SKY Journal of Linguistics is published by the Linguistic Association of Finland (one issue per year).

Notes for Contributors

Policy: *SKY Journal of Linguistics* welcomes unpublished original works from authors of all nationalities and theoretical persuasions. Every manuscript is reviewed by at least two anonymous referees. In addition to full-length articles, the journal also accepts short (3–9 pages) 'squibs' as well as book reviews.

Language of Publication: Contributions should be written in English, French, or German. If the article is not written in the native language of the author, the language should be checked by a qualified native speaker.

Style Sheet: A detailed style sheet is available from the editors, as well as via WWW at http://www.linguistics.fi/skystyle.shtml.

Abstracts: Abstracts of the published papers are included in *Linguistics Abstracts* and *Cambridge Scientific Abstracts*. The published papers are included in *EBSCO Communication & Mass Media Complete*. *SKY JoL* is also indexed in the *MLA Bibliography*.

Editors' Addresses (2012):

Pauli Brattico, Faculty of Information Technology, P.O. Box 35 (Agora), FI-40014 University of Jyväskylä, Finland

Tiina Keisanen, Faculty of Humanities, P.O. Box 1000, FI-90014 University of Oulu, Finland

Saija Merke, Finnish Language, P.O. Box 3, Fabianinkatu 33, FI-00014 University of Helsinki, Finland

Mia Raitaniemi, German Department, Henrikinkatu 2, FI-20014 University of Turku, Finland

Erika Sandman, Department of World Cultures, East Asian Studies, P.O. Box 59, FI-00014 University of Helsinki, Finland

Editors' E-mail: sky-journal(at)helsinki(dot)fi

Publisher:

The Linguistic Association of Finland
c/o General Linguistics
P.O. Box 24 (Unioninkatu 40)
FI-00014 University of Helsinki
Finland
http://www.linguistics.fi, http://www.linguistics.fi/skyjol.shtml

The Linguistic Association of Finland was founded in 1977 to promote linguistic research in Finland by offering a forum for the discussion and dissemination of research in linguistics, both in Finland and abroad. Membership is open to anyone interested in linguistics. The membership fee in 2012 was EUR 25 (EUR 15 for students and unemployed members). Members receive *SKY Journal of Linguistics* gratis.

Cover design: Timo Hämäläinen 1999

SKY Journal of Linguistics
25

Suomen kielitieteellisen yhdistyksen aikakauskirja
Tidskrift för den Språkvetenskapliga föreningen i Finland
Journal of the Linguistic Association of Finland

2012

ISSN-L: 1456-8438
ISSN: 1456-8438 (Print)
ISSN: 1796-279X (Online)

Hansaprint, Turku 2012

Contents

External Reviewers of *SKY JoL* 25 (2012)

In addition to members of the current advisory editorial board (the following scholars have acted as external reviewers for *SKY Journal of Linguistics* in 2012:

Annelie Ädel (Dalarna University), Ante Aikio (University of Oulu), Rajaa Aquil (Georgia Institute of Technology), Larisa Avram (University of Bucharest), Tim Bazalgette (Cambridge University), Monika Bednarek (University of Sydney), Koen Bostoen (Ghent University/ Université libre de Bruxelles), Esther Lynn Brown (University of Colorado Boulder), Mousa A. Btoosh (Al-Hussein Bin Talal University), Rémi Camus (Institut Nationale des Langues et Civilisations Orientales (Inalco), Paris), Candice Chi-Hang Cheung (Chinese University of Hong Kong), Denis Creissels (University of Lyon), Daniel Currie Hall (University of Toronto), Jonathan Fine (Bar-Ilan university), Ángel Gallego (Universitat Autònoma de Barcelona), Daniel Hall (Saint Mary's University), Peter Hallmann (University of Vienna), Ilona Herlin (University of Helsinki), Patrick Honeybone (Edinburgh University), Rasheed Al-Jarrah (Yarmouk University), Vesa Jarva (University of Jyväskylä), Elsi Kaiser (USC Department of Linguistics), Jean-Michel Kalmbach (University of Jyväskylä), Johanna Laakso (University of Vienna), Minna Laakso (University of Helsinki), Yury Lander (Moscow State University), Jouko Lindstedt (University of Helsinki), Terje Lohndahl (Norwegian University of Science and Technology), Severine Maggio (Université Blaise Pascal), Prashant Mishra (Government S.V.P.G. College Neemuch), Hazem Yousef Najjar (Bethlehem University), Tiina Onikki-Rantajääskö (University of Helsinki), Felix Otter (Universität Heidelberg), Sumru Özsoy (Boğaziçi University, Istanbul), Renate Pajusalu (University of Tartu), Chih-hsiang Shu (Academia Sinica), Usama Soltan (Middlebury College), Achim Stein (Universität Stuttgart), Nina Sumbatova (Russian State University for the Humanities), Anne Vainikka (John Hopkins University), Alena Witzlack-Makarevich (Universität Zürich), Islam Yousse (University of Tromsø)

Gabriela Alboiu and Virginia Hill

Early Modern Romanian and Wackernagel's Law

Abstract

Historical linguistic studies consider that the use of enclitics in Early Modern Romanian is due to the presence of Wackernagel's law. This characterization fits in the tradition of Indo-European and Romance historical linguistics, where the presence of Wackernagel's law is determined on the basis of phonological criteria. This paper argues that, when we approach the same data from the perspective of diachronic syntax, there is no support for this claim. We draw a distinction between encliticisation and the second position requirement for clitics, and show that the tendency for encliticisation in Early Modern Romanian is the result of syntactic operations that front the verb or/and the phrasal constituents for reasons that are unrelated to the phonological properties of clitics. We identify the triggers for such movements in discourse driven syntax.

1. Introduction

Wackernagel (1892) established a phonological principle for Proto-Indo-European languages whereby clitics (i.e., phonologically non-accented items for him) occupy the second position in the clause. The element in the first position, hosting the enclitic, could be a word or a phrasal constituent. The second position clitic rule reflects on contemporary ideas regarding the morpho-phonological properties of clitics in Medieval Romance languages, where clitics, which normally preceded the verb, were noted to follow the verb in main clauses in the absence of some other sentence initial constituent (Tobler 1875/1912; Mussafia 1888). Wackernagel's law has been successfully developed for explaining diachronic changes in Slavic languages as well (Slawski 1946 a.o.).

Romanian is genetically Romance but it had intensive language contact with South Slavic languages, in which Wackernagel's law is strongly represented (Bošković 2001; Browne 1974; Franks 2000; Pancheva 2005). While there are no studies where Wackernagel's law is

tested for Romanian, in light of its genealogy and its geography, Early Modern Romanian[1] (henceforth EMR) and the previous stages of the language are assumed, in historical linguistics, to belong to the typological group with Wackernagel's law (Frâncu 2009, following Meyer-Lübke 1890, Sandfeld 1930). From this perspective, the fact that Modern Romanian (henceforth MR) has proclitics instead of enclitics, regardless of clitic position in the clause, shows a diachronic change, whereby the language switches to the typological group that does not obey Wackernagel's law (Frâncu 2009 a.o.).

An immediate question mark arises for this assumption from the fact that EMR (at least in its written form) was under intensive influence from Church Slavonic, which displays Wackernagel's law. In this respect, the justification for Wackernagel's law in EMR could come from language contact. However, South Slavic languages show a diachronic intensification in the second position clitic requirement, compared to Old Church Slavonic (Migdalski 2006; Radanović-Kocić 1988; Tomić 1996), whereas MR displays the opposite result, by abolishing Wackernagel's law. Why would the trend be reversed in Romanian, instead of being intensified?

More precisely, MR displays proclitics on verbs, which may occur clause initially, as in (1a) – clitics in bold; enclitics are also possible with certain verb forms, such as imperatives or gerunds (1b). Note that under the umbrella of 'clitics', Romanian includes not only pronouns, but short adverbs and auxiliary verbs as well (Dobrovie-Sorin 1994), which is also the range of clitics acknowledged in Wackernagel (1892). The clitics that help us test Wackernagel's law are the EMR/MR pronouns and the auxiliaries.

(1) a. ***L-am*** *chemat pe Ion.* // ***L-aş*** ***fi*** *chemat pe Ion.*
him-have called DOM Ion // him-would have called DOM Ion
'I have called Ion.'// 'I would have called Ion.'

b. *cheamă-**l*** // *chemându-**l***
call.IMP-him // calling-him
'call him!' // 'calling him'

[1] The time span for Early Modern Romanian (EMR) starts with mid 16th c. (the time of the first written documents) up to the end of the 18th c. (Densuşianu 1901; Chivu at al. 1997).

Since procliticization is generalized to finite verbs in MR, often resulting in the presence of clitics in clause initial position, as in (1a), it is clear that Wackernagel's law is not operative in this grammar. Encliticization, however, is not prohibited, see (1b).

Contrasting with MR, EMR displays alternating locations for clitics around the finite verb when in clause initial position, either as enclitics, (2a), or proclitics, (2b).

(2) a. ***Află-să*** *această ţară să fie fostŭ lăcuit şi*
happens-REFL this country SUBJ be.SUBJ.3 been lived and
alţii într-însa…
other in-it
'This country happens to have been inhabited by others as well.'
(Ureche/Panaitescu 1958: 67)

b. ***Să vedea*** *că după acest războiu fără noroc, ce făcusă*
REFL saw that after this war without luck that made
leşii cu Stefan vodă, va fi perirea lor.
Poles.the with Stefan king will be destruction their
'You could see that after this unlucky war the Poles had made against king Stefan, this will be their destruction.' (Ureche/Panaitescu 1958: 115)

The alternating locations of the clitic in (2) is seen in historical linguistics as reflecting a transitional phase from a stage of the language with strict application of Wackernagel's law to a stage where such a law is abolished (Chivu et al. 1997, Frâncu 2009, Todi 2001).

2. Questions

The mismatch between the disappearance of Wackernagel's law, on the one hand, and the general diachronic tendencies for Romanian under the impact of the Balkan Sprachbund, on the other hand, has not escaped the attention of historical linguists. The standard explanation in this respect is as follows: it is assumed that older stages of Romanian had only a short pronominal clitic paradigm that conformed to the second position clitic requirement. Note that there are no texts to confirm this hypothesis, since the first documents date from the mid 16th century. In EMR, around the beginning of the 17th century, an innovation occurs in this clitic paradigm, where a prothetic sound *î* [ɨ] (high, central, unrounded) is added to some short clitics, making them stronger (though still accentless). From that

point on, short and reinforced clitics may be used in free alternation, as in (3), where the clitic pair occurs in the same religious text.

(3) *cu pizmă huluiia-**l*** // *cu pizmă **îl** huluiia* (Frâncu 2009: 277)
with hate cursed-him // with hate him cursed

In (3), the short clitic follows the verb, whereas the reinforced clitic precedes the verb.

Thus, the phonological innovation allows the language to by-pass the second position clitic requirement in the relevant contexts, and by analogy, the proclitic use is extended to other contexts. Accordingly, alternations in clitic placement as in (2) and (3) reflect only on changes in the phonological properties of clitics (Chivu at al. 1997, Frâncu 2009, Todi 2001).

This phonologically based hypothesis leaves room for important questions. First, relating the innovation of prothesis in clitics to the change in clitic placement from post- to pre-verbal runs into empirical inconsistencies. One inconsistency is that EMR short clitics are not banned from pre-verbal position before prothesis emerged. For example, in *cumu i-au dat împăratul slobozie* 'as to.him-has given emperor.the freedom' (Scrisoarea lui Neacşu, 1521 in Mareş et al : 51) the clitic cluster *i-au* 'to.him-has' forms a prosodic unit with the verb, not with the relative pronoun on its left, as would be expected under the theory of prothesis. In other words, proclitics were used in the 16th century in the same way they are used in MR, as *l-am* 'him-have' in (1), with the only difference that they do not often occur in the beginning of the clause. Another weak point in the prothetic hypothesis is that only a restricted series of clitics benefited from this innovation. Direct object clitic pronouns for first and second person, reflexives and third person feminine do not show prothesis. For these classes of clitics, nothing has changed in their phonology that would justify a change in their ordering. Rather, it looks like the proclitic use triggers the prothetic innovation, and not the other way around.

Wackernagel's law provides an explanation for the tendency of avoiding clitics in clause initial position, but the data show inconsistencies in this respect as well. First, the writing style in EMR displays a requirement for transition formulae. In particular, *şi* 'and' often begins a new clause, indicating its relevance to what has been said before. Clitics following this particle will then necessarily be in the second position. Note, however, that this type of 'and' is considered neutral for Wackernagel's

law in Romance (Fischer 2003; Rivero 1993), since it doesn't have a syntactic function in the clause. Accordingly, linearizations such as *Şi se-au dus în sus pre Dunăre* 'and REFL-have gone upstream on Danube.the' (Scrisoarea lui Neacşu, 1521 in Mareş et al : 51) would indicate that Wackernagel's law is absent in EMR before prothesis in clitics emerges.

Even if we were to accept *şi* 'and' as a legitimate host for enclitics, there would still be inconsistencies concerning the application of Wackernagel's law. The requirement for second position placement of clitics is systematic and rigid, whereas EMR encliticization displays random locations. For example, in (3), the first constituent is a prepositional phrase, but the clitic occurs on the subsequent verb, hence in the third, not second, position. Actually, the reinforced clitic, between the prepositional phrase and the verb, seems to obey Wackernagel's law better than the old short form. Data such as (3) clearly indicate that the rise of prothesis is independent of Wackernagel's law.

Another ordering issue concerns clitics with imperatives. Interestingly, the same word order as in (3) is maintained in MR with third person short clitics on imperative verbs, as further shown in (4). If the emergence of reinforced third person clitics is sufficient to allow for procliticization, why is this operation disallowed in this same context (4b)? Note that EMR imperatives disallow this alternation as well.

(4) a. *Cu pizmă huluieşte-**l**!*
with hate curse.IMP-him
'Curse him with hate!'

b. **Cu pizmă **îl** huluişte!*
with hate him curse.IMP

In (4), the imperative reading on the verb depends on its position in relation to the clitic: it is successful with encliticization (4a) but not with procliticization (4b). The latter triggers an indicative (assertive) reading instead of an imperative one. The contrast in (4) is not predictable under the phonological hypothesis.

To sum up, leaving aside the inconsistencies in the phonological justification for the clitic distribution in EMR, the main problems arising from classifying EMR as a second position clitic language concern the word order: either the enclitic position is not respected (since the enclitic may surface in third, fourth or other position), or the reinforced proclitic is banned from the predicted environments (i.e., in front of verbs). Since word

order is a syntactic problem, a syntactic approach is necessary to clarify clitic placement in EMR.

3. Wackernagel's law in generative grammar

The syntactic approach we propose will be couched in the framework of generative grammar, although a technical treatment of our findings will be avoided. Wackernagel's law has already been translated to syntactic constraints in this framework, notably in Rivero's studies on Romance and Slavic languages (Rivero 1993 and previous work).

Rivero's main argument is that the position of the clitic reflects not only on phonological requirements (i.e., need of an adequate lexical host), but on morpho-syntactic requirements as well, mainly relating to the nature of complementizers (henceforth, C). The complementizer phrase (CP) is hierarchically higher than the rest of the clause (i.e., Inflectional Phrase – IP), that is, [CP > IP], where IP is the domain for verb inflection. The features of the head C attract either constituent movement (e.g., for operators, topicalization etc.) or verb movement (e.g., as a structure preserving device). The second position clitic requirement has the clitic attached to any of these items, specifically, to whichever constituent is in the first position in the clause (i.e., in CP). This is illustrated with the Bulgarian data in (5), where the clitic *e* 'has' attaches either to the XP *Petur*, (5a), or the verb head (X) *procel* 'read'.

(5) a. *Petur* ***e*** *procel knigata.* **XP/constituent > clitic**
Peter has read book.the
'Peter has read the book.'

b. *Procel* ***e*** *knigata.* **X/verb > clitic**
read has book.the
'He has read the book.' (from Rivero 1991: 323)

Importantly, the two items cannot co-occur in front of the clitic, as seen in (5c), which Rivero mentions is due to a syntactic restriction on doubly filled CP, independent of phonological restrictions on clitics.

(5) c. **Petur procel* ***e*** *knigata.* **XP > X > clitic**
Peter read has book.the

When verb movement to C applies, the verb may by-pass an auxiliary, as long as the auxiliary is a clitic, which is the case in South Slavic, as in (6) for Serbo-Croatian. This is known as Long Head Movement (henceforth LHM).

(6) a. *Ja **sam** citao knjigu.* **- LHM (XP > clitic)**
I have read book.the
'I have read the book.'

b. *Citao sam knjigu.* **+ LHM (X > clitic)**
read have book.the (from Rivero 1991: 330).

Only non-finite verb stems may undergo LHM, as in (6b). That is, past participles in relation to 'have'/'be' auxiliaries (V > pronouns-'have'/'be'), or infinitives in relation to 'will' auxiliaries (V > pronouns-'will'). If a finite (tense inflected) verb stem moves to C, it is argued to do so for reasons having to do with structural requirements that amount to Verb Second (V2), not LHM. In other words, V2 and the second position clitic requirement are in complementary distribution (which is predicted in Wackernagel 1892, and discussed in Anderson 1993).

Rivero identifies two systematic properties for LHM: 1. It is restricted to root clauses only, since subordinate clauses have complementizers, which the clitics can use as phonological hosts; 2. Fronting of phrasal constituents and verb movement are in complementary distribution in LHM, since only one of them can be clause initial.

Technically, for LHM, the word order is either [V > pronouns/Aux > XP], or [XP > pronouns/Aux > V]. EMR shows both orders, as seen in (7).

(7) a. *Pus-**au** şi pe trii boiernaşi de au tras* **V > clitic**
made-has also DOM three lord.like.PL to have push.IND
'he has also made three minor lords to push' (Neculce/Iordan 1955: 106)

b. *Aşe **au** încetat turcii de a fugi* **XP > clitic**
thus have stopped Turks.the of to run.INF
'thus the Turks stopped running' (Neculce/Iordan 1955: 284)

This word order indicates the need to investigate the EMR data from the LHM angle, especially because Rivero (1993) includes Romanian in the LHM group, with the inference that the second position clitic requirement applies in this language.

4. EMR, LHM and Wackernagel's law

Most of our data[2] come from the Moldavian Chronicles written directly in Romanian in the 17th and the 18th centuries. Comparison with the use of clitics in translated documents will be resorted to as necessary to underline the influence of the Slavonic word order.

The main point of this section is that EMR displays LHM, but no evidence for Wackernagel's law. The arguments for absence of a second position clitic requirement in EMR are as follows:

(i) Enclitics are not restricted to the second position in the clause. More precisely, LHM and topicalization may co-occur, as shown in (8). This should be ruled out under Wackernagel's law, since either operation can provide the phonological host for the clitic.

(8) a. *[Într- acei păstori ce au nemeritu locul acesta]*
among those shepherds who have found place.the this
fost-au *şi Dragoş, carileau venitu de la Maramoroş,*
been-has and Dragos who has come from at Maramures
'Among the shepherds who found this place there was also Dragos, who alighted from Maramures.' (Ureche/Panaitescu 1958: 72, 1359)

b. *[Deciia] [Stefan vodă]* ***strîns-au*** *boierii ţării…*
so Stefan king gathered-has lords.the country.the.of
'So king Stefan has gathered the lords of the country…'
(Neculce/Iordan 1955: 91)

c. *Pre acest Hrize foarte îl iubea Costandin-vodă,*
DOM this Hrize much him loved Costandin-king
[şi] [de taină crediincios] ***făcutu-l-au*** *boiarin,*
and of counsel trustful made-him-has lordship
'King Constantin liked this Hrize man a lot, and he made him a lord trusted with counselling.' (*Letopiseţul Cantacuzinesc/* Onu 1970: 168)

(ii) EMR displays LHM in subordinate clauses as well, as in (9a), although the complementizer *că* 'that' is available as host to the clitic. (9b) shows that the complementizer may, indeed, support enclitics. LHM in

[2] The main source of data is the Moldavian chronicles (complete corpus) because they provide the most extensive texts written directly in Romanian. Other sources (religious, official texts) are used as well, as needed. Note that the clitic morpho-syntax is not subject to regional variation in EMR/MR, so, for this reason, the data may come from either the northern or the southern parts of the country.

subordinate clauses goes against the predictions of both Wackernagel and Rivero, and it appears in the early documents, when Wackernagel's law was supposed to be observed more strictly than in late EMR. For example, in Dosoftei (text dating from 1679), *că* 'that' may head matrix clauses, providing the clause initial support, but this seems to be orthogonal to the clitic-verb inversion, which may or may not apply (9c, d).

(9) a. *Scrie letopiseţul nostru* ***[că*** *in anii 6947…* ***intrat-au***
writes chronicle.the ours [that in years 6947 gotten-has
în ţară oaste tătărască
in country army Tatar
'Our chronicle writes that, in 6947, Tartar army has invaded the country.'
(Ureche/Panaitescu 1958: 83, 1439)

b. *Bine face* ***că****-i ocăreşte*
well does that-them scolds
'he does well to scold them' (Neculce/Iordan 1955: 104)

c. ***Că rădica-să-va*** *de pre pământ viaţa Lui, totdeauna,*
for rise-REFL-will of from earth life.the His always
acmu şi pururea şi-n vecii de veci.
now and for.ever and-in eras of eras
'For His life will rise from the earth always, now and for ever.' (Dosoftei/ Ursu 1980: [24])

d. ***Că Ţî să*** *cuvine toată slava, cinstea şi închinăciunea,*
for to.you REFL befits all glory honor and supplication
'For You deserve all the glory, honor and supplication.'
(Dosoftei/Ursu 1980: [44])

(iii) After phrasal constituents, short clitic pronouns may occur simultaneously in preverbal and in post-verbal position (i.e., double spell-out of the clitic in two positions), as in (10a). Alternatively, the clitic pronoun may surface only as an enclitic, whereas the auxiliary is proclitic (10b).

(10) a. *şi* ***i-au*** ***închisu-i****,…*
and them-has jailed-them (Neculce/Iordan 1955: 153)

b. *pe alţi mulţi boieri munteneşti* ***au*** ***prinsu-i****…*
DOM other many lords Wallachian has caught-them
'He caught many other Wallachian lords' (Neculce/Iordan 1955: 150)

The clitic in (10a) has a reinforced alternative with prothetic *î*, but the reinforced form is disallowed in this context. Wackernagel's law cannot accommodate double spelling of clitic pronouns, nor the breaking of the clitic clusters as in (10b). LHM also fails to account for these configurations, since there is encliticization without verb movement to C.

(iv) Translation mistakes from Church Slavonic indicate that Wackernagel's law was foreign to the grammar of EMR writers. Romanian translators strove to keep as close as possible to the word order of the Slavonic original. Enclitics seem to have made this endeavor very difficult. Consider (11): the entire clitic cluster (i.e., pronoun and auxiliary) is repeated, once in enclisis, once in proclisis, in addition to allowing it to co-occur with topicalization.

(11) *părinţii noştri….i-ai mîntuitu-i-ai*
parents.the ours CL.3PL.DAT-AUX.2SG bless.PRTC-CL.3PL.DAT-AUX.2SG
'you blessed our parents' (PH.xxi, 5 *apud* Densuşianu 1997: 707)

Such confusions indicate hyper-corrections and the translator's lack of intuition in handling the enclitics and, presumably, the Wackernagel's law.

(v) Another example comes from the use of negation. Since negation serves as a phonological host for clitics, clitic-verb inversion does not apply in negative clauses, nor does LHM (Rivero 1991), as in (8) or (9) above. EMR translators, however, show confusion in this respect as well, as seen in (12), where encliticization is uncertain. (12a) shows lack of clitic-verb inversion in the presence of negation, as expected (which also holds for MR), but (12b) shows atypical encliticization, probably under the pressure of Slavonic clitic ordering.

(12) a. *nu **vă** temereţi*
not refl fear.IMP.2PL
'don't be afraid' (Chivu at al. 1997: 342)

b. *nu ciudireţi-**vă***
not wonder.IMP.2PL-REFL
'don't wonder' (Chivu et al. 1997: 244)

The examples from (8) to (12) are sufficient to show that Wackernagel's law is not present in EMR, even in its early stages. The way this rule is handled in the century preceding the Moldavian chronicles would be best characterized as hectic. To us, this means that a distinction must be drawn

between Wackernagel's law and encliticization, only the latter being attested in the written documents.

That being said, LHM is present in EMR (7a), and so is V1 (finite verb movement to the first position in the clause), see (2a), which yields the encliticization on finite verbs. However, the reason for LHM and V1 is not Wackernagel's law but some other factor, to be determined in the rest of this paper. What we have established so far is that the change in clitic placement from EMR to MR does not concern the loss of Wackernagel's law but the loss of some other parametric setting that has to do with the location of the verb, not with the location of the clitic.

5. Tests for V2 in EMR

If Wackernagel's law does not apply to EMR, then encliticization is not distributionally constrained beyond the availability of an adequate phonological host, anywhere in the clause. Consequently, the alternation between proclitics and enclitics on finite verbs as in (2) cannot be accounted for by the phonological properties of clitics, but by syntactic triggers that concern the location of the verb in the hierarchical structure of the clause. Thus, the next step in our analysis is to account for the factors that trigger LHM and finite verb fronting to CP. In this section, we discuss the possibility of having V2 in EMR, which would explain why finite verbs move so high at the left periphery of clauses.

The theoretical background for this discussion will be extended from the CP > IP hierarchy to the cartographic representation of the CP. In particular, Rizzi (1997) points out that discourse pragmatics is encoded in the left periphery of clauses, in the same area where complementizers occur for clause typing. Thus, he splits the CP field to accommodate the operations at the left periphery, a shown in (13b), with TopP being potentially recursive. In (13a), we provide an example from MR to illustrate the word order.

(13) a. *Zice [că la mare cu Ion până la urmă să se ducă,*
says that to sea with Ion up to end SUBJ REFL go
nu cu Maria.]
not with Maria
'He says that, in the end, she should go to the seaside with Ion, not with Maria.'

b. ForceP > TopP > FocP > ModP > FinP (>NegP)
that > to the sea > with Ion > in the end > SUBJ > (nu)
> TP > vP
REFL.go > …

The illustration in (13) resorts to an embedded clause[3]. Root clauses may display the same word order, but do not have a specific complementizer in Force in declaratives. However, the existence of ForceP is attested in interrogative clauses when the interrogative word (*wh*-word) is fronted to that level.

With respect to the position of clitics within the cartography in (13b), they are located in T, the head of TP, which equates IP in our theoretical background in the previous sections. This follows the more economical Minimalist system (Chomsky 1995). In this hierarchy, most inflectional features are associated with T (e.g., mood, tense, agreement). Kayne's (1991) location of clitic pronouns in Romance can now be converted to T. This extends to EMR, since clitics are hierarchically lower than negation, and their host is the verb, not the negation (which is not a clitic; see Isac & Jakab 2004). Clitic auxiliaries were shown to be in Agr/I in the Government-Binding clause hierarchy (Dobrovie-Sorin 1994); since [agr] features are associated with T in (13), auxiliaries are also in T. Therefore, in our tests V > clitic (pronoun and/or auxiliary) order indicates verb movement above TP, which could be to Fin or to a higher head.

The word order in (13) applies in the syntax of EMR as well, every time the constructions have proclitics on verbs. This is illustrated in (14).

(14) Matrix clause = TopP > FocP > FinP/IP
a. *Iar VasiIie-vod nici cu un megiieş, precum am apucatu*
but Vasilie-king.TOP not with one citizen.FOC as have witnessed
şi noi aceia domnie, viaţă bună n-au avut,…
and I that reign life good not-has had
'But king Vasilie, there was not one single citizen he pleased, as I myself witnessed during that reign.' (Costin/Panaitescu 1979: 89)

[3] For more tests and information on the cartography of Romanian clauses see Alboiu (2002) and the references therein.

Embedded clause = ForceP > TopP > FocP > FinP/IP

b. *Ţara Muntenească, într-acesta anŭ, vara, la mare*
Kingdom Vallach.top in-this year.TOP summer.TOP at great
răutăţi era de turci, că din doao părţi avîndu oşti
damage.FOC was by Turks for of two sides.FOC having army
Impărăţia Turcului asupra Crăiei Ungureşti, o samă de oşti
Empire Turk.the.of against Kingdom Hungarian a some of army
despre Buda,iară altă oaste asupra Ardealului avè,
towards Buda and another army towards Ardeal had
[că şi împăratul nemţescŭ oştile lui într-acolea
because.FORCE and king.the German.TOP armies his.TOP there.TOP
împrotiva turcilor era orînduite.]
against Turks-FOC were positioned
'That year, in the summer, the Kingdom of Wallachia suffered great damage from the Turks, for the Ottoman Empire had his army against the Hungarian Kindgom, split in two: some of it was directed against Budapest, some of it was directed against Ardeal, because it was against the Turks that the German king had his army there positioned as well.' (Costin/Panaitescu 1979: 18)

Rivero's LHM means verb movement to Force, since LHM occurs in complementary distribution with V2. Therefore, for the alternation between proclitics and enclitics on EMR verbs, (13) tells us that the verb could either stay lower in the hierarchy (in I) or move higher up. MR lost the high verb movement option in most environments.

Let us now consider the possibility that the order [V > clitic] follows from a requirement on V2 in EMR. This exercise is necessary because we ruled out Wackernagel's law, which might suggest the likelihood of V2. In (13), V2 in declarative clauses means movement of the verb to Force, plus the presence of a preceding constituent in ForceP (e.g., in German). Such co-occurrence is possible in EMR (see 15b V2), but it is neither obligatory nor systematic. That is, the verb may also occur as V1 (no preceding constituent) or V3 (two preceding constituents) in similar contexts (15a, b).

(15) a. ***Asemănă-se*** *acel boiar cu Iuda care au vândut pre domnu-său.* V1
liken-REFL that lord with Juda who has sold DOM master-his
'That lord resembles Juda who sold his master.'
(*Letopiseţul Cantacuzinesc*/Onu 1970: 157)

b. *[Şi] [într-acea vreme lăcuind el acolo], [pre 2 din fraţi*
and in-that time living he there DOM 2 of brothers
*carei mersese cu dânsul], **trimise-i** către Dumnezeu,* V3
who went with him sent-them to God
*iar [pre al treilea] **lăsă-l** să meargă în cetatea Solunului.* V2
while DOM the third let-him SUBJ go to fort.the Solun.the.of
'And during that time, when he was living there, he sent to God two of the brothers who came with him, and the third, he let him go to the Solun fort. (*Letopiseţul Cantacuzinesc*/Onu 1970: 161)

A further mismatch between EMR word order and V2 appears in *yes-no* interrogatives, as in (16). In V2 languages, these constructions display the verb in clause initial position, without a preceding constituent. EMR has the same restriction on the location of the verb, but it extends the restriction to non-finite verb stems (i.e., LHM in 16b, c), on a par with the finite stems (16a).

(16) a. ***Cunoşti-mă** pre mine, au ba?*
know.2SG-me DOM me or not
'Do you recognize me or not?' (Neculce/Iordan 1955: 120)

b. *sta în cumpene şi să mira ce or face, **fugi-or**,*
stayed in doubts and REFL wondered what should do run-should
*au **spune-or** lui Grigorie-vodă?*
or say-should to Grigorie-king
'they were in doubt and wondered what they should do: should they run or should they tell king Grigorie?' (Neculce/Iordan 1955: 343)

c. ***Pus-au** oamneii săi şi puşcile au ba?*
put-has men.the his and guns.the or not
'Did he install his men and guns, or did he not?' (Costin/Panaitescu 1979: 124)

V2 acts on finite verbs only, so the free alternation with LHM, as in (16), indicates that despite the location of the verb in the clause initial position, the trigger and, possibly, the level of verb movement must be different from what happens in V2 constructions.

The data in (15) and (16) indicate that the order V > clitics in EMR does not follow from a V2 pattern. In particular, what we see in these data is a general fronting of the verb, on an optional basis (see examples in (2)), irrespective of its finite or non-finite stem. The result of this general verb fronting is encliticization, which may misleadingly suggest the application of Wackernagel's law.

6. Verb movement in EMR

If neither Wackernagel's law nor V2 justify high verb movement in EMR, what other factor can explain it, while also accounting for its optional occurrence, as shown in (2)? The answer we provide in this section is that EMR verb movement targets the Focus head in (13), not Force, as in Rivero (1993). Verb movement to Focus is triggered for discourse purposes, not as a structure-preserving device. That is why the movement is optional, depending on whether certain discourse features (i.e., a focus operator feature optionally associated with the CP field) are present or not in the derivation.

Our analysis will cease to distinguish between LHM and V1, since both verb forms seem to behave similarly for the purpose of movement. Instead, we shall attempt to provide a unified analysis for the instances where the word order is [V > clitic], and also, for the conditions that allow the [clitic > V] order. In this respect, we first survey the distribution of the [V > clitic] order in root clauses, according to the clause type involved.

Declarative clauses display an optional [V > clitic] order, as shown in (2). On the other hand, interrogative clauses differ, as the [V > clitic] order is either obligatory (i.e., with *yes-no* interrogatives, as in (16)) or systematically excluded (i.e., with *wh*-interrogatives, as in (17)).

(17) ***Cum ar hi*** *împăratu să hie drag tuturora?*
how would be king.the SUBJ be dear all.DAT
'How should the king be to be loved by all?' (Costin/Panaitescu 1979: 33)

Negation systematically disallows the [V > clitic] order, even in the contexts where such order is otherwise obligatory (i.e., *yes-no* questions). This is shown in (18): in (18a) *nu* 'not' precedes a proclitic cluster at the beginning of an assertive clause; in (18b) it does the same in a *yes-no* interrogative; in (18c) the negation and clitics follow a *wh*-word.

(18) a. ***Nu i-au mai*** *trebuit istoric strein,*
not to.him-has more needed historianforeign
'He no longer needed a foreign historian' (Neculce/Iordan 1955: 104)

b. ***Nu v-am*** *spus ca acesta om de boierie nu este?*
not to.you-have said that this man of lordship not is
'Haven't I told you that this man is not worthy of lordship?'
(Costin/Panaitescu 1979: 65)

c. *Dară cui nu ieste urît a muri,* ***cine n-ar***
but to.whom not is hateful to die who not-would
pofti *să vieţuiască?*
like SUBJ live
'But who does not hate dying, who wouldn't want to live?' (Ureche/Panaitescu 1958:191)

Constituents with a topic reading may occur in front of the [V > clitic] string in declaratives, as in (15b). In *wh*-interrogatives, such constituents must precede the *wh*-phrase, as shown in (19); however, these constructions always have a [clitic > V] order.

(19) *[după sutele de ani]* ***cum*** *să vorŭ putea şti*
after hundreds of years how REFL will can know
poveştile adevărate, de atîtea vacuri?
stories.the true of so.many centuries
(Costin/Panaitescu 1979: 189)

The evidence considered up to this point allows us to determine the target for verb movement by using the hierarchy in (13). More precisely, the [V > clitic] order indicates that the verb moves out of TP/IP (i.e., it is higher than the location for clitics). Although topic constituents can precede the [V > clitic] string, the landing site for the verb cannot be Top, because V-to-Top entails sequences such as V > *wh*-word, which are ungrammatical in EMR/MR (i.e., 'yesterday came how he?'). That leaves us with two possible targets for verb movement: either Fin or Foc.

FocP in (13) is associated with contrast and other type of foci that involve operators and propositional scope. In this respect, we do not expect information focus in this position. Studies in the semantics of focus identify four types of focus operators: contrastive focus (CF), verum focus (VF), question focus (QF) and emphatic focus (EF) (Höhle 1992; Krifka 2007; Richter & Mehlhorn 2006). All these types of focus are present in EMR, some being realized through constituent fronting, some through verb movement, as we shall see in (20) to (23). The point is that structurally, a constraint that precludes two items to fill out the Focus phrase (because only one item can check the operator feature and link the variable) triggers a configuration where either constituent fronting to FocP or verb movement to Foc head may occur, but not both at the same time. This complementary distribution applies systematically in our constructions, and provides a sure indication that the verb targets the Foc head, not Fin, because the latter

would have allowed for co-occurrence of contrastive focus constituents and [V > clitic] order.

The following examples provide evidence for the way in which FocP is lexicalized in EMR. First, consider *wh*-questions, such as presented in (18c), (19) and further in (20).

(20) *Deci trei domnii căte 500-600 de pungi de bani la*
so three reigns.TOP each 500-600 of purses of money at
înnoituri, tot într-un anu, ***cum*** *au putut hi bine?*
deadlines.top all in-one year.top how.QFOC has could be well
'So during three reigns, 500-600 purses of money for each, per year, at deadlines, how could that be well?'(Costin/Panaitescu 1979: 100)

In (20), topic constituents precede the *wh*-phrase *cum* 'how', the latter checking the QF operator. Since the *wh*-phrase is in FocP, verb movement will not take place. Indeed, [clitic > V] is the only acceptable order in these constructions.

Next, the CF operator involves constituent fronting, as in (21). The word order is the same as in (20), with a topic constituent preceding the constituent with contrastive focus. Predictably, the CF constituent occupies the same structural position as *cum* 'how' in (20) and systematically entails a [clitic > V] order.

(21) *De care lucru cunoscînd Stefan vodă că ajutoriul* ***nu de aiurea***
of which thing knowing Stefan king that help.the not of anywhere.CFOC
i-*au fost, ci numai de la Dumnezeu si de la*
to.him-has been but only of from God and of from
Preacurata Maica sa,
Pristine Mother his…
'Knowing king Stefan from this that help did not come from nowhere, but onlyfrom God and his beloved Mother…' (Ureche/Panaitescu 1958: 95)

The VF operator triggers the derivation of *yes-no* questions. Generally, there is no phrasal constituent fronted to FocP for the purpose of VF, so the derivation resorts to verb movement to Foc, as shown in (16), which is another way of checking the operator. Hence, the [V > clitic] order is systematic, as mentioned for those examples. However, when a compatible constituent is used to check the VF operator, instead of verb movement – e.g., *adeverat* 'truly' in (22) - the word order reverts to [clitic > V]. Such examples are edifying for the mapping of verb movement, since they match

the predictions arising from the syntactic constraints applying to the other types of foci operators.

(22) *în dooă-trei rînduri au trimis să vadză,* ***adeverat*** *au sosit?*
in two-three times has sent subj see truly.VFOC have arrived
'he sent [someone] two-three times to see, did they TRULY arrive?' (Costin/Panaitescu 1979: 118)

Finally, the EF operator may also be labeled *narrative focus*, and occurs in declarative sentences. In general, this happens when a new event is introduced in the story, as in (23a). It may also occur any time the narrator needs to highlight an event, as in (23b). As shown in (23a), the [V > clitic] order occurs in the introduction of the new event, but not in the following coordinated clause. The coordinated verb is only elaborating on the introduced event, and displays the [clitic > V] order[4]. In (23b), LHM is not motivated on a grammatical basis, since there are constituents preceding the clitic, but only on pragmatic grounds: the inverted verb needs highlighting for the interpretatio. In these environments, the verb stem moves to Foc, as argued for the [V > clitic] order in general.

(23) a. *Deciia Stefan vodă* ***strîns-au*** *boierii ţării […]*
therefore Stefan king gathered-has lords country.the.of
şi i-au întrebatu pre toţi
and them-has asked DOM all
'Therefore, king Stefan gathered the lords of the country and asked them all' (Ureche/Panaitescu 1958: 91)

b. *Pentr-acea vrăjmăşie şi groază ce-i împlusă*
for-that enmity and terror which-him poured
inema diiavolul de lăcomia ce avè, ***urît-au*** *toţi pre Duca-vodă.*
heart.the Devil.the of greed.the that had hated-have all DOM Duca-king
'Because of that enmity and terror which the Devil poured into his soul, everybody HATED king Duca.' (Neculce/Iordan 1955: 155)

[4] We have already mentioned that syntacticians consider 'and' to be neutral for Wackernagel's law. One may object that 'and' qualifies as a phonological host, and should, therefore, count for the application of Wackernagel's law or of encliticization in general. This is true for those languages where Wackernagel's law applies and involves the level of the word, in additional to the level of the constituent (e.g., Serbo-Croatian in Browne 1974). There is no evidence that EMR uses the word for encliticization, since, besides *şi* 'and', all the phonological hosts for encliticization are demonstrably phrases (including CP) or the verb itself.

To sum up, all four types of operator foci are realized in EMR. Two of them are realized through constituent movement to FocP (i.e., CF and QF), making the operator visible; the other two (VF and EF) are realized through verb movement to Foc, the operator being null. The negation, illustrated in (18) is able to take over the operator checking function, and move to Foc, instead of the verb. That is why in the presence of the negation the only possible order is [clitic > V]. This analysis shows that the [V > clitic] order in EMR does not involve real LHM or V2, in the sense that such movements target Force, whereas the EMR verb movement targets Foc. The former are justified through structure preserving constraints, the latter through the presence of discourse features with operator properties. The structure preserving movement is obligatory, the discourse driven movement is optional (insofar as the introduction of the respective pragmatic features in the derivation is optional).

7. Conclusions: Diachronic change

This paper aimed to demonstrate that Wackernagel's law is not operative in EMR. What EMR has is encliticization on verbs, arising from syntactic triggers. We identified these triggers as being the focus feature with operator properties, encoded high in the left periphery of clauses which, in certain contexts, trigger verb movement above the location for clitics. A host of peculiarities concerning the word order were accounted for in this way, while it was also shown that the nature of verb movement in EMR is different from LHM, V1 or V2 (which are all structure preserving operations).

In terms of diachronic changes, MR lost the [V > clitic] order in declaratives and in *yes-no* questions, as shown in (24a, b).

(24) a. ***Te- ai dus*** *la mare? vs.* ****dusu-te-ai*** *la mare?*
REFL have gone to sea gone-REFL-have to sea
'Did you go to the sea?'

b. *Din cauza asta, toată lumea* ***l-a URÂT*** *pe voievod.*
from cause this all people him-has hated DOM king
'Because of this everyone hated the king.'

c. ***Bătu-te-ar*** *norocul!*
beat-REFL-would luck.the
'What a scoundrel!' (let the luck beat you)

The loss of the V > clitic order in matrix clauses has nothing to do with Wackernagel's law. This change concerns the loss of verb movement to Foc in the respective configurations, and the interpretation corresponding to the lost syntactic operations is now recuperated from prosody only. Some traces of [V > clitic] order survive in idiomatic exclamatives, as in (24c).

References

Alboiu, Gabriela (2002) *The Features of Movement in Romanian*. Bucharest: EUB.

Anderson, Stephen (1993) Wackernagel's revenge. *Language* 69 (1): 68–98.

Bošković, Željnek (2001) *On the Nature of the Syntax-Phonology Interface*. Amsterdam: Elsevier.

Browne, Wayles (1974) On the problem of enclitic placement in Serbo-Croatian. In Richard D. Brecht & Catherine V. Chvany (eds.), *Slavic transformational syntax*, Michigan Slavic Materials 10, pp. 36–52. Ann Arbor: University of Michigan.

Chivu, Ion & Ion Gheţie (ed.) (1997) *Istoria limbii române literare. Epoca veche (1532–1780)*. Bucharest: Editura Academiei.

Chomsky, Noam (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

Densuşianu, Ovidiu (1997/1901) *Histoire de la langue roumaine*. Bucharest: Editura Grai şi Suflet – Cultura Naţională.

Dobrovie-Sorin, Carmen (1994) *The Syntax of Romanian: Comparative Studies in Romance.* Berlin: Mouton de Gruyter.

Fischer, Susan (2003) Rethinking the Tobler-Mussafia Law. *Diachronica* 20 (2): 259–288.

Frâncu, Constantin (2009) *Gramatica limbii române vechi (1521–1780).* Iaşi: Demiurg.

Franks, Steven (2000) Clitics at the Interface. In Frank Beukema & Marcel den Dikken (eds.), *Clitic Phenomena in European Languages*. Amsterdam: Benjamins.

Höhle, Tilman (1992) Über Verum Fokus in Deutschen. *Linguistische Berichte Sonderheft* 4: 112–141.

Iordan, Iorgu (ed.) (1955) *Ion Neculce, Letopiseţul Ţării Moldovei*. Bucharest: Editura de Stat.

Isac, Dana & Jakab, Edith (2004) Mood and force features in the languages of the Balkans. In Olga Mišeska-Tomić (ed.), *Balkan Syntax and Semantics*, pp. 315–338. Amsterdam: John Benjamins.

Kayne, Richard (1991) Romance clitics, verb movement and PRO. *Linguistic Inquiry* 22: 647–687.

Krifka, Manfred (2007) Basic Notions of Information Structure. In Caroline Fery, Gisbert Fanselow & Manfred Krifka (eds.), *Interdisciplinary Studies of Information Structure 6*, pp. 1–12. Potsdam: Universitätsverlag Potsdam.

Mareş, Alexandru (ed.) (1994) *Crestomaţia limbii române vechi*. Bucharest: Editura Academiei.

Meyer-Lübke, Wilhelm (1890) *Grammatik der romanischen Sprachen*. Leipzig: Reisland.

Migdalski, Krzysztof (2006) *The Syntax of Compound Tenses in Slavic.* Utrecht: LOT.
Mussafia, Alfredo (1888) Enclisi o proclisi del pronome personale atono quale oggeto. *Romania* 27: 145–6.
Onu, Liviu (ed.) (1970) *Cronicari munteni*. Bucharest: Editura Stiintifica.
Panaitescu, Petre P. (ed.) (1958) *Grigore Ureche, Letopiseţul Ţării Moldovei*. Bucharest: Editura de Stat.
—— (ed.) (1979) *Miron Costin, Letopiseţul Ţării Moldovei*. Bucharest: Editura Minerva.
Pancheva, Roumyana (2005) The Rise and Fall of Second-Position Clitics. *Natural Language and Linguistic Theory* 23 (1), 103–167.
Radanović-Kocić, Vesna (1988) The Grammar of Serbo-Croatian Clitics: A Synchronic and Diachronic Perspective. Ms. University of Illinois, Urbana.
Richter, Nicole & Mehlhorn, Grit (2006) Focus on contrast and emphasis: Evidence from prosody. In Valeria Molnar & Susane Winkler (eds.), *The Architecture of Focus*, pp. 347–373. Berlin: Mouton de Gruyter.
Rivero, Maria Luisa (1991) Long Head Movement and Negation: Serbo-Croatian vs Slovak and Czech. *The Linguistic Review* 8: 319–351.
—— (1993) LHM vs V2 and null subjects in Old Romance. *Lingua* 89: 217–245.
Rizzi, Luigi (1997) The fine structure of the left periphery. In Liliane Haegeman (ed.), *Elements of Grammar*, pp. 281–339. Dordrecht: Kluwer.
Sandfeld, Kristian (1930/1968) *Linguistique balcanique: problèmes et résultats*. Paris: Klincksiek.
Slawski, Franciszek (1946) *Miejsce enklityki odmiennej w dziejach jrzyka bulgarskiego* [The Place of the Inflected Enclitic in the History of Bulgarian]. Krakow: Polska Akademia Umiejetnosci.
Tobler, Adolf (1875/1912) *Vermischte Beiträge zur französischen Grammatik*. Erste bis fünfte Reihe. Leipzig: S. Hirzel.
Todi, Aida (2001) *Elemente de sintaxă românească veche*. Iaşi: Editura Paralela 45.
Tomić, Olga Miseska (1996) The Balkan Slavic Clausal Clitics. *NLLT* 14 (4): 811–872.
Ursu, Neculai A. (ed.) (1980) *Dosoftei, Dumnezăiasca liturghie* (*1679).* Iaşi: Mitropolia Moldovei şi Sucevei.
Wackernagel, Jacob (1892) Über ein Gesetz der indogermanischen Wortstellung. *Indogermanische Forschungen* 1: 333–436.

Abbreviations

AUX = auxiliary
CL = clitic
DAT = Dative,
DOM = differential object marker
FOC = focus
IMP = imperative
PL = plural

PRTC = past participle
REFL = reflexive
SG = singular
SUBJ = subjunctive
TOP = topic

Contact Information:

Gabriela Alboiu
Linguistics, DLLL
York University
4700 Keele Street
Toronto, ON, M3J 1P3
Canada
e-mail: galboiu(at)yorku(dot)ca

Virginia Hill
Humanities and Languages
University of New Brunswick
P.O.Box 5050
Saint John, NB, E2L 4L5
Canada
e-mail: mota(at)unb(dot)ca

Pauli Brattico

Structural Case Assignment and Phi-Agreement in Finnish[1]

Abstract

Structural Case assignment, agreement in phi-features, and the EPP-movement are related to each other. However, their exact syntactic relationship remains controversial. The matter is examined here from the point of view of Finnish morphosyntax. Finnish provides close to an ideal language for this purpose, as it has fifteen case forms and full syntactic phi-agreement on verbs (finite and nonfinite), nouns, adjectives and prepositions. In addition, Finnish exhibits certain more exotic Case assignment phenomena, among them the long distance Case assignment, quantificational Case and aspectual Case. It is argued that the recent minimalist theory of Agree provides a sound starting point to explain the phenomena, but requires certain modifications to fully capture the Finnish facts. Specifically, it will be argued that Case is not a reflex of an uninterpretable phi-set probe, as posited in the standard theory, but, instead, it is a reflex of a more abstract phi-specification feature of functional heads. In addition, locality restrictions on Agree posited in the standard theory are argued to be too strong.

1. Introduction

Not long ago Case assignment and phi-agreement (that is, agreement in person, number and gender features) were explained in the generative grammar by relying on strictly local phrase structure configurations, such as the Spec-Head configuration (Chomsky 1993). An example of this specimen is agreement between the grammatical subject and finite verb, as

[1] Acknowledgements. This study was supported by a sabbatical grant from the Academy of Finland for a scholarly visit to the University of Salento and University of Firenze, Italy. I thank Professor Rita Manzini and Professor Mirko Grimaldi for providing me ideal conditions for contemplating problems of Finnish syntax. Saara Huhmarniemi, Jukka Purma and Tiina Keisanen have provided feedback and comments on the draft version(s) of this paper, and my work with Anne Vainikka (as well as her own seminal work on Finnish Case) has been very influential.

shown in the example (1). The subject is assigned the nominative Case by the finite verb, and the verb receives phi-features from the subject.

(1) *He arrives.*
nom↔3sg

According to this view, agreement takes place under strictly local configurations. In addition, Case assignment and agreement are clearly "paired" in some way. The subject that agrees with the verb is assigned the nominative, so that agreement and Case assignment go in tandem (see Adger & Harbour 2008).

A finite verb can nevertheless agree with a phrase that has not been raised to its Spec. English *there*-expletive and Finnish subject-object inversion provide clear examples. In the example (2a), there is agreement between the finite verb and *three men*. In the example (2b), there is agreement between the postverbal thematic subject *Graham Greene* and the finite verb, while the thematic object occupies the grammatical subject position (see Holmberg & Nikanne 2002).[2]

(2) *a. There seem to have arrived three men.*
pl pl

b. Tämän kirjan kirjoitti Graham Greene.
this.ACC book.ACC wrote.3SG Graham.NOM Greene.NOM
'Graham Greene wrote this book.'

Data such as this led linguists to give up the strict locality requirement, while keeping the pairing between Case assignment and phi-agreement (Chomsky 2000, 2001, 2008). Local checking is substituted by an operation Agree, which does *not* rely on the Spec-Head configuration. Movement to Spec follows if probe's EPP-feature is not satisfied otherwise. In the example (2a), it is satisfied by the expletive *there*, in (1) it is satisfied by moving the subject DP and in (2b) it is satisfied by moving the object. These assumptions are further illustrated in (3a–b). In (3a), the finite verb agrees

[2] I will use the following abbreviations in this article: A = A-infinitival; ACC = accusative Case (any variant); ACC(n) = genitive-looking n-accusative; ACC(0) = nominative-looking zero-accusative; ABL = ablative Case; GEN = genitive Case; ELA = elative Case; INE = inessive Case; KO = question clitic; MA = ma-infinitival; NOM = nominative Case; NUT = past participle; PASS = impersonal passive; PRT = partitive Case; PX = possessive suffix/agreement marker; TRS = translative Case; VA = va-infinitival.

with the nominative subject but the subject remains *in situ* inside the *v*P in a postverbal position, while in (3b) the agreeing nominative subject also raises to the subject position Spec,TP (specifier of the tense phrase) or Spec,FP (specifier of the finiteness phrase).

(3) *a. Tämän kirjan kirjoitti Graham Greene.*
this.ACC book.ACC wrote.3SG Graham.NOM Greene.NOM
'Graham Greene wrote this book.'

b. Graham Greene kirjoitti tämän kirjan.
Graham.NOM Greene.NOM wrote.3SG this.ACC book.ACC
'Graham Greene wrote this book.'

An important feature of this model is that once EPP is factored out of Agree, the case assigner (e.g. a finite verb) will c-command[3] the case assignee, and the phi-assigner (e.g. the nominal) will be c-commanded by the phi-assignee (the verb). Chomsky (2000: 122) proposes that structural Case (e.g. nom, acc) is a "reflex" of c-commanding phi-set probe (4).

(4) *Condition on Phi-Agree (CPA)*
Structural Case is a reflex of a c-commanding uninterpretable phi-set probe.

I will argue here that this generalization applies, in a surprisingly general fashion, to structural Case assignment in Finnish and unlocks the relation between phi-agreement and structural Case assignment in this language.

One might be surprised to read an article-length treatment in the defense of (4), since this principle has established itself as part of the mainstream view during the last ten years. However, principle (4) is controversial what comes to the description and explanation of Finnish case, specifically (see Vainikka 1989, 2011). Yet my main concern and the principal motivation for this study lies elsewhere. It is perhaps useful to spell out these reasons before we move on. Finnish is a language with a particularly rich Case and agreement system, and that system gives rise to several rather perplexing Case assignment phenomena, such as long distance case, polarity case, aspectual case, quantificational case and case concord island, among other phenomena. It is these unexplained facts, and not the defense

[3] To a first approximation, X c-commands Y if and only if Y is contained inside the sister of X. Intuitively, the notion designates a certain kind of structural dominance, where X is syntactically more prominent than Y. In Finnish a good heuristic is that, all else being equal, if X c-commands Y, then X precedes Y in the linear order.

of CPA, that motivate the present study. I will argue that all structural and semi-structural case assignment configurations in this language can be explained as a consequence of CPA, independent of whether they concern standard finite clauses such as (3a–b) or more peripheral and exotic constructions, such as long distance Case. If the argument holds, then these diverse facts can be shown to follow from one relatively simple and elegant principle of the UG. There is no need for construction-specific mechanisms. An extra benefit is that we can be sure that CPA is the right way to look at such matters.

The following is a rough blueprint to the structure of this article. Section 2 discusses the theoretical background and delineates CPA more carefully. Section 3 discusses various constructions in Finnish involving Case and Phi. We will not limit ourselves to finite and nonfinite clause structure but, instead, we will examine quantificational Case, negative constructions, accusative alteration, aspectual Case, adpositions, adjective constructions and noun phrases. Section 4 summarizes the data and Section 5 incorporates it all into a revised theory of Agree.

2. Theoretical and methodological background

This work has been done in the context of the minimalist program (MP) (Chomsky 1995). Before explaining and elaborating the research hypothesis itself, a brief introduction to MP is provided (see also Huhmarniemi 2012: Ch. 4). CPA will then be reformulated.

The minimalist program is a research *program*, a set of stated goals for linguistic analysis rather than an empirical hypothesis. One may either adopt or reject these goals, depending on personal preference and interest. The minimalist program can be adopted inside any linguistic framework. Its goal, then, is to explain the properties of human language(s) on the grounds of certain "given", extra-linguistic faculties, properties and processes so that the role of complex, autonomous syntax is reduced to an absolute minimum. The program itself does not claim that such unification is possible, and it does not begin from the axiomatic assumption that the program will succeed; rather, it is an attempt to see how far that goal can be reached. (Perhaps it cannot be reached.)

The minimalist program leads us to ask what we might consider as "given" in the context of linguistics. Perhaps we can agree that language establishes a relation between sound and meaning, and does so by utilizing a discrete, combinatorial system. We might also agree that there is a lexi-

con, a storage of primitive units. If so, we have established the four core components of MP: the basic structure-building operation, *Merge*; a repository of lexical information, *the lexicon*; and the two *interfaces* Logical Form (LF) and Phonetic Form (PF), which mediate between language and the two extra-linguistic systems, meaning-thought and perception-articulation, respectively. Merge builds structured linguistic representations from lexical items such that the representations have a "realization" at the interfaces.[4]

In addition to the four indispensable components, language has a fifth operation that is harder to justify on minimalist grounds: Agree. Agreement in the broadest sense of the term refers to feature covariation between linguistic elements. Another way to think about agreement is feature displacement: some linguistic feature or property is copied or displaced and occurs at another location. For instance, phi-features of a grammatical subject are redundantly copied or displaced to a finite verb in many languages. Furthermore, some grammatical heads impose Case features to the nominal elements they govern (3a–b).

The question of why such formal features, Case and phi in particular, exist, has been debated to some extent in the minimalist literature, without resolution. The general consensus seems to be that they have no minimalist explanation whatsoever and behave like "bogus" features that emerge from the lexicon but lack realization at the interfaces and must, therefore, be deleted. Here we arrive at the heart of the current minimalist technologies, partly inessential to the main goal of the present article but useful in illustrating the background of CPA. Consider again the derivation of (3a). When the tense T is merged, it possesses an uninterpretable phi-set probe, call the feature φ*. That feature is semantically uninterpretable, hence it

[4] Another possibility is to view language as a trivial externalization of thought. According to such an extreme minimalist view, there is no need to posit a linguistic operation such a Merge. The problem with this view is that it is possible to compose words and expressions without meaning: *married bachelor*, for instance. We assume that words can be combined even if no meaning or thought is implied or involved. Language is more creative than thought, and not bound in any way by what can be interpreted semantically or understood by the faculty of thinking. Yet another possibility is to emphasize communication instead of, or in addition to, the cognitive architecture. Communication is an extra-linguistic phenomenon in the sense that there is both communication without language (e.g. in bacteria) and language without communication (keeping a diary). Communication and the cognitive processes supporting it are "given" in the sense required by minimalism. Its role in the description and explanation of morphosyntax and syntax is negligible, however.

does not have a realization at the LF-interface, and must be deleted before the representation is passed on to the LF-interface. Call the uninterpretable feature the *probe*. It will establish Agree with the closest DP, the *goal*, the subject (*Greene*). The configuration is shown in (5), where the '*v* + V' cluster contains the verbal material.

(5) T [*Greene* *v* V *tämä kirja*]
φ* Case=NOM
probe goal

The derivation cannot continue unless something is done with φ*. Agree is, according to the minimalist theory of Chomsky (2000), the mechanism by which something gets done. Upon Agree(φ*, DP), the uninterpretable feature φ* is deleted and, during the same process, structural Case is assigned to the goal *Greene*. Structural Case is, therefore, viewed as a consequence of a minimalist "clean-up" operation. From this the "reflex" condition (CPA) follows. The nominative Case assigned to the subject *Greene* by Agree(φ*, DP) is associated with the c-commanding uninterpretable phi-set probe φ* that triggers the operation (5). In general, DPs obtaining structural Cases are paired with a functional head bearing an uninterpretable phi-set probe: the basic idea underlying the probe-goal system of Chomsky (2000, 2008).

If the uninterpretable features are deleted, why do we see phi-agreement and Case assignment? The additional assumption is that the lexicon provides phi- and Case-features without *value*. That is, a functional head might possess a feature such as Number* and a nominal element may posses a feature such as Case*. Unvalued features have realization neither at the LF-interface nor at the PF-interface. One cannot spell out Case* or Number*. What happens, then, is that such features are deleted from the path constructing a representation for the LF-interface, and they are provided morphological/phonological *values* for the path leading to the PF-interface. The latter operation is called *valuation*. For instance, valuation might lead to value assignments such as Number*=SG and Case*=NOM.

This is, in brief, how the current minimalist theory attempts to derive CPA as a theorem. The theory could be right or wrong, but the matter is inessential for the task at hand. I will attempt to show how CPA itself helps us to rationalize Finnish morphosyntax.

In opposition to much theorizing to the contrary, minimalism (and the generative theory in general) assumes that linguistics belongs to the natural

sciences. It assumes that language is (in part) a natural phenomenon, like vision, insect navigation or bee communication. It can therefore be fruitfully investigated by means of experimental hypothesis testing and rigorous, formal theory-development. There will be no *a priori* restriction on the type of evidence and method that can bear on some empirical question, whether it be acceptability judgment, corpus search, controlled experiment, brain imaging or a genetic study.

3. Case and Phi in Finnish

3.1 Finite verbs

Finite clause constitutes a core example of CPA. Let's consider (2b) and (3) once again, repeated here as (6a–b).

(6) *a.* *Tämän* *kirjan* *kirjoitti* *Graham* *Greene.*
this.ACC book.ACC wrote.3SG Graham.NOM Greene.NOM
'Graham Greene wrote this book.'

b. *Graham* *Greene* *kirjoitti* *tämän* *kirjan.*
Graham.NOM Greene.NOM wrote this.ACC book.ACC
'Graham Greene wrote this book.'

Finite verbs in Finnish are associated with the nominative DP, and they phi-agree with such nominative DPs. The finite element, finite verb in this example, assigns the nominative Case to the DP at Spec,*v*P.[5] That DP will then value its phi-features to the verb. This instantiates Agree, exemplified in (6a). The subject DP may raise to Spec,TP (6b). CPA is satisfied: before the EPP-movement the external argument bearing the nominative is c-commanded by a finite verb that contains an uninterpretable phi-set. The situation is essentially the same if the subject agrees with an auxiliary, modal or with the pre-sentential negation: all these elements c-command the subject before EPP-operations apply.

The probe-goal relation underlying Agree can target an element inside another clause. Raising constructions such as (7) exemplify this phenomenon in Finnish.

[5] Spec,*v*P is the postverbal position of the external argument that is associated with, or constitutes, the thematic role of Agent or Causer. The external argument may then be moved to Spec,TP, but not necessarily.

(7) *Professori näyttää _____ haluavan lopettaa kognitiotieteen.*
professor.NOM seem.T.3SG want.VA terminate cognitive.science
'Professor seems to want to terminate the cognitive science program.'

The uninterpretable phi-set at the raising verb *näyttää* 'seem' constitutes the probe, and its goal is the thematic agent of the complement clause *haluavan …* 'to want …' (there are several types of non-finite complements in Finnish, glossed as A, MA and VA in this paper; see Koskinen 1998; Vainikka 1989; Vilkuna 2005). Agree assigns the nominative to the subject and values the phi-features '3SG' to the raising verb. EPP-movement then brings the subject to its presentential subject position. The c-command condition holds, as can be seen by analyzing the derivation of (7) just before the EPP-movement applies (8).

(8) *näyttää* [*professori haluavan lopettaa kognitiotieteen.*]
seem.T.φ* Professor.NOM want.VA terminate cognitive.science
Agree(φ*, *professori*.nom)

3.2 Numerals

Finnish numerals exhibit mysterious behavior that can be sorted out elegantly with the help of CPA. Many numerals (other than *yksi* 'one') in Finnish assign the partitive to their complement, which contains the noun head and, optionally, an adjective:

(9) *Nuo kaksi pien-tä auto-a varastettiin.*
those.PL two.SG small-SG.PRT car-SG.PRT were.stolen
'Those two small cars were stolen.'

The numeral assigns the partitive irrespective of whether the host DP occurs in a subject or object position. The numeral is in singular and thus it cannot be combined with a *tantum plurale* noun such as *sakset* 'scissors' (10).

(10) **kaksi saks-i-a*
two.SG scissor-PL-PRT
'two scissors'

The numeral itself occurs in a case-less form. This is a marked form, since it is possible to inflect the numeral like an adjective, both in phi-features

and in case features. But the declining numeral does not assign the partitive, as shown in (11).

(11) *Nuo kahdet piene-t auto-t varastettiin.*
Those two-ACC.PL small-ACC.PL car-ACC.PL were.stolen
'Those two small cars were stolen.'

Following Rutkowski (2007), I will call the numeral which declines like adjectives but does not assign Case the "adjective numeral", or A-numeral, and the partitive assigning numeral as the Q-numeral, Q from "quantificational". Finnish therefore has two types of numerals in its lexicon, the Q-numeral and the A-numeral. There are two differences between the phi-features at the Q-numeral and at the A-numeral: whereas in the former the phi-features are uninterpretable, in the latter they are interpretable. For instance, the numeral *kaksi* 'two' in (9) denotes a plurality, but it is marked for singular (Brattico 2008, 2010). Such number marking is semantically uninterpretable. The second difference is that the Q-numeral assigns Case while the A-numeral doesn't. Hence the numeral that is marked (by lexical specification) by an uninterpretable number feature functions as a Case assigner, while the numeral which is not so marked does not function as a Case assigner. The existence of uninterpretable phi correlates with Case assignment, as predicted by CPA (Brattico 2011b) and shown in (12).

(12) *Nuo kaksi pien-tä auto-a varastettiin.*
those.PL two.φ* small-PRT car-PRT were.stolen
→ ↗ ↗
'Those two small cars were stolen.'

The hypothesis that Case assignment by the Q-numeral results from the uninterpretable phi-feature at the numeral makes the following prediction. If the grammatical feature [singular] is attached to a numeral representing singularity, such as *yksi* 'one', Case assignment should not materialize. This prediction is borne out, as shown in (13a–c).

(13) *a. Näin yhden auton.*
saw.1SG one.SG.ACC car.SG.ACC
'I saw one car.'

*b. *Näin yksi autoa*
saw.1SG one.SG.0 car.SG.PRT

c. **Näin* *yhden* *autoa*
saw.1SG one.SG.ACC car.SG.PRT

The numeral *yksi* has only the declining A-numeral form. The hypothesis predicts that a combination of *yksi* 'one' and [plural] should create a Q-numeral. At first it looks as if this prediction is not borne out. There is no way to make the plural form of *yksi* a Case assigner (14a–b); instead, the construction is grammatical if and only if the plural numeral and the noun head decline (14c).

(14) *a.* **yhde-t* *auto-a*
one-ACC.PL car-SG.PRT

b. **yhde-t* *auto-ja*
one-ACC.PL auto-PL.PRT

c. *yhde-t* *auto-t*
one-ACC.PL car-ACC.PL

However, here we have overlooked one important detail, namely the difference in interpretation between 'one car' marked as singular and 'one car' as plural. Reflecting on the meaning of (14c) makes it clear that *yhdet autot* 'one.PL car.PL' does not refer to just one car, but several, and that the plural marked *yhdet* 'one.PL' signals indefiniteness and not a cardinal number. Its syntactic behavior shows that the plural 'one' further belongs to the category of D and not Num, since it can occur together with another numeral (15).[6]

(15) *a.* *Me* *tavattiin* *yhdet* *kolme* *miestä.*
we met one.PL three.0 man.SG.PRT
'We met three indefinite men.'

b. *Yhdet* *kolme* *pistettä* *vaihdoin* *viivaksi.*
one.PL three.SG points changed.1SG into.line
'I changed three points into a line.'

c. *Taas* *yhdet* *kolme* *vuotta* *elämästä* *takana.*
again one.PL three.SG years from.life behind
'Again (one chunk of) three years of my life have elapsed.'

[6] Examples (15b–c) from the Internet.

Returning to the main point again, the combination of *yksi* 'one' and the feature [plural] does not produce a Q-numeral; instead, it produces a plural indefinite determiner. It does not function as Case assigner, as predicted by CPA, because the feature set [plural][indefinite] is interpretable.

In sum, then, several mysterious properties of Finnish numerals follow from CPA. It is useful at this point to state quite explicitly what has been established and what has not. What these facts suggest is that we should put the finite clause and the quantificational construction completely on par what comes to φ* and Case assignment. They exhibit the same regularity, and instantiate the same Case assignment/phi-Agree system, regardless of the fact that two entirely different constructions are at stake. In addition, the observation should increase our confidence towards CPA. The third important thing to notice is that the common regularity between the finite clause and the quantificational construction only comes visible once we have dissociated Agree and EPP and eliminated the idea that Case assignment in a finite clause is a function of a local Spec-Head relation. Case assignment must always be a "downstream relation". Finally, what was said here by no means explains the whole picture; lexical Case assignment, for instance, interacts with structural Case assignment in a manner I do not discuss here (see Brattico 2011b).

3.3 Sentential negation

A negated clause in Finnish is formed by adding a verbal negative particle *e-* to the beginning of the clause, above the TP according to most experts (Holmberg & Nikanne 1993, 2002). The negative word agrees in phi-features with the nominative subject DP (if there is one) that is typically located at its Spec, assigns partitive Case to the direct object(s) of the clause, and selects for a special form of the matrix verb *syönyt* (participle in the past, suffixless form in the present) which shows tense information (16a–b).

(16) *a.* *Pekka* *e-i* *syö-nyt* *leipä-ä* / **leivä-n.*
Pekka.NOM not-3SG eat-NUT.PAST bread-PRT / bread-ACC
'Pekka did not eat a/the bread.'

b. *Pekka* *söi* *leipää* / *leivän*
Pekka.NOM ate bread.PRT / bread.ACC
'Pekka ate some bread/ the bread.'

For those unfamiliar with the syntax of Finnish negation, it can be thought of as "tenseless auxiliary". That the partitive is assigned by the negation and not the by participle verb is shown by two facts. First, the participle verb by itself does not require its complement to be in the partitive Case. This can be shown by using the participle form in its adjectival function (17).

(17) *Leipä-ä / leivä-n syö-nyt koira nukkui.*
bread-PRT / bread-ACC eat-NUT.PAST dog slept
'A dog that ate some bread/the bread slept.'

Second, if the participle verb is complemented with a non-finite verb, or a sequence thereof, the object must still take the partitive Case in a negated sentence (Brattico in press, Ross 1967):

(18) *a. Uskoin Pekan haluavan olla syömässä leivä-n / leipä-ä.*
believed Pekka want.VA be.A eat.MA bread-ACC / bread-PRT
'I believed that Pekka wanted to be eating bread.'

b. En uskonut Pekan haluavan olla syömässä
not.1SG believed.SG Pekka want.VA be.A eat.MA
**leivä-n / leipä-ä.*
*bread-ACC/ bread-PRT
'I did not believe Pekka to want to be eating the bread.'

The relationship between the negation and the object persists despite the intervening nonfinite complements. Neither the main verb nor any of the embedded verbs have a vote on the matter. Therefore, it is a feature of Finnish grammar that the negation requires all direct objects in its path of government to be in the partitive.

CPA suggests that the negation data should be compared with the Q-numeral data. Specifically, the negation assigns partitive Case to an object (or several) it c-commands, but it also bears uninterpretable phi-features. Hence this data confirms CPA. The negation data further demonstrates that it would be wrong to confine Case assignment into a local domain. The negation is able to partitivize direct objects over a distance. This supports the theory of Agree, but it shows at the same time that the standard theory of Agree cannot be the full story. Finnish object Case assignment is not even clause bound, so we need to revisit the issue of locality.

The construction reveals an additional property of Agree that is not part of the standard theory: the negation does *not* phi-Agree with the object DP which it partitivizes; rather, the negation phi-Agrees with the subject, which it also probes into Spec,NegP. Stated in other words, the partitive is a "reflex" of a c-commanding phi-feature probe at the negation as stated in CPA, but there is no upward phi-Agree as assumed in the standard theory. The standard theory will be revised accordingly. The situation is illustrated in (19).

(19) *Pekka* *ei* *syönyt* *leipää*
Pekka.NOM not.3SG eat.NUT bread.PRT
NOM ↔ φ* → PRT

The present analysis can be objected on the grounds that also other negative polarity items, such as negative adverbs, have a similar partitivization effect. This suggests that partitivization is related to polarity, not to agreement. I concur with the hypothesis that polarity does affect Case assignment. On the other hand, these items do not necessitate the partitive, unlike the negation, as they only license it under a verb that does not normally do so (20a–b).

(20) *a.* *Voitan* **kilpailua* / *kilpailun*
win.1SG competition.PRT / competition.ACC(n)
'I will win the competition.'

b. *Tuskin* *voitan* *kilpailua* / *kilpailun*
hardly win.1SG competition.PRT / competition.ACC(n)
'I will hardly win the competition.'

Thus, it still remains that the uninterpretable phi-features at the negation contribute to partitive Case assignment: they make it obligatory in a manner required by structural Case assignment.

Another objection is based on the fact that the partitive-accusative alteration encodes aspect. In brief, the partitive direct object signals the fact that the event described by the sentence is perceived as being incomplete (Vainikka 1989). This opens up the possibility of explaining the partitivization by negation as resulting from the fact that a negative sentence must always be incomplete in its aspect: something that is negated was never completed. This hypothesis is wrong, however. First, because the partitivization by negation is not restricted by clause boundaries, it affects direct

objects in clauses whose aspectual interpretation is not affected by the presence or absence of the negation in *another clause*. Second, the grammatical tests which diagnose the presence of complete/incomplete aspect interpretation can register complete aspect in a sentence that has been negated and whose direct object still occurs in the partitive (Brattico in press).

Both the negation and the quantificational numeral bear uninterpretable phi-features, and they both assign the partitive to a goal they c-command. The negation can assign the partitive to several goals. A finite verb, in contrast, assigns the nominative. This calls for an explanation. The explanation is that the negation and the Q-numeral assign the object Case to an element from which they do *not* get any phi-features, while the finite verb assigns the nominative to an element from which it *does* obtain phi-features. In other words, the nature of the Case assigned depends on whether the goal will donate its phi-features to the probe (see (19)). We will return to this observation later on.

3.4 Accusative alteration

Possibly the most interesting piece of data related to the generalization CPA concerns the Finnish accusative.[7] There are several forms of the accusative in Finnish, two of which are important for present purposes: the n-accusative (ACC(n)) and the 0-accusative (ACC(0)). Vainikka & Brattico (submitted) argue that which of the two accusative suffixes is realized depends on whether the object DP is c-commanded by a predicate (or several) that possesses uninterpretable phi-features φ*. To see this principle at work, consider (21a–b). In (21a), the direct object is c-commanded by a finite verb which agrees with the subject, hence the n-accusative emerges; in (21b), the impersonal passive verb lacks agreement and the 0-accusative is assigned. Both sentences have a nominative subject and their meaning is the same, regardless of the passive morphology at the verb. They are thus

[7] The literature on the Finnish accusative is vast and cannot be reviewed here. Basically, if a direct object DP is not assigned partitive Case, it is assigned the accusative. The distribution of the two is controlled by several factors, still partly controversial (Kiparsky 1998, 2001; Vainikka & Brattico submitted). A consensus has emerged which says that the partitive is assigned if the action denoted by the verb phrase is interpreted as not being completed, while the accusative is assigned if it is completed. The accusative Case is therefore linked with the aspectual properties of the clause (Vainikka 1989).

both active sentences. The passive form is used in colloquial Finnish, while the active form is used in written Finnish.

(21) *a.* *Me* *löysimme* *auto-n.*
we.NOM found.1PL car-ACC(n)
'We found a car.'

b. *Me* *löydettiin* *auto.*
we.NOM found.PASS car.ACC(0)
'We found a car.'

Vainikka & Brattico (submitted) argue that this generalization covers all sentence/phrase types. I will not repeat their arguments here. Why should c-commanding phi-features affect the form of the direct object Case? Although at first a mystery, the answer is provided by CPA: Case is a reflex of c-commanding uninterpretable phi φ*.

The accusative alteration further supports the contention that we should dissociate phi-Agree from Case assignment. The phi-set that regulates the form of the accusative does not come from the direct object bearing the accusative. There is never an agreement between the n-accusative or 0-accusative and the verb. The phi-set that determines the form of the direct object in examples such as (21) emerges from another DP, usually the grammatical subject (21'). A further reason to follow this hypothesis is that the accusative Case assignment, like the partitive of negation, is a long distance phenomenon (Vainikka & Brattico submitted; Brattico in press). There is a remote Case assignment between a probe and its goal such that the two *never* phi-Agree. Brattico (in press) claims that direct object Case assignment and only direct object Case assignment in Finnish is a long distance phenomenon.

