic tendencies, the constraints which bring motivation to the alternations.

References

Alderete, J. 1999. *Morphologically Governed Accent in Optimality Theory.* Ph.D. dissertation. University of Massachusetts, Amherst.

Archangeli, D. and D.T. Langendoen, (eds.) 1997. *Optimality Theory: An Overview*. Malden, MA: Blackwell.

Avanesov, R.I. 1956. *Fonetika Sovremennogo Russkogo Jazka.* Moscow: Izdatelstvo Moskovskogo Universiteta.

Beckman, J. 1997. *Positional Faithfulness.* Ph.D. dissertation. University of Massachusetts, Amherst.

Brown, D., G. Corbett, A. Hippisley, N. Fraser, and A. Timberlake. 1996. "Russian Noun Stress and Network Morphology". *Linguistics* 34:1, 1-55.

Coats, H. S. 1976. *Stress Assignment in Russian*. Edmonton, Canada: Linguistic Research Inc. Vol 1. Inflection

Halle, M. 1973. "The Accentuation of Russian Words". *Language* 49.2, 312-348.

Halle, M. and W. Idsardi. 1995. "General Properties of Stress and Metrical Structure". In J. Goldsmith (ed.) *Handbook of Phonological Theory*. 403-443. Cambridge, MA: Blackwell.

Halle, M. and J.-R. Vergnaud. 1987. *An Essay on Stress.* Cambridge, MA: MIT Press.

Hayes, B. 1995. *Metrical Stress Theory*. Chicago: The University of Chicago Press.
Kager, R. 1994. "Generalized Alignment and Morphological Parsing". Rutgers Optimality Archive: ROA-36-1094.
Kager, R. 1995. "The Metrical Theory of Word Stress". In J. Goldsmith (ed.) *Handbook of Phonological Theory*. 367-402. Cambridge, MA: Blackwell.
Kiparsky, P. and M. Halle. 1977. "Towards a Reconstruction of the Indo-European accent". In L. Hyman (ed.) *Studies in Stress and Accent*. Southern California Occasional Papers in Linguistics. CA: USC. Vol 4, 209-238.
McCarthy, J. and A. Prince. 1993. *Generalized Alignment*. Technical Report No. 7, New Brunswick, NJ: RuCCS.
McCarthy, J. and A. Prince. 1995. "Faithfulness and Reduplicative Identity". In J. Beckman, L. Dickey and S. Urbanczyk (eds.) *Papers in OT*. UMOP 18. UMass, Amherst: GLSA. 249-384.
Melvold, J. 1990. *Structure and Stress in the Phonology of Russian*. Ph.D. dissertation. MIT.
Prince, A. and P. Smolensky. 1993. *Optimality Theory*. Technical Report Number 2. New Brunswick, NJ: RuCCS.
Smith, J. 1998. "Noun Faithfulness". Rutgers Optimality Archive: ROA-242.
Sullivan, W. 1969. *A Stratificational Description of the Phonology and Inflectional Morphology of Russian*. Ph.D. dissertation. Yale University.
Trubetzkoy, N.S. 1971. *Principles of Phonology*. Berkeley and LA: University of California Press.
Zaliznjak, A.A. 1967. *Russkoe Imennoe Slovoizmenenie*. Moscow: Russkij Jazyk.

Program in Linguistics
University of Florida
Box 115454
Gainesville, FL 32611-5454
wiltshir@lin.ufl.edu

Agreement and Case-Marking in Russian: A Psycholinguistic Investigation of Agreement Errors in Production

Janet Nicol and Rachel Wilson
University of Arizona

1. Introduction

A central question in psycholinguistic research is how human beings produce well-formed sentences in a timely fashion. On the surface, this would seem to be a relatively straightforward process: a speaker selects the words that convey particular concepts, arranges them in a syntactically legitimate order, and then pronounces each in turn. This process is complicated, however, by the fact that in most languages, there are discontinuous dependencies, so that the form of certain words depends critically on the nature of prior words in the utterance. This aspect of sentence production requires that speakers keep in mind (at the very least) a number of pieces of a sentence at once.

An example of this type of dependency is subject-verb agreement. In English, for example, the verb is required to agree in number with the subject NP, or more specifically, the head of the subject NP. This head and verb may be contiguous, as in *The **bill was** late*, or the subject NP may be complex, containing intervening phrases, as in *The **bill** for the supplies to the schools **was** late*. One might imagine that in cases where the head NP is separated from the verb by intervening phrases, errors in subject-verb agreement (**The bill for the supplies were late*) would sometimes occur. This is indeed the case; such errors occur in spontaneously produced utterances, in written language, and in the laboratory (e.g., Bock and Miller 1991). Within the psycholinguistics literature, speech errors are assumed to be revealing about the nature of the production system (for example, with respect to stages of processing and interactions among stages). Errors in verb agreement are the focus of this paper.

This paper is organized as follows. In Section 2, we describe the experimental findings. We then connect these findings to a current model of language production. Then we describe a recent study of English which raises a question about the role of case information. In Section 3, we present the methodological details and results of the experiment that we conducted in Russian. This experiment was conducted in order to further explore the issue raised by the results of the English study. In the remaining sections, the results of the Russian and English studies are discussed.

2. Background

2.1. Experimental Results: Factors which Influence Agreement Errors

Bock and Miller (1991) first induced errors in the laboratory by presenting speakers with sentence preambles which were typically complex NPs in which NP number was varied, as in (1), and having the subjects repeat the preamble and complete the sentence.

(1) a. The bill to the school...
b. The bill to the schools...
c. The bills to the schools...
d. The bills to the school...

In experiments of this type, errors do occur. Specifically, they occur (a) when there is a mismatch in number between the two nouns in the sentence, as in (1b,d), and (b) when the head NP is singular and the NP within the modifier is plural, as in (1b). (Bock and Miller 1991; Bock and Cutting 1992; Bock and Eberhard 1993; Bock 1995; Eberhard 1997; Nicol 1995; Vigliocco, Butterworth and Garrett 1996; Vigliocco and Nicol 1998). This pattern is mirrored in spontaneously occurring errors (Bock and Miller 1991). The fact that errors do not arise when the nouns match in number (e.g., *The bill to the school...*) suggests that the mismatching nonhead NP causes interference. The fact that errors

arise only when the head is singular suggests that errors are not simply due to "local agreement", in which the speaker forgets the head and simply makes the verb agree with the most recently occurring NP.[1] The number asymmetry (errors only when the head is singular) has been attributed to the markedness of the plural (Bock and Eberhard 1993). Note that in English in particular, the "conceptual markedness" of the plural is mirrored in the morphology: a plural morpheme is "appended" to a stem. It could be argued that the default or base form is singular, and the derived (and marked) form is plural, and that with respect to subject-verb agreement, only marked forms cause interference.

Further experimentation has suggested that the ordering of NPs within the preamble matters. For example, Nicol, Barker and Vigliocco (in preparation) have found that more errors occur when the interfering plural NP is nearer the head, as in (2a) than when it is nearer the verb, as in (2b). This finding suggests that subject-verb agreement errors occur at a stage of processing at which the syntactic structure of the preamble is represented.

(2) a. The statue which stood in the gardens near the mansion...
b. The statue which stood in the garden near the mansions...

As a final point, note that the agreement process is apparently immune to phonological effects. Using a variant of the experimental procedures, in which preambles were presented visually, Bock and Eberhard (1993) explored whether or not nonheads that sounded plural (e.g., "rose") interfered with agreement in the same way as true plurals. They found that their pseudo-plurals gave rise to no errors whatsoever.

---

[1] Note that the errors that are considered for analysis are produced in the context of a fully correct production of the preamble; if a speaker were to completely forget the head NP, then she would produce an incorrect repetition of the preamble (an error of a different sort, and not included in the tally of "pure" agreement errors).

2.2. Models of Language Production

Models of language production typically hypothesize a number of discrete stages of production. For example, models by Bock and Levelt (1994) and Garrett (1990) include an early stage of processing at which predicate-argument structure is computed, the *Functional Level*. This is followed by a stage at which a syntactic structure is represented, the *Positional Level*.

Lexical information is also partitioned. The semantic and syntactic information associated with lexical items is represented at the Functional level. Lexical items are referred to here as *lemmas*. Lemmas that best represent the concepts to be conveyed are selected. The selected lemmas then interact with the predicate-argument processes. The phonological forms associated with the lemmas, *lexemes*, are retrieved at a later stage of production.

Note that it is possible in principle for agreement to be specified at the Functional level: Nouns would be specified for number during a process in which the speaker's "message" (a nonlinguistic proposition) is mapped to lexical entries and annotated with the number specification appropriate to the message. At the point where predicate-argument relations are computed, there could be a mechanism which locates the NP flagged as the "head of subject" and copies that number specification to the verb. However, the findings by Nicol, Barker and Vigliocco (in preparation) show that, at least for English, agreement is computed at a stage of production where constituents appear in a linear order, i.e., within a syntactic representation. This suggests that agreement is computed at the positional level, conceivably at a point in processing before the phonological forms of lexical items are retrieved.

Assuming that agreement is basically syntactic, one mechanism for agreement is one that copies the plural feature (if there is one) from the head NP to the verb. An error arises if the plural feature from the wrong noun (the nonhead) is copied.

2.3. Pronominal Nonheads and Error Production

Recent work by Nicol, Antón-Méndez and Wilson (in Nicol, Antón-Méndez and Wilson, in preparation) suggests further constraints on the type of representation over which agreement is computed. In one study, they contrasted pronominal nonheads with lexically specified nonheads, as in example (3):

(3) a. The bill from the accountant...
b. The bill from him...
c. The bill from the accountants...
d. The bill from them...

They observed errors in both conditions (both significantly more than the control conditions in which the two NPs matched in number) but they found significantly more errors in the condition with the lexically specified heads than in the pronoun condition. There are a number of possible explanations for this difference. One is that plurals that are marked morphologically are more likely to cause interference (note that lexically specified nouns but not pronouns bear an affix that marks number). However, this is at odds with results by Bock and Eberhard (1993) who found that irregular plurals (e.g., *mice*) elicited just as many errors as regular plurals (e.g., *rats*). Another explanation is that case marking a nonhead (pronouns but not lexically specified nouns) reduces interference. A third possibility is that pronouns have some other attribute (apart from case-marking) that interacts with error production.

The experiment reported here is a preliminary investigation of the second possibility, that case-marking of a nonhead reduces its ability to influence the agreement process. We used Russian because all nouns and pronouns are case-marked. This allows us to determine whether or not case-marking is a critical factor in the occurrence of errors.

3. Agreement and Case-Marking in Russian: A Production Study

3.1. Method

Subjects. Thirty-two native Russian speakers participated in this experiment. The subjects were acquaintances and relatives of acquaintances who are residents of St. Petersburg and Moscow, Russia. Subjects ranged in age from 12 to 65.[2] Each was paid the equivalent of $5 for participating.

Design and Materials. Thirty-two stimulus quadruplets were created. Each stimulus item consisted of a simple noun phrase, followed by a complex phrase consisting of a singular subject noun followed by a prepositional phrase modifier. This modifier contained a NP that was either singular (S) or plural (P) and contained either a common noun or pronoun. The head NP was always singular. Hence, there were two factors, Nonhead Number and Nonhead Type. The initial simple noun phrase was provided to serve as a referent for the pronouns. An example quadruplet appears in Table 1.[3]

Table 1: An Example of the Experimental Quadruplet

	Singular	Plural
Noun	Buxgalter. Sčet ot buxgaltera... 'Accountant. Bill from accountant...'	Buxgaltery. Sčet ot buxgalterov... 'Accountants. Bill from accountants...'
Pronoun	Buxgalter. Sčet ot nego... 'Accountant. Bill from him...'	Buxgaltery. Sčet ot nix... 'Accountants. Bill from them...'

[2] There were three subjects under 18 years of age and three over 55 years of age. These subjects performed similarly to the other subjects in this study.

[3] In Russian, different prepositions require different case-endings on the nouns that follow them. Case and number information are included in one morpheme. Thus, the *-ov* ending on *buxgalterov* indicates both plurality and Genitive case.

The example in Table 1 contains a preposition which requires the Genitive case, but prepositions which require the dative and instrumental were also included. Out of 32 experimental quadruplets, seven contained the preposition *ot* 'from', three contained the preposition *dlja* 'for', one contained the preposition *iz* 'from' (all of which required the Genitive case), two contained the preposition *po* 'along', eight contained the preposition *k* 'towards', (both requiring the Dative case) and four contained the preposition *s* 'with', four contained the preposition *za* 'behind', one contained the preposition *pod* 'under', two contained the preposition *pered* 'in front of' (all of which require the Instrumental case). The experimental items contained an equal number of masculine and feminine items in all four possible combinations: masculine-masculine, masculine-feminine, feminine-masculine, and feminine-feminine.

In addition, 36 filler items and 12 practice items were interspersed with the 32 experimental trials, resulting in a total of 80 trials. Experimental and filler items were assembled to form four versions of the experiment in a fully counterbalanced design. The items in each list were presented in a random order (the same order for each list). The four lists were tape-recorded by a female native Russian speaker.

Procedure. Subjects were tested individually. They heard each preamble, repeated it and finished the sentence. Responses were recorded onto audiotape.

Data Treatment. Sentence repetitions and continuations were categorized as follows (response codes appear in parentheses). *Correct (C)*: These consist of a grammatical sentence of Russian; *Agreement Errors (AGR)*: These consist of correctly uttered preambles with a sensible continuation and an agreement error on the verb. *Other Errors (O):* These consist of correctly uttered preambles with a sensible continuation and an error *other* than an agreement error on the verb; *Repetition Errors (Rep.):* These consist of incorrectly uttered preambles; *Head Misselection Error (Head)*: These consist of correctly uttered preambles but where the

predicate provided in the response clearly indicated that the participant was using the object of the prepositional phrase as the intended referent; *Miscellaneous (Misc.)*: These consist of responses which did not constitute a sentence of Russian.

Results. The distribution of responses is displayed in Table 2.

Table 2. Number of Responses in Each Response Category for Each Condition

	C	AGR	O	Rep.	Head	Misc.
Pronoun-S	186	4	2	8	0	56
Pronoun-P	164	12	3	27	1	49
Noun-S	195	1	1	7	0	52
Noun-P	166	10	1	24	1	54

As is clear from the table, more agreement errors were associated with the plural nonhead conditions. Analysis of variance shows this effect to be significant by subjects (F1) and marginally significant by items (F2): $F1(1,32) = 8.76$, $p < .006$; $F2(1,32) = 3.12$, $p < .087$. In other words, reliably more errors occurred with the plural Nonheads than singular Nonheads, but the type of Nonhead (pronoun vs. NP) had no effect on error rate. Neither the Nonhead Type (pronoun vs. noun) nor the interaction of Nonhead Type and Nonhead Number were significant.

The marginal effect for items (reported just above) suggests that the effect of the plural nonhead (relative to the singular nonhead) varies across items. A subanalysis of the items which grouped them according to the gender of the two NPs shows that there are indeed significant differences within the item set. The mean percentage of errors for each gender combination is shown in Table 3.

Table 3. Percentage of Errors for the Four Different Gender Combinations for Plural Nonheads.

	Mean proportions
Masculine-Masculine	0.005
Masculine-Feminine	0.016
Feminine-Feminine	0.106
Feminine-Masculine	0.006

4. Discussion

The results of this experiment show that plural nonhead pronouns elicit as many errors as lexically-specified nouns, as long as the nouns are also case-marked. It is worth noting that the error percentages in this study are roughly equal to the error percentages for pronouns in the English study (in English, error rates are .06 for pronouns and .15 for nouns). This strongly suggests that case-marking is at the root of the difference between the pronoun and noun conditions in the English study: case-marking of a nonhead reduces the extent to which it may interfere with agreement.

This view is compatible with some recent findings reported by Hartsuiker, Antón-Méndez, and van Zee (submitted). They conducted a study in Dutch which also contrasted pronoun and noun nonheads. Their pronouns were either case-marked (animate) pronouns or non-case-marked (or case-ambiguous) (inanimate) pronouns. Lexically-specified nouns are not case-marked, just as in English. They found equal numbers of errors for the non-case-marked pronouns and their lexically-specified noun counterparts. In the condition in which nouns were compared to case-marked pronouns, there was a significant difference in the error rate, with more errors in the noun condition than the pronoun condition. Again, case-marking appears to reduce interference from the nonhead plural NP.

Let us now consider in some detail how case information might come into play during language production. The process by which it is determined which roles the different NP arguments play within

a clause could make reference to role information such as "subject" (or "head of subject") and "object" or to case information such as "Nominative" and "Accusative". We will assume that both types of notions are encoded at the Functional level, and that this joint information is used to map NPs into particular syntactic positions, as shown in Figure 1.

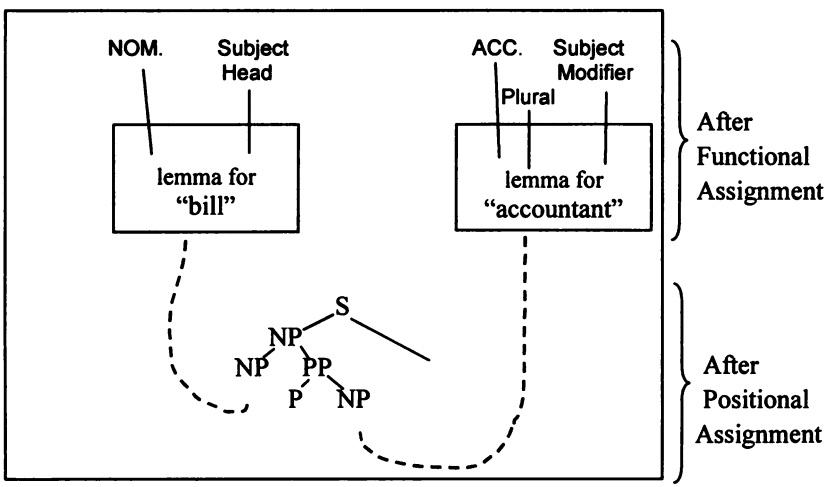


Figure 1. Functional Level-Positional Level mapping of arguments, for a sentence with the complex subject NP *The bill from the accountants...*

We assume that the case information is carried through to the Positional level of representation, and that this case information interacts with the process of phonological spell out. Hence, if the NP within the subject modifier is marked with Genitive case, then the Genitive ending will be produced. We also assume that subject-verb agreement is carried out at the positional level of representation, possibly via a mechanism that copies the number feature from the subject head to the verb.

For the English examples in which the NP is lexically-specified (*The bill to the accountants...*), there would be no case

information. Errors would arise as outlined above: the plural feature from the wrong noun is copied to the verb.

For the Russian examples in which the NP is lexically specified, case information would be represented. When the nonhead is lexically specified, the case information somehow reduces the chance that the plural feature from the wrong noun is copied to the verb. One way this might work is that the presence of additional information about the distinct roles of the nouns reduces the possibility of confusion. In Russian sentences, both the structural information and case information keep the head NP and nonhead distinct. In English, there is only structural information.

Now let us consider the pronoun examples. In English, first and third person pronouns are case-marked, so this information would be available first at the Functional level of representation and then at the Positional level. The combination of structural and case information would help to keep the nouns distinct, and so errors would be reduced. In Russian, the pronoun examples work exactly the same way as the lexically-specified nouns: case information is available and helps to keep the two NPs distinct.

What we are claiming is that case information is represented during syntactic encoding only when it is reflected in the *form* of the word. This presents a puzzle: in the model sketched out in the introduction (Section 2.2), syntactic encoding involved lemmas, not lexemes, so word form is not even available when agreement is carried out. How, then, can word form have an effect on the agreement process? We suggest that the lemma representations of nouns and pronouns carry information about case-markability. In English, the lemmas for first and third person pronouns are so marked; other pronouns and nouns are not. Presumably, lemma representations are developed through comprehension; during language acquisition, children learn that first and third person pronouns have two different forms depending on their positions and roles in a sentence. This information could certainly be part of the lemma representation, even though the actual word forms are not represented at this level.

There is one result that we cannot account for. As Table 3 shows, our Russian study showed that feminine nouns and pronouns were significantly more likely to elicit errors than masculine nouns and pronouns, with the most errors occurring in the feminine-feminine condition (11% vs. less than 2% for the other three conditions). We considered whether the cause of the difference might be due to homophony in case-endings in the feminine in Russian. For example, there is an overlap in case-marking for the nominative plural (-y) and the genitive singular (-y). It is then conceivable that a noun which is in fact singular (genitive singular) but sounds like a plural (nominative plural) could induce errors. However, if homophony were in fact the cause of the gender effect, then we would expect to see relatively more errors in the *SS condition* (since this is where the genitive singular appears) or relatively *fewer* errors in the *SP* condition. As is evident from the table above, there were simply not enough errors in the SS condition for this to be the root of the gender effect. Further research is necessary to illuminate this finding.

5. Conclusions

The research reported here on English and Russian showed that the occurrence of subject-verb agreement errors in spoken sentence production increases with (a) the presence of a plural nonhead NP within the subject NP, and (b) the absence of case-marking on the NPs within the subject NP. We assume that agreement is carried out via feature-copying, and that if the head noun is plural, the plural feature is copied to the verb. An error occurs when a plural feature from a nonhead NP is erroneously copied to the verb. The possibility of such an error is reduced by the presence of case information, which presumably provides the system with additional information for distinguishing the NPs within the complex subject.

This work adds to the body of research which seeks to elaborate the architecture of the production system. Our research shows that case information must be represented at the level of processing at which agreement is computed. Other research

suggests that this level is one in which the syntactic structure is represented.

References

Bock, J.K. 1995. "Producing Agreement". *Current Directions in Psychological Science* 4, 56-61.

Bock, J.K. and J.C. Cutting, 1992. "Regulating Mental Energy: Performance Units in Language Production". *Journal of Memory and Language* 31, 99-127.

Bock, J.K. and K.M. Eberhard, 1993. "Meaning Sound and Syntax in English Number Agreement". *Language and Cognitive Processes* 8, 57-99.

Bock, J.K. and W.J.M. Levelt, 1994. "Language Production. Grammatical Encoding". In M.A. Gernsbacher (Ed.) *Handbook of Psycholinguistics*. San Diego: Academic Press.

Bock, J.K. and C.A. Miller, 1991. "Broken Agreement". *Cognitive Psychology* 23, 35-43.

Eberhard, K. 1997. "The Marked Effect of Number on Subject-Verb Agreement". *Journal of Memory and Language* 36, 147-164.

Garrett, M.F. 1990. "Sentence Processing". In D.N. Osherson and H. Lasnik (eds.), *Language: An Invitation to Cognitive Science, Volume 1*. Cambridge MA: MIT Press.

Hartsuiker, R.J., I. Antón-Méndez,, and M. van Zee, (submitted) Object attraction in subject-verb agreement construction.

Nicol, J. 1995. "Effects of Clausal Structure on Subject-Verb Agreement Errors". *Journal of Psycholinguistic Research* 24, 507-516.

Nicol, J., J. Barker, and G. Vigliocco, (in preparation) "The Effect of Syntactic Structure on Subject-Verb Agreement Errors". Manuscript, University of Arizona.

Nicol, J., I. Antón-Méndez and R. Wilson, (in preparation) "Case Marking in Language Production". Manuscript, University of Arizona.

Vigliocco, G, B. Butterworth, and M.F. Garrett, 1996. "Subject-Verb Agreement in Spanish and English: Differences in the Role of Conceptual Constraints". *Cognition* 61, 261-298.
Vigliocco, G. & J. Nicol, 1998 "Separating Hierarchical Relations and Word Order in Language Production: Is Proximity Concord Syntactic or Linear?" *Cognition* 68, 13-29.

Psychology Department
PO Box 210068
Tucson, AZ 85721-0068.

Monotonicity-Based vs. Veridicality-Based Approaches to Negative Polarity: Evidence from Russian*

Asya Pereltsvaig
McGill University

1. Introduction

Since Ladusaw's (1980) breakthrough work, it has been widely assumed that downward monotonicity[1] is the property responsible for licensing negative polarity items (NPIs) such as English *any* in *Jeff doesn't watch any sitcoms*. However, in recent years it has been argued that a more general property of (non)veridicality is the one that properly describes the distribution of polarity sensitive items (PSIs) (cf. Zwarts 1995, Giannakidou 1997, 1998, Peres 1998). In this paper, I investigate the distribution of PSIs in Russian and compare the monotonicity-based approach (MBA) with the veridicality-based approach (VBA) to polarity sensitivity as a tool to account for Russian data.

Two main claims are made in the paper. Empirically, I argue against the claims made by Haspelmath (1997) with respect to the distribution of *libo*-items in Russian. From the theoretical point of view, I claim that the MBA works better in describing the distribution of Russian PSIs than the VBA, because the latter makes wrong predictions for licensing both *ni-* and *libo*-items.

