UNIVERSITY OF
LATVIA

26–28 June 2025

GRAMMAR AND CORPORA

10th international conference

UNIVERSITY OF LATVIA

FACULTY OF HUMANITIES
DEPARTMENT OF LATVIAN AND BALTIC STUDIES

26–28 June 2025

10th international conference

GRAMMAR AND CORPORA

BOOK OF ABSTRACTS

Riga: University of Latvia Press, 2025

SCIENTIFIC COMMITTEE

Ilze Auziņa (University of Latvia, Riga)
Sanita Bērziņa-Reinsone (University of Latvia, Riga)
Daiki Horiguchi (University of Kyoto)
Andra Kalnača (University of Latvia, Riga, conference director)
Andres Karjus (Tallinn University)
Torsten Leuschner (Ghent University)
Ilze Lokmane (University of Latvia, Riga)
Helle Metslang (University of Tartu)
Nicole Nau (Adam Mickiewicz University in Poznań)
Miina Norvik (University of Tartu)
Jurgis Pakerys (Vilnius University)
Hélène de Penanros (Institut national des langues et civilisations orientales, Paris)
Erika Rimkutė (Vytautas Magnus University, Kaunas)
Inguna Skadiņa (University of Latvia, Riga)
Michal Škrabal (Charles University, Prague)
Beata Trawiński (Leibniz-Institut für Deutsche Sprache, Mannheim)
Jurgita Vaičenonienė (Vytautas Magnus University, Kaunas)
Piotr Wyroślak (Adam Mickiewicz University in Poznań)

ORGANIZING COMMITTEE

Ieva Auziņa (University of Latvia, Riga)
Vanesa Balmane (University of Latvia, Riga)
Anita Butāne (University of Latvia, Riga)
Milan Hoplíček (University of Latvia, Riga)
Laura Paula Jansone (University of Latvia, Riga)
Andra Kalnača (University of Latvia, Riga)
Kristīne Levāne-Petrova (University of Latvia, Riga)
Ilze Lokmane (University of Latvia, Riga)
Paula Miķelsone (University of Latvia, Riga)
Oskars Otomers (University of Latvia, Riga)
Tatjana Pakalne (University of Latvia, Riga)
Inta Urbanoviča (University of Latvia, Riga)
Evelīna Zilgalve (University of Latvia, Riga)

The conference is organized under the auspices of the University of Latvia and the Latvian Council of Science project "Database of Latvian Morphemes and Derivational Models (DLMDM)" (No. lzp-2022/1-0013)

UNIVERSITY OF LATVIA

FLPP
FUNDAMENTAL AND APPLIED RESEARCH PROJECTS

Compiler – Andra Kalnača
Editors – Andra Kalnača, Paula Ozola, Milan Hoplíček
Proof reader – Andra Damberga
Layout designer – Ineta Priga

https://doi.org/10.22364/gcc.10.ba
ISBN 978-9934-36-400-6
ISBN 978-9934-36-401-3 (PDF)

CONTENTS

1. PLENARY TALK ABSTRACTS

2. ALL ABSTRACTS

1. PLENARY TALK ABSTRACTS

ON A COMMON BALTIC TEXTUAL CORPUS (CBTC) AND RELATED GRAMMAR ISSUES: A SHOWCASE FROM THE OLD LATVIAN “FLOOD PRAYER”

Pietro U. Dini

Pisa University

The translation of Martin Luther’s Small Catechism (also Enchiridion) marked the beginning of the three Baltic written traditions – those of Latvian, Lithuanian, and Old Prussian. Within the broader corpus of Baltic-language catechisms, at least two or three denominational sub-genres – Lutheran, Catholic, and Calvinist – can be identified. Among these, the Lutheran Small Catechism forms a distinct corpus that connects the older stages of Lithuanian (Willent 1579) and Latvian (Rivius et al. 1586) with the primary linguistic monument of Old Prussian (Will 1561).

Considering these circumstances, I propose the following initiative: (a) the creation of a Common Baltic Textual Corpus (CBTC) based on the Lutheran Small Catechisms of the three traditions; (b) analysis of their main structural features; and (c) the identification of effective methods for studying them in parallel (cf. Dini 2014). To showcase the value of this proposal, I present examples from the “Flood Prayer” (cf. Dini 2022), in which CBTC data reveal the presence of collective numeral compounds – a grammatical category foreign to the Baltic languages but episodically attested in early texts (cf. Zwoliński 1954). I explore and describe its fate, with special attention to Old Latvian.

References

Dini, Pietro U. 2014. *Ins undevdſche gebracht. Sprachgebrauch und Übersetzungsverfahren im altpreußischen Kleinen Katechismus*. Akademie der Wissenschaft zu Göttingen. Berlin: De Gruyter.

Dini, Pietro U. 2022. Zu apr. Ench. 119.13 *ſubban Aſman* (und alit. *patį aſchmą*) als Kollektiv–zahlwort. A Festschrift for Bonifacas Stundžia on the occasion of his 70th birthday. Vilnius: Vilnius University Press, 117–144.

Rivius et al. 1586. ENCHIRIDION | Der kleine Ca= | techismus: Oder Chriſt= | liche ʒucht fu^{e}r die gemeinen Pfar= | herr vnd Prediger auch Hausueter etc. | Durch | D. Martin. Luther. | Nun aber aus dem Deud= | ſchen ins vndeudſche gebracht / vnd | von wort ʒu wort / wie es von D. | M. Luthero geſetʒet / gefasſ= | ſet worden. | Gedruckt ʒu Ko^{e}nigsperg bey | George Oſterbergern | Anno M.D. LXXXVI.