(21') *a.* *Me* *löysimme* *auto-n.*
we.NOM found.1PL car-ACC(n)
NOM ↔ φ* → ACC(n)

b. *Me* *löydettiin* *auto.*
we.NOM found.PASS car.ACC(0)
NOM PASS ACC(0)

Example (21, 21') shows that the nominative Case of the subject must be available without the presence of uninterpretable phi-features at the finite

element. Vainikka & Brattico (submitted) follow the standard minimalist theory and assume that it originates from the C (complementizer), which is often assumed to contain φ* as well.

The accusative rule deviates from the examples investigated so far in the sense that here the bearer of the uninterpretable phi-set assigns the n-accusative and the 0-accusative, not the partitive or the nominative. The problem here is, however, that it has never been clear what the "n-accusative" and the "0-accusative" really are. The difficulty stems from the fact that the n-accusative is homonymous with the genitive, while the 0-accusative is homonymous with the nominative. It is not irrelevant whether we think these features as the true subject Cases, genitive and nominative, respectively, or whether they are regarded as merely homonyms. I leave this technicality for a further study.

3.5 Adpositions

Many adpositions (PPs here) in Finnish assign the partitive to their complement DP (compare *towards him/*he*). For some adpositions there exists a near-synonymous expression with a genitive DP at Spec,PP that also exhibits phi-agreement with the preposition. Compare (22a–b).

(22) *a. Hän nukkui lähellä minu-a.*
He slept near I-PRT
'He slept near me.'

b. Hän nukkui (minu-n) lähellä-ni.
He slept I-GEN near-PX/1SG
'He slept near me.'

Agreement in number and person is represented by the possessive suffix in (22b). A non-agreeing DP in an object Case occurs after the preposition, while a DP in the subject Case appears before the preposition, presumably at Spec,PP, and triggers agreement.[8] Example (22a) reveals that a head that assigns the partitive does not need to show overt phi-features, contrary to the the negation and the Q-numeral, both which possess an overt phi-set.

[8] Only genitive human pronouns trigger overt agreement in this construction, never full DPs. But regardless, Case and phi are systematically related. My own thinking is that the possessive agreement suffix encodes person, number and a human feature (compare *hänen autonsa* 'his car.PX/3SG', **sen autonsa* 'its car.PX/3SG')

The same phenomenon can be observed in connection with transitive verbs, as shown in the next section. We will return to the explanation of this observation in the last section. In addition, to maintain the hypothesis that Case assignment is a reflex of *c-commanding* phi-set, as stated in CPA, we must assume that the genitive DP reaches the Spec,PP position via movement and is assigned its Case prior movement (Manninen 2003; Brattico 2011a). This analysis is illustrated in (23a–b).

(23) *a.* [$_{PP}$ _____ *lähellä* [$_{DP}$ *minä*]] (first Merge positions, P + DP)
near I

b. [$_{PP}$ *minun* *lähellä-ni* [$_{DP}$ ____]] (Agree + EPP, Spec + P)
I.GEN near-PX/1SG
'near me'

3.6 Prenominal genitives

In this section we will look at prenominal genitive Case assignment. We begin with the basic properties of Finnish deverbal nominals, as illustrated in (24).

(24) *ne* *kaksi* *holtitonta* *isä-n* *auto-n* *ostami-sta*
those two reckless-PRT father-GEN car-GEN buying-PRT
'those two reckless buyings of the car by the father'

The noun head *ostamista* 'buying' is made of a verbal root *osta-* 'buy' and a nominal suffix *-minen* '-ing' in the partitive (*-minen* > *-mista*). In the sentential context, the verbal root *osta-* is used as a transitive verb; in the nominalized context (20) it is used as a transitive predicate in the sense that the DP contains both the agent (father) and the patient (car). In agreement with previous literature (Brattico 2008; Brattico & Leinonen 2009), I assume that the nominalizer *-minen* selects for a VP (*v*P) which contains the verbal root together with the thematic arguments. The verbal root is combined with the nominalizer *n*, by head movement.

This assumption is supported by several facts, worth keeping in mind. First, the *-minen* nominalization is fully productive. This is expected if the process takes place in syntax. Second, as shown in (25a–b), the deverbal nominalizer can be modified by an adverb. If we assume that adverbs are adjoined to VPs, then this fact poses no mystery.

(25) *a. Uude-n auto-n osta-minen liian nopea-sti on virhe.*
New-GEN car-GEN buying-NOM too fast-ADV is mistake
'The buying of a new car too fast is a mistake / It is a mistake to buy a car too fast.'

b. ...lukeminen liian hitaasti voi tarkoittaa, että ymmärtäminen vaikeutuu.
reading too slowly can mean that understanding becomes.more.difficult
'Reading too slowly can mean that understanding becomes more difficult.'

Third, whatever thematic roles are assigned in the corresponding clausal context are mirrored in the nominal context. This parallelism is illustrated in (26a–b).

(26) *a. Isä osta-a auton.*
father.agent buy-3SG car.patient.
'The father buys a car.'

b. isän auton ostaminen
father.agent car.patient buying.NOM
'the buying of a car by the father'

This pattern becomes understandable if both expressions are built from the same underlying VP, distributing the same thematic roles to the argument DPs.

Vainikka (1989) argues that Finnish genitive Case is assigned to Spec,NP (or Spec,*n*P under the current system, see Huhmarniemi 2012; Brattico & Leinonen 2009). The genitive DPs, unlike PP and CP arguments, always occur in a prenominal position. There are three problems in Vainikka's hypothesis. The first puzzle is that there can be from zero to two genitive DPs in the prenominal position, suggesting that one specifier position is not enough.[9] The second puzzle concerns case concord. Finnish shows extensive case concord, as shown in (27a–b).

[9] In a recent study, Huhmarniemi (2012) assumes that *n* selects for an NP. This leaves room for two nominal specifiers, Spec,NP and Spec,*n*P. If the DP is projected originally from a category-neutral root, as I assume here, then Huhmarniemi's hypothesis means that there are two nominalizers *n*-*n**, perhaps analogous to the verbal *v*-*v** system. This hypothesis provides the two specifier positions required in Vainikka's analysis.

(27) *a. Minä katsoin yh-tä pien-tä punais-ta talo-a.*
I looked at [one-PRT small-PRT red-PRT house-PRT]:PRT
'I watched that one small read house.'

b. Minä näin yhde-n piene-n punaise-n talo-n.
I saw [one-ACC small-ACC red-ACC house-ACC]:ACC
'I saw one small red house.'

But genitive DPs constitute an exception to this pattern: they need not show the Case feature assigned to the DP as whole. We say that they constitute a "case concord island" within the DP:

(28) *Todistin sitä kauheaa* ***sikojen*** *teurastamista.*
Witnessed.1SG [that.PRT horrible.PRT pig.PL.GEN butchering.PRT]:PRT
'I witnessed that horrible butchering of the pigs.'

Why genitive DPs are spared from Case concord even if they appear to c-command the nominal head? Vainikka's model can describe these facts by stipulation, but it does not explain them.

The third problem is as follows. We have seen evidence that nominalization takes place in syntax by means of head raising (verbal root => *n*). The thematic arguments of the deverbal nominal, on the other hand, are generated inside the VP. But these operations predict that the genitive DPs should occur in the postnominal position. The two genitive DPs must, therefore, find their way to Spec,*n*P before they are pronounced or spelled out. But if they obtain the genitive Case at that position, then the hypothetical Spec-head Case assignment rule must assign the genitive Case *after* A-movement. This assumption violates the theory of Agree, which says that the EPP-movement follows Case assignment and phi-agreement.

These problems are related to each other. They are all solved if we follow the standard theory of Agree and assume that the genitive Case is assigned *before* movement. If the genitive DPs are based-generated inside the VP, there is no reason why they should be assigned matrix Case. This is because case concord does not reach postnominal elements and will therefore not percolate to the VP. Case concord island phenomenon is explained, too. Suppose that the genitive arguments are first assigned genitive Case *in situ* inside the VP by the nominalizer *n* and are then moved to the prenominal Spec,*n*P position. A Case concord island will be generated. These operations also derive the correct word order facts: genitive DPs appear at the

Spec,*n*P while other constituents, such as PPs and CPs, can appear in the postnominal position.

More evidence for the assumption that prenominal genitive DPs are moved rather than base-generated to their prehead position can be derived from Finnish nominal raising constructions. The relevant construction is shown in (29).

(29) *a. Vauvan näyttäminen nukkuv-a-lta oli helpotus.*
Baby.GEN semblance sleep-VA.SG-ELA was relief
'Baby's looking like he would be sleeping was a relief.'

b. Vauvojen näyttäminen nukku-vi-lta oli helpotus.
Babies.GEN semblance sleep-VA.PL-ELA was relief
'Babies' looking like they would be sleeping was a relief.'

Vauva 'baby' is not the semantic subject of *näyttäminen* 'semblance', but rather the thematic subject of *nukkuvalta* 'like sleeping.SG'. As can be seen by comparing (29a–b), the embedded predicate *nukkuvilta* 'like sleeping.PL' agrees in number with the raised subject *vauva*/*vauvojen* 'baby/babies' (incomplete agreement is associated with a trace/copy left by movement). As expected, it is possible to construct a sentence in which the prenominal genitive represents the embedded object, as in the following:

(30) *Pellon näyttäminen rottien tuhoamalta säikäytti minut*
field semblance rats.GEN destroyed frightened me
'The field's looking like it had been destroyed by rats frightened me.'

The prehead genitive DP *pellon* 'field.GEN' is the thematic object of *tuhoamalta* 'get destroyed'. This evidence shows that the prenominal position, like the subject position of a finite clause, is not a thematic position, and is presumably filled by EPP-movement.

Two further pieces of evidence for the assumption that the genitive DP comes to Spec,*n*P via movement is provided by agreement facts. First, the prenominal genitive pronoun DPs agree in phi-features with the nominal head.

(31) *Minun auto-ni hajosi.*
My car-PX/1SG broke
'My car broke down.'

The possessive suffix agreement does not encode possession or other semantic attribute, but reflects the pronoun that is raised to Spec,*n*P irrespective of its semantic role:

(32) *minun esittely-ni*
my introduction-PX/1SG
'the introduction of me to somebody else' or
'the introduction of somebody by me.'

Second, we have seen that the genitive pronoun agrees with the noun in full phi-features. The genitive DP also agrees with the embedded predicate, but this agreement is incomplete. Incomplete agreement often takes place between traces/copies left behind by movement. For instance, in a Finnish negative clause the negation shows full phi-features but the tensed verb shows incomplete number agreement (in the past). There is evidence that the subject DP has been moved through the Spec,TP position before landing into Spec,NegP (Brattico & Huhmarniemi 2006). Analogously, the incomplete phi-agreement shown on the predicate in a nominal raising construction might signal the presence of movement.

Consider CPA, which says that Case is a reflex of c-commanding uninterpretable phi. The prenominal genitive constructions have exactly these properties. The functional head *n* possesses uninterpretable phi-features and, therefore, counts as a Case-probe. The nominalizer is merged with a VP that contains the (two) argument DP(s), so *n* will c-command these DPs. Therefore, it will assign Case to them. Finally, *n* has the EPP-feature which forces both DPs to move to Spec, *n*P (see Brattico & Leinonen 2009)(33a–b).

(33) *a.* [$_{nP}$ _____ *-y* [$_{VP}$ *esittele- minä*]
n,φ* present- I

b. [$_{nP}$ *minun*$_i$ *esittel*$_j$*-y-ni* [$_{VP}$ __$_j$ __$_i$]
I.GEN present-*n*-PX/1SG
'introduction of me/introduction by me'

3.7 Preverbal genitive DPs

The genitive is also associated with certain verbs. Finnish VA-infinitivals, for one, take prehead genitive DP arguments:

(34) *Pekka näki Merja-n lähte-vän.*
Pekka.NOM saw Merja-GEN leave-VA.present
'Pekka saw Merja leaving.'

Because the VA-infinitival shows tense, it contains at least the TP (tense phrase). The VA-infinitival agrees in phi-features with the embedded subject, provided that the subject is phonologically empty and coreferential with the matrix subject, hence assumed to be represented by PRO (an empty pronominal):

(35) *Minna uskoi PRO lähte-vä-nsä.*
Minna believed leave-VA-PX/3SG
'Minna believed that she would leave.'

This construction can be unified seamlessly with the prenominal genitive data by assuming that the genitive DP has been moved to the Spec,TP, exactly as the prenominal genitive is moved to Spec,*n*P, both triggering pronominal agreement. There is direct evidence for this proposition. It is possible to craft a VA-raising construction and then have an embedded thematic object DP to occupy the same prehead position:

(36) *Minä uskon auton näyttävän hienolta.*
I believe car.GEN seem.VA fine.SG.ELA
'I believe that the car will look fine.'

I therefore assume that the genitive Case is assigned by T, after which the genitive DP is moved to Spec,TP. This analysis agrees with CPA: the genitive Case is assigned by a c-commanding functional head bearing an uninterpretable phi-set (37a), just prior EPP-movement (37b).

(37) *a.* _____ T *näyttävän* [*auton hienolta*]
φ* seem.VA car.GEN fine.SG.ELA
Agree(φ*, *auton*)

b. *auton* T *näyttävän* [_____ *hienolta*]
car.GEN φ* seem.VA fine.SG.ELA
EPP(*auton*)

3.8 Preadjectival genitives

A participle adjective (glossed as MA) crafted from a transitive root may take either the patient or the agent argument within its own projection (but not both). When the agent argument is included, it appears before the adjective and agrees in full phi-features with the adjective:

(38) *Meidän nopeasti löytä-mä-mme pallo oli rikki.*
we.GEN fast find-MA-PX/1PL ball was broken
'The ball found fast by us was broken.'

It is not possible to show that the genitive DP has been moved to the prehead position, because the participle adjective does not have a raising form. On the other hand, as seen in (38), these adjective phrases are compatible with verbal adverbs. This suggests that they are composed out of underlying VP, which hosts the thematic argument before it is raised to Spec,*a*P (*a* = adjectivizer, see Brattico 2005). Is there any reason to believe that the adjective contains the TP? One obvious reason derives from the observation that the participle adjectives in Finnish do show overt tense alteration. If we indeed assume that the adjective phrase is composed out of the T + *a* + *v** structure, namely, because they allow eventive adverb modification, assign object Case, contain an external argument and show tense, then this construction can be unified with other constructions with prehead genitive DPs arguments. The adjectivizer *a* or T (or both) first assign the genitive to the argument it c-commands and phi-agrees with it, then the genitive DP is raised to a left peripheral "subject" position at Spec,TP or Spec,*a*P. Principle CPA applies. The genitive Case is assigned either by T or *a* to the base position of the argument DP, after which the DP raises to Spec,TP or Spec,*a*P. The Case-probe contains an uninterpretable phi-set, as shown by the agreement facts.

3.9 Transitivity and object Case

Transitivity in Finnish is associated with the partitive and the accusative. Transitive verbs assign either Case depending on the aspectual properties of the verb (or the whole VP). This is true of finite and non-finite verbs alike. Deverbal participle adjectives, too, are capable of assigning the object cases and they come with a similar shift in aspectual interpretation (39a–b).

(39) *a. Leivä-n syö-nyt koira löytyi.*
bread-ACC eat-NUT.PAST dog was.found
'The dog that ate the whole bread was found.'

b. Leipä-ä syö-nyt koira löytyi.
bread-PRT eat-NUT.PAST dog was.found
'The dog that ate some of the bread was found.'

I assume that the object Case is assigned by the transitivizer *v**. In this case, it is more difficult to assess directly how well these facts fall under CPA: Finnish doesn't show object agreement. Therefore it is hard to tell whether *v** possesses phi-features. Either *v** constitutes an exception, or an independent factor prevents overt object agreement in Finnish. These theoretical options are discussed in the last section.

3.10 Summary of the data

Several grammatical contexts have been examined which involve Case assignment, phi-Agree and movement. The results are summarized in Table 1.

Table 1. Case assignment, phi-agreement and EPP

Probe	Assigns Case	Phi-probe	phi-Agree	EPP
Finite verb	NOM	yes	yes	yes
Preposition (type 1)	GEN	yes	yes	yes
Noun head	GEN	yes	yes	yes
VA-infinitival	GEN	yes	yes	yes
MA-adjective	GEN	yes	yes	yes?[a]
Q-numeral	PRT	yes	no[b]	no
Negation	NOM/PRT	yes	yes/no	yes[c]
Transitivity (*v**)	PRT(ACC)	no	no	no
Preposition (type 2)	PRT	no	no	no
Quantifier	no	no	no	no
Demonstrative	no	no	no	no
Numeral yksi 'one'	no	no	no	no
A-numeral	no	no	no	no
Adjective modifier	no	no	no	no
Complementizer	no	no	no	no
Conjunction	no	no	no	no
Clitic (-kO,-hAn,-pA)	no	no	no	no

[a] Property not possible to show, because the diagnostic raising construction is illicit.
[b] The numeral is fixed lexically to [singular], agreeing with the noun head.
[c] The negation assigns NOM to the subject and PRT to the direct object. It phi-Agrees with the subject.

The left column lists the Case-probes discussed so far together with a control group containing functional heads for which it is not possible to show that they assign Case. The second column from the left lists the Case features assigned by that particular probe. The middle column labeled as 'phi-probe' is composed so that 'yes' appears if the probe can have phi-features, and the next column tells whether the goal must also agree in its phi-features with the probe. The rightmost column tells whether the probe also attracts the goal to its Spec.

Prepositions occur twice in Table 1, because many of them have two forms: one assigning the partitive and the other the genitive.

4. A theory of Agree

The data shows a near-perfect correlation between Case assignment and the presence of uninterpretable phi-features (second and third rows from the left). In addition, the grammatical head bearing the uninterpretable phi-set c-commands the Case assignee. These properties follow from CPA. It would seem that CPA describes structural Case assignment in Finnish in a cross-categorical or construction-independent fashion. When a functional head comes to possess uninterpretable phi-features, for whatever reason, Case feature always emerges downstream, independent of the grammatical context or construction.

We cannot claim that Case assignment goes in tandem with phi-valuation (assignment of a phi-value). There are too many constructions which violate this condition: the negation, adpositions and transitive verbs being the clearest counterexamples. The negation data is particularly telling, since the partitivization crosses clause boundaries and thus Case assignment obtains between remote elements that can *never* phi-Agree, under any circumstances. On the other hand, Table 1 reveals that the bidirectional feature flow holds quite generally for the subject Cases, the genitive and the nominative. These are regularly associated with overt phi-Agree. To capture this correlation, I will correlate phi-Agree with the subject Case assignment ("subject Case assignment" covering both the nominative and genitive assignment) and not with object Case assignment. We will still sharpen this idea later on, but the underlying idea must be correct.

According to the standard theory, phi-agreement and Case assignment establish a bidirectional relation even if there are no overt phi-features at the probe. But since there aren't always overt phi-features at the probe,

something more has to be said. The usual strategy is to claim that the transitivizer, say, bears abstract, or *covert*, phi-features. Finnish data leads me to question this analysis: there are too many constructions which require the presence of abstract phi-agreement, making the notion suspiciously empty of empirical content. In addition, long distance Case assignment would require abstract phi-agreement over a distance and between elements which can never agree.

It is tempting to correlate object Case assignment with the presence of c-commanding uninterpretable phi-features, as stated by CPA. There is much data which confirms this generalization: the Q-numerals, accusative alteration, and the partitive of negation striking me as particularly interesting. But there are counterexamples, such as the adpositions and transitive verbs. They assign the object Cases but do not bear overt uninterpretable phi-features. If we look closer, it is evident that the criterial property for the object Case assignment is not the *existence* of c-commanding uninterpretable phi-features φ* but the fact that phi-Agree does not and cannot take place. The reason why a c-commanding uninterpretable phi-feature set at some grammatical head is linked with the object Case assignment must be because *those heads cannot phi-Agree the second time*. If a functional head is valued some phi-features, or the phi-features are lexically specified, that functional head must reject further phi-Agree. This condition subsumes also the transitive verbs and adpositions. There is no object agreement in Finnish, so transitive verbs, or rather *v*, must reject phi-Agree with the object DP. The same generalization applies to those adpositions which assign the object Cases: there is never phi-Agree. I propose that object Case assignment results from "rejected phi-Agree".

These generalizations can be unified in the following way. Suppose P is a functional head, a probe, and G is a nominal element, the goal. Then (40) must be part of the theory of Agree:

(40) *Condition of Agree*
a. The goal G is assigned an object Case (partitive, accusative) only if it does not and cannot value its phi-features to P (its phi-features are *rejected*)
b. The goal is assigned a subject Case (genitive, nominative) only if it does value its phi-features to P (its phi-features are *accepted*).

Generalization (40) puts the quantificational Case construction on par with transitive verbs, many adpositions and the negation in that all of these elements *reject* the phi-features from the goal they assign the object Case.

They reject the phi-features for different reasons, however. Transitive verbs and many adpositions reject the phi-features because they cannot inflect for the phi-features in the first place (in any context in this language); the negation rejects them because it had agreed already with another DP, the subject; and the Q-numeral rejects the phi-features because it is lexically marked for the phi-feature [singular].[10]

Generalization (40) makes a distinction between direct object Cases and subject Cases, but there are two varieties of both: the partitive and the accusative are both direct object Cases while the nominative and the genitive are subject Cases. While the partitive and the genitive are the cross-categorically default object and subject Cases, respectively, the nominative and accusative are "special". The nominative is linked with agreement in the finite clause (excluding the possibility that the zero-accusative is actually the nominative), and the accusative Case is associated with a special aspectual interpretation and therefore only occurs in the direct object position of verbal or deverbal predicates. The accusative Case is thus triggered by an aspectual feature residing in the *v*P, and the nominative is associated with the finiteness fin^0-head residing in the CP-layer. Because fin^0 phi-Agrees, a subject Case is assigned; because *v* never phi-Agrees, it assigns the object Case. These properties follow from (40).

According to (40), Case assignment is linked with phi-valuation. Specifically, Case assignment reflects probe's desire or refusal to get phi-features. Let's call this property its *phi-specification*. Functional head's phi-specification is a lexical property which dictates how the element will interact with phi-features nearby. CPA may then be reformulated:

(41) *Condition of Phi-Agree (final version) (CPA*)*
Structural Case is a reflex of c-commanding phi-specification.

Intuitively, structural Case assignment is a reflex of whether a c-commanding functional head will accept or reject its phi-features.

What other properties must hold for Agree? The probe must c-command the goal. The probe can only "probe" into its derivational past, not into its future. Another obvious property is that the goal must be a potential phi-donator. I capture this property by borrowing Chomsky's *feature*

[10] Notice that this generalization only works if we assume that the two variants of the accusative, 0-accusative and n-accusative, are direct object cases and on par with the partitive and the true t-accusative Case. An alternative analysis is to regard them as the subject Cases nominative and genitive.

match condition (Chomsky 2000: 122–123). Probe's phi-specification must match with interpretable phi-features at the goal. This explains why Case is assigned to nominal element: they are the potential phi-donors. Finally, for most situations we must assume that a given DP is assigned its Case only once. In other words, for each probe-goal pair, the decision concerning whether the goal will donate its phi-features to the probe will be made only once; after the decision is made, it is never negotiated again.

As stated in the standard theory of Agree, EPP must be factored out from the mechanism that regulates Case assignment and phi-valuation. It is often true that the EPP follows subject Case assignment and phi-Agree, but not necessarily. Similar, it is often true that EPP-movement does not follow object Case assignment, but this condition, too, is not absolute. The reversed construction discussed by Holmberg & Nikanne (2002) shows this: the agreeing subject remains *in situ* while the direct object satisfies the EPP-condition of the finite verb. In addition, there are constructions where the relation between EPP and Agree is strict. For instance, the genitive DPs inside a noun phrase can never be in the post-nominal position: they must always EPP-move. I have argued elsewhere that the EPP-diacritic regulating such facts is irreducible and must be assumed to be part of the lexical specification (Brattico 2011a). Why it correlates with phi-specification remains to be explained; one possibility is that movement of a whole DP is functionally equivalent to the transfer of its phi-features. This explains why heads which do not agree overtly with a DP in the Spec nevertheless can assign the subject Case (Brattico submitted).

In a recent paper, Vainikka (2011) presents strong arguments towards the conclusion that in Finnish, the genitive can be assigned also into an upward direction, from a head to its Spec. This analysis returns partially to the earlier theory which implements Case assignment under a local Spec-Head relation. She proposes that there are two Case assignment mechanisms, the downstream pointing relation, described here, and another pointing into an upward direction. CPA* rules out the upward directed Case assignment mechanism. Vainikka's main argument derives from examples such as *valtavan paljon ihmisiä* 'enormous.GEN many people', where there seems to be no c-commanding functional head assigning the genitive to the adverbial modifier. An obvious theoretical alternative is to regard the adverb as consisting of a lexical root and a functional head ('adverbializer'), and assume that the modifier is merged between the root and the functional head and assigned its Case by the functional head, exactly as is the case with the verbal, adpositional, nominal and adjective constructions. But the

case is hard to establish convincingly. I will leave the matter for future research, merely noting here where the disagreement lies and how Vainikka's argument can be used as a powerful argument against the present theory.

Agree, and agreement in general, presents a mystery for minimalism. There appears to be nothing that is "given" in the context of linguistics that could explain these properties. The present hypothesis provides the following perspective on this matter. Each functional head makes a stance, its phi-specification in essence, concerning what type of phi-features it will require within its own projection: interpretable in their own terms (e.g. plural marking on nouns) or uninterpretable (e.g. plural marking on a finite verb), and if the latter, whether they are lexical (e.g. singular on the numerals) or borrowed from another location (e.g. singlar on a finite verb). It they are borrowed, it will also require either the transfer of phi-features (Agree), the transfer of a whole phrase carrying those features (EPP), or both. If we assume that language(s) incorporate a hierarchical phi-specification system of this kind, structural Case appears to encode, in a one-to-one fashion, such specification. I propose that phi-specification is a lexically determined feature of functional heads, and structural Case is its phonological exponent. If so, at least one half of the mystery, the origin and function of structural Case, is solved.

5. Conclusions

A number of constructions were examined in Finnish which involve Case and Phi. It was argued that the theory of Agree provides a useful perspective to examine these constructions, allowing us to explain their otherwise mysterious properties. However, some properties do not fall under the standard theory of Agree and it was revised accordingly. More specifically, it was proposed that structural Case is a reflex of uninterpretable *phi-specification feature* of a c-commanding functional head. A phi-specification in turn determines whether a functional head will require phi-features within its own projection. If it does, it may request them by phi-Agree (transfer of phi-features alone), by EPP-movement (transfer of the host together with the phi-features), or by both. If if doesn't, then phi-Agree does not apply and EPP becomes optional.

References

Adger, David & Harbou, Daniel (2008) Why Phi? In Daniel Habour; David Adger & Susan Béjar (eds.) *Phi Theory: Phi-Features across Modules and Interfaces*, pp.1–34. Oxford University Press: Oxford.

Brattico, Pauli (2005) A category-free model of Finnish derivational morposyntax. *SKY Journal of Linguistics* 18: 7–46.

—— (2008) Kayne's model of Case and Finnish nominal phrases. *Nordic Journal of Linguistics* 31 (2): 135–160.

—— (2010) The two-part models and one-part models of nominal Case: Evidence from Case distribution. *Journal of Linguistics* 45: 47–81.

—— (2011a) The diehard extended projection principle. *SKY Journal of Linguistics* 24: 181–188.

—— (2011b) Case assignment, case concord, and the quantificational case construction. *Lingua* 121: 1042–1066.

—— (submitted) *Feature Displacement: A Study of Finnish Case*. Manuscript.

—— (in press) Long distance case assignment in Finnish. *Studia Linguistica*.

Brattico, Pauli & Huhmarniemi, Saara (2006) Finnish negation, EPP and the valuation theory of morphosyntax. *Nordic Journal of Linguistics* 29 (1), 5–44.

Brattico, Pauli & Leinonen, Alina (2009) Case distribution and nominalization: Evidence from Finnish. *Syntax* 12 (1), 1–31.

Chomsky, Noam (1993) A minimalist program for linguistic theory. In Kenneth Hale & Samuel K. Keyser (eds.), *View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, pp. 47–72. Cambridge, MA: MIT Press.

—— (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

—— (2000) Minimalist inquiries: The framework. In Roger Martin; David Michaels & Juan Uriagereka (eds.), *Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik*, pp. 89–156. Cambridge, MA: MIT Press.

—— (2001) Derivation by phase. In Michael Kenstowicz (ed.), *Ken Hale: A Life in Language*, pp. 1–52. Cambridge, MA: MIT Press.

—— (2008) On phases. In Robert Freidin; Carlos P. Otero & Maria Luisa Zubizarreta (eds.), *Foundational Issues in Linguistic Theory: Essays in Honor of Jean-Roger Vergnaud*, pp. 133–166. Cambridge, MA: MIT Press.

Holmberg, Anders & Nikanne, Urpo (1993) Introduction. In Anders Holmberg & Urpo Nikanne (eds.), *Case and Other Functional Categories in Finnish Syntax*, pp. 1–22. New York: Mouton de Gruyter.

—— (2002) Expletives, subjects and topics in Finnish. In Peter Svenonius (ed.), *Subjects, Expletives, and the EPP*, pp. 71–106. Oxford: Oxford University Press.

Huhmarniemi, Saara (2012) *Finnish A'-Movement: Edges and Islands.* Studies in Cognitive Science 2. Helsinki: Institute of Behavioral Sciences, University of Helsinki.

Kiparsky, Paul (1998) Partitive case and aspect. In Miriam Butt & Wilhelm Geuder (eds.), *The Projection of Arguments: Lexical and Compositional Factors*, pp. 265–307. Stanford: CSLI Publications.

—— (2001) Structural case in Finnish. *Lingua* 111: 315–376.

Koskinen, Päivi (1998) *Features and Categories: Non-finite Constructions in Finnish*. Unpublished doctoral dissertation, University of Toronto.

Manninen, Satu (2003) Finnish PPs and the phase impenetrability condition. In Diane Nelson & Satu Manninen (eds.), *Generative Approaches to Finnic and Saami Linguistics*, pp. 295–320. Stanford: Center for the Study of Language and Information (CSLI), Stanford University.

Ross, John Robert (1967) *Constraints on Variables in Syntax*. Unpublished doctoral dissertation, MIT.

Rutkowski, Pavel (2007) Grammaticalization in the nominal domain: The case of Polish cardinals. In Blake H. Rodgers (ed.), *LSO Working Papers in Linguistics: Proceedings of WIGL 2006*, pp. 89–102. Madison: Department of Linguistics, University of Wisconsin–Madison.

Vainikka, Anne (1989) *Deriving Syntactic Representations in Finnish*. Unpublished doctoral dissertation, University of Massachusetts Amherst.

—— (2011) Genitive Case in Finnish reconsidered. *Biolinguistica Fennica Working Papers* 2: 1–32

Vainikka, Anne & Brattico, Pauli (submitted) The Finnish accusative: Long distance Case assignment by phi-agreement.

Vilkuna, Maria (2005) *Suomen lauseopin perusteet* [Fundamentals of Finnish Clause Structure]. Helsinki: SKS.

Contact information:

Pauli Brattico
Institute of Behavioral Sciences
University of Helsinki
Finland
e-mail: Pauli(dot)Brattico(at)helsinki(dot)fi

Brian Fell

Applicatives and Incorporation in Ubykh

Abstract

This article seeks to provide a description of verbal agreement and its relationship with applicatives and other incorporated elements in the Ubykh verb as well as an analysis of these structures grounded in a modern syntactic framework. Ubykh is a Northwest Caucasian language and displays many characteristics of a polysynthetic language, including polypersonal agreement and the incorporation of either nominal roots or postpositions into the verb complex. These latter constructions are analyzed as the incorporation of a complex head – consisting of the root or postposition merged with an agreement head generated in a functional projection dominating the structure out of which the incorporated element moves. Applicatives are treated as a subset of postposition incorporation, with the postpositional phrase generated in the specifier position of the applicative projection.

1. Introduction

The structure and function of applicative constructions have been the subject of a great deal of research in the past several years. Applicatives are responsible for increasing the valence of a verb by licensing an additional argument. This is commonly a benefactive or goal, but may also be an instrument or locative.

Applicatives are variously marked by specific verbal morphology and/or adpositional constructions and double object constructions. Polysynthetic languages in particular use verbal morphology which originates either as an inflection (i.e., the head of a functional syntactic projection) or an incorporated phrase. The Yimas language, of the Lower Sepik family, provides an example of the former while the Iroquoian language Mohawk exemplifies the latter.

SKY Journal of Linguistics 25 (2012), 61–92

Yimas

(1) *awt ŋa-kra-yawra-mpi-waraca-ŋa-n.*
fire IMP.1-1PLD-pick.up-SEQ-return-BEN-IMP
'Bring fire back for us.' (Foley 1991: 309)

Mohawk

(2) *wa-hake-natar-a-kwétar-ʌ-'.*
FACT-3SGS.M>1SGO-bread-LNK-cut-BEN-PUNC
'He cut the bread for me.' (Baker 1996: 427)

Many of the polysynthetic languages which have been studied for their applicatives follow the above examples, in that the applicative is signaled by a morpheme suffixed to the verb stem.

Forms such as the above are analyzed by Baker (1996) with the applicative (*-ŋa-* and *-ʌ-*, respectively) heading its own VP which selects for three theta-roles, an agent, theme and goal. The goal of this applicative verb is a PP with a zero-postposition as its head. The adpositional object is a null pronominal in both the Yimas and Mohawk examples. The applicative verb is unusual in that its theme-role is assigned to a VP – *waraca* and *kwétar* in the above examples. The oddity of this applicative extends further, in that both it and the theme-VP select the same argument on which to discharge their agent theta-role. The theme-VP in both examples also selects for its own theme, the overt *awt* 'fire' in Yimas and the incorporated *natar* 'bread' in Mohawk.

Polysynthetic applicatives also may be, according to Baker (1996), derived through a form of incorporation. In languages where the applicative morpheme appears as a prefix, the construction is formed through postposition incorporation. The applicative postposition begins as an adjunct of a VP and early in the derivation (before the V° begins to move through the syntactic tree) the PP is moved as a whole and adjoined to AspP. The applicative P° then incorporates into Asp°, and the V° raises to form a complex head.

Mayali (Gunwingguan)

(3) *na-mege daboldabolk bandi-marne-ganj-ginje-ng.*
CLI-that old.people 3PLS>3PLO-BEN-meat-cook-PST/PFCT
'They cooked meat for the old people.' (Baker 1996: 439)

The Mayali example is derived in this fashion. First, the benefactive PP is raised and joined to the AspP projection. Next the theme NP (*ganj*) is incorporated into the V° according to the principles of noun incorporation.

The complex [N° + V°] further raises and incorporates into the [P° + Asp°], forming the core of the above example.

O'Herin (2001, 2002) extends Baker's theory of applicative formation through adposition incorporation by including an agreement phrase (AgrP) above the PP – which Baker posits for independent PPs in Mohawk under the title of Func(tional)P – to account for the presence of applicative agreement in Abaza.

Mohawk
(4) *ka-nakt-óku.*
3SGS.N-bed-under
'Under the bed.' (Baker 1996: 406)

Abaza (Northwest Caucasian)
(5) *y-p-s-q'ə-t'.*
3SGA.N-PV-1SGE-break-DYN
'I broke it.' (O'Herin 2001: 482)

Abaza
(6) *yə-l-č^{w}ə-p-s-q'ə-t'.*
3SGA.N-3SGD.F-MAL-PV-1SGE-break-DYN
'I broke it to her disadvantage.' (O'Herin 2001: 482)

The Northwest Caucasian language Ubykh provides an example of a similar construction – one in which the applicative morpheme stands as a prefix and also shows agreement with the applied object.

(7) *sə-laylak'ə-n* *za-g°ará-Ø* *Ø-s-x'á-nə-w-q'a.*
1SGPSG-stork-ERG one-INDF-ABS 3SGA-1SGD-BEN-3SGE-bring-PST
'My stork brought me something.' (Dumézil & Esenç 1978: 25)

Ubykh has a rather productive system of postpositional applicatives similar to that described by Baker (1996) and O'Herin (2001, 2002). There is also what appears to be a limited amount of NI, which follows many of the technical properties of Baker (1996), but as will be shown is not the direct incorporation of a nominal.

This paper seeks to provide a description of the incorporation strategies found in Ubykh, beginning with an outline and analysis of the agreement system before moving on to a description of applicatives and incorporation. The paper closes with a proposal for an analysis of the Ubykh data.

2. Description

Ubykh is a member of the Northwest Caucasian family of languages. It was originally spoken along the Black Sea coast in the area of what is now the Russian city of Sochi, Krasnodar Krai. Virtually the entire population relocated to the then Ottoman Empire following the conquest of their homeland by the Russian Empire in the mid-19th century. The Ubykh assimilated into Turkish culture within the next generation or so, giving up their language in favor of Turkish or one of the other Northwest Caucasian languages spoken by the diaspora, such as Adyghe or Kabardian. Ubykh is now an extinct language, as Tevfik Esenç, the last native speaker, died in 1992.

It is a polysynthetic language having minimal nominal morphology and extensive verbal inflections. Ubykh verbs are polypersonal, and may agree with absolutive, ergative and a variety of dative/oblique arguments. They are also inflected for tense, aspect, mood, number and polarity and a wide variety participials and gerundives.

Ubykh is, for the most part, a head-final language. The only exception to this is determiners such as the definite article, demonstratives and numerals, which are head-initial. Clauses always end with the verb. Matrix clauses must be terminated by a finite verb, and subordinate clauses are ended by a gerundive or other non-finite verb form. There is extensive pro-drop, and nominal arguments may be rather freely omitted. When nominals are present, the subject typically begins the clause and the direct object appears immediately before the verb. If an indirect object is present, the unmarked order places it between the subject and direct object. However, in a sampling of sentences where both the direct and indirect object appear, there is almost an even split between the orders IO-DO and DO-IO.

Much of the previous research on Ubykh has centered on its unique phonology – which has between 80 and 83 consonants and three phonemic vowels, which are differentiated only by their respective height – including work by Colarusso (1988, 1992), Dumézil (1955, 1957), Dumézil and Namitok (1954), Hewitt (1986), Leroy and Paris (1974) and Vogt (1963). A number of people have also prepared descriptive grammars or grammatical sketches of Ubykh, such as Charachidzé (1989), Dirr (1928, 1929), Dumézil (1931), Genko (1928), Hewitt (2005), Kumakhov (1998), and von Mészáros (1934). However, I believe the only work on Ubykh using modern syntactic theory is that of Özsoy (1988) which analyzes the system of relativization in the language.

2.1 Transitivity

Ubykh distinguishes transitivity by both case marking and verbal inflection. The subject of an intransitive verb and the patient of a transitive verb are case-marked with the absolutive, which surfaces as a zero-morpheme with no distinction in number. The agent of a transitive verb is marked by the ergative[1], which has the form *-n* in the singular and *-na* in the plural. Ubykh does not distinguish, in either case marking or verbal agreement, between unergative and unaccusative intransitives.

(8) *a-tə́t-Ø Ø-q̄a-q'á.*
the-man-ABS 3SGA-run-PST
'The man ran.' (Charachidzé 1989: 370)

(9) *a-čədə́-Ø a-χaśá-n.*
the-donkey-ABS 3SGA-be.tired-PRS
'The donkey is tired.' (Charachidzé 1989: 450)

(10) *yə́-śa-k°abž'a-na á-lašąpła-Ø Ø-ą-k°'-áwt°q'a.*
this-three-man-ERG.PL the-army.chief-ABS 3SGA-3PLE-kill-FUT.PFCT
'These three men will have killed the army chief.' (Charachidzé 1989: 390)

There is also a conjugation class of verbs having an absolutive subject and a dative object. The vast majority of these verbs have an incorporated postposition or the translocative prefix governing the dative argument.

(11) *sə-w-bɣ'a-k''a-n.*
1SGA-2SGD-over-go-PRS
'I conquer you.' (Lit. 'I go over you.') (Vogt 1963: 94)

There is a restricted sub-class of this group in which the dative appears without a governing morpheme as in ditransitive verbs.

(12) *a-tə́t-Ø a-x'ə́-n yə́-Ø-ya-q'a.*
the-man-ABS the-prince-DAT 3SGA-3SGD-strike-PST
'The man struck the prince.' (Charachidzé 1989: 439)

[1] Traditional descriptions of Ubykh grammar note only two primary case markers, the zero-marked absolutive and the oblique. The oblique is used to mark ergative, dative, adverbial and genitive cases, without discrimination. Throughout this paper, I have decided to clarify the specific functions of this case marker, especially in relation to the coindexing of the verbal agreement morphemes.

2.2 Agreement

The intraverbal agreement noting the first and second person, singular and plural, is the same for the absolutive, ergative and dative.

	SG	PL
1	*sə-*	*š′ə-*
2	*wə-*	*s°ə-*

The third person has several variants depending on number and syntactic alignment.

	ABS	ERG	DAT
SG	*a-, Ø-, (y)ə-*	*Ø-, n(ə)-*	*ɣa-; Ø-*
PL	*a-, yə-, Ø-*	*ạ-, nạ-*	*aɣa-; ạ-*

The 3rd person absolutive form *(y)ə-/yə-* is used preceding a 3rd person singular dative inflection (which is always Ø-). The *a-* form is found before a 1st or 2nd person or a 3rd person plural dative. The absolutive forms are quite fragile and are often reduced to their zero-variant regardless of the following agreement morpheme.

(13) *(yə́)-Ø-w-t°ə-n.*
3SGA-3SGD-2SGE-give-PRS
'You give it to him.' (Dumézil 1975: 90)

(14) *a-s°ə́-š′-t°-ạn.*
3PLA-2PLD-1PLE-give-PRS.PL
'We give them to you all.' (Dumézil 1975: 91)

The 3rd person ergative forms are conditioned by the valency of the verb. The forms Ø-/*ạ-* are found with transitive verbs which have no dative argument, and the forms *n(ə)-/nạ-* are found elsewhere.

(15) *wə-Ø-byə́-n.*
2SGA-3SGE-see-PRS
'She sees you.' (Charachidzé 1989: 395)

(16) *ə́-məš°a-n a-płaq°'aq'ə́-Ø yə́-Ø-n-t°ə-n.*
the-bear-DAT the-silver-ABS 3SGA-3SGD-3SGE-give-PRS
'He gives the silver to the bear.' (Dumézil 1967: 165)

The dative uses a different series of intraverbal agreement than does the absolutive and ergative. This agreement uses the possessive singular inflections with the 3rd person singular having the variant Ø-/*ạ*- in all circumstances except before the translocative *ạ*-, in which case it has the form *ɣạ-/aɣạ-* (> *ɣa:ạ-/aɣa:ạ-*).

(17) *wə-ɣạ́-ʒɣa-n.*
2SGA-3SGD:TRLOC-question-PRS
'You ask her.' (Dumézil 1975: 76)

(18) *w-aɣạ́-ʒɣa-n.*
2SGA-3PLD:TRLOC-question-PRS
'You ask them.' (Dumézil 1975: 76)

The dative bloc of inflections is located between the absolutive and ergative agreement and serves a variety of functions. Agreement with an indirect object is the simplest, and possibly the most common use.

(19) *Nạzəm beyə-n sɣ°a a-sə-n-q'adaǯ'-q'a.*
Nâzım bey-ERG I.DAT 3SGA-1SGD-3SGE-tell-PST
'Nâzım bey told it to me.' (Dumézil 1965: 43)

A second use is agreement with an applied or adpositional object. This function requires the incorporation of an additional morpheme, depending on the nature of the object. It may be any one of five applicatives (benefactive, malefactive, comitative, translocative or ablative), a postpositional object or the object of an incorporated nominal, which will be covered in more detail below.

2.3 Incorporated Material

There are four types of elements which may be incorporated into the verbal complex. Three of these always appear with dative agreement: incorporated nouns, incorporated postpositions and applicatives. The fourth, directional preverbs, never appear with agreement. These elements may cooccur with one another, but to my knowledge there cannot be multiple occurances of a single type of element in a single verbal complex.

2.3.1 Incorporated Nouns and Postpositions

Ubykh has a productive system of incorporated postpositions. Many of these are derived from obsolete nominals referring to body parts and which may also be used as incorporated nouns. A limited number of these incorporated postpositions may be used productively outside the verb complex as postpositions. Most of the spatial relations in Ubykh are handled by the incorporated postpositions, which may take a range of subtle senses based on the meaning of the verb root to which they are attached. The postpositions may have the sense of location (adessive), motion to (allative), motion from (ablative), or, with the addition of *-ɣa* to the postposition, motion through (perlative).

(20) *á-ɣ°ənə-n* *sə-Ø-baċ'a-q°'át°ə-n.*
the-tree-DAT 1SGA-3SGD-under-stop-PRS
'I stop under the tree.' (Dumézil 1975: 106)

(21) *yə́-šak''aq'a-n* *š'ə-Ø-baċ'a-k'á-naw-mạla...*
this-shelter-DAT 1PLA-3SGD-under-enter.PL-PL-FUT1-GER
'We will be entering this shelter...' (Dumézil 1975: 106)

(22) *á-šanə-n* *Ø-Ø-baċ'á-sə-wt°'ə-n.*
the-table-DAT 3SGA-3SGD-under-1SGE-remove-PRS
'I take it from under the table.' (Dumézil 1975: 106)

(23) *a-q̄°'á-n* *š'ə-Ø-baċ'aɣa-la-x°a-q'an.*
the-cave-DAT 1PLA-3SGD-under:PERL-LOC-pass-PST.PL
'We passed through (under) the cave.' (Dumézil 1975: 106)

Contrasting with the postpositions, Ubykh also has a fossilized system of incorporated nominals. Only one of these nominals (*d°ə*, 'outside') may appear either independent or incorporated and maintain a minimal semantic contrast between the forms. Another independent nominal *č'ə́* 'horse' alternates with the incorporated nominal *č'a* 'knight' in a similar fashion as can be seen in the example below. The remaining incorporated nominals have formed idiomatic compounds from which they can no longer be separated.

(24) *a-d°á-n* *sə-Ø-g'ə-wá-n.*
the-outside-DAT 1SGA-3SGD-in-enter-PRS
'I leave, I go outside.' (Dumézil 1975: 112)

(25) *sə-Ø-d°ə-g'ə-wá-n.*
1SGA-3SGD-outside-in-enter-PRS
'I go outside.' (Dumézil 1975: 112)

(26) *a-č'á-n* *sə-Ø-bɣ'a-wəsə-n.*
the-horse-DAT 1SGA-3SGD-on-sit.down-PRS
'I mount the horse.' (Charachidzé 1989: 431)

(27) *sə-Ø-č'á-bɣ'a-wəsə-n.*
1SGA-3SGD-knight-on-sit.down-PRS
'I mount a horse.' (Charachidzé 1989: 431)

(28) *a-z-g'ə-c°'ac°'ə-n.*
3SGA-1SGD-heart-hurt-PRS
'I pity him.' (Lit. 'He hurts my heart.') (Vogt 1963: 124)

The distinction between incorporated postpositions and nominals is subtle. Many of the roots can be interpreted as either form with no overt marking to differentiate them. There are, however, several points where the two forms do not agree. Most clearly, the incorporated nominal but not the postposition may take agreement in the same person and number as the subject of the verb without using an overt reflexive marker.

(29) *a-s-łą́-sə-ɣa-n.*
3SGA-1SGD-foot:TRLOC-1SGE-hang-PRS
'I put on my shoes.' (Dumézil 1975: 104)

A similar structure with a postposition is ill-formed and requires a synthetic circumlocution to become grammatical. This is the same synthetic form which replaces the intraverbal reflexive.

(30) **a-z-bɣ'á-sə-wt°'ə-n.*
3SGA-1SGD-on-1SGE-remove-PRS
'I take it off of me.' (Dumézil 1975: 105)

(31) *sə-g'á-n* *Ø-Ø-bɣ'á-sə-wt°'ə-n.*
1SGPSG-self-DAT 3SGA-3SGD-on-1SGE-remove-PRS
'I take it off of myself.' (Dumézil 1975: 105)

Another distinction between the postpositions and nominals which can be seen in the above examples is the interpretation of the dative argument in relation to the incorporated root. The dative occurring with the postposition has the sense of a postpositional object, whereas that occurring with the nominal is seen as the possessor of the nominal. Often, this possessive relationship is lost due to the idiomatic nature of Ubykh incorporation, but in many instances the literal sense can be extrapolated.

(32) *a-s-q'ᵊ́-ɣ-Ø.*
3SGA-1SGD-hand:TRLOC-be.hanging-STAT.PRS
'I have it.' (Lit. 'It is hanging from my hand.') (Dumézil 1975: 119)

2.3.2 Applicatives

Ubykh has five types of applicatives: benefactive, malefactive, comitative, translocative and ablative. Each of these is signaled by an incorporated morpheme preceded by an agreement inflection.

The benefactive is marked by *x'a-* in the verb complex and may indicate not only that the action is done for the benefit of or more simply for the applied object, but also a directive aspect of motion towards a goal.

(33) *wəɣ°á pč'arəxá-n sə-w-x'a-š'-áw.*
you.DAT squire-ADV 1SGA-2SGD-BEN-become-FUT1
'That I (may) become a squire for you.' (Dumézil 1975: 140)

(34) *sə́-laylak'ə-n za-g°ará-Ø Ø-s-x'á-nə-w-q'a.*
1SGPSG-stork-ERG one-INDF-ABS 3SGA-1SGD-BEN-3SGE-carry-PST
'My stork brought me something!' (Dumézil & Esenç 1978: 25)

The malefactive is indicated by *c°ə-* and, contrary to the benefactive, denotes that the action is done to the detriment of or against the will of the object. It may also take a directive sense, like the benefactive, and indicates motion from a person or place.

(35) *y-ᵊ́ḇa-Ø Ø-š'-c°ə-dwá-wt.*
this-illness-ABS 3SGA-1PLD-MAL-die-FUT2
'This illness will kill us.' (Dumézil 1975: 141)

(36) *a-s-c°ə-q̄á-n.*
3SGA-1SGD-MAL-run-PRS
'He runs away from me.' (Charachidzé 1989: 428)

The comitative, marked by *ǯ'ə-*, denotes that the action is done with the object, but not in an instrumental sense.

(37) *wə-z-ǯ'ə-k''á-wt.*
2SGA-1SGD-COM-go-FUT2
'You will marry me.' (Dumézil 1975: 139)

The translocative *ạ-*, which Dumézil (1975) called "la particule attributive," and which is unique among the applicative inflections in that the 3rd person agreement shows up as a possessive prefix *ɣa-* instead of Ø-. The exact meaning of the translocative is difficult to establish in most cases, but when combined with an incorporated postposition it takes on a general sense of direction toward.

(38) *sə-ɣạ́-ʒɣa-n.*
1SGA-3SGD:TRLOC-question-PRS
'I ask him.' (Dumézil 1975: 142)

Finally, there is the ablative *ɣa-*, which indicates motion away from its object.

(39) *a-w-ɣá-sə-wt°'ə-n.*
3SGA-2SGD-ABL-1SGE-remove-PRS
'I take it (away) from you.' (Dumézil 1975: 80)

Similar to other languages which use a prefixed applicative morpheme, there is no transitivity restriction in Ubykh as to the addition of the applicative. It may equally appear with intransitive and transitive verbs. The issue of high versus low applicatives will be addressed below.

(40) *a-s-c°ə-z°aps-q'á.*
3SGA-1SGD-MAL-be.evening-PST
'It became evening to my surprise.' (Charachidzé 1989: 429)

(41) *a-nd°aša-Ø a-Nartə-n Ø-Ø-x'a-nạ-ɣa-k'a-q'an.*
the-rope-ABS the-Nart-DAT 3SGA-3SGD-BEN-3PLE-CAUS.PL-carry.PL-PST.PL
'They brought the rope to the Nart.' (Dumézil 1965: 164)

2.3.3 Directional Preverbs

Finally there are two directional preverbs, neither of which are found with agreement. The first is the cislocative prefix *y-* and the second the general locative *la-*.

The cislocative indicates, generally speaking, motion toward the participant. This participant is typically not stated overtly, but may be indicated using a postpositional phrase headed by *-laq* 'at, near (to)'.

(42) *sə-y-k''a-n.*
1SGA-CIS-go-PRS
'I am coming.' (Dumézil 1975: 132)

(43) *səɣ°a sə-laq a-y-n-wə-n…*
I 1SGPSG-at 3SGA-CIS-3SGE-carry-PRS
'He brought it to me.' (Dumézil 1961: 50)

The general locative prefix indicates a static, general location.

(44) *sə-la-g'ət°ə-n*
1SGA-LOC-remain-PRS
'I remain there.' (Vogt 1963: 136)

2.4 Analysis of Transitivity and Ergative/Absolutive Agreement.

The underlying structure of Ubykh case-assignment and agreement may be analyzed along the lines of Baker's (1996) Uniformity of Theta Assignment Hypothesis (UTAH) and Bittner & Hale's (1996) theory of case assignment. According to the UTAH, thematic roles are always assigned to the same structural positions across comparable verbs. The theme appears in the specifier of the minimal VP. The agent likewise appears in the specifier of a VP-external phrase – termed *v*P – which is located immediately above the VP.

The assignment of absolutive case in both transitive and unaccusative verbs is conditioned by the fulfillment of Bittner & Hale's (1996) K-Filter, given below.

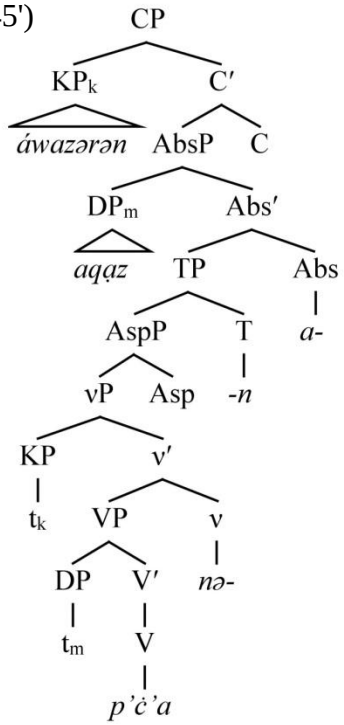
K-Filter
An argument chain headed by a K-less nominal (DP or NP) contains a position that is c-commanded and governed by K or C, and does not contain any Case-bound position. (Bittner & Hale 1996: 12)

The theme is required by this filter to raise to the specifier of the projection immediately below the CP, which in this analysis and that of O'Herin (2002) for Abaza is termed the Absolutive Phrase (AbsP). Once the theme is raised, the unmarked absolutive is licensed by its new structural position. Thus, the two examples below may be diagrammed as in (45') and (46') (the mechanics of the ergative phrase in (45) will be explained below).

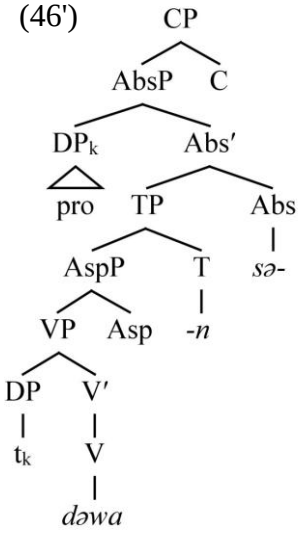
(45) *á-wazərə-n a-qą́z-Ø á-nə-p'ċ'a-n.*
the-vizier-ERG the-goose-ABS 3SGA-3SGE-clean-PRS
'The vizier cleans the goose.' (Dumézil & Esenç 1978: 59)

(46) *sə-dəwá-n.*
1SGA-die-PRS
'I die.' (Vogt 1963: 114)

(45')



(46')



As stated above, the agent theta-role is assigned to the specifier of vP. The agent, if it is a nominal and the subject of a transitive verb, has the form of a DP contained within a KP. Following Bittner & Hale (1996), case is

assigned to this position through case-binding from the Aspect Phrase (AspP) directly above the νP. I take AspP to be the lowest member of the "exploded" IP, which is composed of AbsP, a Tense Phrase (TP) and AspP. Bittner & Hale (1996) propose that in this position, K° must be realized as Ergative by their Direct Case Realizations below. After receiving its case, the agent KP raises to the specifier of CP.

> Direct Case Realizations
> If α Case-binds an overt empty-headed KP β, then the empty K of β is realized as:
> (i) ERG, if α is I (or D);
> (ii) ACC, if α is V (or P) and has an adjoined D.
> (Bittner & Hale 1996: 11)

The agent theta-role is also assigned to the sole argument in unergative verbs. However, this argument does not take the ergative case as one would expect. Instead, it is assigned the absolutive case, as in Samoan (Austronesian) or Warlpiri (Pama-Nyungan).

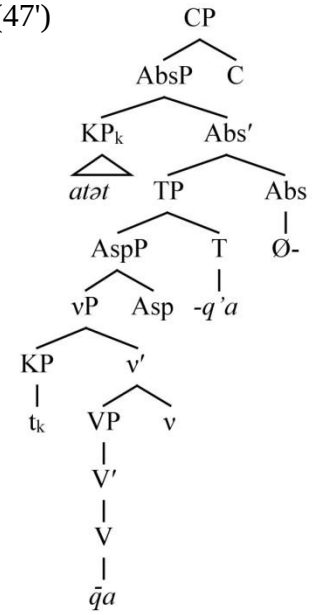
(47) *a-tə́t-Ø* *Ø-q̄a-q'á.*
the-man-ABS 3SGA-run-PST
'The man ran.' (Charachidzé 1989: 370)

Samoan
(48) *sa sola Ø le teine.*
PST run.away ABS the girl
'The girl ran away.' (Bittner & Hale 1996: 29)

Warlpiri
(49) *ngaju ka-rna parnka-mi.*
me.NOM PRS-1SG run-NPST
'I am running.' (Bittner & Hale 1996: 31)

The Ubykh example (47) is diagrammed in (47').

(47')



The assignment of absolutive case to the agent of an unergative is a result of the K-Filter (see above). This filter, ultimately requires the specifier of AbsP to be filled. Normally this position is filled by the theme argument, but as unergatives only project an agent, it must raise in order to fill this position. The verb itself does not case-bind anything because it does not c-command any argument. Furthermore, though the Asp° locally c-commands the trace in the specifier of vP, it does not case-bind this position because there is no case-competition, which is required in order for Asp° to assign ergative case to the specifier of vP. The agent DP, by raising to the specifier of AbsP, fulfills the K-Filter and is assigned absolutive case by virtue of the fact that it is both c-commanded and governed by C°.

2.4.1 An Introduction to Applicatives

Pylkkänen (2002) approaches the question of applicatives from a semantic standpoint. She divides applicatives into two classes, termed high and low. High applicatives denote a thematic relationship between an individual and an event, such as a benefactive. Low applicatives indicate a relation of possession, or transfer of possession, between the applied object and the direct object. The Venda sentence below illustrates the high applicative. The event, or action, of "building" occurs for the benefit of the applied object "father". It is the applied object which registers agreement in the verbal complex and not the direct object. The Korean example shows a

transfer of possession of the direct object “ring” from “Mary” to the “thief”, and is thus indicative of a low applicative construction.

Venda (Bantu)

(50) *ndi mu fhaṱ-el-a nnḓu khotsi.*
SM.1SG OM.CL1 build-APPL-FV hut.CL9 father.CL1
‘I build a house for father.’ (Ziervogel et al. 1972: 113)

Korean (Isolate)

(51) *totwuk-i Mary-hanthey panci-lul hwumchi-ass-ta.*
thief-NOM Mary-DAT ring-ACC steal-PST-DECL
‘The thief stole a ring from Mary.’ (Pylkkänen 2002: 21)

Structurally, high applicatives appear below the vP and above the VP, while the low applicative appears immediately below the VP with the direct object as the complement to Appl°. Since the low applicative denotes a relationship between the applied and direct object, it can only appear attached to transitive or unaccusative verbs because unergative verbs lack a theme argument. High applicatives, on the other hand, are free to combine with unergative verbs as well as transitives and unaccusatives.

Finally, Baker (1996) analyzes polysynthetic applicatives as being split between two structural classes, those where the applicative morpheme appears as a suffix and those where it is a prefix. Languages such as Mohawk and Nahuatl are clear examples of the former group, as may be seen in the Classical Nahuatl (Uto-Aztecan) sentence below.

Classical Nahuatl

(52) *ni-quin-tlaxcal-temo-lia-Ø in no-pil-huan.*
1SGS-3PLO-tortilla-seek-APPL-PRS IN 1SGP-son-PL
‘I seek bread for my sons.’ (Anderson 1973: 104)

According to Baker (1996) – and which was described above in the introduction to this article – suffixed applicative morphemes head their own verbal projection and may assign up to three thematic roles (agent, theme and goal/path) depending of the nature of the verb to which they are suffixed. Often, the goal/path theta-role is assigned to a adpositional phrase headed by a null P° and an unpronounced pro which may be coindexed in the verbal agreement morphology. The theme is given to the higher VP which contains the action or state denoting verb stem. The theme of this higher VP is discharged on its own direct object or unaccusative subject – a form which is often incorporated in the polysynthetic languages described

in Baker (1996). If an agent theta-role is present it is assigned to the agent of the higher verb.

The applicative suffix *-lia* in the above example heads its own VP and assigns a goal theta-role to the phrase "my sons". The theme role is assigned to a higher VP, headed by the verb *temo* 'seek'; the theme of this verb is given to the incorporated nominal *tlaxcal-* (the full form of which is *tlaxcalli*). The agent theta-role of the applicative verb *-lia* and the higher verb *temo* is assigned to the same argument, in this case the phonetically null, 1st person pro.

Prefixed applicatives – such as those in Ainu, Mayali and the Northwest Caucasian languages – are the result of adposition incorporation. The applicative begins as a PP below the verb and raises to be adjoined to a phrase above the νP, which Baker (1996) analyzes as the AspP. The P°, hosting the applicative morpheme, merges with Asp° before the verb begins movement and later joins to the verbal complex as the verb passes through Asp°.

Ainu (Isolate)

(53) *huci* *matkaci* *ko-paskuma.*
grandmother girl APPL-tell.old.stories
'Grandmother told the old stories to the girl.' (Shibatani 1990: 54)

2.4.2 Low Applicatives

According to Pylkkänen (2002), the low applicative structure is responsible for introducing an additional argument below the VP which has a relationship with the direct object indicating possession or a transfer of possession. Following Pylkkänen (2002), Georgala et al. (2007) and Jeong (2006), the low applicative is generated as the complement of V°. The dative object appears in the specifier of ApplLP with the corresponding agreement in ApplL°. The theme object is generated as the complement of ApplL°.

This type of applicative is found with ditransitive verbs such as *t°ə* 'give' which indicate a transfer of possession from the direct object to the indirect object.

(54) *waná-n* *za-tχə́ɬ-g°ara-Ø* *Ø-sə́-n-t°-q'a.*
that-ERG one-letter-INDF-ABS 3SGA-1SGD-3SGE-give-PST
'He gave me a (certain) letter.' (Dumézil & Esenç 1978: 59)

The derivation of the surface form of the verb is the result of the V° raising and joining with ν° to form the complex head [ν° + V°]. At this phase boundary, the ApplL° containing the dative agreement raises and joins with the complex head forming [ApplL° + [ν° + V°]].

2.4.3 High Applicatives

Contrasting with the low applicative, high applicatives introduce an additional argument between the νP and VP which indicate a thematic relationship (such as a benefactive) between the applied object and the event denoted by the verb (Pylkkänen 2002, Georgala et al 2007, Jeong 2006). The Ubykh high applicative (ApplHP) generates the applied object in the specifier of ApplHP in the form of a postpositional phrase.

According to Baker's analysis – which is also used by O'Herin (2001, 2002) to describe his Abaza data – PPs project a functional category immediately above themselves which hosts agreement with their object. Baker terms this FuncP, but following O'Herin (2001) I have simply termed this an Agreement Phrase (AgrP). This analysis is diagrammed in Figure 1.

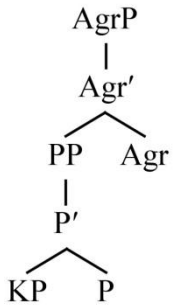


Figure 1. An analysis of preposition phrases according to O'Herin (2001).

The applicative construction fills P° with one of the applicative morphemes. The applicative P° selects for a KP, under which the applied object (either a DP or NP) is placed. The applicative always appears to require the dative case. The AgrP agrees with the applied object. The P° raises to Agr° and forms a complex head with it. This complex [Agr° + P°]

then incorporates into the verb at v°. Unlike the low applicative, the ApplH° remains vacant.

(55) *a-x'ə́-n* *wałá-na* *za-tχəł-Ø* *Ø-ạ-x'á-n-tχ-q'ayt'.*
the-prince-ERG those-DAT.PL one-letter-ABS 3SGA-3PLD-BEN-3SGE-write-PLUP
'The prince had written a letter for them.' (Charachidzé 1989: 429)

While none of the languages displaying prefixed applicatives in Baker (1996) show agreement with the applied object, overt agreement is found throughout the Northwest Caucasian group. The applicatives in the examples below show oblique agreement with their objects.

Abkhaz (Northwest Caucasian)
(56) *a-pħ°ə́s* *a-ẋác'a* *a-ẋárp* *Ø-yə-zə́-lə-ʒ̇°ʒ̇°a-yt'.*
the-woman the-man the-shirt 3SGA.N-3SGD.M-BEN-3SGE.F-wash-AOR
'The woman washed the shirt for the man.' (Hewitt 1989: 67)

Shapsug Adyghe (Northwest Caucasian)
(57) *mə* *xaŝʷe-r* *a-ƒ-a-wəqebzə-ʁ* *ležekʷ'e.pːsewaqʷe-me.*
this field-ABS 3PLD-BEN-3PLA-clear-PST peasant-OBL.PL
'This field was cleared for peasants.' (Lander 2010: 78)

Interestingly, both the high and low applicatives may be found in the same verbal complex, as in example (58) below. Also, multiple high applicatives may surface in a single verbal complex as in (59) below.

(58) *za-m̲á-Ø* *Ø-t-x'a-wə́-s-t°ə-n* *təṭ-Ø.*
one-apple-ABS 3SGA-REL-BEN-2SGD-1SGE-give-PRS man-ABS
'The man for whom I give you an apple.' (Charachidzé 1989: 441)

(59) *a-s-x'á-w-ɣa-nə-wt°'-ạy-awt.*
3SGA-1SGD-BEN-2SGD-ABL-3SGE-take-ITER-FUT2
'He will take it back from you for me.' (Dumézil 1975: 102)

The high applicative construction is also responsible for the assignment of dative case to the direct object of verbs such as *ya* 'strike', *k''ał'a* 'approach', *pła* 'look at' and *məśa* 'call (to)'. This class of verbs comprises approximately one dozen roots, most of which are derived directly from unergative verbs. These have their valence increased using a high applicative indicating a goal of the verbs' action.

(60) *a-tə́t-Ø* *a-x'ə́-n* *yə́-Ø-ya-q'a.*
the-man-ABS the-prince-DAT 3SGA-3SGD-strike-PST
'The man struck the prince.' (Charachidzé 1989: 439)

Compare also the monovalent unergative in (61) with the bivalent applicative in (62).

(61) *wa-z°apsə́ǯ'* *za-p'č''á-g°ara-Ø* *a-məṡa-q'a.*
that-night one-guest-INDF-ABS 3SGA-call-PST
'That night a (certain) guest called.' (Vogt 1963: 148)

(62) *Sə́wsərəq°a-n* *s°ə-Ø-mə́ṡa-n!*
Sə́wsərəq°a-DAT 2PLA-3SGD-call-PRS
'(You all) call to Sə́wsərəq°a!' (Vogt 1963: 148)

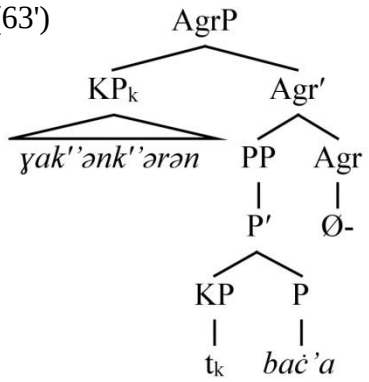
2.5 Ubykh Postposition Incorporation

The incorporation of postpositions in Ubykh is similar to the structure of applicatives, with the various postpositions, rather than an applicative marker, being hosted in P°. This postpositional phrase, instead of being generated in the specifier of a high applicative projection, begins below the VP. Example (63') details the AgrP of example (63).

(63) *a-k''arąx°a-Ø* *ɣa-k''ənk''ərə-n* *Ø-Ø-baċ'a-n-q°'a-q'a.*
the-revolver-ABS 3SGPSG-adam's.apple-DAT 3SGA-3SGD-under-3SGE-stick-PST
'He put the revolver under his throat.' (Alparslan & Dumézil 1964: 352)

(64) *səɣ°á sə-ɣac'a-n* *Ø-Ø-ɣac'á-ɬ-Ø* *a-w-ċ'á-nə-ṡ?*
I.GEN 1SGPSG-inside-DAT 3SGA-3SGD-inside-be-NFIN 3SGA-2SGE-know-PRS-Q
'Do you know what is inside me?' (Charachidzé 1989: 373)

(63')



The PP raises and is adjoined to the νP. Once the verb raises and joins with ν°, the [Agr° + P°] excorporates from the adjoined phrase and merges with the verbal complex in ν°.

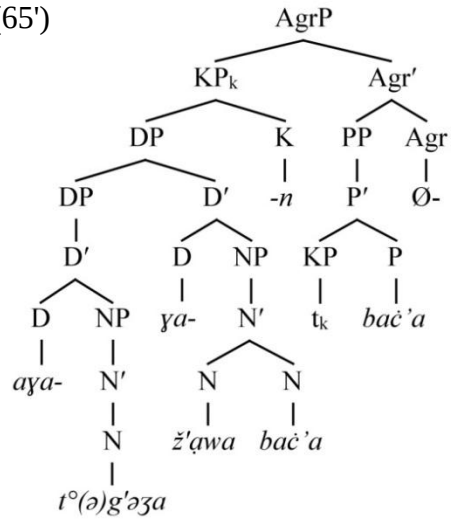
Another matter to be examined is the doubling of the incorporated element. Only the incorporated postpositions may be doubled in Ubykh. There appears to be no discernable difference in meaning between example sentences where both forms are given. This construction is infrequently used and is significantly more complex than its undoubled counterpart.

This construction relies on what may be called "relational nouns" set in a genitive relationship with their object as in example (64). The relational noun appears with a possessive prefix coindexed with the object, and one of three cases – the locative *-ɣa*, the instrumental *-awnə*, or the dative *-n ~ -na*. The selection of this case is dependent on the function of the structure. The locative typically indicates motion, either to or from, the instrumental marks motion by or through and the oblique is used elsewhere.

The relational noun may also appear compounded directly with its object as the example (65) below. The compound, as the structure examined above, may be inflected with the same array of cases in the same situations.

(65) *aɣa-t°(ə)g'əʒa-Ø* *ɣa-ž'ạwa-bạ̀'a-n*
3PLPSG-grandfather-GEN 3SGPSG-shadow-under-DAT
Ø-Ø-bạ̀'a-χa-nan.
3SGA-3SGD-under-be.PL-PRS.PL
'(Sitting) under the shadow of their grandfather.' (Alparslan & Dumézil 1964: 352)

(65')



2.5.1 Baker's Movement Theory of Noun Incorporation

According to Baker (1996), noun incorporation is derived through the head movement of a verbal object – typically a theme argument – from its base position to be joined with the verb stem. This movement is motivated by the Morphological Visibility Condition (MVC), given below.

> Morphological Visibility Condition
> Every θ-role associated with a head Y must be coindexed with a distinct morpheme in the word containing Y. (Baker 1996: 286)

Simply put, the MVC states that if a word assigns a theta-role to another word or phrase, that word or phrase must either show agreement on the theta-role assigner or incorporate into it. Gronemeyer (1996) develops this position and suggests that incorporation also stems from or may be triggered by a morphological defect which disallows certain stems from appearing as independent words, such as West Greenlandic post-bases. The morpheme *-qar* 'have' in the following example is required to be suffixed to another word form in order to be properly realized.

West Greenlandic (Eskimo-Aleut)
(66) *kunngi-p panip-passua-qar-poq.*
king-ERG daughter-many-have-3SG.IND
'There are many king's daughters.' (Malouf 1999: 48)

She also notes, as does Baker (1996), that determiner phrases are defective in polysynthetic languages. If a nominal is dominated by a DP, that noun is prohibited from incorporating because doing so would violate the Head Movement Constraint by skipping over the D°.

> Head Movement Constraint
> An X° category Y can only adjoin to the head of the phrase that immediately dominates the maximal projection of Y. (Baker 1996: 284)

Incorporating first into the D° and then into the verb is also prohibited by the Proper Head Movement Generalization. Thus, noun incorporation can only affect a theta-marked, bare NP.

> Proper Head Movement Generalization
> A lexical category cannot move into a functional category and then back into a lexical category. (Baker 1996: 284)

Baker (1996) continues his analysis by examining why certain theta-roles incorporate and why others typically do not. Goals (and by extension, Paths) will not incorporate because ditransitive verbs in polysynthetic languages call for a PP headed by a null adposition to express the argument. This null adposition blocks the direct incorporation of the goal into the verb stem. Agents are simply never in a structural position where they are capable of incorporating into the verb.

Themes, on the other hand, are base generated as bare NPs, which does allow them to be incorporated into the verb before the verb begins its movement upward through the syntactic tree. Gerdts (1998) points out that certain languages can incorporate constituents such as instruments and passive agents (67) with its unincorporated counterpart in (68).

Southern Tiwa (Tanoan)
(67) *khwien-ide Ø-kan-ẽdeure-ban.*
dog-BAS 3SGS>3SGO-horse-kick.PASS-PST
'The dog was kicked by the horse.' (Gerdts 1998: 87)

Southern Tiwa
(68) *khwien-ide Ø-ẽdeure-ban kan-ide-ba.*
dog-BAS 3SGS>3SGO-kick.PASS-PST horse-BAS-INST
'The dog was kicked by the horse.' (Allen et al. 1984: 302)

Baker (1996), while not delving into the debate for lack of materials, believes that at least in the case of oblique incorporation it is a matter of lexical compounding instead of incorporation driven by head movement.

2.5.2 Ubykh Noun Incorporation.

Noun incorporation in Ubykh is not true noun incorporation (NI) in the traditional sense of the term. Typically, NI involves the head-movement of the theme N° from its base position to one where it is joined to the V°.

Mohawk
(69) *kíkʌ á'shar-e' ka-natar-a-kwétar-ʌs.*
this knife-NSF 3SGS.N-bread-LNK-cut-HAB
'This knife cuts bread.' (Baker 1996: 207)

This movement is triggered by the MVC, as described above. The theme "bread" gets its theta-role from the verb but lacks any coreference in it,

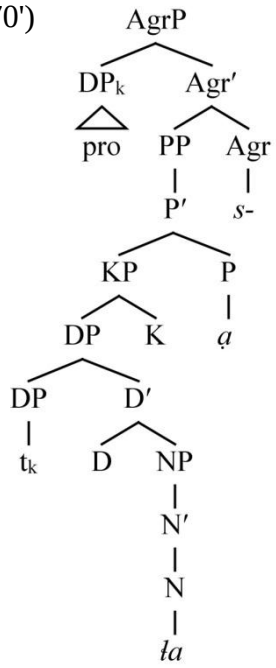
violating the MVC. If object agreement were generated elsewhere, it would receive case from the verb which would mean the overt theme "bread" violates the Case Filter (which simply states that all overt nominal arguments must be case-marked). NI is therefore the only means to produce a grammatical construction in languages like Mohawk.