---

* In working on this paper I have benefited from discussions with Joanna Blaszczak, Jonathan Bobaljik, Anastasia Giannakidou and Shravan Vasishth. I am grateful to the participants in the Syntax Project at McGill and the audiences at the Negation conference in Salford and FASL-8 for their comments and to Lotus Goldberg for proofreading the paper. This research was partially funded by the FCAR project #97 ER 0578 and by the Hydro-Quebec McGill Major Fellowship. The usual disclaimers apply.

[1] Different terms have been used for this property, such as 'downward entailingness', and 'monotone decreasingness'.

2. Data[2]

In this paper, I examine three sets of PSIs in Russian: *ni*-items, *libo*-items, and *nibud'*-items, listed in Table 1. For reasons of space, I will not discuss *to*-items in this paper. These items are positive polarity items, marked for wide-scope interpretation, and are grammatical only in contexts where they can receive such wide-scope interpretation.

Table 1. PSIs in Russian (cf. Bernini and Ramat 1996:152-153)

	ni-series	*libo*-series	*nibud'*-series
anyone	nikto	kto-libo	kto-nibud'
anything	ničto	čto-libo	čto-nibud'
anywhere	nigde	gde-libo	gde-nibud'
to anywhere	nikuda	kuda-libo	kuda-nibud'
from anywhere	niotkuda	otkuda-libo	otkuda-nibud'
ever	nikogda	kogda-libo	kogda-nibud'
anyhow	nikak	kak-libo	kak-nibud'
any amount	niskol'ko	*	skol'ko-nibud'
anyone's	ničej	čej-libo	čej-nibud'
any (Adj.)	nikakoj	kakoj-libo	kakoj-nibud'
preliminary status	strong NPIs[3]	weak NPIs	narrow-scope non-specific PSIs

Unlike English *any*-NPIs, Russian *ni*-items require clausemate sentential negation, as in (1a). In particular, *ni*-items cannot appear

[2] The data in this paper are based on a survey of 34 native speakers of Russian.

[3] For arguments that *ni*-items are NPIs rather than negative quantifiers, see Blaszczak (1998) and Pereltsvaig (1998).

in positive (i.e., non-downward entailing) contexts, as illustrated in (1b). Nor can they appear in so-called weak negative contexts, such as distant negation, conditionals, nuclear scope of *only* and the second argument of *few*, as shown in (1c-f); for more discussion of the syntax of *ni*-items see Brown (1999).

(1) a. CLAUSEMATE SENTENTIAL NEGATION
Ko mne na den' roždenia nikto ne prixodil.
to me for day birth.GEN *ni*-who not came.IMPRF
'Nobody would come to my birthday.' (from a children's song)

b. POSITIVE CONTEXT
*Ja videl nikogo.
I saw *ni*-who
'I saw no one.' (Brown 1999:24, (14))

c. DISTANT NEGATION

*Ja ne dumaju, čto Adam čital nikakoj
I not think that Adam read.PAST *ni*-which.ACC
žurnal.
journal.ACC
'I don't think that Adam read any journal.'

d. CONDITIONALS

*Esli vy vstretite nikogo, pozvonite mne.
if you meet *ni*-who.ACC call me
'If you meet anyone, call me.'

e. NUCLEAR SCOPE OF *ONLY*

*Tol'ko Adam čital nikakoj žurnal.
only Adam read.PAST *ni*-which.ACC journal.ACC
'Only Adam has read any journal.'

f. SCOPE OF *FEW*

*Nemnogie studenty čitali nikakoj .
few students read.PAST *ni*-which.ACC
žurnal.
journal.ACC
'Few students read any journal.'

Libo-items cannot appear in positive non-NPI contexts either, as illustrated in (2).[4]

(2) *Ja videl čto-libo.
I saw what-*libo*
'I saw something/nothing.'

The only contexts where *libo*-items are allowed are the so-called weak negative contexts, illustrated in (3): distant negation, condi-tionals, scope of *only*, second argument of *few*, interrogatives, com-plements of negative predicates, such as *doubt*, complements of *without*, *too*-constructions, as well as both sentential and phrasal comparatives. (*By*, glossed as 'CCP', is a counterfactual conditional particle.)

(3) a. DISTANT NEGATION

Ja ne dumaju, čto Adam čital kakoj-libo
I not think that Adam read.PAST which-*libo*.ACC
žurnal.
journal.ACC
'I don't think that Adam read any journal.'

b. NON-COUNTERFACTUAL CONDITIONAL

Esli vy kogo-libo vstretite, pozvonite mne.
if you who-*libo*.ACC meet call me
'If you meet anyone, call me.'

[4] *Libo*-items are associated with higher styles, and are therefore quite rare in spoken languages and journalistic writings.

c. COUNTERFACTUAL CONDITIONAL

Esli by kto-libo priexal, my ustroili by večerinku.
if CCP who-*libo* arrived we arrange CCP party
'If anyone came, we'd arrange a party.'

d. SCOPE OF *ONLY*

Tol'ko Adam čital kakoj-libo žurnal.
only Adam read.PAST which-*libo*.ACC journal.ACC
'Only Adam has read any journal.'

e. SCOPE OF *FEW*

Nemnogie studenty čitali kakoj-libo žurnal.
few students read.PAST which-*libo*.ACC journal.ACC
'Few students read any journal.'

f. INTERROGATIVE (with either information or rhetorical reading)

Vy čitali kakoj-libo žurnal?
you read.PAST which-*libo*.ACC journal.ACC
'Have you read any journal?'

g. COMPLEMENT OF *DOUBT*

Ja somnevajus', čto kto-libo prijdet.
I doubt that who-*libo* come.FUT
'I doubt that anybody will come.'

h. COMPLEMENT OF *WITHOUT*

Vy spravites' bez kakoj-libo pomošči.
You manage.FUT without which-*libo* help
'You'll manage without any help.'

i. *TOO*-CONSTRUCTIONS

Adam sliškom ustal, čtoby čitat' kakoj-libo
Adam too got-tired in-order-to to-read which-*libo*
žurnal.
journal
'Adam is too tired to read any journal.'

j. SENTENTIAL COMPARATIVE

Ona umnee čem kto-libo predpolagal.
she smarter than who-*libo* supposed
'She is smarter than anyone supposed.'

k. PHRASAL COMPARATIVE

Ona umnee čem kto-libo v ee klasse.
she smarter than who-*libo* in her class
'She is smarter than anyone in her class.'

Contrary to the claims made by Haspelmath (1997:274), *libo*-items cannot appear in irrealis non-specific contexts, such as imperatives (4a) and futures (4b).[5]

(4) a. IMPERATIVES
*Spojte nam kakuju-libo pesnju.
sing.IMPER to-us which-*libo* song
'Please sing us any/some song.'

b. FUTURE

*My vstretimsja gde-libo.
we will.meet where-*libo*
'We'll meet anywhere/somewhere.'

Likewise, *libo*-items are not licensed by modal verbs, as shown in (5).

(5) a. *Vy možete vzjat' kakuju-libo knigu.
you may take which-*libo* book
'You may take any/some book..'
b. *Kto-libo dolžen rešit' ètu zadaču.
who-*libo* must solve this problem
'Anyone/someone must solve this problem.'

[5] According to all of the speakers consulted, sentences such as (4) and (5) are interpretable but sound non-native.

Thus, *libo*-items seem to be weak NPIs; however, they are not licensed in strong polarity contexts (that is, clausemate sentential negation), as shown in (6).

(6) *On kogo-libo ne vstretil.
he who-*libo*.ACC not met
'He didn't meet anyone.'

The third set of items investigated in this paper are *nibud'*-items. These are licensed freely in irrealis non-specific contexts, such as imperatives (7a), futures (7b), and modals (7c). Note that, in examples like (7c), *nibud'*-items take narrow scope with respect to modal verbs.

(7) a. IMPERATIVES
Spojte nam kakuju-nibud' pesnju.
sing.IMPER to-us which-*nibud'* song
'Please sing us a song.'

b. FUTURE

My vstretimsja gde-nibud'.
we will.meet where-*nibud'*
'We'll meet somewhere.'

c. MODAL
Vy možete vzjat' kakuju-nibud' knigu.
you may take which-*nibud'* book
'You may take a/any book.'

Nibud'-items are also licensed in some weak negative contexts, such as the second argument of *few*, conditionals, interrogatives, and complements of negative predicates, though there is a fair amount of inter-speaker variation, as well as variation related to particular examples. Furthermore, *nibud'*-items are not licensed by clausemate sentential negation, as shown in (8).

(8) *Ona ne videla kogo-nibud' utrom.
she not saw who-*nibud'* morning.INSTR
'She saw someone in the morning.'

The distribution of the three series of PSIs is summarized below.

Table 2. Distribution of Russian PSIs

context	*ni*-items	*libo*-items	*nibud'*-items
clausemate sentential negation	OK	*	*
weak negative contexts (conditional, interrogatives, comparatives, etc.)	*	OK	? (subject to variation)
irrealis non-specific contexts (future, imperatives, modals)	*	*	OK

3. Comparison of the Theories

3.1. The Monotonicity-based Approach

This approach to polarity sensitivity originates with Ladusaw (1980), who proposed that the class of licensers for NPIs should be defined as the class of expressions whose meanings are "downward entailing" (or downward monotone). Since then, the notion of monotonicity has constituted a basis for most of the research on polarity sensitivity, even though several problems have been identified with this approach (for discussion, see Heim 1984, Linebarger 1987, Atlas 1996, 1998, and Horn 1996, among others). For the sake of space, I will not attempt a detailed description of the technicalities of this approach here, but will give only a very concise summary of its main ideas.[6]

[6] For a good exposition and a discussion of the MBA as it applies to English and Dutch PSIs, see van der Wouden (1997).

Monotonicity is defined as preserving entailments either from supersets to subsets (downward monotonicity), or from subsets to supersets (upward monotonicity). Thus, a context can be either downward monotone, upward monotone, or non-monotone. Formal definitions of downward and upward monotonicity from van der Wouden (1997:95) are given in (9).

(9) a. Let B and B^* be two Boolean algebras. A function f from B to B^* is *monotone decreasing* [downward monotone] iff for arbitrary elements $X, Y \in B$: $X \subseteq Y \rightarrow f(Y) \subseteq f(X)$.

b. Let B and B^* be two Boolean algebras. A function f from B to B^* is *monotone increasing* [upward monotone] iff for arbitrary elements $X, Y \in B$: $X \subseteq Y \rightarrow f(X) \subseteq f(Y)$.

For example, noun phrases of the form *few N* are downward monotone, noun phrases of the form *many N* are upward monotone, and noun phrases of the form *exactly n N* are non-monotone. Readers are invited to check the entailment patterns for themselves.[7]

NPIs are said to be licensed by downward monotone contexts, as shown in (10).[8]

(10) a. Few students saw anything.
b. * Many students saw anything.
c. * Exactly three students saw anything.

Further distinctions between different classes of NPIs are made on the basis of additional monotonicity-related properties, such as anti-additivity and anti-morphicity. Anti-additivity is not particularly relevant for this paper, so I will not discuss it in any detail here. Antimorphic contexts (functions) are defined by the equations in (11) below.

(11) $f(X \cap Y) = f(X) \cup f(Y)$
$f(X \cup Y) = f(X) \cap f(Y)$

---

[7] For more discussions and examples, see van der Wouden (1997).

[8] For a discussion of examples such as the one in (10c), see Giannakidou (1998:8-11).

Examples of antimorphic contexts are clausemate sentential negation, noun phrases of the form *not definite NP/proper name* (e.g., *not the teacher*, *not Peter*) and complements of the preposition *without*.

3.2. The Veridicality-based Approach

This approach to polarity sensitivity originates with the work done in the Netherlands, such as Zwarts (1995) and Giannakidou (1997, 1998). Their main claim is that different sensitivities, including negative polarity, free choice polarity, and the choice of mood in relative clauses, are instances of a general sensitivity to (non)-veridicality. A definition of (non)veridicality from Giannakidou (1998:106) is given in (12) below.[9]

(12) Let *Op* be a monadic propositional operator. The following statements hold:
(i) *Op* is *veridical* just in case *Op p* → *p* is logically valid. Otherwise, *Op* is *nonveridical*.
(ii) A nonveridical operator *Op* is *antiveridical* just in case *Op p* → ¬ *p* is logically valid.

As illustrated in (13), *yesterday* is a veridical operator, *perhaps* is a nonveridical operator, and sentential negation *not* is an antiveridical operator.

(13) a. Dharma kissed Greg yesterday. → Dharma kissed Greg.
b. Perhaps Dharma kissed Greg. –/→ Dharma kissed Greg.
c. Dharma didn't kiss Greg. → It is not the case that Dharma kissed Greg.

According to the VBA, different PSIs are either licensed or anti-licensed by veridicality, nonveridicality, or antiveridicality. For instance, emphatic items in Greek are licensed by antiveridicality, whereas *any* in English is said to be anti-licensed by veridicality.

---

[9] Giannakidou (1998:110-117) further relativized the notion of (non)veridicality to models. However, this revision need not concern us here.

(14) a. *Dharma kissed anyone yesterday.
b. *Perhaps Dharma kissed anyone.
c. Dharma didn't kiss anyone.

Note that being anti-licensed by veridicality does not entail being licensed in all non-veridical contexts. Thus, *any*, anti-licensed by veridicality, is not licensed in the contexts of *perhaps*, a non-veridical operator. Next I will show that (non)veridicality alone cannot account for the distribution of Russian PSIs.

3.3. Russian PSIs: Monotonicity or Veridicality?

In this section, I investigate which of the two approaches to polarity sensitivity – monotonicity-based or veridicality-based – can better account for Russian PSIs.

Let's begin with *ni*-items. As has been shown in Section 2, these items have a peculiar restriction to clausemate sentential negation contexts. Yet, clausemate sentential negation cannot be described purely in semantic terms in either the MBA or the VBA. Thus, in terms of the MBA, clausemate sentential negation is an antimorphic context, but other contexts, such as noun phrases of the form *not definite NP/proper name* and complements of the preposition *without*, are also antimorphic. In the VBA, clausemate sentential negation is an antiveridical operator, but again, it is not the only antiveridical operator. Other antiveridical contexts include complements of the preposition *without*, rhetorical questions, both sentential and phrasal comparatives, the second argument of quantifiers such as *few*, and counterfactual conditionals, all of which do not license *ni*-items, as shown in (15).

(15) a. COMPLEMENT OF *WITHOUT*
*On spravitsja bez nikakoj pomošči.
he will-manage without *ni*-which help
'He will manage without any help.'

b. RHETORICAL QUESTIONS
*Est' li nikto umnee ee?
is Q *ni*-who smarter her
'Is there anyone smarter than her?'

c. PHRASAL COMPARATIVE
*Ona umnee nikogo / čem nikto.
she smarter *ni*-who.ACC / than *ni*-who
'She is smarter than anyone.'

d. SENTENTIAL COMPARATIVE
*Ona umnee čem nikto mog predpoložit'.
she smarter than *ni*-who could suppose
'She is smarter than anyone would expect.'

e. SECOND ARGUMENT OF *FEW* (=1f)
*Nemnogie studenty čitali nikakoj žurnal.
few students read.PAST *ni*-which.ACC journal.ACC
'Few students read any journal.'

f. COUNTERFACTUAL CONDITIONAL
*Esli by nikto priexal, my ustroili by večerinku.
if CCP *ni*-who arrived we arrange CCP party
'If anyone came, we'd arrange a party.'

Thus, whether one adopts a MBA or a VBA, something special has to be said about *ni*-items. A possible solution in MBA terms has been proposed in Pereltsvaig (1998). According to this proposal, *ni*-items are sensitive not only to the semantic properties of their licensor, but also to its position in the syntactic structure. In particular, *ni*-items require that their licensor be in Polarity Phrase in the complementizer range[10] at LF. The notion that the appropriate configuration should hold at LF is important because I

[10] Following Rizzi (1997), I assume that COMP splits into several distinct functional projections, which I call the "complementizer range". However, I do not follow Rizzi in treating topic and focus as part of that complementizer range. The Polarity Phrase is another functional assumed to be in the complementizer range. Similar proposals have been made by Laka (1990) and McCloskey (1996).

take LF to be the level where polarity sensitivity is licensed in general.[11]

At first glance, it might seem that this proposal can be easily reformulated in the VBA framework. This is not true, however, because several antiveridical contexts that do not license *ni*-items involve an operator in the complementizer range (e.g., rhetorical questions, sentential comparatives and, possibly, counterfactual conditionals). Thus, it is very hard if not impossible to define the right contexts for the *ni*-items within the VBA framework. To sum up so far, the MBA is superior to the VBA in accounting for Russian *ni*-items.

Next, consider *libo*-items. As summarized in Table 3 below, *libo*-items are ungrammatical with clausemate sentential negation, and in imperative, future and modal constructions, as well as in affirmative contexts that do not contain a downward monotone operator (for the data see Section 2 above).

Thus, an analysis in terms of a MBA is readily available: *libo*-items are licensed by a downward monotone operator. This accounts for all the judgments in Table 3 except the star for the clausemate sentential negation contexts. An account of this piece of the data in terms of morphological blocking has been proposed in Pereltsvaig (1998). In brief, *libo*-items are prevented from occurring with clausemate sentential negation by the competition with *ni*-items, which have priority over *libo*-items because they have more highly specified lexical entries (cf. Distributed Morphology framework, Halle and Marantz 1993). *Ni*-items are said to have more highly specified lexical entries for two reasons. First, they can appear in a narrower set of contexts than *libo*-items (antimorphic as opposed to downward monotone); second, *ni*-items but not *libo*-items are specified for the syntactic position of their licenser. Note that this account extends to *nibud'*-items, which are also ungrammatical with clausemate sentential negation.

---

[11] For other proposals that at least some NPI licensing occurs at LF, see Progovac (1988), Mahajan (1990), Hasegawa (1991), Duffield (1993) and Uribe-Etxebarria (1994).

Table 3. Distribution of *Libo*-items in Russian

Context	Monotonicity	Veridicality	*libo*-items
clausemate sentential negation	↓ (AM)	AV	* (6)
without	↓ (AM)	AV	OK (3h)
rhetorical questions	↓	AV	OK (3f)
comparatives	↓	AV	OK (3j,k)
few	↓	AV	OK (3e)
counterfactual conditionals	↓	AV	OK (3c)
non-counterfactual conditionals	↓	NV	OK (3b)
too-constructions	↓	NV	OK (3i)
complements of negative predicates	↓	NV	OK (3g)
imperatives	NM	NV	* (4a)
future	NM	NV	* (4b)
modals	NM	NV	* (5)
episodic past tense affirmative declarative	↑	V	* (2)

Can the VBA account for the distribution of *libo*-items in Russian? The answer is no. It is not possible to say that *libo*-items are licensed by antiveridicality because they can appear in some antiveridical contexts but not in all (e.g., they are banned from clausemate sentential negation contexts). In addition, such a claim would be too strong since *libo*-items are licensed in some contexts

which are not antiveridical, such as non-counterfactual conditionals, *too*-constructions and complements of negative predicates. Further-more, the hypothesis that *libo*-items are licensed in nonveridical contexts fails to account for their ungrammaticality in those nonveridical contexts that happen to be non-downward monotone (e.g., imperatives, futures, and modals).

Giannakidou (1998:95) suggests that Russian *libo*-items are anti-licensed by veridicality. However, her proposal is based on the data from Haspelmath (1997), which I argued to be wrong in Section 2 above.[12] In particular, an account relying on anti-licensing by veridicality fails to explain why *libo*-items are ungrammatical in imperatives, future and modal contexts. Giannakidou's solution is that anti-licensing by veridicality does not entail licensing by non-veridicality. In other words, the 'anti-licensing by veridicality' condition does not imply that the items in question will appear in all contexts which are not veridical. Yet, this seems to be an ad hoc solution since it fails to capture the generalization that the non-veridical contexts that do not license *libo*-items are also non-downward monotone (i.e., either upward monotone or non-mono-tone). In brief, the failure to explain the ungrammaticality of *libo*-items in non-veridical non-downward monotone contexts is a fatal problem for a VBA to polarity sensitivity.

Furthermore, this approach cannot account for the ungrammaticality of *libo*-items with clausemate sentential negation either. Giannakidou (p.c.) rejects the morphological blocking analysis of Pereltsvaig (1998) outlined above. Instead, she proposes to analyze such items as licensed by nonveridicality (or anti-licensed by veridicality; for problems with such solutions see the discussion above) and anti-licensed by antiveridicality (Giannakidou 1998:117, discusses *i*-items in Serbian/Croatian that are similar in distribution to Russian *libo*-items). However, this analysis makes a wrong prediction that *libo*-items should be ungrammatical in all

[12] Martin Haspelmath (p.c.) has admitted that some of his Russian data with respect to *libo*-items are not accurate. This is an unfortunate even if understandable mistake (Haspelmath conducted a survey of 40 languages).

antiveridical contexts. As was mentioned above, *libo*-items are ungrammatical as complements of the preposition *without*, rhetorical questions, comparatives, counter-factual conditionals, and the second argument of *few*. This forces Giannakidou to change her analysis in a later chapter of her book and propose that *i*-items are anti-licensed by negation (p. 162). Of course, this has to be further refined so as not to exclude distant negation from the set of possible licensers of *libo*-items. Yet, this brings us back to the problem of defining the narrow set of contexts, consisting of clausemate sentential negation only, in which *libo*-items are ungrammatical and *ni*-items are grammatical. As has been discussed above, this proves to be no easy task for the VBA.

To sum up, a VBA fails to account for the distribution of *libo*-items in Russian. In particular, it provides no clear explanation for the ungrammaticality of *libo*-items in two sets of contexts: nonveridical non-downward monotone contexts, and clausemate sentential negation.

Lastly, consider *nibud'*-items. To recapitulate, these items appear to be polarity-sensitive since they are not grammatical in episodic past tense affirmative contexts, as shown in (8) above. Moreover, *nibud'*-items are ungrammatical with clausemate sentential negation. The degree of their grammaticality with downward monotone contexts varies: they are fully grammatical in some downward monotone contexts, but not in others. Yet, *nibud'*-items are fully grammatical with non-downward monotone nonveridical contexts, as with imperatives, future, and modals.

Two alternative analyses are possible in terms of a MBA. According to the first analysis, *nibud'*-items are anti-licensed by upward monotonicity (i.e., they are ungrammatical in upward monotone contexts). According to the second analysis, *nibud'*-items are not polarity sensitive items at all. Rather they are sensitive to the presence of an operator with respect to which they can take narrow scope. At present, the second analysis seems to be more promising. First, it allows for a different degree of grammaticality of *nibud'*-items in different downward monotone contexts. Secondly, it accounts for the narrow interpretation that *nibud'*-items obligatorily have. The ungrammaticality of *nibud'*-

items with clausemate sentential negation is explained in terms of morphological blocking, as discussed above for *libo*-items.

A VBA has to rely on anti-licensing by veridicality to describe the distribution of *nibud'*-items. Yet, this account will run into problems with the ungrammaticality of *nibud'*-items with clausemate sentential negation, as well as with the varying degrees of grammaticality in weak negative contexts such as complements of *without* (in which *nibud'*-items are ungrammatical) and counterfactual conditionals (in which *nibud'*-items are grammatical).

At present, both the MBA and the VBA are lacking in the degree of detail in their analysis of *nibud'*-items; therefore, I will not attempt to evaluate the two theories in this respect. More research is needed to fully understand the distribution of *nibud'*-items.

4. Conclusion

In this paper, I have examined the distribution of three sets of PSIs in Russian: *ni-*, *libo-* and *nibud'*-items. I have shown that even though both the MBA and the VBA have problems of their own, the MBA is superior in describing the distribution of PSIs in Russian. In particular, (non)veridicality alone cannot account for the distribution of either *ni-* or *libo*-items. Yet, this should not be taken as a absolute rejection of the VBA, which captures the affinity of negative polarity, free choice, and mood selection phenomena. The present research shows that it is too early to dispense with the MBA because notions such as monotonicity and antimorphicity are essential in describing Russian PSIs.

References

Atlas, J. D. 1996 "*Only* Noun Phrases, Pseudo-Negative Generalized Quantifiers, Negative Polarity Items, and Monotonicity". *Journal of Semantics* 13, 265-328.

Atlas, J. D. 1998 *Adverbial Verb Phrases and Downward Monotonicity*. Paper Presented at the Negation Conference, Salford, UK.

Bernini, G. and P. Ramat 1996 *Negative Sentences in the Languages of Europe - A Typological Approach*. Mouton de Gruyter.

Blaszczak, J. 1998 *The Riddle of N-Words in Polish: Quantificational versus Nonquantificational, Universal versus Existential*. Paper Presented at the Negation Conference, Salford, UK.

Brown, S. 1999 *The Syntax of Negation in Russian: A Minimalist Approach*, Stanford, California: CSLI Publications.