Will Abel 1561. ENCHIRIDION | Der Kleine | Catechiſmus | Doctor Martin Lu= | thers / Teutſch vnd Preuſſisch. | Gedruckt ʒu Ko^{e}nigsperg in Preuſſen | durch Johann Daubman. | M. D. LXI.

Willent 1579. ENCHIRIDION | Catechiſmas | masas / dæl paſpalitu | Plebonu ir Koʒnadiju / | Wokiſchku lieſsuwiu para= | ſchits per Daktara Mar= | tina Luthera. | O iſch Wokiſchka lieſsuwia ant | Lietuwiſchka pilnai ir wiernai pergul= | ditas / per Baltramieju Willentha | Plebona Karalaucʒuie ant | Schteindama. | Iſchſpauſtas Karalau= | cʒui per Iurgi Oſterber= | gera / Metu Diewa M. D. LXXIX.

Zwoliński, Przemysław. 1954. *Liczebniki zespołowe typu „samotrzeć" w języku polskim na tle słowianskim i indoeuropejskim*, Prace językoznawcze: Wrocław.

SYSTEMATICALLY LEVERAGING LARGE LANGUAGE MODELS FOR LINGUISTIC ANNOTATION AND ANALYSIS

Andres Karjus

Tallinn University; Estonian Business School

Keywords: LLM, AI, annotation, mixed methods, quantitizing designs

The increasing capacities of instructible large language models (LLMs) present an unprecedented opportunity to scale up corpus-based analysis in linguistics and grammar research. Of particular interest for corpus linguistics is their use as zero-shot classifiers and inference engines in annotation tasks. While supervised data classification and annotation has long been possible, the necessity to train or fine-tune models on large labelled sets in supervised learning have limited its application. Generative models enable easy "distillation" of linguistic expertise into model inputs (prompts). This talk discusses recent research on applying LLMs to corpus-based inquiries, and argues for formulating such research pipelines as quantitizing designs, where LLMs can be plugged in as (co-)annotators. LLM outputs inevitably contain noise and error (as does human annotation), but these uncertainties can be incorporated into downstream statistical modelling. A bootstrapping framework is introduced to estimate error rates, which in turn supports integration into quantitizing designs. Such an approach allows linguists to leverage the scale and flexibility of LLMs while ensuring replicability, transparency, and theoretical relevance in corpus-driven grammatical research.

CONCESSIVE CONDITIONALS IN (NOT ONLY) GERMAN: GRAMMAR, CORPORA AND THE EMERGENCE OF CONSTRUCTIONS

Torsten Leuschner & Flor Vander Haegen

Ghent University and Queen Mary University of London; & Research Foundation – Flanders and Ghent University

Keywords: concessive conditionals, construction family, constructionalization, constructional change, quantification

An interesting challenge to both corpus-linguistic methodology and grammatical theory are emergent constructions on the discourse/syntax borderline. We will address this challenge through so-called "concessive conditionals" (König 1986; Haspelmath, König 1998; Leuschner 2006), as exemplified below from the German DeReKo corpus:

(1) ***Selbst wenn*** *sich Hunderte von Radfahrern auf den Fähren drängen, bleiben die Mitarbeiter freundlich und hilfsbereit [...].* (Rhein-Zeitung, 02/08/2019)
'Even if the ferries are crowded with hundreds of cyclists, the employees stay friendly and helpful.'

(2) ***Ob*** *die Männer von einer unglücklichen Liebe sangen* ***oder*** *vom Wasser, das am Morgen gestorben ist, dem Publikum ging es kalt den Rücken hinunter.* (St. Galler Tagblatt, 16.12.1997)
'Whether the men were singing about an unrequited love or about water, the public had shivers going down their spies.'

(3) ***Was immer*** *die Regierung macht, sie schafft sich Feinde.* (Salzburger Nachrichten, 24.11.1998)
'Whatever the government does, it makes enemies.'

Concessive conditionals form a family of constructions (Leuschner 2020, 2023) because of their shared character as quantified conditionals on the overlap with concessivity (König 1986). The protasis evokes an exhaustive set of antecedent propositions {p_1, p_2, ..., p_n} which conceivably could, but in fact do not, have a bearing on the consequent *q*. Since there are different ways of quantifying the set, concessive conditionals can vary considerably in terms of surface form. Whereas type (1)a. is based on a prototypical *wenn*/'if'-conditional preceded by a focus particle, the protasis in (1)b. and (1)c. is reminiscent of an embedded interrogative or free relative clause, which however functions as an adjunct rather than an argument and remains on the margins of the apodosis both prosodically and syntactically. (In English, which lacks German verb-second sentence structure, the protasis still tends to be prosodically marginal.) The set of antecedent propositions is evoked by the interrogative-like protasis, whose semantics are employed for quantificational purposes here, and conditionality is present implicitly as part of the constructional meaning of protasis and apodosis together.