Ubykh NI, on the other hand, relies on a construction similar to that of postposition incorporation with an external doubled postposition. The incorporated nominal (IN) begins as the object in a genitive construction which is itself the object of a null postposition (or an overt postposition such as the translocative). The Agr° dominating the PP hosts agreement with the possessor of the nominal, which is most often a 3rd singular pro. This pro (or an overt nominal) raises to the specifier of AgrP in order to be coindexed with the agreement in Agr°. This nominal object raises and incorporates into the P°, and the [N° + P°] complex raises to merge with Agr°. As with the incorporated postpositions, the [Agr° + [N° + P°]] merges with the verb in ν°.

(70) *sə-łabž'a-Ø* *Ø-s-łą́-sə-ɣa-n.*
1SGPSG-shoe-ABS 3SGA-1SGD-foot:TRLOC-1SGE-put-PRS
'I put on my shoes.' (Dumézil 1975: 117)

(71) *š'ə́-Ø-č'a-bɣ'a-k'áž°a-na-n.*
1PLA-3SGD-knight-on-sit.down.PL-PL-PRS
'Let's mount our horses!' (Dumézil 1975: 109)

(70')

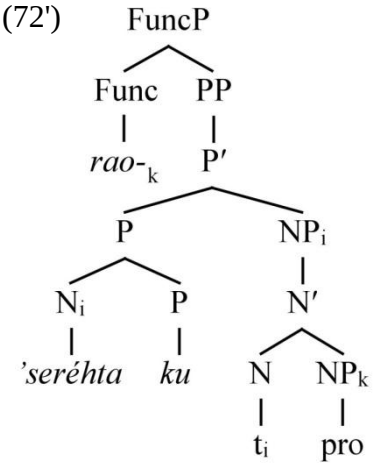


As may be seen in the above chart, the possessed object "foot" is moved out of its object position and forms a complex head with P° (the translocative *ą*-), and the pair raise together to form an even more complex head with Agr°. The form in (71) provides a relatively rare example of an incorporated nominal combined with an overt postposition (excluding the translocative and ablative forms). A variety of INs must occur with the translocative prefix *ą*-, which, if it was not combined directly with the IN, would necessarily occur with a possessive prefix as *sə*-, *wə*-, *ɣa*-, etc.

This structure is akin to that found in Mohawk postpositions having a possessed object. Take, for example, the following sentence, with the postpositional phrase diagrammed below.

Mohawk
(72) *Shawátis* *rao-'seré-ht-a-ku* *wa'-ke-'nerohkw-íta-'.*
Shawatis 3SGP.M-car-NMLZ-LNK-in FACT-1SGS-box-put-PUNC
'I put the box in Shawatis' car.' (Baker 1996: 411)

(72')



The postpositional object "car" is governed by a phonologically null pro which is coreferenced with the agreement prefixed to the postpositional complex and also externally to the adjunct nominal "Shawatis". This object is raised and incorporated into the postposition, and the complex is then raised to Func° to acquire the agreement.

Similar to the cooccurance of the high and low applicatives, the incorporation of a nominal may also be found with a high applicative as in the following example.

(73) *wába-n* *a-š'-x'a-z°a-nə-Ø-t°'a-q'ayła.*
god-ERG 3PLA-1PLD-BEN-sky-3SGE-CAUS-descend-PLUP.PL.PTCP
'Those (things) that god had made descend from the sky for us.' (Dumézil 1975: 34)

Ubykh noun incorporation in many cases has a reading akin to possessor raising. Possessor raising constructions involve the incorporation of a possessed nominal into the verb stem. The possessor then raises to fill the now vacant thematic position left by the possessed object (van Geenhoven 2002).

West Greenlandic
(74) *Nuka-p* *puisi-Ø* *ami-ir-p-a-a.*
Nuka-ERG seal-ABS skin-remove-IND-TR-3SGS>3SGO
i) 'Nuka removed the seal's skin.'
ii) 'Nuka removed the skin from the seal.' (van Geenhoven 2002: 772)

Certain examples of NI in Ubykh seem to fit with van Geenhoven's (2002) description of possessor raising. The absolutive arguments in the following two examples – *ɣatχ* and *ɣáməz*, respectively – can be interpreted as

possessors of the INs. Since the theme argument in each case has been incorporated, there is no argument left which can meet the requirement to fill the specifier of AbsP, save the possessor of the IN. The possessor is then marked by the absolutive case, instead of the genitive (which is typicaly reduced from *-n* to *-Ø* before the 3rd person possessive prefix *ɣa-*).

(75) *ɣa-tχ-Ø* *a-Ø-c°á-z-la-n.*
3SGPSG-back-ABS 3SGA-3SGD-skin-1SGE-remove-PRS
'I flay his back.' (Charachidzé 1989: 431)

(76) *ɣá-məz-Ø* *a-Ø-šá-n-q'əda-q'a-ma.*
3SGPSG-child-ABS 3SGA-3SGD-head-3SGE-cut-PST-NEG
'He did not decapitate his child.' (Charachidzé 1989: 431)

The possessor remains in the genitive (which retains the full form since the phonological conditioning of the possessive prefix has been removed) when there is an argument present to fill the specifier of AbsP.

(77) *a-q̄ąp̄'-Ø* *a-nt°a-n* *Ø-Ø-č''a-č'aw-q'a.*
the-branch-ABS the-door-GEN 3SGA-3SGD-mouth-fall-PST
'The branch fell in front of the door.' (Dumézil & Esenç 1971: 44)

(78) *á-nt°ą̇ʒa-Ø* *Ø-Ø-č''á-nə-wt°'ə-n* *a-č'ə́-Ø*
the-outer.door-ABS 3SGA-3SGD-mouth-3SGE-open-GER the-horse-ABS
a-g°ą́ɣa-n *Ø-Ø-g'ə́-n-ɣa-k'a-q'an.*
the-enclosure-GEN 3SGA-3SGD-in-3SGE-CAUS.PL-enter.PL-PST.PL
'(He) opened the outer door and made the horse enter the enclosure.' (Charachidzé 1988: 4)

3. Conclusion

This paper has sought to provide a straightforward description of the applicative and noun incorporation structures and the accompanying systems of agreement found in the Northwest Caucasian language Ubykh. Additionally, I have proposed an analysis of these data consistent with modern morphosyntactic theory. This analysis treats the four related constructions – dative verbal agreement, appliatives, incorporated postpositions, and incorporated nouns – as being derived from the same underlying structure. A structure which is not uncommon

crosslinguistically, but which is used in an unique way in Ubykh and the greater Northwest Caucasian family.

High applicatives in Ubykh are analyzed based on the theory proposed by Baker (1996) and subsequently extended by O'Herin (2001, 2002) and on the analysis of applicatives by Georgala et al. (2008). These proposals state that in languages where the applicative morpheme appears as a verbal prefix, this prefix may be regarded as the result of adposition incorporation. The applicative PP is generated in the specifier of a high applicative phrase (ApplHP) which is located between the νP and VP. The applicative P° head raises within the PP and merges with an agreement head (Agr°) and the complex form [Agr° + P°] incorporates into the verbal complex at ν° after the verb has raised and merged with the agreement located in ν°.

Low applicatives, on the other hand, are generated below the VP in a low applicative phrase (ApplLP). The applied object is generated in the specifier of this phrase while the corresponding agreement is found in the ApplL°. Similar to the high applicative, the low applicative agreement raises and merges with the verbal complex in ν°. This low applicative construction is responsible for intraverbal dative agreement in the absence of a governing incorporated postposition or noun.

The incorporation of postpositions into the verbal complex is similar in form to that of high applicatives. The PP in this construction, rather than being generated in the specifier of an ApplHP, occurs below the VP. The PP raises and is adjoined to the νP. Once the verb raises and joins with ν°, the [Agr° + P°] excorporates from the adjoined phrase and merges with the verbal complex in ν°. The incorporated postposition may be doubled by using a relational noun with the same form as the P° standing as the complement to P°. The incorporation of the postposition and agreement proceeds without alteration.

Ubykh noun incorporation relies on a construction similar to that of postposition incorporation with an external doubled postposition. The incorporated nominal begins as the object in a genitive construction which is itself the object of a null postposition (or an overt postposition such as the translocative). The Agr° dominating the PP agrees with the possessor of the nominal, which is most often a 3rd singular pro. This pro (or an overt nominal) raises to the specifier of AgrP in order to be coindexed with the agreement in Agr°. This nominal object raises and incorporates into the P°, and the [N° + P°] complex raises to merge with Agr°. As with the incorporated postpositions, the [Agr° + [N° + P°]] merges with the verb in ν°.

References

Allen, Barbara; Gardiner, Donna & Frantz, Donald (1984) Noun Incorporation in Southern Tiwa. *International Journal of American Linguistics* 50: 292–311.

Alparslan, Orhan & Dumézil, Georges (1964) Le parler besney (tcherkesse oriental) de Zennun köyü (Çorum, Turquie), II: Textes Folkloriques. *Journal Asiatique* 252: 327–364.

Anderson, Arthur J.O. (1973) *Rules of the Aztec Language*. Salt Lake City: University of Utah Press.

Baker, Mark (1996) *The Polysynthesis Parameter*. New York: Oxford University Press.

Bittner, Maria & Hale, Ken (1996) The structural determination of case and agreement. *Linguistic Inquiry* 27: 1–68.

Charachidzé, Georges (1988) Derniers textes oubykhs de Hacı Osman köyü avec version abkhaze. *Revue des Études géorgiennes et caucasiennes* 4: 1–20.

—— (1989) Ubykh. In B. George Hewitt (ed.), *The Indigenous Languages of the Caucasus. Vol. 2: The North West Caucasian Languages*, pp. 357–459. Delmar, NY: Caravan Books.

Colarusso, John (1988) *The Northwest Caucasian Languages: A Phonological Survey*. New York: Garland Publishing.

—— (1992) How many consonants does Ubykh have? In B. George Hewitt (ed.), *Caucasian Perspectives*, pp. 145–155. Munich: Lincom Europa.

Dirr, Adolf (1928) Die Sprache der Ubychen. *Caucasica* 4: 5–144.

—— (1929) Die Sprache der Ubychen. *Caucasica* 5: 1–54.

Dumézil, Georges (1931) *La langue des oubykhs*. Paris: Champion.

—— (1955) Les gutturales de l'oubykh. *Bulletin de la Société de Linguistique de Paris* 51: 176–180.

—— (1957) Le vocalisme de l'oubykh. *Bulletin de la Société de Linguistique de Paris* 53: 196–201.

—— (1961) Chah-Meymun: Texte oubykh, traduit et commenté. *Orientalia Suecana* 10: 41–80.

—— (1965) *Documents anatoliens sur les langues et les traditions du Caucase III: Nouvelles études oubykh*. Paris: Institut d'Ethnologie.

—— (1967) *Documents anatoliens sur les langues et traditions du Caucase V: Études Abkhaz*. Paris: Adrien-Maisonneuve.

—— (1975) *Le verbe oubykh: Études descriptives et comparatives*. Paris: Imprimerie nationale.

Dumézil, Georges & Esenç, Tevfik (1971) Structure des racines verbaux de l'Oubykh. *Bedi Kartlisa* 28: 24–52.

—— (1978) Textes oubykhs. *Studia Caucasica* 4: 16–103.

Dumézil, Georges & Namitok, Aytek (1954) Le système des sons de l'oubykh. *Bulletin de la Société de Linguistique de Paris* 50: 162–189.

Foley, William A. (1991) *The Yimas Language of New Guinea*. Stanford: Stanford University Press.

van Geenhoven, Veerle (2002) Raised Possessors and Noun Incorporation in West Greenlandic. *Natural Language & Linguistic Theory* 20: 759–821.

Genko, Anatolii (1928) O jazyke ubyxov [On the language of the Ubykh]. *Izvestija Akademii Nauk SSSR*: 227–242.

Georgala, Effi; Waltraud Paul & Whitman, John (2008) Expletive and Thematic Applicatives. In Charles B. Chang & Hannah J. Haynie (eds.), *Proceedings of the 26th West Coast Conference on Formal Linguistics (WCCFL 26)*, pp. 181–189. Somerville, MA: Cascadilla Proceedings Project.

Gerdts, Donna (1998) Incorporation. In Andrew Spencer & Arnold M. Zwicky (eds.), *The Handbook of Morphology*, pp. 84–100. Oxford: Blackwell Publishers Ltd.

Gronemeyer, Claire (1996) Noun Incorporaion in Hopi. *Lund University Department of Linguistics Working Papers* 45: 25–44.

Hewitt, B. George (1986) The Labialised Sibilants of Ubykh (North West Caucasian). *Revue des etudes georgiennes et caucasiennes* 2: 21–30.

—— (1989) Abkhaz. In B. George Hewitt (ed.), *The Indigenous Languages of the Caucasus. Vol. 2: The North West Caucasian Languages*, pp. 39–88. Delmar, NY: Caravan Books.

—— (2005) North West Caucasian. *Lingua* 115: 91–145.

Jeong, Youngmi (2006) *The Landscape of Applicatives*. Ph.D. dissertation. University of Maryland, College Park.

Kumakhov, Mukhadin (1998) Ubyxckii yazyk [The Ubykh Language]. *Yazyki mira: Kavkazskie yazyki*. Moscow: Academia.

Lander, Yury (2010) Relativization in Shapsug Adyghe. *Rice Working Papers in Linguistics* 2: 75–91. <http://www.ruf.rice.edu/~rls/vol.2/05%20RWPL2_Lander.pdf>

Leroy, Christine & Paris, Catherine (1974) Étude articulatoire de quelques sons de l'oubykh d'après film aux rayons X. *Bulletin de la Société de Linguistique de Paris* 69: 255–286.

Malouf, Robert (1999) West Greenlandic Noun Incorporation in a Monohierarchical Theory of Grammar. In Gert Webelhuth; Andreas Kathol & Jean-Pierre Koenig (eds.), *Lexical and Constructional Aspects of Linguistic Explanation*, pp. 47–62. Stanford, CA: CSLI Publications.

von Mészáros, Julius (1934) *Die Päkhy-Sprache*. Chicago: University of Chicago Press.

O'Herin, Brian (2001) Abaza Applicatives. *Language* 77: 477–493.

—— (2002) *Case and Agreement in Abaza*. Dallas, TX: SIL International.

Özsoy, Sumru (1988) Relative Clause Construction in Oubykh. *Proceedings of the IVth Colloquim on Caucasology*. Paris: Centre National de la Recherche Scientifique.

Pylkkänen, Liina (2002) *Introducing Arguments*. Ph.D. dissertation. Cambridge, MA: Massachusetts Institute of Technology.

Shibatani, Masayoshi (1990) *The Languages of Japan*. Cambridge: Cambridge University Press.

Vogt, Hans (1963) *Dictionnaire de la langue oubykh avec introduction phonologique, index français-oubykh, textes oubykhs*. Oslo: Universitetsforlaget.

Abbreviations

1 = 1st Person
2 = 2nd Person
3 = 3rd Person
A = Absolutive
D = Dative
E = Ergative
F = Feminine
M = Masculine
N = Neuter
O = Object
P = Possessive
S = Subject

ABL = Ablative
ABS = Absolutive
ACC = Accusative
ADV = Adverbial
AOR = Aorist
APPL = Applicative
BAS = Basic
BEN = Benefactive
CAUS = Causative
CL = Noun Class
COM = Comitative
DAT = Dative
DECL = Declarative
DYN = Dynamic
FACT = Factive
FUT = Future
FUT1 = Future 1
FUT2 = Future 2
FV = Final Vowel
GEN = Genitive
GER = Gerundive
HAB = Habitual
IMP = Imperative
IND = Indicative
INDF = Indefinite
INST = Instrumental
ITER = Iterative
LNK = Linker
LOC = Locative
MAL = Malefactive
NEG = Negation
NFIN = Non-finite
NMLZ = Nominalizer
NOM = Nominative
NPST = Non-past
NSF = Noun Suffix
OM = Object Marker
PASS = Passive
PERL = Perlative
PFCT = Perfect
PL = Plural
PLUP = Pluperfect
PRS = Present
PST = Past
PTCP = Participle
PV = Preverb
PUNC = Punctual
Q = Question
REL = Relative
SEQ = Sequential
SG = Singular
SM = Subject Marker
STAT = Stative
TR = Transitive
TRLOC = Translocative

Contact information:

Brian Fell
(No Affiliation)
e-mail: bfell1618(at)gmail(dot)com

Mikko Heikkilä

Kaleva and his Sons from *Kalanti* – On the Etymology of Certain Names in Finnic Mythology

Abstract

This article deals with the origin of the names of certain mythical figures in Finnic pre-Christian mythology. The purpose of the article is to propose an etymology and a dating for the mythical names *Kaleva*, *Kalevanpoika*, *Tiera*, *Niera* and *Liera*, as well as the etymologically related Finnish place-names *Kalanti* and *Torre*. *Kaleva* and *Kalevanpoika*s are ancient mythical powerful giants. There have been many attempts to etymologize the name *Kaleva*, none of which turns out to be both phonologically and semantically satisfactory after critical scrutiny. *Kaleva* belongs to the pre-Christian Finnic worldview, so the word must have existed as early as in the Iron Age. The word's distribution is rather limited. I propose that *Kaleva* is a Proto-Scandinavian loanword whose origin is the Proto-Scandinavian proto-form of the Old Norse sea-god giant *Hlér*. The phonemic Proto-Germanic and Proto-Scandinavian reconstruction of *Hlér* is **χlewaz*, but the phonetic reconstruction of **χlewaz* is **χ^{a}lewaz*, whence *Kaleva* has been borrowed. The sound substitutions are regular. There is plenty of evidence of the existence of and parallels for an epenthetic vowel [a] in consonant clusters /C^{a}C/ both in runic inscriptions and in Germanic loanwords in Finnic. On the basis of the word's sound shape, the date of borrowing can be estimated. *Kaleva* is likely a Roman Iron Age loanword. In conclusion, the names of many mythical beings of the Finnic pre-Christian mythology turn out to be Iron Age loanwords from Proto-Scandinavian.

1. Introduction

Etymological research on the Finnic and Sami lexicon and toponymy has been lively and fruitful over the past few decades. However, remarkably little research has been conducted on the etymology of the names occurring in Finnic and Sami mythology. The most important contributions to this field of Fennistics and Lappology include Haavio (1967), dealing with the origins of Finnish proper names of mythical beings, Turunen (1979) which lists and describes words with a mythical referent, and the comprehensive

Finnish etymological dictionary *Suomen sanojen alkuperä* (hereafter SSA) (1992–2000), which also etymologizes lexemes with a mythical referent. A folkloristic reference book on Finnish mythology is Pentikäinen (1989). Sami mythology is discussed in Pentikäinen (1995). Many names in Finnic mythology, e.g. *Kaleva* and *Ahti*, still lack a satisfactory etymology. However, real world proper names have been studied more thoroughly (see USN 1988; Huldén 2001; SPNK 2007). This article, therefore, deals with the origin of the names of certain mythical figures in Finnic pre-Christian mythology from a linguistic and etymological point of view, also considering folkloristics and cultural history. The primary purpose of my article is to propose an etymology and a dating for the names *Kaleva* and *Kalevanpoika*, as well as the south-western Finnish place-name *Kalanti*, which is situated in one of the two most central settled areas in Finland in the Iron Age. North Germanic pre-Christian mythology in Scandinavia in the Late Iron Age turns out to be of great importance in the search for the origin of *Kaleva* and other related mythical figures in old Finnic folklore, such as *Tiera*, *Niera* and *Liera*.

In the Finnic oral rune tradition, *Kaleva* is a mythical being who is seen as an ancient powerful giant and a heroic forefather (Turunen 1979: 88–89; Saagpakk 1982: 235; Järv 1987: 32, Pentikäinen 1989: 233). The word is both a proper name and a common noun. The Kalevala-inspired Finnish male names *Kaleva* and the Estonian influenced *Kalevi* were introduced in the latter half of the 19th century (USN 1988: 94). The Finnish word *Kaleva* has a known cognate only in the Karelian and Estonian languages, which means that the word seems to have a rather limited distribution in the Finnic languages. The compound *Kalevanpoika* [‘Kaleva’s son’] ‘strong giant’ also occurs in Ingria, including the Votic region of Ingria. The Finnish and Karelian common noun *kaleva* has been attested at least in the following meanings: ‘giant’, ‘tall strong man’, ‘ill-mannered (and notorious) person (who practises witchcraft)’, ‘introverted taciturn old-fashioned person’, ‘arrogant person’, ‘tall tree’ and ‘term of abuse, used in name-calling’ (cf. the Finnish words *hiisi*, originally ‘pagan cemetery’, ‘place of worship’, after the introduction of Christianity ‘spiritual being, the deuce’, *hitto* ‘the deuce’ and the Proto-Sami loanword *pisa* ‘the deuce’ and its Finnish cognate *pyhä* ‘holy’) (SMS 1999: 24–25, 27–28). In old folk stories, *Kaleva* is a giant of the ancient times, or “the general of all giants” as Christfrid Ganander (2003: 48–49), the writer of *Mythologia fennica*, formulates *Kaleva*’s nature (Huurre 2003: 236–237).

The Estonian word *kalev* has two meanings, ‘giant’ and ‘broadcloth, woollen cloth’, the latter of which has most likely arisen elliptically from the phrases *kalevirohi* ‘broadcloth’ [“kalev’s grass”], *kalevipoja hiused* ‘id.’ [“kalev’s son’s hair”] and *kalevid* ‘id.’ [“kalevs”], where *kalev* means ‘giant’ (Turunen 1979: 88). The unchangeable word-final vowel *kalev_* (nom. sg.) : *kalev**i*** (gen. sg.) < OEst **kalev**i*** : **kalev**i**n* suggests that **kalevi* is either a relatively young word in the language or it has emerged through denominal derivation **kaleva-j* > **kalevi* : **kalevin* > *kalev* : *kalevi*. As in Estonian, the Finnish and Karelian word *kaleva* occurs (read: has been preserved until historical times) in many established phrases and compounds such as *kalevanpoika* ‘giant’ (cf. the Estonian cognate *kalevipoeg* ‘kalev’s son’), *puun kaleva* ‘exceptionally tall tree’, *kalevantuli* ‘kaleva’s fire’, *Kalevan tähti* ‘Orion’ (literally ‘Kaleva’s star’), *Kalevan miekka* ‘Kaleva’s sword’ and *Kalevanpojan jäljet* ‘Kaleva’s son’s footprints’. The word *Kalevatar* [‘Kaleva’s daughter’] is attested as well. (Haavio 1967: 268; Turunen 1979: 88–89; Pentikäinen 1989: 155; SMS 1999: 24–25, 27–28.) The attested parallel forms of the Finnish word *kaleva* are *kalevi*, *kalevo*, *kalehva*, *kalehvo*, *kalehvi*, *kalevas*, *Kalava*, *Kalevas*, *Kalevainen* and *Kalavainen* (Turunen 1979: 88–90; SMS 1999: 24–25, 27–28; Huurre 2003: 236–237; Ganander 2003: 48–49). The forms *kalev**i*** and *kalev**o*** prove that the stem form *kalev**a*** really has been augmented with a derivational suffix **-j* (cf. the Estonian word). The *h*-forms *kale**h**va*, *kale**h**vo*, *kale**h**vi* have arisen through a metathesis as follows: *kalevas* : **kalevahan* > **kalehvaan* > *kalehva* (cf. *imeh* > *ihme* ‘wonder’, *Kainus* : *Kainuhun* > *Kaihnuun*, *taivas* : *taivahan* ‘sky, heaven’ > *taihvaan*, *kirves* : *kirvehen* ‘axe’ > *kirhveen*). An implication is that *kaleva**s*** is likely this word’s original stem form.

Many heroes in Finnic folk poetry bear the epithet *Kalevanpoika* at least in some versions of rune songs, including the old and steadfast *Väinämöinen*, the primary divine shaman and hero of Kalevala (Siikala 1987: 18). *Kalevanpoika*s are giants (either constructive or destructive ones) and *kalevanpoika* is a common epithet for a hero in (south-western) Finnish folklore (Huurre 2003: 236, 253). The name *Kaleva* occurs in many (south-western) Finnish toponyms, e.g. *Kalevanharju* ‘Kaleva’s ridge’, *Kalevankangas* ‘Kaleva’s heath’, *Kalevanniitty* ‘Kaleva’s meadow’, *Kalevanmäki* ‘Kaleva’s hill’, *Kalevanhauta* ‘Kaleva’s grave’, *Kalevankallio* ‘Kaleva’s rock’, *Kalevankivi* ‘Kaleva’s stone’, *Kalevanpojankallio* ‘Kaleva’s son’s rock’, *Kalevanpoikainkivi* ‘Kaleva’s son’s stone, i.e. erratic boulder’ etc. (cf. *Hiidenkivi* ‘The devil’s stone’,

Äijänkivi ‘Old man’s stone’ and *Ukonkivi* ‘Old man’s stone’, all terms for erratic boulders) (Names Archive; MapSite; see Appendix). A general observation regarding these toponyms is that they are names of objects of exotic shape and/or size in nature, such as erratic boulders, rocks and crags, whose existence has been explained as the work of giants (Huurre 2003: 250). The *Kalevanpoika* tradition is concentrated in the south-western parts of Finland – the so-called Varsinais-Suomi [‘Finland Proper’] – and there especially in the Laitila-Kalanti region [= the so-called Vakka-Suomi] (Koski 1967: 111; Järv 1987: 32; Siikala 1987: 16; Vahtola 1987: 40; Huurre 2003: 247, 251–252, 434–437). The Laitila-Kalanti region also has a greater density of place-names which refer to Kalevalaic mythic heroes (Anttonen 2003: 226). I will return to the origin of the place-name *Kalanti* later in this article.

2. The earliest historical records of the name *Kaleva*

The father of the Finnish written language Mikael Agricola (ca. 1510–1557) mentions *Caleuanpojat* in his list of Tavastian pre-Christian gods in 1551 (SSA 2001 s.v. *Kaleva*). Finnish farms (and families) called *Kaleva* are found in historical documents from the 15th century onwards. The capital of Estonia, Tallinn, and the homophonic name of a heroic giant is mentioned in Old Russian chronicles and folk tales (bylinas) in the form *Kolyvan* from the year 1223 CE onwards. The town name *Kolyvan* is probably to be read as **Kalevan*(*linna*) ‘Kaleva’s (burg)’ (USN 1988: 94; Huurre 2003: 250).[1] The final vowel in the name *Kolyv**a**n* points to that the Estonian *kalev* goes back to an earlier form **kaleva*, identical with the Finno-Karelian stem form. An even earlier attestation of the name *Kaleva* might be found in a source from a distant country. The Arab geographer Muhammad Al-Idrisi (ca. 1100–1165) from Sicily began the work of

[1] Cf. the name of the city of *Tallinn* (Fin *Tallinna*) derives from **Taninlinna* [sic] ‘the burg of the Danes’ > **Tanilinn* > *Tallinn* (cf. Pajusalu, Hennoste, Niit, Päll & Viikberg 2002: 242). The usually given reconstruction **Taani linn* (e.g. Turunen 1979: 328; Zetterberg 2007: 55, 67–68) is anachronistic for two reasons: Firstly, the Scandinavian etymon *dan* ‘Dane’ of the Estonian word *taani* ‘Dane, Danish’ still had a short vowel at the time of the Danish conquest of Estonia in 1219 CE. Secondly, the reduction of unstressed syllables was just beginning in Old Estonian. Consider *Somelind**e*** (1212 CE), *Lind**a**nise* (1219), *Soo**n**tagana* (1211), *Hergæ**n**pæ* (1241), *Paiu**m**pe* (1241) and *Ude**n**küll* (1323). (Haugen 1976: 258–259; SSA s.v. *linna*; Pajusalu et al. 2002: 235, 271, 278; Lättiläinen 2003: 120, 133, 187.)

drawing a world map with a commentary in 1140 at the court of and by the commission of the Norman King Roger II of Sicily (1095–1154). Al-Idrisi published his famous world map and its commentary (Tabula Rogeriana) in 1154. On the map, a small fortress town spelled *qlwry* is mentioned as situated in the country of *ʹslʹndh* /aslandh/ 'Estonia' (cf. OGut *Aistland* 'Estonia'). (Tuulio 1936: 34–40; USN 1988: 94; Grünthal 1997: 220, 237; Tarvel 2004: 1–9; Salo 2008: 294–298.) Vowels are not written in Arabic so they have to be added, but the first three consonants *qlw* in *qlwry* fit with the ones in the Finnish–Estonian name *Kalev*(*a*). When compared with the Old Russian name *Kolyvan* for Tallinn, it is possible that the burg and town of *qlwry* really refers to Tallinn and contains the earliest known attestation of *Kaleva* less than a hundred years after the end of the Viking Age, which traditionally has been dated as ranging from 793 CE to 1066 CE or from 800 to 1050 (Hentilä, Krötzl & Pulma 2002: 14; Palm 2010: 459–461). However, a more realistic starting point for the Nordic Viking Age would be ca. 760 (Salo 2008: 223; Tarkiainen 2008: 38; Palm 2010: 459). I wonder if *qlwry* actually stands for an Old Estonian compound **Kaleveeri* < **Kalevinveeri* "Kaleva's hill(side)" (cf. the nearby place-names *Randvere*, *Kallavere*, *Kostivere*) (see SSA s.v. *vieri*; Pajusalu et al. 2002: 272–274).

3. Previously posited etymologies of *Kaleva* and their shortcomings

There have been many attempts to etymologize the nationally important name *Kaleva*, none of which turns out to be both phonologically and semantically satisfactory after critical scrutiny, which means that the origin of *Kaleva* has so far remained unknown (Huurre 2003: 249; EES 2012: 121). I will first go through and briefly comment on previous etymologies for *Kaleva* before I present and motivate my own etymology.

The first serious attempt to etymologize *Kaleva* was made by the compiler of *Kalevala* Elias Lönnrot (1802–1884) himself. Lönnrot derived *Kaleva* from the Russian word *golová* [gələvá] 'chief, head' (Turunen 1979: 88). Lönnrot undoubtedly had an eye for etymology, but he did not have access to the results of modern historical linguistics. Lönnrot's etymology is impossible because *Kaleva* clearly belongs to the pre-Christian Finnic worldview, i.e. prehistoric times, and no loanword could have been borrowed from medieval or New Age Russian to prehistoric Finnic. Neither the Proto-Slavic proto-form **golvá* nor the Proto-Balto-Slavic proto-form **golHváH* (> Lith *galvà* 'head') of the Russian word

golová comes into question for phonological reasons (Fraenkel 1962: 131; Kortlandt 1983: 7).

In another etymology, *Kaleva* has been derived from the Old Norse name *kylfingjar* mentioned in Egil's Saga (Turunen 1979: 88; Järv 1987: 31). This etymology is phonologically impossible, as the vowels do not fit. *Kaleva* cannot be derived from either *kylfingjar* or its Proto-Scandinavian proto-form **kulƀingōR* (→ Rus *kolbjagi*).

According to Julius Krohn, J. J. Mikkola and Hannes Pukki, *Kaleva* originally meant '(red) broadcloth' because the Estonian word *kalev* means '(red) broadcloth' besides the meaning 'giant' (Turunen 1979: 88). I find this etymology highly improbable because it contains a major semantic error. The Estonian word *kalev* '(red) broadcloth' has the parallel forms *kalevirohi* and *kalevipoja hiused* (Turunen 1979: 88). The short form *kalev* likely emerged elliptically from the compounds. In other words, *kalev* did not originally mean 'broadcloth' but 'giant', because *kalevirohi* – literally 'kalev's (= giant's) grass' – was a metaphorical expression for broadcloth, comparable with other metaphorical phrases/compounds such as *äijänkivi* 'erratic boulder', and the Finnish word *äijä* definitely does not mean any kind of stone but 'old man', although the compound *äijänkivi* has a completely different meaning. If the word *kalev* had originally meant 'broadcloth', the expressions *kalevirohi* "broadcloth's grass" and *kalevipoja hiused* "broadcloth's son's hair" would be semantically completely absurd. After having shortened elliptically and having consequently acquired the etymologically secondary meaning 'red broadcloth', the word *kalev* started to be used in new compounds such as *kalevitõbi* 'scarlet fever', literally "red broadcloth's illness".

Paul Ariste has proposed that the Estonian noun *kalev* is a derivative from the adjective *kale* 'hard, severe' that has a known cognate in Finnish (*kalea* 'cool, hard, slippery'), Ingrian (*kale* 'cool') and Karelian (*kalie* 'cold weather in the autumn'). This etymology has been favoured by Ants Järv (1987: 31). However, I find it rather unlikely because of morphological and semantic problems. The Finnic suffix *-va*/*-vä* and *-pa*/*-pä*, e.g. *lihava* 'fat', *kätevä* 'handy', *väkevä* 'strong', *juopa* 'small river' and *syöpä* 'cancer' (< EPF **-pa*/**-pä* and **-ŋa*/**-ŋä*) undoubtedly derives adjectives from nouns and present participles (which often develop into adjectives) from verbs, but not adjectives from other adjectives (cf. Fin *nop**ea*** 'quick, rapid' ***nop**eva***) (Sammallahti 1998: 91; Lehtinen 2007: 125). The primary meaning of the Finnic word *kalea* (< LPF **kaleða*) is 'cool, chilly, cold', whereas 'hard' is secondary. The Finnic word itself is a Germanic

loanword. (SSA 2001 s.v. *kalea*; Aikio 2006: 29, 48). Ariste motivates his etymology with the Estonian adjective *kalevine* which according to him means ‘strong’, but if *kalev* were an adjective (indeed it is a noun meaning ‘giant; broadcloth’!) and it already meant ‘strong’, so why should it have been augmented with a suffix in order to acquire the adjectival meaning ‘strong’? Moreover, the adjective *kalevine* means ‘gigantic’ according to Paul F. Saagpakk (1982: 235). So instead, I would claim that the Estonian word *kalevine* (< OEst **kalev-inen*) is an adjectival derivative of *kalev* ‘giant’ and the meaning is explained by the fact that giants were thought to be very strong beings.

August Ahlqvist and Emil Nestor Setälä thought that *kaleva* stems from the Baltic languages. According to Ahlqvist and Setälä, a potential original would be the Lithuanian word *kálvis* ‘smith’ (Turunen 1979: 88). This etymology has been favoured as most probable by Aimo Turunen (1979: 88), and Ants Järv (1987: 31) does not dislike it either. I do not consider it plausible because of phonological and semantic problems (cf. Junttila 2005: 55). Firstly, the primary meanings of *kaleva* do not include ‘smith’ (Junttila 2005: 55). Secondly, the Proto-Baltic word (*)*kalvis* would not have rendered *kaleva* in either Proto-Finnic or Finnic, but it would have become PFin ***kalviš* > Fin ***kalve* or (P)Fin ***kalvis* depending on the date of the borrowing (Heikkilä forthcoming). Furthermore, *Ilmarinen* (< PF (*)*ilma* < Finno-Ugric **i̯lma* ‘air, weather’) was clearly the primary (if not the only) god of the sky and of iron-working in the Finnic pre-Christian religion instead of *Kaleva* (Turunen 1979: 65–66).

4. Has *Kaleva* a cognate in Sami?

In 1918 the famous Finnish linguist and folklorist Toivo Ilmari Itkonen connected *Kaleva* with the North Sami word *gállagas-dolla* ‘phosphorescence of the sea’ and the Lule Sami word *kāllaka jussa* ‘suddenly descending thick fog over a lake’ (Turunen 1979: 88).[2] The uninflected form is *kāllak*. The linguists Uno Harva and Y. H. Toivonen accepted Itkonen’s etymology and developed it by deriving both the Finnic and the Sami word from a common proto-form **kaleɣa*. Their etymology has been accepted (with some reservations) in SSA (s.v. *Kaleva*). In my

[2] The Lule Sami word *jussa* literally means ‘fart of a wild animal’ (Grundström 1946: 123). Thus *kāllaka jussa* is “old man’s fart” (cf. *kalevirohi* p. 98).

view, the supposed etymological connection between *Kaleva* and *kāllak* is impossible because of problematic sound correspondences. These words cannot be derived from a common Proto-Finno-Sami proto-form. The vowel combination /á–a/ in the Sami word *kāllak* is unetymological, which indicates that the lexeme has entered the language after the so-called great Sami vowel shift (on which see Aikio 2006: 13, 44). The Sami word *kāllak* had the sound shape **kālle̮kke̮* in Proto-Sami. The unetymological vowel combination indicates that the Sami word cannot even be traced back to the Early Proto-Sami language, not to mention about the Finno-Sami proto-language, i.e. Early Proto-Finnic (Aikio 2006: 44). If the Finnic word *Kaleva* dated back to the Finno-Sami proto-language or even further back in history, which I do not believe, the Finnic proto-form would have been **kalį-pa* or **kalį-ŋa*. As we can see, the Proto-Finnic and Proto-Sami forms do not fit together. I would therefore propose that this Sami word is a derivative from the common Sami lexeme *gállis* 'old honoured man', which had the sound shape **kāllās* in Proto-Sami. This word itself is a loanword from the Proto-Scandinavian noun **karlaz* '(old) man' (cf. the Greek cognate *gérōn* 'old man'), whence Swe *karl* 'man' and the Germanic male name *Karl* (> Swe/Fin *Kalle*) (Hellquist 2008: 447; Álgu database s.v. *gállis*).[3] A similar change of suffix is demonstrated by the Sami noun *vuotta* 'shoelace' < EPSa **vant-**ik*** ← PGerm **wand**uz*** (> Got *wandus* 'whip, twig') → PF **vant**iš*** > Fin *vanne* 'hoop' (SSA 2001 s.v. *vanne*). SSA (2001 s.v. *Kaleva*) states that *kāllak* also occurs in certain names of stars, such as *Boaris-Gál´lá* 'Sirius' (literally 'Old Man'), which have been folk-etymologically associated with *gállis* (Álgu database s.vv. *boaris*, *gállis*). However, I venture to say that *gállis* and *Gál´lá* belong not only folk-etymologically but also etymologically together since saN nom. *Gál´lá* : gen. *Gállá* 'Sirius' is an analogical formation from *gállis* : *gál´lá*, so actually 'Mythic Old Man' (cf. Germ *Mann* 'man' and *Männchen* 'elf', literally 'little man').

It has also been suggested that *Kaleva* might have been mentioned in the oldest surviving Old English poem called *Widsith* ("Wide-traveller"), where it is said that "Casere weold Creacum ond *Cælic* Finnum", which can be translated as "(Julius) Caesar ruled the Greeks and *Cælic* the Finns". Finns are mentioned three times in the poem. Widsith is a "thoroughly heathen poem" that tells about historical and mythical tribes, chiefs, kings

[3] The same word occurs in the name *Horagalles* '(South) Sami thunder-god' (cf. ON *Þórr* 'thunder-god') (Haavio 1967: 85; Pentikäinen 1995: 233–234; Sammallahti 1998: 36).

and heroes in (Northern) Europe in the early Migration Period (375–449 CE) [Germ Völkerwanderung] before the Germanic conquest of Britain beginning in 449 CE. It has been estimated that Widsith was probably composed in the 7th century in Mercia in Britain, that is to say, in the Merovingian Period, and the transcript was written down in the Viking Age ca. 1000, but the historical events mentioned in the poem took place much earlier. (Alexander 1977: 32–35, 38–42; Turunen 1979: 88; Huurre 2003: 249–250.) However, I would argue that the sound shape *Cælic* cannot be derived from *Kaleva* but might be derived from the Proto-Sami word **kālle̮k* (> SaLu *kāllak*) since the Sami root vowel was somewhat front (cf. SaI *sää̮mi* ~ ON ***sæ****msvein* 'Sami man') and, furthermore, a back root vowel /a/ before a second syllable front vowel /i/ developed into /æ/ in Pre-English already, in the so-called *i*-umlaut. The substitution of the English second syllable /i/ for the Proto-Sami second syllable /e̮/ is predictable since only four vowels, namely /i/, /æ/, /a/ and /u/, were phonotactically possible in non-root syllables in Old English. The geminate /l:/ in the Sami word **kālle̮k* rendered a single consonant in Old English because a geminate sound /l:/ did not occur after a long stressed vowel /æ:/. (Korhonen 1981: 109–114; ODEE 1982: passim; Nielsen 2000: 79; Wójcik 2001: 383; Antonsen 2002: 332.) The possibility that *Cælic* is a phonological blend of *Kaleva* and **kālle̮k* should, however, be left open. Presumably, facts about the Finns and the Sami have repeatedly been intermingled in old foreign sources (see Julku 1985: 85, 1986: 51; Linna et al. 1988: 165; Pentikäinen 1995: 165). As mentioned above, I derive the Late Proto-Sami **kālle̮k* from the common Sami noun which is *gállis* 'old honoured man' in North Sami. However, the Sami names *Gállá* 'Sirius' and *Gállábárdnit* 'Gállá's sons' may well have been semantically influenced by the Finnish *Kalevan tähti* 'Orion' and *Kalevanpojat* 'Kaleva's sons' (see further Pentikäinen 1995: 137). In conclusion, as far as I can see, the answer to the question posed in the heading above is negative.

5. The origin of *Kaleva* – a Proto-Scandinavian loanword?

As mentioned above, the mythical being *Kaleva* belongs to the pre-Christian Finnic worldview, so the word must have existed in the Iron

Age.[4] However, *Kaleva*'s distribution is rather limited. The word has been attested only in a handful of Finnic languages, which implies that it most likely does not stem from "time immemorial" and does not go back to the Uralic proto-language. *Kaleva* can hardly be older than the Iron Age (ca. 500 BCE–1200 CE). This is the case especially if and when the Sami word *kāllak* is not a cognate of *Kaleva*. It is quite natural to seek a loan original for *Kaleva* among the names of the figures in Old Scandinavian mythology since we know that a great deal of the Finnic as well as the Sami vocabulary has been borrowed from Early Proto-Germanic, Proto-Germanic and Proto-Scandinavian, one of the descendants of Proto-Germanic (see LÄGLOS 1991–2012).[5] Indeed, in Old Scandinavian mythology, I have found a very potential etymon which in my opinion is phonologically, semantically and (cultural) historically impeccable and which I will write next about.

I propose that *Kaleva* 'giant' is a Proto-Scandinavian loanword whose origin is the Proto-Scandinavian proto-form of the Old Norse sea-god giant *Hlér*. In Old Norse mythology *Hlér* is the same as *Ægir*, who is described as being "havets jätte, bosatt i en hall under Læsø i Kattegatt" ['the giant of the sea living in a hall under the island of Læsø[6] (< ON *Hlésey* 'Hlér's island') in Kattegat (off the Danish coast)'] (Ohlmarks 1983: 153, 406).[7]

[4] Old Finnish folk stories tell how *Kalevanpoika*s did not like churches and threw big stones at them (although never hitting the target) when the first churches were being built in Finland (Huurre 2003: 244). These stories actually tell about the tension and conflict between the old and the new beliefs.

[5] In addition, several names of figures in the Finnic mythology have been proven to be of Germanic origin. I refer to such words as Kar *kaveh* : *kapehen* 'creature, mythological being/maiden' ~ SaN *guobas* 'witch' (< EPF **kapįš* ← PreGerm **skabis* > PGerm **skapiz* > Swe *skapelse* 'creature'), *menninkäinen* 'troll' (← PGerm **menþiŋgō* > Icel *minning* 'memory of the deceased'), *tur*(*i*)*sas* 'sea-monster, sea-god' (> the Finnic mythical figure *Iku-Turso* 'gigantic sea-monster') (← PGerm **þur*(*i*)*saz* > ON *þurs* 'evil giant', OHG *thur*(*i*)*s*, OE *þyrs*), *Runkoteivas* 'name of a god' (← EPGerm **Teiwaz* 'clear sky' > PGerm **Tīwaz* > ON *Týr* 'name of a Old Scandinavian god') and *halti*(*j*)*a* '(physical) holder; (mythical) holder of a place' (← PScand **halðijaz*, cf. Eng *holder*) (Haavio 1967: 109–110; Hofstra 1985: 331; SSA s.vv. *haltija*, *menninkäinen*, *tursas*; Aikio 2006: 11; Hellquist 2008: 1249). The words' sound shapes and distributions suggest that the cognate set *kave* ~ *guobas* is the oldest among these examples.

[6] *Læsø* is the largest island in Kattegat. The name-form *Ægir* also occurs in two hydronyms (Ohlmarks 1982: 406; Holtsmark 2001: 130).

[7] For more information about *Hlér* and the synonymous name *Ægir*, see Holtsmark (2001: 130–131) and Ohlmarks (1983: 153, 406).

The deity names *Hlér* and *Ægir* are personifications of the Old Norse common names *hlér* and *ægir* respectively, both of which mean 'sea'. The word *hlér* 'sea' is a euphemism, the original meaning being 'good conditions for sea-farers'. (NO 1993: 191, 511.) Since *Hlér* and the synonym *Ægir* are mentioned in two Old Norse sagas, let us now take a look at a couple of extracts, taken from *Fundinn Noregr* ['The Foundation of Norway'] and *Hversu Noregr byggðist* ['How Norway was inhabited'], both of which are included in the *Orkneyinga Saga* (or The History of the Earls of Orkney). The manuscripts are included in the Icelandic *Flateyjarbók*. The surviving manuscript was written between 1387 and 1394 CE, but the Orkneyinga Saga was probably composed in Iceland ca. 1230 and its author was probably the great Icelandic saga author Snorri Sturluson (1179–1241) himself. (Julku 1986: 70; Rowe 2000: 441–454.) Even if the sagas were recorded long after the conversion to Christianity – which officially took place in 1000 CE in Iceland – they must have existed in an oral form (the so-called skaldic poetry) long before because they most often tell about events that happened in the Viking Age before the introduction of Christianity, e.g. *Landnámabók* ['The Book of Settlement'] describes the time ca. 870–930 CE, and reflect a pre-Christian worldview (Häme 1991: 182, 185–186; Hentilä, Krötzl & Pulma 2002: 31; Palm 2010: 460). The extracts in question are:

> From *Fundinn Noregr* in Old Icelandic:
>
> *Fornjótr* hefir konunger heitit; hann réð fyrir því landi, er kallat er *Finnland* ok *Kvenland*; þat liggr fyrir austan hafsbotn þann, er gengr till móts við Gandvik; þat kǫllu vér Helsingjabotn; Fornjótr átti þrjá syni; hét einn *Hlér*, er vér kǫllum *Ægi*, annarr Logi, þriði Kári; hann var faðir Frosta, fǫður Snæs hins gamla. Hans sonr hét Þorri; hann átti tvá syni; hét annarr Nórr, en annan Górr, dóttir hans hét Gói. *Þorri var blótmaðr mikill; hann hafði blot á hverju ári at miðjum vetri; þat kǫlluðu þeir þorrablót*; af þi tók mánaðrinn heiti. (Malling 1860: 219; Julku 1986: 61.)
>
> English translation:
>
> There was a king named *Fornjot*[*r*]; he ruled over [that land] which [is] called *Finland* [*Proper*] and *Kvenland* [= *south-western Finland*]; that is to the east of that bight of the sea which goes northward to meet Gandvik [= the Kandalaksha Gulf]; that we call the Helsingbight [= the Gulf of Bothnia]. Fornjot[r] had three sons; one was named *Hler*, whom we [= Icelanders] call *Ægir*, the second Logi, the third Kari; he was the father of Frost, the father of Snow the old, his son's name was *Thorri*; he had two sons, one was named Norr and the other Gorr; his daughter's name was Goi. *Thorri was a great sacrificer, he had a sacrifice every*

> *year at midwinter; that they called Thorri's sacrifice*; from that the month took its name. (Dasent 1894; Pálsson & Edwards 1978: 23.)[8]
>
> From *Hversu Noregr byggðist* in Old Icelandic:
>
> *Fornjótr hét maðr. Hann átti þrjá sonu*; *var einn Hlér*, annarr Logi, þriði Kári. Hann *réð* fyrir vindum, en Logi fyrir eldi, *Hlér fyrir sjó*. Kári var faðir Jökuls, föður Snæs konungs, en börn Snæs konungs váru þau *Þorri*, Fönn, Drífa ok Mjöll. *Þorri var konungr ágætr. Hann réð fyrir Gotlandi, Kænlandi ok Finnlandi. Hann blótuðu Kænir til þess*, at snjóva gerði ok væri skíðfæri gott. Þat er ár þeira. Þat blót skyldi vera at miðjum vetri, ok var þaðan af kallaðr Þorra mánaðr. Þorri konungr átti þrjú börn. Synir hans hétu Nórr ok Górr, en Góa dóttir. (The Northern Way 2009)
>
> English translation:
>
> *There was a man called Fornjotr. He had three sons*; *one was Hler*, another Logi, the third Kari; he ruled over winds, but Logi over fire, *Hler* [*ruled*] *over the seas*. Kari was the father of Jökull, the father of [K]ing Snow. But the children of [K]ing Snow were these: *Thorri*, Fönn, Drifa and Mjol. *Thorri was a noble king*; *he ruled over Gotland, Kvenland and Finland. To him Kvens sacrificed* that it might be snowy, and that there might be good going on snow-shoon. That was their harvest.[9] (Dasent 1894)

From these extracts we see that pre-Christian Scandinavians believed that the seas were ruled by a sea-god whose name was either *Hlér* or *Ægir*, both meaning 'sea'. Furthermore, we can see that the archaeologically and linguistically indisputable lively contacts between Scandinavia (including Gotland) and Finland (on which see Salo 2003a; Lehtinen 2007: 232) are mentioned in two historical documents. It is also worth noting that Finland Proper and Kvenland must originally have lain close to each other, because they were understood as one land with two names in the eyes of the ancient Scandinavians. This is in accordance with the conception that Kvenland originally lay in the present-day south-western Finland slightly to the north of Finland Proper (Vilkuna 1969: passim; Nuutinen 1989: 21–22; Koivulehto 1995: 93–94; Salo 2003a: 90, 2003b: 36, 59–60, 2008: 158, 161; see also Julku 1986: 36–37; Lehtinen 2007: 256). However, this formulation must preserve memories from a fairly distant past. Namely, when the saga was written down by Snorri Sturluson in ca. 1230, the name

[8] The comments in the square brackets are the present writer's amendments and explanations.

[9] Good skiing conditions were needed on voyages from Kvenland to Lapland and back (Salo 2003b: 34–36).

Kvenland had centuries before begun to be used about a far more northern region than south-western Finland, first about southern Ostrobothnia and later about northern Ostrobothnia (see Julku 1986: 38, 52–57; Koivulehto 1995: 94–95).

Let us now return to the direct traces of *Kaleva*. The phonemic Proto-Germanic and Proto-Scandinavian reconstruction of *Hlér* is **χlewaz*, from which *Kaleva* cannot be phonologically successfully derived, but the situation is changed decisively by evidence from runic inscriptions and loanword study which prove that there was an epenthetic (svarabhakti) vowel [a] in Proto-Scandinavian consonant clusters consisting of one of the sounds /l, m, n, r/ (= sonorants (= R)) and another consonant, that is to say the sequence /CR/ was realized as [C^{a}R] in speech (Antonsen 1975: 15–16, 34–36, 44–45, 53, 56–57, 64, 83–86; Antonsen 2002: 86–87, 89; Bjorvand & Lindeman 2007: 634, 688–689; Runtextdatabas 2008: passim). This vowel epenthesis was the second in sequence of the several vowel epentheses in the history of the North Germanic languages, and a similar vowel epenthesis existed in the attested medieval West Germanic languages as well (Wessén 1968: 59–61; Liberman 1992: 195–196, 205; Ringe 2006: 81, 152).[10,11] For instance, the phonemic Proto-Germanic reconstruction of the Common Germanic word for 'raven' is **χraƀnaz* (> OE *hræfen*, Eng *raven*, OHG (*h*)*raban*, Germ *Rabe*(*n*), Icel *hrafn*, Nor *ravn*, Dan *ravn* and OSwe *rampn*), but it occurs as <*h**a**rab**a**naz*> in the Järsberg (in Värmland, Sweden) runic inscription, dated 520–570 CE (Antonsen 2002: 120–123; Runtextdatabas 2008 s.v. *harabanaz*). Thus, the phonetic reconstruction of **χlewaz* is **χ^{a}lewaz*, whence *Kaleva* has been borrowed. The sound substitutions are regular. The regular Proto-Finnic substitute of the (word-initial) Germanic phoneme /χ/ was /k/ in old loanword strata and /h/ in younger loanwords when the Germanic /χ/ had become /h/ word-initially (Hofstra 1985: 70). The Germanic word-final consonant /z/ was quite irregularly either dropped or borrowed as /s/ or */š/ (> /h/) into Proto-Finnic. In this case, the general dropping of the word-final consonant is predictable and understandable because there are few or

[10] In the Gotlandish dialect of Swedish, a chronologically later epenthetic vowel is /a/, too (Wessén 1966b: 50; Pamp 1978: 78).

[11] There was also a Proto-Germanic consonant epenthesis, where an intrusive plosive /t/ emerged in the consonant cluster */sr/ rendering /str/ and an intrusive plosive /b/ emerged in the consonant cluster */mr/ rendering /mbr/, e.g. PIE **srou-m-os* > PGerm **straumaz* > ON *straumr* 'stream' and PIE **dem-r-om* > PGerm **timbra* > ON *timbr* 'timber' (Hellquist 2008: 1093, 1186).

no three-syllabic words ending in /-vas/ or /-väs/ in the Finnic phonotactics. There are a couple of Finnish words ending in /-vas/ or /-väs/, such as *tai**vas*** ‘sky, heaven’ and ***evä**s* ‘packed lunch’, but they are disyllabic and thus not comparable, since the last syllable /-vA/ in three-syllabic words is almost always a derivational suffix, e.g. *liha**va*** ‘fat’ and *käte**vä*** ‘handy’. Furthermore, we may recall the fact that even the form *kalevas*/*Calewas* has been attested, which certainly does not weaken the plausibility of the word’s Germanic etymology (SMS 1999: 27; Ganander 2003: 49). When *Kaleva*(*s*) was borrowed, the phonemic step-by-step development was (ON *hlér* ‘sea’, *Hlér* ‘sea-god’ <) PScand **χ^{a}léwaz* ‘sea, sea-god giant’ → LPF **Kalevas* ‘mythical giant’ > Fin *Kaleva*(*s*) (Bjorvand & Lindeman 2007 s.vv. *le*, *ly*).

There is plenty of evidence of the existence of and parallels for an epenthetic vowel /C^{a}C/ in both runic inscriptions and Germanic loanwords in Finnic. Consider the attested runic words *h**a**rab**a**naz*, *b**a**riutiþ*, *b**a**rutR* (cf. Swe *bryt**e**r*, ON *briótr* and Icel *brjót**u**r*), *hariwol**a**fR*, *haþuwul**a**fR*, *wita*[*n*]*da-h**a**laiban*, *h**a**razaz*, *haer**a**ma*, *her**a**ma*, *wor**a**hto*, *w**a**ritu* ‘I write’, *w**a**rait* ‘I wrote’ and *asugis**a**las* (Antonsen 1975: passim). We can see that the epenthetic vowel could exist in any syllable. I think that the emergence of an epenthetic vowel in consonant clusters /C^{a}C/ has been enhanced by such trisyllabic Proto-Germanic words as **sumaraz* ‘summer’ (> ON *sumar*), **χamaraz* ‘hammer’ (> ON *hamarr*), **aχanō* ‘husk, glume’ (> Got *ahana*), **aganō* ‘husk, glume’ (> OHG *agana*), **gamalaz* ‘old’ (> ON *gamall*), where the second syllable /a/ is original (Bjorvand & Lindeman 2007: passim). In addition, several Finnish words of Germanic origin show proof of an epenthetic vowel /a/ in the etymon, e.g. *haik**a**ra* ‘stork’ ← PScand **haig^{a}ran* (> OHG *heig**a**ro* ‘stork’, Swe *häger* ‘stork’), *mat**a**ra* ‘bedstraw’ ← LPScand **mað^{a}ran* (> ON *maðra* ‘bedstraw’), *hatt**a**ra* ‘foot cloth’ ← PScand **haþ^{a}rō* (cf. OHG *had**a**ra* ‘rag’ and *el**a**ho* ‘elk’) and *katt**a**ra* ‘weed in oat’ ← PGerm **χaþ^{a}rō* (LÄGLOS 1991–2012 s.vv. *haikara*, *hattara*, *kattara*, *matara*).[12] Evidence of North Germanic vowel epenthesis is even found in the north Norwegian Sami place-name *Máhk**a**rávju* ← PScand **Mag^{a}rauju* > Nor *Magerøya*. Furthermore, more

[12] The words *kattara* and *hattara* were borrowed from the same North Germanic original, but they belong to two different Germanic loanword strata in Finnish. ***K**attara* is an older loanword and ***h**attara* somewhat younger (cf. Fin ***k**elvata* ‘to do well’ and ***k**elpo* ‘good, decent, able’ (← PGerm **χelpana* ‘to help’ and **χelpō* ‘help’) ~ ***h**elppo* ‘easy’ and ***h**elpottaa* ‘facilitate’ (← EPScand **helpō* ‘help’) and Fin ***k**ansa* ‘(a) people’ ~ ***H**ansa* ‘the Hanse’) (LÄGLOS s.vv. *hattara, helppo, kansa, kattara, kelvata*).

loanwords of this kind may be discovered in the Finnic lexicon. Previously unnoticed words of Germanic origin containing proof of an epenthetic vowel in the etyma are Fin *h**a**rakka*, Est *h**a**rakas*[13] 'magpie' < LPF **harakka*(*s*) ← PScand **h*^{a}*rōkaz* (> Icel *hrókr* 'magpie')[14] and Fin *kak*(*k*)***a****ra* 'lump, pancake, oatmeal bread, brat'[15] ← (OGut *hagri* 'oat' <) EPScand **χag*^{a}*ran* < PGerm **χagran* 'oat' → PF **kakra* > Finnic *kakra* and *kaura* 'oat' (LÄGLOS 1991–2012 s.v. *kaura*; SSA 2001 s.v. *kaura*). Another similar Germanic loanword in Finnic is *karhu* 'bear' ← *karhea* 'rough' < PF **karšeta* ← PF **karša* ← EPGerm **skraχā* 'skin, hide' (cf. another sound substitution in EPGerm **skraχā* → PF **raša* 'squirrel skin' > Fin *raha* 'money'). This etymology has unnecessarily been rejected in LÄGLOS (s.v. *karhea*).

Previously unnoticed new evidence of the existence of an epenthetic vowel in Proto-Scandinavian is provided by the sound development of some root words augmented with the unstressed Germanic prefix *ga*-. This prefix was retained in ancient East Germanic (Gothic) and West Germanic languages, but was regularly dropped in North Germanic, e.g. PGerm **gasinþija*/**gasinþijan* 'fellow traveller' > Got *gasinþja*, OHG *gisind*/*gisindi*, OSax *gisīth*/*gisīthi*, OE *gesíð* (cf. *Widsith* pp. 100–101) and ON *_sinni* (Bjorvand & Lindeman 2007: 381; Hellquist 2008: 912). However, there was one exception to this rule: The consonant /g/ in the prefix *ga*- was retained if the prefix was succeeded by one of the sonorants /l, m, n, r/, i.e. the same sounds which took the epenthetic vowel [a] when they occurred as a part of a consonant cluster. Let us consider the following examples:

> EPGerm **garasnán* > PGerm **garáznan* > PScand **g*^{a}*ránnæ* > ON *granni* 'neighbour' (cf. Got *garazna* 'neighbour')
> PGerm **ganṓgaz* > PScand **g*^{a}*nṓgaz* > **g*^{a}*nṓgaR* > ON *gnógr* 'enough' (cf. Got *ganōhs*, OE *ġenog*, OHG *ginuog*)
> PGerm **galī́kaz* > PScand **g*^{a}*lī́kaz* > **g*^{a}*lī́kaR* > ON *glíkr* 'like'

---

[13] I wish to thank Mr. Johan Schalin, who came up with this etymology in our discussions in the seminar *The Viking Age in Finland* in November 2011. However, the precise reconstruction of the word's phonological development and responsibility for its correctness is mine.

[14] The standard Proto-Baltic etymology (cf. Lith *šárka* 'magpie') of the Finnic word *harakka* mentioned in SSA (s.v. *harakka*) is phonologically more problematic since a Proto-Baltic **šarka* should have rendered ***harka* in Finnish.

[15] Cf. Fin *mukula* 'protuberance; kid', *apara* 'beverage of oat, yeast, mash', *äpärä* 'illegitimate child' (SSA s.vv. *apara*, *mukula*, *äpärä*).

PGerm **garáiðiz* > PScand **gᵃráiðiz* > **gᵃréiðiR* > ON *greiðr* ‘ready’ (cf. Got *garaid* ‘ready’)
PGerm **χráƀnaz* > NWGerm **χᵃráƀᵃnaz* > PWGerm **χᵃráƀᵃna* > OHG *hrab**a**n* ‘raven’
PGerm **χléwaz* > PScand **χᵃléwaz* (→ *Kalevas*) (cf. <h**a**laiban>) > **χᵃléwaR* > *Hlér* (→ *Liera*)
PGerm **gasínþija*(*n*) > PScand **gᵃsínþija* > **_sinnija* > ON *_sinni* ‘journey’ (cf. Got *gasinþja* ‘fellow traveller’) (Bjorvand & Lindeman 2007: passim).

The next question is: when was *Kaleva* borrowed from Germanic into Finnic? On the basis of the word’s sound shape, the date of borrowing can be estimated. The dates of the runic inscriptions containing the epenthetic vowel /a/ give us a *terminus ante quem* and possibly also a *terminus post quem* for the borrowing. The epenthetic vowel /a/ occurs in numerous runic inscriptions written in the Elder Futhark between ca. 300 CE and ca. 650 CE (Antonsen 1975: passim). However, since the oldest runic inscriptions have been found within the territory of modern Denmark and Norway, and, because no runic inscriptions older than 300 CE are known from Svealand, Götaland or Gotland, that is to say the very regions in Scandinavia which the Proto-Finns were most in contact with, the vowel epenthesis may be older in these regions, as the Germanic loanword evidence in Finnic suggests (Haugen 1976: 114; Dahl 2001: 224; Salo 2003b: 9). The epenthetic vowel was probably short and somewhat reduced – though clearly audible – since it did not change the Proto-Scandinavian stressing pattern of the words involved. Its quality seems to have stood nearest to the full vowel /a/, because this epenthetic vowel was consistently written with the rune denoting <a> and because it was borrowed as /a/ to Late-Proto-Finnic.[16] Yet, the rather limited distribution of the word *Kaleva* suggests that it is hardly very much older than the runic inscriptions. ***Ka**leva* (not *****Ha**leva* or *****_Levä***) and *tur*(***i***)***s**as* (not ***tur*(***e***)***h**as*) are approximately equally old on the basis of phonological and distributional criteria. The Finnish word *tursas* ‘(evil) sea-monster, sea-deity’ has a cognate in the Karelian, Estonian and Veps languages, so this word’s known distribution is slightly wider than Kaleva’s distribution. (Haavio 1967: 108–115, 118;

[16] In younger runic inscriptions, another epenthetic vowel written with <e> and <i> was recorded, too (Antonsen 2002: 86–87). It has not hitherto been noticed that the Finnish weekday name *p**e**rjantai* ‘Friday’ seems to reflect this younger epenthetic vowel /e/. The word *p**e**rjantai* (< *p**e**rjantaki*) has been borrowed from LPScand **fᵉrīadagr* (> ON *frīadagr* ‘Friday’). The word-medial /n/ is an analogy from *sunnu**n**tai* and *maana**n**tai* where the nasal is regular. My conclusion is that vowel epenthesis was very likely a real phonetic phenomenon in Old Scandinavian.

SSA 2001 s.vv. *Kaleva*, *tursas*.) The word *tu**rs**as* is a Proto-Germanic (← **þur*(*i*)*saz* > ON *þurs* 'evil giant', OHG *thur*(*i*)*s*) loanword. It has probably been borrowed after the Proto-Finnic sound change */š/ > /h/ (cf. Swe *fors* 'rapids' < PGerm **fu**rs**a* → EPF **pu**rš**a* > Fin *pu**rh**a* 'foaming rapids', SaN *borsi* 'foaming rapids in a canyon'). This sound change belongs to the younger ones in the relative chronology of Proto-Finnic sound changes. The substitution pattern of the Germanic /s/ and /z/ changed from */š/ to /s/ in Proto-Finnic after this sound change. (Hofstra 1985: 160–163; Aikio 2006: 19–20; Kallio 2007: 235–237; Heikkilä 2011: 70–73.) In my dissertation (Heikkilä forthcoming), I estimate that it took place in the 2nd century BCE, which is also the *terminus post quem* for the existence of *Kaleva* in Finnic. A *terminus ante quem* can be determined, too. Judging from the initial consonants (i.e. /h/ vs. /k/), the Germanic loanwords ***ha**rtia/o* 'shoulder' (← PScand *****ha**rðijōz*), ***ha**ltija* (← PScand *****ha**lðijaz*) and ***hai**k**a**ra* (← PScand *****hai**g^{a}ran*) are younger than ***Ka**leva* (← PScand *****χ**^{a}lewaz*). The lexemes *hartia* (cf. ON *h**e**rðr* 'shoulder') and *haltija* were borrowed before the Proto-Scandinavian *i*-umlaut and *haikara* before the similarly Proto-Scandinavian sound change /ai/ > /ei/, both of which took place about 500 CE, which is the *terminus ante quem* (Antonsen 2002: 28–29; Heikkilä forthcoming). Thus, *Kaleva* is likely a Roman Iron Age loanword. *Kaleva* and *tursas* belong to an early stratum of mythology-related loanwords of Germanic origin in Finnish. Younger mythic words of Germanic origin will be discussed below.

6. *Tiera, Niera, Liera* and *Torre* – the Scandinavian fellows of *Väinämöinen*

The so-called Kalevalaic folk poetry also knows such mythical characters as *Tiera*, *Niera* and *Liera*, who most often occur together in a formulaic line such as "Iku Tiera, Nieran poika" ["Iku Tiera, Niera's son"] (Haavio 1967: 205–208, 211; SKVR 2007). These variant forms are attested in rune songs in the meaning 'Väinämöinen's helper warrior and friend'. The striking feature about the forms ***T**iera*, ***N**iera* and ***L**iera* is their end rhyme instead of alliteration, which is a key feature of Finnic folk poetry. It seems to me that these names are of Scandinavian origin, too. Their Scandinavian origin has also been supported by Harry Lönnroth and Martti Linna (Messenius 2004: 6). I assume that *Liera* is a younger loan from the same North Germanic word, namely EPScand **χ^{a}lewaz* > LPScand **HlewaR* > ON *Hlér* → EFin **Leera* > Fin *Liera*, as *Kaleva*, and *Niera* can be

successfully explained as an equally old borrowing from *Snær* [snɛ:r] (< PGerm **snaiwaz* 'snow') (cf. OSwe *hœria* 'destroy, devastate, plunder' → Fin *herjata* 'revile'). The place-name *Torre* and *Torren lähde* ['Torre's spring'] in Laitila (cf. the toponym *Torrenkangas* in Kaustinen in Ostrobothnia and the farm name *Torra* in Sastamala in Satakunta) seem to reflect the Old Norse saga name *Þorri*, the mythical king of Finland and Kvenland[17], whose Proto-Scandinavian proto-form was **ÞorRē* (< PGerm **þurzan* 'dry snow' (cf. *Snær* 'snow', *Frosti* 'frost', *Drífa* 'snowdrift')), whence the Finnish toponym was borrowed in ca. 400 CE (Names Archive, Heikkilä forthcoming). The connection between *Torre* and *Þorri* was first made by Unto Salo (2008: 158, 161, 163), although he thinks that *Torre/Þorri* is equal with the Scandinavian god-name *Þórr* (< PGerm **þunraz* > Eng *thunder*), which is chronologically and thus phonologically very unlikely. I suppose that *Torren lähde* was a place of worship dedicated to the deity *Þorri* (cf. what is said in the sagas above). *Tiera* can be seen as formed on the analogy of the original assonant names *Niera* and *Liera* and it has replaced the older form *Torre* in runes, but *Torre* has been preserved in a couple of place-names. The date of borrowing is between ca. 650, when the reduction of the unstressed syllable took place, and 1142, when the already Christianized Swedes and a bishop among them made a raid on Ladoga (Antonsen 1975: 78–88; SVL 1989: 121; Jokipii 2003: 316). I would like to point out that the other Finnish (and Sami) words meaning 'giant' are of (North) Germanic origin as well, e.g. *tursas* (> *Iku-Turso* 'gigantic sea-monster in Finnic folk poetry'), *jatuni*, *jättiläinen* and saN *jiehtanas* (SSA s.vv. *jatulit*, *jätti*, *tursas*, Álgu s.v. *jiehtanas*). One more similar mythical name is *Teuri*, who is a warrior in Kalevalaic runes (Turunen 1979: 336–337; SKVR 2007). It is most likely a Scandinavian loanword, too. The original is the Proto-Scandinavian male name (Swe *djur* 'animal', Germ *Tier* 'animal' < PGerm **deuza* → PScand **deuzan* >) **DeuRē* (> the male names RSwe *TiuRi*, Icel *Dýri*) (Larsson 2002: 127;

[17] The Scandinavian influence and settlement wave to south-western Finland at the beginning of the Common Era may be recorded in the Roman historian Tacitus' book *Germania* (98 CE), where he describes a tribe called *Sitones* living on the coast of the Baltic Sea in the vicinity of *Suiones* (cf. Icel *svíar*, Swe *Svealand*). The Latinate tribal name *Sitones* can be derived from PGerm **sīðōn* (> ON *síða* 'side, coast', Eng *side*). A good translation would be 'Coast people'. In Tacitus' description *Sitones* are said to greatly resemble the Germanic tribe *Suiones* with the only exception that *Sitones* are ruled by a woman (cf. Adam of Bremen's *Terra feminarum* 'Women's land' in ca. 1075 CE) (Tacitus 98: 45; Julku 1985: 85, 1986: 51–52, 84–88; Salo 2008: 129).

Bjorvand & Lindeman 2007 s.v. *dyr*, Hellquist 2008: 146). The word's phonemic shape is very probative as to the date of borrowing. It was borrowed after the sound change /z/ > /R/, but before the sound change /eu/ > /iu/, both of which occurred in Proto-Scandinavian (see Wessén 1966a: 28, 36; Haugen 1976: 154–155). I (Heikkilä forthcoming) estimate that these sound changes took place around the year 400 CE, which is also the date of borrowing. In conclusion, the Old Scandinavian deity name *Hlér* seems to have been borrowed twice into Finnic.

7. *Kaleva's* sons from *Kalanti* – a new etymology for the place-name *Kalanti*

In my opinion, not only the etymon and the date of *Kaleva* but also the place of borrowing can be traced quite exactly thanks to onomastic evidence. I propose that *Kaleva* was borrowed from the Scandinavians in the *Kalanti* region, where there is a strong concentration of old folk stories about *Kalevanpoika*s and many place-names beginning with *Kalevan-* (Huurre 2003: 236; Names Archive; MapSite). A document from the year of 1347 reveals that *Kalanti* was the name of the whole Vakka-Suomi region at that time (Vilkuna 1969: 78; DF No. 521). Furthermore, I propose that the very toponym *Kalanti* is a derivative of *Kaleva*. The last syllable in the word *Kale**va*** was associated with the homophonous derivational suffix which could be removed and replaced with another derivational suffix (see pp. 105–106). I assume that the (wrongly analyzed) root *Kale-* was augmented with the Late-Proto-Finnic place-name suffix **-nδek* (> Fin *-nne* : *-nteen*), rendering (nom. sg.) **Kalenδek* : (gen. sg.) **Kalenteɣen* (cf. the Finnish hydronyms *Päijänne* : *Päijänteen*, *Elänne*, *Älänne* and *Peränne*). Furthermore, I assume that this **Kalenδek* was the name of the ancient long bay of the Laitila-Kalanti region which was later silted up by the isostatic uplift (see Salo 2003a: 20, 2008: 150, 152). At the mouth of this ancient fairway lie the place-names *Kalevankallio* and *Kalevanpojanvaha*, and along this fairway lie the toponyms *Kalevankallio* and *Kalevanpoikainkivi* (Huurre 2003: 247; MapSite). The hydronym **Kalenδek* was borrowed into Elder Old Swedish as **Kalend* 'Kalanti', whence the derivative **kalenn-ing* 'person from Kalanti', which is found in the oldest attestation of the name in the form *Kalenningiatekt* ["outer field of a person from Kalanti"] in a letter dated 23.6.1316. *Kalanti* is *Kaland*

[ka:land] (*Kaland* in 1437, *Anders Kaalandes* in 1435)[18] in documented Swedish, the other official language of Finland, because the name pair **Kalend* : **kalenning* was associated with the common North Germanic place-name pattern *x-l**a**nd* : *x-l**ä**nning*, e.g. *Isl**a**nd* 'Iceland' : *isl**ä**nning* 'person from Iceland' (cf. Icel *Isl**a**nd* : *isl**e**ndingur*) and *Å**l**a**n**d* : *å**l**ä**n**ning*. The modern Finnish form *Kalanti* is a re-loan from Swedish (cf. Swe *Gotland* → Fin *Gotlanti* and Swe *Öland* → Fin *Öölanti*). A more original sound shape is found in the place-name *Kalanteenkorpi* (< **Kalenteɣen-korpi*) in the Rauma municipality in south-western Finland. The existence of the second syllable vowel /a/ instead of the original /e/ can be attributed to two factors: the Swedish model *Kal**a**nd* and the association with the Finnish word *kal**a*** 'fish' (cf. the similar vowel variation in *Kal**e**va* ~ *Kal**a**va*). A Latinate form *Kalandia* is documented as early as 1332 and the oldest attestation of the name in a document written in Finnish is *Calandis* (ines. sg.) from the year 1548 by Mikael Agricola (Huldén 1984: 123; Salo 2003a: 13, 17). In addition, the name *Kalanti* is mentioned in a folk story written down in Vesilahti in Tavastia. The folk story tells about an early foreign missionary called *Hunnun Herra* ["Huntu's mister"] and *Kalannin kala* ["Kalanti's fish"], who converted people in Vesilahti to the Christian faith. (Punkari 2005.)

An archeologically and onomastically visible wave of Scandinavian influence (and quite likely also Scandinavian settlement) came to south-western Finland in about 100 CE and a second wave in the 5th century CE. Probative place-names are e.g. *Hallu* and the hill-name *Hallusvuori* (← PScand **Halluz* < PScand **halluz* > ON *hallr* '(bench of) rock', cf. Got *hallus* 'id.'), *Torre* (← PScand **ÞorRē* > ON *Þorri*), *Kainu* (< *Kainus* < *kainus* 'fairway, passage' ← PScand **gainuz* 'opening, gap, passage'), *Tuuna* (← PScand **Tūna*, cf. the *Tuna*-names in Sweden) and *Tachtoma* (< *tachto-ma* 'outer field' ← PScand **takiþō* > Swe *täkt* 'outer field'). In the light of the archaeological evidence, the Scandinavian settlers became assimilated into the Finns, but the contacts between Finland–Kvenland ['Muinais-Kainuu'] and Scandinavia remained strong during the Merovingian Period and the Viking Age. (Koivulehto 1995: 93; Salo 2003a: 39, 2003b: 26, 2008: 68.) Judging from the sound shape, *Kaleva*

[18] As in many other Germanic languages, including English, originally short vowels were lengthened in stressed open syllables in Late Old Swedish ca. 1400 (Wessén 1968: 90–92; Haugen 1976: 258–259). After the lengthening of the short vowel the Swedish name *Kaland* [ka:land] was folk-etymologically reinterpreted and translated into *Kaalimaa* "Cabbage land" by the Finns (Salo 2003a: 58).

most likely belongs to the older stratum of Scandinavian influence whereas *Torre* represents a younger stratum, and it can probably be dated to ca. 400 CE. *Liera*, *Niera* and *Tiera* are even younger. They were very likely borrowed between ca. 625 and 1142, that is to say, most probably during the Viking Age. The belief in *Kaleva* was introduced into south-western Finland by Iron Age Scandinavians. I would like to point out that place-names containing the name of a mythical pre-Christian being are by no means rare in the Scandinavian toponymy, e.g. *Odensåker*, *Odensala*, *Ulleråker*, *Torstuna*, *Torsåker*, *Frötuna*, *Fröstuna*, *Närtuna* and *Tiveden* (Pamp 1988: 32, 41, 109; SOL 2003: passim).