Brown, S. and S. Franks 1997 "The Syntax of Pleonastic Negation in Russian". In Browne, W., E. Dornisch, N. Kondrashova and D. Zec (eds.) *Annual Workshop on Formal Approaches to Slavic Linguistics. The Cornell Meeting*, 135-164. Ann Arbor: Michigan Slavic Publishers.

Duffield, N. 1993 "On Case-Checking and NPI Licensing in Hiberno-English". *Rivista de Linguistica* 5(2), 215-244.

Giannakidou, A. 1997 *The Landscape of Polarity Items*, Ph.D. Dissertation, Rijksuniversiteit Groningen.

Giannakidou, A. 1998 *Polarity Sensitivity as (Non)Veridical Dependency*. Amsterdam: John Benjamins.

Halle, M. and A. Marantz 1993 "Distributed Morphology and the Pieces of Inflection". In Hale, K. and S. J. Keyser (eds.) *The View from Building 20*, 111-175.

Hasegawa, N. 1991 "Affirmative Polarity and Negation in Japanese". In Georgopoulos, C. and R. Ishihara (eds.) *Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda*, 271-285. Dordrecht: Kluwer.

Haspelmath, M. 1997 *Indefinite Pronouns*. Oxford: Clarendon.

Heim, I. 1984 "A Note on Negative Polarity and Downward Entailingness". *Proceedings of NELS 14*, 98-107.
Horn, L. R. 1996 "Exclusive Company: *Only* and the Dynamics of Vertical Inference". *Journal of Semantics* 13, 1-40.
Ladusaw, W. A. 1980 *Polarity Sensitivity as Inherent Scope Relations*. Ph.D. Dissertation. New York: Garland Publishing.
Laka, I. 1990 *Negation in Syntax: On the Nature of Functional Categories and Projections*. Ph.D. Dissertation, Cambridge, Mass.: MIT Working Papers in Linguistics.
Linebarger, M. C. 1987 "Negative Polarity and Grammatical Representation". *Linguistics and Philosophy* 10, 325-387.
Mahajan, A. 1990 "LF-conditions on Negative Polarity Licensing". *Lingua* 80, 333-348.
McCloskey, J. 1996 "On the Scope of Verb Movement in Irish". *Natural Language and Linguistic Theory* 14, 47-104.
Pereltsvaig, A. 1998 *Negative Polarity Items in Russian*. Paper Presented at the Negation Conference, Salford, UK.
Peres, J. 1998 *On Romance Sensitivity to Non-Veridicality*. Paper Presented at the Negation Conference, Salford, UK.
Progovac, L. 1988 *A Binding Approach to Polarity Sensitivity*. Ph.D. Dissertation, University of South California.
Rizzi, L. 1997 "The Fine Structure of the Left Periphery". in Haegeman, L. (ed.) *Elements of Grammar*. Dordrecht: Kluwer.
Uribe-Etxebarria, M. 1994 *Interface Licensing Conditions on Negative Polarity Items: A Theory of Polarity and Tense Interactions*. Ph.D. Dissertation, University of Connecticut.
Wouden, T. van der 1997 *Negative Contexts: Collocation, Polarity and Multiple Negation*. London: Routledge.
Zwarts, F. 1995 "Nonveridical Contexts". *Linguistic Analysis* 25, 286-312.

McGill University
Department of Linguistics
1085 Dr. Penfield
Montréal, PQ, Canada H3A 1A7
aperel@po-box.mcgill.ca

The Derivation of Yes-No *Li* Questions in Russian: Syntax and/or Phonology?*

Elena Rudnitskaya
CUNY Graduate Center

1. Introduction

According to Wackernagel (1953), enclitics in proto-Indo-European were cliticized to the first prosodic word (a word or string of words with an independent stress, henceforth PWd) in the clause. This mechanism of clitic placement is called 1W placement, whereas the other type of clitic placement available in Slavic languages is placement after the first constituent of the clause (1C placement). Both 1W and 1C clitics are instances of second position enclitics. I will consider Russian questions with the clitic *li*. *Li* is a yes-no question particle; it is a second position enclitic (King (1994)).

In Russian, unlike other Slavic languages, there are just a few clitics. There are no pronominal clitics, and the only clitics are the yes-no question clitic *li* and a couple of discourse clitics, such as *že* "as for, but". The placement of *li* is illustrated in (1):

(1) [Anina (li) sestra] (??li) prišla? (??li)
Ann's.NOM (li) sister.NOM (??li) came
'Is it Ann's sister that came?'

As (1) shows, 1W placement in Russian is strongly preferred over 1C placement. In certain Slavic languages, such as Serbo-Croatian

* I would like to thank Marcel den Dikken, Robert W. Fiengo, Tracy H. King and Yakov G. Testelec for extensive discussions and comments. Earlier versions of this paper were presented at "Dialog-98" in Russia, at SCIL-6 at the U of Texas at Arlington and at WCCFL 18 at the U of Arizona. I am grateful to the audiences of these conferences, as well as the audience of FASL-8 and the FASL-8 reviewers for helpful comments. I would also like to thank Russian native speakers for their help in coming up with accurate judgments.

(SC), 1C placement is possible or even preferred over 1W placement, cf. (2) from Progovac (1996:414).

(2) [$_{DP}$Anina (?im) sestra] (im) nudi čokoladu. [SC]
[$_{DP}$ Ana's (?them) sister] (them) offers chocolate
'Ana's sister is offering them chocolate.'

I will consider only 1W but not 1C *li* questions (the latter are extremely rare in Russian). Data on clitic placement (see Halpern (1992, 1995); Schütze (1994, 1996); Bošković (1998); Wilder & Ćavar (1994)) show that this phenomenon involves not only syntactic but also phonological mechanisms. There are two major points of view on the role of the phonological component in clitic placement. The first point of view, advocated by Bošković (1998), is that the phonological component contains a second position filter which rules in or rules out the output of the syntactic component, but phonology does not involve any movement operations. According to the second point of view (Halpern (1992, 1995); Schütze (1994, 1996)), the phonological component contains a last resort mechanism for reordering elements.

In particular, an enclitic can be initial in the output of the syntactic representation and acquire second position via Prosodic Inversion (PI) (Halpern (1992, 1995)). PI is a phonological operation that inverts the clause-initial enclitic and the first PWd of the clause. It is important that PI does not see any syntactic structure but only prosodic units, such as "clitic" and "PWd".

I will show that PI is involved in 1W *li* questions in Russian and propose a derivation of these questions based on PI.

2. PI is Involved in 1W *Li* Questions

Li questions involve syntactic fronting of a certain head/constituent which is the focus of the question (King (1994)):

(3) Uznal$_1$ li Ivan t_1 Petra?
recognized$_1$ li John.NOM Peter.ACC
'Has John recognized Peter?'

(4) [KNIGU]$_1$ li on prines t_1?
[book.ACC]$_1$ *li* he.NOM brought
'Is it the BOOK that he brought?'

According to King (1994), this movement is driven by the strong [+FOC] feature of *li*. Since, on King's view, *li* is base-generated in C, the focused element would move to the CP projection. King distinguishes between two types of focus fronting: head movement (verb fronting to C, as in the fronting of *uznal$_1$* in (3); cf. (25) in Section 3.2) and an XP narrow focus movement to SpecCP (e.g., object fronting, as in the fronting of the DP *KNIGU* in (4): I capitalize narrow focus). In both (3) and (4), *li* is in second position after focus fronting. The fact that the focus fronts is accounted for by invoking a [+FOC] feature which drives the fronting.

2.1. The Fronted Constituent vs. the Actual Focus

In the following examples, an XP is fronted which consists of more than one PWd, and *li* interrupts this XP. Such examples are pointed out as a problem by King because they cannot be accommodated in her analysis. Consider a fronted two-word XP. Three types of realizing a focus in this XP are possible: (a) the whole XP is the focus (5), (b) only the clitic host is the focus (6) and (c) the clitic host is not the focus but the second word, which follows the clitic, is the focus (7) – cf. a similar argument in Franks (1999a, 1999b).

(5) [$_{DP1}$ ANINU li SESTRU] on vstretil t_1?
[$_{DP1}$ Ann's.ACC li sister.ACC] he.NOM met
'Is it ANN'S SISTER that he met?' [or Mary's mother]

(6) [$_{DP1}$ ANINU li knigu] on prines t_1?
[$_{DP1}$ Ann's.ACC li book.ACC] he.NOM brought
'Is it ANN'S book that he brought?' [or Mary's book]

(7) [$_{DP1}$ Kartinu li VAN GOGA] MoMA
[$_{DP1}$ painting.ACC li Van Gogh.GEN] MoMA.NOM
priobrel t_1?
obtained
'Is it VAN GOGH'S painting that MoMA obtained?' [or Matisse's painting]

The word order in (5)-(7) cannot be achieved via King's analysis. If the DP fronts to SpecCP, and *li* is in C, *li* follows the whole fronted DP but not its first PWd. This yields 1C placement of *li*. In order to achieve 1W placement, additional phonological movement of *li* to the left until *li* reaches second position is needed. In (5), *li* would have to move over *SESTRU*. Since no phonological movement besides PI is possible, the fronted XP must land to the right of *li*, as in (8a). Then, PI (8b) can yield 1W placement of *li*.

(8) a. [$_{CP}$[$_{C'}$[$_{C}$li]][$_{XP\text{-}[+FOC]}$PWd$_1$PWd$_2$...]
[$_{AgrsP}$PWd$_N$ PWd$_{N+1}$...t$_{XP}$]]
b. PWd$_1$ li PWd$_2$ PWd$_3$ PWd$_4$ ...

On the first step (a) of (8), the DP which is/contains the focus is fronted to a position to the right of C (see Section 3.1 for details). After this syntactic movement, the clitic is clause-initial. The (b) step is PI: the first PWd of the clause and the clitic are inverted. For instance, the derivation of (6) will be (9a-b):

(9) a. [$_{CP}$ [$_{C'}$ [$_{C}$ li]] [$_{DP1}$ ANINU knigu] [$_{AgrsP}$ on prines t_1]]
b. [$_{PWd1}$ ANINU] li [$_{PWd2}$ knigu] [$_{PWd3}$ on] [$_{PWd4}$ prines]

For examples such as (6), in which the actual focus is the first PWd of the fronted XP (*ANINU*), another derivation (10a-b) is also possible, which is purely syntactic. Its first step is (10a)=(9a), and the second step (10b) is syntactic movement of the focus (*ANINU*) out of DP$_1$ to SpecCP. (10b) can be regarded as tenable in case of (6) because *ANINU* can undergo independent syntactic movement in other environments (see (11) below):

(10) a. [$_{CP}$ [$_{C'}$ [$_{C}$ li]] [$_{DP1}$ ANINU knigu] [$_{AgrsP}$ on prines t_1]]
b. [$_{CP}$[$_{AP2}$ANINU][$_{C'}$[$_{C}$ li]][$_{DP1}$ t_2 knigu][$_{TP}$ on prines t_1]]

If the derivation (10a-b) is to be adopted instead of (9a-b) for (6), King's claim that the focus moves to SpecCP in syntax can be preserved at least in some cases. However, I would like to maintain a unitary derivation (8)/(9) for all of (5)-(7) and give up (10a-b). Also, data from the so-called "fortress" examples to be presented below prove that the derivation (10a-b) is not tenable.

When an XP fronts in (5)-(7), this XP is the actual focus only in (5) (*ANINU SESTRU*). In (6)-(7), the fronted DP contains the focus as its first (*ANINU* in (6)) or its second (*VAN GOGA* in (7)) PWd. That is, the actual focus pied-pipes the DP when it fronts.[1]

This pied-piping is not obligatory, except for certain cases such as "fortresses". In (11)-(12), only the actual focus fronts:[2]

(11) [$_{AP1}$ ANINU] li on prines [$_{DP}$ t_1 knigu]?
[$_{AP1}$ Ann's.ACC] li he.NOM brought [$_{DP}$ book.ACC]
'Is it ANN'S book that he brought?' [or Tanya's book]

---

[1] The actual focus can pied-pipe the argument/adjunct XP in which it is contained but it cannot pied-pipe the TP of the clause. If this were possible, the verb would front together with its complement yielding a structure like (i). Then, PI would yield the ungrammatical (ii):

(i) [$_{CP}$ [$_{C'}$ [$_{C}$ li]] [$_{TP1-[+FOC]}$ prines ANINU knigu] [$_{AgrsP}$ on t_1]]
(ii) *[$_{TP1}$ Prines li ANINU knigu] on/Ivan t_1?
[$_{TP1}$ brought li Ann's.ACC book.ACC] he.NOM/John.NOM
[cf. (23a-b) in Section 3.2)]

[2] Example (12) shows that the ban on Right Branch extraction, pointed out in Sekerina (1997), does not always hold for right branch complements.

(12) [$_{DP2}$ VAN GOGA] li MoMA priobrel
[$_{DP2}$ Van Gogh.GEN] li MoMA.NOM obtained
[$_{DP1}$ kartinu t_2]?
[$_{DP1}$ painting.ACC]
'Is it VAN GOGH'S painting that MoMA obtained?' [or Dega's painting]

2.2. "Fortress" Examples

An important argument in favor of the PI derivation (8) is the fact that there are no fortresses for *li* placement in Russian. Fortresses are certain types of constituents whose parts cannot undergo syntactic movement. Progovac 1996 and Bošković 1998 claim that since clitic placement in SC is purely syntactic, clitics cannot interrupt a "fortress" in SC. Examples (13)-(14) show that *li* can interrupt a "fortress" in Russian, both when the whole fortress (13) and the clitic host (14) is the focus.

(13) ?[ANNA li I PETR] priezžajut vo
[Ann.NOM li and Peter.NOM] come on
vtornik?
Tuesday.ACC
'Is it ANN AND PETER who are coming on Tuesday?'
[or Mary and John]

(14) a. [LEV li Tolstoj]
[Leo.NOMli Tolstoy.NOM] COP
velikij russkij pisatel'?
great.NOM Russian.NOM writer.NOM
'Is it LEO Tolstoy who is a great Russian writer?'
[or Aleksej Tolstoj]

b. [$_{DP1}$ PRIJATELEJ li tvoej sestry]
[$_{DP1}$ friends.ACC li your.GEN sister.GEN]
my tol'ko čto videli t_1?
we.NOM just saw
'Is it your sister's FRIENDS who we have just seen?'
[or relatives of your sister]

(13)-(14) cannot be derived without PI because the clitic host cannot be moved out of the fortress, as (15) and (16a-b) show.

(15) *[$_{?P2}$ ANNA] li priezžajut vo vtornik
[$_{?P2}$ Ann.NOM] li come on Tuesday.ACC
[$_{DP1}$ t_2 I PETR]?
[$_{DP1}$ and Peter.NOM]

(16) a. *[$_{?P2}$ LEV] li Ø velikij
[$_{?P2}$ Leo.NOM] li COP great.NOM
russkij pisatel' [t_2 Tolstoj]?
Russian.NOM writer.NOM [Tolstoy.NOM]
b. *[$_{?P2}$ PRIJATELEJ] li my tol'ko čto videli
[$_{?P2}$ friends.ACC] li we.NOM just saw
[t_2 tvoej sestry]?
[your.GEN sister.GEN]

Examples (13)-(14) vs. (15)-(16) show that PI is necessary for deriving "fortress" *li* questions. Generalizing, I propose that PI is always involved in 1W *li* placement.

Fortresses in which the clitic host is the actual focus (*LEV* and *PRIJATELEJ* in (14a-b)) are crucial for adopting the PI derivation (9) but not the syntactic derivation (10) for (6). According to (10b), *ANINU* moves from DP_1 to SpecCP in syntax. Even though independent syntactic movement is possible for *ANINU* in (11), it is impossible for *LEV* and *PRIJATELEJ*, as in (16a-b). Then, (10) cannot be extended to "fortress" examples. Since it is important to have a unitary derivation for all cases in which the host of *li* is the actual focus (both (6) and (14)), I adopt the derivation (9).

To conclude, I have shown in this section that the derivation of 1W *li* questions crucially involves PI; these questions cannot be derived via syntactic movement alone. *Li* questions are derived in two steps. First, the focus of the question or the XP containing the focus fronts to a position between CP and Agr_SP (a focus position, see Section 3 for details). Second, PI yields 1W placement of *li*.

3. The Derivation of Yes-No Questions with *Li*

3.1. The Landing Site of the Fronted Focus

According to (8a) in Section 2.1, the fronted XP in *li* questions is in a position right below CP. What position is this? Stjepanović (1995) proposes for SC and Stepanov (1998) for Russian that a narrow focus in a statement is either preverbal or clause-initial; a focus *in situ* can rarely be a narrow focus.

(17) a. Ivan KNIGU kupil. [Stepanov 1998:461]
John.NOM book.ACC bought
b (Èto) KNIGU Ivan kupil.
this.DFT book.ACC John.NOM bought
c. (??)Ivan kupil KNIGU.
John.NOM bought book-A
'It is a BOOK that John bought.'

My judgments are different from Stepanov's judgments. Both the preverbal and the *in situ* positions in (17a) and (17c) are neutral (I place Stepanov's "??" for (17c) in parentheses). The initial position in (17b) is degraded, especially without *èto* "this".[3]

Bošković (1997, 1999) proposes that foci in statements are fronted and adjoined to $Agr_{O/S}P$ (to yield (17a) and (17b)). By contrast, I assume that no focus movement occurs in statements. Instead, the focus status of an XP, for instance of *KNIGU,* can be achieved via focal stress assignment even if this XP stays *in situ*, as in (17c). A focal stress can also be assigned to an XP which is fronted for independent reasons, as in (17a-b).

Li questions differ from statements in that their focus is always in the initial position; the focus *must* be fronted, as (18a-c) show.

---

[3] *ÈTO* highlights the focused status of *KNIGU. ÈTO* is not a real demonstrative since it always has the default form SG.NEUT and does not have to agree with the following noun (see the analysis of *ÈTO* in Junghanns (1997)).

(18) a. *Ivan KNIGU li kupil?
John.NOM book.ACC li bought
b. KNIGU li Ivan kupil?
book.ACC li John.NOM bought
c. *Ivan kupil KNIGU li?
John.NOMbought book.ACC li
'Is it a BOOK that John bought?'

Russian also allows yes-no questions without any question particle, and the focus in such questions does not have to be fronted:

(19) a. Ivan KNIGU kupil?
John.NOMbook.ACC bought
b. ??KNIGU Ivan kupil?
book.ACC John.NOM bought
c. Ivan kupil KNIGU?
John.NOMbought book.ACC
'Is it a BOOK that John bought?'

I assume that the structure of questions without *li* (19) is similar to the structure of statements (17) rather than of *li* questions (18). That is, no focus fronting occurs in these questions, and their interrogative status is yielded via question intonation assignment.[4]

---

[4] A possible argument for this analysis is that yes-no questions without *li* can only be matrix but not embedded. I propose that the C of matrix questions without *li* has no [+Q] feature at all. A matrix statement has a sentential force operator in C, and I assume that this operator can be converted into an interrogative operator; then, a statement (17) turns into a question (19). The C in (19) has no [+Q] feature because there is no *li* particle in either (17) or (19). It is *li* but not the focus of the question which has the [+Q] feature; hence the [+Q] feature of C can only be checked with the [+Q] feature of *li*. Thus, C must have a [strong] [+Q] feature only in questions with *li*. In Section 3.2, I propose that a C with a [+Q] feature subcategorizes for the Foc projection in which *li* is generated; in this case, the [+Q] feature of C will always be checked.

Embedded yes-no questions without *li* are impossible. I assume this is because an embedded clause has no sentential force operator in C, and hence an interrogative operator cannot be obtained. A strong [+Q] feature is generated in

I would like to argue that the landing site of the focus fronting in *li* questions is not the AgrP-adjoined position but a Spec position of a SpecFocP projection. First, focus fronting is only required in *li* questions (18) but not in yes-no questions without *li* (19). This implies that the focus in *li* question fronts not because it has an inherent strong [+FOC] feature (this would be the case if it fronted to an AgrP-adjoined position) but because *li* questions have a functional projection with a strong [+FOC] feature which attracts focus fronting. This projection is absent in yes-no questions without *li*, and therefore foci do not have to front in such questions.

Second, *li* questions cannot contain multiple question foci (this constraint is unexpected if the focus fronts to an adjoined position). Multiple question foci are hardly possible in any yes-no questions for pragmatic reasons, but *li* questions with multiple fronted foci are especially degraded: cf. (20) without *li* and (21).

(20) #IVAN MARII dal knigu,
JOHN.NOM MARY.DAT gave book.ACC
ili Petr Anne?
or Peter.NOM Ann.DAT

(21) #*?/*IVAN li MARII dal knigu,
JOHN.NOM li MARY.DAT gave book.ACC
ili Petr Anne?
or Peter.NOM Ann.DAT
'Is it JOHN, and is it to MARY that he gave the book, or is it Peter and Ann?'

Examples (20) vs. (21) show that only one landing site is available for focus fronting in *li* questions (# is "pragmatically ill-formed").[5]

C instead, and C subcategorizes for a FocP with *li*. Then, [+Q] is always checked.

[5] An alternative explanation for the fact that *li* questions can have only one focus is proposed by King (1994): this fact is derived from the semantics of focus (by assuming that the non-fronted part of a *li* question is presupposed, cf. the discussion of clefting and presupposition in Jackendoff (1972)). The remaining question in King's account is why multiple foci are marginally possible in yes-

I propose the following account of focus fronting in *li* questions. The clause in a *li* question has a Foc head which projects a maximal projection FocP.[6] FocP is subcategorized by a C which has a [+Q] feature (cf. footnote 4). *Li* in Foc has a strong [+FOC] feature that attracts the focused phrase to SpecFocP. SpecFocP cannot be multiply filled because the [+FOC] feature deletes after the first checking operation; another focused XP cannot be attracted. The derivation of a *li* question like (18b) will be as in (22):

(22) a. $[_{CP}\,[_{C'}\,[_{C}$ li]] $[_{FocP}\,[_{Foc'}\,[_{Foc}$ [+FOC]]]]
Ivan kupil KNIGU-[+FOC]]]
b. $[_{CP}\,[_{C'}\,[_{C}$ li]] $[_{FocP}\,[_{DP1}$ KNIGU-[+FOC]]
$[_{Foc'}\,[_{Foc}]]$ Ivan kupil t_1]
c. $[_{PWd1}$ KNIGU] li $[_{PWd2}$ Ivan] $[_{PWd3}$ kupil]

This derivation explains the obligatory fronting of the focus in *li* questions and the ban on multiple foci in *li* questions.

3.2. *Li* is Both a Focus and a Question Particle

Besides the second position restriction on *li* discussed above, there is another important restriction: *the host of li must be (part of) the constituent fronted to SpecFocP*. Compare well-formed (6)-(7) from Section 2.1 and ill-formed (23a-b):

(23) a *Ivan li $[_{DP1}$ ANINU knigu]
John.NOM li $[_{DP1}$ Ann's.ACC book.ACC]
prines t_1?
brought
Impossible: 'Is it ANN'S book that John brought?';
'Is it John who brought ANN'S book?'

---

no questions without *li*, as in (20).

[6] Cf. the proposal in Brody (1990) for Hungarian.

b. *MoMA li [$_{DP1}$ kartinu VAN GOGA]
MoMA.NOM li [$_{DP1}$ painting.ACC Van Gogh.GEN]
priobrel t_1?
obtained
Impossible: 'Is it VAN GOGH'S painting that MoMA obtained?';
'Is it MoMA that obtained VAN GOGH'S painting?'

In all of (23), *li* is in second position. This position can be acquired either (I) without PI via fronting the subject (*Ivan* in (23a)) to SpecCP, or (II) via PI of *li* with *Ivan* after *Ivan* has been moved to SpecXP (which is between CP and FocP). (23) must be ruled out.

Since *li*'s host *kartinu* in (7) is not itself an actual focus, the host of *li* is not always the actual focus. *Kartinu* belongs to the preposed constituent *kartinu VAN GOGA*, which is pied-piped by the focus *VAN GOGA* to SpecFocP, so *kartinu* can host *li*. *MoMA* in (23b) is not an actual focus either but, unlike *kartinu*, *MoMA* is not located in SpecFocP; therefore *MoMA* cannot host *li*.

I propose the following account of the constraint on *li* hosts in (23). First, I stipulate that *li* questions have a null question operator (Q_{op}) in SpecCP. Therefore, SpecCP in these questions is always filled in syntax but no overt material can occupy SpecCP.[7] Thus, *li* is never clause-initial in syntax but since Q_{op} is null, *li*'s second position requirement can be achieved only via PI.

Second, I propose that *li* is both a question particle and a focus particle: *li* is base-generated in Foc with the [+FOC] feature. The [+FOC] feature of *li* attracts preposing of the focused phrase to SpecFocP (cf. King's (1994) account). Then, *li* overtly moves to C because *li* is also a question particle and has the [+Q] feature; this feature must be checked with the strong [+Q] feature of C. Thus, *li* ends up in C, and the focus phrase ends up in SpecFocP immediately to the right of C; the configuration for PI is achieved.[8]

---

[7] SpecCP cannot be doubly filled; thus, the possible derivation (I) of (23) mentioned above is ruled out.