The formal diversity of concessive conditionals in German (see Breindl, Volodina, Waßner 2014, 963–1010 for a survey) and indeed in many other European languages including English is one reason why their functional similarity is often overlooked in the literature. Another reason is the fact that many formal varieties of concessive conditionals show features of emergent grammar (König 1992; Leuschner 2006) and of constructional gradience, making it difficult to distinguish them from related constructions and/or proto-syntactic configurations in discourse. Citing recent research (Bossuyt, De Cuypere, Leuschner 2018; Vander Haegen, Bossuyt, Leuschner 2022; Vander Haegen 2023), we will highlight empirical evidence both of incipient constructionalization and of post-constructionalization changes (cf. Traugott, Trousdale 2013) in some types of concessive conditionals in German, followed by theoretical conclusions from the perspective of usage-based Construction Grammar. At the end of our talk, we will show how concessive conditionals in German (and, *inter alia*, English) fit into patterns of typological variation (Haspelmath, König 1998; Bossuyt 2023), explaining why their surface forms in German and related languages are subject to such dynamic, but also relatively inconspicuous changes.

References

Bossuyt, Tom. 2023. Concessive conditionals beyond Europe: A typological survey. *Studies in Language*. 47(1), 1–31. https://doi.org/10.1075/sl.20068.bos

Bossuyt, Tom, De Cuypere, Ludovic, Leuschner, Torsten. 2018. Emergence phenomena in *W immer/auch*-subordinators. In Fuß, Eric, Konopka, Marek, Trawiński, Beata, Waßner, Ulrich H. (eds.). *Grammar and Corpora 2016*. Heidelberg: Heidelberg University Publishing, 97–120.

Breindl, Eva, Volodina, Anna, Waßner, Ulrich H. 2014. *Handbuch der deutschen Konnektoren 2. Semantik der deutschen Satzverknüpfer*. Berlin, Boston: De Gruyter.

Haspelmath, Martin, König, Ekkehard. 1998. Concessive conditionals in the languages of Europe. In van der Auwera, Johan (ed.). *Adverbial Constructions in the Languages of Europe*. Berlin, New York: De Gruyter, 563–640.

König, Ekkehard. 1986. Conditionals, concessive conditionals and concessives: Areas of contrast, overlap and neutralization. In Traugott, Elizabeth C., ter Meulen, Alice, Snitzer Reilly, Judy, Ferguson, Charles A. (eds.). *On Conditionals.* Cambridge: Cambridge University Press, 229–246.

König, Ekkehard. 1992. From discourse to syntax: The case of concessive conditionals. In Tracy, Rosemary (ed.). *Who Climbs the Grammar Tree*. Tübingen: Niemeyer, 423–433.

Leuschner, Torsten. 2006. *Hypotaxis as Building-Site: the Emergence and Grammaticalization of Concessive Conditionals in English, German and Dutch.* Munich: Lincom.

Leuschner, Torsten. 2020. Concessive conditionals as a family of constructions. *Belgian Journal of Linguistics*. 34(1), 235–247. https://doi.org/10.1075/bjl.00049.leu

Leuschner, Torsten. 2023. Die Familie der Irrelevanzkonditionale im Deutschen. Von der funktionalen Sprachtypologie zur gebrauchsbasierten Konstruktionsgrammatik. In Mollica, Fabio, Stumpf, Sören (eds.). *Konstruktionsgrammatik IX. Konstruktionsfamilien im Deutschen*. Tübingen: Stauffenburg, 327–351.

Traugott, Elizabeth Closs, Trousdale, Graeme. 2013. *Constructionalization and Constructional Changes.* Oxford: Oxford University Press.

Vander Haegen, Flor. 2023. Ein neues Mitglied der irrelevanzkonditionalen Konstruktionsfamilie. Universale Irrelevanzkonditionale des Typs [IRR *w*-] im Gegenwartsdeutschen. In Mollica, Fabio, Stumpf, Sören (eds.). *Konstruktionsgrammatik IX. Konstruktionsfamilien im Deutschen*. Tübingen: Stauffenburg, 353–388.

Vander Haegen, Flor, Bossuyt, Tom, Leuschner, Torsten. 2022. Emerging into your family of constructions: German [IRR *was*] 'no matter what'. *Constructions and Frames*. 14(1), 150–180.

MEANINGFUL REPETITION IN LATVIAN

Nicole Nau

Adam Mickiewicz University in Poznań

Keywords: repetition, reduplication, cognate constructions, Latvian, corpus study

Latvian has a rich inventory of repetitive constructions, which are an excellent object for corpus studies. This paper will give an overview of the main construction types, and their characteristics based on a study of Latvian corpora. Investigated features include differences among parts-of-speech, preferences for certain tokens, frequency and regularity of patterns. Repetition in language has been studied by linguists at least since Pott's (1862) seminal book and continues to be discussed in cross-linguistic research, e.g., by Stolz et al. (2015), Stolz (2018), Finkbeiner, Freywald (2018), Mattiola, Masini (2022). A bottom-up, data-driven approach, as followed in this paper, is a necessary complement to theoretical and typological studies in this domain.

The proposed typology of repetitive constructions starts with the kind of element that is repeated: an exact word form, a lexeme in different word forms, a root in different lexemes, or a meaningless part such as a syllable. Characteristic for Latvian is the repetition of meaningful elements, while repetition of word parts (as in typical morphological reduplication) is very rare and does not form productive patterns. Word-form repetition is sometimes called *iteration* (Stolz 2018), while lexeme and root repetition across wordforms are known as *figura etymologica* or *cognate constructions* (Nau 2019). The second criterion is a distinction between asyndetic and syndetic repetition, the latter including special constructions with additional material.