8. Conclusions

We have seen that many mythic beings of the Finnic pre-Christian mythology and their names have turned out to be Iron Age loans from Proto-Scandinavian. This is not at all peculiar since so many lexemes with a real world referent are of the same origin, too. The Finnic name *Kaleva* seems to have been borrowed from Proto-Scandinavians in south-western Finland, from where it spread to Karelia with the western Finnish migration in the 7th and 8th centuries, in other words at the same time as the Vikings' famous *austrvegr* ['eastway'] was established at the latest (see Huurre 2003: 250; Vahtola 2003: 24; Harrison 2009: 108, 112). In theory, *Kaleva* could have been borrowed independently from Proto-Scandinavians in northern Estonia (Est Virumaa), but it is also rather likely that it was borrowed across the Gulf of Finland with the intensive contacts in the Middle Iron Age, prior to the Viking Age when contacts became less [sic] intensive (see Huurre 2003: 253–254; Salo 2008: 135, 173, 195, 232). The folk poems and tales about *Kaleva* (ON *Hlér*), *Niera* (ON *Snær the old*), *Liera* (ON *Hlér*), *Torre* (ON *Þorri*) etc. were a part of people's worldview in the Late Iron Age both in Scandinavia and Finland, and even in Karelia and Estonia, in short, in Northern Europe. In the light of my linguistic and folkloristic article, the Late Nordic Iron Age can in many ways be seen as the last period of fully preserving the old – the climax of the Iron Age and pre-Christian Nordic culture – before a new era. Relics of the old era have been retained in the numerous Germanic loanwords in Finnish, for instance, and *Kaleva* is one of them. What further happened during the Late Iron Age was that knowledge about Finland and Kvenland, initially very little, began gradually to spread to new geographical regions, namely Britain and Iceland.

References

Aikio, Ante (2006) On Germanic-Saami contacts and Saami prehistory. *Journal de la Société Finno-Ougrienne* 91: 9–55.

Alexander, Michael (1977) *The Earliest English Poems*, second edition. London: Penguin Books.

Álgu = Etymological Database of the Saami Languages (2002–) Helsinki: The Institute for the Languages of Finland. <http://kaino.kotus.fi/algu> (1 March 2012).

Antonsen, Elmer H. (1975) *A Concise Grammar of the Older Runic Inscriptions*. Tübingen: Max Niemeyer Verlag.

—— (2002) *Runes and Germanic Linguistics*. Berlin/New York: Mouton de Gruyter.

Anttonen, Veikko (2003) Mitä *pyhä*-alkuiset paikannimet kertovat Kalannin menneisyydestä? [What do place-names beginning with *Pyhä*- tell about the past of Kalanti?]. In Veijo Kaitainen; Esa Laukkanen & Kari Uotila (eds.), *Muinainen Kalanti ja sen naapurit. Talonpojan maailma rautakaudelta keskiajalle* [Ancient Kalanti and its Neighbours: The Peasant's World from the Iron Age to the Middle Ages], pp. 225–235. Helsinki: The Finnish Literature Society.

Bjorvand, Harald & Lindeman, Frederik Otto (2007) *Våre arveord: Etymologisk ordbok* [Our Inherited Words: Etymological Dictionary]. Instituttet for sammenlignende forskning. Oslo: Novus Forlag.

Dahl, Östen (2001) The origin of the Scandinavian languages. In Östen Dahl & Maria Koptjevskaja-Tamm (eds.), *The Circum-Baltic languages. Volume* I: *Past and Present*, pp. 215–236. Amsterdam: John Benjamins B. V.

Dasent, George W. (1894) *The Orkneyingers' Saga* (*Icelandic Sagas, Vol. III*). Internet Sacred Text Archive. <http://www.sacred-texts.com/neu/ice/is3> and <http://books.google.fi/books?id=aBx_K__w7mMC&pg=PA197&dq=The+orkneyingers+saga> (1 March 2012).

DF = *Diplomatarium Fennicum*. Archivum Finlandiae Publicum. <http://extranet.narc.fi/DF/index.htm> (24 August 2012).

EES = Iris Metsmägi, Meeli Sedrik, Sven-Erik Soosaar (eds.) (2012) *Eesti etümoloogiasõnaraamat* [Estonian Etymological Dictionary]. Tallinn: Eesti Keele Sihtasutus.

Fraenkel, Ernst (1962–1965) *Litauisches etymologisches Wörterbuch*. Heidelberg: Carl Winter Universitätsverlag.

Ganander, Christfrid (2003) [1789] *Mythologia fennica*. Juha Pentikäinen (ed.), translated by Brita Löflund. Jyväskylä: Gummerus.

Grundström, Harald (1946) *Lulelapsk ordbok. Lulelappisches Wörterbuch*. Vol. I. Uppsala: Lundequistska Bokhandeln.

Grünthal, Riho (1997) *Livvistä liiviin. Itämerensuomalaiset etnonyymit* [From *Livvi* to *Livonian*. Finnic Ethnonyms]. Castrenianumin toimitteita 51. Helsinki: Suomalais-ugrilainen laitos.

Haavio, Martti (1967) *Suomalainen mytologia* [Finnish Mythology]. Porvoo: Werner Söderström Osakeyhtiö.

Harrison, Dick (2009) *Sveriges historia 600–1350* [Sweden's History 600–1350 CE]. Stockholm: Norstedts.

Haugen, Einar (1976) *The Scandinavian Languages: An Introduction to their History*. London: Faber and Faber Limited.

Heikkilä, Mikko (2011) Huomioita kantasaamen ajoittamisesta ja paikantamisesta sekä germaanisia etymologioita saamelais-suomalaisille sanoille [Observations about the dating and locality of Proto-Sámi and Germanic etymologies for Sámi and Finnic words]. *Virittäjä* 115: 68–84.

—— (forthcoming) *Bidrag till Fennoskandiens språkliga förhistoria* [Contributions to Fennoscandia's Linguistic Prehistory]. Doctoral dissertation. University of Tampere.

Hellquist, Elof (2008) [1939] *Svensk etymologisk ordbok* [Swedish Etymological Dictionary]. Malmö: Gleerups.

Hentilä, Seppo; Krötzl, Christian & Pulma, Panu (2002) *Pohjoismaiden historia* [The History of the Nordic Countries]. Helsinki: Edita.

Hofstra, Tette (1985) *Ostseefinnisch und Germanisch: Frühe Lehnbeziehungen im nördlichen Ostseeraum im Lichte der Forschung seit 1961*. Rijksuniversiteit te Groningen. Groningen: Drukkerij van Denderen B.V.

Holtsmark, Anne (2001) *Fornnordisk mytologi: Tro och myter under vikingatiden* [Old Scandinavian Mythology: Belief and Myths during the Viking Age]. Translated by Henrik Williams. Lund: Studentlitteratur.

Huldén, Lars (1984) Namnet Kaland – ett tolkningsförsök [The name Kalanti – An attempt at interpretation]. In Lars Huldén & Carl-Eric Thors (eds.), *Festskrift till Åke Granlund 28.4.1984* [Festschrift for Åke Granlund Apr 28 1984]. Skrifter utgivna av Svenska litteratursällskapet i Finland 517, pp. 123–129. Helsingfors: Svenska litteratursällskapet i Finland.

—— (2001) *Finlandssvenska bebyggelsenamn: Namn på landskap, kommuner, byar i Finland av svenskt ursprung eller med särskild svensk form* [Finnish Swedish Settlement Names: Names of Provinces, Municipalities, Villages in Finland of Swedish Origin or with a Special Swedish Form]. Skrifter utgivna av Svenska litteratursällskapet i Finland 635. Helsingfors: Svenska litteratursällskapet i Finland.

Huurre, Matti (2003) Kalevanpojat: Tarua vai totta [Kalevanpoikas: Tale or true]. In Veijo Kaitanen; Esa Laukkanen & Kari Uotila (eds.), *Muinainen Kalanti ja sen naapurit. Talonpojan maailma rautakaudelta keskiajalle* [Ancient Kalanti and its Neighbours: The Peasant's World from the Iron Age to the Middle Ages], pp. 236–254. Helsinki: The Finnish Literature Society.

Häme, Mikko (1991) Saagoista ja muinaisista kuninkaistamme [On sagas and Finland's ancient kings]. *Faravid. Pohjois-Suomen historiallisen yhdistyksen vuosikirja* 15: 181–186.

Julku, Kyösti (1985) Keskiaika [The Middle Ages]. In Kyösti Kulku (ed.), *Faravidin maa: Pohjois-Suomen historia* [Faravid's Land: The History of Northern Finland], pp. 82–199. Oulu: Kustannusosakeyhtiö Pohjoinen.

Julku, Kyösti (1986) *Kvenland – Kainuunmaa*. Studia historica septentrionalia 11. Oulu: Kustannusosakeyhtiö Pohjoinen.

Junttila, Santeri (2005) *Thomsenista Kalimaan: Vanhimpien itämerensuomalais-balttilaisten kielikontaktien tutkimushistoria vuoteen 1936* [From Thomsen to

Kalima: The History of Reserach into the Oldest Finno-Baltic Language Contacts up to 1936]. Master's thesis. <https://www.doria.fi/bitstream/handle/10024/4184/thomseni.pdf> (1 March 2012).

Järv, Ants (1987) Kalevipoeg, Kalevala, Kalevanpojat. In Martti Linna (ed.), *Muinaisrunot ja todellisuus. Suomen kansan vanhojen runojen historiallinen tausta* [Ancient Poems and Reality: The Historical Background of the Old Poems of the Finnish People], pp. 29–33. Jyväskylä: Historian aitta XX.

Kaitanen, Veijo; Laukkanen, Esa & Uotila, Kari (eds.) (2003) *Muinainen Kalanti ja sen naapurit: Talonpojan maailma rautakaudelta keskiajalle* [Ancient Kalanti and its Neighbours: The Peasant's World from the Iron Ages to the Middle Ages]. Helsinki: The Finnish Literature Society.

Kallio, Petri (2007) Kantasuomen konsonanttihistoriaa [On the consonant history of Proto-Finnic]. In Jussi Ylikoski & Ante Aikio (eds.), *Sámit, sánit, sátnehámit: Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007* [The Sami People, Words and Word-forms: Festschrift for Pekka Sammallahti on May 21 2007]. Mémoires de la Société Finno-Ougrienne 253, pp. 229–249. Helsinki: Finno-Ugrian Society.

—— (2012) The non-initial-syllable vowel reductions from Proto-Uralic to Proto-Finnic. In Tiina Hyytiäinen; Lotta Jalava; Janne Saarikivi & Erika Sandman (ed.), *Per Urales ad Orientem: Iter polyphonicum multilingue: Festskrift tillägnad Juha Janhunen på hans sextioårsdag den 12 februari 2012* [Across the Urals to the East: a Polyphonic Multilingual Journey: Festschrift dedicated to Juha Janhunen on his 60th Anniversary Feb 12 2012]. Mémoires de la Société Finno-Ougrienne 264, pp. 163–175. Helsinki: Finno-Ugrian Society..

Koivulehto, Jorma (1995) Ala-Satakunnan *Kainu* ja pohjoisen *Kainuu* [*Kainu* in Ala-Satakunta and *Kainuu* in the North]. In Päivi Rintala (ed.), *Kielen ja kulttuurin Satakunta. Juhlakirja Aimo Hakasen 60-vuotispäiväksi 1.11.1995* [The Language and Culture of Satakunta: Festschrift for Aimo Hakanen on his 60th Anniversary Nov 1 1995]. Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja 51, pp. 71–104. Turku: Turun yliopisto.

Korhonen, Mikko (1981) *Johdatus lapin kielen historiaan* [Intoduction to the History of the Sami Language]. Helsinki: The Finnish Literature Society.

Kortlandt, Frederik (1983) From Proto-Indo-European to Slavic. *Journal of Indo-European Studies* 22: 91–112.

Koski, Mauno (1967) *Itämerensuomalaisten kielten hiisi-sanue: Semanttinen tutkimus I* [The Word Family *hiisi* in Finnic: Semantic Study I]. Turku: Turun yliopisto.

Larsson, Patrik (2002) *Yrrunan. Användning och ljudvärde i nordiska runinskrifter* [The *ýr*-rune: Its Use and Phonetic Value in Scandinavian Runic Inscriptions]. Runrön 17. Institutionen för nordiska språk. Uppsala: Uppsala universitet.

Lehtinen, Tapani (2007) *Kielen vuosituhannet: Suomen kielen kehitys kantauralista varhaissuomeen* [The Millennia of a Language: The Development of the Finnish Language from Proto-Uralic to Early Finnish]. Tietolipas 215. Helsinki: The Finnish Literature Society.

Liberman, Anatoly (1992) Vowel Lengthening before Resonant + another Consonant and Svarabhakti in Germanic. In Irmengard Rauch; Gerald F. Carr & Robert L.

Kyes (eds.), *On Germanic Linguistic: Issues and Methods*. Trends in Linguistics: Studies and Monographs 68, pp. 163–214. Berlin: Mouton de Gruyter.

Linna, Martti (ed.) (1987) *Muinaisrunot ja todellisuus. Suomen kansan vanhojen runojen historiallinen tausta* [Ancient Poems and Reality: The Historical Background of the Old Poems of the Finnish People]. Jyväskylä: Historian aitta XX.

Linna, Martti; Lagerstedt, Jorma & Palmén, Erkki (transl. and eds.) (1988) *Johannes Messeniuksen Suomen, Liivinmaan ja Kuurinmaan vaiheita sekä tuntemattoman tekijän Suomen kronikka* [Chronologia Finlandiae, Livoniae et Curlandiae & Chronicon Finlandiae]. Helsinki: The Finnish Literature Society. (The original works published in 1636 and 1728 respectively.)

LÄGLOS = A. D. Kylstra; Sirkka-Liisa Hahmo, Tette Hofstra & Osmo Nikkilä (eds.) (1991–2012) *Lexikon der älteren germanischen Lehnwörter in den ostseefinnischen Sprachen*. Amsterdam: Rodopi.

Lättiläinen, Henrik (2003) *Henrikin Liivinmaan kronikka* [Henrici Chronicon Livoniae]. Translated by Maijastina Kahlos & Raija Sarasti-Wilenius. Helsinki: The Finnish Literature Society.

Malling, P.T. (1860) *Flateyjarbok. En samling av norske konge-sagaer* [Flateyjarbok. A Collection of Norwegian King-sagas]. Christiania. <http://books.google.fi/books?id=UmgJAAAAQAAJ&printsec=frontcover&hl=fi&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false> (1 March 2012).

MapSite. The National Land Survey of Finland. <http://kansalaisen.karttapaikka.fi> (5 May 2012).

Messenius, Johannes (2004) *Suomen riimikronikka* [Finland's Rhyme Chronicle]. Translated and edited by Harry Lönnroth & Martti Linna. Helsinki: The Finnish Literature Society. (Original work published in 1637.)

Names Archive. Helsinki: The Institute for the Languages of Finland.

Nielsen, Hans Frede (2000) *The Early Runic Language of Scandinavia: Studies in Germanic Dialect Geography*. Heidelberg: Universitätsverlag C. Winter.

NO = Leiv Heggstad; Finn Hødnebø & Erik Simensen (eds.) (1993) *Norrøn ordbok. 4. utgåva av Gamalnorsk ordbok* [Old Norse Dictionary]. Oslo: Det Norske Samlaget.

Nuutinen, Olli (1989) Satakunnan synty ja Kainuun kato. Tutkimus Suomen *sata-*, *sato-* ja *satama*-paikannimistä [The birth of the province of Satakunta and the disappearance of Kainuu]. *Virittäjä* 93: 11–49.

ODEE = C. T. Onions (ed.) (1982) *The Oxford Dictionary of English Etymology*. Oxford: Clarendon Press.

Ohlmarks, Åke (1983) *Fornnordiskt lexikon* [Old Scandinavian Lexicon]. Stockholm: Tidens förlag.

Pajusalu, Karl; Hennoste, Tiit; Niit, Ellen; Päll, Peeter & Viikberg, Jüri (2002) *Eesti murded ja kohanimed* [Estonian Dialects and Place Names]. In Tiit Hennoste (ed.), *Eesti murded ja kohanimed* [Estonian Dialects and Place Names]. Tallinn: Eesti Keele Sihtasutus.

Palm, Rune (2010) *Vikingarnas språk 750–1100* [The Language of the Vikings 750–1100]. Stockholm: Norstedts.

Pálsson, Hermann & Edwards, Paul (1978) *Orkneyinga Saga: The History of the Earls of Orkney*. London: Penguin Books.

Pamp, Bengt (1978) *Svenska dialekter* [Swedish Dialects]. Natur och kultur. Stockholm.

—— (1988) *Ortnamnen i Sverige* [Place-names in Sweden], fifth edition. Studentlitteratur.

Pentikäinen, Juha (1989) *Kalevala Mythology*. Translated and edited by Ritva Poom. Bloomington/Indianapolis: Indiana University Press.

—— (1995) *Saamelaiset: Pohjoisen kansan mytologia* [The Sami People: The Mythology of a Northern People]. Helsinki: The Finnish Literature Society.

Punkari, Yrjö (2005) *Kristinuskon ja kirkon tulo Vesilahteen* [The Arrival of Christianity and the Church in Vesilahti]. <http://www.narva.sci.fi/amaijala/vti_kris.html> (10 March 2012).

Ringe, Don (2006) *From Proto-Indo-European to Proto-Germanic. A Linguistic History of English*. Volume I. Oxford: Oxford University Press.

Rowe, Elizabeth Ashman (2000) Origin Legends and Foundation Myths in Flateyjarbók. In Geraldine Barnes & Margaret Clunies Ross (eds.), *Old Norse Myths, Literature and Society*. Proceedings of the 11th International Saga Conference 2–7 July 2000, University of Sydney, pp. 441–454. Centre for Medieval Studies, University of Sydney.

Saagpakk, Paul F (1982) *Eesti-inglise sõnaraamat* [Estonian-English Dictionary]. New Haven/London: Yale University Press.

Salo, Unto (2003a) Oliko Kalanti muinaismaakunta [Was Kalanti an ancient province?]. In Veijo Kaitanen, Esa Laukkanen & Kari Uotila (eds.), *Muinainen Kalanti ja sen naapurit. Talonpojan maailma rautakaudelta keskiajalle* [Ancient Kalanti and its Neighbours: The Peasant's World from the Iron Age to the Middle Ages], pp. 13–91. Helsinki: The Finnish Literature Society.

—— (2003b) Faravidin retket ja Satakunnan organisoituminen rautakaudella [Faravid's journeys and the organisation of the province Satakunta in the Iron Age]. *Faravid. Pohjois-Suomen historiallisen yhdistyksen vuosikirja* 27: 5–62.

—— (2008) *Ajan ammoisen oloista: Satakunnan ja naapurimaakuntien esihistoriaa* [On the Conditions of Time Immemorial: On the History of Satakunta and Neighbouring Regions]. Helsinki: The Finnish Literature Society.

Sammallahti, Pekka (1998) *The Saami Languages: An Introduction*. Kárášjohka: Davvi Girji.

Samnordisk runtextdatabas [Common Nordic Rune Database] (2008) <http://www.nordiska.uu.se/forskn/samnord.htm> (1 March 2012).

Siikala, Anna-Leena. 1987. Myytti ja historia eeppisessä kansanrunoudessa [Myth and History in Epic Folk Poetry]. In Martti Linna (ed.), *Muinaisrunot ja todellisuus. Suomen kansan vanhojen runojen historiallinen tausta* [Ancient Poems and Reality: The historical Background of the Old Poems of the Finnish People], pp. 11–19. Jyväskylä: Historian ystäväin liitto.

SKVR = *Suomen Kansan Vanhat Runot* [The Old Poems of the Finnish People] (2007) The Finnish Literature Society. <http://dbgw.finlit.fi/skvr> (25 March 2012).

SMS = Matti Vilppula (ed.) (1999) *Suomen murteiden sanakirja* [Finnish Dialect Dictionary], sixth volume. The Research Institute for the Languages of Finland. Helsinki: Edita.

SOL = Mats Wahlberg (ed.) (2003) *Svenskt ortnamnslexikon* [Swedish Place-name Lexicon] Utarbetad inom Språk- och folkminnesinstitutet och Institutionen för nordiska språk vid Uppsala universitet. Uppsala.

SPNK = Sirkka Paikkala (ed.) (2007) *Suomalainen paikannimikirja* [Finnish Place-name Book]. Helsinki: Karttakeskus & The Research Institute for the Languages of Finland.

SSA = Erkki Itkonen & Ulla-Maija Kulonen (eds.) (2001) [1992–2000] *Suomen sanojen alkuperä* [The Origin of Finnish Words]. Proceedings of The Finnish Literature Society 556/Proceedings of The Research Institute for the Languages of Finland 62. Helsinki: The Research Institute for the Languages of Finland.

SVL = Martti Linna (ed.) (1989) *Suomen varhaiskeskiajan lähteitä* [Sources about Finland in the Early Middle Ages]. Historian aitta. Historian ystäväin liitto XXI. Jyväskylä: Historian ystäväin liitto.

Tacitus, Cornelius Publicus (98) *Germania*. <http://www.thelatinlibrary.com/tac.html> (23.8.2012).

Tarkiainen, Kari (2008) *Sveriges Österland: Från forntiden till Gustav Vasa: Finlands svenska historia 1* [Sweden's Eastland: From Ancient Times to King Gustav Vasa: Finland's Swedish History 1]. Helsingfors: Svenska litteratursällskapet i Finland.

Tarvel, Enn (2004) Idrisi und Reval. In Bernhard Jähnig & Klaus Militzer (eds.), *Aus der Geschichte Alt-Livlands. Festschrift für Heinz von zur Mühlen zum 90. Geburtstag. Schriften der Baltischen Historischen Kommission. Band* 12, pp. 1–9. Münster: Lit Verlag.

The Northern Way (2009) *Frá Fornjóti ok hans ættmönnum: Hversu Noregr byggðist* [On Fornjótr and his Relatives: How Norway was inhabited]. The New Northvegr Center. <http://www.northvegr.org> (1 March 2012).

Turunen, Aimo (1979) *Kalevalan sanat ja niiden taustat* [Kalevala's Words and their Backgrounds]. Joensuu: Karjalaisen kulttuurin edistämissäätiö.

Tuulio, Oiva Johannes (1936) Satakunnan vanhin tunnettu paikannimi [Satakunta's Oldest Known Place-name]. *Satakunta. Kotiseutututkimuksia* X: 34–43.

USN = Vilkuna, K. & Mikkonen, P. (eds.) (1988) *Uusi suomalainen nimikirja* [New Finnish Name Book]. Helsinki: Kustannusosakeyhtiö Otava.

Vahtola, Jouko (1987) Kansanrunojen nimistön ja todellisen nimistön välisistä yhteyksistä [On the connections between the names in folk poems and real world toponymy]. In Martti Linna (ed.), *Muinaisrunot ja todellisuus. Suomen kansan vanhojen runojen historiallinen tausta* [Ancient Poems and Reality: The Historical Background of the Old Poems of the Finnish People], pp. 37–50. Jyväskylä: Historian aitta XX.

Vahtola, Jouko (2003) *Suomen historia: Jääkaudesta Euroopan unioniin* [A History of Finland: From the Ice Age to the European Union]. Helsinki: Kustannusosakeyhtiö Otava.

Vilkuna, Kustaa (1969) *Kainuu – Kvänland. Ett finsk-norsk-svenskt problem* [Kvenland: A Finnish-Norwegian-Swedish Problem]. Acta Academiae Regie Gustavi Adolphi XLVI. Uppsala: Almqvist & Wiksells boktryckeri aktiebolag.

Wessén, Elias (1966a) *Isländsk grammatik* [Icelandic Grammar], second edition. Stockholm: Svenska bokförlaget/Norstedts.

—— (1966b) *Våra folkmål* [Our Dialects], seventh edition. Lund: Fritzes.

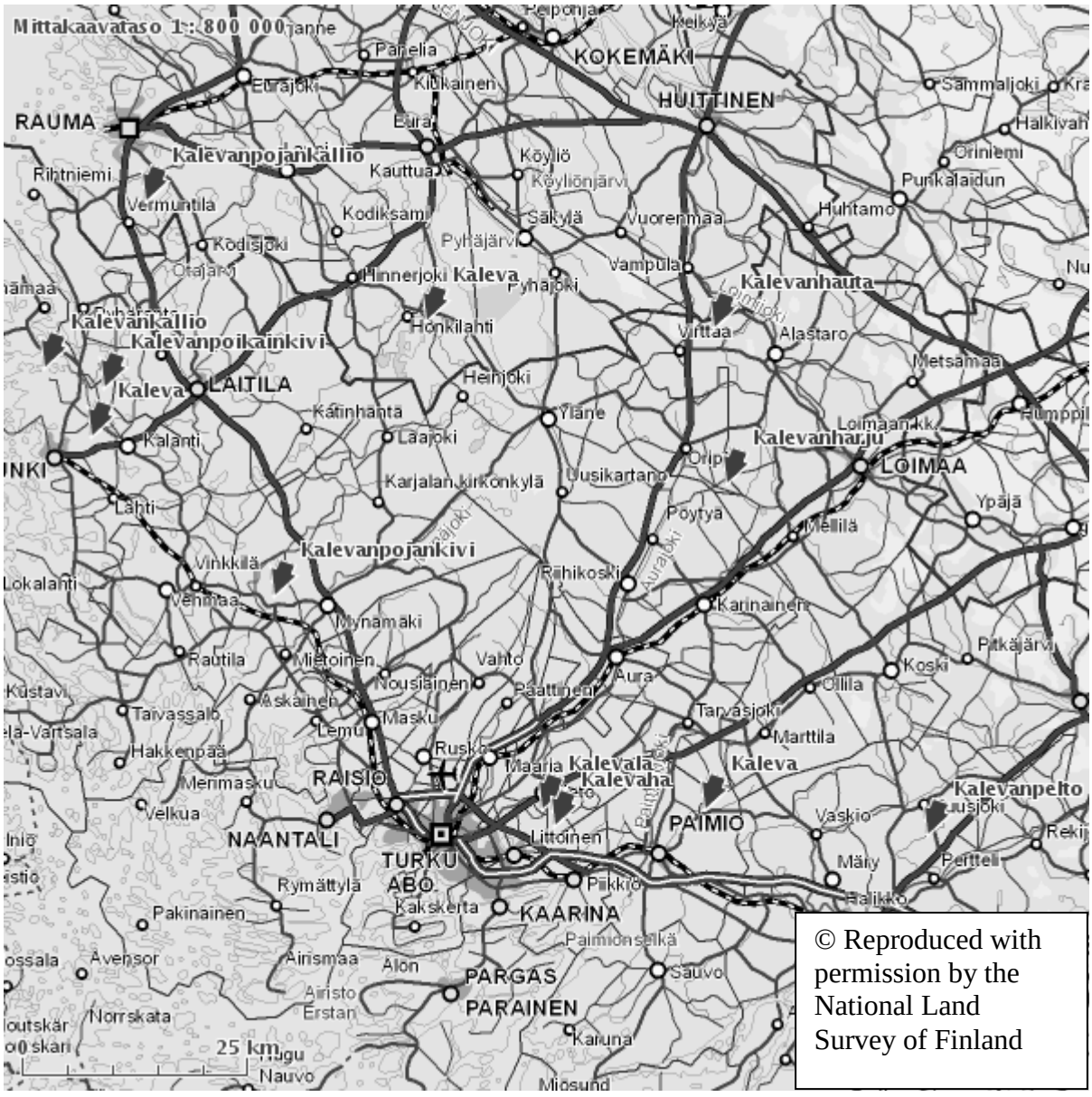
—— (1968) *Svensk språkhistoria I: Ljudlära och ordböjningslära* [History of Swedish 1: Phonology and Morphology]. Stockholm: Almqvist & Wiksell.

Wójcik, Jerzy (2001) Old English fricatives: lenition and licensing. In Katarzyna Dziubalska-Kołaczyk (ed.), *Constraints and Preferences*. *Trends in Linguistics*, pp. 359–390. Berlin: Mouton de Gruyter.

Zetterberg, Seppo (2007) *Viron historia* [Estonia's History]. Helsinki: The Finnish Literature Society.

Appendix

A map of some place-names containing the word *Kaleva* in south-western Finland



Abbreviations

A → B = B is borrowed from A, derivation
A > B = A develops into B
* = reconstructed sound shape
** = impossible or non-existing sound shape

C´C = the overlong grade in the Sami consonant gradation
Dan = the Danish language
EFin = Early Finnish (varhaissuomi) (ca. 1000–1543 CE)
Eng = the English language
EPF = Early Proto-Finnic (ca. 1500–500 BCE)
EPGerm = Early Proto-Germanic (ca. 1000–600 BCE)
EPSa = Early Proto-Sami (ca. 600–1 BCE)
EPScand = Early Proto-Scandinavian (ca. 160–500 CE)
ESa = Early Sami (ca. 600–1000 CE)
Est = the Estonian language
Fin = the Finnish language
gen. = genitive
Germ = the German language
Got = the Gothic language
Icel = the Icelandic language
ines. = inessive
Kar = the Karelian language
Lith = the Lithuanian language
LPF = Late Proto-Finnic (the first millennium of the Common Era)
LPScand = Late Proto-Scandinavian (ca. 500–800 CE)
MPF = Middle Proto-Finnic (ca. 500–1 BCE)
nom. = nominative
Nor = the Norwegian language
NWGerm = North-West-Germanic (ca. 1–200 CE)
OE = Old English (ca. 700–1100 CE)
OEst = Old Estonian (ca. 1000–1500 CE)
OGut = Old Gutnish (spoken in Gotland) (ca. 900–1500 CE)
OHG = Old High German (ca. 750–1100 CE)
ON = Old Norse (fornvästnordiska = norrønt) (ca. 800–1350 CE)
OSax = Old Saxon (ca. 750–1100 CE)
OSwe = Old Swedish (1225–1526 CE)
PF = Proto-Finnic (ca. 500 BCE–1000 CE)
Pgerm = Proto-Germanic (ca. 600–1 BCE)
PIE = Proto-Indo-European (ca. 4000 BCE)
PreGerm = Pre-Germanic (ca. 1800–600 BCE)
Pscand = Proto-Scandinavian (urnordiska) (ca. 160–800 CE)
PWGerm = Proto-West-Germanic (ca. 200–450 CE)
RSwe = Runic Swedish (ca. 800–1225 CE)
Rus = the Russian language

Sa = Saami
SaI = Inari Sami
SaLu = Lule Sami
SaN = North Sami
sg. = singular
Swe = Swedish
Vot = the Votic language

Contact information:

Mikko Heikkilä
Kanslerinrinne 1
Pinni B 4096
33014 University of Tampere
Finland
e-mail: Mikko(dot)K(dot)Heikkila(at)uta(dot)fi

Nivedita Kumari and S. Devaki Reddy

(In)directness of Requesting in Hindi

Abstract

The present study examines whether the (in)directness of requests correlates with the degree of politeness and whether the use of honorifics influences the degree of directness of linguistic forms in requests, in Hindi. Since honorifics play a significant role in expressing politeness, using honorifics appropriately is integral to effective communication; yet, indirectness contributes to the mitigation of a face-threatening act (Brown & Levinson 1987). Taking Blum-Kulka's (1989) theory on categorization of requests, the study attempts to describe how a language with an inbuilt honorific system uses direct and indirect request forms. The study is based on three kinds of data gathered from respondents: (1) the responses to an open-ended question aimed to gather the canonical request forms in the language in different request situations; (2) responses to a closed question in which the respondents indicated their choice of the level of politeness they would use with addressees in different role-relationships; (3) respondents' ranking of the forms from the most polite to the least polite. The results show that direct and conventionally indirect forms of request are used frequently in Hindi. Further, the need for clarity determines the degree of directness of requests, as honorifics serve the purpose of indicating politeness.

1. Introduction

The literature on the theories of politeness and speech acts considers requesting to be inherently threatening to the negative face of the addressee as it involves an expectation of a favor that a speaker demands from the addressee. In order to mitigate this Face-threatening Act, languages have a set of politeness strategies (Brown & Levinson 1987). Focusing on requests as speech acts, Blum-Kulka et al. (1989) categorize the linguistic forms based on the degree of explicitness of the illocution in

the speech acts. The present study describes the request forms in Hindi, employing this categorization, and brings out the culture-specific nuances of requesting in Hindi, which has an inbuilt system of honorifics.

The study is based on a set of written discourse completion tests and two scale questions. Both questions have been answered by a total of 162 respondents on a university campus from different age-groups, gender, and occupations including students, teachers and non-teaching staff.

The paper examines how explicit and implicit a request can be in Hindi in a common request situation with addressees of different social distance (friends, family, strangers) and age group (younger and older). The paper examines whether there is a correlation between directness/indirectness and politeness in Hindi requests. As indirect requests seem to be more polite than the direct requests due to their mitigating effect, the paper seeks to answer whether a language with a built in system of honorifics, such as Hindi, views the levels of (in)direct request strategies differently?

The paper is divided into six sections, including the introduction as the first section. Section 2 briefly outlines the literature on politeness and honorifics. Section 3 gives an overview of data collection whereas section 4 describes the gathered data and the methods of analysis used. Section 5 lists the findings of the study and section 6 ends with the discussion and conclusion based on the study.

2. Literature background

Drawing on Goffman's (1967) concept of 'face', Brown & Levinson (1987) describe the politeness strategies used to mitigate a face-threatening act (FTA). The strategies are listed under five broader categories:

1. Do the FTA, without redressive action (bald on record)
2. Do the FTA, with redressive action (positive politeness)
3. Do the FTA, with redressive action (negative politeness)
4. Do the FTA (off record)
5. Don't do the FTA

The most direct form, 'bald on record', is the least mitigating and least polite by this categorization. The degree of politeness is said to increase with the degree of indirectness. Blum-Kulka et al. (1989), in their goal to study the cross-cultural realization of requests as speech acts, rank the

Head Acts of requests (based on illocutionary[1] force) from most direct to least direct, as the following list illustrates:

1. Mood Derivable
2. Explicit Performative
3. Hedged Performative
4. Locution derivable
5. Want Statement
6. Suggestory Formula
7. Query Preparatory
8. Strong Hint
9. Mild Hint

One of the basic distinctions between direct and indirect forms, according to Blum-Kulka (1989: 2) is that in "direct speech acts the speaker says what he means but in indirect speech acts, he or she means more than or something other than, what he or she says."

However, the correlation of (in)directness with politeness is contested in Blum-Kulka (1987), where the author finds that indirectness does not always indicate more politeness, and points out the importance of 'clarity' and the explicitness of a speech act. This discussion highlights the significance of clarity (Grice 1967) and politeness in 'pragmatic competence' (Lakoff 1973). The variation in the use of (in)directness in requesting, in different cultures, is highlighted by Fukushima (1996) and Marti (2006).

Politeness theories in the literature can be divided into two kinds. The first category describes universal aspects of politeness (Lakoff 1973; Brown & Levinson 1987; Leech 1983, 2005). The second category focuses on the culture-specific aspects of politeness strategies (Wierzbicka 1985; Blum-Kulka et al.1989, 1987).

[1] Illocutionary acts are described in Austin (1962), Searle (1969) and further explained in Levinson (1983), as the conditions that are necessary for the act (here request) performed.

2.1 Theories of politeness in requesting and honorifics

The act of requesting, categorized as an FTA, threatens the negative face of the hearer (Brown & Levinson ibid: 66) as a Speaker (S) indicates that he wants the Hearer (H) to do or refrain from doing, some act A.

One of the criticisms of the universal nature of the theory of politeness suggested by Brown & Levinson (ibid) is the inclusion of honorifics in the negative politeness strategy, because honorifics are normally used even when there are no face-threatening acts involved. Fukada & Asato (2004), however, suggest that if honorifics are not used, it would lead to a face-threatening situation. Culpeper (2010) argues (im)politeness to be both inherent in an expression and also to get its meaning from its use in a particular context.

Blum-Kulka et al. (1989) in CCSARP (Cross Cultural Speech Act Realization Project) build up a nine-level categorization of Head Acts of requests (see table 5) and also describe the non-Head Acts such as alerters and supportive moves. Based on the (in)directness of these Head Acts Blum-Kulka et al. (1989) have further clubbed these nine levels of requests into direct, conventionally indirect and unconventionally indirect.

2.2 Honorification and politeness strategies in Hindi

Politenesss in Hindi may be expressed by varying the pronouns, verbs, titles of the addressee and reference, and distinct lexical forms to refer to things and actions. Politeness may also be indicated by using different types of sentences like, passive and imperative (Pandharipande 1973). Thus, in Hindi, the level of politeness can be expressed by various morpho-syntactic means. Pronouns in Hindi vary according to number (singular/plural) or social distance. The second person pronoun ‘you’ has a three way distinction in Hindi, *tu* ‘you1’, *tum* ‘you2’ and *aap* ‘you3’ (Misra 1977). The forms ‘you1’ and ‘you2’ are used among intimate interactants, whereas, ‘you3’ is the honorific form, used to show respect to the addressee (Table 1). Honorifics may also be suffixed to proper nouns or titles of the addressees.

Table 1. Honorifics in Hindi noun, pronoun and verb

	Proper Noun	Noun	Pronoun	Verb
Neutral 1	Name	Name	*tu* 'you'	*do* 'give'
Neutral 2	Name	Name	*tum* 'you'	*de* 'give'
Polite/ Honorific form	name + -ji *Mohan-ji* 'name-HON' *Sharma-ji* 'surname-HON'	profession + -ji / -sahab, *master-ji* 'teacher-HON' *doctor sahab* 'doctor HON'	*aap* 'you.HON'	*dijiye* 'give.HON'

The verbs in Hindi are inflected to agree with singular or plural subjects as follows.

(1) *tum* *do*
you.neutral1 give.neutral1
'you give'

(2) *aap* *dijiye*
you.HON give.HON
'you give'

Apart from these there are some lexical forms in Hindi-Urdu[2] (Jain 1969:83–84) which show verbalization of respect. The author shows how in asymmetrical and symmetrical relationships the speaker uses neutral forms, humble forms and exaltation forms. For example, *ghar 'house'* is neutral, *garibkhana* 'humble-house, hut' is the humble form, and *daulatkhana* 'wealth house/palace' is an exaltation form. Verbs like *to give* also have three forms such as *dena* 'to give', *pesh karna* 'to offer', *ada karna* 'to grant'.

Pandharipande's hierarchy as shown in Srivastava & Pandit (1988: 190,199) described the hierarchy of degree of politeness in different forms of sentence constructions in Hindi such as passive, simple present,

[2] Hindi-Urdu is used among the Indian linguists to refer to Hindi which is not different from Urdu. However, the lexical terms in Urdu have more polite forms but are mutually intelligible to Hindi speakers.

optative, future imperative and imperative. However, Srivastava & Pandit (1988:203) state that the degree of politeness associated with different sentence construction also depends on the context. Thus, the social context and the social meaning of a language structure need to be taken into consideration to determine the degree of politeness of the linguistic form. These various ways of showing varying degree of politeness, including, honorifics and sentence construction, do not form a part of the nine-way categorization of (in)directness model offered by Blum-Kulka et al. (1989). In order to widen the description of verbal (in)directness and politeness to fit languages like Hindi as well, the present study focuses on the lacunae in the description of the use of linguistic politeness on the basis of (in)directness of a speech act without considering the inbuilt honorification system.

2.3 Purpose of the study

Building on the available literature, the paper describes the request forms in Hindi and categorizes them based on their (in)directness. Further the requests are categorized based on the use of honorifics in order to examine whether the degree of directness or indirectness is influenced by the use of honorifics.

3. Data

3.1 Instrument

The present study is based on a questionnaire, which had three sections (See Appendix 1 for the English equivalent of the questionnaire used). The first section of the questionnaire contained an open-ended question. The respondents were asked to write the request forms they would use in the given situations to three categories of addressees of varying social distance (friends, family members, strangers). Table 2 below lists three request situations and the role-relationship between the interactants.

Table 2. The request situations in the questionnaire

Degree of imposition → Addressee ↓	**Situations**
To friends (younger/older, male/female friend)	Ask money
To family members (younger/older, sister/brother, mother, father)	Ask money
To strangers (younger/elder male/female stranger)	Ask to move ahead in a queue

The second section of the questionnaire had a three point scale question aimed to determine the level of politeness that the respondents would claim to use with each category of addressee of different age groups (younger, older) and social distance (friend, family, strangers). The question contained the labels 'very polite', 'polite' and 'casual,' and the respondents were asked to indicate which level of politeness they would use. The respondents had to tick one of the options.

The last section of the questionnaire contained nine request forms used in the context of asking for money. In this context, the addressee was a friend belonging to the same age-group and gender. The respondents were asked to rank the nine request forms on a scale of 1 to 9, where 1 was the least polite and 9 the most polite.

3.2 Respondents

The respondents chosen for the study were native speakers of Hindi, who were studying or working on a university campus in Delhi[3]. There were a total of 163 respondents. The questionnaire was in written form and it was distributed manually to the respondents after having their personal consent. The average time to fill up the questionnaire was 15 minutes.

[3] The field of the study was Jawaharlal Nehru University, Delhi as a representative field that has a heterogeneous population of Hindi speakers.

4. Data description

4.1 Request Acts in Hindi elicited from the respondents

The responses to the open-ended question were categorized based on CCSARP (Cross Cultural Speech Act Realization Project) as in Blum-Kulka (1989) into the following categories:

1. Alerters
2. Supportive Moves
3. Head Acts
4. Internal Modifications

For the purpose of analysis only the Head Acts were considered. The Head Act is the kernel of a request whereas the other parts of the speech act are optional and may be present in different orders and combinations. To illustrate an example from Hindi data,

(3) *suniye! Please aapke paas kuch paise hoNge?*
listen.HON please you.GEN with some money have.FUT.INT?
baat ye hai ki mere paise khatm ho gaye haiN
reason this be.PRES COMP I.POSS money over be.PRES go.PRES be.PRES.PL
aur mujhe ghar jana hai.
CONJ I.DAT house go.INF be.PRES
'Listen! Please do you have some cash on you? The thing is that my money is over and I need to go home'

- *Suniye* =**Alerter** (attention getter),
- *Please* = **Internal Modification** (politeness marker),
- *Aapke paas kuch atirikt paise honge* = **Head Act** (Conventionally Indirect Request),
- *Baat ye hai ki mere paise khatm ho gaye hain aur mujhe ghar jana hai.* = **Supportive Moves** (Giving reason and explaining the situation)

The Head Acts were categorized into nine request forms based on Blum-Kulka's categorization. Table 3 below illustrates the nine categories of Head Acts with the examples as given in Blum-Kulka and the Hindi examples taken from the data of the present study:

Table 3. Categorization of Hindi requests in the DCT based on Blum-Kulka et al. (1989)

Categories	**Sub-categories based on illocutionary force**	**English Examples (as given in Blum-Kulka)**	**Hindi Examples as found in the responses for the request situations in Table 3**
A Direct Request	A1 Mood Derivables	Clean the kitchen	*ghar jane ke liye paise do* 'give (me) money to go home'
	A2 (Explicit) Performative	I am asking you to clean the mess	*maiN tere paise le rahi huN* 'I am taking your money'
	A3 Hedged Performative	I would like to ask you to give your presentation a week earlier	*yadi aap itna de sake to aabhari rahuNgi* 'if you are able to give this much I would be obliged'
	A4 Obligation statement	You'll have to move the car	*maaf kijiyega aapko aage baRhna chahiye* 'sorry but you should move ahead in the queue'
	A5 Want Statement	I really wish you'd stop bothering me	*is kaam ke liye kuchh paisoN ki zarurat hai* 'I need some money for this work'
B Conven-tionally Indirect Request	B6 Suggestory Formula	How about cleaning up?	*aage chaleN* 'shall we go ahead'

	B7 Query Preparatory	Could you clean up the kitchen, please?	*kya tum mujhe kuch paise de sakti ho?* 'can you lend me some money?'
C Unconven-tionally Indirect Request	C8 Strong Hint	You have left the kitchen in a right mess	*kuch paise kam par gaye* 'falling short of cash'
	C9 Mild Hint	'I am a nun' in response to a persistent hassler	*aaj to maiN baRi musibat me paR gayi* 'today I am in a big trouble' (making the addressee ask about the problem and offer for help, like lending money to go home)

Table 4. The mean percentage of request types used by respondents with friends, family and strangers.

	A1	A2	A3	A4	A5	B6	B7	C8	C9	Won't ask
Friends	27.3	0.9	7.37	0.15	6.87	0	29.12	6.27	0	12.55
Family	31.08	0.6	8.8	0	18.28	0	19.95	5.43	0	4.7
Strangers	69.15	0	3.1	0.6	0	3.1	12.57	6.4	0	0.6

The table 4 above shows that the Mood Derivable (A1), Hedged Performative (A3), Want Statement (A5), Query Preparatory (B7), Strong Hint (C8) and Won't ask[4] are the more frequently used forms amongst all nine types. The frequency of these forms varies with the social distance between the interactants as the use of forms such as Query Preparatory (B7) and Won't ask options were relatively more amongst friends whereas, with family members Mood Derivable (A1), Query Preparatory (B7), Want Statement (A5) and Hedged Performative (A3) were used more. The most frequent use of Mood Derivable (A1) even with strangers is strikingly evident in the data.

[4] Won't Ask was included in the coding of the data to capture the complete withdrawal of making the request. According to Brown & Levinson (ibid) it would fall in the fifth category of 'Don't do an FTA.'

4.2 Use of honorifics in the request acts

The data show a three way division in the level of politeness based on honorifics. The Direct Request form, A1 (Mood Derivable) and Conventionally Indirect Request form B7 (Query Preparatory) clearly had this three levels of politeness: honorific, neutral1 and neutral2. This sub-section shows the use of honorifics with the (in)directness in requests. Table 5 below enlists the verb and pronoun forms and their level of politeness based on honorification.

Table 5. Honorific verb and pronoun forms in Hindi as in the open ended responses

CATEGORIZATION BASED ON CCSARP	**VERB FORMS** for *do*'give' and *baRho*[5] 'move ahead'	**Honorific**	**Neutral1**	**Neutral2**
A1 (Mood Derivable)	**Imperative**	*deN*, *dijiye, baRheN*	*dena, do*	*de*
B7 (Query Preparatory)	**Interrogative**	*denge/ dengi?, dijiyega? baRhiyega?*	*doge/dogi?*	*degi?*
	Interrogative+ Modal+Aux (*Kya + sak/pa* 'capability modals' + *hai* 'be verb form')	*de sakti/sakte haiN (kya)?, sakenge (kya)?, de payenge?*	*de sakti/sakte ho, kar sakoge/ sakegi/ paogi (kya)? le sakti ho?*	*de sakti/sakta hai*
	PRONOUNS			
All request forms	Hearer-oriented Second person pronoun 'you'	*aap*	*tum*	*tu*
A5 (Want Statement), A3 (Hedged performative)	Speaker-oriented	*main* 'I'		
C8 (Strong Hints)	Impersonal	*line* 'Queue'		

[5] Hindi data showed a number of serial verb constructions used in Hindi, like baRh jaayen 'move.go.HON.' As the discussion of these complex predicates falls outside the scope of the present study, only simple verb forms were presented in the Table 5.

4.2.1 Honorification in verbs as per the data

The Table 5 above shows the frequently used forms of *do* 'give' in Hindi when asking for money and *baRho* 'go ahead' when asking to move ahead in the queue. The variety of forms of 'give' calls for an explanation for each of them. To begin with, the imperative form 'de' is a form used only with younger friends or siblings as in,

(4) *chal* *paise* *de*
ALERTER money give.neutral2
'C'mon give money'

It sounds like a demand on an addressee. *de do*, a serial verb construction of the form V1V2, and *dena*, a non-finite form of the verb, which are also other intimate and neutral2 forms. But the level of politeness shown by *de* is a little lower than the latter forms. For example,

(5) *behen* *paise* *de do/dena*
sister (address form) money give-give/give.INF.neutral1
'Sister! Give money'

deN is a typical formal honorific form, which is used with customers or for general public. *dijiye* however, is a more general form of honorific used in day-to-day life. The example sentences below illustrate this.

(6) *kripya* *paise* *deN*
please money give..HON
'Please give money'

(7) *mummy* *paise* *dijiye.*
mother money give..HON
'Mother! Give money'

The Query Preparatory forms of 'give' have only interrogatives like *dijiyega* and *dengi* that are honorific forms of implying 'will you give', whereas, *dogi/doge* is the neutral1 form. *degi/dega* belong to the neutral2 form.

Likewise, the modal interrogatives use the capability modal *sakna* or *paana* 'to be able to' and have a three way division as shown in Table 5 above.

(8) *kya kuch paise de sakte haiN / de payenge?*
INT some money give MOD(can) be.HON / give MOD(able to).INT.HON
'Can you give some money?' / 'Will you be able to give some money?'

Similarly, *de sakti ho* is one of the neutral1 and *de sakti hai* is the neutral2 form.

4.2.2 Honorification in pronouns as per the data

Most of the pronoun forms in request data were second person pronouns as the request was addressed to a hearer and these pronouns had a three way distinction in the levels of honorification as shown in Table 6. However, there were two other cases of pronominal usage, which did not show any distinction in the levels of honorification. First, the pronouns used for 'I' in Speaker-oriented requests as shown in the example (15) below. This request form falls under the category A5 (Want Statement).

(9) *mujhe paise chahiye*
I.DAT money want
'I want money'

Second, the data contained impersonal forms of requests where neither the Hearer nor the Speaker was mentioned. For example, in (16) the Speaker just mentions that 'the queue' is moving ahead. This request form falls under the category C8 (Strong Hint).

(10) *line aage barh gayi hai*
queue ahead move go.PERF be.PRES
'The queue has moved ahead'

4.3 Rating of level of politeness based on the role-relationship

The second part of the questionnaire is a role-relationship question where the respondents had to choose the level of politeness used with addressees of different age and social distance. The Table 7 below shows the

hierarchy in the level (very polite, polite, and casual) of perceived politeness based on the mean percentage of the choice of the levels of politeness of the respondents for each category of addressee. As all the three levels of politeness were found to be used by addressees of all the three groups, the Table 6 below presents the only the variations in the rating of levels of politeness used by the respondents.

Table 6. Hierarchy in politeness level varying with social distance and age. C = Casual, P = Polite, VP = Very Polite

Group of addressees I: Family	Group of addressees II: Friends	Group of addressees III: Strangers
Parents VP > P > C	Elder Friends P > C > VP	Elder Strangers VP > P > C
Elder Siblings P > VP > C	Young Friends C > P > VP	Younger Strangers P > VP > C
Younger Siblings C > P > VP		

The hierarchy shown in Table 7 shows that the speaker's relationship with the addressee as well as their age decide the level of politeness perceived to be used by the respondents. With elder addressees a higher use of 'Very Polite (VP)' level of politeness was rated by the respondents, as with parents and elder strangers. But with elder friends Polite (P) level of politeness were used more frequently. With younger addressees, Casual (C) level of politeness were chosen more by the respondents, as with younger friends and siblings. However, with younger strangers, Polite (P) level of politeness was chosen by a higher number of respondents.

4.4 Ranking of nine types of request based on degree of politeness

The third part of the questionnaire has the nine canonical types of Head Acts used in the requests for borrowing money from a friend. The respondents were asked to rank the requests from 1 to 9 to indicate what they perceived as the hierarchy of politeness, from the least polite (1) to the most polite (9). Since the interactants were supposed to be of the same age and gender, honorifics were not included in the sentences given for

ranking, which helped in eliciting the attitude of the respondents towards request forms of various degrees of directness.

Table 7. Mode distribution of ranking of request forms in Hindi

Request forms	**Ranking in Hindi** (percentage of respondents giving it the ranking given in brackets)
A1 (Mood Derivable)	1 (41.7%)
A2 (Performative)	4 (16.0%)
A3 (Hedged Performative)	5 (16.6%)
A4 (Obligation Statement)	2 (17.2 %)
A5 (Want Statement)	8 (20.9%)
B6 (Suggestory Formula)	7 (17.2%)
B7 (Query Preparatory)	9 (38.7%)
C8 (Strong Hint)	6 (17.2%)
C9 (Mild Hint)	3 (12.3%)

Table 7 shows that A1 (Mood Derivables) was ranked the least polite request form (41.7 % of the respondents gave it the lowest value, 1) and B7 (Query Preparatory) was ranked the most polite form by a large percentage of respondents (38.7 % of the respondents gave it the highest value, 9).

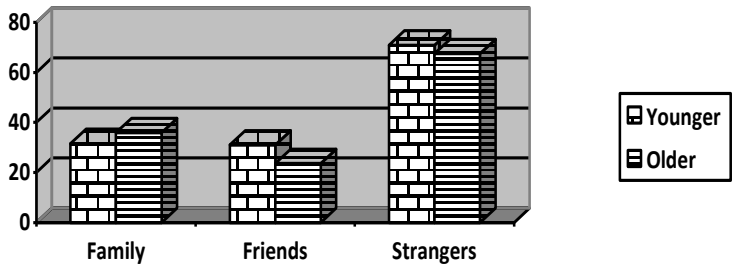
5. Results and discussion

5.1 Direct requests (A1 – C9) and honorifics

As shown in table 4 the Mood Derivable (A1), Hedged Performative (A3) and Want Statement (A5) are used in Hindi requests, and use of Performative (A2) and Obligation Statement (A4) is negligible (the percentage use is less than 1 % as shown in Table 4). As the examples in Table 3 and Table 5 suggest A3 and A5 are self-oriented where the action and need of self is highlighted. Hence these do not necessitate the use of honorifics. However, A1 has a set of verb forms that vary in the degree of politeness. The coding by CCSARP could not include the variation in the use of different verb forms with the age of the addressee even though the level of politeness used for elders is always higher than that used for younger addressees (as in Table 6).

The illustration of the attitude towards the request forms as in Table 7 shows that A1 (Mood Derivable) without honorifics is considered to be the least polite by 41 % of the respondents. Yet the responses about usage as per the responses in open-ended question show that A1 is used with strangers by nearly 70 % of the respondents as shown in Table 4. It is also used with friends and family members. The use of A1 (Mood Derivable) with younger and older addressees shows a relatively similar distribution.

Figure 1. The mean percentage of use of Mood Derivable (A1) with younger and older addressee with varying social distance.



The results on the use of Mood Derivable (A1) differ from that of the perception and attitude of the native speakers towards the use of Mood Derivable. All of these results indicate that direct requests, especially A1, reflect different levels of politeness as they are used along with honorifics. Direct requests, therefore, were not considered to be less polite because of the mitigating effect of honorifics.

5.2 (In)direct requests (A1–C9), their use and perception

As shown in Table 7, Query Preparatory (B7) is considered to be the most polite form by the native speakers. This is supported by the higher use of B7 among friends than with family members (as in Table 4), where A1 is used relatively more. However, Want Statement (A5), where only the speaker's need is articulated and the addressee is left with a better option to turn down the request, are considered to be next in the degree of politeness (in Table 7) but are not so often used as A1 (as in Table 4). Similarly, even though Suggestory Formula (B6) is perceived to be only next in the degree of politeness, these forms are not used at all with friends

and family members. In addition to this, only 3.1 % of the respondents use it with strangers. Forms like Mild Hint (C9), Performative (A2) and Obligation Statements (A4) are in the list of non-preferred forms in both use and perception. Also, the use and perception of Hedged Performative (A3) and Strong Hint (C8) is relatively less frequent.

Table 8. Use and perception of (in)direct request forms from least used to most used and least polite to most polite.

	USE	**PERCEPTION**	
Least used	C9 (Mild Hint)	A1 (Mood derivable)	Least polite
↑	A2 (Performative)	A4 (Obligation Statement)	↑
	A4 (Obligation Statement)	C9 (Mild Hint)	
	B6 (Suggestory formula)	A2 (Performative)	
	C8 (Strong Hint)	A3 (Hedged Performative)	
	A3 (Hedged Performative)	C8 (Strong Hint)	
	A5 (Want statement)	B6 (Suggestory formula)	
↓	B7 (Query preparatory)	A5 (Want statement)	↓
Most used	A1 (Mood derivable)	B7 (Query preparatory)	Most polite

The use and perception hierarchy is shown in Table 8 above. Even though a triangulation of perception and use of polite forms would be possible in a very prescriptive endeavor, their patterns of dissimilarities and similarities show that hints (C8, C9), though the most indirect forms, were neither used nor were perceived as polite. Conventionally indirect requests, although indirect, are a more preferred because they are perceived as clear, based on convention. The direct request, A1 (Mood Derivable) was the most used form but was perceived as the least polite form.

5.3 Discussion

The results in 5.2 and 5.3 show that (A1) Mood Derivable, which is the most direct request form based on CCSARP is also the most frequently used of all the nine request forms. An examination of the direct request forms in the data reveals that the use of honorifics in the direct forms serves to mitigate the requests. This accounts for the difference or lack of correlation between what the respondents believe to be polite and the forms that they actually use. The researchers argue that in order to understand the use of linguistic politeness, one needs not only to look at the directness of requests but also consider the use of honorifics, especially in languages such as Hindi that have an inbuilt system of honorification. This study highlights the need to study honorification, in cross-cultural studies on the use of speech acts and politeness in various languages.

6. Conclusion

In the present paper we started with a discussion on the theories on politeness and their relation to the illocutionary force evident in a speech act. With the discussion on the honorific forms in Hindi and the use of linguistic politeness for requesting among Hindi speakers, the paper has touched on three related issues in politeness. The paper has examined i) speakers' use of request forms in different situations based on social distance and age of the addressee, ii) their claims about the degree of politeness they would use with different categories of people and iii) their perception of politeness, based on their ranking of the given request forms.

The results show that the direct and conventionally indirect forms of request are the most frequently used. One of the reasons for the less frequent use of Hedged Performative (A3) or need statements (A5) and a greater use of Mood Derivable (A1) in the direct requests could be the inbuilt system of honorifics in Hindi, as the use of these honorific forms renders the request a polite tone even though they are direct. The use of Query Preparatory (B7) among speakers of different role-relationships also implies that clarity of requests is given more priority as far as (in)directness in the request forms is concerned. The politeness norms are fulfilled either by using honorifics when addressing the hearer or using Hedged Performative (A3) and Want Statement (A5) that refer only to

speaker's need or action. Also, the infrequent use of hints suggests that hints, though implicit and more indirect, may not be considered to be polite always. Clarity of the request is also important, this being the reason hints are not used for the request with strangers at all even when the addressee is at a greater social distance.

A further step of the research could be to test how the level of politeness varies with honorifics in languages like Hindi. The findings show that a further detailed analysis of the influence of honorifics on the (in)directness of requests can be done to see how the speakers prioritize and combine the use of honorifics with (in)directness in requesting. This would lead to a clearer understanding of the act of requesting in the languages that have an inbuilt honorific system.

The limitations of the study are that it is based on a written discourse completion test and is the constructed view of the speakers. In a way it shows what the speakers consider best to use in the situation and not the actual situation. However, it does capture the canonical forms of requesting in Hindi and the triangulation with the scale data leads to a wholesome argument.

References

Austin, John Langshaw (1962) *How to Do Things with Words*. Oxford: Claredon Press.

Blum-Kulka, Shoshana (1987) Indirectness and politeness in requests: Same or different? *Journal of Pragmatics* 11: 131–146.

Blum-Kulka, Shoshana; House, Juliane & Kasper, Gabriele (1989) Investigating Cross-Cultural Pragmatics: An Introductory Overview. In Shoshana Blum-Kulka, Juliane House & Gabriele Kasper (eds.), *Cross-Cultural Pragmatics: Requests and Apologies*, pp. 1–34. Norwood: Ablex.

Brown, Penelope & Levinson, Stephen C. (1987) *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Culpeper, Jonathan (2010) Conventionalised impoliteness formulae. *Journal of Pragmatics* 42: 3232–3245.

Fukada, Atsushi & Asato, Noriko (2004) Universal politeness theory: application to the use of Japanese Honorifics. *Journal of Pragmatics* 36: 1991–2002.

Goffman, Erving (1967) *Interactional Ritual: Essays on Face-to-face Behaviour*.. Garden City, NY: Anchor Books.

Grice, Paul (1975) Logic and Conversation. In Peter Cole & Jerry Morgan (eds.), *Syntax and Semantics 3: Speech Arts*, pp. 41–58. New York: Academy Press.

Jain, Dhanesh (1969) Verbalization of Respect in Hindi. *Anthropological Linguistics* 11: 79–97.

Kasper, Gabriele (1990) Linguistic Politeness: Current research issues. *Journal of Pragmatics* 14: 193–218.

Lakoff, Robin (1973) The logic of politeness; or minding your p's and q's. *Chicago Linguistic Society* 8: 292–305.

Leech, Geoffrey (1983) *Principles of Pragmatics*. London: Longman.

—— (2005) Politeness: Is there an east-west divide? *Journal of Foreign Languages* 28: 1–30.

Levinson, Stephen (1983) *Pragmatics*. Cambridge: Cambridge University Press.

Misra, K. (1977) *Terms of Address and Second Person Pronominal Usage in Hindi: A Sociolinguistic Study*. New Delhi: Bahri Publications.

Pandharipande, Rajeshwari (1979) Passive as an optional rule in Hindi, Marathi and Nepali. In Braj B. Kachru (ed.), *South Asian Language Analysis*, pp. 88–106. Urbana, IL: University of Illinois.

Searle, John (1979) *Speech Acts*. London: Cambridge University Press.

Srivastava Ravindra Nath & Pandit, Ira (1988) The pragmatic basis of syntactic structures and the politeness hierarchy in Hindi. *Journal of Pragmatics* 12: 185–205.

Wierzbicka, Anna (1985) Different cultures, different languages, different speech acts. *Journal of Pragmatics* 9: 145–178.

Appendix 1

The English equivalent translation of Questionnaire

Hello!
This is a PhD student from IIT Madras. For my research on 'Linguistic politeness', I need to collect this questionnaire from some Hindi speakers. All the information that you give in this questionnaire will be used only for the purpose of my research. Please fill it only according to your opinion and without consulting with others.

Thank you for your cooperation.

Nivedita Kumari

Age: ……
Gender: M/F
Occupation:

Student	Teacher	Non-teaching staff	Shopkeeper

Mother Tongue(s):

Hindi	Japanese	If any other, please specify _____

Medium of Instruction in school:

Hindi	Japanese	English	If any other, please specify.....

English known:
_____
Native of:
_____
Currently living in:
_____
Name:
Email:

I. Please fill in the blanks with the appropriate sentence that you would use in the given place, situation with the given person to make the given kind of request.

S.No.	Place	Situations	Kind of Request	Request directed to	Please state how would you request
1.	In the Market	You have done a lot of shopping and now left with less cash. You need money to go home.	Borrow money	Younger female friend	
				Older female friend	
				Younger male friend	
				Older female friend	
2.	At home	For a forthcoming event, you do not have enough cash.	Borrow money	Younger sister	
				Elder sister	
				Younger brother	
				Elder brother	
				Mother	
				Father	
3.	In a queue	The person standing in front of you in the queue has forgotten to move forward as he is reading something. You need to remind him to move ahead.	Ask to move forward	Younger female stranger	
				Older female stranger	
				Younger male stranger	
				Older female stranger	

II. Based on the role-relationship with the addressee choose the degree of politeness you would use with them and tick (√) in the appropriate box:

S.No.	Person you are requesting to	Very Polite	Polite	Casual
Example	**Person A**		√	
1.	Father			
2.	Mother			
3.	Younger sister			
4.	Elder sister			
5.	Younger brother			
6.	Elder sister			
7.	Younger male friend			
8.	Older male friend			
9.	Younger female friend			
10.	Older female friend			
11.	Younger male stranger			
12.	Older male stranger			
13.	Younger female stranger			
14.	Older female stranger			

III. You have done a lot of shopping and now left with less cash. You need money to go home. There are 9 requests given below. Please put these 9 requests on a 1–9 scale. Give 1 for the least polite request and 9 for the most polite request.

Least polite 1 ⟷ 9 most polite

- *Main ghar jaane ke liye paise chahiye.* 'I want money to go home.'
- *Kya main ghar jane ke liye paise mang sakti/sakta hun?* 'Can I ask money to go home?'
- *Dost! Mujhe ghar jane ke liye tumko paise dene parenge.* 'Friend! You will have to give me money to me to go home.'
- *Ghar jane ke liye paise do.* 'Give me money to go home.'
- *Sare paise kharch kar diye. Ab ghar kaise jaungi/jaunga?* 'I have spent all my money. How will I go home?'
- *Main ghar jane ke liye paise mangti/mangta hun.* 'I ask money to go home.'
- *Ghar jane ke liye paise milenge kya?* 'Will I get money to go home?'
- *Mujhe ghar jane ke liye paise mangne parenge.* 'I will have to ask money to go home.'
- *Tumhare paas paise honge kya?* 'Do you have cash on you?'

Thanks a lot for your cooperation.

Abbreviations

ACC = accusative
COMP = complimentizer
CONJ = conjunction
DAT = dative
FUT = future
GEN = genitive
HON = honorific
IMP = imperative
INF = infinitive
INT = interrogative
MOD = modal
NEG = negative
PERF = perfect
PL = plural
POSS = possessive
PRES = present

Contact information

Nivedita Kumari
Research Scholar
Department of Humanities and Social Sciences
Indian Institute of Technology
Madras 600036
India
e-mail: compoundverbs(at)gmail(dot)com

Dr. S. Devaki Reddy
Associate Professor
Department of Humanities and Social Sciences
Indian Institute of Technology
Madras 600036
India
e-mail: dre(at)iitm(dot)ac(dot)in

Helena Metslang

On the Case-Marking of Existential Subjects in Estonian[1]

Abstract

It has been suggested in typological descriptions that there are three kinds of factors that condition the syntactic expression of core arguments (such as case marking of subjects): referential properties, the predicate or the whole clause. This paper outlines the system of differential subject marking in Estonian existential sentences and systematises the bulk of variables, their interplay and prominence relationships. It shows that in Estonian existential subject marking, all three conditions apply but they are not equally important. Also two additional factor types count: construction type (existential clause, characterised by topicality and inclusivity effects) and other pragmatic factors. The paper suggests that the dominant variables co-defining the Estonian existential subject case are the subject's divisibility-based referential properties, the referent's situational inclusivity determination and the use of a quantifier in the subject phrase. The paper proposes a new and simpler binary division conditioning the case of Estonian divisible subjects. It relies on the distinctive pragmatic implicatures arising from situational uses.

1. Introduction

In the Baltic language area the differential subject and object marking are wide-spread phenomena and the factors conditioning them partly overlap. In the coastal Finnic, Baltic and East Slavic languages, differential subject marking (DSM) is more characteristic of the subjects with fewer prototypical properties, especially the subjects of existential clauses which

[1] I thank Tuomas Huumo, Balthasar Bickel, Alena Witzlack-Makarevich and two anonomous reviewers for their comments and feedback and Erika Matsak for her help with data processing. This study has been funded by the programme DoRa (European Social Fund / Archimedes Foundation) and the University of Tartu research projects SF0180084s08 "Morphosyntactic structure and development of Estonian" and SF0180056s08 "Language and meaning: semantics and grammar in a cognitive perspective".

brings them closer to the objects (Koptjevskaja-Tamm & Wälchli 2001: 656, 665). The examples (1) and (2) are from Estonian:

(1) *Peenra-l* *kasva-vad* *lille-d.*
flowerbed-ADE grow-3PL flower-N.PL
'There are flowers growing on the flowerbed.'[2] (Erelt et al. 1993: 14)

(2) *Peenra-l* *kasva-b* *lill-i.*
flowerbed-ADE grow-3SG flower-P.PL
'There are some flowers growing on the flowerbed.' (Erelt et al. 1993: 14)

Existential subjects (e-subjects) are not active agents and they do not function as actors of transitive propositions. Koptjevskaja-Tamm & Wälchli (2001) have described these non-canonical grammatical elements as a grey area between typical objects and subjects that permits different sub-divisions (2001: 656, 666).

The purpose of this paper is to give a comprehensive overview of the factors influencing DSM in the existential clauses (ECs) in Estonian. The paper assesses and systematises the main subject marking conditions pointed out in earlier research on Estonian (especially Nemvalts 2000), as well as closely related Finnish (Vilkuna 1992; Vähämäki 1984; Huumo 2001, 2010). The paper also simplifies the internal organisation of the varied set of factors influencing DSM and weighs the salience of each factor. In this process new facts from corpus data are interpreted and also rarer phenomena of the system are described. It is necessary to give a new account of the Estonian e-subjects' case choice factors due to the difficulty in applying the large bulk of conditions introduced so far on the analysis of real texts.

The rest of the introduction of the paper (Section 1) gives an overview of the study, of Estonian EC and the main notions relating to DSM. The main part of the work (Section 2) first proposes an account of the DSM system in Estonian and presents numeric corpus data and a flow chart of the relative ordering of DSM rules. It then outlines the order of prominence, usage frequencies and implementation principles of the subject case-marking restrictions. Section 3 summarises the account of Estonian e-subjects' case-alternation. The conclusion (Section 4) indicates

[2] See the list of abbreviations in the Appendix.

some implications of the Estonian data on the typological description of arguments' realisation.

1.1 Data and method

In this study, 279 ECs from the syntactically annotated part of the Corpus of Written Estonian (SAC) were analysed. The genre of the corpus texts is fiction: narratives about various aspects of human lives. By using the filtering options of MS Excel, first all clearly unsuitable SAC sentences (containing the object and predicative tags) were removed and thereafter the existential sentences were found manually by applying the criteria defined in Section 1.2. I call the final data set of ECs the existential clause corpus (ECC).[3] When studying less frequent phenomena, the author's native-speaker introspection was also used, as well as examples from the larger Balanced Corpus of Estonian (BCE; contains fiction, journalistic and scientific texts) and the internet.[4] These clauses are not included in the frequency charts and tables of Section 2. Sentences with coordinated subjects were included more than once because the subjects can have different case-marking conditions. Where necessary, the sample clauses have been shortened.

1.2 Defining the Estonian existential sentence

One can distinguish the following basic clause types in Estonian: the unmarked (multifunctional) clauses, existential, possessive, source-marking resultative and experiential clauses (Erelt & Metslang 2006: 254). DSM is observed in the ECs and the possessive clauses which can also be considered as a subtype of existentials (Nemvalts 2000: 45); in other clause types the subject is invariably in the nominative. ECs are not frequent: in the data, out of 2818 BCE clauses 10% were clearly ECs and 2% marginal ECs.

[3] The corpus is available upon request from the author.

[4] Despite the relative drawbacks of the use of Google in linguistic research (e.g. the representativeness, comparability and verifiability of the data), it was sometimes necessary to use this source due to the rarity of some phenomena that belong to the Estonian e-subject case-marking system.

It is difficult to give a universal definition of Estonian e-subjects, however they share some semantic, information structural and formal features, some of which are grammatical rules and some statistical preferences. **Semantically**, EC is used to present some referent in a spatial or temporal location (or the whole situation or the whole world) in order to *characterise* the location (see also Milsark 1979: 170). The situation in EC is structured from the perspective of the presupposed location (Partee & Borschev 2007: 156). In Estonian ECs, the location is usually mentioned or activated in the discourse first and then a discourse new referent, the e-subject, is presented in the situation (cf. Huumo & Perko 1993: 391 on Finnish). As has been claimed about Finnish (Helasvuo 1996: 352), and as it also shows in the ECC, the e-subject is seldom used to introduce salient discourse participants.

ECs have a statistical preference for their **information structure**: a vast majority of ECs have the locative phrase as the topic and the predicate and subject belong to the pragmatic assertion (see Section 1.3). As mentioned above, the e-subject is usually new in discourse. Among the 279 ECs in the ECC 93% of e-subjects are non-topical. A smaller subgroup of ECs has marked information structure with a subject that has been mentioned in the discourse earlier, see example (3). The corresponding sentence with the nominative pre-verbal subject, like (3b), and similarly (4), should be considered non-existential intransitive clauses: the subject is not the element being presented by the clause; the role of the clause is to characterise the subject referent, not the location, and there are no formal features of an EC (see below). In (3b) the subject is also definite. Definiteness is not a criterial feature of e-subjects but statistically indefinite e-subjects and definite non-existential intransitive subjects are extremely common (see Section 1.3). In both examples the subject is the topic of the clause – that is not a criterial e-subject feature either but still very uncommon among them.

(3) *(Kus kõik mu sokid on?)*
(Where are all my socks?)

a. *Nei-d* *on* *vannitoa-s* *ja* *magamistoa-s.*
they-P be.3 bathroom-INE and bedroom-INE
'Some of them are in the bathroom and (some of them are) in the bedroom.'
(cf. Vilkuna 1992: 53)

b. *Nee-d* *on* *vannitoa-s* *ja* *magamistoa-s.*
they-N.PL be.3 bathroom-INE and bedroom-INE
'They are in the bathroom and bedroom.' (cf. Vilkuna 1992: 53)

(4) *Põhjus* / *vahe* *on* *selle-s,* *et ...*
reason.N.SG / difference.N.SG be.3 this-INE that
'The reason / difference is that...' (BCE)

ECs share some **characteristic formal features**: the possibility of the partitive subject in the affirmative clauses and requirement in the negative clauses; the lack of number agreement on the predicate if the subject is in the partitive plural; the subject's preferred post-verbal position (XVS order). Also prototypical direct objects have the same properties, though the objects lack agreement with the verb in *every* situation. Nevertheless, the dominating view in Estonian linguistics is that the argument permitting case alternation in ECs is a subject and not an object (Nemvalts 2000: 47–48, see also Hakulinen et al. 2004 § 894 on Finnish).[5] As in Finnish, most of the Estonian intransitive verbs can serve as the predicate of an EC. In ECs the verbs' existential meaning is foregrounded and the rest of the lexical content backgrounded in this use (cf. Huumo 1999: 41). If the negative counterpart of an EC has a non-partitive subject, it is normally an unmarked clause.

Hence when evaluating whether a particular clause is an EC one has the choice of semantic, information structural and formal criteria. In this study, **when selecting clauses for the final data set**, I used the following formal and semantic criteria.

1. The subject was in the partitive (which sometimes co-occurred with the lack of agreement).
2. The function of the clause was to present some referent in a discourse (in a location or the whole situation). If the function of the clause was to say something about the location or situation, not the subject referent, I regarded it an EC. Sometimes it was necessary to use discourse context to identify this.
3. The verb had a foregrounded existential meaning.

When the first condition was fulfilled (see (5)) I did not look at the criteria 2 and 3. The latter were used with nominative subjects, see example (6).

[5] Recent research on Finnish (Helasvuo & Huumo 2010) suggests to separate this argument from both subject and object and call it the e-NP (existential NP).

(5) *Pole pääsu.*
be.NEG escape.P
'There is no other way out.' Lit. 'There is no escape.'

(6) *Kui jõud-sid kätte suur-ima väsimuse hetke-d, ...*
when arrive-PST.3PL at.hand great-SUP.G tiredness.G moment-N.PL
'When the moments of greatest tiredness arrived, ...' (ECC)

Possessive clauses were also included in the corpus (adessive NP + 'be' + e-subject).

(7) *Su-l on kogemus-i.*
you-ADE be.3 experience-P.PL
'You have some experience.' Lit. 'You have some experiences.' (ECC)

Properties like word order, definiteness, givenness, discourse salience and the partitive in the corresponding negative clause were not used for identifying ECs as they rather represent statistical tendencies. See also Section 2.2.2 on using the criteria for determining e-subjects.