[8] When *li* moves from Foc to C, no other head X can intervene: in footnote 4, I propose that a C with the [+Q] feature subcategorizes for FocP. Adjunction to

Thus, the derivation of a *li* question like (18b) is (24a-b) instead of (22a-b); the last PI step remains the same:

(24) a. [$_{CP}$ Q_{op} [$_{C'}$ [$_{C}$ [+Q]]] [$_{FocP}$ [$_{Foc'}$ [$_{Foc}$ li-[+Q]-[+FOC]]]
Ivan kupil KNIGU-[+FOC]]]
b. [$_{CP}$ Q_{op} [$_{C'}$ [$_{C}$ li]] [$_{FocP}$[$_{DP1}$ KNIGU-[+FOC]]
[$_{Foc'}$ [$_{Foc}$ t_{li}]] Ivan kupil t_1]]

The remaining problem is cases in which a verb is fronted, as in (3) in Section 1. I propose that no PI is required in these cases (see Rudin & al. (1999) for an alternative solution). First, V head-moves to Foc and left-adjoins to *li* instead of moving to SpecFocP, and the [+FOC] feature of *li* is checked; then, V moves to C with *li*. Since V is left-adjoined to Foc, V is clause-initial, and only V precedes *li*. Thus, *li* is in second position.

Several PWd-s cannot precede *li* in V-fronting cases. In principle, the fronted V might be a complex head which consists of an auxiliary and a main verb; then, two PWd-s would precede *li*. However, such a situation never emerges in actual examples:

(25) a. Budet li Ivan est' maslo?
will li John.NOM to-eat butter.ACC
b. ??Budet li est' (est' li budet)
will li to-eat (to-eat li will)
Ivan maslo?
John.NOM butter.ACC
'Will John eat the butter?'

The contrast between the well-formed (25a) and the degraded (25b) shows that only the auxiliary but not both the auxiliary and

---

FocP is ruled out because it will create another focus (the [+FOC] feature will percolate to the adjoined phrase) and violate the ban on multiple foci in *li* questions (see (20)-(21) and footnote 5). Thus, no material can intervene between *li* and SpecFocP, and the derivation (II) of (23a) mentioned above is impossible.

the lexical verb undergoes fronting to Foc. Thus, both of these verbs can not precede *li* after the movement of Foc to C.

4. Conclusions

Syntactic movement alone cannot derive 1W *li* placement in Russian, and an additional phonological operation PI is required. This implies that in other Slavic languages with clitics, 1W placement cannot be achieved via syntactic movement at least in some cases (e.g., in yes-no questions with *li*).

The analysis of focus in this paper implies that focus can be licensed *in situ* in Russian. A C with a [+Q] feature subcategorizes for FocP, and a [+FOC] feature of *li* in Foc, but not an interrogative operator or a [+Q] feature in C, can license syntactic focus fronting.

References

Bošković, Ž. 1997. "Fronting W*h*-Phrases in Serbo-Croatian". In M. Lindseth and S. Franks (eds.) *Formal Approaches to Slavic Linguistics: The Indiana Meeting*, 86-107. Ann Arbor: Michigan Slavic Publications.

Bošković, Ž. 1998. "Second Position Cliticization: Syntax or Phonology?". To appear in Beukema, F. and M. den Dikken (eds.) *Clitic Phenomena in European Languages*. Amsterdam: John Benjamins.

Bošković, Ž. 1999. *What is Special About Multiple Wh-Fronting?* Ms., UCONN.

Brody, M. 1990. "Remarks on the Order of Elements in The Hungarian Focus Field". In I. Kenesei (ed.) *Approaches to Hungarian, Vol. 3: Structures and Arguments.* Szeged: JATE.

Franks, S. 1999a. "Optimality Theory and Clitics at PF". In K. Dziwirek, H. Coats and C. M. Vakareliyska (eds.) *Formal Approaches to Slavic Linguistics: The Seattle Meeting 1998,* 101-123. Ann Arbor: Michigan Slavic Publications.

Franks, S. 1999b. "Clitics at the Interface". To appear as Introduction to Beukema, F. and M. den Dikken, (eds.) *Clitic*

Phenomena in European Languages. Amsterdam, John Benjamins.

Halpern, A. L. 1992. *Topics in the Placement and Morphology of Clitics*. Ph.D. Dissertation, Stanford University.

Halpern, A. L. 1995. *On the Placement and Morphology of Clitics.* Stanford: CSLI Publications.

Jackendoff, R. 1972. *Semantic Interpretation in Generative Grammar.* Cambridge, MA: MIT Press.

Junghanns, U. 1997. "On the So-Called *Èto*-Cleft Constructions". In Lindseth, M. and S. Franks (eds.). *Formal Approaches to Slavic Linguistics: The Indiana Meeting*, 166-190. Ann Arbor: Michigan Slavic Publications.

King, T. H. 1994. "Focus in Russian Yes-No Questions". *Journal of Slavic Linguistics* 2.1, 92-120.

Progovac, L. 1996. "Clitics in Serbian/Croatian: Comp and the Second Position". In A. L. Halpern and A. M. Zwicky (eds.) *Approaching Second: Second Position Clitics and Related Phenomena,* 411-428. Stanford: CSLI publications.

Rudin, C., C. Kramer, L. Billings and M. Baerman. 1999. "Macedonian and Bulgarian *li* Questions: Beyond Syntax". To Appear in *Natural Language and Linguistic Theory 17.3*.

Schütze, C. 1994. "Serbo-Croatian Second Position Clitic Placement and the Phonology-Syntax Interface". *MIT Working Papers in Linguistics* 21, 373-473.

Schütze, C. 1996. "Serbo-Croatian Clitic Placement: An Argument for Prosodic Movement". In J. Toman (ed.) *Formal Approaches to Slavic Linguistics: the College Park Meeting,* 225-248. Ann Arbor, Michigan Slavic Publications.

Sekerina, I. A. 1997. *The Syntax and Processing of Scrambling Constructions in Russian.* Ph.D. Dissertation, The City University of New York..

Stepanov, A. 1998. "On W*h*-Fronting in Russian". In Tamanji, P. and K. Kusumoto (eds.) *Proceedings of the North Eastern Linguistic Society* 28. Amherst, University of Massachusetts, 453-467.

Stjepanović, S. 1995. "Short-Distance Movement of Wh-Phrases in Serbo-Croatian Matrix Clauses". Ms., University of Connecticut. Wackernagel, J. 1953

Wackernagel, J. 1953. "Über Ein Gesets Der Indogermanischen Wortstellung". In *Kleine Schriften,* Göttingen, B.1.

Wilder, C. and Ćavar, D. 1994. "Long Head Movement? Verb Movement and Cliticization in Croatian". *Lingua* 93, 1-58.

Linguistics Program
CUNY Graduate Center,
365 Fifth Avenue
New York, NY 10016-4309
erudnitskaya@gc.cuny.edu

Porting A Rapid Deployment Morphological Analyzer Across Slavic Languages

Svetlana Sheremetyeva and Sergei Nirenburg
New Mexico State University

1. Introduction

Most current systems of computational morphology rely, in addition to word inflection paradigms, on large lexicons of stems. Morphological modules of larger natural language processing (NLP) systems are expected to provide near-perfect coverage. Therefore, the completeness of lexicons is an ongoing concern especially if one considers that, according to some estimates, on average five new words enter a major language, such as English or Russian, every day. It is often the case that large on-line dictionaries are not available to NLP developers. They either don't exist or simply cannot be reached at the moment, as was the case with the Corelli project *http://crl.nmsu.edu/Research/Projects* at the Computing Research Laboratory (CRL), New Mexico State University (NMSU). It would be a clear advantage to be able to carry out morphological processing without the reliance on a large dictionary of stems and, therefore, without a need for a large-scale knowledge acquisition effort. Our system attempts just that.

This paper presents a model of morphological analysis which, unlike practically all other approaches to computational morphology, does not rely on a large dictionary of stems. For the treatment of inflectable open-class items several relatively small-scale lexicons are used in place of the usual large lexicon. The model also covers proper names and hyphenated words. The system is robust and can be used both for lemmatizing (base form generating) and determining the values of morphosyntactic features of word forms. It can be used both in text analysis and generation.

The Rapid Deployment Morphology (RDM) model described here has been implemented in the framework of the Corelli project at CRL, NMSU. It was originally developed for Russian and then successfully applied to Serbo-Croatian. We expect it to be extendable to a variety of inflecting and agglutinating languages.

The central differences between our model and those used in practically all other approaches to computational morphology—two-level "KIMMO" systems (Koskenniemi 1984; Antworth 1990), DECOMP (Allen et al. 1987), MORPHOGEN (Pentheroudakis and Higinbotham 1991) or the morphological module in the ETAP-2 system (Apresyan et al. 1989), to name just a few — lie in a) the lack of reliance on a comprehensive stem dictionary and b) in the choice of the basic units of description. Two-level rules are based on the correspondence between lexical and surface forms; in other models rules are described in terms of abstract *morpheme* categories (MORPHOGEN) or *morphs*, the orthographic representation of morphemes (DECOMP, ETAP-2) and inflection paradigms. In our model, a crucial unit of description, the *quasi-root* (see definition below) is not a regular morpheme, as it is not expected to carry meaning. This decision has been made for reasons of assuring coverage, efficiency and economy of knowledge acquisition. It does not advance theoretical connections between morphology and semantics. This approach is one of those "bags of tricks" which, in the opinion of Yehoshua Bar-Hillel, seconded by Yorick Wilks, are essential components of all practical NLP systems and "will advance computational linguistics in the future" (Wilks 1996:62).

One of the requirements for our model was the absence of reliance on dictionaries of base forms, roots or stems. In an attempt to fulfill this requirement, we have observed that many Russian words belonging to the same inflectional paradigm have identical strings of characters just before their endings. For example, a Russian character string *-ani*, associated with the following declension paradigm: *-e -ja -ja -j -ju -jam -e -ja -em -jami -i -jax* is common to a large number of nouns such as *zavoev**ani**e* 'gain', *obzalov**ani**e* 'appeal', *osnov**ani**e* 'foundation', *sozd**ani**e* 'creation', *sobr**ani**e* 'meeting', *pit**ani**e* 'nourishment', etc. In fact, The

Grammatical Dictionary of Russian (Zaliznjak, 1980) lists on the order of 1,500 words of this type. We hypothesized that this shared string (we call it "quasi-root" to distinguish it from the more traditionally defined "root" morpheme) can be used as a determinant of the morphological information for the entire class. This paper reports the results of implementing and testing this hypothesis. The article is organized as follows. In Section 2 we describe the morpheme classes recognized by RDM and the lexicons constituting the knowledge base of the Russian implementation of the model. The process of morphological analysis in RDM is described in Secrion 3. Section 4 walks the reader through the RDM analysis of a Russian example. Section 5 gives specifics concerning porting the RDM to Serbo-Croatian. The enhancements of the RDM approach when an on-line dictionary is available are described in Section 6. Section 7 presents evaluation results. Conclusions are in Section 8.

2. RDM For Russian

2.1. Morphotactics

The RDM model defines the following types of sub-word strings which play the role of morphemes:

reflexive: a word-final character string whose values are one of: {*sja, s'*};

ending: a word-final or pre-reflexive character string (possibly, empty), as found in the lexicon of endings. Endings recognized in our model do not always coincide with those in traditional grammar. We posit that all the substrings at the end of the word, which change during declension, are endings. Our choice of endings was informed by the desire to bypass special treatment of alternations. This is why the ending in *ugol* 'angle' is *ol* (because of the existence of *uglom* (singular instrumental), in which *lom* is the ending. Theoretically, *ugol* has a null ending, while the ending in *uglom* is *om*;

suffix: a character string (possibly, empty) immediately preceding the ending, as found in the lexicon of suffixes;

quasi-root: a three-letter character string immediately preceding the ending or suffix, if present;

prefix: a word-initial character string (possibly, empty), as found in the lexicon of prefixes;

body: a compound character string (never empty) starting after the prefix (if present) and ending before a suffix, if present, or ending;

stem: a word-initial compound character string (never empty) stretching to the ending, as found in the lexicon of stems.

The morphotactics of Russian in our model is defined by the following very simple grammar:

(1) word ::= stem ending reflexive;
stem ::= prefix [string quasi-root | body] suffix,

where "string" stands for any character string including the null string.

For example, the correct analysis of the Russian word *sdelala* ('*did*', feminine, singular) leads to the following assignments of morph types:

(2) *sdelala* = stem: *sdela* ending: *la* reflexive: Ø
stem: *sdela* = prefix: *s*[string: quasi-root: *ela*|body: *dela*]
suffix: Ø

2.2. The Knowledge Base

The knowledge base in the RDM for Russian includes the following lexicons.

The lexicon of endings (about 900 entries) which consists of all the experimentally derived inflections for nouns, verbs, participles and adjectives. The entries contain information from inflection paradigms for verbs, adjectives and participles. For example, the entry of the ending *la* is as follows:

(3) *la*
noun, 25, masculine, singular, genitive
verb, imperfective, past, feminine, singular,
verb, perfective, past, feminine, singular

The number in the nominal reading is that of the declension paradigm. This entry is matched by such words as *ugla* ('*angle*', noun, masculine, singular, genitive), *sdelala* ('did', verb, perfective, past, singular, feminine), *bezala* ('was running', verb, imperfective, past, singular, feminine), etc.

The lexicon of suffixes (58 entries) which contains only suffixes of full participles and adjectives in the superlative degree. A sample entry is as follows:

(4) *enn*
participle, full, perfective, passive

This suffix can be identified in such words as *rassmotrennyj* 'considered', *uvidennaja* 'seen', *dobavlennoe* 'added'.

The lexicon of quasi-roots (about 2,900 entries) in which every entry is marked as a nominal, verbal or adjectival quasi-root. The adjectival and verbal entries include endings of their base forms. The nominal entries include a set of declension paradigm numbers with the endings of their base forms. This lexicon is used to resolve categorial homonymy among verbal, nominal and adjectival candidates as well as to construct base forms in case of verbs, participles and adjectives. Sample entries of the quasi-root lexicon are:

(5) *ani*
Noun, 3 ja; 11 e

(6) *ack*
Adjective, ij

(7) *ata*
Verb, t'

The lexicon of prefixes (53 entries) which contains only prefixes of perfective forms of the verbs.

The lexicon of bodies which contains bodies of perfective verbs and is used for homonymy resolution between perfective and imperfective verbs (quasi-roots of such verbs are homonymous) and for finding base forms of perfective participles and perfective verbs.

(8) *blokir*
ovat'

(9) *vra*
tit'

(10) *smotr*
et'

The lexicon of stems (7,000 entries) which include a small number of special-treatment items; this lexicon is quite small and is used to resolve homography in morphological analysis. For example, the lexicon of stems contains the entries

(11) *peremenn*
Adjective *yj*; Noun, feminine, *aja*

(12) *poverenn*
Participle *yj;* Noun, masculine, *yj*

3. Analysis in the RDM Model

The model accepts an inflected word form as input and returns all its legal base forms with all legal morphosyntactic feature value sets. The analysis process starts at the end of the word and proceeds backwards. Segmentation is carried out simply by matching the characters of a word against lexicon entries or by chopping off a certain number of characters to the left of the current segmentation point. We use a bottom-up, depth-first parsing algorithm.

The main analysis procedure is Analyze-Word:

(13) **procedure** *analyze-word*
find-candidates;
process-candidates.

The procedure *find-candidates* searches the lexicon of endings for all possible matches on the input word, returning all possible pairs of endings and their feature values which are stored with the entries of the lexicon of endings. Each such pair is called a *candidate*. Several candidates may be returned for a single word. For instance, for the word *rassmotrennymi* ('considered', participle, perfective, passive, plural, instrumental) *find-candidates* will return 72 candidates some of which are given below:[1]

(14)	a	*-ymi*	Noun, 17, neuter, plural, instrumental
	b	*-ymi*	Noun, 18, feminine, plural, instrumental
	c	*-ymi*	Noun, 48, masculine, plural, instrumental
	d	*-ymi*	Noun, 49, masculine, plural, instrumental
	e	*-ymi*	Participle full, plural, instrumental)
	f	*-ymi*	Adjective plural instrumental
	g	*-mi*	Verb imperative perfective singular
	h	*-i*	Verb imperative perfective singular
	i	*-i*	Verb imperative imperfective singular
	j	*-i*	Adjective short, plural
	k	*-i*	Noun, 2, feminine, plural, nominative
	l	*-i*	Noun, 11, neuter, singular, locative

Further processing will, at least partially, disambiguate the homonymy among the readings.

1. This example involves the most ambiguous ending in RDM, *-i*; we list only five candidates for this ending. Only very few, if any, of these candidates undergo a full treatment in RDM, see Section 4.

It is carried out by the following procedure:[2]

(15) *process-candidates*
process-nominal-candidates;
IF no legal nominal readings
THEN begin
process-participle-candidates;
IF no legal participle candidates
THEN begin
process-verbal-candidates;
IF no legal verbal candidates
THEN *process-adjective-candidates*
end
end
ELSE *process-verbal-candidates*
UNLESS stems of legal nominal readings are in lexicon of stems;

The input to every component procedure in *process-candidates* is the input word and a set of corresponding candidates (nominal, participial, verbal and adjectival). The order of the calls to component procedures in the above algorithm is established to minimize the processing time and effort, as it is possible to avoid calling each of the component procedures with each input word. The ordering is based on a set of heuristics such as the following. Noun candidates are processed first because a vast majority of words in a Russian text are nouns and nouns can relatively seldom be also analyzed as adjectives, and can very seldom be verbs or adverbs. If a participle candidate has been analyzed successfully, there is no need to test any verbal candidates because in Russian participles and finite forms of verbs are practically never homographic. This order is clearly language-dependent (though for Serbo-Croatian, the order developed originally for Russian

2. We describe the analysis process conceptually. Implementations can vary.

worked). If the ordering heuristics are not available, the control structure of *process-candidates* can be easily modified to have each subroutine called in any order. We understand that a more fashionable approach is to use a universal rule interpreter and not to be concerned with heuristics-based control issues. However, universal interpreters sometimes make it more difficult to write processing rules, and, in any case, in the absence of such an interpreter, the concern for minimization of effort led us to use linguistic heuristics for control.

4. An Example of Processing in the RDM Model

We describe the procedures for processing the candidates through a Russian example. We trace the processing of the participle *rassmotrennymi 'considered'*. This example is quite complex and we selected it in order to illustrate as many of the types of processing performed by the system as possible.

The sample set of candidates for this word is given above in (14), see Section 3. Procedure *process-candidates* stipulates that the nominal candidates are tested first. Procedure *process-nominal-candidates* first chops off the quasi-root of the input word. The quasi-root in *rassmotrennymi* with the ending *ymi* is *enn*; the quasi-root with the ending *i* is *nym*. Next, these quasi-roots are looked up in the lexicon of quasi-roots. No match with the part-of-speech feature value *noun* is found for *nym* in the quasi-root lexicon, so all the candidates (14j)-(14l) are immediately discarded. A nominal match is found for *enn*. The corresponding entry in the lexicon of quasi-roots states that the noun belongs to Declension Paradigms 17, 18, 48 or 49 (which are all paradigms for deadjectival nouns). (As an example, Paradigm 18 is as follows: *-aja -ye -oj -yh -oj -ym -uju -ye -oj -ymi -oj -yh.)*

Process-nominal-candidates next checks whether any of these paradigm numbers appear in the candidates. If it is not the case, the candidate is discarded. If it is, the base form of a noun is identified in the quasi-root lexicon. The input word with its base form and the

feature value list is then output as a result of the analysis process.

Some declension paradigms (including Paradigms 17, 18, 48 and 49) are applicable to adjectives and full participles. For these paradigms, the procedure finds the stem of the input word and checks whether this stem is listed in the stem lexicon as a nominal stem. The stem *rassmotrenn* is not in this lexicon. Therefore, Candidates (14a)-(14b) are discarded.[3]

Next, *process-participle-candidate* is called for Candidate 18. The procedure first attempts to match a string to the left of the ending to an entry in the lexicon of suffixes. Our input matches the suffix *enn*, as a result of which the feature values *passive* and *perfective* are added to the list of feature values in the candidate. At this point, Candidate (14e) is as follows:

(16) *ymi* Participle full, passive, perfective, plural, instrument.

It remains to determine the base form of the word. For participles we require both the participial and the verbal base form. The former is derived graphotactically — if the participial suffix ends in a sibilant, the base form ending is *ij;* otherwise, it is *yj*. The latter is determined using the lexicon of quasi-roots or, in case of prefixed perfective forms, in the lexicon of bodies. In our case the prefix *ras* is identified and the body *smotr* (*ras-smotr-enn-ymi*) is extracted. The body lexicon has the entry *smotr-et'*. The final output is:

(17) *rassmotrennymi rassmotrennyj rassmotret'* Participle full, passive, perfective, plural, instrumental.

3. The lexicon of stems is only accessed for a subset of each set of candidates that share a nominal quasi-root or when a nominal candidate belongs to a shared nominal/adjectival paradigm. This means that even if no match occurs in the lexicon of stems, there are always remaining candidates to select for output.

5. Porting the RDM from Russian to Serbo-Croatian

5.1. The Knowledge Base

Both Russian and Serbo-Croatian have the three-way gender opposition of masculine, feminine, and neuter, and two numbers, singular and plural. They share six cases, but Serbo-Croatian, similarly to Ukrainian, also features an extra case, vocative. Almost all nouns decline. The morphology of both languages is predominantly fusional; gender correlates strongly with declension class. Like Russian, Serbo-Croatian is rich in alternations and suppletive forms. For this reason, the lexicon of endings for Serbo-Croatian was constructed the same way it was done for Russian— using non-traditional definitions of endings and (in a small number of cases) describing a word with the help of two suppletive quasi-roots.

Serbo-Croatian has a multi-tense system of verbal morphology which differs from that of Russian. It has a simple present tense which, unlike Russian, includes present tense for perfective verbs distinct from the future tense. The two remaining simple tenses are the imperfect and the aorist. Other tenses (future, past, pluperfect, conditional and past conditional) are compound and are formed using non-finite forms of verbs (participles and infinitives) and various kinds of enclitics[4]. This naturally leads to a different set of features associated with every ending.

The knowledge base for Serbo-Croatian does not contain the lexicon of suffixes (strings preceding endings), as in Serbo-Croatian participles and superlative adjectives are identified by their endings.

The aspectual system of Serbo-Croatian is similar to that of Russian. In terms of morphology, perfectives are typically derived from imperfectives by prefixation. This led to creating a lexicon of prefixes of perfective verbs. One more factor which affected the

4. The model recognizes simple tenses and constituents of compound tenses which are to be recognized as analytical forms at a higher (syntactic) level of analysis for which RDM is not responsible.

structure of the knowledge base for the Serbo-Croatian RDM was the availability of an on-line dictionary (which was not the case for Russian). This made it possible to greatly reduce the need for the lexicon of stems. For Russian, this lexicon was constructed to deal with ambiguity and, hence, with overgeneration. An on-line dictionary can be used for the same purpose, filtering out those readings in the output whose base forms have no match in the dictionary. In the RDM-based analyzer of Serbo-Croatian, the availability of an on-line dictionary allowed us to eliminate a) the lexicon of bodies, b) the largest component of the lexicon of stems, the one which includes stems of nouns and adjectives.

The knowledge base for the Serbo-Croatian RDM analyzer, thus, consists of an off the shelf online dictionary (20,000 entries) a lexicon of endings (about 900 entries), a lexicon of quasi-roots (about 4,000) entries, a lexicon of perfective verb prefixes (53 entries) and a lexicon of stems (about 1,000 entries) which only includes stems of bi-aspectual verbs and prefixless perfective verbs.

5.2. RDM Morphotactics for Serbo-Croatian

It should already be clear from the above that the RDM morphotactics for Serbo-Croatian can be simpler than that for Russian. The RDM morphotactics for Serbo-Croatian is, in fact, a subset of the Russian RDM morphotactics:

(18) word ::= stem ending;
stem ::= prefix string quasi-root ,

where "string" stands for any character string including the null string. For example, the analysis of the Serbo-Croatian word *izostali* ('fallen out', participle, active, past, perfective, masculine, plural) leads to the following assignments of morph types:

(19) word: *izostali* = stem: *izost* ending: *ali*
stem: *izostali* = prefix: *iz* string: Ø quasi-root: *ost* ,
where the quasi-root *ost* identifies a verb form, the prefix *iz*

signals perfective form of a verb and the ending *ali* identifies participle, past tense, masculine, plural.

5.3. Analysis in the RDM Model for Serbo-Croatian

The processing algorithm in the RDM for Serbo-Croatian, though consisting of the same main procedures, is much simpler than that for Russian. It involves predominantly look up (in the given order) in the lexicons of endings, quasi-roots, prefixes and stems. In the general case, the simplicity of Serbo-Croatian morphotactics and processing algorithms leads to much more significant base form[5] overgeneration than that for the Russian RDM model. This is dealt with by matching all output base form/part-of-speech pairs against the on-line dictionary. If there is no match, the reading is discarded.