Word-form repetition is most frequent with verbs and adverbs but rare with nouns. Adjectives show diverse behaviour in syndetic and asyndetic patterns and preferences for certain forms and lexemes. Verbs in asyndetic and syndetic types may be repeated more than once (asyndetic verb, verb, verb, syndetic verb (and) verb and verb). A special construction is the frame *kā _____, tā _____*, e.g., *kā gaidām, tā gaidām* literally 'as we are waiting, so we are waiting', meaning 'we are still waiting'.

Lexeme repetition is found with verbs and nouns in various constructions. For nouns, most productive are two constructions where a noun is preceded by the same noun in genitive plural. The first has the meaning 'the ultimate' (*idiotu idiots* 'the ultimate idiot'), the second, with several subtypes, indicates a large number (*kilometru kilometri* 'many kilometers', 'miles and miles').

Root repetition is found with verbs, adjectives, and adverbs in various types. A typical Baltic pattern is the formation of an adverb which precedes the word with the root in question. For verbs, this adverb is formed with the suffix *-in* (Škrabal, Veckalne 2019), e.g., *klientu skaits* ***augtin auga*** 'the number of clients grew steadily' (literally 'grow.adv grew', "growingly grew"). Adjectives use diverse adverbial suffixes in this construction, e.g., with *garš* 'long', we find *garum garš, garu garš* and *garin garš*. With adjectives and adverbs, root repetition may occur within one word (*gargarš* 'long', *ļotļoti* 'very'). In this type, the root almost always has only one syllable, preferably of the form CVC, which makes the result look like typical reduplication. However, it is still a meaningful element (morpheme) that is repeated, not a syllable as such.

References

Finkbeiner, Rita, Freywald, Ulrike (eds.). 2018. *Exact Repetition in Grammar and Discourse*. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110592498

Mattiola, Simone, Masini, Francesca. 2022. Discontinuous reduplication: a typological sketch. *STUF Language Typology and Universals* 75.2, 271–316. https://doi.org/10.1515/stuf-2022-1055

Nau, Nicole. 2019. The Latvian continuative construction *runāt vienā runāšanā* 'talk in one talking' = 'keep talking'. *Baltic Linguistics* 10, 21–63. https://doi.org/10.32798/bl.360

Pott, August Friedrich. 1862. Doppelung (Reduplikation, Gemination) als eines der wichtigsten Bildungsmittel der Sprache. Beleuchtet aus Sprachen aller Weltteile. Lemgo, Detmold: Meyer'sche Hofbuchhandlung.

Stolz, Thomas, Urdze, Aina, Nintemann, Julia, Tsareva, Marina. 2015. When some dots turn a different color ...: Thoughts on how (not) to determine whether or not reduplication is universal. *Studies in Language*, 39(4), 795–834. https://doi.org/10.1075/sl.39.4.07sto

Stolz, Thomas. 2018. (Non-)Canonical reduplication. In Urdze, Aina (ed.). *Non-Prototypical Reduplication*. Berlin, Boston: De Gruyter Mouton, 201–279. https://doi.org/10.1515/9783110599329-007

Škrabal, Michal, Veckalne, Aiga. 2019. Bērtin bēru valodiņu: par kādu latviešu valoda verbālu konstrukciju korpuslingvistikas skatījumā. *Language: Meaning and Form.* 10, 217–231.

THE GENITIVE OF NEGATION IN LITHUANIAN – A COMPLEX PHENOMENON THAT CANNOT BE TAUGHT? ON THE IMPORTANCE OF CORPORA FOR RESEARCH AND TEACHING

Hélène de Penanros

Inalco (Institut National des Langues et Civilisations Orientales, Paris) and SEDYL (UMR8202)

Keywords: genitive of negation, Lithuanian, microlinguistics, didactics, corpus

The general trend in didactics today remains the direct heir to the communicative method born in the 1970s and 1980s. The beginnings of this approach can be traced back to the works of Noam Chomsky (e.g., 1965), whose Language Acquisition Device posits the existence of a 'universal grammar'. Transposed to language teaching, the idea is that we can encourage the 'natural' learning of any language by minimally guiding the learner through concrete communication situations.

This method has certainly made it possible to break with the old structuralist approach based on rote learning of given grammatical structures and the mechanical repetition of sentences out of context, by placing the emphasis on authentic language as it appears in real communication situations, but it has had the glaring drawback, mainly at the outset, of neglecting grammar: the role of grammar is kept to a minimum, and its presentation does not evolve much from the old out-of-context rules. Indeed, research in didactics focuses on the external conditions of communication, in particular on its socio-cultural specificities, but rarely deals with the language itself.

However, the structuring character of linguistics for language learning should no longer need to be demonstrated. The importance of metalinguistic reflection in the acquisition of the first or an additional language has been widely described (Thomas 1988, Gombert 1990, Chini 2009, or De Angelis 2007), but if the place of linguistics in current didactics remains marginal, it is probably due to the fact that the linguist has not resolved the major difficulty of transposing his research to the field of teaching. Clearly, a direct import of his theoretical analyses is doomed to failure and that it is essential to find a simple and effective way of shedding light on the phenomena, using rudimentary conceptual tools and accessible discourse.