To describe the distribution of **DSM** in Estonian it is necessary to recognise two main types of e-subjects: NP subjects and quantifier phrase (QP, see Section 2.4.2.2) subjects. NP subjects have as a head singular count nouns (nominative), plural count nouns (nominative and partitive), singular mass nouns (nominative and partitive) and singular nouns whose categorisation is unclear (nominative).[6] The subjects' number and case-marking are also influenced by an active recategorisation process between these types, compare example (7) with (8).[7]

(8) *Su-l on kogemus-t.*
you-ADE be.3 experience-P.SG
'You have some experience.'

[6] Although it is semantically hard to draw a line between discrete and indiscrete objects, different languages are thought do this through the means of grammar (cf. Lyons 1977: 42). However, in the case of this noun group Estonian grammar does not seem to make this distinction, see Section 2.3.1.

[7] It has been noted that 22% of Estonian simple nouns are polysemous (Langemets 2009: 5); therefore different meanings of the same lexeme play a considerable role in the Estonian DSM and differential object marking.

In example (7) the plural indicates that *kogemus* 'experience' is a count noun while in example (8) the singular partitive indicates that it is a mass noun, see also Section 1.3. In the affirmative clauses, the case-marking of Estonian e-subjects depends on semantic, syntactic and pragmatic factors. The following list contains the **factors in the order of occurrence frequency** (based on the number of subjects in ECC whose case is determined by the factor):

1. the subject noun's lexical semantics (including countability and quantitative definiteness);
2. inclusivity of the subject referent's quantity in the situation;
3. the lack of inclusivity determination on divisible e-subjects in the situation;
4. the predicate verb's lexical semantics or the construction as a whole.

In most cases these factors overlap with each other. The list above indicates the dominant and decisive factors upon which the e-subject case depends in the ECC. Most of these factors can trigger both the nominative and the partitive case-marking – this will be shown in the rest of this paper.

1.3 Relevant notions

Contemporary typological approaches to **differential subject marking** tend to be rather wide and involve different layers of language: "In a broad sense, a language may be said to have DSM if some subjects have a different [c]ase, agree differently, or occur in a different position than others." (Woolford 2009). DSM is a split in subject-marking that is caused by referential, predicate-related, clausal, pragmatic, morphological or phonological factors (Dixon 1994: 70–110; Witzlack-Makarewich 2010: 65–157; Woolford 2009). In the literature the term DSM (also non-canonical subject marking) has been used with several kinds of splits in the marking of transitive and intransitive subjects in both dominatingly accusative and ergative languages. This approach allows the incorporation of indexing and discrimination, split and fluid intransitivity and several other phenomena. Narrower approaches restrain this notion to the marking of semantically lower subjects (Aissen 2003) or marking caused by subject features alone (Woolford 2009: 17) or to the typologically common splits between different lexical predicate groups. For this study of Estonian e-subjects narrower approaches are not suitable as e-subjects' differential case-marking is caused by referential properties of the NP, semantics of the predicate verb or the whole construction (in the sense of Goldberg 1992),

as well as by clausal and pragmatic properties. In this paper I will only look at the case-marking split of e-subjects and will not address the differences between the e-subject and the intransitive subject.

Divisibility plays a major role in the e-subject marking in both Estonian (Nemvalts 2000: 60) and Finnish (Csirmaz 2005; Vilkuna 1992; Vähämäki 1984: 404). Divisibility can be described as follows: if x is part of gold and x' is part of x then also x' is part of gold unless we have information that x' is not part of gold (Krifka 1989:41). It separates both mass nouns and plural count nouns from singular count nouns (Krifka 1989: 39–41). In the position of e-subject (and elsewhere) the divisibles, i.e. mass nouns and plural count nouns, have the same case-marking motivations (Nemvalts 2000: 71, 104, 147–148). These nouns are also characterised (in the case of Finnish, Vilkuna 1992: 39–41 and in the case of Estonian, Rajandi & Metslang 1979: 11–12) by the following properties:

– they are additive in the sense that their number marking does not change when you add to them something that belongs to the same category (adding a portion of *sand* to another portion of *sand* still gives *sand* as a result, adding some *boys* to *boys* still gives a *boys* as a result);
– their referent does not have inherent shape (the word *sand* does not have the feature [Shape] in its lexicogrammatical meaning, sand's shape depends on its vessel or location; also the word form *boys* lacks the feature [Shape], the referent group of boys can stand in a row or be randomly located);
– as e-subjects, they can occur in the partitive in the affirmative;
– they combine with the quantifier *palju* 'a lot' but this combination procedure does not change their original number marking (*liiv* 'sand.N.SG' + *palju* 'a lot' > *palju liiva* 'a lot of sand.P.SG', *poisid* 'boy.N.PL' + *palju* > *palju poisse* 'a lot of boy.P.PL'). This is not true in the case of singular count nouns that, in combination with *palju*, need to be used in the plural and not in the singular (*poiss* 'boy.N.SG' + *palju*> *palju poisse* 'a lot of boy.P.PL' (**palju poissi* 'a lot of boy.P.SG'));
– they do not occur directly with numerals (**kaks liiva* 'two sand.P.SG', **kaks poisi-d* 'two boy-N.PL').[8]

Definiteness is a category that includes the interplay of the following factors: identifiability, including familiarity, on the one hand (quality-related notions) and inclusiveness, including uniqueness, on the other

[8] In special cases the plural count nouns can be quantified by a numeral with plural marking but then they take the meaning 'plurality of bounded sets': *viie-d teatripileti-d* [five-N.PL theatre.ticket-N.PL] 'five sets of theatre tickets' (cf. Vilkuna 1992). Sometimes mass nouns can take the plural but then they obtain the meaning 'different kinds'.

(quantity-related notions) (Lyons 1999: 2–13). Lyons brings the following examples. *Put these clean towels in **the bathroom** please* (familiarity: *the bathroom* is definite because it is known from the immediate situational context). *They've just got in from New York. **The plane** was five hours late* (identifiability: the referent is definite through the association with the situation). *Beware of **the dog*** (uniqueness: the dog is definite in the context when the proposition can be found on a warning sign and the passer-by has actually never seen the specific dog mentioned. The sign says that there is only one dog in the vicinity and the reader is being warned against that unique dog). *Beware of **the dogs*** (inclusiveness: *the dogs* is definite because it refers to all the dogs, i.e. inclusive amount relevant in this context. The dogs are definite even if the reader has never seen the ones mentioned on the warning sign).

Lyons (id.: 3–11) describes the semantics of definiteness as follows. In the case of familiar NPs the referent is definite because the speaker presents it as familiar to both the speaker and hearer. If an NP is definite due to identifiability, it is because the speaker signals that the hearer is in the position to identify the referent (knows it or is able to work it out). In the case of uniqueness the definite NP signals that "there is just one entity satisfying the description used", relative to the particular context. If an NP is definite due to inclusiveness, the reference is to the totality of the objects or mass in the context which satisfies the description.

The more basic and comprehensive category of qualitative definiteness is identifiability that also embraces the expressions that are definite due to familiarity; the more basic category of quantitative definiteness is inclusiveness that also involves the expressions that are definite due to uniqueness (Lyons 1999: 13). Givón shows that the most typical environment for definite NPs is in referring expressions of factual, realis contexts. "Definite" may be viewed as a further sub-specification of "referring" (2001: 441–442).

Hawkins' speech acts based definiteness theory called *location theory* (1978) brings all the aforementioned aspects of definiteness together. In definite reference the speaker introduces a referent to the hearer, locates the referent in some shared set of objects and refers to the totality of the objects or mass within this set (for denoting these actions, Vilkuna (1992: 16) uses the term 'location instruction' in accordance with Hawkins' theory). An expression is definite if its (potential) referent can be **uniquely located in the listener's discourse model** of the moment (it has to have a location instruction in the context or the interlocutors' knowledge, e.g. in the earlier

discourse). In indefinite reference the speaker just introduces a referent to the hearer and refers to a proper subset of the referring expression (Hawkins 1978: 167, 187).

ECs' subject realisation is better characterised by quantification than definiteness (Milsark 1979: 196–208). Also **the Estonian DSM largely depends on inclusiveness** and not on identifiability (as does Finnish DSM; Hakulinen et al. 2004: §1241).[9] In ECs, inclusive quantity, i.e. universal quantification over a set (Milsark 1979: 204), is marked with the nominative:

(9) *Puu-lt* *lange-sid* *lehe-d.*
tree-ABL fall-PST.3PL leaf-**N**.PL
'(All) the leaves fell down from the tree.' (Nemvalts 2000: 126)

Often the partitive NP stands for a part or sub-quantity of a specific, potentially bigger entity that can exist – a **contextual boundary,** location instruction. The phenomenon can be explained in terms of the relationship between two embedded entities or sets. In the same way, as a bounded larger (or standard) set for *some chess pieces* is the whole *chess set* (see also Koptjevskaja-Tamm & Wälchli 2001: 665). However, in the data, the sentences having a bounded larger set is rather an exception than a rule: often this bigger entity is unclear, unspecified in the context (Koptjevskaja-Tamm 2001: 525):

(10) *Õue-s* *mängi-b* *laps-i.*
outside-INE play-3SG child-P.PL
'There are some children playing outside.' (translated from Vilkuna: 1992: 47)

The sentence claims a non-inclusive amount of children to be playing outside but it does not refer to the existence of any major group of children which the ones who are playing outside may be part of. If the NP referent is non-inclusive, i.e. it has indefinite quantity, it does not necessarily mean that the quantity in question is smaller than the total quantity (cf. Vilkuna 1992: 46). Therefore quantitatively indefinite NP means *some quantity* and not necessarily *partial quantity* (see more on inclusivity in Section 2.4.2.).

The definition of ECs depends on **information structure**. According to Lambrecht (1994: 52, 206–219, 335) propositions consist of the

[9] Usually neither nominative nor partitive e-subjects are in Lyons' sense identifiable.

following parts: pragmatic presupposition, pragmatic assertion, topic and focus. Pragmatic *presupposition* is the set of propositions lexicogrammatically evoked in a sentence which the speaker assumes the hearer already knows or is ready to take for granted at the time the sentence is uttered. Pragmatic *assertion* is the proportion of the sentence which the hearer is expected to know or believe or take for granted as a result of hearing the sentence uttered. The *topic* relation is the relation of aboutness between a proposition and a discourse entity. It is the matter of current interest with respect to which a proposition is to be interpreted as relevant. It is part of the presupposition. *Focus* is a pragmatic relation that is part of pragmatic assertion, an element whereby the presupposition and assertion differ from each other, the unpredictable and unrecoverable element that makes an utterance into an assertion. The novelty of the focus is in the fact that a particular denotatum is chosen as a particular semantic relation. For example, in the sentence (*– What is growing in the flowerbed?*) *– There are flowers growing in the flowerbed* (see example (1)), the presupposition is *There is X growing in the flowerbed*, the topic *is in the flowerbed*, the assertion is *what grows in the flowerbed is flowers* and the focus is *X=flowers.* Topic is an obligatory part of every sentence although sometimes it is implicit and not overtly expressed, e.g. time and place (Erteschik-Shir 2007: 13–16).

2. Conditions of subject case alternation in the existential clause

Cross-linguistically, non-canonical argument realisation (non-canonical case, agreement and syntactic behaviour) often depends on semantic features, like for example volitionality or stativity (Onishi 2001: 23–40). Onishi demonstrates that any such feature can (sometimes simultaneously) be bound to different levels of language: the predicate's lexical meaning or one of its sub-meanings, verbal affixes, choice of auxiliary, etc.

Estonian DSM is a complex hierarchical system of case motivations and the main case-assignment factor that underlies all the levels is quantitative definiteness, i.e. inclusivity. In Estonian ECs there is often a mix of different competing motivations for e-subject's case-choice that can potentially play a role in determining the case of an e-subject. For example, certain nouns occur in the nominative as e-subjects (existential nominatives) but negation causes partitive marking. One can ask which factor is dominant and overrules the other one in this use. Will the sentence

get a nominative or partitive marking? In the next pair of examples, only the partitive is grammatical.

(11) *Kassi vaate-s ei ol-nud mingi-t märguanne-t.*
cat.G look-INE NEG be-PST.PTCP any-P signalling-P
'There was no signalling (signal) in the cat's look.'

(12) **Kassi vaate-s ei ol-nud mingi märguanne.*
cat.G look-INE NEG be-PST.PTCP any.N signalling.N
Intended: 'There was no signalling (signal) in the cat's look.'

It is evident in the data that e-subject's case factors form a layered system of dominance where each dominating factor applies to a certain sub-part of the data (determines the e-subject's case in certain ECs). In some contexts it is also impossible for the nominative and partitive case to express some meanings that they convey in other contexts (for example in the negative clauses the partitive cannot express non-inclusive quantity as it can in affirmative clauses, see example (3) above and Rajandi & Metslang (1979: 3) on a similar issue with Estonian direct objects).

I divide the process of subject case assignment in ECs in **four levels**: A (the highest level: see Section 2.1), B (Section 2.2), C (Section 2.3) and D level (the lowest level; Section 2.4). The factors on the higher levels are the dominant ones that override the levels closer to the bottom: although a particular e-subject may for example have properties relevant for both A and B level factors, then according to this approach, its case is governed by the A level factor. The factors within each level are equal to each other. The prominence order of the factors affecting the subject marking of Estonian ECs, as suggested in this paper, is presented in Table 1. It only involves the decisive factors that primarily influence the e-subject's case.

Table 1. The prominence order of subject marking factors.

Case assignment level	**Case factor type**	**Pre-conditions of the case factor type**	**Case factors**
A	**Polarity**	ECs	A1 Negation
B	**Clausal construction**	Affirmative ECs	B1 Clause level constructions with a partitive e-subject

			B2 Clause level constructions with a nominative e-subject
C	**Head noun semantics**	Affirmative ECs with clausal contructions permitting both Nom and Part	C1 The noun belongs to the group "Existential nominatives" (singular count nouns, set nouns, some abstract nouns, *pluralia tantum*) C2 The noun belongs to the group "Existential partitives" (some abstract nouns)
D	**Situational inclusivity of divisible nouns**	Affirmative ECs with clausal contructions permitting both Nom and Part, the e-subject's head noun is a divisible	D1 Subject case alternation is based on the opposition of the presence or lack of inclusivity meaning (PLI) D2 Subject case alternation depends on the opposition of inclusive-non-inclusive meaning (IN)

In addition to these factors the paper discusses a few other potential case-choice conditions that appear not to have primary influence, for example tense and aspect related case-marking distinctions just co-occur with the inclusive-non-inclusive semantics opposition.

The reader might ask whether this view really holds that there is a hierarchy between the case factor levels. Table 2 shows that the higher level factors dominate over the lower level ones in the corpus. In the table, the grey boxes indicate the number of e-subjects whose case is determined by the *dominating* case factor. For example, 8 e-subjects in ECC have their case determined by nominative e-subject constructions. The non-coloured boxes indicate the other, simultaneously competing case motivations that *are overridden* by the dominating factors in the corpus. These are the potential alternative conditions determining e-subjects' case-marking.

Table 2. Frequencies of dominating (grey) and non-dominating (uncoloured) case-choice factors present among the existential clause corpus e-subjects.

Case factor type	**Case factors**	**A1**		**B1**		**B2**		**C1**		**C2**		**D1.1**		**D1.2**		**D2.1**		**D2.2**	
		N	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**
A Polarity	A1 Negation	**1**	**38**																
B Clausal construction	B1 Partitive e-subject constructions	0	2	**0**	**2**														
	B2 Nominative e-subject constructions	0	1	-	-	**8**	**0**												
C Head noun semantics	C1 Existential nominatives	0	18	0	0	6	0	**103**	**0**										
	C2 Existential partitives	0	9	0	1	0	0	-	-	**0**	**20**								
D Inclusivity of divisible nouns	D1.1 Subject case marks non-inclusivity (PLI; Part)	0	-	0	1	0	0	-	-	-	-	**0**	**31**						
	D1.2 Subject case depends on the lack of inclusivity meaning (PLI; Nom)	1	-	0	0	1	0	-	-	-	-	-	-	**41**	**0**				
	D2.1 Subject case marks non-inclusivity (IN; Part)	0	-	0	0	0	0	-	-	-	-	-	-	-	-	**0**	**4**		
	D2.2 Subject case marks inclusivity (IN; Nom)	0	-	0	0	1	0	-	-	-	-	-	-	-	-	-	-	**31**	**0**
	D Divisibles in total	1	11	0	1	2	0	-	-	-	-	**0**	**31**	41	0	0	4	31	0

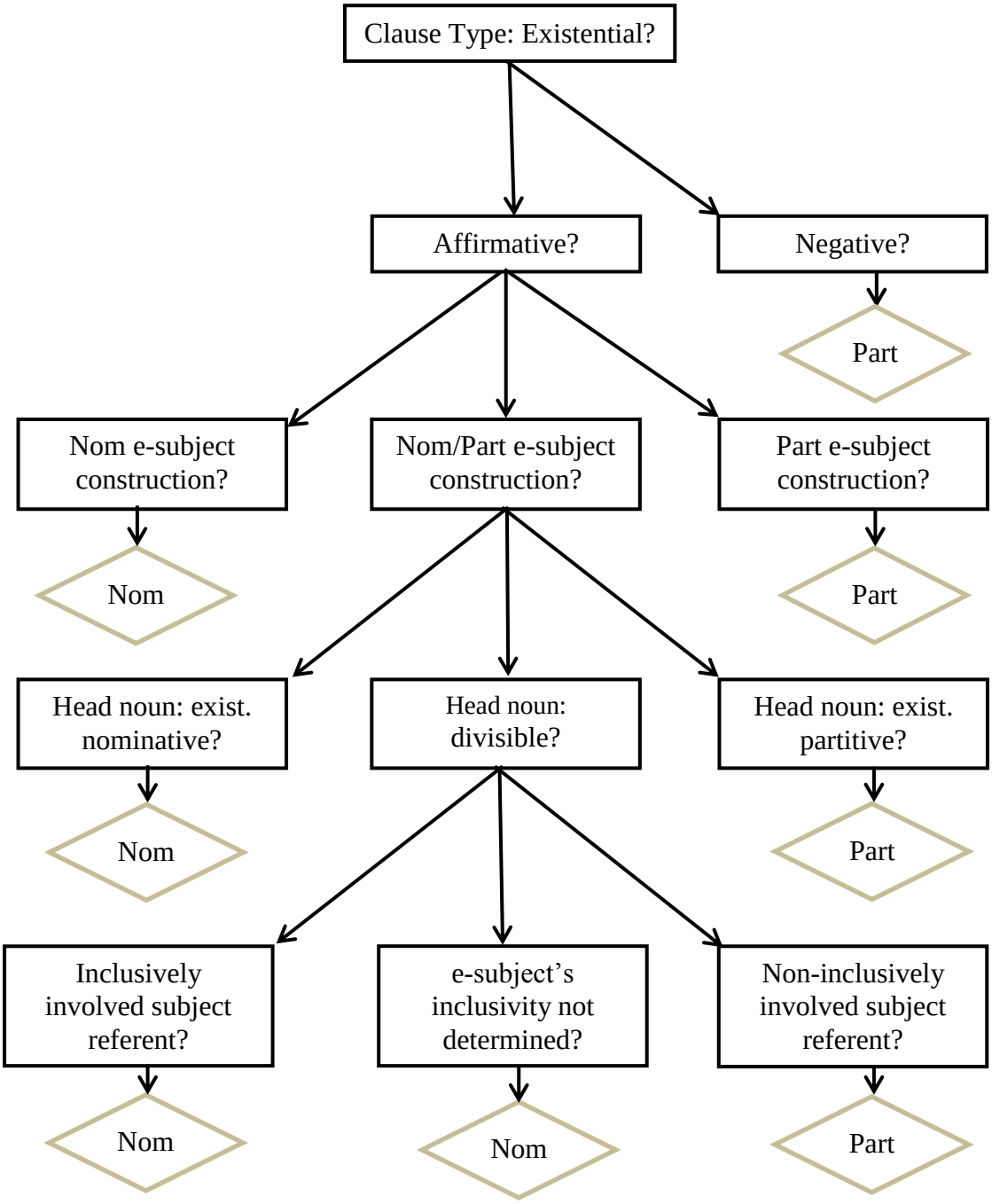
If you look at the columns in the table then it is always the top cells of each column that are grey. This shows that the factors in the top rows decide the case-choice of those columns' e-subjects. For example, 18 negative ECs have an e-subject that belongs to the lexical noun group Existential nominatives (factor C1). Although the e-subject's head noun's type is also a case factor in the hierarchy, it gets overriden by the negation factor (A1). Although the e-subject's head noun's type is also a case factor in the hierarchy, it gets overriden by the negation factor. The numbers in the grey boxes add up to 279 – the amount of all e-subjects in ECC. The numbers in the non-coloured boxes do not add up to the numbers in the grey boxes because there are some incompatibilities between factors and also additional overlappings among them. For example, in negative ECs there are 2 clauses containing a partitive e-subject construction and at the same time they have existential partitive e-subjects (all together there are three competing case conditions in these clauses).

The two leftmost columns of the table are negative clauses, the rest of the table depicts affirmative clauses. Some boxes have not been filled (marked with a "-") again due to incompatibility of some factors. For example, it is not possible to analyse the negative clauses' referents' inclusivity because negation in ECs negates the whole situation (Nemvalts 2000: 163) and the question of e-subject referent's inclusivity is not relevant here. Levels C and D are mutually exclusive. If an NP is not a divisible but has an existential nominative or existential partitive head noun, it's case cannot express PLI or IN. Also conditions C1 and C2 are mutually exclusive, as well as D1.1, D1.2, D2.1 and D2.2 with each other. The subjects of level D are divisibles, see Section 2.4 for their distribution.

To better demonstrate the relationships in the hierarchy, a flowchart is used (Figure 1) which is an effective tool for analysing and documenting complex processes and dependencies. The figure depicts the order of the factors that should be followed while making decisions about an e-subject's case-determining factors in particular sentences. It allows moving from the level A to level D and, in this process, discarding unsuitable factors one by one until we have reached the matching case-assignment factor. Flowcharts have been used before for illustrating the hierarchical nature of differential object marking in Estonian (e.g. Rajandi & Metslang 1979: 14) and Finnish (Vilkuna 1996: 119). As it is widely known, differential subject and object marking largely depend on similar phenomena in Estonian. The building blocks of the case-formation chart of the present study are partly similar to

Rajandi and Metslang (predicate-related and referential properties), however, the structure is adapted for ECs and new nodes are added.

Figure 1. Factors of differential e-subject marking in Estonian.



The chart shows for each rule which subject case it determines or which further restriction it needs. According to this treatment, negative e-subjects

are first assigned partitive marking (level A of subject case assignment, see Table 1). On the next levels, as mentioned above, only affirmative clauses are under consideration. If the EC contains a clausal construction that requires either a nominative or a partitive e-subject, its subject gets its respective case-marking on level B. One should move a level down (to level C) if the clausal construction permits the use of both the nominative and the partitive. Then the analysis should pay attention to the referential properties of the e-subject's head noun. If the head noun under question either belongs to the group "Existential nominatives" or "Existential partitives" it will get its case accordingly. If the noun belongs to neither group, it is a divisible noun and hence receives its case-marking on level D. Divisible nouns are marked by the nominative if their referents participate in the situation inclusively (IN), by the partitive if their referents participate in the situation non-inclusively (IN, PLI). If the referent is neutral with respect to inclusivity, it gets the nominative case (PLI). In the following chapters, all conditions will be outlined in detail.

2.1 Level A of subject case assignment

Negation is the strongest e-subject case factor. If the subject NP is in the scope of negation, it takes the partitive, regardless of its noun type and context, see example (5) above. As the exception to the rule the e-subject gets a nominative marking in negative ECs when it is out of the scope of negation, when it occurs in a contrasted sentence or when it is presupposed. Of the latter options, the ECC only contains one nominative e-subject that occurs with negation due to presupposition:

(13) *Maa-s* *ei* *ol-nud* *mitte* ***rohi***, *vaid* *Ø* *muld.*[10]
ground-INE not be-PST.PTCP not grass.N.SG but be.PST.3SG soil.N.SG
'There was not grass but (was) soil on the ground.' (ECC)

Unlike in prototypical ECs, the subject in this example is not given but still accessible in the discourse and the focus of the clause. It is accented due to contrast.

Presuppositions of referents are usually either created in the previous discourse or the speaker believes the listener to have them. Direct

[10] The verb of such adversative coordination clause can either be in the affirmative or in the negative – neither the meaning of the clause nor the subject case is changed by that.

presupposition of an e-subject's referent in a given location can be unspecific or contrasting (something is previously said or expected to exist in a particular location – that is actually not there). The EC specifies the expected referent, or in the case of incorrectly expected referent, replaces it. The nominative e-subject in the example is a divisible noun neutral from the point of view of inclusivity. Therefore one could doubt whether it really is presupposition that brings about the nominative or perhaps the nominative case is determined by PLI. Nevertheless, the latter is unlikely: both Nemvalts (1996: 43) and Erelt et al. (1993: 42, 196) suggest that presupposed subjects occur more naturally with nominative marking. The environments that normally require the partitive subject (negation) loosen this requirement in the case of a presupposed subject (Erelt et al. 1993: 45, 158–159). Therefore one could also suggest that in the factors ordering schema, above level A, Negation, there is another level, Direct presupposition. I prefer to treat this phenomenon as an exception and have not included this in the hierarchy because it is very rare: presupposition does in general not go well with the role of the EC.

Vilkuna (1992: 94–95) has demonstrated in Finnish that unrealised contexts (where the event or situation is only hypothetical: future, conditional, doubt, negation, etc. as opposed to realised affirmative contexts) can affect argument interpretation and marking. However, in ECC, only negation is affecting the e-subject case whereas the interrogative and hypothetical contexts usually follow the case-marking rules of affirmative e-subjects, see Table 3. Only 4 interrogative sentences have a content that triggers the use of the partitive: the e-subject referent is either negated or evaluated as highly questionable by the speaker.

Table 3. Negative and other clauses with unrealised contexts in ECC.

	Negative clause	**Interrogative clause**	**Other unrealised situation**
Nominative	1	12	23
Partitive	38	4	10

2.2 Level B of subject case assignment

On level A, the subjects of all negative ECs receive their case-marking. From here on, on the B, C and D level, only the subjects of affirmative ECs will be discussed.

The possibilities of the existential interpretation of a clause and non-canonical (partitive) subject marking have been claimed to depend on both the verb and the construction in Estonian (Rätsep 1978; Nemvalts 2000). The same question is widely discussed in studies on Construction Grammar (e.g. Goldberg 1992, 1995, 1997). Different uses of a verb and combinations of argument realisation are claimed to be attributable to the construction (i.e. independent argument structure) itself (Goldberg 1995: 19) or both to the construction and verbal polysemy (Nemoto 2001: 119–133). According to construction grammar, the term 'construction' has a very wide meaning and it involves primitive grammatical units that are pairings of form and meaning which may be atomic or complex, schematic or substantive (Croft 2001). The inventory of constructions in a language varies by degree of schematicity and the taxonomic links or relationships between them (e.g. case and agreement constructions, lexical items, idioms, control constructions, non-finite constructions, subcategorisation frames, word order and sentence type constructions; e.g. Bickel 2010; Croft 2001). However, in the context of this argument realisation discussion, mostly only clause level constructions are relevant.

Section 2.2 analyses the Estonian e-subject's case-marking from the constructions' point of view. On the basis of the existent research by Rätsep and Nemvalts, I will focus on the issue from a more form-related viewpoint and, instead of semantic argument structures, I look at formally determined constructions (however these constructions are not independent of the verb's lexical semantic features (cf. Rätsep 1978: 258)). I will show in Section 2.2.4 that Estonian e-subjects enter into nominative and partitive case-frames sometimes due to the verb's and sometimes due to the whole construction's influence. First, Sections 2.2.1–2.2.3 categorise these constructions according to the subject case(s) found in them.

2.2.1 Constructions with the partitive e-subject

Earlier, Estonian e-subject marking has been described throughout specific constructions where the lexical predicates determine the subject case (Rätsep 1978; Nemvalts 2000). Both sources call such constructions *verb-governed sentence patterns*. Rätsep's (1978) verb-governed sentence pattern (also called *formal syntactic structures* (Talmy 2000: 265)) is a generalised abstraction that unites a set of grammatically similar simple

sentences that share the number, form and order of the verb's arguments and obliques.[11] Rätsep attempts to provide a comprehensive list of lexical predicates involved in each of the 1272 (existential and non-existential) constructions. The sentence patterns have no direct connection with clause types, like experiencer clauses or existential clauses – in the sense of construction grammar, they are more specific than the highly schematic clause types.

An example from Rätsep's monograph on Estonian simple clauses (1978) is the construction that has the obligatory elements of a noun in the partitive case, a predicate verb and the optional oblique – a PP (with the meaning 'for someone/something') and it only permits one argument in a grammatical case – in the partitive:

(14) NOUN$^{\text{PARTITIVE}}$ VERB (NOUN$^{\text{GENITIVE}}$ +jaoks).[12]
Klaasi *piisa-b* *tööstuse* *jaoks.*
glass.P.SG suffice-3SG industry.G for
'There is enough glass for the industry.' Lit. 'Glass suffices for industry.' (Rätsep 1978: 106, pattern 7.2)

The construction only occurs with three verbs: *jätkuma*, *piisama*, *jaguma* which all have the (sub) meaning 'suffice' and all the clauses in it are ECs. In Rätsep's collection the submeanings (in the sense of Langemets et al. 2009) of these verbs also occur in constructions with a nominative subject. For example:

(15) **N$^{\text{N}}$** V (DE) (DT) (DI)[13]
Põõsas-te *rida* *jätku-s* *laua* *juurest* *trepi* *juurde.*
bush-G.PL line.N.SG continue-PST.3SG table.G.SG from staircase.G.SG to
'The line of bushes continued from the table to the staircase.' (Rätsep 1978: 97, pattern 2.2.102)

[11] Rätsep's work decribes non-contextual constructed sentences. The main problems that appeared when applying these constructions in the analysis of contextual ECs in ECC were emphasis and text coherence related word order divergences. If the reason for different word order was detectable I regarded the corpus sentence matching the pattern.
[12] *jaoks* 'for' – a postposition requiring a genitive complement.
[13] *DE* – 'extralocal directionals' (e.g. the elative, the ablative, various PPs and adverbs with the meaning 'from'), *DT* – 'translocal directional' (path semantics, usually marked by PPs), *DI* – 'intralocal directional' (a place where some action is directed, marked by e.g. illative, allative and PPs; id.: 44–46).

Nemvalts (2000) has identified a number of constructions that involve ECs from Rätsep's collection. After complementing his selection on the basis of ECC I have in the **EC constructions list:**

> – 19 constructions that permit just one subject case (12 partitive[14] and 7 nominative subject EC constructions), see the examples in this section and the next one;
> – 76 constructions that allow the alternation of nominative and partitive subject marking; see the examples in Section 2.2.3 and onwards.[15]

These constructions are tied to different verbs that can re-occur throughout them and that can also occur in non-existential constructions. Verbs that can occur with partitive subject constructions but also in other constructions include *olema* 'be', *sadama* 'come down, rain', *sisalduma* 'be in', *immitsema* 'seep, exude', etc. I am not aware of verbs that can only occur in constructions that exclusively permit partitive subjects.[16]

An example of how the requirements of the constructions **override the other subject case-marking conditions** is the obligatory partitive case-marking of singular count noun subjects, see example (16). Under all other circumstances, singular count nouns take the nominative in affirmative ECs (see also Section 2.3.1).

(16) $\underline{N^P\ V\ DI}$. Example with the verb *jätkuma* 'suffice'.
Meistri-t *jätku-s* *iga-le poole.*
master-P suffice-PST.3SG everywhere.ALL
'The master could help out everywhere.' Lit. 'Master sufficed (was) everywhere.'
(Rätsep 1978: 154, pattern 114)

In the ECC there are no occurrences of singular count nouns taking the nominative in the partitive e-subject constructions. ECC only contains two subjects whose partitive case is determined by a construction. The other example is (17), nominative would be ungrammatical here.

[14] Rätsep's patterns 1.3, 7.1, 7.2, 7.3, 8.0, 29, 40, 79, 109, 114, and 287.
[15] The full list is available upon request from the author.
[16] The verb *piisama* is exceptional: it only takes a partitive e-subject or an oblique argument but not a nominative subject.

(17) Construction $\underline{N^{P}\ V\ (N^{G}\ \textit{jaoks})}$
Lauda-s looma jaoks ruumi jätku-b.
shed-INE animal for space.P.SG suffice-3SG
'There is enough space for the animal in the shed.' Lit. 'The space in the shed suffices for the animal.'[17] (ECC) (Rätsep's pattern 7.2)

2.2.2 Constructions with the nominative e-subject

In the list of EC constructions used for this analysis, there are eight that only allow nominative e-subjects.[18] They involve verbs like for example *koitma* 'dawn' and *valitsema* 'rule'. These constructions often have **properties of both EC and non-existential intransitive sentence**, for example:

(18) $\underline{(N^{ADE})\ V\ N^{N}.}$ Suitable with the verbs *algama* 'start', *hakkama* 'begin', *käima* 'be going on'
(Rukkipõld lõppes ja)
juba alga-s=ki luht.
already start-PST.3SG=CL water.meadow.N
'(The rye field ended and (as we walked on)) the water meadow already started.' (ECC) (Rätsep's pattern 3.2)

(19) $\underline{(DE)\ V\ N^{N}.}$ Suitable with the verbs *liginema, lähenema* 'approach'
(Aeglaselt jalutades lähenes Rasmus tänavanurgale.)
Sama-l aja-l lähene-s ristmiku poole üks
same-ADE time-ADE approach-PST.3SG junction.G towards one.N
elusolend.
living.being.**N**.SG
'(Rasmus neared the corner at a slow pace.) At the same time a living being was approaching the junction.' (ECC) (Rätsep's pattern 3.3)

Although these clauses are similar to unmarked intransitive clauses they correspond to the following EC criteria. The role of the clause is rather to characterise the situation or the locative phrase referent, than the subject referent, compare (19) with the non-existential (20). Secondly, the

[17] This particular example does not clearly show that the subject case depends on the pattern: *ruumi* 'space.P' is an existential partitive noun (Section 2.3.2.). However, in BCE there occur many examples with, for instance, normal mass nouns: *Töö-d jätku-s* [work-P be.enough-PST.3SG] 'There was enough work.'
[18] Rätsep's patterns 3, 3.1, 3.2, 3.3, 3.4, 25, and 107.

foregrounded lexical meaning of this construction's predicate verbs, including the ones exemplified above, is either existence (*käima*, submeaning 'be going on') or coming to existence (*algama* 'start', *hakkama* 'begin', *koitma* 'dawn', *lähenema* 'approach'). The remaining lexical content of these verbs is insignificant for the general purpose of these clauses (e.g. the meaning of volitional activity in the case of *käima*). The clauses in these eight constructions also have other prototypical EC properties: the sentence introduces a new entity in the discourse and the subject's post-verbal position is unmarked and neutral. It is not caused by emphasis or contrast.

As the predicate verb is not 'be' in these constructions, partitive subjects would not feel natural in the negative counterparts of these clauses. This is also common among ECs with less typical existential predicates (see also Huumo 1999 on Finnish).

What is mainly common between these sentences and non-existential intransitives as in (20) is that in the affirmative clause, the subject can only occur in the nominative.

(20) *Uurija* *lähene-s* *küsimuse-le* *oskuslikult.*
researcher.N.SG approach-PST.3SG question-ALL skilfully
'The researcher approached the question skilfully.' (Rätsep 1978: 185)

On the basis of the properties listed above, the clauses like (18) and (19) should be classified as ECs (see also Partee & Borschev 2007 on a properties based approach in distinguishing Russian ECs from locative sentences).

In ECC there are 6 clauses with 8 subjects that belong to 'nominative subject only' constructions, see for example (6), (21) and (22).

(21) *Elektriliini* *all* *vii-s* *purre* *üle* *jõe.*
power.line.G under lead-PST.3SG foot.bridge.N.SG over river.G
'Under the power line there was a foot-bridge that goes across the river.' Lit. 'Under the power line a foot-bridge lead across river.'[19] (ECC)

[19] The verb *viima* 'take, lead' is prototypically transitive but it is used intransitively if the sub-meaning is 'point somewhere, be located towards a direction, enable to go somewhere' (source: Kaalep & Muischnek 2002).

(22) *Sealt* *vaata-b* *vastu* *mu* *mehe* *nägu.*
there.ABL look-3SG back my husband.G face.N.SG
'My husband's face is looking back from there.' (ECC)

Unfortunately, there are only 2 clear instances where the subject case-marking is determined by the construction, e.g. (6). They have divisible referents that can in other constructions also occur in the partitive (see Section 2.3.3). However, in this construction partitive would be ungrammatical. In constructions like (22) the semantics of the predicate has bleached: the original volitional meaning has given ground to a mere existential one. This can be seen especially well in the next example.

(23) *Köögi-st* *vaata-s* *vastu* *segadus.*
kitchen-ELA look-PST.3SG back mess.N.SG
'There was a mess in the kitchen.' Lit. 'There was a mess facing (me) in the kitchen.' (BCE)

2.2.3 Constructions with the nominative/partitive e-subject

So far, only these level B factors (constructions) have been discussed where the e-subject's case is specifically determined. However, 76 out of the 95 EC constructions permit both subject cases and therefore their DSM has to be explained by other, lower level factors. They contain a vast number of verbs, for instance *esinema* 'occur', *leiduma* 'be found, reside', *raksama* 'crack' and *värvuma* 'take colour'. The universal EC verb *olema* 'be' is especially common in them:

(24) $\underline{N^{G}\ ees\ V\ N^{N/P}}$
Meie ees *ol-i-d* *uue-d* *ülesande-d.*
we.G in.front be-PST-3PL new-N.PL task-N.PL
'There were new tasks waiting for us.' (Rätsep 1978: 141, pattern 59)

(25) *Meie* *ees* *oli* *uus-i* *ülesande-i-d.*
we.G in front be.PST.3SG new-P.PL task-P-PL
'There were some new tasks waiting for us.' (ibid.)

(26) $\underline{N^{N/P}V\ A^{TR}}$. Example with the verb *värvuma* 'take colour'.
Puu-d *värvu-sid* *kollase-ks.*
tree-N.PL take.colour-PST.3PL yellow-TR
'The trees turned yellow.' Lit. 'The trees coloured yellow.' (Rätsep 1978: 136, pattern 47.0)

(27) *Pu-i-d* *värvu-s* *kollase-ks.*
tree-P-PL take.colour-PST.3SG yellow-TR
'Some trees were turning yellow.' Lit. 'Some trees were colouring yellow.' (Rätsep 1978: 136)

Example (26) is rather categorisable as a non-existential intransitive sentence as it is hard to imagine it being used for the prototypical purpose of the ECs: presenting a new subject referent to the discourse.

2.2.4 Influence of the construction or verbal semantics?

Earlier studies have taken different viewpoints on whether it is the lexical predicates or the whole construction that determines the Estonian e-subject's case. Rätsep (1978: 258) suggests that the arguments' case depends on the verb's lexical meaning but Nemvalts disagrees with this and argues that at least the case-marking of e-subjects depends on the whole clause (2000: 109). It is important to keep in mind though that this is a minority of e-subjects in ECC whose case gets determined on this level B of case-assignment factors – most of the e-subjects' exact case is not determined by neither the predicate verb nor the construction, as shown in Section 2.2.3.

When discussing what influences the case of these e-NPs that get their case-specification on the level B, as shown in Sections 2.2.1 and 2.2.2, it is necessary to distinguish the verb's core semantics from the semantics of the expression (cf. Goldberg 1997: 384). Goldberg demonstrates that the verb only designates the elaboration of the constructional meaning in the most prototypical and common cases (like *He **gave** me an apple*). There are also constructions whose meaning does not depend on the verb's meaning. For example in *The bus **rumbled** down the alley* the core meaning of the sentence involves motion. However, this is not part of the core meaning of the verb (ibid.: 385–386). In the following I will show that the formal realisation of the sentential elements can depend on both verbal and constructional semantics.

Among other groups, there are the following groups that contain e-subjects:

- "nominative e-subjects only" constructions (7 different);
- "partitive e-subjects only" constructions (11);
- "nominative e-subjects only" verbs (a large open class).

There is no "partitive e-subjects only" group of verbs: the verbs that occur in partitive constructions have sub-meanings that occur in other constructions, see Section 2.2.1. Table 4 demonstrates on the basis of the EC sentence constructions list (see above) the possible combinations of EC constructions and the verbs used in ECs.

Table 4. The distribution of existentially used verbs among constructions ("x" marks that there are occurrences in the EC constructions list, "-" marks that there are not.)

Verbs/Constructions	**Nom e-subject cn-s**	**Part e-subject cn-s**	**Nom/Part e-subject cn-s**
Nom subject verbs	x	-	-
Part subject verbs	-	-	-
Nom/Part subject verbs	x	x	x

The lexical predicates view would be supported if verbs clustered together in constructions according to their relevant lexical restrictions. The verbs that belong to the **"partitive subject only" constructions**, should then have two properties:

- compatibility with the existential meaning – as it is the case with all the verbs occurring in ECs;
- their subject referents have obligatorily unbounded (non-inclusive) quantity.

There are over 17 verbs (like *jätkuma* 'suffice, continue', *jaguma* 'suffice, be divisible by', *tunduma* 'seem, smell', *sadama* 'come down, rain', *sisalduma* 'inhere in', etc.; see Section 2.2.1)[20] that can be found in 'partitive subject only' constructions. However, they can also occur in other constructions. Therefore this second meaning component can only be specified by the construction they occur in. In their case it is obvious that the partitive subject requirement cannot only come from the meaning of the verb.

The situation is different in the case of the verbs and constructions with **nominative e-subjects**. All the verbs in 'nominative subject only' constructions are 'nominative subject only' verbs. Although it is theoretically possible, none of the verbs that occur in 'nominative subject only' constructions occur in 'partitive subject only' constructions. Nominative subject-marking shows that the quantity of the e-subject

[20] Rätsep (1978: 77) does not specify the exact number of verbs in the 'partitive subject only' pattern 1.3.

referent is inclusive – the subject denotes e.g. a singular count noun or a bounded set – or is just unmarked (see Sections 2.3.1, 2.4.1 and 2.4.2). When determining whether this is due to the construction or due to the verb's semantics I follow Goldberg's judgement that in prototypical cases it is the verb whose semantics determines the arguments' realisation (see above in this section).

To conclude, both predicate verbs and whole constructions influence Estonian e-subjects' case. In the case of the nominative e-subjects that receive their case-marking on level B it rather seems to be the verb that determines the case-marking. In the case of the partitive e-subjects whose case is assigned on level B, it is the influence of the whole construction. For brevity, in the following I will refer to both factors as construction-influenced subject-marking.

2.3 Level C of subject case assignment

In the affirmative ECs occurring in the constructions "indifferent" to the subject case, its case depends on the lexical semantics of the head noun: in the existential construction these nouns' case-marking is pre-determined (noun groups in the Sections 2.3.1 and 2.3.2). I call it level C of e-subject's case assignment. One noun group (outlined in 2.3.3.) permits subject case alternation. The nouns that belong there get their case-assignment on level D. In the following analysis of the nouns' countability and other properties determining their case, I use the BCE and Google examples, as well as the Explanatory Dictionary of Estonian (Langemets et al. 2009) in addition to the ECC.

2.3.1 The noun belongs to the group "Existential nominatives"

In the description of grammatical behaviour of lexical sub-classes, more fine-grained distinctions than just 'count nouns' are needed (Croft 2001: 60). In the following I use a more detailed classification for the groups of nouns that occur as nominative e-subject heads in affirmative ECs. I use the term 'Existential nominatives' as a general denomination for all nominative e-subject heads whose case is lexically determined. In ECC this was the largest group, 103 (37%) subjects belong here. The subgroups of the Existential nominatives are singular count nouns, abstract nominatives, set nouns and *plurale tantum* nouns.

Of 279 subjects in ECC, 86 are clear instances **singular count nouns**. From these, 5 get their case marking on level B and 81 on level C. For example:

(28) *Ja korraga torka-s mu-lle pähe veider **mõte**.*
and suddenly strike-PST.3SG I-ALL head.ILL strange.N.SG thought.N.SG
'And suddenly I got this strange idea.' Lit. 'And suddenly stroke into my head a strange idea.' (ECC)

In this study, I also treat proper names under this category. A possible explanation of the frequent use of count nouns as e-subjects is that they are more frequent in language in general (for comparison: there were 33 clear mass noun subjects in ECC). In the Frequency Dictionary of Written Estonian (Kaalep & Muischnek 2002) there are about 10 times more count nouns than mass nouns among the 400 most frequent word forms.

Abstract nominatives are a considerable group of abstract nouns that have a rather limited use in the affirmative ECs: they do not permit plural and, as e-subjects, they can only occur in the nominative (cf. Nemvalts 1996: 39–41). As these nouns are often not referential, the further classification between mass and count nouns is in general irrelevant for them. In ECC there are 21 nouns having these limitations of use: one of them had its case assigned on level B and 20 on level C.

(29) *Sugene-s pisut piinlik vaikus*
appear-PST.3SG a.bit embarrassing.N.SG silence.N.SG
'A bit of an embarrassing silence appeared.' (ECC)

There is a group of nouns that doesn't match either the description of count nouns or mass nouns. Often they have the derivational affix *-stik* ('a set of something'), e.g. *sulestik* 'plumage' (Lit. 'feather'+*stik*), *lehestik* 'leafage' (Lit. 'leaf'+*stik*), *lihastik* 'bulk of muscles' (Lit. 'muscle'+*stik*). As e-subjects they occur in the nominative (Nemvalts 1996: 41). For instance:

(30) *Teise-s Ø kahise-s ja turtsu-s torustik.*[21]
other-INE (room.INE) rustle-PST.3SG and splutter-PST.3SG piping.N.SG
'In the other (room), the piping was rustling and spluttering.' (ECC)

[21] The example (30) resembles the "nominative subject only" construction ((N^{ADE}) V N^{N}) outlined in Section 2.2.2 but its predicate verb *olema* 'be' does not occur in that construction.

Here the subject case assignment depends on the semantics of the subject noun. What connects these hardly definable nouns with count nouns is that usually they can be characterised by the feature [+Shape]. They have a definite outline (cf. Rijkhoff 2002: 54), e.g. *lehestik* 'leafage' does not denote an indefinite quantity of leaves of a tree but all the leaves; *sulestik* 'plumage' denotes the whole plumage of a bird. What brings these nouns closer to mass nouns is the feature [+Homogeneous] when looked at their inner structure – if you add to a bird's plumage one feather, you will still get a plumage (cf. Rijkhoff 2002: 53; see also Divisibility in Section 1.3). The noun still preserves its singular marking then. However, at the so called "macro level" of their usage they appear to be neutral to the property [Homogeneous]: for denoting the plumages of several birds, usually the singular form is still used; *lehestik* of several trees is still a singular *lehestik*. The singular *stik*-noun stands both for singleton and multiple sets.

Rijkhoff calls the nouns that possess the property [+Shape] and are neutral to the property [Homogenous] **set nouns**. He characterises them as follows: set nouns are not marked for number, they can contain just one individual or consist of more individuals; they can be directly combined with a numeral and they don't take plural in this case (2002: 46–50).

These criteria apply to the Estonian data in part. In Estonian these nouns are usually both semantically and morphologically unmarked for number. They can consist of one or more individuals. However, the Estonian analogue does not combine with numerals. In ECC, BCE or in Google such examples don't occur:

(31) **Aeda kaunista-sid kaks dekoratiivse-t lehestikku.*
garden.P embellish-PST.3PL two.N decorative-P.SG leafage.P.SG
'The garden was embellished by two decorative leafages.'

In some cases the nouns combine with the plural, but similarly to mass nouns, this brings along a meaning change: from a specific entity to different types or kinds of this entity. Hence nouns like *sulestik* 'plumage' form a clearly separate group in Estonian. A possible solution is to add them to the fringe of Rijkhoff's *set nouns* category. However, not all Estonian *stik*-nouns behave this way: the nouns that tend to have clearer boundaries and more discrete structure like *sõnastik* 'dictionary' (Lit. 'word'+*stik*), *saarestik* 'archipelago' (Lit. 'island'+*stik*), behave syntactically like proper count nouns.

Another kind of existential nominatives is the ***plurale tantum*** nouns that have an inherent boundary for their quantity that requires nominative marking in the affirmative ECs (see Vilkuna 1992: 59–60). ECC contained one example of *plurale tantum*:

(32) *Järgne-sid* *õhtuvärvi-d* *järve-l.*
follow-PST.3PL twilight.colour-N.PL lake-ADE
'(Then) there followed the twilight colours on the lake.' (ECC)

2.3.2 The noun belongs to the group "Existential partitives"

The group Existential partitives contains mass nouns like *aimu* 'vague.idea.P.SG', *ruumi* 'space.P.SG (submeaning: sufficient space for fitting something in)', *üh-t-teis-t* 'something-P.SG (a thing or two)' and *jõudu* 'strength.P.SG'. They are used non-inclusively in the ECs (cf. Nemvalts 2000: 64–67). In ECC there are 20 uses of existential partitives,[22] for example:

(33) *Endal=gi* *Ø* *ruumi* *vaevalt* *ringi* *pööramise-ks.*
self=CL (be.3) space.P.SG merely around turning-TR
'We ourselves (have) only just (enough) space for turning around.' (ECC)

In many such e-subjects, the semantics of the partitive case is fading and the non-inclusive meaning is often not essential to this form any more, as is in the case of the expressions *juttu tule-ma* [talk.P.SG come-INF] 'come to discussion' and *tegemis-t* / *tegu ole-ma* [doing-P.SG be-INF] 'be dealing with', and the construction *something*.P + adjective.P: *midagi ilusa-t* [something.P.SG nice-P.SG] 'something nice'. For example:

(34) *Tegemis-t* *on* *huvitava* *isiksuse-ga.*
dealing-P.SG be.3 interesting.G personality-COM
'Here we are dealing with an interesting personality.' (ECC)

[22] The existential partitives (8) that occurred in a negative sentence are not counted in here.

2.3.3 The noun belongs to the group "Divisibles"

The case of divisible e-subjects is determined by situational inclusivity-related variables and pragmatic implicatures. The rest of this paper describes and explains their case-marking principles.

2.4 Level D of subject case assignment

From now on only the divisible e-subjects (i.e. mass nouns and plural count nouns, see Section 1.3) will be discussed. Unlike in the case of the other noun groups, the existential nominatives and existential partitives, the case of divisibles is not assigned on level C. When on the first three levels the subject case constraints were relatively simple and rigid then among the divisibles several choices and meaning nuances comes into play. I will discuss the following semantic and pragmatic factors: inclusivity, the effect of determiners, pragmatic implicatures and specificity. All of these factors are dependent on the particular situational context in the sentence or discourse, not just on the lexical semantics of the head noun. In ECC, 122 e-subjects are divisibles: 12 of them had their case assigned under negation, 3 on level B and 107 (38% of all ECC subjects) on level D.

The DSM factors discussed in 2.4 belong to three main groups. The subject case depends on the presence or lack of inclusivity meaning (Section 2.4.1) or on the inclusive-non-inclusive meaning of the subject referent (2.4.2). In 2.4.3 unspecific reference will be discussed as a potential DSM factor.

2.4.1 Subject case alternation is based on the opposition of the presence or lack of inclusivity meaning

The e-subjects' inclusivity meaning is a quantitative definiteness phenomenon (see Section 1.3) dependent on the situational context and not on lexical semantics. It operates on two levels in Estonian. Primarily it is the question, whether inclusivity is relevant for the e-subject and whether it is marked at all. In most divisible e-subjects, the nominative-partitive opposition stands for the overt marking of non-inclusivity vs. the lack of it (the topic of this section). Less often, there is the proper inclusivity opposition of inclusive and non-inclusive quantity (see Section 2.4.2). The

difference between these oppositions is in the existence or non-existence of a contextual boundary.[23]

The opposition of presence and lack of inclusivity determination (the PLI opposition) is the second most frequent factor influencing the e-subject marking in ECC. 72 subjects (26%) in ECC have this case-assignment motivation. Out of these 41 are nominative and 31 partitive, 20 are mass nouns and 52 plural count nouns. The following sentence pairs exemplify the nominative and partitive marking of plural count nouns ((35), (36)) and mass nouns ((37), (38)). The subject referents in (35) and (37) are concrete and in (36) and (38) abstract.

(35) a. *Pilti-de-l* *ol-i-d* *koduloom**a**-d.*
picture-PL-ADE be-PST-3PL domestic.animal-N.PL
'There were domestic animals in the pictures.' (ECC)

b. *Pilti-de-l* *oli* *koduloom-i.*
picture-PL-ADE be.PST.3SG domestic.animal-P.PL
'There were (some) domestic animals on the pictures.'

(36) a. *Alatihti* *juhtu-si-d* *tema-ga* *õnnetuse-d.*
almost.always happen-PST-3PL he-COM accident-N.PL
'He often got into accidents.' Lit. 'Almost always accidents happened with him.'

b. *Alatihti* *juhtu-s* *tema-ga* *õnnetus-i.*
almost.always happen-PST.3SG he-COM accident-P.PL
'He often got into accidents.' Lit. 'Almost always accidents happened with him.' (ECC)

(37) a. *Maa-s* *oli* *rohi.*
ground-INE be-PST.3SG grass.N.SG
'There was grass on the ground.'

b. *Maa-s* *oli* *rohtu.*
ground-INE be-PST.3SG grass.P.SG
'There was (some) grass on the ground.' (ECC)

[23] The more precise term *inclusivity opposition* is more accurate here than *quantity opposition* which has sometimes been used too but can also be used to mark the number category and other meanings.

(38) a. *Lähikonna-s* *kost-is* *kolin.*
vicinity-INE sound-PST.3SG clatter.N.SG
'Clatter was heard in the vicinity.' Lit. 'In the vicinity clatter sounded.'

b. *Lähikonna-s* *kost-is* *kolina-t*
vicinity-INE sound-PST.3SG clatter-P.SG
'(Some) clatter was heard in the vicinity.' Lit. 'In the vicinity clatter sounded.' (ECC)

The difference between the partitive and nominative subjects can be better understood with the help of the following paraphrases: *there are some X-s there* (the partitive) vs. *there are such X-s there* (the nominative), hence the distinction is not between inclusive and non-inclusive quantity. The ECs subject to this case-assignment factor do not have a **contextual boundary** (location instruction) that would help to identify the inclusive quantity of the subject referent (see also Section 1.3). The lack of a contextual boundary prohibits the reading of inclusiveness but permits non-inclusive (indefinite quantity) reading. The partitive counterparts of (35)–(38) have the non-inclusive reading and the nominative counterparts' inclusivity is unspecified.

This meaning difference between these ECs with the nominative and partitive subjects is only subtle and it often becomes apparent in a particular context (Erelt et al. 1993: 44). According to Nemvalts (2000: 150) the plural nominative (e.g. *inimese-d* 'people-N.PL') identifies the referent as a set that has qualities characteristic to the class (e.g. the notional CLASS OF PEOPLE). The partitive identifies the referent as an entity belonging to the class, i.e. as a member of the class (see also Vähämäki 1984: 73–75 on Finnish and Section 2.4.3 in this paper).

The choice of the nominative – a canonical subject case – also gives the sentence a more transitive feel (in the sense of the transitivity continuum of Hopper & Thompson 1980) and the partitive brings about more intransitive connotations.

2.4.2 Subject case alternation is based on the opposition of inclusive and non-inclusive meaning

The opposition of inclusive and non-inclusive meaning (the IN opposition) is different from PLI in that both the nominative and partitive e-subject are marked for inclusivity. The nominative marks inclusively involved and the partitive non-inclusively involved subject referents. In ECC there are 35

examples whose decisive case choice factor is the IN opposition; the examples are in the nominative and in the partitive, in the singular and plural. It has been suggested that one of the following conditions has to be met, for a divisible subject to be able to express inclusive quantity. Firstly, the context sets a boundary for the subject referent's inclusive quantity (Vilkuna 1992: 60–61; Nemvalts 2000: 108; see Section 2.4.2.1 below); secondly, the subject NP includes an adjective or determiner (see Section 2.4.2.2), or, thirdly, DSM can be used to express aspectual differences (Section 2.4.2.3) (Nemvalts 2000: 87–89, 100–101, 151–152).

In the Sections 2.4.2.1–2.4.2.3 I argue that only the first constraint is decisive and significant for subject case-marking (this posits a minor meaning difference between two otherwise similar types of partitive divisibles – indefinite quantity e-subjects whose case is assigned by PLI or IN: the existence/non-existence of a contextual boundary in the situation. However, partitive case-marking neutralises this distinction).

The opposition types PLI and IN illustrate the well-known distinction between semantic content and **pragmatic implicatures** (what is said vs. what is being communicated) (cf. Haspelmath 2006). In different situational uses, a word with a more general meaning (e.g. 'lion') can get distinctive pragmatic implicatures: 'lion in general' (semantically unmarked) vs. 'a male lion' (semantically marked, minus-meaning). Both of them can be opposed to the semantically marked 'lioness' (plus-meaning) (ibid.). In Estonian ECs, the semantically unmarked nominative gets in some contexts the specific pragmatic implicature 'inclusive quantity'. As the nature of the partitive case is the marking of non-inclusive quantity, the semantic content of the partitive NP is always specified for inclusivity and may be analysed as a hyponym of the general meaning of the nominative NP.

2.4.2.1 Contextual boundary on the referent's quantity

This group of ECs involves NPs that are similar to *plurale tantum* nouns (see Section 2.3.1) but they get their ground for bounding their quantity from situational context, and not from their lexical semantics like the *pluralia tantum* do, for example:

(39) *Selle-l kase-l on juba lehe-d.*
this-ADE birch-ADE be.3 already leaf-N.PL
'This birch has leaves already.' Lit. 'On this birch is already leaves.' (adapted from Vilkuna 1992: 61.)

The e-subject *lehe-d* 'leaf-N.PL' can have an inclusive reading due to the existence of a contextual boundary, a typical quantity of leaves in the tree: the amount of all the leaves in the leafage. When analysing the case-choice factors of the divisibles in the ECC I considered that (unlike the *pluralia tantum*) the contextually bound NPs also permit the non-inclusive use that brings about the partitive marking:

(40) *Selle-l kase-l on juba leht-i.*
this-ADE birch-ADE be.3 already leaf-P.PL
'This birch has some leaves already.' Lit. 'On this birch is some leaves.'

In the next example pair, the contextual boundary is the inclusive amount of bonfires that the speaker knows to be lit in the forest in the evening under consideration. The nominative subject stands for all the bonfires that were lit and the partitive denotes some of these bonfires.[24]

(41) *Veel õhtupimeduse-s lõõma-sid tule-d metsa all.*
still dusk-INE flame-PST.3PL bonfire-N.PL forest.G under
'At dusk, the bonfires were still flaming in the forest.' (ECC)

(42) *Veel õhtupimeduse-s lõõma-s tule-sid metsa all.*
still dusk-INE flame-PST.3SG bonfire-P.PL forest.G under
'At dusk, some bonfires were still flaming in the forest.'

2.4.2.2 Adjective or determiner in the NP

Nemvalts (2000: 77–80) claims that when the e-subject NP is modified by an adjective the nominative brings about inclusive reading and the partitive the non-inclusive one. The following sentence expresses the possibility of IN opposition difference:

(43) *Hobuse keha-l on punase-d ja musta-d träpsu-d.*
horse.G body-ADE be.3 red-N.PL and black-N.PL spot-N.PL
'There are red and black spots on the horse's body.' (Google)

In general, the nominative NPs state that *all* the head noun referents in this particular location are of the kind specified by the adjective cluster. (43)

[24] In this story the narrator talks about the forest owned by their family and about the forest work done by her husband and children. In the narrative, all the bonfires lit in this forest have definitely been lit by them.

implies that there are no other colour spots on the horse's body than red and black. The use of the partitive e-subject, on the contrary, would carry non-inclusive meaning: i.e. the state that (at least) some of the spots on the horse's body are red and black (cf. Nemvalts 2000: 81). Similar IN opposition can be attested in the case of non-coordinated subjects and mass nouns:

(44) a. *Kahe-l* *pool* *kasva-s* *kõrge* *vili.*
two-ADE side grow-PST.3SG high.N.SG crop.N.SG
'High crop was growing on both sides.' (ECC)

b. *Kõrge-t* *vilja* *kasva-s* *kahe-l* *pool.*
high-P.SG crop.P.SG grow-PST.3SG two-ADE side
'Some high crop was growing on both sides.'

The nominative use states that the only kind of crop growing in the location is high (later on in the narrative, this is also overtly expressed). The partitive use, on the other hand, implies that in addition to high crop, the other kind *can* also be growing, for example low crop.

In ECC, not all adjective-modified e-subjects show IN opposition in their case-alternation. For example in the next sentence pair, neither option expresses inclusive quantity. The nominative e-subject is neutral with respect to inclusivity.

(45) a. *Ootamatult* *hakka-s* *mu-lle* *sigine-ma* *uus-i*
unexpectedly start-PST.3SG I-ALL appear-INF new-P.PL
naistuttava-i-d.
female.acquaintance-PL-P
'Unexpectedly I started to get new female acquaintances.' Lit. 'Unexpectedly new female acquaintances started to appear for me.'(ECC)

b. *Ootamatult* *hakka-sid* *mu-lle* *sigine-ma* *uue-d*
unexpectedly start-PST.3PL I-ALL appear-INF new-N.PL
naistuttava-d.
female.acquaintance-N.PL
'Unexpectedly I started to get new female acquaintances.' Lit. 'Unexpectedly new female acquaintances started to appear for me.'

If in the case of the examples (43) and (44), the subject case alternation reflects the inclusive – non-inclusive quantity distinction (the "IN-group") then in the case of the example (45) it reflects PLI distinction (the "PLI-

group"). The difference between the two groups is again in the existence of a contextual boundary of the subject referent. In the case of the IN-group clauses, the referent of the subject NP, like *black spots*, has a definite larger (background) entity or set where it belongs to: only these spots are being discussed that are located within the boundaries of the horse's body.[25] On the other hand, in the case of *some new acquaintances* there is no standard size of the larger set of new acquaintances.

Of the 103 subjects in ECC that are modified by an adjective only 9 have their case assigned by the IN opposition. To summarise, an adjective in the NP is not an independent factor permitting IN opposition distinction of the e-subject's case-alternation. Instead, the case-alternation of adjective-modified NPs that shows inclusiveness distinction can be explained by the same conditions as outlined in 2.4.2.1.

Nemvalts (2000: 87–89, 100–101) has shown that there are some **determiners** that condition or release the ban of nominative marking in the affirmative ECs. These include determiners like *terve* 'whole', *mõlemad* 'both', *osa* 'part' when used as a modifier, *oma* 'specific', *teatav* 'certain' and *just see* 'exactly this', *mingi* 'some' and *mõningane* 'some, a certain'. For example:

(46) *Seal on mõlema-d lapse-d / *mõlema-i-d laps-i.*
there be.3 both-N.PL child-N.PL / both-P-PL child-P.PL
'There are both children there.' (Google)

(47) *Se-l seiga-l on oma tähtsus /*
this-ADE event-ADE be.3 its.specific importance.N.SG /
**oma tähtsus-t.*
its.specific importance-P.SG
'This event has its specific importance.' (Nemvalts 2000: 87)

(48) *Ta-l on selle-st mingi aim. /*
he-ADE be.3 this-ELA some.N.SG vague.idea.N.SG /
**mingi-t aimu*
some-P.SG vague.idea.P.SG
'He has some idea about this.' (Google)

[25] Although at first sight, in the case of the examples of (44) there does not seem to be any bounded larger entity in the context, the larger entity is actually the immediate vicinity visible for the speaker walking along the lane (in this narrative, a child whose line of vision is blocked by the high crops).

Also numerals and measure nouns (*kaks* ‘two’ and *kamalutäis* ‘handful’) bound the NP quantitatively and occur in the nominative (49):

(49) *Aia-s* *seis-i-d* *kaks* *õunapuu-d.*
garden-INE stand-PST-3PL two.N.SG apple.tree-P.SG
‘There were two apple trees in the garden.’ (ECC)

Such e-subjects are a borderline case that does not actually go well neither under existential nominative nouns nor under the inclusivity effects of divisible nouns. Numerals can be treated as the determiners of nouns (Krifka 1996: 583–584). Although the nouns with numeral and quantifier determiners have singular form, they have plural semantics – therefore this e-subject type is discussed here under divisibles.

In sum, the prototypical function of determiners is to give the subject referent a contextual boundary that enables inclusive quantity interpretation – which can be regarded a manifestation of the IN opposition. For example (46) and (47) rather talk about the inclusive amount of ‘children’ and ‘importance’ than some indefinite or partial amount. In ECC there are 15 examples (5%) where the determiner conditions the e-subject’s nominative marking: in 10 cases a numeral and also *mõni* ‘some’ and *ükski* ‘one’.

2.4.2.3 Subject case and aspect

The case alternation of e-subjects can reflect aspectual distinction only in the case of divisibles, provided that the predicate is an accomplishment or achievement verb (i.e. telic, cf. Vendler 1967; Dowty 1979; see also Nemvalts 2000) and the verb is used in the imperfect or future tense (Nemvalts 2000: 126–130). For example, in the case of the perfect tenses, the aspect is perfective and the aspect alternation is not possible.

If these conditions are met, the nominative can stand for perfective and the partitive for the imperfective reading (the interpretation of aspectual opposition is usually optional). Clauses (50) to (53) exemplify accomplishments and the examples (54) and (55) juxtapose a punctual (54) and an iterative achievement (55).

(50) *Silla* *taha* *kogune-s* *rämpsu.*
bridge.G behind.ILL gather-PST.3SG litter.P.SG
‘Litter was gathering behind the bridge.’ (ECC)

(51) *Silla taha kogune-s rämps.*
bridge.G behind.ILL gather-PST.3SG litter.N.SG
'Litter gathered behind the bridge.'

(52) *(Raputa-si-n põõsas-t, nii et)*
shake-PST-1SG bush-P.SG so that
must-i marj-u ema-le pähe sada-s.
black-P.PL berry-P.PL mother-ALL head.ILL fall-PST.3SG
'(I shook the bush so that) black berries kept falling on my mother's head.' (ECC)

(53) *...musta-d marja-d ema-le pähe sada-si-d.*
black-N.PL berry-N.PL mother-ALL head.ILL fall-PST-3PL
'(I shook the bush so that) black berries fell on my mother's head.'

(54) *(Õhk hakka-s põru-ma ning)*
air.N start-PST.3SG shake-INF and
üle pihlaka-te kihuta-sid reaktiivlennuk-i-d.
over rowan.tree-G.PL sweep-PST.3PL jet.plane-P-PL
'(The air started to shake and) jet planes swept by over the rowan trees.' (ECC)

(55) *...üle pihlaka-te kihuta-s reaktiivlennuke-id.*
over rowan.tree-G.PL sweep-PST.3SG jet.plane-P.PL
'...jet planes were sweeping by over the rowan trees.'

The verb in (50) and (51) can be replaced by another verb *voolama* 'flow' – and the clauses can still reflect aspectual opposition then. However, the use of this verb in the next example pair gives an imperfective result.

(56) *Tema pilgu-st voola-s armastus-t.*
he.G gaze-ELA flow-PST.3SG love-P.SG
'There was love flowing from his gaze.' (ECC)

(57) *Tema pilgu-st voola-s armastus.*
he.G gaze-ELA flow-PST.3SG love.N.SG
'There was love flowing from his gaze.'

These examples show that for aspectual opposition it is necessary that the situation can occur both perfectively (bounded) and imperfectively (unbounded): there must be a potential boundary in the verb meaning or situation, i.e. a change of state point. In the examples (56) and (57) there is no potential boundary and therefore the verb is used as an activity verb; the perfective-imperfective opposition is impossible (the case-alternation

depends on the PLI instead). The ECs with the aspectual opposition possibility express incremental or potentially iterative meaning or dynamic (serial) conceptualisation of the subject (cf. Huumo 2007 on Finnish). Often these ECs contain an adverbial with the lative meaning: an NP in illative or adessive case or a PP, compare (50) and (51) to (56) and (57). A special type of ECs that permit aspectual opposition, have the target state expressed by a translative adjective, see examples (26) and (27). In ECC, the only ECs that permit aspectual opposition and at the same time contain a non-lative adverbial are the ones with the potential punctual-iterative meaning opposition, compare the locative adverbial examples (54) and (55) (see also Huumo 2001 on Finnish).

I analysed the aspectual case-alternation possibility in 92 ECs with divisible subjects (whose case is not assigned on the higher levels; I also discarded from the analysis the NPs with a quantifier determiner that can never take the partitive case in the affirmative). Among these EC there are:

– 22 telic predicate verbs (verbs that enable both perfective and imperfective use) and 70 atelic (verbs that only allow imperfective use);
– 63 uses of aspect-prone tense (imperfect, semantic future) and 29 uses of a tense in which aspect-alternation cannot occur (perfects, present);
– 11 clauses where the mere change in subject case brings about the possibility of the opposite aspectual interpretation[26] and 81 clauses whose subject case-alternation cannot do it.

In ECC, including the examples (50)–(53), the aspectual opposition was always accompanied by the potential of IN distinction, just as outlined in the Sections 2.4.2.1 and 2.4.2.2. If the situation is perfective, the quantity of the subject referent is inclusive, and if the situation is imperfective its quantity is non-inclusive. Aspect is not a decisive factor determining the e-subject's case (but an attendant meaning that sometimes occurs with the IN meaning opposition). Huumo (2010: 89) concludes the same about Finnish.

2.4.3 Unspecific reference: a potential motivation for subject case alternation

Nemvalts (2000: 150) claims that in the case of mass noun and plural count noun e-subjects the nominative identifies a class and the partitive identifies

[26] In the case of coordinated subjects the verb is counted more than once (as many times as there are subject NPs).

the members of a class via their quantity. Following Vähämäki's (1984: 73–75) set theoretical approach Nemvalts means by *class* an abstract concept (e.g. THE CLASS OF BOYS) whose members are not abstract but ontological real world entities.[27] This implies that the plurality of classes means different kinds of classes (e.g. CLASSES OF BOYS includes the CLASS OF SMALL BOYS, etc.).

This section analyses whether class reference, as proposed by Nemvalts, has any effect on the divisible e-subject's case. It is important to keep in mind that this class-related nominative function discussed in this section is different from the type of reference outlined in 2.4.1. (here: "identify a class", in 2.4.1: "identifies the referent as a set that has qualities characteristic to the class"). I start with an example of plural count nouns:

(58) *Aia* *taga* *kasva-sid* *riisika-d.*
garden.G behind grow-PST.3PL milk.mushroom-N.PL
'Behind the garden milk mushrooms were growing.' (ECC)

(59) *Aia* *taga* *kasva-s* *riisika-i-d.*
garden.G behind grow-PST.3SG milk.mushroom-P-PL
'Behind the garden (some) milk mushrooms were growing.'

These clauses have a generic-like meaning if used to state that there were mushrooms growing in this location over several years. The meaning of the partitive subject is very similar to the corresponding nominative but the case overtly marks the NP for non-inclusive quantity. If the nominative plural subject of this EC identifies the class, then, according to the definition above, the nominative noun (unlike the partitive one) must denote an abstract concept, the CLASS OF MILK MUSHROOMS.

For assessing mental concepts, the empirical corpus analysis method of this study is not suitable. However, class semantics has many overlaps with the more accessible linguistic phenomenon of **unspecific reference** which is a complex set of different meaning possibilities. To check whether it is true that such nominative nouns are denoting a class, i.e. an abstract generalisation, I juxtapose these uses with various descriptions of

[27] The class reference vs. quantification opposition has also been attested in differential subject and object marking phenomena in the neighboring language Russian (Paykin & van Peteghem 2002).

unspecific reference and evaluate whether specificity can cause DSM in Estonian.

Chesterman (1991: 188–190) notes that specificity can be viewed as a gradient in which "the cut-off point between the two poles of 'referential' and 'non-referential' would be difficult to place with any precision". Vilkuna (1992) has presented a systematic description of the elaborate nature of unspecific reference in Finnish. In this discussion only the types appearing in affirmative contexts are relevant: reference to a category, quantified contexts, generic NP reference and sort and species reference (*lajireferenssi* in Finnish).

I argue that often these instances overlap with Vähämäki's and Nemvalts' logical class reference (e.g. the CLASS OF BOYS). Although Nemvalts' explanation includes the identification process and Vilkuna's one is rather just centred on the nominals' referential properties, I find that the notions are mostly comparable.

2.4.3.1 Reference to a category

Vilkuna (1992: 106–125) explains the difference between expressing a category or an individual entity with referentiality: when only the category is mentioned, the referent is unspecific; reference to an individual is specific. When the reference is to a category, the focus is on the description which is more important for the communicative purpose than referring to a specific individual. For example: *Now you are asking* ***the wrong person****!* The sentence loses its communicative function if the NP is replaced with the specific *me.*

Vähämäki's characterisation of classes is compatible with Vilkuna's examples: the reference is an abstract concept characterised by identicalness of its members. Among the level D subjects, category reference only occurred in 3 (3%), for example:

(60) *On ju* ***elukutse-i-d*** *nagu tuletõrjuja ja autojuht,*
be.3 MDA profession-PL-P like fire-fighter.N and driver.N
mille-ga laps on varase-st ea-st tuttav.
that-COM child.N be.3 early-ELA age-ELA familiar.N
'Obviously, there are professions like the fire-fighter and driver that children are familiar with from an early age.' (ECC)

In (60) it is clearly the class reference that is relevant to the story: it is *professions* (and not directly fire-fighting or car-driving) that is a generalisation of the particular jobs listed afterwards. The divisible referent

is in the plural partitive and not (as Nemvalts proposes) in the nominative. Reference to a category cannot distinguish nominative divisible subjects from the partitive ones.

2.4.3.2 Quantified contexts

In quantified contexts the situation is recurrent and it is not bound to one point in time or to one place. The referent is unspecific and not unique as the situation is quantified by adverbials like *alati* 'always', *kord aasta-s* 'once a year (INE)', etc. or by the plurality of some NP (Vilkuna 1992: 96). In such ECs there is a mismatch between the grammatical number-marking of the subject NP and its quantity on the semantic level. For instance:

(61) *Kõigi-st padriku-te-st kost-is nen-de mahe hääl.*
all.PL-ELA scrub-PL-ELA sound-PST.3SG they-G.PL mellow.N.SG voice.N.SG
'From all the scrubs their mellow voice was heard.' Lit. 'From all the scrubs sounded their mellow voice.' (ECC)

In (61) the subject referent *hääl* 'voice.N.SG' is quantified by the plural forms of *kõigi-st* 'all.PL-ELA', *padriku-te-st* 'scrub-PL-ELA' and *nen-de* 'they-G.PL'. The quantified context effect only comes forward when *singular count nouns* refer to a number of instances, as in (61). Then the reference is unspecific. However, the NPs that are already in the plural bear the quantified meaning independently from the quantified contexts. The quantification effect does not work with mass nouns because they preserve their homogeneous and shapeless structure in these contexts. In conclusion, unspecific marking in the quantified contexts is close to Vähämäki's class reference but it does not occur with divisible e-subjects.