For example, for the Serbo-Croatian word form *treba* (*trebati,* 'it is necessary') the RDM analyzer outputs the following intermediate readings:

(20) *treba*
treba (N1 N1 C1 G2)
treba (N1 N2 C2 G2)
treba (N7 N1 C1 G2)
treb (N15 N1 C2 G1)
treb (N15 N2 C2 G1)
treb (N24 N1 C2 G1)
treb (N24 N1 C4 G1)
trebati (V S1 T3 N1 P3)

Among the above are 3 nominal readings with the base form *treba*, 4 nominal readings with the base form *treb* and one verbal reading with the base form *trebati.* There are no dictionary matches

5. Overgeneration due to the homonymy of word forms in a language is to be dealt with at the level of syntactic analysis.

for the base forms of any nominal readings, so they are discarded. The verbal reading base form is found in the on-line dictionary and the system final output is

(21) *treba*
trebati (V S1 T3 N1 P3) verb, imperfective, present, singular, third person

In the case when no match is found for any of the output base forms, the input word is marked as "unknown." This is the price we must pay for the simplicity of the knowledge base and the processing procedure. An on-line dictionary imposes limitations on coverage and should be as large as possible. The advantage is that a dictionary is easy to incorporate into the system. Any off-the-shelf machine readable dictionary (MRD) can be used. This resource can be added or augmented once it is connected by non-specialists who have no knowledge of how the analyzer works. We have painlessly added a Russian-English MRD to our Russian analyzer. The dictionary is used to further reduce overgeneration and to look for translation equivalents.

6. RDM When an On-Line Dictionary is Available

The RDM approach allows us to minimize development effort even if an on-line dictionary is available. It is important to realize that it is impossible to generate a complete lexicon of traditional stems automatically from an on-line dictionary on the basis of traditional paradigms, because the stems thus derived do not cover alternations and suppletive forms. The knowledge to cover these phenomena must be acquired manually in any approach, and the problem of treating suppletive forms and alternations requires (as was pointed out by many developers, see, e.g., Mikheev and Liubushkina 1995) the delineation of sets of transformational classes of stems and development of complex control strategies for selecting the correct form. Our approach, in addition to minimizing the knowledge

acquisition effort, allows one to bypass the need to create such sets of transformational classes and control strategies by simplifying treating the alternations using non-traditional definitions of endings and (in rare cases) describing a word by two suppletive quasi-roots.

In our model, the lexicon of quasi-roots (in contrast to the lexicon of stems) can be constructed automatically provided the RDM paradigms have been derived manually. The assignment of paradigm numbers to the words can be done semi-automatically on the basis of the list of base form endings with subsequent checking by a human, for instance, to correct the numbers for irregular paradigms. The next step is then for each word to (automatically) chop off three characters preceding the ending. Inversely sorted word lists for every part of speech is a great help.

The steps of knowledge acquisition for Russian and Serbo-Croatian have been the same but while for Russian all the acquisition work was done manually for Serbo-Croatian the level of automation in knowledge acquisition was as follows: manual acquisition of the lexicon of RDM endings on the basis of a published grammar, automatic construction of individual lexicons for every part of speech from the on-line dictionary; manual extraction of perfective verb prefixes, automatic construction of inverse dictionaries (words are sorted in the alphabetical order of their last letters to cluster the words belonging to the same paradigms to further facilitate the assignment of paradigm numbers) for every part of speech, manual assignment of RDM paradigm numbers to the groups of words in the inverse dictionaries, automatic extraction of the quasi-root lexicon and automatic extraction of stems for bi-aspectual or prefixless perfective forms of verbs on the basis of RDM model endings.

7. Evaluation

The Russian morphological analyzer was tested on a randomly selected raw text from the *Moscow News* newspaper. The corpus comprised about 10,000 usages of 2,562 different lexemes. The results are as follows: for 2.6% of the word usages the RDM mod-

ule overgenerated, i.e. for these words incorrect readings were generated together with the correct ones; for 0.1% of the word usages the system undergenerated, which means that fewer than the full complement of the possible homographic readings was generated; for 1.1% of word usages no fully correct output was generated (in 0.8% of the cases both the base form and the features were incorrect; in 0.2% of the cases the system output an incorrect base form but correct feature sets; and in only 0.08% of the cases an incorrect base form but a correct part of speech feature was returned). We thus may claim that approximately 99% of the output is correct.

We were unable to directly compare these results with those of other Russian morphological analyzers by running them on our test text due to the unavailability of these analyzers. In the descriptions of most systems no evaluation results are given. The issue of evaluation is typically mentioned in passing, if at all. For example, the system described in Bolshakov (1993) is said not to provide enough morphosyntactic information about words and could not resolve rampant overgeneration.

The morphological analyzer of Segalovich (1995) is reported to be optimized for speed at the expense of accuracy and extensibility. Mikheev and Liubushkina (1995) mention that their system is implemented for 1,500,000 word-tokens; other publications did not mention the subject of evaluation at all (e.g., Malkov et al. 1983, Ashmanov 1995).

The Serbo-Croatian morphological analyzer underwent preliminary[6] testing on a raw corpus of about 10,000 words of newspaper text. More than 99% of the output is correct, the rest is divided between the cases of undergeneration (not all the legal readings are output) and failures to output any set of features for a word. The latter situation is explained by the unsuccessful matching against the on-line dictionary, which is clearly incomplete.

---

6. The Serbo-Croatian analyzer does not yet include a module for processing proper names, so they were not counted in the preliminary evaluation cited in the article. The module is under development.

8. Conclusion and Discussion

The objective of this work was to develop a methodology for a quick ramp-up morphological analyzer when the resources are scarce. The main distinguishing features of this approach are a) its lack of reliance on large traditional lexicons of stems, which is crucial in the absence of large on-line resources and b) a novel treatment of stem alternations. Both features allow for great economies in the development effort. Pursuing this objective, we originally created and implemented a rapid-development morphological analyzer, RDM, for Russian.

This model was successfully ported to Serbo-Croatian. Owing to the availability of an on-line dictionary of this language, several lexicons of the analyzer were constructed automatically after the RDM paradigms had been derived. The on-line dictionary and the fact that, unlike in Russian, some of the Serbo-Croatian word features are realized syntactically allowed us to exclude some of the lexicons used in the Russian RDM and to use a subset of the Russian RDM morphotactics as the Serbo-Croatian RDM morphotactics.

The RDM model is robust in that it can process unknown words. For example, a system based on the lexicon of stems of 100,000 words included in the Zaliznjak dictionary (Zaliznjak 1980) would fail to analyze such regular and frequent newly-coined Russian words as *diskoteka* 'disco', *evroremont* 'European style apartment remodelling', etc. Though in RDM the lexicon of quasi-roots has been also exracted using Zaliznjak (1980), RDM can process these words. The quasi-roots *tek* and *ont* are found in our lexicon of quasi-roots extracted, for example, from such words as *biblioteka* 'library' and *remont* 'repair'. The entire selection and compilation work on the RDM lexicons has been accomplished in about ten person-weeks.

Of course, we had to include additional lexicons to cover specific problems which the lexicon of quasi-roots could not solve, and we expect the lexicons to grow as we process the larger corpora. However, the sum total of effort for knowledge acquisition in

RDM has been negligible compared with that necessary for compiling a standard lexicon for morphological analysis. We fully understand that "standard" lexicons will be needed if this model is used as a component of a larger NLP system, such as a syntactic parser or semantic analyzer. However, the results of this work can be immediately incorporated in systems for tagging large corpora and other useful tasks.

In the immediate future we intend to add to this system some context-based rules for selecting from among the multiple outputs of the system.

References

Allen, J., M.S. Hunnicutt and D. Klatt. 1987. *"From Text To Speech: The MITalk System"*. Cambridge: Cambridge University Press.

Antworth, E. 1990. "PC-KIMMO: A Two-Level Processor For Morphological Analysis". *Occasional Publications In Academic Computing 16,* Summer Institute Of Linguistics. Dallas.

Apresyan, Ju.D., I.M. Boguslavsky, L.L. Iomdin, A.V. Lazursky, N.V. Pertsov, V.Z. Sannikov, L.L. Tsinman. 1989. *"Lingvisticheskoe obespechenie sistemy ETAP-2"*. Moscow: Nauka.

Ashmanov I. 1995. "Grammar and Style Checker for Russian Texts". A.Narinjani (ed.) *Proceedings of Dialog'95 International Workshop on Computational Linguistics and its Applications.* Kazan, Russia, 81-89.

Belonogov G. and Ju. Zelenkov. 1989. "An Algorithm for Morphological Analysis of Russian Words". *Voprosy Informatsionnoj Teorii i Praktiki.* 35-42.

Bolshakov I. 1993. "A Multifunctional Thesaurus for Computer Editing of Russian Texts". *NTI, s2. No. 7.* 24-26.

Koskenniemi, K. 1984. "A general computational model for word form recognition and production". *Proceedings of COLING-84.* Stanford. 178-181.

Malkov, M., Volkova I. and Gratsianova T. 1983. "Linguistic Processor TULIPS-2: Morphological Component". *Razrabotka i Primenenie Lingvisticheskih Protsessorov"* Novosibirsk, 34-42.

Mikheev A. and Liubushkina L. 1995. "Russian Morphology: An Engineering Approach". *Natural Language Engineering* 1(3). 1995. 235-260.

Pentheroudakis, J.E. and D.W. Higinbotham. 1991. "Morphogen: A Morphology Grammar Builder and Dictionary Interface Tool." *Proceedings of the 1991 Meeting of the Desert Language and Linguistics Society.* 42-51.

Ritchie, G.D., A.W. Black, S.G. Pulman and G.J. Russell. "The Edinburgh/Cambridge Morphological Analyzer and Dictionary System. Version 3.0.". Unpublished manuscript.

Segalovich, I. S. 1995. "Indexing of Large Russian Texts with a Dictionary Built around the Sparse Hash Table". A.Narinjani (ed.) *Proceedings of Dialog'95 International workshop on Computational Linguistics and its Applications.* Kazan, Russia.

Wilks, Y. 1996. "Guest Editor's Introduction". Y.Wilks (ed.) *Special Issue on NLP, Communications of ACM.* 39:1, 62.

Zaliznjak, A.A. 1980. "Grammaticheskij slovar' russkogo jazyka". Moscow: Russkij Jazyk.

New Mexico State University
Las Cruces, NM 88003-8001
lana@crl.nmsu.edu
sergei@crl.nmsu.edu

Multiple Sonority Thresholds*

Draga Zec
Cornell University

1. Introduction

The central claim of this paper is that prosodic constituency is organized as a prominence hierarchy. This, I will argue, is reflected in sonority threshold restrictions associated with each level of the prosodic hierarchy – the mora, the syllable, the foot, and possibly the prosodic word. Given the prosodic constituency in (1) (cf. Nespor and Vogel, 1986; McCarthy and Prince, 1986, among others), it will be shown that each constituent within the hierarchy establishes a direct relation with the level of segments by imposing a minimal sonority threshold:

(1)		
	ω	prosodic word
	ϕ	foot
	σ	syllable
	μ	mora
s s s s		segments

Taking Optimality Theory as the frame of reference (Prince and Smolensky, 1993; McCarthy and Prince, 1993a, 1995), I will first demonstrate that each prosodic level is indeed associated with a sonority threshold restriction, basing the argument, primarily, on the complex phonological alternations in Old Church Slavonic. Second, I will address the issue of the mutual relations among sonority thresholds at different prosodic levels, and will propose a

* I am grateful to Niken Adisasmito Smith, Christina Bethin, Wayles Browne, Abby Cohn, Lisa Lavoie, Marek Przezdziecki, Tracy King, and Young-mee Yu Cho, for invaluable comments and suggestions.

mechanism for expressing what is immutable in this relation. Finally, I discuss cases of sonority threshold lowering, and show that they are readily accounted for within the general framework developed in this work.

In order to illustrate multiple sonority thresholds imposed by the prosodic hierarchy, I will turn to a well-studied case: English possesses three distinct sonority thresholds, each associated with a distinct prosodic level, as in (2):[1]

(2) **English**: *sonority thresholds imposed by the prosodic hierarchy*

ϕ	foot	vowels and *r*	[–consonantal]
σ	syllable	vowels, liquids and nasals	[+sonorant]
μ	mora	all segments	——

No specific restriction is imposed at the lowest prosodic level. The sonority threshold at the level of the mora admits any member of the English segment inventory: any vowel or consonant may occupy the weight-bearing position within the syllable, which is structurally associated with the position dominated by the second mora of a heavy syllable. As a result, all CVC and CVV syllables in English are heavy, that is, bimoraic.[2] The level of the syllable is more restrictive, imposing a sonority threshold which admits only sonorant segments, vowels as well as consonants. This is illustrated by disyllabic forms such as *butter, murky, bottle, cuddle,* whose first syllable nucleus corresponds to a vowel, and second to a liquid; and by forms like *sudden, button, sedentary*, which, in addition, also admit nasals as syllabic segments. In sum, any segment belonging to the [+sonorant] natural class may act as syllabic in English.

---

[1] The sonority scale assumed here is: vowels, liquids, nasals, obstruents. This scale will be further elaborated as the cases to be presented here call for it.

[2] Evidence for this comes in part from stress: in nouns, stress falls on the penultimate syllable if it is heavy, as in *agénda, Arizóna,* and on the antipenult, if the penult is light, as in *lábyrinth* (Hayes, 1982).

That the weight-bearing and nuclear properties of segments are associated with sonority thresholds imposed by the mora and the syllable, respectively, has already been noted in the literature (Zec, 1995 and the references therein). But it has not been noted that the level of the foot also imposes a sonority threshold. In English, crucial evidence for this comes from two classes of syllable nuclei: while any [+sonorant] segment in English may act as a syllable nucleus, we can clearly distinguish between the distribution of those syllables whose nuclei correspond to *l* or a nasal, and those whose nuclei are either a vowel or *r*. The former class of syllables has a severely restricted distribution, as stated in (3):

(3) *Distribution of syllables with l or a nasal in the nucleus* (L = *l* or nasal):
   a. CL and CLC syllables are never stressed
   b. There are no monosyllabic CLC words, or disyllabic CLCL words
   c. CLC syllables are restricted to positions within the word inflicted by "extraprosodicity" (*student, prudent*)

The restricted distribution in (3) follows from the metrical organization of English, which has the standard system of trochaic feet, with the foot inventory corresponding to two light or one heavy syllable. The collocations in (3), from which syllables with a nasal or *l* in the nucleus are excluded, are precisely those in which they would also act as foot nuclei. First, because stressed syllables have to be foot heads, a CL(C) syllable may not bear stress, as stated in (3a). Next, a prosodic word must, minimally, contain a foot, yet a syllable with a nasal or *l* in the nucleus is incapable of sustaining a foot, which accounts for the absence of CLC monosyllables or CLCL disyllables, stated in (3b). Finally, CLC syllables, which like other heavy syllables should correspond to feet, occur in only those positions in which they cannot be granted foot status, that is, in positions inflicted by extraprosodicity, as stated in (3c) (see Hayes, 1982).

No such restrictions are associated with syllables headed by a vowel or *r*, as stated in (4), since both these segments meet the sonority threshold imposed by the foot.

(4) *Distribution of syllables with r in the nucleus (CR(C)), which is identical with the distribution of CV(C) syllables*:
   a. (CR(C)) syllables may be stressed, as in *murky*, *covert*, *furnace;*
   b. CR and CRC may occur in monosyllables: *fur*; *turf*, *curl*;
   c. Disyllables of the CRCL shape are possible, as in: *curtain, curdle.*

In other words, foot nuclei in English may not correspond to segments that are less sonorous than vowels and *r*. But, do *r* and vowels, the only foot-bearing syllable nuclei, form a natural class? Kahn (1978) shows that *r* and *l* pattern differently with respect to flapping and release, the former patterning with glides and the latter with obstruents.[3] Therefore, in the spirit of Kahn's (1978) proposal, I assume that *r* is a [–consonantal] segment, and that it exhibits both glide-like and vowel-like behavior. This natural class demarcates the class of foot-bearing segments in English, which corresponds to the [–consonantal] set, and is more restrictive than the [+sonorant] set which captures the class of English syllabic segments.

We now turn to the formal part of our proposal. Sonority thresholds associated with constituents in the prosodic hierarchy will be captured by positing the **SON**(ority) family of constraints, with a separate constraint associated with each prosodic level, as in (5):

---

[3] According to Kahn (1978), flapping occurs after vowels, glides and *r*, as in *potting*; *loiter* and *shouting*; *forty* and *parting*. Flapping does not occur after *l* and *n*: **malted*, **shelter*, **winter*. (We are not considering here the flapped variant of *winter*, in which *n* is either shortened or completely elided.) Also, voiceless stops in pre-pausal position are unreleased after vowels and glides (*sit*, *height*), and after *r* (*heart*), but are released after obstruents and *l* (*apt*, *list*, *belt*).

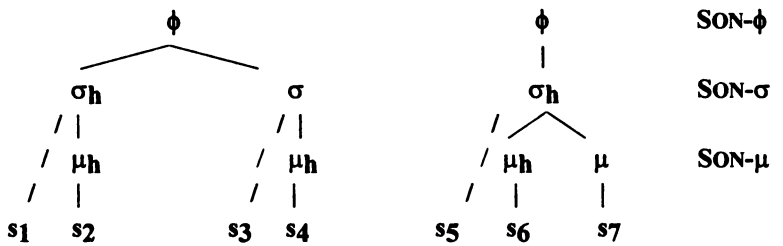
(5) *The SON(ority) family of constraints*:
SON-μ
SON-σ
SON-ϕ

In (6) is given the instantiation of **SON**(ority) constraints in English. As will be shown in Section 3, the mutual relation among the **SON** constraints within any language is fixed with **SON-ϕ** being more restrictive than **SON-σ**, which in turn is more restrictive than **SON-μ**.

(6) *English*
SON-μ admits all segments as weight-bearing
SON-σ[+sonorant]
SON-ϕ[–consonantal]

For prosodic constituents to establish a link with the level of segments, and impose their sonority thresholds, the following crucial assumption needs to be made: each prosodic constituent possesses a head, corresponding to one constituent at the next lower level (marked by the *h* subscript), as in (7). It is by virtue of the head relation that members of the prosodic hierarchy "communicate" with the segmental level. The **SON** constraint at each prosodic level imposes a sonority threshold on the segment at the bottom of its "head" path. In (7), **s7** is under the jurisdiction of **SON-μ**, **s4** is under the jurisdiction of **SON-μ** and **SON-σ**, and **s2** and **s6** are under the jurisdiction of all three sonority threshold constraints.

(7) a. b.



SON-μ, **SON-σ** and **SON-ϕ** are thus responsible for the minimal sonority threshold imposed, respectively, by the mora, the syllable, and the foot.

In sum, prosodic constituency is viewed here as a hierarchy of sonority peaks. It is regulated by several principles which, if viewed as constraints, will have to be undominated constraints. One is the principle of strict layering (Selkirk, 1984, 1995), which excludes any dominance relations other than those in (1), and additionally, calls for the postulate stated in (8), whereby each prosodic constituent possesses one subconstituent marked as the head, i.e., as its most perspicuous element (Itô and Mester, 1993; Zec, 1988; Selkirk, 1995).

(8) **HEADEDNESS**:
A prosodic constituent must contain a head, i.e., constituent *n* must immediately dominate exactly one constituent *n-1* designated as its most prominent element, and marked by the *h* subscript. ("No skipping of prosodic levels")

The headedness relation, captured in (8), has already been invoked in (7) above, where it plays a crucial role in characterizing the hierarchy of prosodic constituents as a prominence hierarchy.

Thus far, we have illustrated how multiple sonority thresholds are manifested in English, and proposed a general framework for capturing this phenomenon. In the following section we turn to Old Church Slavonic (henceforth OCS), a language which provides ample evidence for multiple sonority thresholds.

2. Case Study: Old Church Slavonic (OCS)

OCS exhibits several phonological interactions directly related to sonority threshold constraints.[4] Before turning to these alternations, we will first look into the general phonological properties of OCS and their relevance for multiple sonority thresholds.

Listed in (9) and (10) are the vocalic segments of OCS. The high lax vowels ***I*** and ***U*** in (9) are the so-called jer vowels. Following Lightner (1972), we distinguish them from the other OCS vowels by designating them as [–tense] segments, the other vowels in the language being [+tense].

(9) *Historically short vowels*:

I	**U**	high	**jers:** [–tense]
e	o	non-high	

The vowels in (9) correspond to short vowels in Common Slavic, while those in (10) correspond to what clearly used to be long vowels. It is not obvious, however, that vocalic length was phonologically relevant in the OCS dialect described here. For reasons of accuracy, we state that the vowels in (9) historically correspond to short, and those in (10), to long Common Slavic vowels. But since we have no clear evidence regarding the status of

[4] Old Church Slavonic is the language of the earliest Slavic documents translated by Cyril and Methodius, and their disciples, between the 9th and 11th centuries and, most likely, spoken in the region of Thessaloniki (Lunt, 1959 and the references therein). The documents that this analysis is based on are Codex Zographensis, a 12th century copy of an early manuscript, which according to Lunt (1959:6) is "phonetically... nearest to the language of Cyril and Methodius;" and Ostromirovo Evangelie, which dates from 1056, and possesses certain East Slavic traits. I am grateful to Wayles Browne for invaluable help with the OCS data.

vocalic length at the stage we focus on, the analysis is free of any assumptions in this regard. [5]

(10) *Historically long vowels*:

i	y	u	high
ę		õ	non-high
ě	a		

A note is in order regarding special qualities of some of the historically long vowels in (10): ę and õ are nasal vowels, and ě is an open mid vowel which may have been a diphthong in some dialects.

The system of OCS consonants includes, standardly, obstruents (stops, fricatives, and affricates) and sonorants (nasals and liquids). It is characterized by an abundance of palatal, or palatalized, consonants. Most importantly, consonants occupy only the onset position within the syllable, with the exception of liquids which, in certain dialects, may appear in the coda. This is due to the open syllable strategy in Common Slavic, which led to the elimination of most coda consonants in Slavic languages. (The nasal vowels ę and õ in (10) result from the loss of coda nasals.)

OCS exhibits a distinct **SON** constraint at each level of the prosodic hierarchy. Starting with the level of the mora, we note that, in addition to vowels, liquids also belong to the set of weight-bearing segments. Evidence for this comes in part from Late Common Slavic which, due to a series of processes whose effect was to eliminate closed syllables, inherited only one class of such syllables, those closed with a liquid, as noted above. In Late Common Slavic, syllables closed with a liquid were transformed in various ways, strongly suggesting their moraic status (see Shevelov, 1965; Bethin, 1998, and the references therein). South Slavic was characterized by the process known as *liquid*

[5] Vocalic length is entirely absent from the modern counterparts of the dialects which serve as basis of our analysis – Russian as well as Bulgarian and Macedonian.

metathesis: in a closed syllable, a vowel-liquid sequence was transformed, by virtue of metathesis, into a liquid-vowel sequence. This process was characterized, further, by the concomitant lengthening of the vowel, which can be interpreted as compensatory in nature, and as such, directly related to the mora associated with the liquid. In Old Russian, syllables closed with a liquid were subject to the process known as *pleophony*: closed syllables were eliminated by creating a vowel copy, that is, by converting a CVL sequence into CVLV. Only one class of syllables closed with liquids, those with a jer vowel in the nucleus, survived until later historical stages. Those syllables acted as heavy, as we will argue below; the liquid was thus clearly moraic.[6] In sum, the following constraint imposes the sonority restrictions on the set of weight-bearing segments in OCS:

(11) OCS: **SON-μ**[+sonorant, –nasal]

Turning to the next higher level of the prosodic hierarchy, the syllable, we note that any vowel listed in (9) and (10) may act as the syllable nucleus (although in South Slavic liquids may also be interpreted as syllabic). The relevant **SON** constraint is

(12) OCS: **SON-σ**[–consonantal]

In addition to the mora and the syllable, the foot in OCS imposes a distinct sonority threshold as well. This will be demostrated by focusing our analysis on the phonological alternations exhibited by the jer vowels, known as *Havlik's Law*, which are to be captured in terms of foot structure. The proposed analysis accounts both for cases that fall under Havlik's Law, and those that are seen as its exceptions.

---

[6] The facts about the surviving closed syllables, as well as the evidence regarding their weight, are presented later in this section.

The jer vowels exhibited a characteristic pattern of lowering in certain collocations, a phenomenon, known as Havlík's Law (based on Isačenko, 1970):

(13) Havlík's Law:

a. Word-final jers and jers in syllables followed by vowels other than jers become "weak" ("weak" jers are subsequently dropped);
b. Jers in syllables followed by a "weak" jer become "strong"
c. "Strong" jers ***I*** and ***U*** merge with ***e*** and ***o*** respectively.