Taking as an example a complex linguistic fact, namely the genitive of negation in Lithuanian (cf. Švambarytė 1998, Menantaud 1999, 2007, Semenienė 2005, Kalėdaitė 2008, Aleksandravičiūtė 2013, Arkadiev 2016, Kozhanov 2017), I will show how to highlight the fine distinctions that exist between competing expressions, the precise conditions of use of which pose problems for students (cp. (1) and (2)).

(1)	***Labai***	***gaila,***	***bet***	***ne-atsirado***	***žmog-aus,***
	Very	pity	but	NEG-be_found.PST3	man-GEN.SG

kuris būtų galėjęs pakeisti rungtynių eigą.

'**It's a pity, but there was no man** who could have changed the course of the match.'

https://www.atviraklaipeda.lt/2018/11/14/dragunas-patyre-netiketa-pralaimejima/

(2)	***Tiesiog***	***ne-atsirado***	***žmog-us,***
	simply	NEG-be-found.PST3	man-NOM.SG

su kuriuo galėčiau pasidalinti savo gyvenimo džiaugsmais ir rūpesčiais.

'**There just hasn't appeared a man** with whom I could share the joys and sorrows of my life.'

https://lnk.lt/video/kk2-laidoje-ingrida-simonyte-atviraus-apie-savo-gyvenima-vyrus-ir-pomegius/56199

The method combines a microlinguistic analysis based on a large corpus of attested occurrences and focusing on the study of lexicon and forms in context with a data-driven presentation, where the student's progress through the data is designed by the teacher so as to bring out the regularities that his prior research has highlighted. The student, immersed in first-hand, attested data with a linguistic and situational context, can then identify himself the principles that account for the alternations in the observed linguistic forms.

This method, used in the 4th year of Lithuanian at Inalco, appears to be a fruitful combination of the latest developments in didactics and corpus linguistics.

References

Aleksandravičiūtė, Skaistė. 2013. The semantic effects of the subject genitive of negation in Lithuanian. *Baltic Linguistics*. 4, 9–38. https://doi.org/10.32798/bl.407

Angelis, Gessica De. 2007. Third or additional language acquisition. *Second Language Acquisition*. 24. Dublin: Multilingual matters. https://doi.org/10.21832/9781847690050

Arkadiev, Peter. 2016. Long-distance genitive of negation in Lithuanian. In Holvoet, Axel, Nau, Nicole (eds.). *Argument Realization in Baltic*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 37–81. https://doi.org/10.1075/VARGREB.3.01ARK

Chini, Danielle. 2009. Linguistique et didactique : où en est-on? Quelle place pour une approche conceptualisante de la construction de la langue dans la perspective actionnelle? *Recherches en didactique des langues et des cultures. Les Cahiers de l'Acedle*, 6–2. https://doi.org/10.4000/rdlc.1958

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, London: The MIT Press.

Gombert, Jean-Émile. 1990. *Le développement métalinguistique*. Paris: PUF.

Kalėdaitė, Violeta. 2008. Language-specific existential sentence types: a case study of Lithuanian. *Kalbotyra*. 59(3), 128–137. https://doi.org/10.15388/Klbt.2008.7600

Kozhanov, Kirill. 2017. Studying variation in case marking: The genitive of negation in Aukštaitian dialects of Lithuanian. *Baltic Linguistics*. 8, 97–114. https://doi.org/10.32798/bl.378

Menantaud, Henri. 1999. La négation comme catégorie grammaticale en polonais et en lituanien. *Cahiers de linguistique de l'Inalco*. 1, 43–57.

Menantaud, Henri. 2007. Note sur une alternance morphologique induite par la négation dans les langues baltes modernes (letton et lituanien). *La Négation. Cercle linguistique d'Aix-en-Provence*. 20, 91–99.

Semenienė, Loreta. 2005. Intranzityvinio subjekto žymėjimas vardininku ir/arba kilmininku. *Acta Linguistica Lithuanica*. LII, 67–82.

Švambarytė, Janina. 1998. Objekto galininko ir kilmininko linksių kaita prie neiginio J. Basanavičiaus publikuotuose pasakojamosios tautosakos rinkiniuose. *Lituanistica*. 3(35), 53–60.

Thomas, Jacqueline. 1988. The role played by metalinguistic awareness in second and third language learning. *Journal of Multilingual and Multicultural Development*. 9(3), 235–246.

GRAMMAR AND CORPORA IN A CROSS-LINGUISTIC PERSPECTIVE. CASE STUDIES AND METHODOLOGICAL CONSIDERATIONS

Beata Trawiński

Leibniz Institute for the German Language, Mannheim (Germany)

Keywords: comparability, contrastive grammar, cross-linguistic research, multilingual corpora, (morpho)syntactic annotation

The question of which types of corpora are most appropriate for addressing specific research questions has posed an ongoing challenge for linguists since the empirical turn in the discipline. This issue is particularly pronounced in cross-linguistic domains such as contrastive linguistics (Altenberg, Granger 2002; Granger 2010; Enghels et al. 2020), language typology (Cysouw, Wälchli 2007), and translation studies (Granger et al. 2003). One of the key challenges of using corpora (both mono- and multilingual) for cross-linguistic research is ensuring their comparability, not only with regard to size, modality, and text type, but also in terms of content (see Kupietz et al. 2020; Trawiński, Kupietz 2021). In grammar-oriented studies, the availability of comparable morphological and syntactic annotations is particularly critical (see the Universal Dependencies (UD) framework; https://universaldependencies.org). Additionally, there is an urgent need for multilingual corpora that facilitate both semasiological and onomasiological access to linguistic data. Consequently, many cross-linguistic studies draw on multiple corpus resources in parallel – an approach that can enhance analytical depth, but which also introduces significant methodological challenges, particularly in quantitative contexts. Therefore, investigating grammatical phenomena in cross-linguistic corpus research continues to require innovative solutions in both corpus infrastructure and methodological design.