2.4.3.3 Generic NPs

The generic reference is similar to the above mentioned ones but the contexts are not time-bound or factual (Vilkuna 1992: 149). Prototypical generic sentences refer to laws, norms and principles and allow making predictions based on them (Dahl 1975). In the context of assessing the grammatical function of the nominative e-subject it is relevant to consider that the genericity of the predicate can occur separately from the genericity of the NP. For instance, in the next sentence the predicate is generic but the subject NP is specific: *Riitta's new car drives at 160 km/h* (Vilkuna 1992: 142–144). Vilkuna describes generic NPs as expressions that stand for abstract, intentional entities and kinds like the NPs in the following example: *The child who loses the dummy too early starts sucking her*

thumb, pillow corner or teddy-bear's ear (1992: 151). In the following I rely on Andrew Chesterman's (1991: 131) claim about Finnish generics that states that the generic meaning does not fit with the partitive but does combine with plural NPs.

(62) *Metsässä ulvoo susia.*[28]
forest-INE howl-3SG wolves-P
'In the forest (some) wolves howl.' (non-generic; Chesterman 1991)

(63) *Sudet ovat petoeläimiä.*
wolves-N are beasts-of-pray-P
'Wolves are beasts of pray.' (generic; Chesterman 1991)

This implies that the generic NPs are not compatible with non-inclusive meaning but they do fit with the inclusive collective meaning that the plural nominative can bring about. A difference between the generic NPs and Vähämäki's conceptual class notion is that if a class involves all its members then the generic reference may only roughly do it. Generic sentences can be construed by adding modifiers or quantifiers like *most, all, generally* etc. (Chesterman 1991: 34–35).

ECC contains 18 e-subjects (6%, incl. 9 divisible e-subjects) that occur in sentences that state some regularity or a general principle and can therefore be considered generic, e.g.:

(64) *(Kell kuus hommiku-l on niisugune aeg, kui)*
o'clock.N six.N morning-ADE be.3 such.N.SG time.N.SG when
peatänava-l=gi liigu-vad ainult varase-d rongiletõttaja-d.
main.street-ADE=CL move-3PL only early-N.PL rusher.on.the.train-N.PL
'(At 6 o'clock in the morning it is such a time when) even on the main street there are only early rushers on the train around (there is no-one else, so the street is pretty empty).' (ECC)

If (64) the situation refers to a general norm and plural nominative e-subject can be considered unspecific. On the one hand it would be possible to state now that the nominative-partitive distinction can signify the generic reference vs. specific reference opposition. However, the subject of (64) can easily be replaced with a partitive one and the referent gets a non-inclusive, that is non-generic reading:

[28] I have kept the original glossing in examples (62) and (63).

(65) *...kui peatänava-l=gi liigu-b ainult varase-i-d rongiletõttaja-i-d.*
when main.street-ADE=CL move-3SG only early-P-PL rusher.on.the.train-P-PL
'...when even on the main street there are only some early rushers on the train around.'

The miniscule difference in the interpretation of the two case options rather leads to an alternative analysis: that both the nominative and partitive subject clauses represent a generic *predicate* but specific indefinite subject referent. The subject case-marking then depends on the PLI.

In conclusion, the borders of genericity are vague but (depending on the interpretation) some e-subjects can occur in these contexts. In ECC there are no cases where the e-subject marking unquestionably depends on genericity.

2.4.3.4 Sort and species reference

Also sort and species reference is similar to generic reference, it differs from it in that it is time-bound.

(66) *Anne pühenda-s oma uurimuse lehetäi-le.*
Anne.N devote-PST.3SG her research.G blackfly-ALL.SG
'Anne devoted her research to the blackfly.' (Vilkuna 1992: 157)

The sort and species NPs are non-referential because they are characterised by identicalness of their members and there can be a mismatch between the semantics and coding of the NP's number (Vilkuna 1992: 157–158). Sort and species reference can occur with singular count nouns (non-divisible) and plural count nouns (divisible, reference is to different subspecies). There are no divisible e-subjects of this reference type in ECC. Sort and species reference is compatible with Vähämäki's class reference but it has either little or no compatibility with nominative divisible e-subjects.

2.4.3.5 Interim conclusion

Section 2.4.3 assessed the occurrence of unspecific or class reference in the affirmative ECs and evaluated its influence on subject case-marking. It tested Nemvalts' (2000: 150) hypothesis that *the function of nominative case on divisible subjects is to identify a class* (unlike the partitive that identifies the members of the class through their quantity).

The theoretical notion of class reference matches at large with unspecific reference. Among the 122 divisible subjects in ECC there are 12 clearly unspecific divisible e-subjects that get their case-marking on the

level D (see Table 5) and their case-marking depends on the PLI. In conclusion, I have found no evidence of unspecific reference having an influence on the subject case.

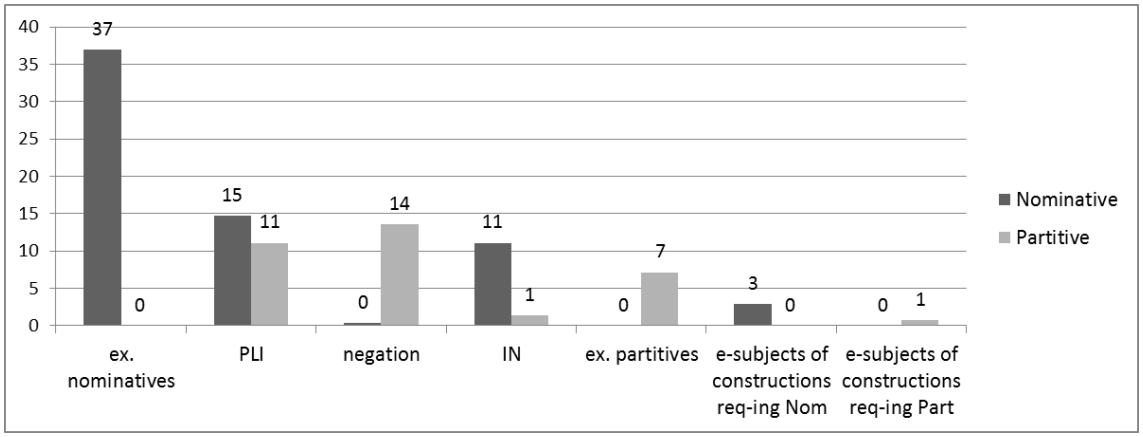
Table 5. The occurrence of different unspecific reference types in the existential clause corpus.

Unspecific reference type / Level of case assignment	Case assignment on level D	Case assignment on the other levels
Reference to a category	3 pl count nouns	1 C level subject
Quantified context	0	5 C level subjects
Generic NP reference	9 pl count nouns	4 C level subjects 5 A level subjects
Sort and species reference	0	0

3. Towards a comprehensive account

The assignment of Estonian e-subject case is a process that takes place on up to four levels. In the system the most influential rule is negation (the A level factor) that overrides the lower level conditions. The B level conditions only override the C and the D level ones, etc. The factors on each level are equal to each other. To determine the decisive rule of each e-subject case use one has to check the factors on every level until he finds the suitable one (starting from polarity, see Figure 1 in Section 2). If the subject case fails to get assigned on level A, it gets assigned somewhere below; if it also fails to get assigned on level B, it gets assigned on the levels below, etc. Figure 2 presents the frequencies of e-subjects divided by their case and level where it gets assigned in the corpus.

Figure 2. Frequencies of e-subjects in the existential clause corpus according to the level of their case assignment (%).



On level A **negative** clause e-subjects of the corpus get partitive marking. Exceptionally, directly presupposed e-subjects also occur in the nominative in the negative context. The case of affirmative ECs' subjects is determined on one of the next levels.

Level B of subject case assignment concerns specific constructions and lexical predicates. In the list of **EC constructions** used in this study, there are 95 different constructions but only 12 require a partitive subject (see Section 2.2.1) and 7 a nominative one (2.2.2). 76 of them leave the subject case choice free (2.2.3). Section 2.2.4 showed that it can be either just the verb or the whole construction that determines particular case assignment. If the verb-governed construction permits both cases, the subject gets its case assigned on one of the next levels.

On level C, DSM depends on **the subject's noun group**: whether the subject noun belongs to the existential nominatives, existential partitives or divisibles (Sections 2.3.1, 2.3.2 and 2.3.3 respectively). The first two groups were introduced in this paper; if these nouns occur in affirmative ECs their case is pre-determined. Existential nominatives is a general term that embraces singular count nouns, *plurale tantum* nouns and two groups introduced in this study that I call abstract nominatives and (on the basis of Rijkhoff's nominal aspect theory) set nouns. Abstract nominatives, e.g. *lootuskiir* 'ray of hope', is a relatively large group of abstract nouns that have a limited use in the affirmative ECs: they tend not to be categorisable neither as count nouns nor as mass nouns. Similarly to count nouns they occur in the singular nominative in the ECs (the partitive is not allowed) but similarly to mass nouns they cannot occur in the plural.

Set nouns are for example *sulestik* 'plumage' and *lehestik* 'leafage' that occur in the singular despite the number of entities they denote (e.g. the plumage of one or many birds). The group existential partitives involves different kinds of abstract nouns whose use in the ECs is fixed to the partitive singular (e.g. *ruumi* 'space.P.SG' and *tegemis-t* 'dealing-P.SG'). None of the ECs whose e-subject case gets assigned on level C contain constructions that specifically determine the e-subject's case.

Level D concerns the marking of **divisible e-subjects** (mass nouns and plural count nouns) that have a more complex meaning-based case alternation. This level involves the opposition of the *presence and lack of inclusivity meaning* (PLI, see Section 2.4.1) and the opposition of *inclusive and non-inclusive meaning* (IN, Section 2.4.2).

In Estonian, the e-subjects' inclusivity meaning operates on two stages. In the first case, the opposition of the **presence and lack of**

inclusivity meaning, only the partitive subject is marked for inclusivity (more specifically, for non-inclusive quantity). The opposition can be better understood with the help of the following paraphrases: *there is some amount of these referents in this location* (partitive subject), or: *there are such referents in this location* (nominative subject). The second opposition, of **inclusive and non-inclusive meaning**, can be better understood by the following paraphrases: *there is some amount of these referents in this location* (non-inclusive quantity, partitive subject marking), or: *all these referents are in this location* (inclusive quantity, nominative subject marking). In the second opposition, both the nominative and the partitive are marked for inclusivity. Hence, the primary difference between these two oppositions is in the meaning of the nominative counterparts. It comes from the existence of a contextual boundary for the subject referent: if there is a contextual boundary, the nominative subject means *all the referents*; if there is no boundary, the nominative stands for *such referents* (inclusivity is irrelevant here). The difference between the partitive e-subjects whose case gets assigned under PLI or IN is minor, dependent on the contextual boundary. The opposition types PLI and IN illustrate the semantic content and **pragmatic implicatures** distinction (Haspelmath 2006; see Section 2.4.2). A special group of e-subjects whose case is assigned under the IN factor is NPs with a numeral determiner (10 subjects out of 35, see Section 2.4.2.2). None of the ECs whose e-subject's case gets assigned on level D neither contain *constructions* that determine the e-subject's case nor specific subject nouns whose case is determined by the noun *lexeme*.

4. Conclusion

This article attempts to give a comprehensive account of subject case alternation in Estonian existential clauses. The paper reassesses the findings of earlier research on Estonian (especially Nemvalts 2000), mainly by using corpus and dictionary analysis and the studies on closely related Finnish.

In this study, the following subject case assignment factors were considered: referential properties of the subject noun, the subject referent's inclusivity and specificity in context; lexical predicates and particular constructions; other clause level and pragmatic properties (polarity, pragmatic implicatures and presupposition).

The Estonian existential subject's case-marking system abundantly exemplifies the interplay of all three variable types indicated by Witzlack-

Makarevich (2010). In typological studies pragmatic factors and the specific nature of morphological forms have also been discussed as factors of argument realisation (Bickel & Nichols 2008: 320). This is also the case with the Estonian e-subjects. The paper proposes an ordered four-level system of grammatical case-assignment rules (in the order of domination it consists of polarity, lexical predicates and particular constructions, the subject noun's lexical properties, situationally determined inclusivity and pragmatic properties). Among the studied 279 ECs, the most frequent subject case-assignment factors are the nominative taking noun type (the 'Existential nominatives' group), the NP referent's inclusivity and negation. The function of the partitive case (as a form marking unbounded meaning) has a major influence on the subject case in affirmative existential clauses. However, the crucial precondition for the option of subject case alternation is usually the existential construction environment itself whose defining feature is the topicality effect (the topic of the sentence is the locational adverbial and the subject tends to serve as the focus of the clause).

Cross-linguistically, non-canonical argument realisation often depends on semantic features, e.g. volitionality, which can (sometimes simultaneously) be bound to different levels of language: the predicate's lexical meaning or one of its sub-meanings, verbal affixes, choice of auxiliary, etc. (Onishi 2001: 23–40). This paper shows that the Estonian existential subject realisation depends in affirmative clauses on one fundamental semantic feature underlying all the case-assignment levels: quantitative definiteness (inclusivity). On these levels, the obligatory or optional marking of the subject referent's inclusive or non-inclusive quantity is either linked to certain verbs, noun lexemes, NPs or constructions. Concerning the situationally triggered inclusivity meaning on the NPs, the paper proposes a new functional explanation based on a binary division: the case alternation either depends on the presence or lack of inclusivity meaning or on the opposition of inclusive – non-inclusive quantity. The distinction between these two oppositions is in different pragmatic implicatures arising from situational uses.

Onishi also demonstrates that in the case of non-canonical arguments, the case-marking, agreement and syntactic behaviour usually differ from the properties of canonical arguments quite significantly (2001: 23–40). As it can be seen from this paper, that while there is some overlap between the case-marking and agreement of the Estonian existential clause argument and the canonical subject there are also considerable differences. Although

it has been suggested in the literature that the properties of Estonian existential subjects bring it close to the object, it can be concluded that it satisfies the conditions of non-canonical argument realisation and can therefore be regarded a non-canonical subject.

References

Aissen, Judith (2003) Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory* 21: 435–483.

Bickel, Balthasar (2010) Grammatical relations typology. In Jae Jung Song (ed.) *The Oxford Handbook of Language Typology*, pp. 399–444. Oxford: Oxford University Press.

Bickel, Balthasar & Nichols, Johanna (2008) Case-marking and alignment. In Andrej Malchukov & Andrew Spencer (eds.), *The Oxford Handbook of Case*, pp. 304–321. Oxford: Oxford University Press.

Chesterman, Andrew (1991) *On Definiteness: A Study with Special Reference to English and Finnish*. Cambridge Studies in Linguistics 56. Cambridge: Cambridge University Press.

Croft, William (2001) *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Csirmaz, Aniko (2005) *Semantics and Phonology in Syntax*. Unpublished doctoral dissertation. Massachusetts Institute of Technology. <http://hdl.handle.net/1721.1/33695> (20 May 2012).

Dahl, Östen (1975) On generics. In Edward Keenan (ed.), *Formal Semantics of Natural Languages*, pp. 99–111. Cambridge: Cambridge University Press.

Dahl, Östen & Koptjevskaja-Tamm, Maria (eds.) (2001) *Circum-Baltic Languages. Vol. 2: Grammar and Typology*. Studies in Language Companion Series 55. Amsterdam: Benjamins.

Dixon, Robert M. W. (1994) *Ergativity*. Cambridge: Cambridge University Press.

Dowty, David (1979) *Word Meaning and Montague Grammar*. Dordrecht: Reidel.

Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina; Saari, Henn; Tael, Kaja & Vare, Silvi (1993) *Eesti keele grammatika II. Süntaks, lisa: kiri* [Estonian Grammar II. Syntax. Appendix: Script]. Tallinn: Estonian Academy of Sciences, Institute of Language and Literature.

Erelt, Mati & Metslang, Helle (2006) Estonian clause patterns – from Finno-Ugric to standard average European. *Linguistica Uralica* 42 (4): 254–266.

Erteschik-Shir, Nomi (2007) *Information Structure. The Syntax-Discourse Interface*. Oxford: Oxford University Press.

Givón, Talmy (2001) *Syntax. Volume I.* Amsterdam/Philadelphia: John Benjamins.

Goldberg, Adele E. (1992) The inherent semantics of argument structure: The case of the English ditransitive construction. *Cognitive Linguistics* 3 (1): 37–74.

—— (1995) *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

—— (1997) The relationships between verbs and constructions. In Marjolijn Verspoor; Kee Dong Lee & Eve Sweetser (eds.), *Lexical and Syntactical Constructions and the Construction of Meaning: Proceedings of the Bi-annual ICLA Meeting in Albuquerque, July 1995*. Current Issues in Linguistic Theory, pp. 383–398. Amsterdam: John Benjamins.

Hakulinen, Auli; Vilkuna, Maria; Korhonen, Riitta; Koivisto, Vesa; Heinonen, Tarja Riitta & Alho, Irja (2004) *Iso suomen kielioppi* [The Large Grammar of Finnish]. Helsinki: Suomalaisen Kirjallisuuden Seura.

Haspelmath, Martin (2006) Against markedness (and what to replace it with). *Journal of Linguistics* 42: 25–70.

Hawkins, John, A. (1978) *Definiteness and Indefiniteness: A Study in Reference and Grammaticality Prediction*. London/Worcester: Croom Helm.

Helasvuo, Marja-Liisa (1996) Ollako vai eikö olla – eksistentiaalilauseen subjektin kohtalonkysymys [To be or not to be – The question of the existential subject's destiny]. *Virittäjä* 100 (3): 340–356.

Helasvuo, Marja-Liisa & Huumo, Tuomas (2010) Mikä subjekti on? [What is the subject?]. *Virittäjä* 114 (2): 165–195.

Hopper, Paul J. & Thompson, Sandra A. (1980) Transitivity in grammar and discourse. *Language* 56 (2): 251–299.

Huumo, Tuomas (1999) Kieltolauseen partitiivisubjektin semantiikkaa [The semantics of partitive subjects in negative clauses]. *Sananjalka* 41: 21–42.

—— (2001) Subjektin kvantiteetti ja verbin aspektitulkinta paljaan partitiivisubjektin konstruktiossa [The quantity of the subject and the aspectual interpretation of the verb in the bare partitive subject constructions]. In Reet Kasik (ed.), *Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks* [A Festschrift for Mati Erelt on the Occasion of His 60th Birthday], pp. 113–131. Tartu: Tartu University Press.

—— (2007) Kvantiteetti ja aika II: nominaalinen aspekti ja suomen predikatiivin sijanvaihtelu [Quantity and time II: Nominal aspect and case-alternation of Finnish predicative]. *Virittäjä* 111 (1): 3–23.

—— (2010) Nominal aspect, quantity, and time: The case of the Finnish object. *Journal of Linguistics*, 46 (1): 83–125.

Huumo, Tuomas & Perko, Jari (1993) Eksistentiaalilause lokaalisuuden ilmaisijana [Existential sentence expressing locality]. *Virittäjä* 97 (3): 380–399.

Kaalep, Heiki-Jaan & Muischnek, Kadri (2002) *Eesti kirjakeele sagedussõnastik* [Frequency Dictionary of Written Estonian]. Tartu: Tartu University Press.

Koptjevskaja-Tamm, Maria (2001) "A piece of the cake" and "a cup of tea": Partitive and pseudo-partitive nominal constructions in the Circum-Baltic languages. In Östen Dahl & Maria Koptjevskaja-Tamm (eds.), pp. 523–568. Amsterdam: Benjamins.

Koptjevskaja-Tamm, Maria & Wälchli, Bernhard (2001) The Circum-Baltic languages. An areal-typological approach. In Östen Dahl & Maria Koptjevskaja-Tamm (eds.), pp. 615–750. Amsterdam: Benjamins.

Krifka, Marfred (1989) *Nominalreferenz und Zeitkonstitution. Zur Semantik von Massentermen, Pluraltermen und Aspektklassen*. München: Fink.

—— (1996) Parameterised sum individuals for plural reference and partitive quantification. *Linguistics and Philosophy* 19: 555–598.

Lambrecht, Knud (1994) *Information Structure and Sentence Form. Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.

Langemets, Margit (2009) *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus keelevaras* [Systematic Polysemy of Nouns in Estonian and its Lexicographic Treatment in Estonian Language Resources]. Unpublished doctoral dissertation, Tallinn University.

Langemets, Margit; Tiits, Mai; Valdre, Tiia; Veskis, Leidi; Viks, Ülle & Voll, Piret (2009) *Eesti keele seletav sõnaraamat* [Explanatory Dictionary of Estonian]. Tallinn: Estonian Language Foundation.

Lyons, Christopher (1999) *Definiteness*. Cambridge: Cambridge University Press.

Lyons, John (1977) *Semantics*. *Vol. 2*. Cambridge: Cambridge University Press.

Milsark, Gary (1979) *Existential Sentences in English*. New York: Garland.

Nemoto, Noriko (2001) Verbal polysemy and frame semantics in construction grammar. Some observations on the locative alternation. In Mirjam Fried & Hans Christian Boas (eds.), *Grammatical Constructions: Back to the Roots*. Constructional Approaches to Language 4, pp. 119–136. Amsterdam: John Benjamins

Nemvalts, Peep (1996) *Case marking of subject phrases in modern standard Estonian*. Unpublished doctoral dissertation, Uppsala University.

—— (2000) *Aluse sisu ja vorm: alusfraasi käändevaheldus tänapäeva eesti kirjakeeles* [The Subject's Meaning and Form: Case-alternation of the Subject Phrase in Contemporary Written Estonian]. Tallinn: Estonian Language Foundation.

Onishi, Masayuki (2001) *Introduction: Non-canonically marked subjects and objects: Parameters and Properties*. In Robert. M. W. Dixon; Alexandra Y. Aikhenvald & Masayuki Onishi (eds.), *Non-Canonical Marking of Subjects and Objects*, pp. 1–51. Amsterdam/Philadephia: John Benjamins Publishing Company.

Paykin, Katia & Peteghem, Marleen van (2002) Definiteness in a language without articles: a case-study of Russian. *Recherches Linguistiques de Vincennes* 31: 97–112.

Partee, Barbara H. & Borschev, Vladimir (2007) Existential sentences, BE, and the genitive of negation in Russian. In Ileana Comorovski & Klaus von Heusinger (eds.), *Existence: Semantics and Syntax*. Studies in Linguistics and Philosophy, pp. 147–190. Dodrecht: Springer.

Rajandi, Henno & Metslang, Helle (1979) *Määramata ja määratud objekt* [Indefinite and Definite Object]. Ars Grammatica. Tallinn: Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Rijkhoff, Jan (2002) *The Noun Phrase*. Oxford Studies in Typology and Linguistic Theory. Oxford: Oxford University Press.

Rätsep, Huno (1978) *Eesti keele lihtlause tüübid* [Estonian Simple Clause Types], ENSV TA Emakeele Seltsi toimetised 12. Tallinn: Valgus.

Talmy, Leonard (2000) *Toward a Cognitive Semantics*. *Vol. 1*. Cambridge: MIT Press.

Vendler, Zeno (1967) Verbs and times. In Zeno Vendler (ed.), *Linguistics in Philosophy*, pp. 97–121. Ithaca, New York: Cornell University Press.

Vilkuna, Maria (1992) *Referenssi ja määräisyys suomenkielisten tekstien tulkinnassa* [Reference and Definiteness in the Interpretation of Finnish Texts]. Helsinki: Suomalaisen Kirjallisuuden Seura.

—— (1996) *Suomen lauseopin perusteet* [Foundations of Finnish Syntax]. Helsinki: Edita.

Vähämäki, Börje K (1984) *Existence and Identity: A Study of the Semantics and Syntax of Existential Sentences in Finnish*. Publications of the Research Institute of the Åbo Akademi Foundation 99. Turku: Åbo Akademi.

Witzlack-Makarevich, Alena (2010) *Typological Variation in Grammatical Relations*. Unpublished doctoral dissertation, Universität Leipzig.

Woolford, Ellen (2009) Differential subject marking at argument structure, syntax, and PF. In Helen de Hoop & Peter Swart (eds.) *Differential Subject Marking.* Studies in Natural Language and Linguistic Theory 72, pp. 17–40. Dordrecht: Springer-Verlag New York Inc.

Abbreviations

1 = first person
3 = third person
A = adjective
ABL = ablative
ADE = adessive
ALL = allative
BCE = Balanced Corpus of Estonian
CL = clitic
COM = comitative
COMP = comparative
DE = substitution class “Extralocal directional”
DI = substitution class “Directional”
DSM = differential subject marking
DT = translocal directional
e-NP = existential noun phrase
e-subject = subject of an existential clause
EC = existential clause
ECC = existential clause corpus
ELA =elative
G = genitive
ILL = illative
IN = inclusive-non-inclusive meaning opposition
INE = inessive
INF = infinitive

MDA = modal adverb
N, Nom = nominative
N = Noun
NP = noun phrase
P, Part = partitive
PL = plural
PLI = presence/lack of inclusivity determination
PP = preposition phrase, postposition phrase
PST = past
PTCP = participle
QP = quantifier phrase
SAC = syntactically annotated corpus
SG = singular
SUP = superlative
TR = translative
V = verb

Contact information:

Helena Metslang
Institute of Estonian and General Linguistics
University of Tartu
Jakobi 2
51004 Tartu
Estonia
e-mail: helena(dot)metslang(at)gmail(dot)com

Johanna Viimaranta

Analogy or Conceptual Metaphor? Coming Concretely and Abstractly Close in Uses of the Russian Prefix *pod-*

Abstract

This article discusses the notion of analogy as compared to conceptual metaphor and how the theoretical difference between the two affects the way in which the different meanings of polysemic linguistic units are seen to be related. This is illustrated through the meanings of concrete approaching and abstract approaching that the Russian prefix *pod-* adds to a verb stem. On the basis of the analysis it is claimed that the notion of analogy alone is not enough, as abstract approaching is not only an extension of the meaning of concrete approaching but includes features that can be explained only by the cooperation of both domains. It is suggested that this kind of approach will also be applicable in discussing the polysemy of other verbal prefixes.

1. Introduction

This article discusses the way that polysemy can be explained through the notions of analogy and conceptual metaphor. This is illustrated by studying certain meanings of the Russian prefix *pod-*. By way of an introduction I will now briefly discuss the importance of prefixation in Russian and the way that polysemy has been dealt with so far. I will then present the prefix *pod-* to the reader.

1.1 Prefixation in Russian

Russian describes action very precisely with mere verbs, due to both the system of verbal aspect[1] and the additional meaning components provided by prefixes. Russian has twenty verbal prefixes, and most of them have

[1] On the category of verbal aspect in Russian see Comrie (1976).

several meanings.[2] This exactness is especially noticeable for verbs of motion.[3] Prefixes also have a close relationship to the verbal aspect of Slavic languages. The meanings discussed here are not, however, dependent on grammatical aspect, since the prefix discussed here, *pod-*, very rarely functions as a pure means of aspectual word formation (as a semantically empty prefix that merely forms the perfective from the imperfective).[4] The same prefix therefore occurs with verbs of both imperfective and perfective aspect.

1.2 Ambiguity in prefixes and how to describe it

The different meanings of a prefix can be discussed as cases of either homonymy or polysemy, and besides classifying different meanings we can focus on the vagueness of the meanings (Zalizniak 2007: 95–98). In the case of *pod-* the connection has, on the one hand, been obvious enough for the meanings to be seen as polysemous when they involve various concrete notions having to do with 'under'. On the other hand, the abstract meanings of *pod-* have commonly been considered completely separate and simply coinciding, i.e. homonymous.

Previous studies on Russian prefixes have differed both in the metalanguage used for describing the meanings and in their theoretical backgrounds. They have described the meanings of different Russian prefixes with the help of words belonging to natural languages. Other ways of describing these semantic connections include different forms of metalanguage such as formal semantics (Dobrušina, Mellina & Paillard 2001), structuralist formal models (Flier 1975, 1985; Gallant 1979), or cognitive schemata (Janda 1986, 1988). A detailed investigation into Russian prefix studies up to the late 1990s can be found in Krongauz (1998: 55–98). During the last fifteen years, various methodological and theoretical issues in the study of prefixation in Russian have been discussed

[2] In this paper the notion of word meaning is not problematized. It is taken for granted that a word or other meaning component has a meaning or several meanings that can be listed and compared.

[3] In Slavic languages verbs of motion are regarded grammatically as a separate group, since they differ from other verbs in their grammatical behaviour.

[4] Adding a prefix may bring in redundant information, which has led to the interpretation that some uses of prefixes are semantically empty. The existence of empty prefixes has been denied by several contemporary scholars (see e.g. Janda 1986; Krongauz 1998).

by Dobrušina, Mellina & Paillard (2001), Norwegian scholars from Tromsø (Nordlyd 2004), Gehrke (2008), Plungjan (2001), and by both Russian and Western scholars in (Krongauz & Paillard 1997). What these approaches to prefixes have in common is that they hardly mention analogy as a way of explaining correspondences in the meanings of prefixes.

New tendencies in polysemy studies in general include different ways of explaining the place of polysemy in the language system and describing the cognitive system behind it (see Evans 2009; Rakova 2003; Rakova, Pethő & Rákosi 2007). This reinforces the development that started in the 1980s (Janda 1986). There is, however, a need for new, more precise ways of describing the polysemy of Russian prefixes.

1.3 The meanings of the prefix *pod-*

The prefix *pod-* has many meanings that are not self-evidently related to one another. Plungjan (2001) offers the most detailed listing of the various meanings. The concrete meanings of *pod-* have to do with either approaching or being under or down ('being under', 'being down', 'going under', 'from under', 'upwards', or 'downwards', depending on the verb the prefix is connected to and on the context in which it is used). The abstract, non-spatial meanings of *pod-* present a curious combination. Plungjan (2001) divides these meanings into three categories: (1) describing supplementary action (additional, simultaneous, or lesser action); (2) causing harm or damage or doing something in secret; and (3) approaching an object (modifying the action according to someone else's needs or to fit into something; imitating something).

The meaning of a prefix can be determined by comparing the meaning of a verb with and without the prefix. One should not, however, be too simplistic about this: in actual fact, the meaning of a prefixed verb is not always the meaning of the prefix plus that of the core verb, but is also affected by other characteristics of both the prefix and the verb and by the interaction between the two (see Dobrušina, Mellina & Paillard 2001). These relevant characteristics include the grammatical properties of both (valence relations), and the semantic properties affecting them. For example, the prefix *pod-* has the following grammatical properties: (1) It combines both with verbs of motion and with other verbs; (2) It combines with both transitive and intransitive verbs; (3) If the verb with *pod-* is used with a preposition, the preposition is *k* (+ dative) with the verbs of motion and *pod* (+ accusative) with other verbs. The verb stem *pevat'-pet'* 'to

sing', which will appear later in example 11 (c), has the following grammatical properties: (1) It is transitive; (2) Its valence relations include the possibility of telling not only who sang and what, but also what the song was about and who wrote it. Semantically, in the combination *podpevat'-podpet'*, *pod-* brings in the meaning of singing along. The grammatical characteristic realized in this verb is that it requires the dative case without a preposition, the word in the dative indicating what we sing along to. Thus, in this particular case adding the prefix not only specifies the meaning of the verb − singing along is a particular kind of singing − but also changes its grammatical behaviour.

In this paper I discuss two meanings of *pod-*: the concrete spatial meaning of approaching something and the meaning(s) that I call abstract approaching. Concrete approaching means physical movement towards, while abstract approaching is coming closer physically and mentally, changing one's opinion in the direction of somebody else's, or approaching a person in an abstract way so as to be or seem to be mentally closer to him/her (in order to achieve something). As for their grammatical behaviour, prefixed verbs with *pod-* in the meaning of concrete or abstract approaching are often followed by the preposition *k* 'towards' + a noun in the dative indicating the person or thing that the approaching is directed at.

I start by introducing the verbs in question through some examples of prefixed verbs with these meanings. The concrete meaning of approaching is commonly acknowledged for *pod-*. This can, first of all, be movement towards a dominating object (the concrete cases of adding *pod-* to motion verbs), for example (1a)–(1b).

(1) a. *pod-beža-t'*
under[5]-run-INF[6]
'to run (close) to'

[5] The prefix pod- has been glossed throughout as 'under'. I interpret the concrete meaning of approaching that the prefix expresses with the verbs of motion as a case of being under, since horizontal and vertical movement are conceptually related (see Luodonpää-Manni & Viimaranta 2010). This can also be interpreted as having to do with the notion of domination (Viimaranta 2012).

[6] The abbreviations used in the glosses: INF = infinitive; REP = repetition, REFL = reflexive

b. *pod-kovylja-t'*
under-stumble-INF
'to hobble up to'

Second, the verbs can also be causatives, as in (2a)–(2c). These verbs mean bringing someone or something close to something and in this way causing the approaching. Some of them (3a)–(3b) are, besides being causative, also reflexive. The one that summons or lures someone else causes the other person to come close to him/her.

(2) a. *pod-ves-ti*
under-bring-INF
'to bring up to'

b. *pod-voloči-t'*
under-drag-INF
'to drag up to'

c. *pod-kati-t'*
under-roll-INF
'to roll up to'

(3) a. *podo-zva-t'*
under-call-INF
'to call up', 'to summon'

b. *pod-mani-t'*
under-allure-INF
'to beckon'

Let us now look at examples of abstract approaching. The notion that I will refer to as abstract approaching has been described by Plungjan (2001) in terms of three categories: "approaching the object", "imitation of a model", and "ingratiating". In the first of these two groups a non-physical action is performed in order to provoke a development that is favourable to the speaker either as such or because of its results. This involves a process of either stimulation (for example, (4a)–(4b)) or adjustment (for example, (5a)–(5b)).

(4) a. *pod-zadori-t'*
under-provoke-INF
'to egg on'

b. *podo-gna-t'*
under-chase-INF
'to urge on'

(5) a. *pod-gotovi-t'*
under-prepare-INF
'to prepare for'

b. *podo-j-ti*
under-walk-INF.
'to be suitable for'

After giving examples of the different meanings of *pod-* it should be clear on what basis I see all of these cases as representatives of a single meaning category. What the various types of abstract approaching have in common is the idea that when the subject moves closer to the object, it adjusts itself in such a way as to get mentally closer to the object. This adjustment is a central notion for what I call abstract approaching. Adjustment means here conforming to other peoples' expectations and changing one's behaviour accordingly.

2. Analogy and conceptual metaphor

How should we deal with polysemy interpreted as a meaning change from concrete to abstract, proven by the use of the same linguistic units? Two competing explanations are offered in this article. The first of these is that concrete and abstract notions are expressed by the same means because of analogy. The second alternative is based on conceptual metaphor theory, according to which the concrete domains in language influence the way the abstract domains are talked about because concrete concepts offer source domains for abstract concepts.

2.1 Analogy

Let us start with the term *analogy*. Itkonen (2005: 15) defines analogy as structural or functional similarity. Humans constantly seek analogical

models and classify them on the basis of their characteristics as compared to other models. This happens unconsciously. (Blevins & Blevins 2009.) Analogy is important because it involves a process of pattern-seeking (the word used by Blevins & Blevins (2009: 1)), which plays a crucial role in human behaviour and conceptualization. What is meant by pattern-seeking is the unconsciously working human capacity to find similarities, repeating patterns, and causal connections everywhere. This capacity is developed in humans in early childhood, and its most primitive forms, found in children from 10 months onwards, consist of noticing changes in one's environment and making connections between events and objects (Goswami 2001).

On the system level, analogy means that the functioning of a system is modelled by another system. More concretely, in analogy the functioning of two systems is similar because of the structural correspondences in the systems. The way that analogy works has been illustrated with various models that all describe analogy as making a connection or reference from one entity to another, drawing conclusions about the similarity in the functioning of the parts of different systems.

In linguistics the notion of analogy is traditionally used especially often in phonology and morphology, both of which include many examples of cases in which the tendency to become more similar on the basis of imitation is very clear. Analogy has also been used in language typology as a means of explaining similarity relationships between languages. When it comes to means of description, the linguistic notion of analogy has been described with the help of different kinds of equations and tables in which the corresponding notions (and sometimes also their functional units) are compared.[7] This kind of description assumes that the analogy relation, based on perceived similarity, is by its nature straightforward. The similarity relationship described in terms of analogy can be of different kinds, depending on the number of factors taking part in it. The similarity results in a change in the members of the paradigm for a certain part of the word. No matter what the concrete form of the description for analogy is (a list, feature matrix, or table), it does not assume change in the resulting form (except for the part under analogy). In this way, analogy does not change the original model that contributed to the analogy.

The importance of analogy has been seen differently in different times and disciplines.

[7] The information provided in the tables can be a result of very different kinds of operations on data, see Skousen, Lonsdale & Parkinson 2002.

Linguistic analogy has long been recognized as a major factor in language acquisition and language learning. According to Anttila (1977: 23), analogy is so important for linguistics because the very structure of language is analogical. It has also been seen (along with association and categorization) as a major factor in semantic change (Győri 2002). The study of analogy from the (cognitive) linguistic viewpoint has a strong link with the study of artificial intelligence and computer models (see papers in Helman 1988; Holyoak, Gentner & Kokinov 2001) and the psychological study of reasoning and memorizing (see Vosniadou & Ortony 1989). Analogy also has its place in the philosophy of science (Hesse 1988).

The role of analogy among different processes of conceptualization and learning has also provoked discussion. It has been claimed that analogy is equal to all reasoning and that it is the core of human cognition (Anttila 1977; Hofstadter 2001; Penn, Holyoak & Povinelli 2008; Blevins & Blevins 2009), the main mechanism of memory construction that ultimately explains all learning (Kokinov & Petrov 2001), a process used all the time in learning, reasoning, decision-making, creation of culture, and scientific reasoning (e.g. Gentner & Gentner 1983; Niiniluoto 1988; Holyoak & Thagard 1995), or an age-old phenomenon that was already in use thousands of years ago in poetry, philosophy, and religion and continues to be used constantly (Holyoak & Thagard 1995; Holyoak, Gentner & Kokinov 2001). Quite to the contrary, it has also been claimed that analogy is not a single cognitive process and because of this it cannot be given such a central position as has been done in the studies mentioned above (Hoffman, Eskridge & Shelley 2009). Itkonen (2005: 199) criticizes the notion that analogy is the basis for all thinking; in his view, the concept of "human thinking" would have to be defined too generally in order to come up with such a simplification. As a cognitive process, analogy has also been explained in relation to other processes such as categorization (Turner 1988).

The idea of different stages of analogy has also been used in creating various computer models for analysing analogy. One such model is that of Holyoak & Thagard (1995: 116–137), who have discussed the stages of selection, mapping, evaluation, and learning. These stages are unconscious cognitive processes that explain the importance of analogy in our conceptualization. There exists independent evidence for each of these stages.

Seeing the relationship between interconnected concrete and abstract things through the notion of analogy suggests that a concrete meaning has

been extended to cover abstract notions that have some kind of connection or similarity relationship with it. Analogy is the process that functions as a combining notion between several categories. It is based on structural similarities.

2.2 Conceptual metaphor theory and blending theory

Our alternative explanation, based on conceptual metaphor theory (Lakoff & Johnson 1980, 1999), assumes that metaphor, i.e. conceptualizing a thing through something else, is a central mechanism in human conceptualization. In this article I make use of the way that the blending theory (e.g. Fauconnier & Turner 1998, 2002) describes conceptual metaphors and other conceptual blends, since in considering how conceptualization is formed this is more precise than the traditional metaphor notation (e.g. SIMILAR IS CLOSE).

Conceptual metaphor theory assumes that in metaphorical conceptualization, which is one of the basic mechanisms of our conceptualization, a certain conceptual domain is understood via another domain, for example similarity in terms of proximity. In the process of conceptualization selected features of one thing are projected onto another thing as a result of the process called blending that involves two different input spaces (which correspond to conceptual domains in conceptual metaphor theory if we assume the blend to be conventionalized). The characteristics of the different input spaces form a new mental space, the blend, into which selected features of the different inputs are projected with the help of the correspondences set by a separate mental space known as the generic space, which makes the blend possible. The generic space contains the possible common features of the two input spaces.

The use of the notion of blends does not in itself contradict the possibility of assuming that the blend also involves analogy as a way of forming the connections. Nevertheless, since the notion of blend, as is clear from its name, always includes the idea of mixing different categories, a pure analogy is, in my opinion, not a blend, because it does not involve the mixing of domains and it does not bring about change in the original domains or mental spaces. Another issue altogether is that pure analogy may be much rarer than one would think, and in many cases where analogy is seen there are actually blends or different forms of conceptualization processes following one another.

2.3 Metaphor and/or analogy

The difference between metaphor and analogy requires further commentary. Both terms are often used without specifying the meaning of the term and sometimes interchangeably. They are also described in a very similar manner. For example, the notions of embodiment, projection, and mapping that are central especially in the metaphor studies of 1980s are used in analogy literature of the same period as well. The same goes for forms of description used in those studies such as the use of image schemata in describing both. See, for example, Johnson (1988).

Fauconnier & Turner (2002) deal with the question by explaining that in analogical projections the source domain is mapped onto a target domain so that inferences easily available in the source are exported to the target and we can thus reason about the target on the basis of a structure-mapping only, while in conceptual blending the blend consists not only of the correspondences between structures, but forms a new domain that is in use when we think and talk about the notions involved. Further, they use the term *analogy* to describe a vital relation that depends upon role-value compression in a blend (ibid: 98–99). In this way it is obvious that Fauconnier & Turner (ibid) do not see analogy and conceptual blending as mechanisms of the same level, but regard the role of analogy as much narrower.

Many scholars have seen analogy and metaphor as separate but related concepts, since both are cases of cross-domain mapping. Holyoak & Thagard (1995) think that the mental processes behind the two are partly the same, but the phenomena themselves are not. For them metaphor is a figurative device linked to analogy. Gentner et al. (2001) also deal with metaphor and analogy as (partly) different phenomena. For them, metaphor, analogy, and similarity are related concepts that cannot be fully differentiated from one another. They distinguish between novel and conventional metaphors, and for them only novel metaphors are mappings between different domains and as such similar to analogy. The difference between metaphor and analogy as they see it is that many metaphors are also analogies, but not vice versa, because metaphor can be, besides analogy, based on shared features, or on both analogy and shared features. Therefore Gentner et al. (2001) call conceptual metaphors "extended analogical mappings".

Analogy and metaphor can also be seen as representing the same phenomenon. According to Blevins and Blevins (2009: 7–8), metaphor is

simply a semantic analogy. Itkonen (2005: 35–44) sees metaphor as a kind of analogy.[8] Onikki-Rantajääskö (2001: 34–38) sees analogy as schematic and metaphor as one form of analogy. This interpretation means that analogy is a very common tendency in language, explaining much of its formal regularity and change.

Thus, there is no commonly accepted truth on the relationship between analogy and conceptual metaphor. My view of analogy is that it is a pattern-seeking process of conceptualization in which a cross-domain mapping takes place. This mapping is based on structural similarities in the systems. These similarities provoke in our system of conceptualization the unconscious need to deal with different issues as the same on the basis of their similarity. Conceptual metaphor, for its part, is also a cross-domain mapping, but in it the similarity that explains the mapping is formed in the process whereby features of the domains change into something new in blending. If we wish to use the term *analogy* for a process which is seen to create objects that have characteristics not fully explainable by the predecessor in the analogical chain, we need to find more precise ways of describing this kind of analogy. The notion of blend is useful especially for the way that it helps to explain the changing power of analogical processes.

3. Material and methods

The material analysed consists of a list of 73 verb infinitives. They are part of a corpus consisting of the 501 prefixed verbs with *pod-* found in the entry for *pod-* and in the separate entries for words with *pod-* in *Bol'šoj Tolkovyj Slovar'* (Kuznecov 1998), the most comprehensive single-volume dictionary of contemporary Russian. The 73 verb infinitives were selected on the basis that only they represent the meaning categories of concrete and abstract approaching. The verbs included are listed in the Appendix, which also shows whether each of them can be used only in the meaning of abstract approaching, only in the meaning of concrete approaching, or both. Some of the prefixed verbs included have other meanings in addition to concrete and/or abstract approaching, but in this paper those meanings are not commented upon or even included in the glosses.

[8] Itkonen also criticizes the way that blending theory is sometimes used in a very vague manner, in the sense that could in his opinion simply be called combining (2005: 43–44).

The material illustrates two related meaning categories whose interrelationship is easy to prove because these same prefixed verbs are so systematically used in both meanings. The purpose of this material is to illustrate the way that differentiating between analogy and conceptual metaphor affects the interpretation of the difference between meaning categories and the functioning of polysemy.

The infinitive forms in themselves show us with which verbs the prefix can be combined and what meanings result from this combination. I use examples from the Russian National Corpus (www.ruscorpora.ru) to further illustrate the meaning of the verbs. These examples have not been used in classifying the prefixed verbs: the classification is based solely on the infinitive forms. This article uses the notions of Landmark (LM) and Trajector (TR) to describe the way in which one object (TR) moves in comparison to another (LM) (see Langacker 1987). These notions have been used in describing the meanings of prefixes (especially the concrete spatial ones). (For example, Janda 1986.)

4. Results: Concrete and abstract approaching and explaining them by analogy or metaphor

The analysis of the material shows a number of tendencies in the meanings of concrete and abstract approaching. Coming close in concrete terms means that the thing moving comes up to a certain limit but does not touch the landmark. Touching can still be the intention of the actor or even presupposed to happen immediately after reaching the destination. The combination of *pod-* with a verb does not, however, assume it. See examples (6)–(7).

(6) *pod-gres-t-is'*
under-row-INF-REFL
'to row oneself up to'

(7) *podo-dvi-nu-t'*
under-move-little-INF
'to move closer'

In (6) *pod-* is used to indicate rowing up to the point of destination. It is reflexive, so the subject is assumed to row him/herself where needed.[9] In this way, (6) demonstrates a case in which the subject is the moving trajector and the destination is the landmark. For example, ***podgrestis'*** *k beregu* 'row to the shore, get to the shore by rowing'. In (7), the destination will be given with the prepositional construction *k* + dative. The object that is moved is a trajector that moves closer to something; the extent to which it moves can be expressed, but this is optional. The main point seems to be that after the movement the trajector will be closer to the landmark. For example, *Kazalos' by, prosto:* ***pododvinut'*** *stul k rojalju i igrat'*. 'It would seem to be an easy task − to **move** the stool **closer** to the piano and play'.

Whereas in concrete approaching the movement brings the object closer, in abstract approaching the coming closer is not physical, but nevertheless has a physical basis in our conceptualization since abstract closeness is related to the need for some kind of communication or exchange. Adjustment is a central notion for abstract approaching. The action involved is modified or adjusted in one way or another to make it more suitable or profitable. Different kinds of adjustment include both such notions as toadying and bribery and more positive things such as inspiring and making more suitable. For example:

(8) *pod-kupi-t'*
under-buy-INF
'to bribe', 'to make an impression on'

(9) *pod-ygr-yva-t'*
under-play-REP-INF
'to toady'

In (8) the person can be 'bought from under', i.e. bribed. This approaching comes with a concrete reward. It can also mean the kind of buying that does not involve money, attracting someone to one's side by actions that are so overwhelmingly impressive or inviting that the object of 'buying' is won over. The achievements of the subject in this case are estimated to be equivalent to a substantial amount of money in their power to win others over. For example *Volodja okončatel'no* ***podkupil*** *moe serdce vkusnymi,*

[9] The same verb can also be used without the reflexive postfix -sja (realized in this case as -s').

mjagkimi buločkami. 'Volodja finally **won** my heart with delicious soft buns.'

In (9), toadying in order to get some kind of benefit from approaching has to do with the way that friends and family favour one another even unconsciously because of their physical closeness. Trying to get oneself into such a position, i.e. approaching in opinion or pretending to do so, is expressed by combining the verb with the notion of approaching, signalled by the prefix *pod-*. For example, *Poetomu my **podygryvaem** toj auditorii, kotoraja važna našemu reklamodatelju.* 'That's why we **play up to** the audience that is important for our advertiser.' There are actually 11 different verbs in the material with the meaning of toadying or ingratiation. Many of them, for example (10a)–(10c), can be translated as 'to lick someone's boots'.

(10) a. *pod-lasti-t'-sja*
under-make up to-INF-REFL
'to lick someone's boots'

b. *pod-liza-t'-sja*
under-lick-INF-REFL
'to lick someone's boots'

c. *podo-l'sti-t'-sja*
under-tempt-INF-REFL
'to lick someone's boots'

In these verbs approaching takes the form of putting oneself (seemingly) under someone else in order to please him/her. This means adjusting one's behaviour in a way that pleases the other person. The idea of doing it from the downside where there is less power is also included, as is the idea of putting oneself down by licking things that are considered to be dirty. For example *Dela svoji ispolnjal ispravno, v durnom ne zamečalsja…razve čto **podol'stit'sja** ne umel.* 'He did his work and did not give reason for complaint…the only problem was that he did not know how to **lick the others' boots**.'

Adjustment can also take the form of incitement or inspiration. The example (4a) *podzadorit'* involves abstract approaching in the form of provoking someone into doing something. In this case the provocation means supporting the decision to do something, bringing closer to doing it. For example *Nu, čego že vy stoite? **podzadorila** ona detišek. – Begite,*

darite mamam… ‘Why are you standing there? she **urged** the children. – Go and give the presents to your mums’.

Coming close can also mean being suitable. The example (5b) *podojti* means concretely going close to something. In its abstract uses it indicates suitability. For example *Edinstvennoe, čto ja mogu posovetovat’, čitat’ kak možno bol’še po sootvetsvujuščej tematike, iskat’, vybirat’ to, čto možet* ***podojti*** *Vam i Vašemu rebenku*. ‘The only advice I can give is to read as much as possible on this subject, look specifically for something that **suits** you and your child.’ The idea of coming close as being suitable for something involves presenting the notion of usability and suitability as reachability. Circumstances that make things possible are things that are close enough.

Podgotovit’ (5a) means preparing something with adjustments and slight changes. If the same verb stem is used without the prefix, the action involved is made in a more general manner, without paying attention to the concrete details of adjustment. For example: *Moja zadača –* ***podgotovit’*** *reformu sejčas.* ‘My job is to **prepare** the reform now.’ This indicates that when the reform is planned the preparations involve taking a close look at the concrete adjustments needed in this concrete case.

Another kind of adjustment takes place when a secondary action is formed in the shadows of another action. These cases involve an adjustment in order to fulfil the expectations having to do with a particular role. For example: (11a)–(11d).

(11) a. *pod-da-k-iva-t’*[10]
under-yes-say-REP-INF
‘to say yes to’, ‘to nod along’, ‘to echo’

b. *pod-mah-iva-t’*
under-wave-REP-INF
‘to wave in the same rhythm’

c. *pod-pe-va-t’*
under-sing-REP-INF
‘to sing along’

[10] Folk etymology finds here also the stem *-kiv-* meaning ‘to nod’.

d. *pod-tveržd-a-t'*
under-assure-REP-INF
'to confirm'

In these verbs the secondary action is made abstractly close to another action that it accompanies. The same movement (a), rhythm (b), melody (c), or model (d) that is being echoed or imitated gives the common ground for the connection between the actions. For example: *A ved' on, sobstvenno govorja, daže ne trepalsja, a prosto glupo i slepo* ***poddakival*** *svoemu sobesedniku*. 'And he was, strictly speaking, not even fluttering, but just **echoed** his interlocutor stupidly and blindly.' Another example of a similar meaning with 7c) could be *Vkjučala svoju kassetu i* ***podpevala***. 'She put her cassette in the recorder and **sang along**.'

Even if the case of confirming may seem different, what it has in common is the way the action is being modelled by another kind of action. For example *Ih naznačenie –* ***podtveržđat'*** *podlinnost' elektronnoj podpisi*. 'Their purpose is to **confirm** the authenticity of the electronic signature'. Confirming, as opposed to just affirming (*utveržđat'*), includes the notion of imitating, since the thing confirmed is approached through similarity.

In meanings involving abstract approaching, as in the examples above, the trajector that moves in comparison with the placement of the landmark is the person that approaches another in opinion, while the other person is the landmark. The notion of approaching, moving towards something is in itself not physical, and as such the increase in proximity cannot be measured in absolute physical terms: it can only be relative to the previous situation. Since the motion takes place between people (although this is sometimes presented as ideas coming closer), both trajector and landmark are capable of moving, but in reality they do not necessarily move at all. Both meanings can take both animate and inanimate actors.

Cross-domain mappings happen both analogically and due to conceptual blending. When interpreted as analogy, the different parts of separate systems form connections in such a way as to make us see the similarities involved. When talking about conceptual blending, the relationship consists not merely of (unconsciously) noticing the similarities in functioning, but the resulting blend shows how in our minds the two previously separate domains form a new domain that includes characteristics from both original domains. On the one hand, if we assume the relationship between concrete and abstract approaching to be analogical, we see that the classifying machine in our heads automatically notices the corresponding parts of these two systems and because of this

allows us to use the same means − in this case the prefix *pod-* − for describing both. If, on the other hand, we assume the relationship between concrete and abstract approaching to be metaphorical (a blend), we see the parts of these systems not only as having functional correspondences, but also as forming a new domain, which blends some characteristics from both original domains but has an independent existence. Furthermore, if we want to explain the existence and use of a certain meaning category, in this case the use of the same prefix for both concrete and abstract approaching, the thing being looked at is not the relationship between concrete and abstract, but the relationship between concrete approaching and something else that makes possible the creation of the meaning category (and thus also the mental space) of abstract approaching.

Interpreted as an analogy, the corresponding notions of concrete and abstract approaching are listed in Table 1.

Table 1. The analogy of concrete and abstract approaching

CONCRETE APPROACHING	ABSTRACT APPROACHING
physically close	mentally close
going closer physically	changing to be more similar-minded, adjusting
able to touch	able to come into contact with
able to hear	able to be influenced
the one that approaches	flatterer/imitator/adaptable person
the thing being approached	the one imitated/flattered/dominating
doing "close to"	doing in a similar manner
under	influenced

In thinking about how abstract approaching has acquired the meaning that it conveys in modern language use, one could first assume that this is a case of analogy. Coming concretely and abstractly closer have several things in common − many features that make abstract approaching seem like concrete approaching. This impression is reinforced by the use of the same verbs in both meanings. The things involved in abstract approaching, such as becoming more similar, have a concrete basis in the sense that people have a tendency to wish for more intimacy with people they agree with and have most in common with. In this way, the wish for concrete closeness is blended in our understanding with the wish to benefit from abstract closeness. Nevertheless, the shift is not total − concrete meanings have

retained their significance, and the notion of abstractly coming closer (in opinion) has only been added to them.

This description of an analogy shows how concrete and abstract approaching have corresponding functional parts. The analogy interpretation would assume that when talking about abstract approaching the functional similarities between these parts − such as the way we go physically closer in the domain of concrete approaching and go mentally closer in the domain of abstract approaching − make it possible for us to use the same prefix for both notions. The table showing analogical correspondences between the two meanings cannot describe the way that this analogy has developed. Historical linguistics can provide further evidence of the order and timing of analogical processes.[11]

The metaphor interpretation suggests that our conceptualization really blends conceptual domains when we use *pod-* for abstract approaching. This entails the assumption that concrete approaching is more primitive as a notion than abstract approaching. In this case the complex metaphor (blend) involved is Adjustment Is Approaching. Relating to the list of primary metaphors presented by Lakoff & Johnson (1999: 50–54), the primary metaphors involved here include at least Intimacy Is Closeness, Similarity Is Closeness,[12] Control Is Up. Primary metaphors are not blends because they are based on mappings between single concepts that draw on our everyday bodily experience and not on entire conceptual domains (or mental spaces). Primary metaphors can, nevertheless, function as inputs to blending. The primary metaphors mentioned here are based on evidence provided in earlier studies. The primary metaphor Similarity Is Closeness justifies the notion that approaching indicates an increase in similarity. Intimacy Is Closeness, for its part, makes understandable the way that abstract approaching uses approaching as a means for imitating and ingratiating. The primary metaphor Control Is Up explains the way that ingratiating and imitating work. Since the person in control is metaphorically above the others, attempts to get closer to that level – whether by imitating, ingratiating, or adjusting one's actions − are conceptualized as involving vertical movement (see Lakoff & Johnson 1980; Luodonpää-Manni & Viimaranta 2010; Viimaranta 2012).

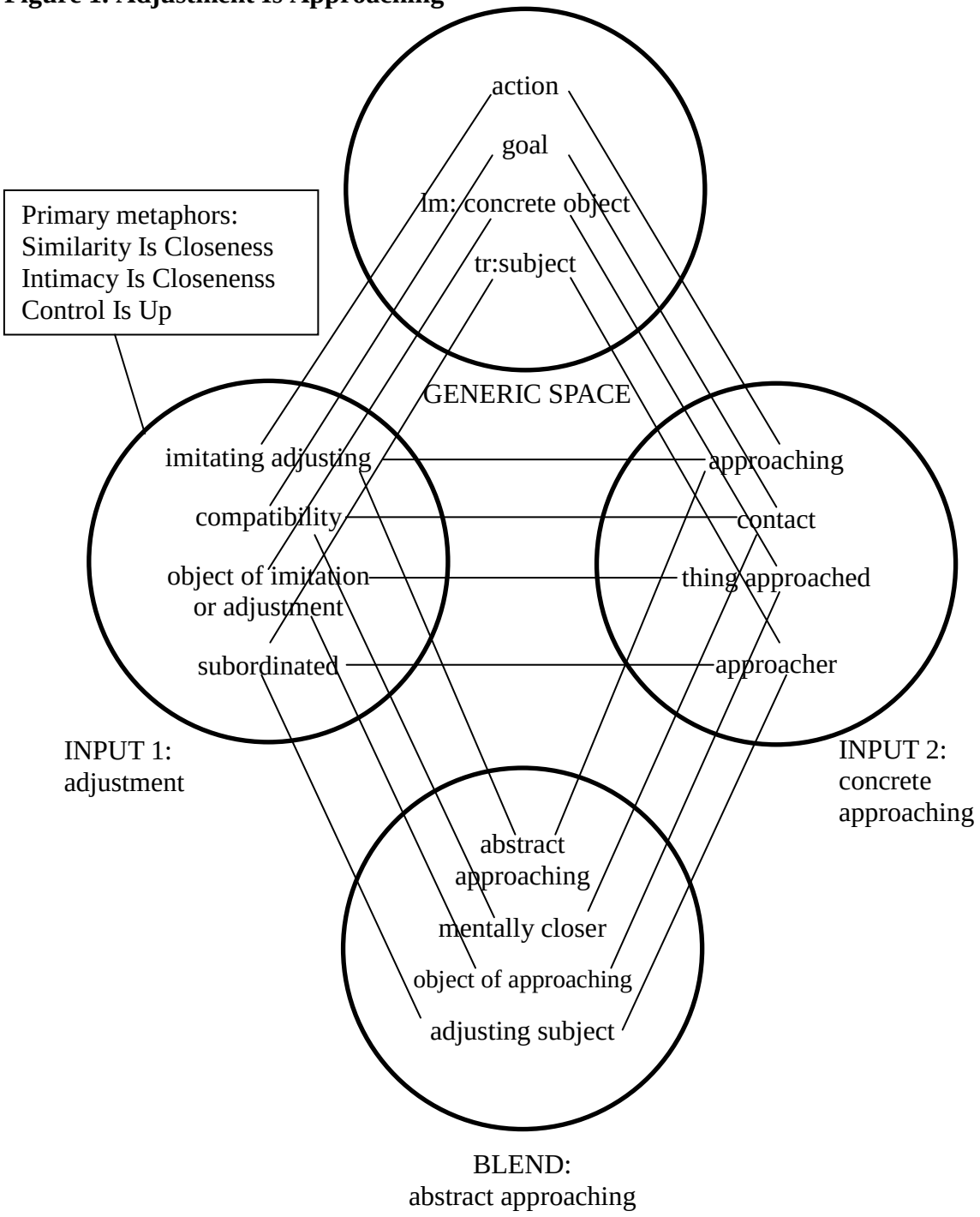
[11] As noted earlier, these processes and their stages can also be tested by experimental means (Holyoak & Thagard 1995), but these experiments have concentrated on the creation of new analogies as evidence for how the process works.

[12] The same conceptual metaphor has been called, at least, SIMILAR IS CLOSE, SIMILARITY IS CLOSENESS, and SIMILARITY IS PROXIMITY.

Designating the conceptual metaphor in question as Adjustment Is Approaching requires further explanation. The notion of adjustment is the target domain that provides us with the features needed to talk about abstract approaching. Concrete approaching is the source domain involved when talking about abstract approaching. The blending theory (for example, Fauconnier & Turner 1998, 2002) illustrates a conceptual metaphor as combining different mental spaces rather than having a one-way transfer from one conceptual domain (source) to another (target). The domains combined are described as input spaces, the things that they have in common (that make the blend possible) as the generic space, and the metaphor resulting from all this as the blend. The form of blend in this type of case that represents a metaphor is a single-scope network. This means that both inputs contain distinct frames and the blend is structured by one of the input frames.

In this case (see Figure 1) the input frames involved are Input 1 (Adjustment) and Input 2 (Concrete approaching). This means that the frame of adjustment is seen in terms of concrete approaching. As the evidence from Russian shows, this blend is conventionalized. It is a conceptual metaphor that has to do with our permanent ways of thinking. The blend Adjustment Is Approaching is a complex metaphor. In the case of both concrete and abstract approaching, the landmark is a concrete object, either animate or inanimate. In the blended space the landmark is the object of imitation or some other kind of abstract approaching. The action involved in Input 1 is adjusting or imitating; and it involves in Input 2 concrete approaching, which results in the action of abstract approaching in the Blend. The primary metaphors Similarity Is Closeness, Intimacy Is Closeness and Control Is Up affect the way that Input 1 is built. In both the input spaces 1 and 2 the action has a goal, which is compatibility in Input 1 and contact in Input 2. These goals are blended into being mentally closer in the Blend. The landmark involved is the object of imitation or source of adjustment in Input 1 and the thing being approached in Input 2, resulting in the object of abstract approaching. The trajector that moves is, in the case of concrete approaching, the subject of the sentence, concretely going or ending up closer to the landmark. In the case of verbs expressing a reflexive meaning, the trajector can also be the object that the subject calls (etc.) to his sphere of influence. In Input 1 the trajector is a subordinated subject and in the Blend the corresponding notion is the adjusting subject.

Figure 1. Adjustment Is Approaching



In all, this particular case does not contradict the other things known about the difference between analogical and metaphorical mappings. If the meaning of abstract approaching that the Russian prefix *pod-* has could be wholly explained in terms of pure analogy, this would mean that the central

(or "diagnostic", in the terminology of analogy literature) features of concrete approaching would have been mapped as such onto abstract approaching and new features would not be necessary. For example, what in the domain of concrete approaching is the physical approaching itself, would be analogously mapped into the domain of abstract approaching without any additional features. So abstract approaching would simply be seen as concrete approaching since the analogy created between those quite different actions (moving towards and adjusting or imitating) would emphasize their similarity.

Dealing with the notions of concrete and abstract approaching as a case of analogy enables us to explain in the form of a table (see Table 1) the way that the meaning shift from concrete to abstract is based on certain correspondences between the fields. These correspondences can be seen to indicate the way that the human mind constantly seeks for similarity in patterns. The evidence from language, the way that the Russian prefix *pod-* is used both for concrete and abstract approaching, proves the point. The analogy interpretation is not, however, able to give any detailed information on either the content or the origin of the correspondences between concrete and abstract approaching. In this respect the metaphor interpretation is more illustrative. The two interpretations are, in the end, compatible in many ways.

5. Discussion

An ideal model of describing polysemy in Russian prefixes would be able to describe the connections between different uses (meanings) of the same prefix. Historically, prefixes have been formed from prepositions (or rather, the two have a common ancestor). The different uses of a certain prefix are related in a way that reflects both the historical development and the constant conceptualization processes of humans. As processes central to our conceptualization, both conceptual metaphor (and metonymy) and analogy are worthy of attention. Analogy is a form of cross-domain mapping in which the correlations noted are based on functional correspondences between the parts of the analogous systems.

The difference between two possible assumptions − that the relationship between concrete and abstract approaching is metaphorical, or that it is just one case of analogy widely found in language − is not only theoretical. This difference has to do with the whole issue of how we understand human conceptualization. If we assume that abstract

approaching in its different forms is merely a case of widening the meaning by analogy, we recognize in our conceptualization the ability to see similarities and make generalizations. On the other hand, if we assume this conceptualization to be metaphorical, we assume that human conceptualization is based on processes that are built not only on recognizing similarities, but also on making conceptual blends in which analogy is just one of the means employed. In this particular case, it is tempting to see the relationship between those meanings of *pod-* that I call concrete and abstract approaching as a case of analogy. The relationship between the meanings seems straightforward enough.

Analogy is a very broad notion used both in linguistics and in other fields. It is hard to deny its importance. If the description includes only the corresponding functional parts of different systems, the concrete functioning of the process is left undescribed. Conceptual metaphor theory is able to explain why we do not even notice that we use the same prefix for concrete and abstract notions. This is a result of conceptual blending that creates a whole new conceptual domain. Conceptualization of the abstract through something concrete is a widespread phenomenon. Conceptual metaphor theory traditionally sees the difference between abstract and concrete notions as a central reason for metaphorical conceptualizations. Coming concretely close can be perceived with our senses and concrete closeness can be measured unambiguously. Nevertheless, even concrete closeness can have abstract consequences – what is close enough or too close can be different for different people. Especially when the approaching objects are human beings, this aspect of the question can be of the utmost importance.

6. Conclusions

The starting point for this article was the way that languages (in this case Russian) can use the same linguistic units when expressing different meanings. The specific case studied was the use of verbs with the prefix *pod-* indicating approaching in concrete and abstract meanings. Analogy and conceptual metaphor were discussed as possible explanations for concrete and abstract meanings of the same combination of a verb with a prefix. The notion of analogy seems to offer a good explanation of how the functioning of concrete approaching has widened into abstract approaching. This relationship assumes that the characteristics of concrete approaching, and more specifically the way that the system of correspondences works in

it, has been analogously transferred into another system, that of abstract approaching. Dealing with some rather exotic meaning categories for *pod-* under the combined notion of abstract approaching helped to see the similarities between these three different abstract meanings and their relationship to concrete approaching.

According to my analysis it is nevertheless impossible, at least in this case, to explain all aspects of the relationship between concrete and abstract approaching with the notion of analogy only. The reason for this is that the cross-domain mappings involved include more complex cases of blending than mere analogy. In this way, although analogy can (also in this case) explain a great deal more than conceptual metaphor theory gives it credit for, it is not quite able to account for all the notions involved.

Thus, the notion of conceptual metaphor is necessary for explaining this case of polysemy, and it can be used to describe the meaning of other verbal prefixes as well. Analogy alone, although an important factor in language in general, is not sufficient for explaining polysemy and the widening and changing of meaning. It is one of a number of processes involved in conceptual blending.

References

Anttila, Raimo (1977) *Analogy*. The Hague: Mouton.

Blevins, James P. & Blevins, Juliette (2009) Introduction: Analogy in grammar. In James P. Blevins & Juliette Blevins (eds.), *Analogy in Grammar. Form and Acquisition*, pp. 1−12. Oxford: Oxford University Press.

Comrie, Bernard (1976) *Aspect: an Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: Cambridge U.P.

Dobrušina, E.R.; Mellina, E.A. & Paillard, Denis (2001) *Russkie pristavki: mnogoznačnost' i semantičeskoe edinstvo* [Russian Prefixes: Ambiguity and Semantic Unity]. Moscow: Russkie slovari.

Evans, Vyvyan (2009) *How Words Mean. Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford: Oxford University Press.

Fauconnier, Gilles & Turner, Mark (1998) Conceptual integration networks. *Cognitive Science* 22 (2): 133–187.

—— (2002) *The Way we Think. Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Flier, Michael S. (1975) Remarks on Russian verbal prefixation. *Slavic and East European Journal* 19 (2): 218−229.

—— (1985) Syntagmatic constraints on the Russian prefix *pere-*. In Michael S. Flier & Richard D. Brecht (eds.), *Issues in Russian Morphosyntax*, pp. 138−154. Columbus, Ohio: Slavica Publishers.

Gallant, James (1979) *Russian Verbal Prefixation and Semantic Features: An Analysis of the Prefix vz*. München: Verlag Otto Sagner.

Gehrke, Berit (2008) Goals and sources are aspectually equal: Evidence from Czech and Russian prefixes. *Lingua* 118: 1664−1689.

Gentner, Dedre; Bowdle, Brian F.; Wolff, Phillip & Boronat, Consuelo (2001) Metaphor Is Like Analogy. In Dedre Gentner, Keith J. Holyoak & Boicho N. Kokinov (eds.), pp. 199−253. Cambridge: MIT Press.

Gentner, Dedre & Gentner, Donald R. (1983) Flowing waters of teeming crowds: Mental models of electricity. In Dedre Gentner & Albert L. Stevans (eds.), *Mental Models*, pp. 99−129. Hillsdale: Lawrence Erlbaum.

Gentner, D. Holyoak & K. J. & Kokinov, B. N. (eds.) (2001) *The Analogical Mind. Perspectives from cognitive science*. Cambridge: MIT Press.

Goswami, Usha (2001) Analogical reasoning in children. In Dedre Gentner, Keith J. Holyoak & Boicho N. Kokinov (eds.), pp. 437−470. Cambridge: MIT Press.

Győri, Gabor (2002) Semantic change and cognition. *Cognitive Linguistics* 13 (2): 123–166.

Helman, D. H. (ed.) (1988) *Analogical reasoning. Perspectives of Artificial Intelligence, Cognitive Science, and Philosophy*. Dordrecht: Kluwer.

Hesse, Mary (1988) Family resemblances and analogy. In Dadid H. Helman (ed.), *Analogical Reasoning. Perspectives of Artificial Intelligence, Cognitive Science, and Philosophy*, pp. 317−340. Dordrecht: Kluwer.

Hoffman, Robert R.; Eskridge, Tom & Shelley, Cameron (2009) A naturalistic exploration of forms and functions of analogizing. *Metaphor and Symbol* 24: 125–154.

Hofstadter, Douglas R. (2001) Epilogue: Analogy as the core of cognition. In Dedre Gentner; Keith J. Holyoak & Boicho N. Kokinov (eds.), pp. 499−538. Cambridge: MIT Press.

Holyoak, Keith J.; Gentner, Dedre & Kokinov, Boicho N. (2001) Introduction: The place of analogy in cognition. In Dedre Gentner, Keith J. Holyoak & Boicho N. Kokinov (eds.), pp. 1−19. Cambridge: MIT Press.

Holyoak, Keith J. & Thagard, Paul (1995) *Mental Leaps. Analogy in Creative Thought*. Cambridge, Massachusetts: The MIT Press.

Itkonen, Esa (2005) *Analogy as Structure and Process*. Amsterdam: John Benjamins.

Janda, Laura A. (1986) *A Semantic Analysis of the Russian Verbal Prefixes za-, pere-, do-, and ot-*. (Slavistische Beiträge 192). München: Verlag Otto Sagner.

—— (1988) The mapping of elements of cognitive space onto grammatical relations: An example from Russian verbal prefixation. In Brygida Rudzka-Ostyn (ed.), *Topics in Cognitive Linguistics*. Current issues in linguistic theory 50, pp. 327−343. Amsterdam/Philadelphia: John Benjamins.

Johnson, Mark (1988) Some constraints on embodied analogical understanding. In Dadid H. Helman (ed.), *Analogical Reasoning. Perspectives of Artificial Intelligence, Cognitive Science, and Philosophy*, pp. 25−40. Dordrecht: Kluwer.

Kokinov, Boicho N. & Petrov Aledander A. (2001) Integrating memory and reasoning in analogy-making: The AMBR model. In Dedre Gentner, Keith J. Holyoak & Boicho N. Kokinov (eds.), pp. 59−124. Cambridge: MIT Press.

Krongauz, M.A. (1998) *Pristavki i glagoly v russkom jazyke: Semantičeskaja grammatika* [Prefixes and Verbs in Russian: Semantic Grammar]. Moscow: Jazyki russkoj kul'tury.

Krongauz, M.A., & Paillard, D. (eds.) (1997) *Glagol'naja prefiksacija v russkom jazyke. Sbornik statej* [Verbal Prefixation in Russian. A Collection of Articles]. Moscow: Russkie slovari.

Kuznecov, S.A. (1998) *Bol'šoj tolkovyj slovar' russkogo jazyka* [Comprehensive Dictionary of Russian]. St. Petersburg: Norint.

Lakoff, George & Johnson, Mark (1980) *Metaphors We Live By*. Chicago (IL): University of Chicago Press.

—— (1999) *Philosophy in the Flesh. The Embodied Mind and Its Challenge to Western Thought*. New York: Basic books.

Langacker, Ronald W. (1987) *Foundations of Cognitive Grammar. Vol. 1, Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Luodonpää-Manni, Milla & Viimaranta, Johanna (2010) Metaphoric expressions on vertical axis revisited: An empirical study of Russian and French material. *Metaphor and Symbol* 25 (2): 74−92.

Niiniluoto, Ilkka (1988) Analogy and similarity in scientific reasoning. In Dadid H. Helman (ed.), *Analogical Reasoning. Perspectives of Artificial Intelligence, Cognitive Science, and Philosophy*, pp. 271−298. Dordrecht: Kluwer.

Nordlyd (2004) *University of Tromsø working papers on language and linguistics* 32 (2).

Onikki-Rantajääskö, Tiina (2001) *Sarjoja. Nykysuomen paikallissijaiset olotilanilmaukset kielen analogisuuden ilmentäjinä* [Series. Finnish Locative Expressions of State as Evidence for the Analogical Nature of Language]. Helsinki: Suomalaisen Kirjallisuuden Seura.

Penn, Derek C.; Holyoak, Keith J. & Povinelli, Daniel J. (2008) Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences* 31: 109–178.

Plungjan V.A. (2001) Pristavka pod- v russkom jazyke: k opisaniju semantičeskoj seti [The prefix pod- in Russian: a description of a semantic net]. *Moskovskij lingvističeskij žurnal* 5: 95−124.

Rakova, Marina (2003) *The Extent of the Literal. Metaphor, Polysemy and Theories of Concepts*. Basingstoke: Palgrave Macmillan.

Rakova, M. & Pethő, G. & Rákosi, Cs. (eds.) (2007) *The Cognitive Basis of Polysemy*. Metalinguistica 19. Frankfurt am Main: Peter Lang.

Skousen, R. & Lonsdale, D. & Parkinson, D. B. (eds.) (2002) *Analogical Modeling. An exemplar-based approach to language*. Philadelphia: John Benjamins Publishing Company.

Turner, Mark (1988) Categories and analogies. In David H. Helman (ed.), *Analogical Reasoning. Perspectives of Artificial Intelligence, Cognitive Science, and Philosophy*, pp. 3−23. Dordrecht: Kluwer.

Viimaranta, Johanna (2012) The metaphors and metonymies of domination: explaining the different meanings of the Russian prefix *pod-*. *Russian linguistics* 36 (2): 157−174.

Vosniadou, S. & Ortony, A. (eds.) (1989) *Similarity and Analogical Reasoning*. Cambridge: Cambridge University Press.

Zalizniak, Anna A. (2007) The phenomenon of polysemy and ways to describe it. In Marina Rakova; Gergely Pethő & Csilla Rákosi (eds.), *The Cognitive Basis of Polysemy*. Metalinguistica 19, pp. 93−121. Frankfurt am Main: Peter Lang.