To paraphrase, jers in so-called "strong" positions lower to the corresponding non-high vowels, while those in "weak" positions remain unaltered.[7] Crucial for our argument is the fact that jers in "strong" positions are immediately followed by another jer, which strongly suggests that this phenomenon is related to foot structure.

The examples in (14) – (16) illustrate the alternations described in Havlik's Law.[8] We first focus on the behavior of jers in disyllabic forms, given in (14), which demonstrate that the jer vowels are subject to lowering only when the underlying form contains consecutive jers, as in (14c). In the other cases, namely, (14a) and (14b), in which the jer is either followed or preceded by a "regular" vowel, jer lowering does not take effect.

---

[7] This is the case at the historical stage we are capturing here (represented in Zographensis and Ostromirovo Evangelie, see Footnote 4). At a later historical stage, jers in "weak" positions were lost, but this aspect of Havlík's Law is outside the scope of this paper (cf. (19a)).

[8] In the schematized representations of OCS examples, we use the following abbreviations: c = any consonant, v = any non-jer vowel, U = jer vowel.

(14) Forms with two syllables:

underlying	surface	
a. **cvcU**		
stolU	stolU	'throne.NOM'
b. **cUcv**		
sUnu	sUnu	'dream.DAT'
dIne	dIne	'day.GEN.SG'
c. **cUcU**		
sUnU	sonU	'dream.NOM'
dInI	denI	'day.NOM.SG'

This strongly suggests that a cUcU sequence is footed as (cUcU). Because it is the leftmost jer that lowers, yielding (cocU), we conclude that the foot system is trochaic, and that the leftmost syllable within the foot is subject to **SON-ϕ**, whose effect is manifested as jer lowering.

The trisyllabic forms in (15) further corroborate what we already observed in (14). Forms which contain only one jer vowel, as in (15a), (15b), and (15c), or those that contain jers interspersed with non-jer vowels, as in (15d), do not exhibit any alternations. Alternations are found in only those forms that underlyingly contain consecutive jers. Those in (15e) and (15f), each containing two consecutive jers, are footed as cv(cUcU) and (cUcU)cv respectively, with the leftmost jer lowering. The crucial case is (15g), with three consecutive jer vowels, which clearly shows that the directionality of footing is right-to-left: a cUcUcU form is footed as cU(cUcU), yielding cU(cocU), rather than *(cUcU)cU, yielding *(cocU)cU.

(15) Forms with three syllables:

underlying	surface	
a. **cUcvcv**		
vUzglasi	vUzglasi	'announce'
b. **cvcUcv**		
tetUka	tetUka	'aunt.NOM.SG'
otIca	otIca	'father.GEN.SG'
c. **cvcvcU**		
bolitU	bolitU	'ache.3PER.SG.PRES'
d. **cUcvcU**		
mUnogU	mUnogU	'many'
e. **cvcUcU**		
otIcI	otecI	'father.NOM.SG'
rabU tU	rabo tU	'this slave'
f. **cUcUcv**		
čItIca	četIca	'reader.GEN.SG'
tImIna	temIna	'dark.NOM.SG.FEM'
sUnIma	sonIma	'gathering.GEN.SG'
vU tIme	vo tIme	'in the darkness'
g. **cUcUcU**		
čItIcI	čItecI	'reader. NOM.SG'
tImInU	tImenU	'dark.NOM.SG.MASC'
sUnImU	sUnemU	'gathering.NOM.SG'

In (16) are listed forms containing four syllables. Again, those forms that contain no consecutive jers, as in (16a) – (16e), do not exhibit any alternations. Forms that do contain consecutive jers, those in (16f), (16g), (16h) and (16i), exhibit jer lowering. All cases of jer lowering are accounted for under the assumption that footing proceeds from right to left.

(16) Forms with four syllables:

underlying	surface	
a. **cvcUcvcU**		
žitInicU	žitInicU	'granary.GEN.PL'
b. **cvcUcvcv**		
žitInica	žitInica	'granary.NOM.SG'
c. **cUcvcUcv**		
mUnožIstvo	mUnožIstvo	'multitude.NOM.SG'
d. **cUcvcvcU**		
vUzglasitU	vUzglasitU	'announce.3PER.SG.PRES'
e. **cUcvcvcv**		
vUzglasite	vUzglasite	'announce.2PER.PL.PRES'
f. **cUcUcvcU**		
vUskrIsnetU	voskrIsnetU	'resurrect.3PER.SG.PRES'
g. **cUcUcvcv**		
sU mUnojo	so mUnojo	'with me'
h. **cUcUcUcv**		
pIpIrIca	pIperIca	'pepper.GEN.PL'
vU sUnIme	vU sonIme	'in (the) gathering'
i. **cUcUcUcU**		
pIpIrIcI	pepIrecI	'pepper.NOM.SG'
vU sUnImU	vo sUnemU	'to (the) gathering'

To conclude, the distribution of "strong" and "weak" jers results from binary footing, assuming the system of moraic trochees. This is captured by the following foot-related Optimality Theory constraints (following McCarthy and Prince, 1993a, 1993b, 1995; Prince 1990, Prince and Smolensky, 1993):

(17) **FTBIN:** Feet are binary.

(18) **ALIGN-R:** Align every foot with the right edge of the prosodic word.

While **FTBIN** ensures that feet are maximally, and minimally, bimoraic, the alignment constraint in (18) mimics the effect of

right-to-left footing, guaranteeing that a cUcUcU sequence becomes cU(cvcU) rather than *(cvcU)cU.

Jer lowering is explained as a sonority threshold restriction on feet in OCS: the vowel heading the foot may not be a jer, that is, has to belong to the class of [–consonantal, +tense] segments. Jers are thus sufficiently sonorous to be syllabic, but are insufficiently sonorous to be foot-bearing, as captured by the following constrains:

(19) OCS:

a. minimal foot sonority: **SON-ϕ**[–consonantal, +tense]
b. minimal syllable sonority: **SON-σ** [–consonantal]

We now turn to the complexities of OCS foot structure, which is characterized by opaque constraint interactions. **SON-ϕ** plays a crucial role in evaluating optimal foot structure, but for this purpose, crucially, dominates **FAITH**. The evaluation of optimal foot nuclei, however, calls for the reversed ranking, with **FAITH** dominating **SON-ϕ** (to allow for jer lowering).

Opaque constraint interactions in OCS will be resolved by positing two phonological levels: Level 1, at which **FAITH** dominates **SON-ϕ**; and Level 2, at which this dominance relation is reversed.[9] This move is necessary because concomitant jer lowering and footing would lead to undesirable results. At Level 1, at which all footing takes place, no jers are lowered due to the ranking **FAITH** >> **SON-ϕ**. Tableau in (20) shows that satisfying **ALIGN-R** is more important than selecting a foot nucleus which meets the **SON-ϕ** requirement.

---

[9] We would avoid introducing multiple levels by invoking a sympathy-based evaluation in addition to input-output evaluation, following McCarthy's (1998) proposal. We opt for a level-based analysis in order to keep the focus on the central concern of this study; the question of the status of levels in phonology is outside its scope.

(20) **Level 1: ALIGN-R >> SON-ϕ**

CvcUcU	ALIGN-R	SON-ϕ
☛ cv(cUcU)		*
(cvcU)cU	*!	

At this point we introduce an additional foot-related constraint, **FTFORM** in (21), which requires that trochaic feet containing a jer and a non-jer vowel may not be headed by the jer. As a result, cUcUcv is footed as (cUcU)cv rather than *cU(cUcv).

(21) **FTFORM:** Feet of the (cUcv) shape are prohibited.

Tableau (22) provides evidence that **FTFORM** dominates **ALIGN-R**, and Tableau (23) shows that this constraint also dominates **FTBIN**.[10]

[10] Crucially, **FTFORM** takes effect only if **FAITH** >> **SON-ϕ** [–consonantal, +tense], thus preventing jer lowering. Under the reversed ranking, **SON-ϕ** >>**FAITH**, the input cv(cUcv) in (22) would yield *cv(cvcv) as the optimal ouput; and the input cUcv in (29) would yield *(cvcv) as optimal. This provides pivotal evidence that the evaluation of OCS forms needs to proceed in two steps.

(22) **Level 1: FTFORM >> ALIGN-R**

CvcUcv	FTFORM	ALIGN-R
☛ (cvcU)cv		*
cv(cUcv)	*!	

(23) **Level 1: FTFORM >> FTBIN**

cUcv	FTFORM	FTBIN
☛ cU(cv)		*
(cUcv)	*!	

The optimal candidate in (23) contains a monomoraic foot, in violation of **FTBIN**. This analysis, which admits monomoraic feet in OCS, is supported by the forms in (24): a jer found in monosyllabic, and monomoraic, forms is subject to lowering, and thus constitutes a foot head.

(24) Monosyllables:

a. **cv**

to	to	'that'

b. **cU**

nU	no	'but'

In (24b), **FTBIN** is violated under the pressure of the constraint in (25), which requires that each prosodic word possesses a foot.

(25) **HEAD-ω:** A prosodic word must have a head.

Because a prosodic word must contain a foot, the footing of a monomoraic form proceeds in violation of **FTBIN**, as shown in (26).

(26) **Level 1: HEAD-ω >> FTBIN**

cU	**HEAD-ω**	**FTBIN**
☛ (cU)		*
cU	*!	

We now turn to the interactions at Level 2. Once all the footing is in place, the jer vowels that occupy the positions of foot nuclei are converted to the corresponding non-jer vowels, in order to meet the requirements of **SON-ϕ**. In other words, the pattern of jer lowering in strong positions calls for **SON-ϕ >> FAITH**. The input form in (27), which is a level 2 interaction, is the output from Level 1.

(27) **Level 2: SON-ϕ >> FAITH**

cv(cUcU)	**SON-ϕ**	**FAITH**
☛ cv(cocU)		*
cv(cUcU)	*!	

Thus far, we have only documented disyllabic feet, those in (7a) above, and this is precisely what Havlík's Law focuses on, if interpreted in terms of foot structure. An obvious further line of investigation is to look for monosyllabic feet, those in (7b), which correspond to a single heavy syllable. This brings us to the so-called exceptions to Havlík's Law: in *Ostromirovo Evangelie*, an OCS manuscript which exhibits certain East Slavic traits, jers in syllables closed with liquids do not follow the "strong"/"weak" pattern predicted by Havlík's Law. According to Kiparsky

(1979:100), "...U and I in the groups tUrt, tIrt, tUlt were always treated as if 'strong'."

Late Common Slavic is characterized by a strong tendency towards open syllables. By the time of the earliest written records, most closed syllables had been eliminated from the language. No closed syllables are documented in South Slavic, and only one type is documented in Late Common Slavic. The phenomenon known as pleophony is part of the general open syllable strategy in Late Common Slavic, and is characteristic of East Slavic (cf. Bethin, 1998). As shown in (28), all syllables closed with liquids other than those with a jer nucleus were eliminated by copying the nuclear vowel.

(28)		oLC	→	oLoC	
		eLC	→	eLeC	
	but	ULC	→	ULC	(with ULUC in certain dialects)

As a result, CUL is the only closed syllable in the East Slavic dialect of OCS, and the only heavy syllable headed by jer. This closed syllable, being bimoraic, forms a trochaic foot.

And indeed, in later East Slavic manuscripts, jers in syllables closed with liquids are lowered (Borkovskij and Kuznecov, 1965; Kiparsky, 1979; Shevelov, 1965; Bethin, 1998). In (29) – (31) we compare the forms found in *Ostromirovo Evangelie*, from the eleventh century, and their correspondents in thirteenth century manuscripts:

		11th c. (Ostr)	13th c.	
(29)	a.	tUrgU	torgU	'square'
	b.	sUmIrtI	smertI	'death
	c.	mUlnIji	molnIja	'lightening'
(30)	a.	tUrgovati	torgovati	'trade'
	b.	vIrba	verba	'willow'
	c.	gUrlo	gorlo	'throat'
	d.	vUlna	volna	'wave'

(31) a. gUrnIcI gornecI 'jar'

All the forms in the left-hand column contain syllables closed with a liquid whose nucleus is a jer vowel. In the thirteenth century forms, the jer lowers in (29), where there is a "weak" jer in the immediately following syllable; in (30), where the lowered jer is the sole jer in the word; and in (31), where there is a "strong" jer in the following syllable. Thus, jer lowering in closed syllables does not fall under Havlík's Law, which is why this case had been treated as an exception.

But under the account of jer lowering proposed here, we can explain why a jer in a CUL syllable behaves like a "strong" jer: because it occupies the head position within a foot. We also explain why the jer in a CUL syllable does not adhere to Havlík's Law: this law refers to jers in monomoraic syllables, but not to those in bimoraic syllables, that is, to disyllabic, but not to monosyllabic, feet. The proposed analysis thus provides a unified account, covering both the cases that fall under Havlik's Law, and those that are seen as its exceptions: we expect jer lowering in all types of trochaic feet, both monosyllabic and disyllabic.

3. Mutual Relation Among Sonority Thresholds

Now that we have established the need for multiple **SON** constraints within any given language, we turn to the issue of their mutual relation. What we note, focusing on individual cases, is that the **SON** constraints exhibit a stable relation: **SON-σ** is either equally or more restrictive than **SON-μ**, and **SON-ϕ** is either equally or more restrictive than **SON-σ**. This is what we see in the cases listed in (32). In English and OCS, **SON-ϕ** is more restrictive than **SON-σ**, which in turn is more restrictive than **SON-μ**. But in the next three cases, two of the three **SON** constraints pick the same

sonority classes: SON-ϕ and SON-σ in Serbian and Piro, and SON-σ and SON-μ in Mordwin.[11]

(32)

English:	SON-ϕ[-cons]	SON-σ[+son]	SON-μ [unrestricted]
OCS:	SON-ϕ [-cons, +tense]	SON-σ[-cons]	SON-μ[+son]
Serbian:	SON-ϕ [+son, +cont]	SON-σ [+son, +cont]	SON-μ[+son]
Piro:	SON-ϕ[-cons	SON-σ[-cons]	SON-μ[+son]
Mordwin:	SON-ϕ [-cons, -high]	SON-σ[-cons]	SON-μ[-cons]

The stable relation among the SON constraints within any given language is expressed in (33), as a principle of *Prosodic Peak Transparency*. By virtue of this principle, the association of sonority classes with SON constraints in (34) constitutes an impossible case, and as such needs to be excluded on a formal ground.

(33) **Prosodic Peak Transparency:** the sonority threshold associated with prosodic constituent *n* may not be less restrictive than the sonority threshold associated with its head.

(34) **Impossible case:**
SON-ϕ[– cons]; SON-σ[+son]; SON-μ[– cons]

But a grammar built out of a set of constraints, and their mutual ranking, is not capable of achieving the effect of Prosodic Peak

[11] The relevant references are: for Piro, Zec (1998a) and the references therein; for Mordwin, Kenstowicz (1994) and Zec (1998b); and for Serbian, Zec (1988).

Transparency.[12] In other words, under the view characteristic of OT, that the interaction among constraints is free, we encounter the following problem: how is the principle in (33) to be implemented, that is, on what ground can the impossible case in (34) be excluded?

The formal device we will invoke to achieve the effect of (33) is the mechanism of local constraint conjunction, proposed in Smolensky (1995), and further elaborated in Itô and Mester (1996) and Hewitt and Crowhurst (1996). The gist of the proposal is that constraints belonging to a local domain may be conjoined to form a "macro" constraint, as stated in (35a); that the observance of a "macro" constraint is computed from the observance pattern of the participating constraints, as in (35b); and that, by stipulation, the "macro" constraint always ranks higher than any of its constituent parts, as in (35c).

(35) **Local Conjunction of Constraints**
(Smolensky, 1995; Itô and Mester, 1996):

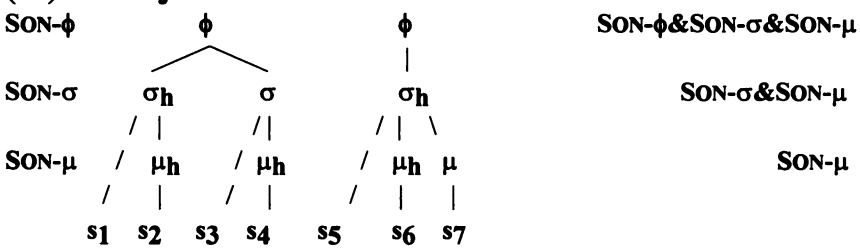
a. "Macro" constraint generation: if **P** and **Q** are members of the constraint set CON, so is the derived constraint $\mathbf{P\&_{l}Q}$ (i.e., **P** locally conjoined with **Q**);
b. Interpretation: $\mathbf{P\&_{l}Q}$ is violated if and only if there is some domain **D** in which either **P** or **Q** is violated (following Hewitt and Crowhurst, 1996);
c. Ranking (universal): $\mathbf{P\&_{l}Q >> P, Q}$.

My proposal is that the set of **SON** constraints is subject to constraint conjunction. In addition to "regular" **SON** constraints, the grammar also includes "macro" **SON** constraints, which are the product of local conjunction. At each level of the prosodic hierarchy, the sonority of segments at the bottom of the head path is regulated not only by the "regular" **SON** constraints, but also by

[12] Note also that the association of sonority classes with any of the **SON** constraints is best assumed to be random, and thus not regulated by any mechanism in the grammar.

conjoined, "macro" SON constraints, as shown in (36). The "macro" constraints result from conjoining "regular" SON constraints, in the following fashion: a SON constraint at each level is conjoined with the SON constraints at all lower prosodic levels. Thus, at the level of the syllable, in addition to SON-σ, operative also is the conjoined constraint SON-σ & SON-μ, while at the level of the foot, in addition to SON-ϕ, there is also the "macro" constraint resulting from conjoining all three SON constraints, SON-ϕ & SON-σ & SON-μ.

(36) **Conjunction of SON Constraints**



The ranking of the SON constraints, both "regular" and "macro," follows the ranking universal in (35c) above, whereby a "macro" constraint always ranks higher than any of its conjunct constraints:

(37) **Ranking:**

SON-ϕ & SON-σ & SON-μ >>
SON-σ & SON-μ >>
SON-ϕ, SON-σ, SON-μ

The conjunction of SON constraints creates local domains in which all SON constraints present in the domain jointly impose their requirements. This creates a cumulative effect whereby all locally relevant sonority requirements are imposed simultaneously. This enriched set of SON constraints is capable of achieving the effect of Prosodic Peak Transparency in (33), as well as of excluding the impossible case in (34).

Tableaux in (38) and (40) show how the presence of conjoined constraints in the grammar ensures the validity of Prosodic Peak Transparency. It is due to the "macro" SON constraints that the situation in (34) is interpreted as indeed impossible. Crucial at this point is the interaction with FAITH. We will present two relevant cases, one with FAITH dominating, and the other with FAITH dominated by, the set of SON constraints. The first case, with FAITH as the dominant constraint, is presented in Tableau (38). Given the hypothetical input form *tnta*, the winning candidate satisfies FAITH, at the cost of violating SON-σ & SON-μ as well as the lower ranked SON-μ[-cons]. The only SON constraint obeyed by the winning candidate is SON-σ[+son].

(38) Exclusion of the impossible case:
SON-σ[+son], SON-μ[– cons], with FAITH >> SON

tnta	FAITH	SON-σ & SON-μ	SON-σ[+son]	SON-μ[– cons]
☞ tnta		*		*
tata	*!			

The effect in (38) is achieved as long as FAITH ranks above the highest SON constraint, which by virtue of (37) is the highest relevant "macro" constraint.[13]

The effect of constraint interaction in (38) is a lowered sonority threshold imposed by the mora. The output form *tnta* violates SON-μ[–consonantal] while satisfying SON-σ[+sonorant]. In other words, with FAITH as the dominant constraint, the impossible case in (34) is interpreted as, and is thus indistinguishable from, the case in (39), which is in full compliance with Prosodic Peak Transparancy:

---

[13] Note that (38) would yield the same result with FAITH ranking above SON-ϕ & SON-σ & SON-μ, which would be violated due to the SON-μ violation.

(39) **SON-ϕ[– cons]; SON-σ[+son]; SON-μ[+son]**

Under the ranking **SON >> FAITH**, the hypothetical input form *tnta* yields as optimal the output form *tata*, which violates **FAITH**, but satisfies all **SON** constraints.[14]

(40) Exclusion of the impossible case:
SON-σ[+son], SON-μ[– cons], with SON >> FAITH

tnta	SON-σ & SON-μ	SON-σ[+son]	SON-μ[– con]	FAITH
☞ tata				*
tnta	*!		*	

This time, constraint interactions result in a raised sonority threshold. Due to the dominated status of **FAITH**, the optimal output form *tata*, while (minimally) departing from the segmental setup of the input *tnta*, meets both **SON-σ**[+son] and the more restrictive **SON-μ**[–con], and thus also satisfies the "macro" constraint **SON-σ**[+son] **& SON-μ**[–con].

Given the interactions in (40), with **FAITH** as a dominated constraint, the impossible case in (34) is indistinguishable from the following case:

(41) **SON-ϕ**[–cons]; **SON-σ**[–cons]; **SON-μ**[+son]

[14] It is sufficient for **FAITH** to rank below *any* **SON** constraint for achieving the effect in (40). If it ranks below the highest ranked **SON** constraint, the "macro" constraint **SON-ϕ & SON-σ & SON-μ** (see (37)), the effect is the same as in (40), since this constraint is violated due to the **SON-μ** violation. And this effect is, of course, preserved under any lower ranking.

Again, the case in (34), which fails to comply with Prosodic Peak Transparency, is interpreted as a case in full compliance with this principle.

We have seen that, by virtue of SON constraint conjunction, an impossible case such as (34) is interpreted as a standard case. To summarize, what needs to be resolved in (34) is the mutual relation between SON-σ and SON-μ: the latter may not be more restrictive, as is the case in (34). Under the FAITH >> SON ranking, the impossible case is resolved by lowering the sonority threshold imposed by the mora, with SON-μ[–consonantal] interpreted as SON-μ[+sonorant], as in (42):

(42) FAITH >> SON
* SON-ϕ[–cons]; SON-σ[+son]; SON-μ[–cons]
interpreted as
SON-ϕ[–cons]; SON-σ[+son]; SON-μ[+son]

But under the SON >> FAITH ranking, it is the SON-σ constraint that is subject to reinterpretation, specifically, to threshold raising: SON-σ[+son] is thus construed as SON-σ[–cons], as in (43).

(43) SON >> FAITH
* SON-ϕ[–cons]; SON-σ[+son]; SON-μ[–cons]
interpreted as
SON-ϕ[–cons]; SON-σ[–cons]; SON-μ[+son]

What I argue with regard to impossible cases such as (34) has serious ramifications for learnability: any case which fails to adhere to Prosodic Peak Transparency is subject to reinterpretation and, in a somewhat altered form, ultimately adheres to this principle, following either of the two reinterpretation paths given in (42) and (43).

4. Sonority Threshold Lowering

While either sonority threshold lowering or sonority threshold raising may be called upon to resolve what we refer to here as impossible cases, only threshold lowering has a broader relevance. The mechanism of constraint conjunction, applied to the **SON** family of constraints, provides a tool for handling cases of minimal threshold lowering, that is, minimal violation of a **SON** constraint, simply by generalizing the situation in (38). In order to illustrate this, we turn to Serbian (formerly Serbo-Croatian, Zec, 1988), which provides evidence for the minimal violation of **SON-σ**, and therefore, for a minimal lowering of the sonority threshold imposed by the syllable.

Serbian possesses the following **SON** constraints: **SON-μ**[+sonorant] picks out all sonorants, vowels, glides, liquids and nasals. Evidence for this comes from pre-sonorant lengthening (Zec 1988): stems such as *slav-*, realized in the nominative singular form in *slava* 'glory'are subject to vowel lengthening when followed by certain suffixes, as in *sla:vna*, *sla:van*. Further examples are: *sila* 'might' vs. *si:lna*, *si:lan*; *vera* 'faith' vs. *ve:rna*, *ve:ran*; *pijan* 'drunk' vs. *pija:nstvo*. The additional mora associated with the vowel is donated by the sonorant consonant; if the stem ends in an obstruent, no lengthening occurs, as in *čudo* 'miracle' vs. *čudna*, *čudan*; *rat* 'war' vs. *ratna*, *ratni*.

At the level of the syllable, the relevant constraint is **SON-σ**[+son, +cont]: vowels and *r* are syllabic, as in ***kotrljati*** 'to roll', ***valovi*** 'waves'. The same degree of restriction holds at the level of the foot, with **SON-ϕ**[+son, +cont] as the relevant constraint: vowels and *r* are foot-bearing, as in monosyllabic forms *val* 'wave', *vrt* 'garden'. Crucially, syllables with *r* in the nucleus, just like those with vocalic nuclei, can bear the pitch accent, as in ***vrt***, ***prvi***, ***kotrljati***. In sum, while any sonorant segment, vowel as well as consonant, is moraic in Serbian, only vowels and *r* are syllabic (as well as foot-bearing).