In this talk, I will explore the potential and challenges of corpus-based cross-linguistic grammar research by presenting four case studies that draw on a range of corpus types, including national and reference corpora, small, UD-annotated monolingual corpora, parallel/translational corpora, as well as mixed corpus resources. Conducted within the framework of the project *German Grammar in European Comparison* (*GDE*) at the Leibniz Institute for the German Language (IDS) in Mannheim, these studies address the following phenomena: (i) imperatives in Germanic (German and English) and Slavic languages (Czech and Polish); (ii) clause-embedding predicates in Germanic languages (German, Dutch, and Swedish); (iii) clausal subjects across German, English, Italian, Polish, and Hungarian; and (iv) negation raising in Germanic (German, English) and Slavic languages (Polish and Russian).

I will also introduce two novel corpus resources that have been initiated and are being implemented at the IDS: the European Reference Corpus EuReCo (Kupietz et al. 2017, 2020, 2024) and the Collection of Multilingual Parallel Sequences CoMParS (Bański et al. 2017, Trawiński et al. 2021). EuReCo is an open, long-term initiative that aims to provide and utilize virtual and dynamically definable comparable corpora based on existing national, reference or other large corpora. CoMParS, on the other hand, is a small multilingual treebank containing parallel sequences of German and other European languages. These are automatically annotated with UDs while preserving language-specific annotations and are then manually reviewed and annotated with cross-linguistic functional-semantic information (Functional Domains). Functional-semantic annotation enables onomasiologically oriented searches. Currently, six languages are systematically represented in CoMParS: German, English, French, Hungarian, Italian and Polish. I will demonstrate how both CoMParS and EuReCo provide new perspectives on fine-grained, empirically grounded cross-linguistic grammar research, drawing on large-scale corpora as well as high-quality, richly annotated small corpora.

References

Altenberg, Bengt, Granger, Sylviane (eds.). 2002. *Lexis in Contrast: Corpus-Based Approaches*. Amsterdam: Benjamins (Studies in Corpus Linguistics). https://doi.org/10.1075/scl.7

Bański, Piotr, Kamocki, Paweł, Trawiński, Beata. 2017. Legal canvas for a patchwork of multilingual quotations: the case of CoMParS. *Corpus Linguistics International Conference 2017*. Birmingham: University of Birmingham, 78–81.

Cysouw, Michael, Wälchli, Bernhard. 2007. Parallel texts: using translational equivalents in linguistic typology. *Language Typology and Universals*. 60(2), 95–99. https://doi.org/10.1524/stuf.2007.60.2.95

Enghels, Renata, Defrancq, Bart, Jansegers, Marlies (ed.). 2020. *New Approaches to Contrastive Linguistics. Empirical and Methodological Challenges*. Berlin, Boston: De Gruyter Mouton (Trends in Linguistics). https://doi.org/10.1515/9783110682588

Granger, Sylviane. 2010. Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Contemporary Foreign Language Studies*. 10(2), 14–21.

Granger, Sylviane, Lerot, Jacques, Petch-Tyson, Stephanie. 2003. *Corpus-based approaches to contrastive linguistics and translation studies*. Amsterdam & Atlanta: Rodopi.

Kupietz, Marc, Witt, Andreas, Bański, Piotr, Tufiş, Dan, Cristea, Dan, Váradi, Tamás. 2017. EuReCo – Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In Bański, Piotr, Kupietz, Marc, Lüngen, Harald, Rayson, Paul, Biber, Hanno, Breiteneder, Evelyn, Clematide, Simon, Mariani, John, Stevenson, Mark, Sick, Theresa (eds.). *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section*. Birmingham, 24 July 2017. Mannheim: Institut für Deutsche Sprache, 15–19. Available at: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-62580.

Kupietz, Marc, Diewald, Nils, Trawiński, Beata, Cosma, Ruxandra, Cristea, Dan, Tufiş, Dan, Váradi, Tamás, Wöllstein, Angelika. 2020. Recent developments in the European Reference Corpus EuReCo. In Granger, Sylviane/Lefer, Marie-Aude (eds.). *Translating and Comparing Languages: Corpus-based Insights. Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference. Louvain-la-Neuve: Presses universitaires de Louvaint*, 257–273.

Kupietz, Marc, Bański, Piotr, Diewald, Nils, Trawiński, Beata, Witt, Andreas. 2024. EuReCo: Not Building and Yet Using Federated Comparable Corpora for Cross-Linguistic Research. In Zweigenbaum, Pierre, Rapp, Reinhard, Sharoff, Serge (eds.). *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*. Paris: ELRA Language Resource Association, 94–103.