APPENDIX: The verbs studied

Concrete/abstract approaching (9)

podgonjat'-podognat' 'to urge on'
podhodit'-podojti 'to come close, be suitable'
podklikat'-podkliknut' 'to call, to hail'
podmanivat'-podmanit' 'to beckon'
podstupat'-podstupit' 'to approach'
podstupat'sja-podstupit'sja 'to get near'
podtjagivat'-podtjanut' 'to pull up to, tighten'
podvodit'-podvesti 'to lead up to'
podvozit'-podvezti 'to give a lift, to haul'

Concrete approaching (23)

pod"ezžat'-pod"ehat' 'to drive up to'
podbežat-podbegat' 'to run up to'
podbrodit'-podbresti 'to roam up to'
podčalivat'-podčalit' 'to moor to'
podgrebat'-podgresti 'to row up to'
podgrebat'sja-podgrestis' 'to row (oneself) up to'
podkatyvat'-podkatit' 'to roll up to'
podkovyljat' 'to hobble up to'
podkradyvat'sja-podkrast'sja 'to sneak up to'
podletat'-podletet' 'to fly up to'
podnosit'-podnesti 'to carry to'
podnosit'sja-podnestis' 'to bring up to'
pododvigat'-pododvinut' 'to move up to'
pododvigat'sja-pododvinut'sja 'to move oneself up to'
podpolzat'-podpolzti 'to crawl up to'
podpuskat'-podpustit' 'to allow to approach'
podskakivat'-podskakat' 'to gallop up to'
podtaskivat'-podtaščit' 'to drag up to'
podtaskivat'sja-podtaščit'sja 'to drag oneself up to'

podvalivat'-podvalit' 'to steam in to'
podvalivat'sja-podvalit'sja 'to join (the others)'
podvolakivat'-podvoločit' 'to drag up to'
podzyvat'-podozvat' 'to call up to'

Abstract approaching (40)
poddabrivat'-poddobrit' 'to cajole'
poddabrivat'sja-poddobrit'sja 'to cajole'
poddakivat'-poddaknut' 'to say yes, to assent'
podgotovljat'-podgotovit' 'to prepare'
podgotovljat'sja-podgotovit'sja 'to get prepared'
podgovarivat'-podgovorit' 'to incite'
podgovarivat'sja-podgovorit'sja 'to obtain by dropping hints'
podhalimničat' 'to toady'
podhalimstvovat' 'to toady'
podkupat'-podkupit' 'to bribe; to win over'
podlaživat'-podladit' 'to adapt'
podlaživat'sja-podladit'sja 'to adapt oneself; to humour'
podleščat'sja-podlestit'sja 'to toady'
podležat' 'to be liable to'
podlipat' 'to toady'
podlizyvat'-podlizat' 'to toady'
podlizyvat'sja-podlizat'sja 'to get benefit by licking someone's boots'
podmahivat'-podmahnut' 'to sign (hastily), to scribble'
podmaslivat'-podmaslit' 'to toady'
podmazyvat'-podmazat' 'to toady'
podmazyvat'sja-podmazat'sja 'to get benefit by licking someone's boots'
podnačivat'-podnačit' 'to incite'
podobstrastničat' 'to toady'
podol'ščat'sja-podol'stit'sja 'to toady'
podospevat'-podospet' 'to arrive in time'
podoždat' 'to wait for'
podpevat'-podpet' 'to sing along'
podravnivat'-podravnjat' 'to align, to level'
podravnivat'sja-podravnjat'sja 'to align oneself'
podščëlkivat' 'to click in tune'
podsluživat'sja-podslužit'sja 'to fawn upon'
podstykovyvat'-podstykovat' 'to attach'
podstykovyvat'sja-podstykovat'sja 'to attach oneself'

podtveržda t’-podtverdit’ ‘to confirm’
podvarivat’-podvarit’ ‘to heat up again; to boil up more’
podvyvat’-podvyt’ ‘to howl (a little)’
podygryvat’-podygrat’ ‘to make an impression on’
podyskivat’-podyskat’ ‘to try to find (a suitable one)’
podzadorivat’-podzadorit’ ‘to egg on’
podzuživat’-podzudit’ ‘to incite’

Contact information:

Johanna Viimaranta
Department of Russian Studies
20014 University of Turku
Finland
e-mail: johanna(dot)viimaranta(at)utu(dot)fi

Inka Wissner

Les grands corpus du français moderne : des outils pour étudier le lexique diatopiquement marqué ?[1]

Abstract

In modern French dialectology, the critical usage of text data for the appropriate description of diatopic (regional) elements of national or regional varieties of French – though crucial – is hardly commented in scientific discourse. The paper offers an overview on the main corpora that are, or could be used for the lexical study of diatopicisms in contemporary French in the northern part of the francophone world, whether it be traditional, mostly literary corpora, journalistic texts, recent web corpora, or 'oral' corpora. The author examines their suitability for identifying diatopicisms with their spatial distribution in a francophone perspective, presents current methods of text analysis, and points out various problems related in particular to text location, regional annotation, discourse genres and treated subjects, but also to formal compatibility, text size and practical access. The article therewith underlines the need, in a close future, of comparable data for the different varieties in the francophone world in order to enhance the description of diatopic variation in modern French.

1. Introduction

La lexicologie française dispose depuis le tournant des XX^{e}/XXI^{e} siècles de grands dictionnaires différentiels de régionalismes, ou *diatopismes* (comme DSR, DHFQ, DRF et DictBelg) outre une base de données lexicographiques panfrancophone (BDLP) – conçus comme compléments des dictionnaires généraux. Leur préparation ayant impliqué la constitution et l'exploitation de bases de données textuelles, plusieurs questions

[1] Notre travail a bénéficié du regard critique de l'éditeur et de nos collègues France Lagueunière et Christophe Benzitoun, spécialistes de la variation diatopique pour l'une, et de corpus 'oraux' pour l'autre, ainsi que d'Alain Polguère pour le projet RLF, présenté ci-dessous.

s'imposent pour l'étude du lexique du français contemporain sous l'angle de sa variation dans l'espace, c'est-à-dire dans sa dimension diatopique :

> – Quels sont les grands corpus existants et exploités en ce domaine ?
> – Quels sont les problèmes liés à leur consultation, en termes pratiques d'accès aux bases, des données et métadonnées qu'ils fournissent, et des types de recherche qui sont possibles ?
> – Quels sont les besoins en lexicologie variationnelle pour une meilleure identification et description des diatopismes du français contemporain à travers la francophonie ?

Pour tenter de répondre à ces questions, on rappellera d'abord quels sont les corpus traditionnels de la lexicographie différentielle. On s'interrogera ensuite sur la possibilité et les besoins d'intégrer d'autres corpus, existants ou en préparation, qui pourraient permettre de mieux identifier les diatopismes dans une optique panfrancophone. On se concentrera surtout sur les plus grands corpus du français moderne de la francophonie du Nord, où l'on dispose déjà de corpus d'une certaine envergure.

La question du rôle des corpus est largement discutée dans de nombreux domaines de la linguistique contemporaine, et prend une ampleur particulière dans les travaux sur les réalisations médiatiques orales. Dans le champ disciplinaire de l'étude de la variation diatopique du français, bien établi en philologie romane depuis un peu plus de deux décennies (cf. Rézeau 2007), l'apport des corpus est toutefois peu thématisé. Les publications des spécialistes qui ont largement fait avancer les recherches en ce champ d'étude portent essentiellement sur l'exploitation de ressources lexicographiques. La problématique des corpus est notamment abordée du point de vue de leur constitution par des équipes en Belgique et au Québec et, du point de vue de leur exploitation – avec des optiques différentes – dans les articles de Queffélec (1997) et de Thibault (2007), surtout.

On a pu confirmer en lexicographie différentielle que l'exploitation de corpus est « de plus en plus sentie comme indispensable par la plupart des lexicographes et marque un renouveau de la méthodologie lexicographique » (Vézina 1998 : 228). Les corpus comme ensembles textuels complètent ainsi les matériaux traditionnels de la lexicologie variationnelle – qui sont surtout lexicographiques et très variables d'une aire de la francophonie à l'autre. L'accès croissant à des ensembles textuels permet en outre d'envisager à moyen terme d'identifier un diatopisme non pas (exclusivement) par rapport aux ressources métalinguistiques – surtout

des dictionnaires généraux, jouant le rôle de point de comparaison (cf. Poirier 2005 : 497) – mais aussi par rapport aux réalisations discursives effectives dans les diverses variétés diatopiques du français, par l'intermédiaire d'une analyse comparative des corpus.

Leur analyse présuppose bien entendu d'être familiarisé avec la situation sociolinguistique dans les aires de la francophonie dont on souhaite rendre compte. Nous mettons ici de côté la discussion de l'apport d'informateurs et de lexicographes locuteurs des différentes variétés diatopiques pour nous concentrer sur le problème de l'exploitation de corpus. Les réflexions proposées se situent dans le cadre de l'évaluation des possibilités d'une prise en compte de la variation diatopique au sein d'un projet dictionnairique qui porte sur le français contemporain, le Réseau Lexical du Français, préparé depuis l'été 2011 au CNRS (ATILF) sous la direction scientifique d'A. Polguère (cf. Lux-Pogodalla & Polguère 2011).

Si le terme de *corpus* désigne un ensemble d'éléments sur lequel se fonde l'étude du phénomène linguistique, il est utilisé en lexicographie différentielle pour renvoyer à un ensemble de ressources (méta-)linguistiques – d'où la notion de 'corpus d'exclusion' (cf. Francard 2001 : 228–230) – mais aussi à des ensembles de réalisations discursives. Ces derniers peuvent avoir été constitués à des fins d'analyse spécifiques, dans le respect de critères de sélection particuliers, comme la base FRANTEXT, mais aussi à d'autres objectifs que la recherche linguistique, comme EUROPRESSE[2]. Les corpus peuvent accueillir des textes dans leur intégralité (comme SCIENTEXT), ou rassembler des extraits d'un nombre limité de propositions phrastiques, pour ce qui est des corpus de citations (comme le FLI, au Québec).

La plupart des corpus existants du français n'ont pas été conçus pour l'analyse de diatopismes, et leur constitution ne s'appuie souvent pas sur une conception claire et actualisée de cette dimension de la langue – d'où résulte toute une série de difficultés d'exploitation pour les analyses à visée diatopique.

Rappelons ici que nous considérons que le français représente une langue historiquement standardisée (bien distincte de variétés historiquement liées comme des créoles à base lexicale française). Elle dispose d'un système à plusieurs modalités, dont la variabilité est

[2] Pour les sigles des corpus, banques de données, plateformes et portails, voir la bibliographie et l'annexe.

inhérente. En accord avec les avancées de la recherche en linguistique variationnelle des langues romanes, nous considérons aussi que la variation diatopique, dans l'espace – qualitative et quantitative – est enchevêtrée avec d'autres facteurs diasystémiques, qu'ils soient diastratiques (socioculturels), diaphasiques (situationnels), mais aussi diachroniques (temporels), diacodiques (dépendant du canal communicatif), ainsi que diamésiques, selon la conception des énoncés (cf. Gadet [2]2007 : 23, 47–49).

Si nous envisageons l'étude du français dans une optique francophone, la francophonie – au sens linguistique – est ici considérée comme l'ensemble de l'espace où le français est langue d'usage, et où il est souvent langue officielle – que ce soit en Europe, en Amérique du Nord et du Sud, en Afrique subsaharienne et au Maghreb, ou encore dans l'Océan Indien, l'Extrême Orient et le Pacifique. Nous ne nous limitons pas à l'étude du français comme langue 'maternelle', terme qui renvoie à un concept inadapté selon de nombreux chercheurs, notamment en dehors de la Métropole (Gadet et al. 2009 : 152*sq*.). Dans l'espace francophone, l'usage de la langue varie en effet fortement selon une diversité de facteurs, y compris la situation sociolinguistique de chaque communauté – et donc les réseaux sociaux et les modes d'acquisition du français – mais aussi sa fréquence d'emploi et son statut (dominant ou dominé) par rapport à des variétés en contact, ou les contraintes systémiques internes de chaque variété diatopique (cf. Gadet & Jones 2008 : 244*sq*.).

En se focalisant sur les variétés diatopiques les mieux dotées en matière de corpus, dans la francophonie du Nord (v. ci-dessous), la présente étude porte toutefois avant tout sur des aires où le français a un statut de langue dominante, représente une langue par tradition, et est un outil de communication quotidien.

2. Des sources traditionnelles

Pour étudier la dimension diatopique du lexique du français contemporain, on dispose de corpus traditionnels qui portent sur les variétés les mieux décrites, c'est-à-dire le français en Belgique (Wallonie, Bruxelles), en France hexagonale, au Québec, et en Suisse romande.

2.1 Des corpus constitués dans le cadre de projets dictionnairiques

Parmi les corpus traditionnels, on dispose avant tout de FRANTEXT – « incontestablement le plus grand corpus de français » (Gadet 2007 : 20), c'est-à-dire parmi les collections ordonnées de textes français, monolingues, publics, et accessibles à la communauté scientifique. Constitué pour l'étude du français général (ou de ce qui est considéré comme 'le français tout court'), il a donné lieu à la publication du dictionnaire de référence de français sur corpus, le TLF.

S'y joignent des corpus inspirés de FRANTEXT, élaborés pour l'étude de variétés diatopiques spécifiques dans le cadre de la préparation des dictionnaires différentiels de diatopismes en Europe et en Amérique du Nord, depuis le dernier tiers du XX^{e} siècle. Il s'agit tout d'abord de QUEBETEXT de Québec, qui contient une petite partie des textes qui ont été dépouillés pour la préparation du dictionnaire DHFQ, et en tant que source textuelle brute complète un très riche corpus de citations (FLI), outre la base métalinguistique ILQ. On dispose également de corpus pour chacune des trois grandes variétés diatopiques en Europe : BELTEXT (cf. Delcourt et al. 1993) élaboré pour la Belgique à Liège, Mons et Louvain-la-Neuve (ayant donné lieu à l'ouvrage de Delcourt 1998–1999), SUISTEXT, conçu à Neuchâtel pour le français en Suisse (DSR), puis REGION, élaboré pour l'étude de diatopismes en France à Nancy pour la publication du DRF. À ces bases traditionnelles s'ajoutent pour l'Amérique la base BDTS (cf. Cajolet-Laganière et al. 2008, Masson et al. 2007) et MCVF (ci-dessous), outre un petit corpus pour un espace entre l'Amérique du Nord et du Sud, les Antilles (ZOBEL).

Toutefois, l'envergure des bases varie énormément, allant de la totalité de l'œuvre de quatorze écrivains romands dans la base suisse, et de 7500 citations de 220 ouvrages littéraires dans la base REGION (ou des onze romans d'un écrivain dans ZOBEL), à 248 millions de mots dans FRANTEXT – en passant par plus de 52 millions de mots dans la récente base BDTS. Si dans d'autres domaines de la linguistique de corpus, il est habituel de faire des comparaisons d'usages à partir de tranches égales (p.ex. mille ou dix mille mots), on ne dispose pas de la possibilité technique de constituer des sous-corpus par nombre de mots pour les corpus traditionnels, ni du nombre de mots pour toutes ces bases (taille p.ex. non communiquée pour BELTEXT). Une telle comparaison serait aussi problématique vu l'hétérogénéité chronologique des corpus, qui portent sur le français d'époques différentes – du XIX^{e} au XXI^{e} siècles (REGION, SUISTEXT,

ZOBEL), du XVI^e^ au XXI^e^ siècles (FRANTEXT et QUEBETEXT) ou du XI^e^ au XIX^e^ siècles (MCVF) – ou exclusivement des années 1960 à nos jours (BDTS).

Pour ce qui est de l'aspect énonciativo-discursif, on sait que de tels corpus recueillent essentiellement du discours à dominante littéraire, élaboré, écrit. Les plus petites bases sont exclusivement littéraires (REGION, SUISTEXT, ZOBEL), d'autres plus diversifiés, tous exclusivement écrits, et relèvent par exemple des domaines administratifs, juridiques, ou scientifiques. La base québécoise BDTS a même une ambition de 'représentativité des usages', qui doit se comprendre au sens de l'intégration de plusieurs types de textes. Si la question de la représentativité se pose pour tout corpus (p.ex. Kleiber 1978 : 65), on sait qu'elle part d'une notion complexe et délicate (v. Cappeau & Gadet 2007). Qu'est-ce qui est représentatif par rapport à quoi – en fonction de quels critères ?

En ce qui concerne la dimension géolinguistique, la base FRANTEXT porte implicitement surtout sur la langue française dans la Métropole. Parmi les écrivains hexagonaux, certains utilisent inévitablement des diatopismes eux aussi, en particulier (quoique non exclusivement) les auteurs qui sont réputés régionalistes, comme H. Vincenot ou R. Bazin. En outre, ce corpus contient aussi – surtout depuis récemment – des textes d'ailleurs, comme ceux dus à des écrivains québécois (comme Guèvremont), belges, réunionnais, ou antillais, tels que Chamoiseau ou Zobel. Les autres corpus, puisque constitués pour l'analyse du français dans sa variation diatopique, fournissent a priori des textes ancrés dans des aires spécifiques. Ils ne sont pas pour autant diatopiquement homogènes, le problème de l'hétérogénéité se posant pour tout discours. À titre d'exemple, le corpus québécois BDTS contient aussi des textes métalinguistiques qui portent sur le traitement lexicographique de diatopismes d'ailleurs (Cajolet-Laganière et al. 2008 : 22*sq.*).

Ces divers types d'hétérogénéité ne sont pas gênants en tant que tels, si l'on procède à une analyse philologique de chaque attestation d'une unité sous étude, et de chaque corpus individuel, comme ceci est habituel pour l'exploitation des dictionnaires (cf. Wissner 2010 : 104–149). De telles divergences deviennent toutefois problématiques dans le cadre de requêtes automatiques et fréquentielles, pour l'analyse d'unités en grand nombre ou de diatopismes qui affichent un nombre très élevé d'attestations, ainsi que pour une exploitation comparative de corpus pour identifier des diatopismes. Celle-ci présuppose en effet une certaine comparabilité selon

plusieurs critères – y compris quantitatifs, chronologiques, discursifs, et géolinguistiques.

Il y a également toute une série de problèmes qui sont liés à l'annotation des textes, à l'accès aux métadonnées (lorsqu'il y en a), et aux types de requêtes par critères que permettent les logiciels d'exploitation des bases qui accueillent les corpus. Les possibilités d'exploitation des corpus de textes intégraux et de citations sont différentes pour des recherches de type énonciativo-pragmatique et syntaxique surtout, quoiqu'à une moindre mesure pour des analyses diatopiques qui portent sur le lexique. En effet, les logiciels d'exploitation qui donnent accès aux corpus contenant l'intégralité de textes sous la forme de bases de données permettent de visualiser non pas ces ensembles, mais seulement des extraits textuels. La taille de ces extraits varie en outre d'une base à l'autre ; à titre d'exemple, pour FRANTEXT, elle est de l'ordre de 700 signes depuis 2012 (300 auparavant). On n'oubliera pas non plus qu'on n'accède pas, ontologiquement parlant, au même discours si on lit un texte comme un roman, en tant qu'ouvrage imprimé destiné à être lu et à divertir et intriguer le public (plus qu'à informer), ou si on y accède par l'intermédiaire d'un corpus pour des fins de recherche.

2.2 L'appartenance géographique des locuteurs : un attribut textuel traditionnel

L'identification géographique de locuteurs est un des paramètres classiques de la sociologie, et l'annotation des énoncés rassemblés dans des corpus selon la région de rattachement des locuteurs est un attribut textuel traditionnel. Il est très peu explicité en lexicographie française, contrairement à la lexicographie anglophone (Atkins et al. 1992 ; Atkins & Rundell 2008 : 89) – avec des paramètres comme le type de source (p.ex. 'informatif'), la catégorie de domaine (du type 'sport et loisir') ou le sous-domaine (*ib*.). Outre l'origine géographique des énonciateurs des textes, les lexicographes anglophones jugent également utile de préciser leur statut de locuteur maternel (« natif ») (*ib*.). Toutefois, la variabilité diatopique d'une langue n'est pas tant liée à son usage en tant que langue 'maternelle' (ou 'officielle'), qu'aux types de fonctions que lui réserve la communauté sociolinguistique (v. 1).

D'une manière générale, l'identification géographique de locuteurs s'appuie traditionnellement sur leur lieu de naissance et/ou de résidence, mais aussi sur les régions où ils ont vécu et qui les ont marqués

linguistiquement. Au cas où l'utilisateur de corpus (intéressé par la diatopie ou non) souhaiterait connaître les rattachements multiples des locuteurs des énoncés qu'il analyse, les corpus ne fournissent le plus souvent aucun balisage ou retiennent seulement un rattachement principal – qui, à lui seul, n'est alors pas nécessairement pertinent. En même temps, l'annotation diatopique de corpus implique des choix qui influent nécessairement sur l'interprétation des données. Les catégories géolinguistiques doivent en outre s'appuyer sur une répartition à jour des aires linguistiques du français – et non pas sur des limites politiques ou géographiques, ni sur des aires d'autres variétés de langue en contact. Or, si l'on prend l'exemple de la France, les catégories géolinguistiques dans la base REGION sont de type politique (p.ex. 'Maine-et-Loire'). Pour une catégorisation à jour, on pourrait désormais s'appuyer sur les quatre aires lexicales principales qui ont été décrites à partir du DRF : une aire occidentale et une aire orientale (passant dans le Sud par l'embouchure du Rhône), puis une aire septentrionale et une aire méridionale, qui inclut le Bordelais à l'ouest et le Lyonnais à l'est (cf. Wissner 2010 : 29) – donc schématiquement 'Nord-ouest', 'Sud-ouest', 'Nord-est' et 'Sud-est'.

Le balisage géolinguistique du discours est bien fourni dans les corpus qui visent explicitement l'étude de diatopismes, comme ce petit corpus REGION, mais aussi dans de nombreux corpus oraux (v. 3.3), dans des manuscrits anciens ou le récent corpus MCVF, et dans les plus grands corpus de référence d'autres langues européennes (v. 4). Au contraire, dans les grands corpus de français comme FRANTEXT, ce type de balisage n'a pas été entrepris – d'où l'impossibilité, pour l'utilisateur, de constituer des sous-ensembles diatopiquement pertinents. Pourtant, le critère géographique est nécessaire pour une interprétation adéquate des données – et ce pour toute étude linguistique, mais aussi pour d'autres disciplines, comme l'histoire ou la sociologie. Dans le cadre d'analyses à visée diatopique, le recours à de tels corpus implique que chaque utilisateur décrypte le discours pour dépister les diatopismes. Ainsi, dans FRANTEXT – que le lexicologue diatopicien n'utilise pas seulement pour identifier ce qui relève de l'usage général (v. Thibault 2007 pour les possibilités de son exploitation) – il faut alors recourir aux procédés traditionnels que sont, schématiquement,

– l'identification des locuteurs à l'aide de biographies et d'autres travaux critiques, et
– l'analyse discursive du cotexte immédiat et du contexte plus large.

On trouve alors des indices d'une possible diatopicité si une forme recherchée figure surtout dans le discours de locuteurs qui sont identifiables comme étant marqués par une variété diatopique d'une région donnée, tout en évaluant les liens historiques de cette variété avec d'autres variétés de français. L'analyse du cotexte et de la situation d'énonciation plus large permet, quant à elle, de tirer profit des mises en relief métalinguistiques dans le discours, comme les commentaires métalinguistiques, et d'autres indices qui peuvent expliciter la valeur géolinguistique d'un énoncé étudié – que ce soit par son attribution au discours d'un locuteur, par la nature du référent, ou par la localisation d'un récit. Bien entendu, l'attribution d'un énoncé par un locuteur à une communauté linguistique est à évaluer avec prudence : il s'agit de simples indices.

Dans le cadre d'analyses à visée diatopique, l'annotation géolinguistique des grands corpus faciliterait considérablement les requêtes automatiques pour des analyses qualitatives et quantitatives. Elle ne permet toutefois pas de faire l'économie d'une analyse philologique sérieuse du discours, en présence de l'hétérogénéité et de la polyphonie du discours – où un diatopisme peut aussi être utilisé par un locuteur exogène, ou être attribué aux propos d'autrui. En effet, en présence de ces caractéristiques inhérentes à tout discours, l'annotation géolinguistique du discours est à doubler d'un balisage des passages de discours rapporté. Ceci concerne autant les unités du discours tels que les discours direct ou indirect, que les extraits qui sont présentés comme cités à l'aide de mises en relief métalinguistiques. De tels passages renseignent en outre indirectement sur la conceptualisation que se fait l'énonciateur de discours d'autrui, passés ou futurs, effectifs ou possibles, qu'ils soient associés à l'oral ou non.

Ce type de balisage, bien fourni pour certains corpus de textes d'états anciens du français, n'a toutefois pas été proposé pour les corpus traditionnels présentés ci-dessus sauf pour la plus petite de ces bases, ZOBEL, où le texte est annoté selon des critères à la fois géographiques, discursifs et métadiscursifs : selon les unités de discours, les notes et les indices typographiques (italique, gras), et la langue des séquences (passages en créole, en anglais, etc.).

L'annotation (géolinguistique) du discours et de ses unités discursives est en effet coûteuse en temps, même sur des corpus écrits (raison pour laquelle ladite base n'a pas été élargie pour devenir un véritable corpus antillais). En outre, le balisage du discours rapporté est problématique : son identification s'appuie en partie sur une interprétation, surtout dans le cas

de superpositions de voix – difficulté qui est particulièrement aigüe dans les textes d'enregistrements de réalisations orales. Si certaines séquences qui sont présentées comme rapportées sont aisées à identifier, et donc à annoter – essentiellement le *discours cité direct* – on rencontre des cas de polyphonie hautement complexe lorsque l'attribution de la responsabilité aux différents énonciateurs impliqués est difficile, voire impossible, comme dans le *discours indirect libre*. Ce sont les avancées de la linguistique à l'intersection de l'analyse du discours et du traitement informatique de la langue qui faciliteront à moyen terme l'annotation du discours cité.

2.3 L'accès aux données et aux métadonnées

Pour être exploitable, le critère géographique doit bien entendu être accessible lors de la consultation du corpus à l'aide du logiciel d'exploitation, et figurer parmi les options de constitution de sous-corpus. Or, le balisage diatopique de bases comme REGION, pour la France, est seulement visible lors d'une consultation manuelle, et certaines bases comme SUISTEXT ou ZOBEL sont actuellement consultables seulement en format texte dans des fichiers séparés (tout comme TCOF ; v. 3.3). Parmi les corpus traditionnels, seulement FRANTEXT, QUEBETEXT et MCVF sont consultables avec des logiciels d'exploitation performants, et seul le dernier permet un accès direct à l'ensemble des textes qu'il contient ; il est aussi le premier à permettre de créer des sous-corpus pour mener des recherches dans des textes relevant d'aires spécifiques dans la francophonie.

Si les corpus BELTEXT, QUEBETEXT, REGION et SUISTEXT relèvent bien, en théorie, des sources de référence traditionnelles pour les lexicologues diatopiciens, c'est surtout pour des problèmes de diffusion qu'ils ne sont pas exploités de façon systématique pour étudier des diatopismes. Traditionnellement, les corpus étaient en effet constitués par des équipes pour usage interne ; pour consulter des corpus ouverts à la communauté scientifique, les chercheurs devaient alors se déplacer. Ainsi, le corpus REGION est consultable sur place uniquement, à Nancy, tandis que les corpus BELTEXT et SUISTEXT ne sont toujours pas rendus accessibles : pour le second, la demande de droits d'auteurs pour une diffusion libre n'a pas abouti. Au contraire, d'autres bases sont consultables sur Internet – librement, pour QUEBETEXT, qui ne contient que des textes libres de droits (complété du Fichier lexical FLI) ; sous abonnement pour FRANTEXT ; et sur demande auprès des auteurs, pour la récente base MCVF.

La consultation de corpus en ligne est en effet devenue un standard en linguistique moderne même si les modalités d'accès varient, et certains corpus récents sont conçus comme des bases de données dynamiques sur Internet – qui permettent aussi la gestion, l'archivage et la consultation des métadonnées. Ainsi, la diffusion en ligne est annoncée pour la BDTS, mais est effective surtout pour des bases autres que les corpus traditionnels à dominante littéraire, comme des ensembles journalistiques tel qu'EUROPRESSE, des corpus tirés du web, ou le corpus oral belge VALIBEL (v. 3).

3. Quelles sources (nouvelles) pour l'étude lexicale de diatopismes dans une approche comparative, panfrancophone ?

À côté de ces corpus de référence traditionnels, constitués de discours essentiellement écrits, littéraires, et élaborés, le lexicologue diatopicien peut s'appuyer sur d'autres genres de discours s'il veut tirer des conclusions valables sur l'usage au sein d'une variété diatopique donnée (pour la notion de genre, encore largement débattue, v. p.ex. Adam & Heidmann 2006). En linguistique variationnelle, on souhaiterait en outre disposer d'un ensemble de données discursives qui soit à la fois équilibré et panfrancophone (Dister et al. 2008 : 296–298 pour l'oral). Pour l'étude du français dans sa dimension diatopique, c'est ce qui permettrait d'identifier et de décrire des diatopismes à partir de données textuelles (v. 1), dans une approche comparative, proprement francophone.

Les critères à retenir pour l'établissement de corpus sont largement discutés dans les travaux de spécialistes d'enregistrements de réalisations discursives orales (p.ex. Baude 2006 ; Benzitoun & Cappeau 2010). Pour l'étude du lexique diatopiquement marqué, les principaux critères à stabiliser dans des corpus qui portent sur différentes variétés diatopiques du français me semblent être de type géolinguistique et chronologique, mais aussi thématique et discursif – deux catégories qui semblent plus pertinentes que la distinction oral/écrit, selon des analyses lexicales ciblées de spécialistes de corpus 'oraux' (p.ex. Benzitoun & Cappeau 2010 : 1390 et 1396). La comparabilité des corpus présuppose aussi une certaine homogénéité des méthodes et des problématiques visées lors de leur constitution, ainsi que des choix d'annotation et de logiciels de consultation – ces facteurs influant directement sur les possibilités d'exploitation des corpus.

Si l'on retient l'aspect discursif, en visant un certain équilibre des genres, on pourrait compléter les corpus traditionnels de corpus récents – comme POLITEXT, pour le discours politique français du XX^{e} siècle (donnant accès à dix millions de mots), ou encore SCIENTEXT, pour le discours scientifique, avec plusieurs millions de mots de textes de huit disciplines (parmi lesquelles la médecine, la linguistique et l'électronique). Ces bases sont toutefois franco-centrées, et non conçues pour l'analyse de diatopismes, même si elles peuvent apporter des renseignements utiles. On se concentrera ici sur des corpus qui existent pour plusieurs variétés, et sur des genres discursifs qui se prêtent à l'établissement de corpus pour différentes variétés diatopiques de français dans la francophonie tout en permettant d'atténuer le penchant des ressources traditionnelles vers le discours littéraire.

3.1 Les corpus journalistiques

Ce sont tout d'abord les corpus journalistiques qui sont déjà exploités avec profit en lexicographie différentielle. Les journaux ne sont pas communément considérés comme des 'viviers' de diatopismes, leur emploi y étant en principe peu intense, par rapport aux romans régionalistes traditionnels. On y en trouve toutefois bel et bien, notamment dans les chroniques de gastronomie, de loisirs ou de tourisme (cf. Thibault 2000 : 554 au sujet du *Monde*).

Ces corpus, d'abord rendus disponibles sous la forme de cédéroms puis aussi sous forme de bases de données, ne sont toutefois pas conçus pour des analyses linguistiques poussées : les cédéroms *Le Monde* et *Le Monde diplomatique*, comme les sites commerciaux en ligne EUROPRESSE ou EUREKA, ne permettent pas de faire des requêtes de lemmes ou de locutions, ni d'établir des sous-corpus. À l'exception d'EUREKA, ils ne fournissent pas non plus de métadonnées géolinguistiques, imposant au diatopicien de recourir aux procédés d'identification géolinguistique traditionnels (v. 2.2) – souvent plus difficile lorsqu'il s'agit d'énoncés de journalistes, puisqu'on ne dispose pas souvent de renseignements suffisants à leur sujet. Si le corpus EUREKA est consultable par langue, région, date de publication et domaine thématique, la requête par suites de mots est en partie inopérante – semblant fournir les attestations de la première entité graphique saisie – et les critères d'établissement des catégories géographiques sont de nature politique, et non pas linguistique (v. 2.2).

De façon générale, on observe toujours, pour l'utilisation de tels types d'outils, l'absence de concordances et de fonctions de recherche statistique, l'aspect « trop limité » des recherches de co-occurrences, et la « difficulté de désambiguïser rapidement et efficacement les résultats obtenus », mais aussi dans certains (comme EUROPRESSE) la redondance des résultats, en raison de la reprise de dépêches, qui posent problème en particulier pour des études fréquentielles (Thibault 2007 : 479). Pour ce qui est des archives de journaux – lorsqu'elles sont accessibles en ligne – elles sont dotées de logiciels d'exploitation encore moins adaptés aux besoins du linguiste.

Le discours journalistique est toutefois aussi désormais mis à profit dans plusieurs corpus qui sont conçus et annotés pour des recherches linguistiques, tels que le corpus hexagonal CERF, qui comporte dix tranches d'un million de mots de la presse nationale et autant de la presse régionale. Toutefois, le corpus de presse le plus grand de la langue française (complété d'un corpus littéraire d'Afrique noire) est actuellement la *Kölner romanistische Korpusdatenbank*, en finalisation à Cologne, dont la diffusion libre est prévue. Ce méga-corpus (80 millions de mots environ) conçu pour des recherches linguistiques – qualitatives et quantitatives, y compris de collocations – accueille des extraits de journaux publiés dans les années 2000, répartis en parts égales entre la presse francophone européenne, surtout de France (*Le Figaro*, *L'Est Républicain* et *Sud-Ouest*), et la presse nationale des pays francophones d'Afrique, surtout d'Afrique noire (v. Annexe).

Le discours journalistique étant en effet relativement facile d'accès grâce à l'informatisation, il est en principe possible, et souhaitable, de constituer un corpus journalistique avec des données comparables pour les différentes zones de l'espace francophone. En France, la plateforme CNRTL met actuellement à disposition un corpus journalistique régional de France – des extraits de trois années des éditions régionales et locales de *L'Est Républicain* (1999, 2002 et 2003) – conçu lui aussi pour des analyses linguistiques. Pour établir un corpus journalistique équilibré du français en France se pose alors le problème de l'équilibre géolinguistique, chronologique (et donc partiellement thématique), et quantitatif. En s'appuyant sur une répartition à jour des aires linguistiques, il faudrait disposer de corpus pour les quatre grandes aires en France, grosso modo le Nord-ouest, Nord-est, Sud-ouest, et le Sud-est (v. 2.2). En se focalisant sur les quotidiens, on pourrait alors retenir (comme pendant de l'*Est Républicain*) la presse régionale *Ouest-France* pour l'Ouest septentrional, qui a une très large couverture (Bretagne, Normandie, Pays de la Loire, qui

comprend la Vendée mais non les Charentes), pour la France méridionale surtout occidentale *La Dépêche du Midi*, et le journal du Sud-est : *Le Dauphiné Libéré*, couvrant les régions Rhône-Alpes et Provence-Alpes-Côte d'Azur (incluant donc le Lyonnais). Bien entendu, ces quatre quotidiens ne couvrent pas toutes les régions de France, et on garderait une certaine non-adéquation dans les résultats puisque la couverture de chacun d'entre elles ne se recoupe pas exactement avec la répartition des aires lexicales.

3.2 Les corpus construits à partir du web

Vu les difficultés générales d'accès aux corpus de référence traditionnels de la lexicographie différentielle, et vu les problèmes d'exploitation efficace des corpus journalistiques actuellement disponibles pour l'analyse de diatopismes, de nombreux lexicologues diatopiciens recourent de plus en plus à GALLICA et Google Recherche de Livres (GRL) – qui donnent accès à des textes numérisés – mais aussi au web lui-même, le plus souvent par l'intermédiaire du moteur de recherche Google.

On sait toutefois que « récupérer du corpus sur le net soulève des écueils » (Cappeau & Gadet 2007 : 108), que la taille des corpus ne résout pas tous les problèmes, et que l'utilisation de tels corpus est discutable. Le statut du web comme corpus est ainsi largement débattu surtout en linguistique computationnelle, et du point de vue lexicographique gagne peut-être plutôt à être conçu comme une source de textes, à partir de laquelle peuvent être établis des corpus (Atkins & Rundell 2008 : 78). Par ailleurs, si le web a pour avantage d'être a priori exploitable pour tout chercheur ayant accès à Internet, le corpus évolue très vite, et les données fournies par le moteur de recherche varient d'un ordinateur à l'autre – en fonction des habitudes de requêtes de ses utilisateurs.

L'exploitation critique du web ou des bases GALLICA et GRL présuppose bien entendu de prendre les mesures de prudence nécessaires dans l'analyse philologique des textes, en analysant exclusivement les énoncés dont la localisation – à l'aide des procédés classiques (v. 2.2) – est suffisamment probable, et en tenant notamment compte du contexte de production des textes que l'on exploite, mais aussi des implications ontologiques et techniques de leur consultation via le Net. Les difficultés d'exploitation des bases évolutives GALLICA et GRL, largement discutées ailleurs, relèvent entre autres de problèmes de traçabilité des formes dans les textes (notamment anciens), dont certaines ne sont pas reconnues par les

logiciels de traitement utilisés, et donc inaccessibles pour l'utilisateur ; d'autres problèmes sont liés à la fiabilité des métadonnées mises à disposition, comme la datation des textes (GRL retenant p.ex. souvent, pour les journaux, l'année de publication du premier fascicule). Il est alors de mise de critiquer, croiser et contrôler les données recueillies (Brochard 2012 pour différentes techniques).

GRL et GALLICA permettent de créer des sous-corpus selon des critères comme l'époque de publication, mais ne tiennent pas compte du paramètre diatopique – tout comme le moteur de recherche Google, pour le web. Si le groupe de Google vise à enregistrer un maximum de renseignements (y compris géographiques) sur les utilisateurs du web à partir de données fournies par les usagers eux-mêmes (comm. pers. de Mairi Mc Laughlin, Berkeley University, le 07/09/2011), se pose non seulement le problème de la fiabilité de ces métadonnées, mais aussi celui des catégorisations géolinguistiques, et des appartenances multiples des locuteurs (v. 2.2).

Cependant, aussi bien GALLICA, GRL que le web en général ont pour grand avantage de donner accès à des textes de diverses régions de la francophonie et d'époques différentes. Leur exploitation semble alors justifiée si l'on y cherche justement ce qui est introuvable ailleurs : notamment un ensemble textuel de très grande taille, mais aussi et, avant tout, des usages mal attestés dans les sources traditionnelles. Le web contient en outre des types de discours propres (comme les blogs), ainsi que des textes variés en termes de sujets abordés et de situations d'énonciation qui y sont représentées – formelles et informelles, privées, publiques autant que professionnelles (Atkins & Rundell 2008 : 80 pour le web en anglais).

Il existe en outre depuis peu des méga-corpus construits à partir du web pour plusieurs langues européennes, dans le cadre du projet italien *WaCky wide web*, y compris pour le français : le FRWAC, basé sur les URL en « .fr », d'une taille d'environ 1,6 milliard de mots-occurrences. Les extraits de cette base – constituée de l'image du web à un moment donné (en 2009 ? ; date non précisée) – ne sont pas nécessairement accessibles ultérieurement par l'intermédiaire d'un moteur de recherche comme Google. Par rapport au web, ce corpus est plus restreint, ce qui évite de noyer l'utilisateur sous des quantités de résultats souvent faramineuses, et est non-évolutif – et donc consultable par d'autres chercheurs pour des fins de vérification, ou des études comparatives.

Toutefois, comme le web lui-même, le FRWAC inclut des redondances (aussi 3.1), et est hétérogène dans le sens où les domaines (URL)

accueillent des sites de tout ordre, y compris en d'autres langues – outre les passages en discours cité (v. 2.2). En outre, s'agissant d'un corpus constitué par interrogation du web avec des paires de mots à l'aide d'un logiciel de traitement informatique (cf. Sharoff 2006), la recherche de lemmes avec le moteur de recherche permet d'accéder seulement aux formes dans le corpus qui sont reconnues par les dictionnaires sur lesquels se sont appuyés les auteurs du corpus. Le FRWAC est aussi désormais disponible sous une version conviviale catégorisée par degré de normativité orthographique et grammaticale, conçue et exploitée en lexicographie française générale, pour le projet du Réseau Lexical du Français en cours à l'ATILF (v. 1). Reste à voir si ce corpus sera exploité de façon plus générale, comme le web, y compris en lexicologie variationnelle.

Vu la taille de FRWAC et sa nature hétérogène, il est quasiment impossible de l'annoter (et indexer) selon les paramètres traditionnels qui sont pertinents pour une exploitation critique, en particulier les critères énonciativo-discursifs (genres, thèmes) et le facteur diatopique (rattachement géographique des locuteurs). Les auteurs de FRWAC ont certes essayé de construire des corpus « relativement homogènes » (Baroni et al. 2009 : 15), en se limitant par exemple au domaine « .uk » pour l'anglais britannique, afin d'exclure

> « les problèmes théoriques et méthodologiques concernant l'inclusion ou l'exclusion de variétés où une langue a un statut officiel mais non maternel » (*ib.*).

Toutefois, la distinction entre des variétés à statut officiel *vs* 'maternel' n'est pas nécessairement pertinente (v. 1 et 2.2), et le choix exclusif d'un seul domaine ne permet pas d'assurer l'homogénéité géolinguistique du corpus, puisque les données peuvent ne pas relever de la variété de rattachement officielle du site où elles figurent.

Pour des analyses lexicales à visée diatopique, il serait possible de recourir au corpus francophone I-FR de Leeds, constitué par la même équipe que FRWAC et selon les mêmes procédures, mais contenant environ 200 millions de mots, de tous les domaines francophones. Il aurait pour avantage d'avoir été annoté thématiquement, selon des critères qui comprennent des renseignements géographiques – même s'il ne s'agit pas de catégories proprement géolinguistiques. L'un des objectifs pour l'analyse du lexique diatopiquement marqué étant de disposer de corpus comparables pour différentes zones de la francophonie, il serait aussi possible d'établir assez rapidement des corpus à côté de FRWAC (pour la France) – par exemple pour la Belgique (domaine .be), la Suisse (.ch), le

Québec (.qc.ca), voire le Canada (.ca), ou pour d'autres provinces (comme .on.ca pour l'Ontario). Ceci permettrait de disposer d'ensembles de données plus importants pour une exploitation à visée comparative. Se pose toutefois la question de savoir si les tendances d'usages quantitatives et qualitatives dans les corpus respectifs qui seraient observables peuvent fournir des indices valables, vu l'hétérogénéité géographique des données textuelles qu'ils contiennent.

En outre, comme dans le web, l'identification géolinguistique des énoncés auxquels donnent accès autant l'I-FR que le FRWAC est en général plus délicate que dans le cadre du discours romanesque et même journalistique – et souvent impossible. Le diatopicien peut certes glaner des indices au sein de l'adresse URL ou dans les textes en présence de commentaires métalinguistiques, de marqueurs typographiques et d'autres dispositifs de citations ; toutefois, une analyse philologique poussée y est impossible.

Pour l'étude sémantique de la phraséologie, même des corpus soigneusement annotés comme le BNC pour l'anglais britannique (rassemblant 100 millions de mots) seraient « too small to reveal the meaning-based patterning of collocations » (comme l'angl. *deep sense* ou *real sense*) même dans une perspective sans visée diatopique, et sont à utiliser en même temps que le Web (Wierzbicka 2009 : 125). En effet, même le méga-corpus qu'est le FRWAC ne fournit en fin de compte que peu d'attestations de diatopismes du français, comme de l'expression verbale *tomber en amour* (français de référence *tomber amoureux*) (dix attestations), tout en fournissant cependant certains indices pour la localisation des énoncés (Wissner à paraître).

3.3 Les corpus 'oraux'

En présence du changement de paradigme en linguistique contemporaine vers une conceptualisation de la langue avec ses réalisations non seulement à l'écrit, mais aussi à l'oral, nombreux sont les linguistes qui considèrent qu'il est nécessaire d'exploiter des « collections ordonnées d'enregistrements de productions linguistiques orales et multimodales », communément appelées *corpus oraux* (déf. d'apr. Baude 2006). Les questions portant sur l'exploitation et la constitution de corpus oraux sont

ainsi largement discutées par les spécialistes du domaine (p.ex. *ib.* ; Cappeau & Gadet 2007 ; Gadet et al. 2009)[3].

Il faut ajouter que l'oralité des productions langagières enregistrées et transcrites dans de tels corpus est de nature diacodique, dépendant du médium de communication, et non pas diamésique (v. 1). Les corpus oraux donnent accès non pas à une 'langue' orale spontanée et authentique, mais à des enregistrements et transcriptions d'énoncés produits oralement, qui ont des modalités fonctionnelles et communicatives diverses.

Malgré « la grande vague de l'oral » (Blanche-Benveniste & Jeanjean 1986 : 43–46), et malgré l'intérêt de phonéticiens, pragmaticiens et morphosyntacticiens pour les corpus oraux, ces derniers sont très insuffisamment exploités en lexicographie française – autant dans les années 1990 (Queffélec 1997 : 353), que de nos jours. Cet état de choses s'explique surtout par la primauté traditionnelle de l'écrit (autant en termes diacodique que diamésique), et par le faible rendement de corpus oraux : le coût élevé pour des enregistrements, et le besoin pour les lexicographes de corpus quantitativement importants, sachant que l'exploitation de corpus oraux (de petite taille) pour l'étude de variétés de français comme en Belgique et en Afrique s'est avérée très coûteuse et peu rentable (*ib.* : 353–356). De fait, « la quasi-totalité des corpus oraux de français » ont été « constitués pour des buts autres que lexicologiques » (Queffélec 1997 : 356). Pourtant, le recours à des corpus oraux, y compris de conversations spontanées, est usuel en lexicographie anglophone (Atkins & Rundell 2008 : 77). Pour la description de diatopismes du français, il est fondamental non seulement lorsque les pratiques langagières se manifestent très majoritairement à l'oral, comme dans la plupart des régions francophones en Afrique (Queffélec 1997 : 365*sq*.), mais aussi pour tenir compte de l'usage sous ses diverses facettes en général.

Il est vrai que l'équipe québécoise préparant leur dictionnaire différentiel DHFQ avait déjà utilisé les corpus oraux de Sherbrooke et de Montréal avec des données recueillies dans les années 1960/70 : le *Corpus de l'Estrie*, *Centre-Sud* et *Sankoff-Cedegren* (DFQPrés 1985 : XVII [Poirier])[4]. En effet, c'est surtout depuis les années 1960/70 au Québec, et

[3] On exclut de la discussion des corpus 'oraux' le méga-corpus bilingue Hansard, disponible en ligne, qui fournit des données pour le français au Canada des débats parlementaires canadiens, mais est constitué des comptes-rendus de ces débats, dont une partie est traduite de l'anglais.

[4] Les deux premiers rassemblent chacun autour de 100 000 mots, le dernier plus d'un million (v. le site du TLFQ pour un descriptif de ces corpus : http://www.tlfq.ulaval.ca).

depuis les années 1980 en Afrique Noire puis au Levant et dans l'Océan Indien qu'ont été réalisés des enregistrements, tous concentrés sur une variété ou une comparaison de deux variétés (Dister et al. 2008 : 296) – avant l'arrivée du PFC (v. ci-dessous). Toutefois, les premiers corpus oraux à objectifs général et sociolinguistique comme le *Corpus Sankoff-Cedegren* ou le corpus hexagonal *Français Fondamental* de 1953, désormais effacé (cf. Baude 2006 : 26), sont difficilement comparables aux corpus de la nouvelle génération – les possibilités d'enregistrement, de transcription, d'annotation et d'exploitation à l'aide de logiciels étant passées par une importante (r)évolution informatique.

Parmi la multitude de petits corpus oraux en Europe relevés par Cappeau & Seijido (2005), seize corpus explicitent le rattachement géographique des textes et/ou des locuteurs enregistrés (pour ce qui est des corpus dépassant une taille correspondant à approximativement 200 000 mots) : trois pour la Suisse, autant pour la Belgique, et huit pour la France continentale (Orléans, Tours, Auvergne ; Alsace ; Perpignan – conçu pour l'étude de régionalismes ; Aquitaine, Toulouse et Grenoble). En France – où J.-M. Debaisieux évaluait la taille de l'ensemble des données orales disponibles en 2005 à environ quatre ou cinq millions de mots (d'apr. André & Canut 2010) – aucun des corpus oraux rendus disponibles à la communauté scientifique ne vise toutefois spécifiquement à permettre des recherches de diatopismes. En outre, toutes ces données sont éparpillées (difficultés d'accès), et pour une exploitation comparative à visée diatopique, ils sont trop hétérogènes. En effet, à l'heure actuelle,

> « de nombreuses contraintes pèsent sur la linguistique française sur corpus, notamment au niveau de l'existence et de la disponibilité des données et des outils » (Benzitoun & Cappeau 2010 : 1384)

En outre, le français ne dispose toujours pas d'un « corpus de référence » (*ib.* : 1385) – même si le CERF, le CRFP (cf. DELIC 2004) et le PFC (cf. Durand & Lyche 2003, Durand et al. 2009) s'y portent candidats. Contrairement aux corpus écrits, les corpus oraux posent en outre tous des problèmes de transcription, et donc d'accessibilité des données enregistrées. L'urgence est ainsi actuellement de faire converger les pratiques, comme au niveau des annotations et des transcriptions, et de systématiser les métadonnées (v. Bruxelles et al. 2009 : 13 et Baude 2006), puis de rendre les corpus disponibles conformes aux standards

internationaux, et accessibles[5]. C'est aussi pour améliorer la description du lexique diatopiquement marqué qu'on plaide depuis longtemps pour une homogénéisation des méthodes d'enregistrement et des normes de transcription, pour une mise en commun des corpus oraux, et pour leur diffusion auprès de la communauté scientifique (Queffélec 1997 : 364–366). Toutefois, une codification minimale qui serait commune pour toutes les visées de recherche (même au sein de la linguistique) n'existe pas, « même si le mirage en est régulièrement soulevé » (Cappeau & Gadet 2007 : 107). L'exploitation efficace de corpus oraux pour l'analyse du lexique diatopiquement marqué implique ainsi de retenir des corpus en transcription orthographique standard.

Même si *big is not necessarily beautiful* (v. 3.2), dans le domaine des corpus oraux, la taille d'un corpus est un critère de valeur pour permettre des analyses qualitativement fructueuses – à côté de la 'rareté' des données qu'ils accueillent, et des types de recherche qu'ils permettent grâce aux possibilités de sélection de sous-corpus (p.ex. Dister et al. 2009 : 121 et 123). Un corpus oral « de taille significative » comporterait deux millions de mots, un « vaste corpus » plusieurs millions de mots (Benzitoun & Cappeau 2010 : 1383 et 1384).

Dans les conditions actuelles où les corpus oraux de français sont de la grandeur d'un million de mots, ceux-ci fournissent déjà une base suffisante pour mener des études lexicales qualitatives et fréquentielles valables (Benzitoun & Cappeau 2010 : 1396). D'autres spécialistes de corpus au contraire estiment que vu leur taille, « il n'est guère possible de faire des recherches lexicales, ni d'établir des statistiques fiables sur les usages » (Baude 2006 : 29) – même du français général[6]. L'apport des

[5] Ainsi pour le français hexagonal, un regroupement de corpus oraux est en train de se faire à l'ATILF par Ch. Benzitoun pour usage au sein du laboratoire. Il accueille des transcriptions d'enregistrements de plusieurs villes en France, totalisant plus de 2,3 millions de mots (dont un million environ sont actuellement exploitables). L'archive intègre les données de l'équipe nancéenne (TCOF) et des autres projets d'envergure qui sont compatibles avec ce dernier, donnant accès à des enregistrements de dialogues pour la plupart : le Corpus de Français Parlé Parisien (CFPP2000) (v. 3.3), CORPAIX (version 2000), le Corpus de Référence du Français Parlé (CRFP, cf. DELIC 2004) – corpus de plusieurs villes en France rassemblant plus de 400 000 mots – et enfin la partie des discussions libres de PFC-France (Benzitoun/Cappeau 2010 : 1385sq.).

[6] En effet, même des verbes courants (du français général) comme *causer* n'affichent qu'une fréquence très limitée dans des corpus d'un million de mots, comme le corpus oral CORPAIX (cinq attestations) ou la tranche journalistique du corpus de textes écrits

corpus oraux actuel est d'autant plus limité pour l'étude de diatopismes (par rapport aux larges corpus écrits traditionnels).

Actuellement, les banques de données les plus consistantes en Europe sont PFC, VALIBEL (ci-dessous) et CLAPI (Dister et al. 2008 : 296), ainsi que désormais TCOF (ci-dessous) et ESLO – le plus grand des corpus oraux hexagonaux, avec une taille approchant les sept millions de mots selon des approximations, mais dont la diffusion annoncée n'est pas effective à ce jour. Si CLAPI forme un ensemble textuel de taille significative – la quarantaine de corpus qu'il contient ont chacun une taille allant jusqu'à 200 000 mots (d'apr. Benzitoun & Cappeau 2010 : 1398 note 2) – cet ensemble n'est pas librement accessible et, comme il est constitué pour l'analyse de l'interaction (cf. Bert et al. 2010), il n'est pas adapté pour des recherches lexicales dans l'optique diatopique. À l'heure actuelle, on compte seulement trois corpus oraux de français contemporain qui pourraient selon nous être exploités avec profit pour l'analyse du lexique diatopiquement marqué dans une optique panfrancophone, en répondant aux conditions fondamentales suivantes :

> – être disponibles en transcription orthographique standard pour permettre des recherches lexicales automatisées, avec un moteur de recherche permettant la requête de lemmes, voire de co-occurrences, dans un seul fichier, avec un concordancier permettant de trier les résultats de requête par contexte gauche/droite pour isoler les sens,
> – fournir l'annotation du discours selon le critère géolinguistique en s'appuyant sur des catégories linguistiquement pertinentes, et rendre ce critère disponible lors de recherches au sein du corpus, et enfin
> – être rendus accessibles à la communauté scientifique, en ligne.

On ne dispose pas actuellement de plusieurs corpus qui satisfassent à ces divers critères, mais plusieurs corpus se portent candidats. Ainsi, pour le français continental, on trouve le corpus TCOF, complété du Corpus de Français Parlé Parisien (CFPP2000, cf. Branca-Rosoff et al. 2012), qui sont accessibles à la communauté scientifique et accueillent des transcriptions d'enregistrements avec des locuteurs hexagonaux. Ayant distingué tous les locuteurs enregistrés dans les métadonnées d'après leur 'appartenance régionale dominante' (André & Canut 2010), ils permettent à l'utilisateur d'identifier les énoncés selon ce critère géolinguistique et de créer des sous-corpus selon ce critère pour faire des requêtes dans tous les énoncés

---

CERF (38 attestations) – sachant que ce décalage est aussi lié aux genres et thèmes des deux corpus exploités (Benzitoun/Cappeau 2010 : 1384).

de locuteurs d'une région donnée. Le premier, TCOF, qui contient des parties d'échanges adulte/enfant et entre adultes (surtout des entretiens, des conversations et des réunions de travail), est un corpus essentiellement lorrain dans le sens où les locuteurs enregistrés sont pour la plupart originaires du Nord-est (Lorraine). Cet ensemble totalise approximativement quatre millions de mots (dont un demi-million sont exploitables en ligne, sans restriction, actuellement sous la forme de fichiers séparés). Le CFPP2000, pour sa part, recueille des énoncés de locuteurs originaires de la région parisienne, et rassemble un ensemble de 400 000 mots environ.

Contrairement aux plus grands corpus franco-français, le corpus de français en Belgique, VALIBEL, constitué à partir d'enquêtes sociolinguistiques et d'entretiens (cf. Dister et al. 2009 : 117), a été explicitement conçu pour l'analyse d'une variété diatopique spécifique, et a permis la préparation du DictBelg (2010). Parmi les corpus portant sur une variété diatopique spécifique du français, il s'agit du corpus oral moderne le plus vaste, avec environ 4 millions de mots. Il est accessible en ligne sur demande, sauf pour les tous derniers enregistrements qui n'ont pas encore donné lieu à des publications pour un objet de recherche précis (cf. Dister et al. 2009 : 126*sq.*). S'il s'agit d'un corpus unique, il n'est pas pour autant systématiquement exploité en lexicographie différentielle pour l'étude de diatopismes en dehors de la Belgique. Inspiré des corpus du Groupe Aixois de Recherche en Syntaxe GARS en France et de la sociolinguistique québécoise, VALIBEL inspire à son tour d'autres corpus dans la francophonie, comme VALIRUN, pour le français et le créole à l'île de la Réunion[7].

Les autres variétés diatopiques du français dans la francophonie du Nord ou du Sud ne disposent à l'heure actuelle pas, à ma connaissance, de corpus oraux d'une taille équivalente qui soient rendus accessibles à la communauté scientifique. Pour l'Amérique du Nord, on dispose tout de

[7] D'autres corpus portent exclusivement sur des créoles à base lexicale française, mais affichent des alternances codiques entre créole et français, comme le corpus de Ludwig et al. (2001). En Europe, un projet parallèle à VALIBEL, VALISUISSE pour le français en Suisse (annoncé sur le site de VALIBEL), a été initié en 2004 par Anne Grobet (Genève) mais n'a pas pu aboutir, faute de moyens (comm. pers. de l'auteure du 24/10/11). On attend donc avec impatience le très récent projet OFROM, en préparation à Neuchâtel.

même déjà du *Corpus de français parlé au Québec* (CFPQ) de Sherbrooke, constitué d'un demi-million de mots environ[8].

Le troisième corpus qui est actuellement exploitable pour une analyse du lexique dans une optique diatopique est le premier des projets panfrancophones qui visent explicitement à rassembler des données comparables pour plusieurs variétés diatopiques de la francophonie, recueillies selon une même méthode : le corpus du projet Phonologie du Français Contemporain (PFC). Celui-ci visant d'abord une exploitation phonologique, puis tardivement des analyses à visée syntaxique et sociolinguistique, ce sont ses entretiens qui sont exploitables avec profit pour une analyse lexicale. L'ensemble textuel actuellement disponible correspond à un million de mots transcrits (comm. pers. de B. Laks du 25/09/11), accessibles en ligne sur demande. Comme pour VALIBEL et TCOF, le discours a été entièrement annoté selon le rattachement géolinguistique des locuteurs dont il recueille les énoncés transcrits. L'interface d'exploitation n'est pas conçue pour des recherches de co-occurrences, mais permet la création de sous-corpus, y compris selon le critère diatopique. Même si le PFC reste un corpus ouvert, ses transcriptions seront à moyen terme consultables pour tous 'les français' qu'il couvre à l'échelle francophone, dans trente-trois zones géographiques.

À plus long terme, les données orales du projet PFC pourront en outre être complétées par des données comparables du projet CFA, pour l'Afrique et l'Océan Indien, dont la méthodologie d'enquête s'appuie sur celle de PFC (cf. Dister et al. 2008). Deux autres grands projets internationaux en cours visent également à permettre des comparaisons d'usage dans la francophonie : le Corpus International et Écologique de la Langue Française (CIEL-F) – initiative franco-belgo-allemande qui rassemble des données pour le français de vastes aires, comme en Suisse, en Acadie et en Égypte (cf. *ib* ; Gadet et al. 2009, 2012) – et le projet *Dynamiques des français périphériques*, lancé en 2008 au laboratoire MoDyCo, qui porte en partie sur les mêmes aires que CIEL-F, mais aussi sur des zones comme la Tunisie et les Îles anglo-normandes.

[8] Le projet international d'Ottawa intitulé *Le français à la mesure d'un continent* (2011–2018), mené sous la direction de F. Martineau, visant à établir un corpus de discours non-conventionnel, écrit et oral (en particulier pour mener des analyses en syntaxe), fournira pour sa part des données pour le français dans l'Ontario, ainsi qu'a priori des données comparables pour des variétés diatopiques ailleurs dans la francophonie.

Si l'usage de corpus oraux tend à être moins 'rentable' que celui de corpus écrits, les corpus oraux sont importants pour une analyse plus juste des diatopismes du français, y compris de son lexique – vu les caractéristiques énonciativo-discursives des énoncés transcrits, et vu que le travail d'identification diatopique des locuteurs y est déjà souvent fourni.

4. Perspectives

À l'heure actuelle, c'est surtout par manque de moyens que l'analyse lexicale des diatopismes s'appuie sur des données hétérogènes, qui sont encore en grande partie lexicographiques. En complément des ressources traditionnelles, l'utilisation de grands corpus 'écrits' et 'oraux' est pourtant nécessaire pour s'appuyer sur des données discursives des diverses variétés diatopiques du français dans la francophonie – sous condition de s'astreindre à l'analyse critique de chaque corpus.

Peut-on raisonnablement imaginer la mise en place d'une interface commune qui soit incontournable pour toute étude lexicologique à visée diatopique ? Un outil partagé pourrait se présenter sous la forme d'un portail commun qui centraliserait la description des corpus, et fournirait des hyperliens vers les sites où leur exploitation serait possible. Il est vrai que le souhait de la constitution d'une « bibliothèque de données de corpus » multifonctionnelle, ouverte et partageable, quoique « dans l'air du temps », est irréaliste (Cappeau & Gadet 2007 : 108 et 107–109), mais un *Inventaire des corpus de français hors de France* est bel et bien en préparation (Gadet à paraître). On pourra regrouper les corpus qui sont exploitables selon des visées de recherche comparables, comme ceux qui visent l'analyse de la phonologie ; de l'interaction ; ou de la morphosyntaxe et du lexique. Il existe déjà pour d'autres langues de grands corpus de référence qui sont exploitables pour des recherches lexicales et morphosyntaxiques, annotés diatopiquement aussi bien que morphosyntaxiquement, et équilibrés selon plusieurs critères (y compris discursif) : notamment pour l'anglais britannique, avec le BNC, et pour l'espagnol de l'Hispania, avec le CREA (constitués de 100 million de mots pour l'un, 160 pour l'autre) – outre des corpus de plusieurs centaines de millions de mots pour l'anglais contemporain (tels OEC ou COCA).

Vu l'organisation essentiellement étatique des plateformes existantes qui hébergent des corpus multifonctionnels du français, exploitables à l'aide de logiciels compatibles – le *Réseau des corpus lexicaux québécois*, au Canada, ou en France le SLDR (anciennement CRDO-Aix, pour

'l'oral') et le CNRTL (pour l'écrit, où se trouve toutefois aussi le TCOF) – c'est dans le cadre du projet Open Resources and TOols for LANGuage (ORTOLANG) que l'on peut instaurer à moyen terme une infrastructure commune, tout au moins pour la France. Le projet (Aix, Nancy, Orléans/Tours, Paris 2012–2019) vise en effet à l'élaboration d'une infrastructure en réseau et de ressources et d'outils, pour rassembler, diffuser et archiver les corpus écrits, oraux et patrimoniaux du français et des autres langues de France (cf. http://www.cnrtl.fr/ortolang/). Il se réalisera dans le cadre de deux consortiums linguistiques (IRCOM et 'Corpus Écrits'), mis en place en 2011 (cf. http://www.corpus-ir.fr/).

Pour mener des études comparatives à une échelle proprement panfrancophone, il faudra certes attendre qu'il existe des ensembles textuels suffisamment larges et comparables pour les diverses variétés diatopiques du français de la francophonie, et qui suffisent aux critères formulés ici. Si tout corpus est nécessairement restreint, un ensemble textuel de référence pour la lexicologie variationnelle contenant des réalisations écrites et orales de variétés diatopiques de la francophonie du Nord mais aussi du Sud serait un pas important. Un tel ensemble contribuerait considérablement à l'identification de diatopismes en tant que tels, et à en retracer les caractéristiques diverses – tout particulièrement des emplois 'rares' et des locutions, et des particularismes qui se distinguent justement par une différence de statut ou de fréquence. Il permettrait aussi de dépasser deux raccourcis majeurs : l'explication de la variation du français par les contacts de langue (du type 'germanisme' ou 'anglicisme'), et l'étiquetage géographique hâtif d'un emploi donné (du type 'belgicisme' ou 'québécisme'), par omission d'une étude sociolinguistique et aréologique sérieuse. De nouveaux corpus adaptés à l'analyse lexicale à visée diatopique ou un éventuel corpus de référence pour la lexicologie variationnelle renseigneraient non seulement sur les diatopismes du français contemporain, mais aussi sur le fonctionnement de la langue, vu la diversité des situations sociolinguistiques au sein de la francophonie.

Références

Adam, Jean-Michel & Heidmann, Ute (2006) Six propositions pour l'étude de la généricité. *La Licorne* 79 : 21–34.

André, Virginie & Canut, Emmanuelle (2010) Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français). *Pratiques : linguistique, littérature, didactique* 147/148 : 35–51.

Atkins, Sue ; Clear, Jeremy & Ostler, Nicholas (1992) Corpus design criteria. *Literary & Linguistic Computing* 7/1 : 1–16.

—— & Rundell, Michael (2008) *The Oxford Guide to Practical Lexicography* (coll. Oxford Linguistics). Oxford : Oxford University Press.

Baroni, Marco ; Bernardini, Silvia ; Ferraresi, Adriano & Zanchetta, Eros (2009) The WaCky wide web : A collection of very large linguisticially processed web-crawled corpora. *Language Resources and Evaluation* 43/2 : 209–226.

Baude, O. (éd.) (2006) *Corpus oraux, guide des bonnes pratiques*. Paris : Éditions du CNRS/Orléans : PU d'Orléans.

BDLP. Base de données lexicographiques panfrancophone constituée de vingt bases nationales, hébergée par le TLFQ, Laval (Québec) : AUF/TLFQ, en accès libre sur Internet <http://www.tlfq.ulaval.ca/bdlp>, aussi via le CNRTL, rubrique « Portail lexical » <www.cnrtl.fr>.

Benzitoun, Christophe & Cappeau, Paul (2010) Description sur corpus. Quelques réflexions autour des données et des instruments pour le français (parlé) à travers la description de *cause* et *causer*. In Franck Neveu ; Valelia Muni Toke ; Thomas Klingler ; Jacques Durand ; Lorenza Mondada & Sophie Prévost (éds), *Congrès Mondial de Linguistique Française – CMLF 2010 [La Nouvelle-Orléans, 12–15 juillet 2010]*, pp. 1383–1398. Paris : Institut de Linguistique Française. Publié en ligne le 12 Juillet 2010 <http://dx.doi.org/10.1051/cmlf/2010237> (20/06/2011).

Bert, Michel ; Bruxelles, Sylvie ; Étienne, Carole ; Mondada, Lorenza ; Teston, Sandra ; Traverso, Véronique ; Jouin-Chardon, Émilie & Justine, Lascar (2010) Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL). *Pratiques : linguistique, littérature, didactique* 147/148 : 17–34.

Blanche-Benveniste, Claire & Jeanjean, Colette (1986) *Le français parlé, transcription et édition*. Paris : Didier-Érudition.

BNC. *British National Corpus*. Corpus textuel de l'anglais britannique de la fin du XX^{e} siècle d'environ 100 millions de mots dont 10 millions de mots de textes 'oraux' (de conversations spontanées et autres enregistrements), d'une grande variété de sources et de genres, annoté linguistiquement et doté de métadonnées contextuelles selon les recommandations de TEI ; dernière version (de 2007) distribuée librement en XML (sous licence).

Branca-Rosoff, Sonia ; Fleury, Serge ; Lefeuvre, Florence & Pires, Mat (2012) *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)*, 23 pp. Paris : Université Sorbonne Nouvelle Paris 3. Publié en ligne <http://cfpp2000.univ-paris3.fr/> (24/07/2012).

Brochard, Marie-José (2012) Les ressources numériques en lexicologie historique. In Dörr, Stephen & Thomas Städtler (éds), Ki bien voldreit raisun entendre. *Mélanges en l'honneur du 70^{e} anniversaire de Frankwalt Möhren*, pp. 27–42. Strasbourg : ELiPhi.

Bruxelles, Sylvie ; Mondada, Lorenza ; Simon, Anne Catherine & Traverso, Véronique (2009 [2007]) Grands corpus de français parlé. Bilan historique et perspectives de recherches. *Cahiers de Linguistique* 33/2 : 1–14.

Cajolet-Laganière, Hélène ; Labrecque, Geneviève ; Martel, Pierre ; Masson, Chantal-Édith ; Mercier, Louis & Théoret, Michel (2008) Dictionnaires usuels du français

et Banque de Données Textuelles de Sherbrooke (BDTS) : convergence et divergence des nomenclatures. In Brigitte Horiot (éd.), *Français du Canada, Français de France.* Actes du septième Colloque international de Lyon, du 16 au 18 juin 2003 (coll. Canadiana romanica 22), pp. 9–28. Tübingen : Niemeyer.

Cappeau, Paul & Gadet, Françoise (2007) L'exploitation sociolinguistique des grands corpus. Maître-mot et pierre philosophale. *Revue Française de Linguistique Appliquée* 12/1 : 99–110.

—— & Seijido, Magali (2005) *Inventaire des corpus oraux en langue française*, document avec l'inventaire version 1.1 en annexe, accessible en ligne <www.dglflf.culture.gouv.fr>.

COCA. *Corpus of Contemporary American English*. La plus grande base textuelle de l'anglais américain contemporain, en accès libre, constitué de 425 millions de textes (de 1990 à 2011) et d'un large spectre de textes et de genres.

CREA. *Corpus de Referencia del Español Actual*. Corpus de référence de l'espagnol actuel du tournant des XX^e^ et XXI^e^ siècles constitué par la Real Academia Española, rassemblant environ 160 millions de mots (état de 2005) de types de textes divers annoté selon les genres et thèmes abordés, de toutes les variétés hispanophones (Argentine, Bolivie, Chili, Colombie, Costa Rica, Cuba, Le Salvador, Équateur, Espagne, États-Unis, Guatemala, Honduras, Mexique, Nicaragua, Panama, Paraguay, Pérou, Philippines, Porto Rico, République Dominicaine, Uruguay, Venezuela) – de 50% de textes espagnols, 50% hispano-américains, et de 90% de textes écrits, pour moitié journalistiques, et 10% de 'textes oraux', de 1975 au présent – en accès libre en ligne <http://www.rae.es>.

Delcourt, Christian (1998–1999) *Dictionnaire du français de Belgique*, vol. I–II. Bruxelles : Le Cri.

—— ; Francard, Michel & Moreau, Marie-Louise (1993) Une banque de données textuelles sur la langue française en Belgique. In Danièle Latin ; Ambroise Queffélec & Jean Tabi-Manga (éds), *Inventaire des usages de la francophonie : nomenclatures et méthodologies* (coll. Actualité scientifique), pp. 313–331. Paris : J. Libbey-Eurotext.

Delic (collectif) (2004) Présentation du Corpus de Référence du Français Parlé. *Recherches sur le français parlé*, 18 : 11–42.

DFQPrés. Poirier, Cl. (éd.) (1985) *Dictionnaire du français québécois. Volume de présentation*. Sainte-Foy (Québec) : PU Laval.

DHFQ. Poirier, Cl. (éd.) (1998) *Dictionnaire historique du français québécois. Monographies lexicographiques de québécismes*. Sainte-Foy (Québec) : PU Laval.

DictBelg. Francard, Michel ; Geron, Geneviève ; Wilmet, Régine & Wirth, Aude (2010) *Dictionnaire des belgicismes*. Bruxelles : De Boeck-Duculot.

Dister, Anne ; Francard, Michel ; Hambye, Philippe & Simon, Anne Catherine (2009 [2007]) Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de la banque de données textuelles orales VALIBEL (1989–2009). *Cahiers de Linguistique* 33/2 : 113–129.

—— ; Gadet, Françoise ; Ludwig, Ralph ; Lyche, Chantal ; Mondada, Lorenza ; Pfänder, Stephan ; Simon, Anne Catherine & Skattum, Ingse (2008) Deux nouveaux corpus internationaux du français : CIEL-F (Corpus International et

Écologique de la Langue Française) et CFA (Français contemporain en Afrique et dans l'Océan Indien). *Revue de Linguistique Romane* 72 : 295–314.

DRF. Rézeau, P. (éd.) (2001) *Dictionnaire des régionalismes de France (DRF).* Bruxelles : De Boeck-Duculot.

DSR. Thibault, André ([2]2004) *Dictionnaire suisse romand* [1997]. Genève : Zoé.

Durand, Jacques & Lyche, Chantal (2003) Le projet Phonologie du Français Contemporain (PFC) et sa méthodologie. In Elisabeth Delais-Roussarie & Jacques Durand (éds), *Corpus et variation en phonologie du français : méthodes et analyses*, pp. 212–276. Toulouse : PU du Mirail.

—— ; Laks, Bernard & Lyche, Chantal (2009) Le projet PFC : une source de données primaires structurées. In *id*. (éd.), *Phonologie, variation et accents du français*, pp. 19–61. Paris : Hermès.

Francard, Michel (2001) Le français de référence : formes, normes et identité. In *id*. (éd.), *Le français de référence. Constructions et appropriations d'un concept*, pp. 223–240. Louvain-la-Neuve : Institut de linguistique de Louvain.

Gadet, Françoise (à paraître) *Inventaire des corpus de français hors de France*. Site de la DGLFLF, <www.dglf.culture.gouv.fr>.

—— ([2]2007) *La Variation sociale en français* (coll. L'Essentiel français) [2003]. Paris : Ophrys.

—— & Jones, Mari (2008) Variation, contact and convergence in French spoken outside France. *Journal of Language Contact*, série THEMA, n° 2, 238–248.

—— ; Ludwig, Ralph ; Mondada, Lorenza ; Pfänder, Stephan & Simon, Anne Catherine (2012) Un grand corpus de français parlé : le CIEL-F. Choix épistémologiques et réalisations empiriques. *Revue française de linguistique appliquée* 17/1 : 39–54.

—— ; Ludwig, Ralph & Pfänder, Stefan (2009) Francophonie et typologie des situations. *Cahiers de linguistique* 34/1 : 143–162.

ILQ. *Index lexicologique québécois* : Inventaire des mots du français québécois ayant fait l'objet d'un commentaire ou d'une étude depuis 1750 jusqu'à nos jours. Fonds documentaire réalisé de 1979 à 1986 dans le cadre de la préparation du DHFQ sous la direction de Claude Poirier et de Louis Mercier, consultable sous forme informatique dans sa version actuelle reprogrammée en 2003 par Jean-François Smith, dans la rubrique « Fonds documentaires » du site du TLFQ, en accès libre <http://www.tlfq.ulaval.ca/ilq/>.

Kleiber, Georges (1978) *Le Mot « ire » en ancien français (XI[e]–XIII[e] siècles). Essai d'analyse sémantique* (coll. Bibliothèque française et romane). Paris : Klincksieck.

Ludwig, R. ; Telchid, S. & Bruneau-Ludwig, F. (éds) (2001) *Corpus créole. Textes oraux dominicains, guadeloupéens, guyanais, haïtiens, mauriciens et seychellois : enregistrements, transcriptions et traductions* (coll. Kreolische Bibliothek). 2 Cédérom, Hamburg : H. Buske.

Lux-Pogodalla, Veronika & Polguère, Alain (2011) Construction of a French Lexical Network : Methodological issues. In Benoît Sagot (éd.), *Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011 (Ljubljana, Slovenia, August 1–5 2011)*, pp. 55–62. Disponible en ligne <http://alpage.inria.fr/~sagot/woler2011/WoLeR2011/Program_%26_Proceedings.html> (25/01/12).

Masson, Chantal-Édith ; Cajolet-Laganière, Hélène & Martel, Pierre (2007) La BDTS-concordances : un outil d'enrichissement de la pratique lexicographique. *JADT 2004 : 7es Journées internationales d'Analyse statistique des Données Textuelles*, pp. 764–775. Disponible en ligne <http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_073.pdf>.

OEC. *Oxford English Corpus*. Corpus textuel de la langue anglaise le plus vaste de son genre avec plus de deux milliards de mots en format XML, tirés de textes de genres divers, annotés selon leurs types de texte, le type de variété d'anglais et l'identité des locuteurs, utilisé dans le cadre du *Oxford English Dictionary* et par le projet de recherche linguistique des Presses universitaires d'Oxford.

Poirier, Claude (2005) La dynamique du français à travers l'espace francophone à la lumière de la base de données lexicographiques panfrancophone. *Revue de Linguistique Romane* 69 : 483–516.