But, with loan words, the situation is somewhat different. In the forms in (44), word-final *l* and nasals are syllabic, in violation of the sonority threshold constraint SON-σ[+son, +cont]:

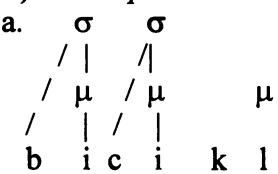
(44) Word-final *l* and nasals in loan words:
bicikl 'bicycle', *dirižabl* 'dirigible', *monokl* 'monocle'; *ak(t)n tašna* 'briefcase'

Yet, word-final obstruents do not become syllabic, even if they belong to "difficult" clusters:

(45) Word-final obstruents in loan words:
kolaps 'collapse', *katarakt* 'cataract', *projekt/projekat* 'project'

While a word-final *l* or a nasal may be moraic, as in (46a), in accordance with SON-μ[+sonorant], if syllabic, as in (46b), it violates SON-σ[+son, +cont]. But the representation selected as the optimal output is (46b), despite the SON-σ[+son, +cont] violation.

(46) *The prosodic status of word-final l or nasal*:

a.


b.


In other words, the optimal output form exhibits threshold lowering. But the form in (46a), in which all SON constraints are duly observed, violates at least one constraint which ranks higher than the SON family, that is, higher than its highest ranked member (see (37) above). It violates EXHAUSTIVITY, which requires the inclusion of prosodic constituents into prosodic structure (and acts in tandem with constraints on the prosodic hierarchy discussed in Section 1).

(47) **EXHAUSTIVITY**: Prosodic constituent *n* must be dominated by prosodic constituent*n+1*, that is, by a prosodic constituent which belongs to the next higher level.

In sum, the sonority threshold imposed by the syllable, **SON-σ**[+son, +cont] is lowered to the threshold imposed by the mora, **SON-μ**[+sonorant], in order to satisfy **EXHAUSTIVITY**, yielding (46b) as the optimal candidate; candidate (46a) fatally violates this constraint, because its final mora is not included into a syllable. This is explicitly shown in Tableau (48):

(48) SON-σ Threshold Lowering

bicikl	**EXHAUS**	**SON-σ & SON-μ**	**SON-σ**	**SON-μ**
☛ σ σ σ / | / | / | b i^{μ} c i^{μ} k l^{μ}		*	*	
σ σ / | / | b i^{μ} c i^{μ} k l^{μ}	*!			

Sonority threshold lowering has to be minimal, precisely because constraint violation has to be minimal. Thus, forms like *bicikl*, as syllabified in (46b) (the winner in (48)), may not bear stress on the syllable with a lowered sonority threshold, because that would automatically involve lowering the sonority threshold imposed by the foot, and incurring the additional violation of **SON-ϕ**. Likewise, word-final obstruents in (45) cannot be syllabic because this would involve the violation not only of **SON-σ**[+son, +cont], but also of **SON-μ**[+sonorant]. The preferred situation would be that the sonority threshold at any level of the hierarchy is lowered at most one level down.

5. Concluding Remarks

We have shown that the distribution of segments within linguistic sequences, rather than being random, is governed by prosodic constituency, whose perspicuous structure has sonority as its driving force. Note however that the imposition of sonority thresholds by the mora, the syllable, and the foot on segments at the bottom of their "head" path constitutes a case of restricted occurrence of segments in *prominent*, that is, nuclear, positions within prosodic constituents. If positional restrictions on the occurrence of segments are viewed as neutralization (under Trubetzkoy's interpretation), then the requirement that prosodic nuclei may only be occupied by segments satisfying a certain minimal sonority threshold should also fall under this rubric.

However, under the theory of positional markedness developed in Steriade (1995) and Beckman (1998), restricted distribution is expected in positions lacking prominence, but not in their prominent counterparts. Under this view, prosodic nuclei, as salient positions within prosodic constituents, should be associated with a free, rather than restricted, distribution of segments. Yet the cases we have presented in this paper provide ample evidence for a restricted distribution of segments in nuclear prosodic positions, clearly demonstrating that this view is not tenable. Moreover, these cases can be characterized as neutralization towards a subclass of segments, notably, those associated with a greater degree of sonority; characterized as a designated minimal sonority threshold imposed by the mora, the syllable, or the foot.

If a higher degree of sonority is seen as a higher degree of "markedness", then the cases we have studied here strongly suggest that the result of neutralization could be a "marked," as well as an "unmarked," element. We could tentatively conclude that, if neutralization takes effect in a prosodically prominent position, its result will be increase, rather than decrease, in "markedness."

References

Beckman, J. N. (1998). *Positional Faithfulness*. Ph.D. Dissertation. University of Massachusetts, Amherst.

Bethin, C. Y. (1998). *Slavic Prosody. Language Change and Phonological Theory*. Cambridge University Press.

Borkovskij, V.I. and P. S. Kuznecov. (1965). *Istoričeskaja grammatika russkogo jazyka*. Nauka.

Hayes, B. (1982). "Extrametricality and English Stress." *Linguistic Inquiry* 13, 227 – 276.

Hewitt, M.S. and M. J. Crowhurst. (1996). "Conjunctive Constraints and Templates in Optimality Theory." *Proceedings of the North East Linstuistic Society 26*, 101 - 116.

Isačenko, A.V. (1980). "East Slavic Morphophonemics and the Treatment of the Jers in Russian: A Revision of Havlík's Law." *International Journal of Slavic Linguistics and Poetics* 13, 73 - 124.

Itô, J. and A. Mester. (1993). *Licensed Segments and Safe Paths*. Linguistics Research Center. Cowell College, UCSC, Santa Cruz.

Itô, J. and A. Mester. (1996). *Rendaku I: Conjunction and the OCP*. Paper Presented at the Kobe Phonology Forum, Kobe University.

Kahn, D. (1978). *Syllable-Based Generalization in English Phonology*. Ph.D. Dissertation, MIT.

Kenstowicz, M. (1994). *Sonority-Driven Stress*. Ms., MIT.

Kiparsky, V. (1979). *Russian Historical Grammar. Volume 1. The Development of the Sound System*. (Translated from the German by J.I.Press.). Ann Arbor: Ardis.

Lightner, T.M. (1972). *Russian Phonology and Turkish Phonology. Problems in the Theory of Phonology*. Edmonton, Alberta: Linguistic Research, Inc.

Lunt, H. (1959). *Old Church Slavonic Grammar*. The Hague: Mouton.

Matteson, E. (1965). *The Piro (Arawakan) Language*. Berkeley: University of California Press.
McCarthy, J. and A. Prince. (1986). *Prosodic Phonology*. Ms., University of Massachusetts, Amherst and Brandeis University.
McCarthy, J. (1998). *Sympathy and Phonological* Opacity. ROA-252-0398.
McCarthy, J. and A. Prince. (1993a). *Prosodic Morphology 1. Constraint Interaction and Satisfaction*. Ms., University of Massachusetts, Amherst and Rutgers University.
McCarthy, J. and A. Prince. (1993b). *Generalized Alignment*. Ms., University of Massachusetts, Amherst and Rutgers University.
McCarthy, J. and A. Prince. (1995). "Faithfulness and Reduplicative Identity." *UMOP* 18. GLSA. UMass, Amherst.
McCarthy, J. (1998). *Sympathy and Phonological Opacity*. ROA-252-0398.
Nespor, M. I. and Vogel. (1986). *Prosodic Phonology*. Dordrecht: Foris
Prince, A. (1990). "Quantitative Consequences of Rhythmic Organization." In Ziolkowski, M., M. Noske, and K. Deaton (eds.) *Papers from the 26th Regional Meeting of the Chicago Linguistic Society. Vol. 2. The Parasession on the Syllable in Phonetics and Phonology*.
Prince, A. and P. Smolensky. (1993). *Optimality Theory. Constraint Interaction in Generative Grammar*. Ms., Rutgers University and University of Colorado, Boulder.
Selkirk, E. O. (1984). *Phonology and Syntax*. MIT Press.
Selkirk, E. O. (1995). "Prosodic Structure of Function Words." *UMOP 18*. GLSA. UMass, Amherst.
Shevelov, J. (1965). *A Prehistory of Slavic*. The Hague: Mouton.
Smolensky, P. (1995). *On the Internal Structure of the Constraint Component Con of UG*. Rutgers Optimality Archives.
Steriade, D. (1995). "Underspecification and markedness." In Goldsmith, J. (ed.) *The Handbook of Phonological Theory,* 114–174. Oxford: Basil Blackwell.
Zec, D. (1988). *Sonority Constraints on Prosodic Structure*. Ph.D. Dissertation. Stanford University.

Zec, D. (1995). "Sonority Constraints on Syllable Structure." *Phonology* 12, 85–129.
Zec, D. (1998a). "Coda Constraints and Conditions on Moraic Structure." *Working Papers of the Cornell Phonetics Laboratory* 12.
Zec, D. (1998b). *Prosodic Weight*. Paper Presented at the Workshop on The Syllable: Typology and Theory. University of Tübingen, Germany. July, 1998.

Morrill Hall
Cornell University
Ithaca, NY 14853
dz17@cornell.edu

Experimental Evidence for Focus Structure in Russian

Gerhild Zybatow and Grit Mehlhorn
University of Leipzig

1. Introduction

Word order and prosody are the linguistic correlates of Information Structure (IS). The constituents marked with the feature [TOP] (Topic) or [FOC] (Focus) occupy a privileged position in the clause and their prosodic characteristics include a special pitch accent, fundamental frequency (F0), and time structure.

The aim of the present paper is to show how prosodic parameters and their acoustic realization interact with discourse-related aspects of utterances. In particular, we will investigate the influence that focus structure as a part of IS has on Russian declarative sentences. To that end, we conducted an experiment that dealt with different types of focus within appropriate contexts.

Whereas in traditional Russian studies on the subject only isolated sentences without context were recorded and interpreted, (Bryzgunova 1980:100), the sentences in our experiment have a predictable and controllable information structure which allows us to describe intonation contours used in response to the communicative intention of the speaker.

Bryzgunova's (1980) concept of intonational constructions reflects only a part of the contour of the sentence: namely the location of the main accent in the centre of the pitch movement, i.e., only the stressed syllable and the syllables immediately preceding and following it. However, different types of focus cannot be distinguished exclusively on the basis of the local tonal contour. It is necessary to compare intonational constructions with tonal movement of adjacent accents. Our approach allows us to analyse relational pitch, i.e., the relation to the global F0 course of the sentence.

This article is organized as follows. In Section 2 we will outline the theory of information structure that we presuppose. Section 3 introduces the different types of foci in Russian. Section 4 describes our experiment and presents acoustic data

that reveal the tonal and durational behaviour of different foci in Russian declarative sentences. A summary and a short discussion conclude the paper (Section 5).

2. Information Structure and Grammar

The term *Information Structure* (IS) refers to the adequacy of an utterance in a text depending on the communicative situation and the discourse structure. In Russian, means of expressing IS include word order, placement and type of pitch accents, and lexical items. The IS as a pragmatically determined principle of ordering can be described within the grammar of the sentence by means of syntactic rules. These rules organize the constituents of a sentence into surface structures with regard to their communicative importance. IS is a cover term for different discourse functions which certain syntactic elements fulfil in an utterance.

An adequate description of IS by means of a sentence grammar requires a corresponding theory of grammar. We assume a generative model of grammar, specifically, a Minimalist one (Chomsky 1993). Within this theory syntactic features play an important role. But contrary to the standard assumptions, we assume that there are two kinds of syntactic features, i.e., morphosyntactic features and IS features. According to the null hypothesis, the IS features are freely assigned to the corresponding constituent in the syntactic tree.

The association of syntactic constituents with IS features has an effect on the word order, phonological properties, and interpretation of the sentence. The influence on word order is based on the fact that IS features can force the movement of a constituent or disallow it. With regard to phonology, the IS features determine placement and types of pitch accents. In a sentence that is well-formed with regard to its context, the IS features are assigned to those constituents which reflect the intentions of the speaker, i.e., his assumptions about the knowledge of the hearer according to discourse function. Therefore, we speak about the adaptation of the sentence to the context. This adaptation is achieved according to several discourse functions. Following Rosengren (1993), Steube (1997) and Molnár (1998) we proceed from the assumption that there are at least four different discourse functions which

determine two dimensions of IS: on the one hand, Focus-Background Structure and on the other hand, Topic-Comment Structure. Focus is the most salient component of the Focus-Background dimension, and Topic is the most salient element of the Topic-Comment division. These discourse functions are determined as follows:

(1)

Focus:	the most important information from the speaker's perspective in a given context
Background:	the less important information from the speaker's perspective in a given context
Topic:	starting point of the sentence; what the sentence is going to be about
Comment:	predication about the topic

Several different types of topics and foci can be distinguished. Topics can be external or internal. Internal topics can be concrete or abstract. Overt topics are realized at the left periphery of the sentence.[1] Alter and Junghanns (1997) showed in an experiment that all topics in Russian have the same tonal pattern of a rise-fall sequence.

3. Different Focus Types in Russian: Prosodic Structure and Interpretation

The focus feature is phonologically more prominent than the topic feature. This is because of the function of the focus, namely to emphasize the important information in the given context. Unlike the topics, the various types of focus have different acoustic parameters.

[1] Space does not allow more than a cursory description. For a comprehensive discussion see Junghanns and Zybatow (1997), Zybatow and Junghanns (1998).

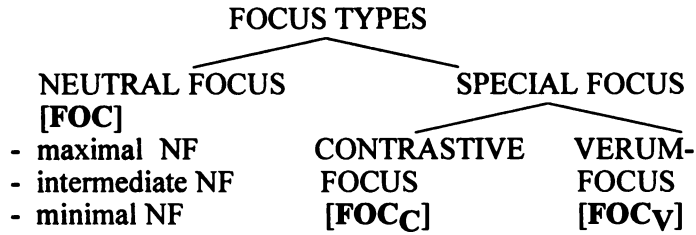


Figure 1. Focus Types

3.1. Neutral Focus

The neutral focus, or the new information focus (NF), is realized at the right periphery (Junghanns and Zybatow 1995) and is characterized by a falling pitch accent. As we indicate below, the NF can be maximal (2a), intermediate (2b) or minimal (2c).

(2) a. A: V čem delo?
in what matter
'What happened?'
B: [FOC[CP [TOPMiroslavai][VP ti uexala v JALtu]]]
Miroslava left for Jalta
'Miroslava left for Jalta.'

b. A: Čto s Miroslavoj?
what with Miroslava
'What about Miroslava?'
B: [TOPMiroslavai][FOC[VP ti uexala v JALtu]]

c. A: Kuda uexala Miroslava?
where left Miroslava
'Where did Miroslava leave for?'
B: [TOPMiroslavai][VP ti uexala [FOC[PPv JALtu]]]

In example (2a), the whole sentence is a possible answer to the question *V čem delo? Čto slu ilos'?* 'What happened?' and we have maximal focus. In sentence (2b), *Čto s Miroslavoj?* 'What about Miroslava?' is answered by *uexala v JALtu* 'left for Jalta'. The neutral focus is assigned to the VP. The focus is non-maximal, and *Miroslava* is a background constituent. As the

neutral focus in (2b) is also non-minimal, the two constituents remain in the VP. Example (2c) illustrates a minimal focus, where we have a question about the argument-PP of *uexat'*: 'Where did Miroslava leave for?' But in all three cases, the focus exponent[2] is the same, namely the syllable JAL- in the word *Jaltu*. We find focus ambiguity in the sentence *Miroslava uexala v Jaltu* which can be disambiguated with the help of the context. A neutral word order is potentially ambiguous with respect to the focus domain.

Alter (1997a,b) has shown that there is a significant difference with respect to prosodic qualities between maximal and minimal focus. A minimally focused constituent which syntactically represents a maximal phrase constitutes its own prosodic domain.

3.2. Contrastive Focus

The contrastive focus (CF) differs from the NF with regard to acoustic realization and discourse function. With respect to its acoustic realization, we assume that the CF starts off higher than the non-contrastive accent and therefore has a greater excursion of the pitch accent.

With respect to meaning, CF also marks important information, but additionally it has emphatic meaning indicating the contrast with other potential assumptions. CF always entails the correction of an explicit utterance or of a presupposition of the situational context. The speaker corrects false information because otherwise this information would be interpreted as a valid background for the sentence by the hearer. The speaker corrects his utterance either if the given sentence involves values that can be misinterpreted by the hearer or if he assumes that the hearer has got the wrong background information.

In contrast to NF, CF does not correspond to a fixed position. In Russian, CF can be assigned to any constituent *in situ* or after movement. Consequently, CF can be realized on

[2] The focus exponent – indicated by capitals – is the syllable carrying the main stress (sentence accent). The findings in our experiment refer to what happens on the accented syllable of the focused constituent.

the left (3a) or right periphery of the clause (3c) or in between (3b).

(3) a. A: Ljuda uže uexala v Jaltu?
'Has Ljuda already left for Jalta?'
B: [$_{FOCc}$MiroSLAva] uexala v Jaltu.
Miroslava left for Jalta
'It's MIROSLAVA who left for Jalta.'
Ljuda tak zanjata, čto daže ob étom i ne mečtaet.
'Ljuda is so busy that she doesn't even dream of going there.'

b. V Jaltu [$_{FOCc}$MiroSLAva] uexala.
for Jalta Miroslava left
'It's MIROSLAVA who left for Jalta.'

c. V Jaltu uexala [$_{FOCc}$MiroSLAva].
for Jalta left Miroslava
'It's MIROSLAVA who left for Jalta.'

Due to a special pitch accent associated with the CF, it is unnecessary to anchor it to any fixed position within the clause.

3.3. Verum-Focus

The third type of focus, the Verum-Focus (VF), is realized on the finite verb in Russian (4a-c) and emphazises the truth of the predicate:

(4) a. A: Mne bylo očen' smešno, kogda ja uznala, čto tvoj djadja zaxotel kupit' sebe dom v Italii. On že vsegda ele-ele svodil koncy s koncami.
'I was quite amused when I heard that your uncle was going to buy himself a house in Italy. As we know, he has hardly enough money to get by on.'
B: No èto tak. [$_{FOCv}$KuPIL] moj djadja ètot dom.
'But it is true. My uncle DID buy this house.'
On neožidanno polučil v nasledstvo mnogo deneg.
'Unexpectedly, he inherited a lot of money.'

b. A: Ploxo, to Jura takoj uprjamyj. On nikogda ne soglasitsja pererabotat´ svoju stat´ju.
'Unfortunately, Jura is so stubborn. He won't ever agree to rewrite his paper.'
B: Jura [$_{FOCv}$BUdet] pererabatyvat´ stat´ju.
Jura will rewrite article
'Jura WILL rewrite the article.'
Ja na ètom nastaivaju. A tem bolee, čto ète v ego interesax. Esli on ne soglasitsja, to stat´ja nikogda ne budet opublikovana.
'I insist on that. Particularly as it is in his own interest. If he doesn't agree, the article won't ever be published.'

c. Moj djadja ètot dom [$_{FOCv}$kuPIL].
my uncle this house bought
'My uncle DID buy this house.'

Considering the use of VF in a given context, it can be seen that usually the verb bearing the VF has already been mentioned. This observation corresponds to Höhle's (1992) definition of VF. According to him, VF emphasizes the truth of a contextually known proposition. With VF, the speaker coreferentially resumes an established proposition and rejects the hearer's doubts about the truth of the proposition.

The verum-focused verb, similar to a contrastively focused constituent, can occur in any position of the clause (Zybatow 1997b). Since the VF pitch accent resembles the CF accent in perception, we assume VF to be a special kind of CF. Therefore, we included the VF in our experiment.

3.4. Prosody and Information Structure

In speech production, speakers make use of intonation in order to signal the relevant information in the sentence. In speech processing, the hearer uses intonation in order to interpret the speaker's intentions in the right way. Intonation gives the sentence the appropriate communicative interpretation relative to the situational context and the speaker's intention. In this function intonation is a means of realizing the IS of the

clause and, in connection with this, a means of realizing accentuation.

The physical correlate of pitch is the fundamental frequency F0 of the voiced parts of the speech signal. The F0 reflects the progress in time of the periodic closing of the vibrating vocal cords. For oral communication, the form of the tonal pattern in relation to the frequency range of the speaker is much more important than the absolute tonal height.

Prosody can be described by means of accents. On the one hand, 'accent' refers to lexical or word stress, i.e., the stressing of a certain syllable. On the other hand, there is sentence accent – the stressed syllable which is normally characterized by different phonetic parameters. Stress usually changes the F0, the loudness of sound, the length of the vowel, and – especially in Russian – the exactness of articulation.

4. Focus in Production: An Experimental Study

This experiment was designed to investigate three specific questions. First of all, we wanted to find out in what way a typical contrastive accent is prosodically realized in Russian. Secondly, we wanted to find possible differences in the prosodic realization of the different focus types. In particular, we expected a higher vertical excursion on the focus exponent and a greater lengthening of the accented syllable of CF as compared to NF. Thirdly, we investigated whether VF and CF differ with respect to any prosodic parameter. Since both VF and CF include a kind of correction and since not every little shade of meaning results in a distinct intonation contour, we did not expect a prosodic difference between CF and VF. The prosodic parameters we are interested in are intonation contours as well as the type, location and duration of pitch accents.

4.1. Method

The present experiment was a controlled production study, i.e., the subjects read experimental sentences within their contexts from a list. All texts were recorded.

Materials and Design. We tested categorical[3] sentences with NF domains of different sizes as well as CF and VF on different constituents and in different positions in the sentence. We had 60 declarative sentences as experimental items. In addition, 40 fillers (external topics, thetic sentences as well as sentences without topics) were added to the 60 experimental items, resulting in a total of 100 items. All 100 sentences were embedded in appropriate contexts to form colloquial dialogues. The context was necessary to guarantee a coherent and controllable IS and to ensure that the discourse functions which were to be investigated could be recognized in the text. The length of the texts varied from 2 to 5 sentences. The subjects did not know which items were experimental, which were context and which were fillers. In order to obtain simple intonation contours, the experimental items were short and contained the minimal number of constituents. No such restrictions were imposed on the contexts.

The following parameters were varied systematically in the experiment:

(5) a. the focus type (NF vs. CF vs. VF);
b. the pitch accent position (clause-initial, clause-final and in the middle of the clause) and
c. the type of accented constituents (subjects, direct objects, indirect objects, verb complements in the instrumental case, adverbs, and adjectives).

Subjects. Eight female subjects were tested. They were native Russian speakers between 20 and 30 years of age. They were paid for their participation in the experiment. Professional speakers were not chosen intentionally. The experiment was conducted at the University of Leipzig.

Procedure. The subjects did not know what exactly was being tested. They saw the sentences in written form. The subjects were orally instructed to read all the texts first in order to familiarize themselves with the content. Then they read each individual text silently and were supposed to imagine themselves in the roles of the speakers. Only then did they read

[3] On the categorical/thetic distinction see Junghanns and Zybatow (1997).

the text aloud – as naturally as possible. All 100 sentences were recorded by all subjects. In general, the subjects took 90 minutes to complete the experiment.

Data Treatment. The texts were recorded in a soundproof booth using a DAT recorder with a frequency of 44.1 kHz and a 16 bit sampling rate. The speech signals were digitized and the experimental items were extracted from their contexts. From these the intonation contours were drawn by pitch tracking.

WinPitch software program was used to analyze the extracted intonation contours. In this program, the following measurements were taken. At three points of each syllable F0 values were extracted: (a) onset (the first one at the beginning of the syllable), (b) peak (the highest from the F0-peak) and (c) offset (the final one at the end of the syllable). There were two important values per measuring point: the fundamental frequency F0 in Hz and the time in msec. On the accented syllable of the topic and focus more points were measured in order to describe exactly what happens on the accented syllable of the focused constituent. In total, approximately 28 points were measured per sentence. The data were transferred to an Excel table and a median value from the data of the eight subjects for each measuring point was calculated. From these values an intonation contour averaged from all eight subjects was produced for each sentence.

The standard transcription system ToBI (for *Tone and Break Indices*) was used for the prosodic labelling of the data. The pitch variation can be described with two separate tones; H and L.[4] H stands for a high target of pitch variation and L for a

[4] Pierrehumbert (1980) was the first to describe the interpretation of discourse as a dynamic two-level process. Based on that, a multi-site Labelling system ToBI was developed. The criterion for describing F0 contours which is emphasized in the ToBI system is that only those intonation events should be labelled which are distinctive in the sense that they can be assigned a function in the domain of discourse interpretation. Later, a consensus prosodic transcription system for German, based on the English ToBI system, was developed (Mayer 1995; Reyelt et al. 1996; Benzmüller and Grice 1997). Guidelines for ToBI Labelling can be found at *http://www.ling.ohio-state.edu/phonetics/E_ToBI/etobi_homepage.html.* No ToBI system has been developed for Russian, yet.

low one.[5] Pitch accent tones can be monotonal (H*; L*) or bitonal (HL*, H*L, L*H, LH*). The reanalysis of the contour is realized through a regular interpolation between these targets.