Trawiński, Beata, Kupietz, Marc. 2021. Von monolingualen Korpora über Parallelund Vergleichskorpora zum Europäischen Referenzkorpus EuReCo. In Lobin, Henning, Witt, Andreas, Wöllstein, Angelika (eds.). *Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch*. Berlin, Boston: De Gruyter, 209–234. https://doi.org/10.1515/9783110731514-012

Trawiński, Beata, Schlotthauer, Susan, Bański, Piotr. 2021. CoMParS: Eine Sammlung von multilingualen Parallelsequenzen des Deutschen und anderer europäischer Sprachen. In Lobin, Henning, Witt, Andreas, Wöllstein, Angelika. *Deutsch in Europa – sprachpolitisch, grammatisch, methodisch. Jahrbuch des Instituts für Deutsche Sprache 2020*. Berlin, Boston: De Gruyter, 301–310. https://doi.org/10.1515/9783110731514-016

2. ALL ABSTRACTS

TRACING THE RISE OF *NO*(*T*)-FRAGMENTS: COLLOQUIALISATION IN WRITTEN PRESENT-DAY ENGLISH

Laura Abalo-Dieste

Universidade de Vigo

Keywords: colloquialisation, fragments, Construction Grammar, British National Corpus, negation

This paper investigates the role of *no(t)*-fragments in contemporary British English as indicators of colloquialisation. While *no*- and *not*-fragments, exemplified in (1) and (2), respectively, have been recognised for their stylistic informality (Cappelle 2021), their distribution across registers and over time remains underexplored, particularly in relation to colloquialisation. Colloquialisation, as defined by C. Mair (1997), refers to the growing preference for speech-like style in written language, which can be observed in the frequency increase of informal linguistic devices and the corresponding avoidance of formal alternatives. E.g., the declining frequency of *no*-negation in favour of *not*-negation in clausal contexts was claimed to represent a negative sign of colloquialisation, "where colloquialisation predicts a choice strongly associated with written texts to be on the retreat" (Leech et al. 2009, 241). This study claims that fragments, categorised as "stand-alone constructions which, despite their reduced, non-canonical, fragmentary structure, are still semantically, discursively and pragmatically equivalent to a complete clause" (Fernández-Pena 2021, 136), specifically *no(t)*-fragments, offer a valuable lens for examining colloquialisation.

(1) *A picture is worth thousand words!* ***No doubt!*** *(Collins 2023, 17)*

(2) *Not us going to Starbucks for the second time today* (Pereira 2023, 1)

While fragments have been claimed to be typical of spoken discourse, their distribution in written registers remains a significant gap in current research. Previous studies have highlighted the prevalence of fragments in special registers, such as diaries (Haegeman 2007), note-taking (Janda 1985), headlines and cooking recipes (Paesani 2006), thus showing that fragments often serve functional and stylistic roles in written formats. However, despite evidence of their presence in written discourse, these studies provide limited insight into their distribution. Building on Fernández-Pena (2021: 151)'s observations that fragments are favoured in speech and informal written registers (see also Biber et al. 1999: 225) and "not as uncommon as has been thought" in writing, this investigation addresses this research gap by determining whether *no(t)*-fragments, specifically, exhibit a bias towards these domains and whether their occurrence has increased over time in contemporary British English, focusing on their role as potential markers of colloquialisation.

To that aim, this paper draws on diachronic data from the *British National Corpus* – the BNC1994 (BNC Consortium, 2007) and the BNC2014 (Brezina et al. 2021; Love et al. 2017) –, accessed through the software #LancsBox X (Brezina, Platt 2024), to analyse the frequency and textual distribution of *no(t)*-fragments. Three research questions are addressed: (i) Has the frequency of *no(t)*-fragments increased between the BNC1994 and the BNC2014? (ii) Do *no(t)*-fragments occur predominantly in informal speech and informal written registers? (iii) Do patterns of register distribution of *no(t)*-fragments show a bias towards special registers? The findings demonstrate that *no(t)*-fragments are pervasive across registers. Specifically, a statistically significant increase in the use of *no(t)*-fragments over time is observed in registers that bridge spoken and written norms, such as "written-to-be-spoken". Moreover, by showcasing a statistically significant spread of these reduced constructions into previously formal domains, such as official documents, these findings evince that *no(t)*-fragments serve as markers of ongoing colloquialisation in Present-Day English.

References

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, Finegan, Edward (eds.). 1999. *Longman grammar of spoken and written English*. Edinburgh: Pearson Education.

BNC Consortium. 2007. *British National Corpus: XML edition*. Oxford Text Archive. Available at: http://www.natcorp.ox.ac.uk/cpr.xml?ID=reference

Brezina, Vaclav, Hawtin, Abi, McEnery Tony. 2021. The written British National Corpus 2014 – design and comparability. *Text & Talk*. 41(5–6), 595–615. https://doi.org/10.1515/text-2020-0052

Brezina, Vaclav, Platt William. 2024. *#LancsBox X* [Software v.5.0.3]. Lancaster (UK): Lancaster University. Available at: https://lancsbox.lancs.ac.uk/

Cappelle, Bert. 2021. *Not*-fragments and negative expansion. *Constructions and Frames*. 13(1), 55–81. https://doi.org/10.1075/cf.00047.cap

Collins, Peter. 2023. Variation in world Englishes through the lens of negation. *World Englishes*. 43(1), 1–24. https://doi.org/10.1111/weng.12638

Fernández-Pena, Yolanda. 2021. Towards an empirical characterisation and a corpus-driven taxonomy of fragments in written contemporary English. *Revista Electrónica de Lingüística Aplicada*. 20(1), 136–154.