Queffélec, Ambroise (1997) Le corpus textuel oral. Constitution, traitement et exploitation lexicographique. In Claude Frey & Danièle Latin (éds), *Le Corpus lexicographique : méthodes de constitution et de gestion* (coll. Champs linguistiques), pp. 353–368. Louvain-la-Neuve : Duculot.

Rézeau, Pierre (2007) Des variétés dialectales gallo-romanes aux variétés régionales du français : la constitution d'un champ disciplinaire. In David Trotter (éd.), *Actes du XXIVe Congrès international de linguistique et de philologie romanes* (Aberystwyth, 1–6 août 2004), vol. IV, pp. 263–275. Tübingen : Niemeyer.

Sharoff, Serge (2006) Creating general-purpose corpora using automated search engine queries. In Marco Baroni & Silvia Bernardini (éds), *WaCky ! Working papers on the Web as Corpus*, pp. 63–98. Gedit : Bologna. Publié en ligne <http://wackybook.sslmit.unibo.it/> (25/01/2012).

Thibault, André (2000) Trois nouveaux dictionnaires différentiels de français : présentation et considérations méthodologiques. In Annick Englebert ; Michel Pierrard ; Laurence Rosier ; Dan Van Raemdonck (éds), *Des mots aux dictionnaires* (XXIIe ACILPR, Bruxelles, 23–29 juillet 1998), vol. IV, pp. 551–561. Tübingen : Niemeyer.

—— (2007) Banques de données textuelles, régionalismes de fréquence et régionalismes négatifs. In David Trotter (éd.), *Actes du XXIVe Congrès international de linguistique et de philologie romanes* (Aberystwyth, 1–6 août 2004), vol. I, pp. 467–480. Tübingen : Niemeyer.

TLF (1971–1994). *Trésor de la langue française. Dictionnaire de la langue du XIXe et XXe siècle (1789–1960)*, vol. I–XVI, publié sous la direction de P. Imbs (vol. I–X). Paris : CNRS Éditions, et de B. Quemada (vol. XI–XVI). Paris : Gallimard. Dictionnaire en accès libre avec ses mises à jours sur le site du laboratoire ATILF, CNRS/Université de Lorraine <http://atilf.atilf.fr/tlf.htm>.

Vézina, Robert (1998) Compte-rendu de : DSR. *Langues et Linguistique* 24 : 225–230.

Wierzbicka, Anna (2009) Exploring English phraseology with two tools : NSM semantic methodology and Google, *Journal of English Linguistics* 37/2 : 101–129.

Wissner, Inka (2010) *Les diatopismes du français en Vendée et leur utilisation dans la littérature : l'œuvre contemporaine d'Yves Viollier*, Thèse : Bonn/Paris :

Philosophische Fakultät der Universität Bonn, Abteilung für Romanistik / Université Paris-Sorbonne Paris IV, ÉD Concepts et langages. Disponible comme ressource électronique à l'URN urn:nbn:de:hbz:5-24001.

—— (à paraître [2013]) *Tombé en amour avec les corpus ?* What text corpora add to the study of expressions in modern French. In Henry Tyne ; André, Virginie ; Boulton, Alex et al. (éds), *Ecological and Data-Driven Perspectives in French Language Studies*. Cambridge : Cambridge Scholars Publishing.

Annexe

Pour une meilleure vue d'ensemble, on présentera les corpus, bases de données et plateformes en ligne pour l'étude du français contemporain dans une section à part. On a choisi de fournir un maximum de renseignements qui sont pertinents pour l'étude du lexique diatopiquement marqué. Toutefois, les informations qu'on a pu obtenir varient d'une ressource à l'autre, celles-ci étant tributaires des priorités des différents auteurs des corpus, et des renseignements qui sont mis à la disposition de la communauté scientifique (description en date du 22/08/2012).

I. Plateformes et portails

CLARIN. *Common Language Ressources and Technology Infrastructure of the Institute for Language and Speech Processing*. Infrastructure de recherche européenne synthétisant la description de corpus de nombreuses équipes internationales pour diverses langues du monde (890 corpus) et fournissant les hyperliens aux sites respectifs ; contenant 55 corpus [surtout] écrits hétérogènes de langue française (dont quinze uniquement français), parmi lesquels FRANTEXT, GALLICA, Les Voisins du Monde, Morphalou et Ananas (état 09/09/2011), conçus pour des recherches variées, et contenant des données allant du XVI^{e} au XX^{e} siècles <http://www.clarin.eu/view_resources>.

CNRTL. Portail du CNRS fédérant un ensemble de ressources linguistiques informatisées et d'outils de traitement de la langue, donnant accès à des corpus de français écrit, notamment littéraire et journalistique, et à un corpus oral transcrit et homogénéisé à Nancy (TCOF), Nancy : ATILF, CNRS/Université de Lorraine, en accès libre <www.cnrtl.fr>.

Réseau des corpus lexicaux québécois. Plateforme mise en place par le Secrétariat à la politique linguistique SPL du Québec, présentant

quinze corpus de français québécois exploitables en ligne, et fournissant une interface d'interrogation par mot-clé (sans recherche de co-occurrences) ainsi que les liens aux sites permettant d'accéder aux corpus provenant de cinq universités québécoises, dont FLI, ILQ, BDLP et QUEBETEXT de Laval, LEXIQUM de Montréal, Corpus du Témiscouata de Rimouski, et BDTS. En accès libre en ligne <http://www.spl.gouv.qc.ca/languefrancaise/corpuslexicaux/description/>.

SLDR. *Speech & Language Data Repository* (anciennement CRDO-Aix) : Banque de données parole et langage archivées au CINES et distribuées par le CC-IN2P3, constituée sous la coordination scientifique de Philippe Blache et Daniel Hirst visant à pérenniser et partager des corpus de parole et à les enrichir à l'aide de transcriptions et d'annotations, en accès libre sur Internet (Open Archives, CLARIN), sous condition (certaines catégories d'utilisateurs, après acceptation d'une licence non-commerciale, selon le Code du patrimoine), hébergée par l'Université de Provence <http://sldr.org>.

II. Corpus et bases de données

BDTS. Banque de données textuelles de Sherbrooke, exploitable avec l'outil BDTS-concordances, recueillant plus de 52 millions de mots tirés de quelque 15 000 'textes représentatifs des différents usages du français en usage au Québec' couvrant les années 1960 à 2000 – textes littéraires, journalistiques, didactiques, et spécialisés (notamment techniques, scientifiques, sociopolitiques, administratifs, juridiques et culturels), et un sous-corpus oral d'environ 2 millions de mots (transcriptions d'enquêtes sociolinguistiques orales et de discours plus formels comme des téléromans et des textes radiophoniques et télévisés) ; corpus préparé dans le cadre des travaux du Centre d'analyse et de traitement informatique du français québécois (CATIFQ) pour servir de base à l'élaboration de la nomenclature du dictionnaire du projet de dictionnaire du français québécois de l'équipe FRANQUS en préparation sous la direction éditoriale d'Hélène Cajolet-Laganière et Pierre Martel, Sherbrooke (Québec) : Université de Sherbrooke. En accès limité aux collaborateurs, diffusion prévue sur le Site de l'Université où un sous-ensemble contenant plus de 16 millions de mots est accessible <http://www.usherbrooke.ca/catifq/accueil/>, rubrique « Recherches ».

BELTEXT. Base de données textuelles sur la langue française en Belgique et sur la littérature belge de langue française constituée de corpus écrits et oraux, transcrits en orthographe conventionnelle et accompagnée des enregistrements originaux, permettant des recherches littéraires et linguistiques pour des exploitations multiples (lexicale, morphologique, syntaxique, pragmatique, etc.) ; contribution de la Communauté Wallonie-Bruxelles au projet international BDLP, dans le cadre de la préparation d'un dictionnaire de belgicismes par une équipe interuniversitaire dirigée à l'origine par Delcourt (Liège), Francard (Louvain-la-Neuve) et Moreau (Mons) ; renseignements non obtenus au sujet des possibilités de consultation auprès de Christian Delcourt.

CERF. Corpus Évolutif de Référence du Français. Corpus de l'équipe DELIC (sous la direction de Jean Véronis) comportant 9 millions de mots, répartis en neuf tranches (à 1 million de mots chacune) et équilibrés discursivement entre neuf types de discours tirés de l'écrit (récupérés en partie sur Internet), y compris de magazines, textes scientifiques et de la presse nationale et régionale, ainsi qu'une tranche orale (CORPAIX) d'environ 1 million de mots ; corpus présenté en ligne <http://sites.univ-provence.fr/delic/corpus/index.html>.

CFA (*en cours*). Corpus de Français contemporain en Afrique et dans l'Océan Indien, projet d'une équipe internationale menée sous la direction d'Ingse Skattum initié en 2006 à Oslo pour compléter le projet PFC, portant sur huit zones (Burkina Faso, Cameroun, Centrafrique, Côte d'Ivoire, La Réunion, Mali, Maurice, Sénégal), recueillant des productions formelles et des entretiens semi-directifs (comme le PFC), des entretiens sur les usages et attitudes des locuteurs, et des conversations libres selon les contextes locaux, avec douze locuteurs par point d'enquête ; ensemble de données comparables pour les variétés considérées, complété de données de genres variés (cours magistral, prêche, débats télé, etc.) qui sera soumis à des outils d'indexation phonologique et syntaxique pour permettre des analyses phonologiques, syntaxiques et sociolinguistiques, transcrit, étiqueté, codé et aligné texte/son/métadonnées ; diffusion en accès libre sur Internet prévue.

CFPP2000. Discours sur la ville. Corpus de Français Parlé Parisien des années 2000. Corpus composé d'un ensemble d'interviews sur les quartiers de Paris et de la proche banlieue (Sonia Branca-Rosoff,

Serge Fleury et Florence Lefeuvre). Corpus en accès libre en ligne, sous licence <http://cfpp2000.univ-paris3.fr/>.

CFPQ. Corpus de français parlé au Québec préparé depuis 2006 dans le cadre des travaux du Centre d'analyse et de traitement informatique du français québécois (CATIFQ) de l'Université de Sherbrooke sous la responsabilité de Gaétane Dostie ; corpus multimodal visant à 'refléter le français québécois' en usage dans les années 2000 qui rassemble des enregistrements effectués sur support audiovisuel de plus de 45 heures et les transcriptions alignées, réalisées à l'aide du logiciel Transana (382.181 mots le 24/05/12, statistique non mise à jour) ; support audiovisuel consultable sur place uniquement ; transcriptions en accès libre sur le Site de l'Université de Sherbrooke <http://pages.usherbrooke.ca/cfpq/corpus.php>.

CIEL-F (*en cours*). Corpus International et Écologique de la Langue Française : données audio et vidéo établies dans des situations comparables dans différents pays choisis selon une typologie des aires et situations d'usages (Acadie, Algérie, Antilles, Belgique [Bruxelles-Brabant Wallon], Cameroun, Côte d'Ivoire, Égypte [Le Caire], France, La Réunion, Liège, Maurice, Québec, Ontario, Sénégal, Suisse) ; initiative franco-belgo-allemande (Lyon, Paris, Freiburg, Halle, Louvain) menée sous la responsabilité de Lorenza Mondada (Lyon) et Stefan Pfänder (Freiburg), en partenariat avec des équipes internationales ; diffusion libre sur Internet prévue via les interfaces [moca] et CLAPI <www.ciel-f.org>, copyright 2008–2012.

CLAPI. Base de données orales contenant 46 corpus, constitués d'enregistrements audio ou vidéo (non média) pour permettre l'analyse de l'interaction (Lyon 2) <http://clapi.univ-lyon2.fr/>.

CORPAIX. Corpus d'enregistrements d'oral réalisés par le Groupe Aixois de Recherche en Syntaxe (GARS), dont la version de 2000 (CORPAIX2000) comporte environ un million de mots , cf. <http://sites.univ-provence.fr/delic/corpus/index.html>.

CRFP. Corpus de Référence du Français Parlé. Corpus présentant des situations variées avec une majorité d'entretiens, dont les enregistrements ont été effectués dans diverses villes de l'Hexagone et selon trois situations de parole (privée, publique et professionnelle) (CRFP-1). Corpus d'environ 440 000 mots transcrits par des transcripteurs multiples selon les Conventions de transcription orthographiques (conventions DELIC), constitué sous la responsabilité de Jean Véronis (Aix-en-Provence) en vue d'études

linguistiques, notamment en syntaxe <http://sites.univ-provence.fr/delic/crfp/> ; sa diffusion par la Délégation DGLFLF est prévue.

Dynamiques des français périphériques (*en préparation*). Projet inter-composantes du laboratoire MoDyCo (Modèles, Dynamiques, Corpus) lancé en 2008 sous la direction de Françoise Gadet et Colette Noyau en partenariat avec des équipes de trois continents, visant la mise en réseau de chercheurs et de laboratoires pour rassembler des données pertinentes pour des comparaisons d'usage dans les différentes variétés diatopiques du français (extraction et mise en forme de corpus disponibles ou collecte coordonnée de données spécifiques) ; les travaux et/ou corpus des partenaires actuels portent sur le français aux Îles anglo-normandes, en Acadie, Ontario et Louisiane, à l'Île Maurice et à la Réunion, au Cameroun, au Gabon, au Sénégal, au Togo, en Côte d'Ivoire, et en Tunisie ; présentation prévue sur le site *Aspects acquisitionnels et sociolinguistiques des dynamiques des français* (ATRADY).

ESLO. Enquête Sociolinguistique à Orléans. Corpus oral constitué au Laboratoire Ligérien de Linguistique (LLL, ex-CORAL) de l'université d'Orléans en collaboration avec d'autres laboratoires, constitué à partir de 700 heures d'enregistrements, dont 300 heures de 1968 à 1971 (ESLO-1, à visée didactique) et 400 heures d'enregistrements comparables dans les modalités de collecte, depuis les années 2000 (ESLO-2, à visée variationniste) ; accès libre sur Internet annoncé pour 01/07/2012, non effectif en date du 21/08/2012 <http://www.univ-orleans.fr/eslo/>.

L'Est Républicain, corpus constitué d'articles de toutes les éditions régionales et locales du quotidien régional de l'est de la France de 1999, 2002 et 2003, consultable via le CNRTL en format XML-TEI P5 en trois fichiers séparés par année <http://www.cnrtl.fr/corpus/estrepublicain/> ; corpus en projet d'extension pour rassembler les textes des années 2005 à 2010.

EUREKA. Corpus de presse multilingue (anciennement *Biblio Branchée*) depuis les années 1980 surtout (sans indication exacte) couvrant l'actualité internationale, nationale, régionale et locale en douze langues dont le français, permettant des requêtes thématiques par mots-clés et en principe par suites de mots (option en partie inopérable), régulièrement enrichi, rassemblant actuellement 6023 sources référencées (état du 06/01/12, sans indication de taille),

consultables par langue, région, date de publication, domaine thématique et parties de texte par une interface affichant les 300 premiers documents ; corpus comportant des parts égales de textes de la presse généraliste et spécialisée – surtout des journaux (comme *Le Monde* et *Le Monde diplomatique*, pour la France), publications spécialisées, fils de presse, émissions télévisées et radiophoniques transcrites, blogues triés sur le volet, sites Web référencés, rapports). Accès sous licence, Copyright CEDROM-SNi inc. 2012 <http://www.eureka.cc/Default.aspx>.

EUROPRESSE. Presse d'information francophone et anglophone en texte intégral qui réunit le texte des publications suivantes : *AFP Général*, dep. 19/03/01 ; *L'Express*, dep. 07/01/93 ; *L'Humanité*, dep. 16/11/1999 ; *La Croix*, dep. 01/09/95 ; *Le Figaro*, dep. 31/10/1996 ; *Le Monde*, dep. 01/01/1997 ; *Le Monde diplomatique*, dep. 01/01/1984 ; *Le Parisien*, dep. 02/05/1998 ; *Le Point*, dep. 07/01/1995 ; *Les Échos*, dep. 02/01/1991 ; *Libération* dep. 02/01/1995 ; accès payant aux archives sur Internet <www.europresse.com>.

FLI. Fichier lexical informatisé. Base de données du français québécois du XVI^e^ au XX^e^ siècles comportant 400 000 fiches avec une ou plusieurs citations chacune, essentiellement d'emplois québécois (état 02/2010), tirées d'un corpus de citations papier constitué pour l'essentiel entre 1975 et 1990 en vue du DHFQ à l'Université de Laval (Québec), contenant plus de 1,2 millions de fiches manuscrites de sources diverses : récits anciens, documents d'archives manuscrits ou imprimés, documents administratifs, journaux et magazines, littérature et textes de création, études et textes spécialisés, manuscrits de radioromans et de téléromans, enregistrements oraux, relevés d'enquêtes sur le terrain. En accès libre sur Internet avec deux modes de recherche, « dans les entrées » et « dans les citations » <http://www.tlfq.ulaval.ca/fichier/>.

FRANTEXT. Base de données textuelles de la littérature française conçue dans le cadre de la préparation du TLF à l'ATILF, CNRS/Université de Lorraine ; corpus à dominante littéraire constitué de quelque 248 millions de mots, de 4.084 références de textes français (état du 09/09/11) du XVI^e^ au XXI^e^ siècles – appartenant aux domaines des sciences, des arts, de la littérature, et des techniques – consultable au format TEI avec le logiciel Stella en ligne sur abonnement, permettant de visualiser des extraits textuels de 700 signes

<http://www.frantext.fr/> ; un corpus réunissant 500 œuvres de la littérature française du XVIII^e^ au XX^e^ siècles libres de droits est en accès libre <http://www.cnrtl.fr/corpus/frantext/>.

FRWAC. Corpus textuel d'une tranche du Web (domaine « .fr ») d'environ 1,6 milliards de mots, construit dans le cadre du projet WaCky Wide Web (Trente/Bologne) par interrogation du web avec des paires de mots et par filtrage et nettoyage des pages, indexé avec CorpusWorkBench, étiqueté et lemmatisé avec TreeTagger. En accès libre sur demande <http://wacky.sslmit.unibo.it/doku.php?id=download> ; également exploitable sous une version catégorisée par degré de normativité orthographique et grammaticale, conçue à l'ATILF, CNRS/Université de Lorraine (Nancy) en collaboration avec Druide Informatique Inc. (Montréal). Sa mise à disposition à la communauté scientifique est envisagée par l'intermédiaire des chercheurs à l'origine du projet WaCky Wide Web, sauf si une diffusion via le site CNRTL est accordée et juridiquement sûre ; pourra y figurer au moins un lien vers la base.

GALLICA. Bibliothèque numérique de la Bibliothèque nationale de France (Paris) : collections de manuscrits occidentaux et orientaux d'époques diverses numérisées depuis 2007, en accès libre sur Internet <http://gallica.bnf.fr/>.

GRL. Google Recherche de Livres : Base de données textuelles scannées et traitées par Google, en accès libre sur Internet <http://books.google.fr/>, consultée en mode de Recherche « Livres entiers ou en aperçu limité » [données à vérifier sur l'original].

HANSARD. Corpus en français au Canada des *Débats de la Chambre des communes* constitué des comptes rendus *in extenso* des débats ayant lieu à la Chambre et en comité plénier du Parlement du Canada, depuis le 35^e^ Parlement de janvier 1994 au jour d'aujourd'hui (41^e^ Parlement) ; textes préparés à partir des enregistrements sonores des délibérations ainsi que des renseignements fournis par le personnel du Service des comptes rendus en poste sur le parquet de la Chambre. Corpus publiés en français et anglais et doté d'un référencement minimal, saisi, enregistré et diffusé à l'aide du logiciel PRISME. En partie en accès libre sur Internet (diffusion de l'ensemble prévue ; taille inconnue), avec une possibilité de requête en ligne par mots-clés et séquences exactes et exclusion d'autres mots-clés <http://www.parl.gc.ca/HouseChamberBusiness/ChamberSittings.asp>.

I-FR. Corpus francophone de Leeds tiré de 50 000 pages Internet du Web francophone en 2006 sans limitation de domaines, d'environ 200 millions de mots, établi au *Centre for Translation Studies* de l'Université de Leeds avec un traitement de recherche automatisé par interrogation du web avec des paires de mots, indexation avec CorpusWorkBench, filtrage et nettoyage des pages, et étiquetage/lemmatisation avec TreeTagger (selon la même procédure que FRWAC), avec annotation automatique par thèmes selon des groupes thématiques (*thematic clusters* ; sans vérification manuelle), qui comprennent des informations de type géographique (comme 'congolais', 'Liban' ou 'Québec') ; corpus consultable via un logiciel d'exploitation permettant des recherches de concordances et de collocations. En accès libre sur Internet (Corpus.Leeds.ac.uk/internet.html).

Kölner romanistische Korpusdatenbank. Corpus écrit préparé par Sascha Diwersy sous la responsabilité de Peter Blumenthal (Cologne) constitué de textes journalistiques des années 2000 (totalisant environ 80 millions de mots), pour moitié environ d'Europe, surtout de la France continentale (presse régionale et nationale : *Le Figaro, L'Est Républicain, Sud-Ouest*), et pour plus de la moitié de presse nationale d'Afrique, en particulier du Cameroun (*Cameroon Tribune, Mutations*), du Sénégal (*Le Soleil*) et de Côte d'Ivoire (*Fraternité Matin*) ; corpus complété d'environ 100 romans publiés depuis 1950 d'auteurs de tous les états francophones d'Afrique noire, en particulier du Cameroun, du Sénégal et de Côte d'Ivoire ; corpus annoté conçu pour des recherches qualitatives et quantitatives de collocations ; plus de 5 millions de mots lemmatisés et syntaxisés ; consultable sur place à l'Université de Cologne, diffusion libre prévue pour 2012.

Le Monde (2003). *Le Monde : L'histoire au jour le jour 1939–2002*, cédérom, Coedrion : Le Monde, Emme et IDM.

Le Monde diplomatique (2011). *Le Monde Diplomatique. 43 années d'archives sur DVD-ROM (1968–2010)*, nouvelle édition. Ressource informatique contenant plus de 40 000 documents, de l'édition française (1968–2010) ainsi que de cinq des éditions étrangères en version originale et sous forme de traductions, Coedrion : Le Monde, Emme et IDM.

MCVF. Modéliser le changement : les voies du français. Corpus constitué de textes en grande partie intégraux de genres discursifs divers tels que des correspondances, romans et documents administratifs, du

français de diverses variétés diatopiques (notamment d'Amérique du Nord), 'structuré de façon dialectale, sociale et historique' et couvrant quatre périodes historiques : ancien, et moyen français, français du XVI^{e} siècle, et français classique (France et Nouvelle-France) (XVII^{e} et XVIII^{e} siècle) (documents 'sources' du XI^{e} au XIX^{e} siècle) ; corpus constitué sous la direction de France Martineau dans le cadre des « Grands Travaux de Recherché concertée » de l'Université d'Ottawa, copyright 2009 ; corpus en format XML-TEI permettant des requêtes sur des mots et des co-occurrences de mots à l'aide d'un concordancier en accès libre sur demande en cédérom et en ligne <http://www.voies.uottawa.ca/corpus_pg_fr.html>.

PFC (*en finalisation*). Phonologie du Français Contemporain : usages, variétés et structure. Base de données ouverte préparée dans le cadre d'un projet international sous la direction de Jacques Durand, Bernard Laks et Chantal Lyche (MoDyCo), recueillies en 60 points d'enquête dans l'ensemble de la francophonie avec dix locuteurs échantillonnés par point selon une méthodologie commune comportant une lecture de mots, une lecture de texte, une conversation soutenue et un dialogue informel, représentant un total de 900 heures d'enregistrements transcrits, dont un ensemble d'environ un million de mots consultable par la communauté scientifique sur demande auprès d'Atanas Tchobanov (état 25/09/2011) <http://www.projet-pfc.net>, copyright 2004–2008.

POLITEXT. Base de données construite à l'Université de Nice rassemblant cinq cents discours des grands hommes politiques français (de Jaurès à Jospin, de Poincaré à Chirac) couvrant le XX^{e} siècle et correspondant à 10 millions d'occurrences, disponible en format word (actuellement indisponible sur Internet) <http://www.unice.fr/ILF-CNRS/politext/>.

QUEBETEXT. Base de données textuelles du Trésor de la langue française au Québec. Contribution québécoise au projet international du « Trésor des vocabulaires francophones » (TVF) – qui visait à créer un ensemble de fonds textuels compatibles avec la base FRANTEXT – élaborée depuis les années 1990 dans le cadre des travaux du TLFQ de l'Université de Laval, Québec pour la préparation du DHFQ, réunissant des textes de genres discursifs différents, à dominante littéraire, d'auteurs québécois des XIX^{e} et XX^{e} siècles (textes publiés depuis 1837, les années 1960 étant les mieux représentées). Corpus consultable à l'aide du logiciel TACT, permettant des recherches

d'occurrences et de co-occurrences de mots selon divers paramètres ; accès individuel libre à quatre corpus de texte intégral libres de droits sur Internet : « Littérature québécoise (1837 à 1919) », « Textes sur l'anglicisme (1826–1930) », « Témoignages des voyageurs (1651–1899) », « Préfaces de répertoires lexicaux (1841–1957) » <www.tlfq.ulaval.ca/quebetext/>.

REGION. Banque de données de régionalismes du français hexagonal conçue dans le cadre de la préparation du DRF à Nancy, réunissant 7500 contextes de 220 ouvrages dus à 156 auteurs de France métropolitaine (le Centre-ouest exclu, sauf *Sentiers d'eau*, Mathé 1978), annoté par région d'appartenance des écrivains ; corpus consultable sur place au centre du FEW, ATILF, CNRS/Université de Lorraine.

SCIENTEXT. Base de données textuelles d'écrits scientifiques, totalisant 4,8 millions de mots de 219 textes en français, de plusieurs disciplines (surtout linguistique et science de l'éducation) (version 1.3.9, état du 01/06/2012) ; conçue à l'Université Stendhal (laboratoire LIDILEM, Grenoble) en collaboration avec les Universités Bretagne Sud (LI-CORN) et de Savoie (LLS), pour permettre l'étude des caractéristiques linguistiques des textes scientifiques à travers des structures phraséologiques et des marques lexicales et pour des requêtes sémantiques et syntaxiques ; corpus annoté syntaxiquement (*Treetagger*) et structurellement (TEI Lite) avec isolation des parties textuelles, consultable en trois modes – sémantique, simple guidé (par lemme, catégorie et relations syntaxiques), et complexe (par expressions régulières). En accès libre sur Internet <http://scientext.msh-alpes.fr>.

SUISTEXT. Base de données textuelles sur la littérature suisse romande contemporaine de langue française, contenant la totalité de l'œuvre de quatorze écrivains romands contemporains du XX^{e} siècle ; contribution suisse au projet international du TVF élaborée dans le cadre de la préparation du DSR sous la responsabilité de Pierre Knecht à l'Université de Neuchâtel ; corpus numérisé de 129 textes consultable par les collaborateurs du Centre de dialectologie et d'étude du français régional à l'Université de Neuchâtel.

TCOF. Traitement de Corpus Oraux du Français. Base de données de corpus oraux de français contemporain hexagonal, avec alignement texte/son (avec le logiciel *JTrans* ou transcriber, format .wav) constitué depuis 2005 à l'ATILF, CNRS/Université de Lorraine, comportant des

enregistrements de données orales numérisées et transcrites en TRS d'échanges adultes/enfants et entre adultes (surtout des entretiens, conversations et réunions de travail), de locuteurs français originaires de Lorraine pour la plupart ; sur approximativement quatre millions de mots enregistrés, environ un demi-million sont consultables avec un concordancier en accès libre (sous forme de fichiers séparés) via la plateforme CNRTL <http://www.cnrtl.fr/corpus/tcof/>.

VALIBEL. Banque de données textuelles du français oral en Belgique constituée par le Centre de recherche VALIBEL à Louvain-la-Neuve (1989–2009) visant à documenter la variation linguistique en Belgique francophone (Wallonie et Bruxelles), rassemblant un ensemble d'environ 4 millions de mots transcrits et alignés (gérée par le logiciel Praat) dans une base de données dynamique et évolutive en ligne, diffusé (sauf les tous derniers enregistrements) à l'aide du système [moca] sur demande avec mot de passe, pour une durée limitée <http://www.uclouvain.be/valibel-corpus.html>.

VALIRUN. Banque de données textuelles orales du français et créole réunionnais établi sous la responsabilité de Gudrun Ledegen, transcrites selon des conventions convergentes au corpus VALIBEL, totalisant environ 1 million de mots ; diffusion libre sur Internet prévue.

ZOBEL. Banque de données textuelles de littérature antillaise francophone comprenant l'œuvre en prose de l'écrivain martiniquais Joseph Zobel établie par André Thibault – onze ouvrages publiés de 1946 à 2002 (en onze fichiers .rtf) – totalisant 2044 pages, numérisées à l'aide d'Omnipage SE de Scansoft et annotées en XML par Patrick Drouin, consultable sur demande auprès des auteurs.

Coordonnées :

Inka Wissner
CNRS, ATILF
44 Avenue de la Libération
B.P. 30687
F-54063 Nancy Cedex
France
e-mail: inka(dot)wissner(at)atilf(dot)fr

Squibs

Kalyanamalini Sahoo

Telicity vs. Perfectivity: A Case Study of Odia[1] Complex Predicates

Abstract

This paper studies the telic feature of light verbs in Odia complex predicates. Odia V-v sequences, carrying a main verb and a light verb, both of which are semantically predicative and hence form a complex predicate, denote telicity of the action/event. Such light verbs are compatible with the morpho-syntactic perfective aspectual morpheme, which provides evidence that the perfective morpheme and the light verbs are different subtypes of telic features.[2]

1. Introduction

In the linguistic literature, clearly there is a distinction between morpho-syntactic tense and aspect markers on the one hand, and the notion of semantic aspect, on the other. 'Morpho-syntactic tense and viewpoint aspect, which reflects the perfective/imperfective distinction, are determined by the Reference time system based on the relations established between Reference time, Speech time, and Event time' (Borik & Reinhart

[1] 'Odia', formerly known as 'Oriya', is an Indo-Aryan language, spoken in Odisha (formerly known as 'Orissa'), an eastern state in India.

[2] I am grateful to the anonymous reviewer of this article for the constructive comments and suggestions. My sincere thanks to Mia Raitaniemi for her comments as well. All the errors are mine, of course.

2004). They clearly distinguish 'perfectivity' from semantic aspect, such as *telic* and *atelic* aspects. A verb or VP that presents an action or event as being *complete* or having an endpoint, is said to be *telic*, and an action or event being *incomplete or lacking an endpoint* is *atelic*. Thus, telicity indicates a reached goal or action completed as intended.

Telicity and *boundedness* have often been treated as kindred terms (Kamp & Reyle 1993; de Swart 1998; Kratzer 2004; Giorgi & Pianesi 2004) as both the terms denote the culmination of events. Although (a)telicity and (un)boundedness are frequently used as synonyms, still quite a number of researchers have argued for separating telicity and boundedness (Verkuyl 1993; Depreaetere 1995; Klein 1995; Borik 2006). 'Telicity' is considered by these researchers as an extra-linguistic parameter, mostly determined by pragmatic factors, while 'boundedness' is defined as a lexico-grammatical property of the linguistic predication. Following Depraetere (1995), we draw a distinction between the notions (a)telicity and (un)boundedness:[3]

> (A)telicity has to do with whether or not a situation is described as having an inherent or intended endpoint; (un)boundedness relates to whether or not a situation is described as having reached a temporal boundary (Depraetere 1995: 2–3).

In terms of Depraetere, telicity relates to the "potential actualization" of a situation, whereas the boundedness parameter measures the "actual realization" of the situation. The telicity parameter is thus based on the attainment of a terminal point, which may be "inherent" or "intended".

In this article, we study the telic feature of light verbs in Odia complex predicates. The article is structured as follows: section 2 gives a short overview of Odia complex predicates, section 3 deals with the impact of light verbs for the telicity of a construction. Section 4 compares and contrasts telicity with perfectivity, and finds out that semantic aspect like (a)telicity and morphological aspect like (im)perfectivity are two independent systems. In section 5, we investigate the completive aspectual feature of light verbs indicated through time adverbials and negation, and establish that light verbs contribute towards telicity in V-v complex predicates.

[3] In this article, we do not consider the boundedness property of light verbs, the focus here is only on telic features of light verbs.

2. Odia complex predicates

Complex predicates or asymmetrical serial verb constructions in Odia (Sahoo 2001) consist of a sequence of two verbs. Such verbal complexes consist of an uninflected verb and an inflected verb: the uninflected one is a combination of the verb stem and a conjunctive morpheme, and the inflected one is a verb stem carrying Tense, Aspect, and Modality (TAM) features. They form a complex predicate as they involve the combination of two predicates in a single clause. In such V-v complexes, usually the first verb is a main verb and the second one is a fully or partially bleached light verb, where the main verb carries the lexical semantic information; the sequence as a whole determines the argument structure, e.g.:

(1) a. *mun gata dui ghanTaa-re dui-Taa kaahaaNi* ***lekh-il-i***,
I last two hours-in two-CL[4] story write-PAST-1SG
kintu puraapuri nuhen
but completely be.NEG
'In the last two hours I wrote two stories, but not completely.'

b. *mun gata dui ghanTaa-re dui-Taa kaahaaNi* ***lekh-i*** ***de-l-i***
I last two hours-in two-CL story write-CONJ give-PAST-1SG
*(*kintu puraapuri nuhen)*
(*but completely be-NEG)
'In the last two hours I wrote two stories (*but not completely).'

In (1b), the main verb *lekh* 'write' and the light verb *de* 'give' form a sequence. Although the sentence can be formed by using a single verb predicate as in (1a) or a V-v sequence like (1b), semantically both the constructions denote different meanings. (1b) entails the natural culmination of endpoint, while (1a) entails the arbitrary culmination as well as possible cancellation of endpoint. Note that there is no aspectual morpheme in either of the constructions. So, following Depraetere (1995), both the predicates in (1a) and (1b) are 'bounded' and also can be considered as 'telic', though (1a) still gives a chance to consider it atelic (with the conjoined clause representing the terminal point is not achieved and the action can be continued further), while (1b) is telic through. The

[4] The abbreviations used in this article are as follows: classifier (CL), conjunctive (CONJ), infinitival (INF), subjunctive (SUBJN), imperative (IMP), causative (CAUS), gerundive (GER), auxiliary (AUX), perfective (PERF), progressive (PROG), etc.

light verbs, thus, carry the semantic feature of telicity, and such a telic feature is not inherent in the full verbs of these constructions.

In Odia, while the regular verbs require the (morpho-syntactic) (im)perfective aspectual morpheme to mark (a)telicity, as in example (2); the light verbs mark it by default, as in example (3).

(2) a. *skul jibaa purbaru mun goTie*[5] *kabitaa* ***lekh-i-th-il-i***
school going before I a poem write-PERF-AUX-PAST-1SG
'Before going to school, I had written a poem.'

b. *skul jibaa purbaru mun goTie kabitaa* ***lekh-u-th-il-i***
school going before I a poem write-PROG-AUX-PAST-1SG
'Before going to school, I was writing a poem (and not finished it yet).'

(3) a. *skul jibaa purbaru mun goTie kabitaa*
school going before I a poem
lekh-i de-i-th-il-i
write-CONJ give-PERF-AUX-PAST-1SG
'Before going to school, I had already written a poem.'

b. *skul jibaa purbaru mun goTie kabitaa*
school going before I a poem
lekh-i de-u-th-il-i
write-CONJ give-PROG-AUX-PAST-1SG
'Before going to school, I used to write a poem (and finished it).'

In the above examples, (2a) and (2b) indicate telicity and atelicity, respectively, because of the perfective and progressive aspectual morphemes. However, in example (3), because of the V-v sequence, both the constructions are telic, irrespective of the presence of a progressive aspectual morpheme in (3b). Note that here, the progressive morpheme does not refer to the continuation of the situation, rather it refers to a meaning of repetition of the completed action ['the number of times I went to school, I used to write a poem before going']. So, on the basis of the above examples, we can say that the aspectual semantics of Odia is more complex than only distinguishing perfective and progressive actions/events. The telic feature seems to be an additional semantic feature that needs to be analyzed here to clarify what the relation of all these aspectual features is.

[5]The upper-case letters in the transcription stand for retroflex phonemes.

Traditional studies on aspectual system in Odia (Nayak 1987; Pradhan et al. 1995; Mohapatra & Das 1962; Sahoo 2001) refer to two types of aspects:

> 1. The perfective aspect which expresses completeness, referring to completely done action or conceptualized to be complete, action drawn to the last point, totality or the result of actions.
> 2. The imperfective (progressive) aspect which expresses only actions in progress or repetition of action.

However, the traditional viewpoint does not explain why imperfective verb forms are used to express telic actions in some contexts, while in others their use is restricted. Also, it does not discuss how telicity is marked in the absence of a morphologically realized aspect marker. In addition, the traditional viewpoint does not sufficiently explain the use of aspect in infinitive, imperative and subjunctive constructions. In this study, we address all these phenomena and discuss how in the case of complex predicates, telicity is realized in the absence of a morphologically realized aspect marker.

Complex predicates often play the role of a morpho-syntactic gap-filler, expressing directional, aspectual or case-marking functions by lexical rather than morphological means. The light verbs, although form a closed class, manifest themselves in a wide range of constructions. The light verbs that occur in Odia as second element in V-v constructions can be grouped according to their original lexical semantic value as follows (Lemmens & Sahoo in progress):

- MOTION verbs: *jaa* ‘go’, *chaal* ‘walk’, *paD* ‘fall’, *pakaa* ‘drop’, *uTh* ‘rise’, *aas* ‘come’
- STATIVE verbs: *bas* ‘sit’, *rah* ‘stay’
- TRANSFER verbs: *de* ‘give’, *ne* ‘take’

Obviously, these verbs cannot co-occur randomly with any verb, but show clear semantic constraints, as well as differences in frequency.

The terms *telic* and *atelic* are not traditionally used in Odia grammatical description; instead, it is customary to speak of *completion* of the event/action denoting a result as in example (1). The resultative construction as in (1b) *necessarily* requires finishing the story writing; while in the irresultative/unbounded construction as in (1a), it is not expressed whether the situation has really ended, thus giving a chance for

the repetition of the action in future. This matches with Depraetere's notion of (a)telicity:

> A clause is telic if the situation is described as having a natural or an intended endpoint which has to be reached for the situation to be complete and beyond which it cannot continue. Otherwise it is atelic. (Depraetere 1995: 3)

Comparing it with Comrie's definition of a telic situation:

> (…) "a telic situation is one that involves a process that leads up to a well-defined terminal point, beyond which the process cannot continue" (Comrie 1976: 45).,

we can say that (2b) is a telic construction.

3. Light verbs denoting telicity

Almost all light verbs can have an aspectual function indicating telicity or completion of an event. Consider the following examples.

The light verb *pakaa* 'put' denotes that the action is seen as happening suddenly or abruptly, with a telic effect:

(4) *hari khusire liLaa-ku* ***kunDhaa-i*** ***pakaa-il-aa***
Hari happily Lila-ACC embrace-CONJ drop-PAST-3SG
'Being happy, Hari embraced Lila.' [Hari felt so happy that, he embraced Lila immediately]

The light verb *paD* 'fall' indicates telicity in addition to suddenness or abruptness:

(5) *se sandhyaa na-heuNu* ***so-i*** ***paD-il-aa***
he evening NEG-happen sleep-CONJ fall-PAST-3SG
'He fell asleep before evening.'

The light verb *ne* 'take' is used for situations where the agent profits from the action (self-benefactive), adding a telic reading of the event:

(6) a. *se jibaa purbaru taara sabu kaama* ***kar-i*** ***ne-l-aa***
he going before his all work do-CONJ take-PAST-3SG
'He got all his work done before he left.'

b. *se raajaa-dwaaraa taara sabu kaama* ***kar-aa-i*** ***ne-l-aa***
he Raja-by his all work do-CAUS-CONJ take- PAST-3SG
'He got all his work done through Raja.'

The light verb *uTh* 'rise' entails telicity, denoting suddenness of the event. E.g.

(7) *mun kathaa-Taa kahi na-saaruNu se jorre* ***has-i*** ***uTh-il-aa***
I tale-CL say NEG-finish he loudly laugh-CONJ rise-PAST-3SG
'He started laughing loudly before I finished saying that.'

Usually such constructions, as in (4)–(7) are used/preferred with a V-v sequence than a single verb sequence.

In the case of infinitival, subjunctive and imperative constructions as in examples (8)–(10), respectively, the presence of a light verb denotes telicity of the event:

(8) a. *mun sei kaama-Taa* ***kar-ibaa-ku*** *chaah-u-ch-i*
I that work-CL do-INF-to want-PROG-AUX-1SG
'I want to do that work.' [I want to initiate that work]

b. *mun sei kaama-Taa* ***kar-i*** ***de-baa-ku*** *chaah-u-ch-i*
I that work-CL do-CONJ give-INF-to want-PROG-AUX-1SG
'I want to do that work.' [I want to complete that work]

(9) a. *se jadi naaga saapa* ***dhar-ant-i***, *taahele taanku dhar-ibaa*
he if cobra snake catch-SUBJN-3SG, then him catch-INF
paain di-aa-j-ib-a
for give-CAUS-go-FUT-3SG
'In case he catches cobra snake, then he will be given the offer to catch cobra.'

b. *se jadi naaga saapa* ***dhar-i*** ***di-ant-i***,
he if cobra snake catch-CONJ give-SUBJN-3SG,
taahele taanku dhar-ibaa pain puraskaara di-aa-j-ib-a
then him catch-INF for prize give-CAUS-go-FUT-3SG
'In case he catches cobra snake, then he will be awarded for catching cobra.'
['In case he caught a cobra snake, then he will be awarded for catching cobra.']

Note that the subordinate clauses cannot be exchanged in these two constructions.

(10) a. *sei kaama-Taa* ***kar-a*** *!*
that work-CL do-IMP
'Do that work!' ['initiate that work.']

b. *sei kaama-Taa* ***kar-i di-a*** *!*
that work-CL do-CONJ give-IMP
'Do that work!' ['Complete that work.']

In the above examples, the single verb sequence constructions as in (8a), (9a) and (10a) are atelic; while the presence of light verbs in the corresponding two verb sequence constructions as in (8b), (9b) and (10b), makes them telic.

4. Perfectivity vs. telicity

Often telicity is superficially similar to the perfective aspect, and one can find descriptions such as "roughly perfective/imperfective". However, does perfective aspectual morpheme denote telicity? Consider the following examples.

In certain cases in pluperfect, the presence of an aspectual morpheme does not suffice and a light verb becomes obligatory to denote telicity. E.g.

(11) a. *1995 banyaa-re Cuttack-ra sabu ghara*
1995 flood-in Cuttack's all house
bhaas-i jaa-i-th-il-aa/ ****bhaas-i-th-il-aa***
flow-CONJ go-PERF-AUX-PAST-3SG/ *flow-PERF-AUX-PAST-3SG
'In the flood of 1995, all the houses of Cuttack had flowed away.'

b. *1995 banyaa-re Cuttack-ra sabu ghara*
1995 flood-in Cuttack's all house
bhaas-i ga-l-aa/ ****bhaas-il-aa***
flow-CONJ go-PAST-3SG/ *flow-PAST-3SG
'In the flood of 1995, all the houses of Cuttack got flowed away.'

In (11), both the constructions are possible/acceptable in the presence of a light verb only: in (11a), in the presence of a perfective aspectual morpheme and in (11b), without any aspectual morpheme. Note that both the constructions are telic irrespective of the perfective aspectual morpheme. This indicates that telicity can just be attributed to the presence

of light verbs.[6] The single verb-variant (with or without a perfective morpheme) cannot be used in all constructions, and the light verb construction is another way for expressing telicity. This certainly distinguishes "perfectivity" from "telicity". Also, as we found in example (3b), because of the presence of a light verb, the co-occurrence of the light verb and the progressive morpheme denotes "repetition of the completed action" rather than the "action in progress". So, we can claim that, perfective aspect ≠ telicity. Aspectual markers like perfective, imperfective/progressive are grammatical phenomena, while (a)telicity refers to semantic features.

5. Completive aspectual feature of light verbs

In terms of temporal sequence of events, light verbs indicate a completive aspectual feature. The aspectual V2 position seems to be pre-programmed for that purpose. Temporal sequences of events are expressed by juxtaposition of clauses or VPs without any intervening conjunctions.

5.1 Light verbs and adverbials

Consider the following constructions (V-v sequences containing light verbs, and their corresponding single verb variant constructions) along with time adverbials:

(12) a. *mun klaas-re pahanch-ibaa purbaru semaane*
I class-in arrive-GER before they
pahanch-i jaa-i-th-il-e
arrive-CONJ go-PERF-AUX-PAST-3PL
'They had already arrived in the class before I arrived.'

b. **mun klaas-re pahanch-ibaa purbaru semaane*
I class-in arrive-GER before they
pahanch-i-th-il-e
arrive-PERF-AUX-PAST-3PL
'They had arrived in the class before I arrived.'

[6] Telicity can be attributed to the presence of light verbs, as illustrated by the grammatical constructions in (11), unless we propose that there is a null telic marker in light verb constructions. However, there is no real support for a null telic marker.

(13) a. **mun ghar-e pahanch-ibaa pare semaane*
I home-in arrive-GER after they
pahanch-i jaa-i-th-il-e
arrive-CONJ go-PERF-AUX-PAST-3PL
['They had already arrived after I arrived at home.'] / 'They arrived after I had arrived at home.'

b. *mun ghar-e pahanch-ibaa pare semaane* ***pahanch-i-th-il-e***
I home-in arrive-GER after they arrive-PERF-AUX-PAST-3PL
'They arrived after I had arrived at home.'

(14) a. *mun ghar-e pahanch-ibaa beLa-ku semaane*
I home-in arrive-GER time-PP they
pahanch-i jaa-u-th-il-e
arrive-CONJ go-PROG-AUX-PAST-3PL
['They used to arrive when I was arriving at home.'] / 'They arrived when I arrived/was arriving at home.'

b. *mun ghar-e pahanch-ibaa beLa-ku semaane* ***pahanch-u-th-il-e***
I home-in arrive-GER time-PP they arrive-PROG-AUX-PAST-3PL
'They arrived when I arrived/was arriving at home.'

In (12), the light verb *jaaithile* ascribes completive meaning, i.e., the action of their reaching the class is completed before the action of my doing the same. The temporally last action ('me reaching the class') cannot be expressed by a light verb construction (**mun klaas-re pahanch-i-j-ibaa)* denoting a completive action. Thus, only the temporally previous sequence of action is acceptable with the light verb denoting a completive action. The acceptability is reversed in (13) because of the nature of the adverb *pare* 'after'. In (14), because of the simultaneity nature of the time adverbial, the event represented by the light verb 'at the same time' does not have to follow any temporal sequence as both the events are done simultaneously. Thus, we can say that in the above examples, the light verb emphasizes the completion of event. Note that, light verbs interact with all tenses and aspects.

5.2 Light verbs and negation

As light verbs denote completion of events, the interaction of negation and light verbs does not go well. This is shown in the following examples:

(15) a. *mun taaku* ***de-l-i,*** *kintu se ne-l-aa-ni*
I him give-PAST-1SG but he take-PAST-3SG-NEG
'I gave him, but he didn't take.' [means, 'I offered him, but he didn't take.']

b. *mun taaku* ***de-i*** ***de-l-i,*** *(*kintu se ne-l-aa-ni)*
I him give-CONJ give-PAST-1SG (*but he take-PAST-3SG-NEG)
'I gave him, (*but he didn't take).' [means, the 'giving' act is completed.]

(16) a. *se chiThi-Taa* ***paDh-il-aa,*** *kintu paDh-i-paar-il-aa-naahin*
he letter- CL read-PAST-3SG but read-PERF-can-PAST-3SG-NEG
'He read the letter, but could not.' [means, 'He tried to read the letter, but could not read it.' More precisely, 'he started reading the letter but was not able to continue reading'.]

b. *se chiThi-Taa* ***paDh-i*** ***de-l-aa,***
he letter-CL read-CONJ give-PAST-3SG
*(*kintu paDh-i-paar-il-aa-naahin)*
(*but read-PERF-can-PAST-3SG-NEG)
'He read the letter, (*but could not).'

The ungrammaticality because of the conjoined clauses in (15b) and (16b) indicates that there is a completive meaning/ telicity associated with the light verb, which necessarily denotes the completion of the action/event and thus, does not allow the event to be shown 'undone', hence, not prone to negation. This supports our claim that light verbs have [+telic] feature and hence, indicate the telicity of the VP/sentence they occur in.

6. Conclusion

The study provides insight into the aspectual semantics of Odia. The semantic features of (a)telicity can be expressed by several linguistic means in Odia. Here, we have taken up the telic feature of light verbs, which has previously been neglected, as the progressive/perfective aspectual morphemes have reserved the main attention. The lexico-semantic properties of light verbs, which denote telicity in this language, function like a sort of aspectual feature. Aspectual markers like perfective, imperfective/progressive are grammatical phenomena, while (a)telicity refers to semantic features. The point we put forward in this paper is: the perfective morpheme and the light verbs are different subtypes of telic features.

References

Borik, Olga (2006) *Aspect and Reference Time*. Oxford: Oxford University Press.

Borik, Olga & Reinhart, Tanya (2004) Telicity and Perfectivity: Two Independent Systems. In László Hunyadi; György Rákosi & Enikő Tóth (eds.) *Proceedings of LOLA 8 (Symposium on Logic and Language)*, pp. 13–34. Debrecen, Hungary.

Comrie, Bernard (1976) *Aspect*. Cambridge: Cambridge University Press.

de Swart, Henriette (1998) *Introduction to Natural Language Semantics*. Stanford: CSLI Publications.

Depraetere, Ilse (1995) On the necessity of distinguishing between (un)boundedness and (a)telicity. *Linguistics and Philosophy*. 18: 1–19.

Giorgi, Alessandra & Pianesi, Fabio (2004) On the Speaker's and the Subject's Temporal Representation: The Case of the Italian Imperfect. In Jacqueline Guéron & Jacqueline Lecarme (eds.), *The Syntax of Time*, pp. 259–298. Cambridge: MIT Press.

Kamp, Hans & Reyle, Uwe (1993) *From Discourse to Logic*. Dordrecht: Kluwer.

Klein, Wolfgang (1995) A time-relational analysis of Russian aspect. *Language* 71: 669–695.

Kratzer, Angelika (2004) Telicity and the meaning of objective case. In Jacqueline Guéron and Jacqueline Lecarme (eds.) *The Syntax of Time*, pp. 389–424. Cambridge: MIT Press.

Lemmens, Maarten & Sahoo, Kalyanamalini (in progress). Something's gotta *go*, something's gotta *give*: Oriya light verb constructions. University of Lille 3, France.

Mohapatra, Pandit Narayana & Das, Sridhara (1962) *Sarbasaara byaakaraNa* [The Best Grammar]. Cuttack: M/s New Student's Store, Binod Bihari.

Nayak, Rath (1987) Non-finite Clauses in Oriya. Doctoral dissertation, CIEFL, Hyderabad.

Pradhan, Sarat Chandra; Dash, Gaganendra Nath; Mohanty, Panchanan; Sadangi, Krusnachandra & Srichandansingh, Subhendra Mohan (1995) *oDiaa byaakaraNa* [Odia Grammar]. Cuttack: Board of Secondary Education, Orissa.

Sahoo, Kalyanamalini (2001) *Oriya Verb Morphology and Complex Verb Constructions*. Doctoral dissertation, Norwegian University of Science & Technology, Trondheim, Norway.

Verkuyl, Henk (1993) *A Theory of Aspectuality: The Interaction between Temporal and Atemporal Structure*. Cambridge: Cambridge University Press.

Contact information:

Kalyanamalini Sahoo
Department of Linguistics & Contemporary English
English & Foreign Languages University, Hyderabad
Andhra Pradesh 500007
India
e-mail: kalyanirs(at)gmail(dot)com

La Fauci, Nunzio & Mirto, Ignazio M. (2010 [2003]) *Fare. Elementi di sintassi* [Faire. Éléments de syntaxe]. Pisa : Edizioni ETS. Pp. 108.

Compte rendu de Samuel Bidaud

Nous profitons de la réédition de ce petit livre initialement sorti en 2003 et réimprimé cette année pour en faire à la fois le compte rendu et développer quelques idées sur le sujet traité par les deux auteurs.

Nunzio La Fauci et Ignazio M. Mirto se proposent d'analyser le verbe *fare* 'faire' dans une perspective essentiellement syntaxique et fonctionnelle. C'est de ce point de vue qu'ils étudient quatre constructions du verbe *fare* en italien : la construction causative, du type *Ugo fa sorridere Pia* 'Ugo fait sourire Pia', la construction où *fare* sert de support à un nom, comme dans *Ugo fa una serenata a Pia* 'Ugo fait une sérénade à Pia', la construction où *fare* est accompagné d'un nom de métier, comme dans *Ugo fa l'avvocato* 'Ugo fait l'avocat (fr. Ugo est avocat)', et celle où *fare* se combine avec un nom et signifie une représentation scénique, comme dans *Ugo fa il moro di Venezia* 'Ugo fait le Maure de Venise (fr. Ugo interprète le Maure de Venise)' (2010 : 8).

Les auteurs proposent une analyse de chacune de ces constructions d'un point de vue syntaxique et fonctionnel, au sens où ces constructions ne correspondent pas à un nombre d'arguments et de propriétés identiques, ce qu'ils prouvent parfaitement et avec rigueur par toutes sortes de tests de combinaisons. On admirera au passage l'élégance du raisonnement p. 78 et suivantes à propos de la construction *Ugo fa l'avvocato*, *Lia fa la ballerina russa*, etc. (littéralement 'Ugo fait l'avocat', 'Lia fait la danseuse russe').

Mais là où nous sommes en désaccord avec les auteurs, c'est dans la conclusion qu'ils tirent : *fare* est un mot dont le signifié serait uniquement fonction de son environnement syntaxique et n'aurait pas de sens à proprement parler (c'est également le cas d'après eux des auxiliaires que sont *essere* 'être' et *avere* 'avoir') : la preuve en est, disent-ils en conclusion à leur ouvrage, que *fare* n'a pas moins de 63 sens différents et ils en concluent que « quando qualcosa pare significare tanto, forse troppo, […] è proprio perché non significa nulla » ('lorsque quelque chose semble signifier tellement, peut-être trop, […] c'est justement parce qu'il ne signifie rien'). Et les auteurs précisent que « (è) la sintassi, in altre parole,

interamente responsabile di *fare* » ('C'est la syntaxe, en d'autres termes, qui est entièrement responsable de *fare*').

Or, il nous semble, et ce n'est là en aucun cas un reproche, qu'il conviendrait d'ajouter à une perspective uniquement syntactico-fonctionnelle et donc en grande partie pragmatique une perspective qui permette de voir que l'ensemble des effets de sens étudiés par les auteurs peuvent en réalité se ramener à quelques valeurs d'où ils dérivent.

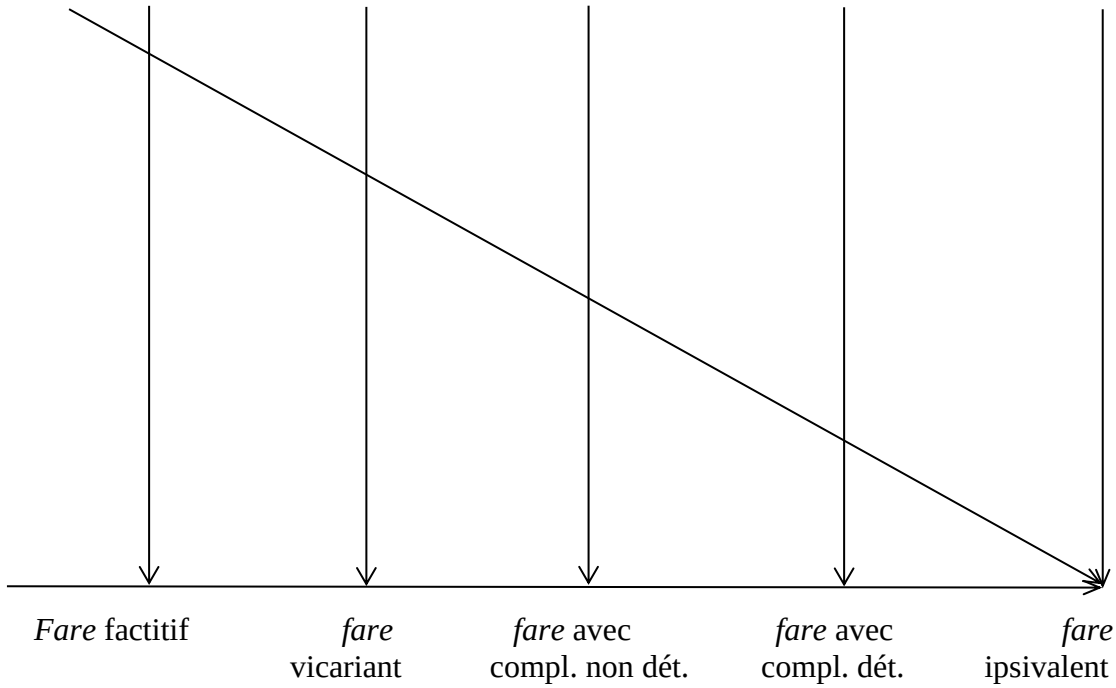
C'est là que nous proposons de faire intervenir une autre école qui ne remettrait pas en cause leur travail mais qui permettrait de faire de la syntaxe fonctionnelle un élément d'explication situé au niveau du discours : nous pensons à la psychomécanique du langage de Gustave Guillaume, pour qui la notion de mot vide est proprement impensable.

Que peut-on tirer de la psychomécanique ?

On peut réduire le signifié du verbe *fare* à ce que la psychomécanique appelle un cinétisme sur lequel plusieurs saisies vont s'inscrire. Mais ces saisies sont limitées et elles donnent lieu à des signifiés d'effet différents (les 63 sens du Dictionnaire) : en effet, on a bien un signifié de langue qui correspond à une idéogénèse et qui reconstitue le mouvement de pensée du verbe *fare*. C'est cette idéogénèse qu'a reconstruite, pour l'ancien français, Thierry Ponchon (1994), avec de nombreuses saisies qu'il conviendrait de discuter, ou ce qu'a fait Samir Bajrić d'une manière plus succincte (2008). Le fait est que les deux auteurs reconstituent en réalité, dans une perspective guillaumienne, le mouvement de pensée de *faire* qui conduit à l'idée de fabriquer et sur lequel plusieurs saisies peuvent intervenir. Mais le signifié n'en reste pas moins un, quoique cinétique (c'est ce que la psychomécanique nomme le *signifié de puissance*) : on a un nombre réduit de saisies sur l'idéogénèse qui donnent lieu à des effets de sens multiples au niveau du discours, effets de sens qui sont justement (et c'est là que nous rejoignons les auteurs), l'effet de la syntaxe fonctionnelle.

En d'autres termes, nous ne nions pas du tout la portée des travaux des auteurs, loin de là (les preuves alléguées sont d'ailleurs fort convaincantes), mais nous voudrions les compléter avec une idéogénèse dont le cinétisme est constitué par plusieurs valeurs qui trouvent justement leur effet de sens en syntaxe. Il est d'ailleurs possible de situer sur une idéogénèse les quatre constructions observées par les auteurs, par exemple sur la suivante, qui n'est certes que provisoire dans la mesure où nous menons actuellement des recherches sur le sujet et où Samir Bajrić doit prochainement faire paraître un livre où il donne des résultats qui seront sans aucun doute plus convaincants que ceux que nous donnons là :

Figure 1. Idéogénèse du verbe *fare*



L'idéogénèse du verbe *fare* conduit à l'idée de fabrication et peut être décrite comme allant d'un *fare* factitif où le sujet ne participe pas à l'action mais la commande à une idée de fabrication pure. Donnons quelques exemples en italien :

Fare factitif :

(1) *Ha fatto fare un vestito al sarto.*
'Il a fait faire un costume au tailleur'

Fare vicariant :

(2) *Poi si fece portare la corrispondenza in arrivo, la lesse scambiando qualche parola col capo ufficio e indicando la soluzione delle pratiche normali come il vice direttore aveva fatto con lui.*
'Puis il se fit porter la correspondance de la réception, il la lut en échangeant quelques paroles avec le chef de bureau et en indiquant la solution des dossiers communs comme le vice-directeur l'avait fait avec lui.' (Radice, *Un matrimonio mancato*, p. 45).

Fare avec complément non déterminé :

(3) *Giunto davanti allo scrittoio, il vice direttore gli fece cenno di sedere.*
'Juste devant le bureau, le vice-directeur lui fit signe de s'asseoir.' (Radice, *Un matrimonio mancato*, p. 42).

Fare avec complément déterminé :

(4) *Ugo fa una serenata a Pia.*
'Ugo fait une sérénade à Pia'

Fare ipsivalent :

(5) *Il SERVITORE con una cioccolata.*
'LE SERVITEUR avec un chocolat.
Cavaliere. (...) Fanne subito un'altra. (al Servitore)
'Le Cavalier (au Serviteur) : (…) Fais-en tout de suite un autre.' (Goldoni, *La Locandiera*, p. 795).

L'idéogénèse du verbe *fare* est porteuse de plusieurs sèmes en réalité : elle correspond à la fois à un sème d'activité du sujet, mais également, qui dit activité, dit participation du sujet à l'événement. C'est-à-dire que c'est à partir du moment où les deux sèmes sont réunis que l'idée d'activité se trouve la plus avancée. Mais si l'un des sèmes manque, alors nous sommes dans une saisie plus précoce. C'est le cas de l'emploi factitif, où le sujet ne participe pas à l'activité du tout, et qu'il est possible de situer en haut de l'idéogénèse, puisqu'elle est plus subduite que la vicariance. Dans : *Il a fait faire un costume au tailleur*, le sujet n'a en réalité rien fait d'autre que demander à quelqu'un d'autre d'effectuer l'activité à sa place, autrement dit, il n'y a pas participé, sinon en donnant un ordre, en effectuant une demande, ou ce que l'on voudra. On voit la différence avec le cas de la vicariance, où le sujet est à la fois participant et actif, même si l'activité est parfois minime : il y a bien participation et activité du sujet, alors qu'il n'en va pas de même avec le verbe *fare* factitif, où le sujet délègue une activité et voit réduire sa participation à un sémantisme d'activité minime, celui de déléguer une action. Nous sommes donc bien en face d'une saisie plus précoce. *Fare* avec complément non déterminé est quant à lui moins avancé, du fait de la soudure du complément à son verbe (type *fare cenno*), que le verbe *fare* avec un complément déterminé qui par l'autonomie qu'il prend redonne à *fare* un sens plus plein, qui ne se réalise complètement que dans le signifié plénier.

Or, les constructions si bien décrites par les auteurs sont des *effets de sens* de certaines saisies. *Ugo fa sorridere Pia* est un effet de sens de *fare* factitif : en effet Ugo est cause du sourire de Pia mais ne sourit pas lui-même, ce qui n'est qu'un des effets de sens conditionnés par le factitif ; et les autres constructions sont toutes des effets de sens de *fare* avec complément déterminé, qui ont toutefois chacune leur spécificité mais qui toutes dérivent de la même saisie : à chaque fois un complément nominal déterminé est associé à un *fare* qui a perdu une partie de son sens plein alors que les noms déterminés lui apportent un complément de matière prédicatif nécessaire pour former un entier de discours (voir Moignet, 1960 : 32).

C'est en tout cas un livre d'un grand intérêt qu'ont écrit Nunzio La Fauci et Ignazio M. Mirto, un livre qui complète la perspective psychomécanique et que la psychomécanique complète également. Aussi nous nous permettons d'en recommander la lecture à tous ceux qui travaillent sur le verbe *faire* et à tous ceux pour qui ce verbe « caméléon » suscite la curiosité.

Références

Bajrić, Samir (2008) Le verbe faire en français contemporain : syntaxe et sémantique. In *Suvremena Lingvistika* 66 (2) : 143–197.

Goldoni, Carlo (1945) *La locandiera*, in Commedie di Carlo Goldoni (4). Milano: Mondadori.

Moignet, Gérard (1961) L'adverbe dans la locution verbale. Etude de psycho-systématique française. *Cahiers de psychomécanique du langage.* Québec : Presses de l'Université Laval.

Ponchon, Thierry (1994) *Sémantique lexicale et sémantique grammaticale : le verbe faire en français médiéval*. Genève : Droz.

Radice, Raul (1947) *Un matrimonio mancato*. Milano: Mondadori.

Coordonnées :

Samuel Bidaud
1, rue Ernest Petit
21000 Dijon
France
e-mail: samuel(dot)bidaud(at)aliceadsl(dot)fr

Response to Maria Kela's Response to the Review of Idström, Anna & Sosa, Sachiko (eds.) (2009) *Kielissä kulttuurien ääni* [The voice of cultures is in the languages], reviewed by Kalle Korhonen in *SKY Journal of Linguistics*, vol. 23 (2010), pp. 351–355. Kela's Response published in *SKY Journal of Linguistics*, vol. 24 (2011), pp. 199–200.

Response written by Kalle Korhonen

Reviewers of collective volumes usually face two options: they must either focus on a small number of articles that they find the most interesting, or they can try to say something on each article. In my review of the important collective volume *Kielissä kulttuurien ääni*, published in vol. 23 of this Journal, I chose the second option because I did not want to omit any contribution. This is a way of showing appreciation of every single contribution and ensuring that the interested readers will be guided to the original articles. My review was highly favourable.

However, a passage of my review has prompted Maria Kela to make two critical points in vol. 24, p. 199. I quote my original passage in full:

> The Bible belongs to the most widely translated works in the world, and Maria Kela's article discusses how biblical metaphors that originated in Hebrew have been translated into different languages. Bible translations have in the course of time created several strange expressions or misunderstandings (e.g., the "horns of Moses"). Kela focuses on the metaphors "God's son", "Son of Man" and "the right hand" and explains in a lucid manner why the phrase "Ihmisen Poika" ("the son of a human being", Finnish for "Son of Man") has disappeared from the most recent Finnish Bible translation. The element "son" was in the original Hebrew undergoing grammaticalization and the phrase was a dead metaphor with the meaning "a human being". (*SKY Journal*, vol. 23 (2010), 352)

Kela correctly points out that I used an imprecise expression when writing that "the phrase "Ihmisen Poika" ("the son of a human being", Finnish for "Son of Man") has disappeared from the most recent Finnish Bible translation". Kela is right: the expression is still used several times in the newest translation. However, it has disappeared from the translation of the passage Kela focuses on on p. 91 of her article, namely Daniel 7:13. I should have written "from certain passages", not "from the … translation".

According to Kela, I claim that she "took the metaphor 'horns of Moses' as an example of a difficult biblical metaphor." However, I did not

make such a claim. The point was to illustrate a well known mistranslation from the context of the Bible, which I thought could be familiar to readers.

Contact information:

Kalle Korhonen
Department of World Cultures
University of Helsinki
Finland
e-mail: Kalle(dot)Korhonen(at)helsinki(dot)fi

Rice, Curt and Sylvia Blaho (eds.) (2009) *Modeling Ungrammaticality in Optimality Theory*, Advances in Optimality Theory. London: Equinox Press. Pp. 297.

Reviewed by Michael T. Putnam

Optimality Theory (hereafter OT), much like the Minimalist Program (MP), is, as argued by Broekhuis and Vogel (2010), highly programmatic in nature; serving as a cover term for a host of a sub-varieties of OT that all hold to the central concept of violable constraint interaction and their harmonic ranking which determines whether a given output is grammatical in a particular language (thus sufficing the requirements of descriptive adequacy). OT's focus on output, as pointed out by Rice and Blaho, "emphasizes a central architectural property of the theory, namely that every input is associated with some output" (p. 2). The notion of *ineffability* (coined by Pesetsky (1997) and more common in OT-syntax) or *absolute ungrammaticality* (more common in OT-morphophonology circles) presents a significant challenge for OT-grammars for the following reasons: First, as mentioned above, is OT's focus on output (in an input-output correspondence system), which forces each input to have some output. The notion of ineffability, i.e., the fact that sometimes no winner can be derived from a given competition, has proven to be a confounding problem for linguists who use OT. Secondly, and closely related to the first point, is the concept of constraint violability – a core component of OT. Due to the fact that there are no "perfect" candidates in the generated output candidate set (CON), i.e., any winner can violate an infinite number of lowly ranked constraints, it remains unclear to some (see e.g. especially Legendre's contribution in this volume) on the surface how and why a null parse candidate (commonly marked as ⊙ in OT-tableaux) solves this puzzle (a puzzle, I might point out, that was recognized as early as the seminal work on OT by Prince and Smolensky (1993/2004)). In sum, OT does not generate perfect candidates, it generates optimal, "good enough for the situation at hand," structures, which marks the fact that sometimes no possible output is given for a particular input as a serious challenge to the ontology of the framework. The papers in the volume take on the task of modeling ungrammaticality in OT, presenting novel theoretical solutions to this problem and often forcing a re-evaluation of core, long-assumed

axioms of the theory. The contributions of the volume are grouped into three sub-sections: *Architecture, Paradigms,* and *Ineffability in Syntax*.

The section on *Architecture* begins with Matthew Wolf and John McCarthy's contribution entitled *Less than zero: correspondence and the null output*. The purpose of Wolf and McCarthy's contribution is to rationalize the properties of the null parse (or null output as they refer to it). The central guiding question to their paper is: How is it possible for this candidate to violate only the constraint MPARSE while it (apparently) satisfies all other faithfulness and markedness constraints in a grammar? Their answer to this question rests in a revision of the Correspondence Theory (McCarthy and Prince 1995, 1999), where strings rather than segments are the formal objects that stand in correspondence. According to this revision, well-behaved unfaithful mappings do not alter ℜ's (ℜ = relation between segments in an input string *i* of an output string *o*) as a total bijective function. In their own words, "candidates with a less orderly ℜ violate MPARSE; among these candidates there is one that harmonically bounds all others, the null output ⊙" (p. 60). In *Dutch diminutives and the question mark*, Marc van Oostendorp breaks away from a contemporary view of Correspondence Theory in favor of Containment Theory, which embodies a different theoretical understanding of faithfulness. The view of faithfulness in Containment Theory is a return to the "original" understanding of faithfulness, a version of OT that was strictly monostratal, with the constraints targeting only the output (including faithfulness constraints). The remainder of Oostendorp's contribution sketches out how Containment Theory is a step forward in the discussion of ineffability in OT. This section concludes with Ohran Orgun and Ronald Sprouse's contribution entitled *Hard constraints in Optimality Theory*, where the authors present the CONTROL approach to morphological gaps. According to this approach, constraints can belong to the more conventional EVAL component or CONTROL. The harmonic constraint ranking present in EVAL functions as one would expect in all variants of OT, proposing an optimal candidate. With the addition of CONTROL, the candidate must also satisfy all of the constraints in CONTROL, with failure to meet this requirements resulting in an ungrammatical output. In contrast to EVAL, CONTROL only evaluates one candidate (the optimal/winning candidate resulting from EVAL), this preventing any sort of repair mechanism or alternative candidate possibilities. In this system, CONTROL determines grammaticality, a component of the grammar that lacks any sort of grammatical capacity.

The second set of papers devoted to the topic of Paradigms seeks to provide insight into modeling phonological, morphological, and lexical gaps from an OT-perspective. Adam Albright's *A Lexical and morphological condition of paradigm gaps* focuses on gaps that only affect certain words, while other, seemingly parallel words surface as expected. In this paper, Albright attempts to predict which parts of a paradigm, and in particular which lexical items, may be affected by paradigm gaps. Albright concludes that "it is hypothesized that gaps affect only those forms that are computed with reference to another base form in the paradigm, and occur only in cases where the mapping between the base and the derived form requires an inference over small amounts of possibly conflicting data" (p. 160). Outi Bat-El explores data in Hebrew, a language that often avoids surface forms where a string of suffixes has identical consonants. In her contribution entitled *A gap in the feminine paradigm of Hebrew: a consequence of identity avoidance in the suffix domain*, Bat-El presents various strategies employed in the grammar of Hebrew to amend inputs with identical consonants in the suffix domain, with one of the possible solutions being the null output candidate. Finally, Peter Rebrus and Miklós Törkenczy investigate the notion of defectiveness (i.e., the occurrence of paradigmatic gaps) from a bifurcated perspective, looking separately at *phonologically motivated gaps* in juxtaposition to *arbitrary gaps* and distinguishing *overt* vs. *covert defectiveness*. Rebrus and Törkenczy's contribution illustrates that not all instances of defectiveness should be analyzed at the same level of grammar, for example forcing the input of certain paradigm gaps to allomorphs rather than morphemes in the case of their Hungarian data.

The third and final subset of papers addresses the topic of *Ineffability in Syntax*. Gélardine Legendre takes a critical look at the notion of null parse in her paper, *The neutralization approach to ineffability*. Legendre provides arguments against the null parse candidate in OT-syntax, opting rather for a neutralization approach to ineffability, "whereby different inputs (interpretations) neutralize to and the same optimal output because specific input features ([wh]; operator scope) may be underparsed. The optimal candidate is close to the input but not identical: "Sometimes it's best to say something else"" (p. 240). Ralf Vogel's *Wh-islands: a view from Correspondence Theory* closes out this section and the volume. Vogel focuses on the notion of *expressive ineffability* in the syntax of languages and language families, i.e., the notion that a certain meaning cannot be expressed by using a particular structure S in some language. Vogel

employs a Correspondence-based version of OT to take a closer look at two particular cases of syntactic ineffability: 1.) the impossibility of a particular structure S in some language(s), and 2.) the impossibility to express a particular meaning M by using a particular structure S in some language(s). Vogel concludes that instances of wh-islandhood result in an unparsimonious correspondence between conflicting requirements (formalized as OT-constraints) enforced at the syntax-semantics interface.

This collection of research on the notions of ineffability and absolute ungrammaticality is the first step at filling a long-standing gap (pun intended) in the literature pertaining to our understanding of modeling ungrammaticality in OT. The various contrastive approaches to model ungrammaticality presented in this collection of paper, e.g., null parse, underspecification of the input, the addition of a post-EVAL level of constraints (CONTROL), etc., provides evidence that this volume is but a first step towards understanding how best to model ungrammaticality from an OT-perspective. I say this not as a mark against this book, but rather to simply point out future work that needs to be done in this area. As work progresses on this topic, this volume will undoubtedly serve as an invaluable starting point for such research ventures.

References

Broekhuis, Hans & Vogel, Ralf (2010) Crash-proof syntax and filters. In Michael T. Putnam (ed.), *Exploring crash-proof grammars*, pp. 248–265. Amsterdam: John Benjamins.

McCarthy, John & Prince, Alan (1995) Faithfulness and reduplicative identity. In Jill N. Beckman, Laura Walsh Dickey & Suzanne Urbanczyk (eds.), *University of Massachusetts Occasional Papers in Linguistics* 18, pp. 249–384. Amherst, MA: GLSA Publications (ROA 103).

—— (1999) Faithfulness and identity in prosodic morphology. In René Kager, Harry van der Hulst & W. Zonnenveld (eds.), *The Prosody-Morphology Interface*, pp. 218–309. Cambridge: CUP.

Pesetsky, David (1997) Optimality theory and syntax: movement and pronunciation. In Diana Archangeli & D. Terence Langedoen (eds.), *Optimality Theory, an overview*, pp. 134–170. London: Wiley-Blackwell.

Prince, Alan & Smolensky, Paul (1993) *Optimality Theory. Constraint interaction in generative grammar, Technical Report #2*. Piscataway, NJ: Rutgers University Center for Cognitive Science. ROA 537; (Published in 2004 by Blackwell).

Contact information:

Prof. Michael T. Putnam
Pennsylvania State University
Department of Germanic & Slavic Languages & Literatures
427 Burrowes Building
University Park, PA 16802
U.S.A.
e-mail: mtp12(at)psu(dot)edu