4.2. Results and Discussion

Neutral Focus. We compared sentences with different focus domains. In the case of maximally focused sentences, e.g. *[Miroslava uexala v JALtu]*, there is a single intonation phrase which was pronounced without breaks. The pitch accent on the focus exponent is a fall labelled with HL* (Figure 2).[6] The target of the pitch movement is a low-tone in the intonation contour which is indicated by a star (*).

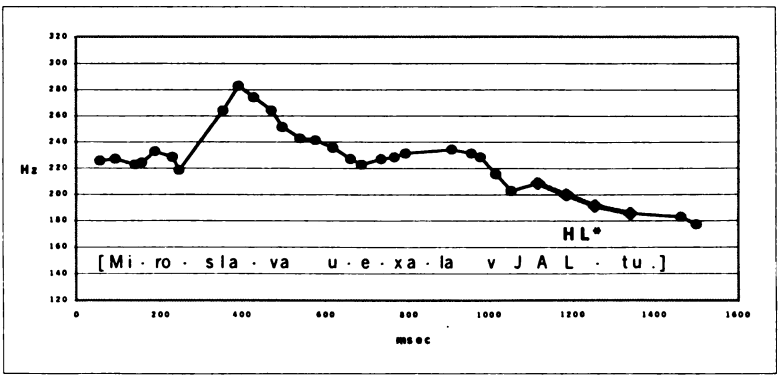


Figure 2. Maximal NF[7]

Both the sentences with intermediate focus of the type *Miroslava [uexala v JALtu]* and maximally focused sentences exhibit a sequence of accents which descend in stepwise fashion. This so called *downstepping* begins with the topic constituent, continues with the first lexical stress of the intermediate focus – in this case the verb – and ends with a pitch accent on the focus exponent which is again pronounced as a fall: HL* (Figure 3).

[5] The inventory of ToBI also comprises boundary tones and phrase accents, but for convenience we use only the labels relevant for our purposes.
[6] The focus exponent is marked by the diamonds on the intonation contour.
[7] In each case the curves indicate the median value across all subjects.

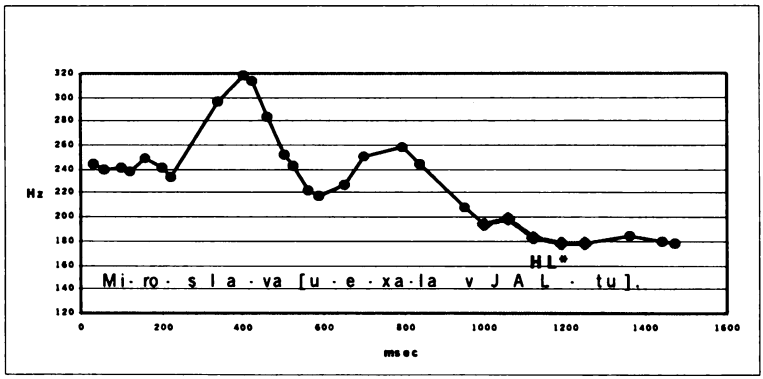


Figure 3. Intermediate NF

In sentences with a minimal NF, e.g. *Miroslava uexala [v JALtu]*, there is a somewhat stronger fall on the F0 on the focus exponent than in the cases of maximal and intermediate foci. We again label this pitch accent with HL* (Figure 4).

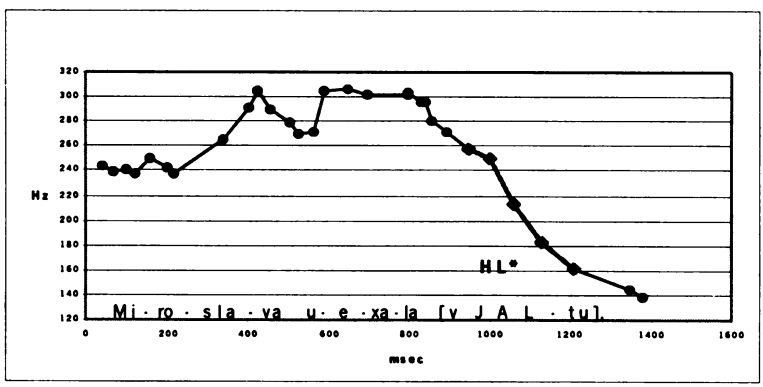


Figure 4. Minimal NF

For all sentences with a NF, the focus exponent is pronounced with a greater lengthening than any other syllable in the sentence (Figure 11). Regardless of the extension of the focus domain, the topic accent in categorical sentences with NF shows a rise in the F0 (LH*) which is reversed on the following syllables.

Contrastive Focus. In the case of sentence-initial CF, e.g. *[$_{FOCc}$ MiroSLAva] uexala v Jaltu*, a combination of a strong rise and a beginning fall of the F0 on the accented syllable is observed. A low tone precedes the accent, i.e., the pitch falls until the position of the accented syllable. On the focus exponent there is a strong rise. The target of the pitch movement is a high tone which is indicated by the star. The pitch range is extended. Still on the same syllable the pitch movement is reversed in a beginning descent which is continued on the following unaccented syllables until the end of the sentence. We label this pitch accent with LH*+L.[8] In this case, there is no topic accent in the sentence, i.e., topic and contrastive accent occur on the same constituent.[9] The fall is continued to the end of the sentence. The remaining contour of the sentence is flat; there are no other pitch accents (Figure 5).

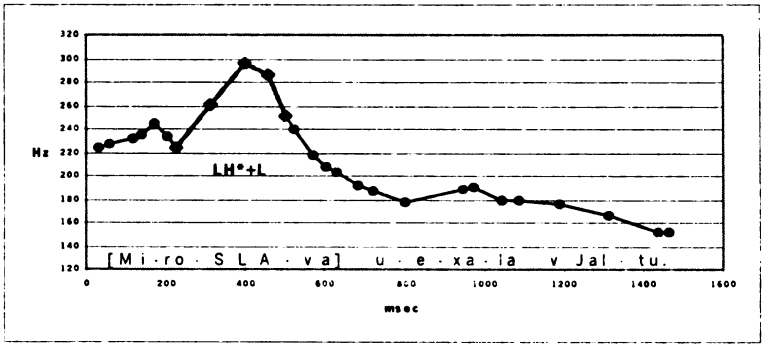


Figure 5. CF on the Left Periphery

The form of a CF on the second syntactic constituent in the sentence, as in *V Jaltu [$_{FOCc}$MiroSLAva] uexala*, resembles the sentence-initial CF. Again, there is a rise and a fall on the

[8] In ToBI, a rise of F0 is transcribed as LH*. Although a combination of no more than two tones is allowed in ToBI, we label this contrastive accent with LH*+L, in order to be able to capture the beginning fall of the F0 on the accented syllable.

[9] Nevertheless, there is an important difference between a topic accent in a neutral sentence and a sentence-initial CF. Whereas on the stressed syllable of a topic there is only a rise of the F0 and the fall begins later on the postaccentual syllables, the focus exponent of the CF combines the rise and fall of the F0 on one syllable.

accented syllable and the fall of the F0 is continued to the end of the sentence. However, in contrast to a sentence-initial CF, there is a kind of topic accent, albeit one which does not reach such a large vertical excursion as in sentences with NF.

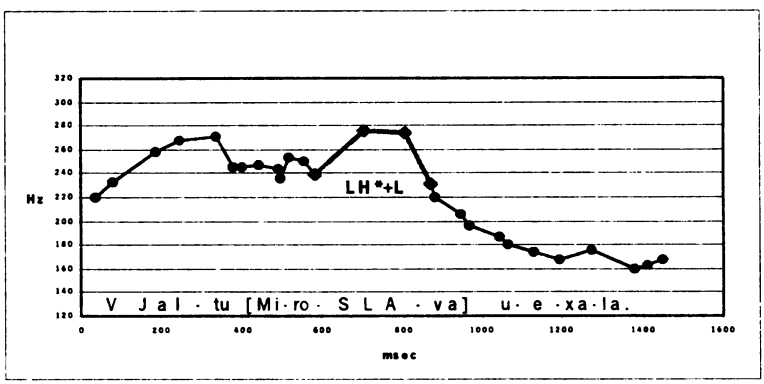


Figure 6. CF in the Middle of the Utterance

Sentences with a CF at the right periphery exhibit a normal topic accent. If the contrastive pitch accent is assigned to the last syntactic phrase in the sentence, as is the case in *V Jaltu uexala [*$_{\text{FOCc}}$ *MiroSLAva]*, the focus exponent is realized with a strong fall of the F0 which is preceded by a small rise on the same syllable which intensifies the vertical excursion of the fall. Again, we can label this pitch accent with LH*+L.

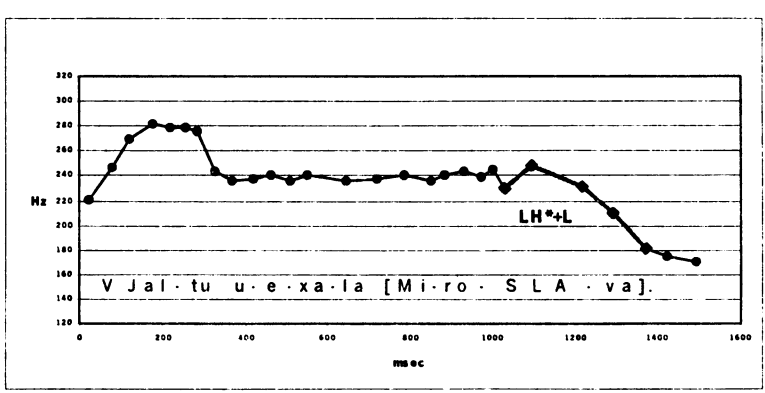


Figure 7. CF on the Right Periphery

Verum Focus. In its prosodic behaviour the VF shows parallels with CF. If the VF falls on the sentence-initial constituent as in the sentence *[$_{FOCv}$KuPIL] moj djadja ètot dom*, it does not differ from an intonation contour of a CF on the first constituent (compare Figures 5 and 8).

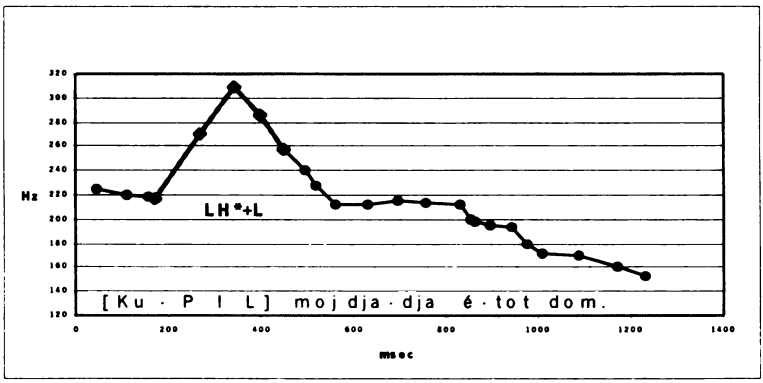


Figure 8. VF on the Left Periphery

A VF on the second position, e.g., on the auxiliary in the sentence *Jura [$_{FOCc}$BUdet] pererabatyvat' stat'ju*, is also marked by a large rise and a beginning fall of the focus exponent (LH*+L). The fall is continued on the syllables after the focus exponent, and the remaining declination of the sentence is totally flat (Figure 9).

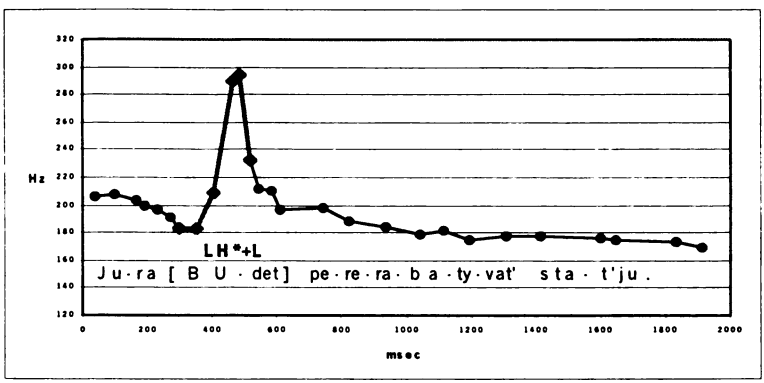


Figure 9. VF on the Second Position

If the sentence-final constituent is verum-focused as in the sentence *Moj djadja ètot dom* [$_{FOCv}$*kuPIL]*, the target – a high tone – is located on the accented syllable immediately before the fall. Because of this rise on the same syllable, the fall on the focus exponent is clearly evident. Again, we label this tonal sequence with LH*+L (Figure 10).

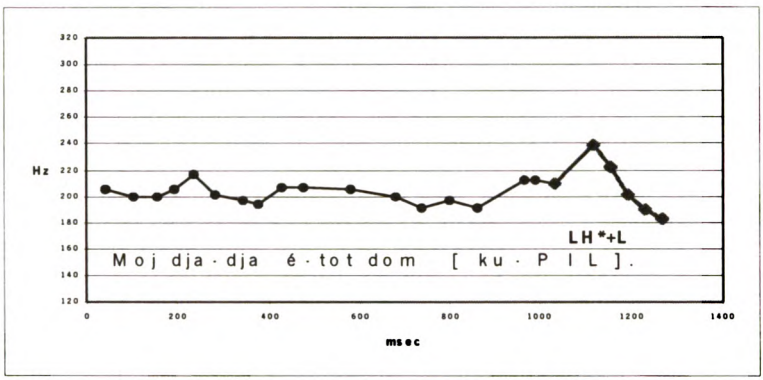


Figure 10. VF on the Right Periphery

In all sentences with CF and VF, the syllable bearing the focus accent is characterized by a very strong lengthening which is even larger than the one on the nuclear accent of a NF (compare Figures 11, 12 and 13). This salient lengthening on the accented syllable in sentences with CF or VF seems to be an additional contribution to the perception of a contrast.

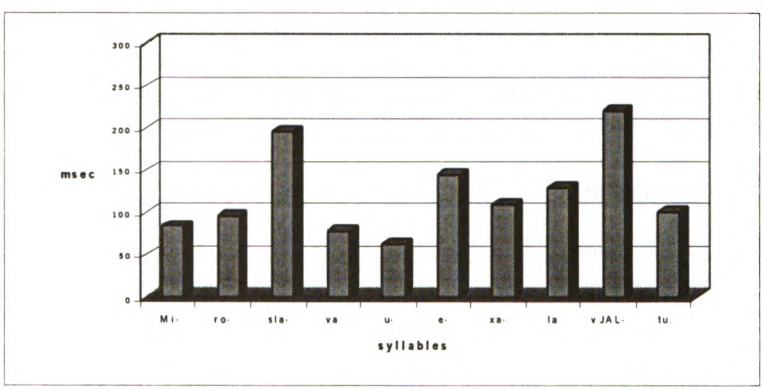


Figure 11. Duration of the Syllables in a Sentence with NF

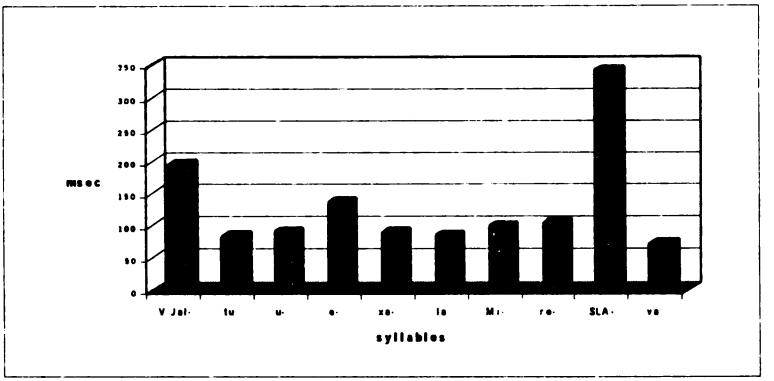


Figure 12. Duration of Syllables in a Sentence with CF

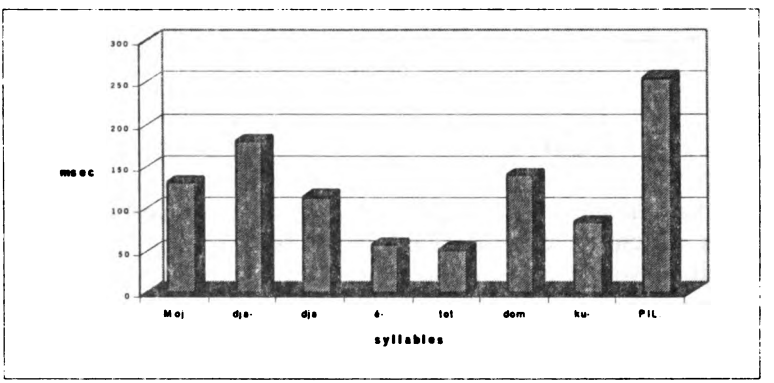


Figure 13. Duration of Syllables in a Sentence with VF

As before, we are dealing with the median value across all subjects. Generally, the vertical excursion of CF and VF will be higher the closer it is to the beginning of the sentence.

5. Conclusions

The experiment reported in Section 4 above provides empirical evidence for the focus structure in Russian as reflected in placement and prosodic characteristics of pitch accents. The subjects placed the sentence accents at the expected position in the sentence, i.e., on the focus. The context and the IS of the sentence in question forced this intonation.

The focus exponent of the NF is usually located on the last lexical accent in the sentence and is produced by a fall of the F0 (Table 1). Sentences with NF are, in terms of intonation, relatively independent from the lexical stressed syllables before the sentence accent. They can be rhythmically subdivided.

The sentence accent in contrastive contexts and sentences with a VF exhibits somewhat more distinct parameters in the speech signal: the vertical height on the accented syllable is essentially higher than that on the NF, and the duration of the nuclear accent is clearly increased. The local tonal pattern, a combination of rise and fall of the F0, dominates the whole utterance. The contour of the remaining declination is smoothed out, and the whole sentence is pronounced more quickly. Such sentences cannot be rhythmically divided. They have strong-centered contours. The typical contrastive accent can be described as a combination of a strong rise and a beginning fall of the F0 which occur on the same syllable – the focus exponent. This kind of accent we label with LH*+L. It occurs on both CF and VF which gives evidence to consider the VF as a special case of CF.

Table 1. Prosodic Characteristics of the Different Focus Types

Type of Focus	Ex.	Figure	F0 on Focus Exponent	Length of Focus Expon.
Neutral				
maximal	(2a)	2.	HL*	
intermed.	(2b)	3.	HL*	lengthening
minimal	(2c)	4. & 11.	HL*	
Contrastive				
initial	(3a)	5.	LH*+L	substantial
middle	(3b)	6.	LH*+L	lengthening
final	(3c)	7. & 12.	LH*+L	
Verum				
initial	(4a)	8.	LH*+L	substantial
middle	(4b)	9.	LH*+L	lengthening
final	(4c)	10. & 13.	LH*+L	

The shape of a typical contrastive accent is created by the interaction of different prosodic parameters:

(7) a. the accent type (combination of a rise and a beginning fall on the accented syllable)
b. a higher vertical excursion of the F0 (extended pitch range)
c. an obvious lengthening on the accented syllable as compared with all other syllables in the sentence
d. the flat remaining sentence declination which signals that there was important information later in the sentence.

We were able to show that disambiguation of different focus types in sentences with identical word order is possible with the help of the context and intonation contours. IS features correspond to a particular position in the sentence and particular prosodic characteristics. This allows us to describe intonation contours on the basis of the communicative intention of the speaker.

Finally, with the different types of focus and various concepts of contrast we are able better to distinguish what in previous intonation research under the term "special accentuation" (*akcentnoe vydelenie*) remained quite unclear. With the help of discourse functions a new relationship to intonation contours is possible, contrary to traditional Russian studies where intonation constructions only referred to sentence mood and to the presence or absence of emphasis.

With this experiment, prosodic correlates of different focus types in the Russian declarative sentence were found. Thus, it was proved that Information Structure is reflected in intonation and that focussing of syntactic constituents corresponds to a certain pitch contour.

References

Alter, K. (1997a). "Russian Prosody: Phrasing and Tonal Structure." In Lindseth M. and S. Franks (eds.): *Formal Approaches to Slavic Linguistics 5: The Indiana Meeting.* 1-17. Ann Arbor: Mich. Slavic Publications.

Alter, K. (1997b). "Fokusprosodie im Russischen: Phonologische und akustische Korrelate von Informationsstrukturierung." In Junghanns U. and G.

Zybatow (eds.) *Formale Slavistik*. 399-414. Frankfurt am Main.

Alter, K. and U. Junghanns. (1997). "Topic-Related F0 Patterns in Russian." To appear in Kosta, P., M. Unger et al. (eds.) *Formal Description of Slavic Languages 2*.

Benzmüller, R. and M. Grice (1997). "Trainingsmaterialien zur Etikettierung Deutscher Intonation mit GToBI." In *Phonus 3*. Institute of Phonetics, University of the Saarland, 9-34.

Bryzgunova, E. A. (1980). "Intonacija." In *Russkaja Grammatika*. Moscow: Nauka, Vol 1.

Chomsky, N. (1993). "A Minimalist Program for Linguistic Theory." In Hale, K. and S.J. Keyser (eds.) *The View from Building* 20. Cambridge, Mass.: MIT Press.

Höhle, T. N. (1992). "Über Verum-Fokus im Deutschen." In Jacobs, J. (ed.) *Informationsstruktur und Grammatik. Linguistische Berichte* S4, 112-141.

Junghanns, U. (1997). "Features and Movement." In Alexiadou A. et al. (eds.) *ZAS Papers in Linguistics*. Berlin: Vol 9, 74-88.

Junghanns, U. and G. Zybatow. (1995). "Fokus im Russischen." In Kohlhof I. et al. (eds.) *Proceedings of the Goettingen Focus Workshop*. 17. DGfS, 113-137.

Junghanns, U. and G. Zybatow. (1997). "Syntax and Information Structure of Russian Clauses." In Browne, W. et al. (eds.) *Formal Approaches to Slavic Linguistics 4: The Cornell Meeting*. 289-319. Ann Arbor: *Mich. Slavic Publications.*

Mayer, J. (1995). *Transcription of German Intonation: The Stuttgart System*. Technischer Bericht, 14 pp., Stuttgart.

Molnár, V. (1998). "Topic in Focus. On the Syntax, Phonology, Semantics and Pragmatics of the So-Called 'Contrastive Topic' in Hungarian and German." In *Acta Linguistic Hungarica* 45 (1-2), 89-166.

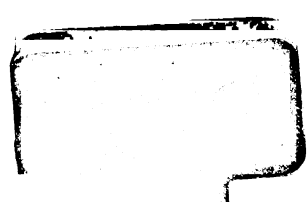
Pierrehumbert, J. (1980). *The Phonetics and Phonology of English Intonation*. Ph.D. Disseratation, MIT.

Reyelt, M., M. Grice, R. Benzmüller, J. Mayer, and A. Batliner. (1996). "Prosodische Ettikettierung des Deutschen mit ToBI." In Gibbon D. (ed.) *Natural Language and Speech Technology*. 144-155. Berlin: Mouton de Gruyter.

Rosengren, I. (1993). "Wahlfreiheit mit Konsequenzen – Scrambling, Topikalisierung und FHG im Dienste der

Informationsstrukturierung." In Reis M. (ed.) *Wortstellung und Informationsstruktur*. Tübingen, 251-312.
Steube, A. (1997). *Ein kognitionswissenschaftliches Modell für Informationsstrukturierung*. Ms., Leipzig.
Zybatow, G. (1997a). "Determinanty Informacionnoj Struktury." In Junghanns U. and G. Zybatow (ed.) *Formale Linguistik*. 359-370. Frankfurt am Main.
Zybatow, G. (1997b). "Über Verum-Fokus im Russischen." To appear in Kosta, P., M. Unger et al. (eds.) *Formal Description of Slavic Languages 2*.
Zybatow, G. and U. Junghanns. (1998). "Topiks im Russischen." *Sprache und Pragmatik*, Lund.

Universität Leipzig
Institut für Slavistik
Augustusplatz 9
D – 04109 Leipzig
Federal Republic of Germany
zybatow@rz.uni-leipzig.de
mehlhorn@rz.uni-leipzig.de



Babby
Boeckx
Brala
Caink
Cetnarowska
Chew
Corbett
Crosswhite

Helmreich
Jakab
Kazenin
Kupść
Lambova
Leko
McShane
Mehlhorn
Milnes
Nicol
Nirenburg
Pereltsvaig
Raskin
Rudnitskaya
Sheremetyeva
Stjepanović
Wilson
Wiltshire
Zec
Zybatow