Haegeman, Liliane. 2007. Subject omission in Present-Day written English: On the theoretical relevance of peripheral data. *Rivista di Grammatica Generativa*. 32, 91–124.

Janda, Richard D. 1985. Note-taking English as a simplified register. *Discourse Processes*. 8(4), 437–454. https://doi.org/10.1080/01638538509544626

Leech, Geoffrey, Hundt, Marianne, Mair, Christian, Smith, Nicholas. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511642210

Love, Robbie, Dembry, Claire, Hardie, Andrew, Brezina, Vaclav, McEnery Tony. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*. 22(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Mair, Christian. 1997. The spread of the going-to-future in written English: A corpus-based investigation into language change in progress. In Raymond Hickey & Stanislaw Puppel (eds.). *Language History and Linguistic Modelling. A Festschrift for Jacek Fisiak on His 60th Birthday*. Berlin: De Gruyter, 1537–1543.

Paesani, Kate. 2006. 6. Extending the nonsentential analysis: The case of special registers. In Ljiljana Progovac, Kate Paesani, Eugenia Casielles & Ellen Barton (eds.). *Linguistik Aktuell/Linguistics Today*. 93. Amsterdam: John Benjamins, 147–182. https://doi.org/10.1075/la.93.08pae

Pereira, Guilherme M C. 2023. Not me getting with the times: A new kind of *not*-fragment in English. *Yale Working Papers in Grammatical Diversity*. 5(1), 1–22.

ANALYSING CROSS-LINGUISTIC BORROWING IN BILINGUAL FIRST LANGUAGE ACQUISITION: INTERFERENCE AND COMPARTMENTALIZATION

Marco Angster, Jakov Proroković, Mia Batinić Angster, Gordana Hržica, Marijana Kresić Vukosav & Metka Bezlaj

University of Zadar, University of Zadar, University of Zadar, University of Zagreb, University of Zadar & University of Zadar

Keywords: cross-linguistic borrowing, corpus-based research, bilingual first language acquisition, transfer, interference

Bilingual first language acquisition (2L1) is a rich field of psycholinguistic study that has been extensively explored, particularly in relation to its cognitive implications, with recent systematic reviews having challenged earlier claims that bilingual children universally benefit cognitively from early exposure to multiple languages (Gunnerud et al. 2020, Sebastian-Galles, Santolin 2020). This study aims to provide insights into the types of transfer that occur in 2L1, focusing on authentic data collected from Italian-Croatian bilingual children. The sample, collected by the parents (also researchers) in the form of diary notes, covers the period between 3 and 7 years for the older, and between 15 months and 4 years and 9 months – for the younger child. The children live in Croatia but are not surrounded by a bilingual environment outside their family. In the first phase, the diary notes were collected using a note-taking mobile phone app, ensuring timely written down context-rich observations. Later, the data collection continued using an instant messaging app which exploited the built-in meta-data (sending user, data stamp of the individual messages) allowing for faster harvesting and more fine-grained information retrieval.

The study examines children's spoken productions in relation to the hypothesis that simultaneous bilingualism from birth can impact morphosyntactic and lexical development (cf. Van Dijk et al. 2022; Bylund et al. 2023). While this language combination is not necessarily unexplored in the context of cross-linguistic influence and borrowing, it has been until recently considered mainly in the adult population of the Italian minority in Croatia. Besides, there still remain a number of open research avenues pertaining to the degree of permeability across different linguistic components such as phonology, morphology, and syntax (cf. Gardani 2022, 2024), which are directly related to issues concerning the potential compartmentalization strategies, but also the psycholinguistic motivation behind specific types of transfers, especially in the context of morphological development (Gardani 2021). The divergences form the target languages are evaluated in view of main levels of linguistic analysis – morphology (both inflection and word-formation), syntax, lexico-semantics –, the strength of language interference (accounting for the certainty of the allogenous source of the divergences observed) and the direction of transfer/borrowing (from the societal to the non-societal language, or vice versa). Finally, the systematized account of the phenomena observed is put in relation to the age of production and the situation-specific context in which they were produced. In this context, the notion of bidirectionality in language transfer is useful in exploring the types of language features that are transferred without addressing the role of dominance which appear as controversial in 2L1 context (cf. Pavlenko, Jarvis 2002). On the one hand, language transfers in 2L1 development are expected to occur bidirectionally across all domains, influencing both the societally dominant language and the non-dominant counterpart. While societal language dominance may exert a stronger influence on the non-societal language in some cases, ample exposure to the non-societal language can lead to reverse transfer, making the process bidirectional (cf. Bernardini, Van de Weijer 2017, Engemann 2022 etc.).

References

Batinić Angster, Mia, Angster, Marco. 2024. Building MaLi, a Croatian-Italian bilingual child corpus. *Strani jezici: časopis za primijenjenu lingvistiku*. 53(1), 69–88. https://doi.org/10.22210/strjez/53-1/4

Bernardini, Petra, Van de Weijer, Joost. 2017. On the direction of cross-linguistic influence in the acquisition of object clitics in French and Italian. *Language, Interaction and Acquisition*. 8(2), 204–233. https://doi.org/10.1075/lia.16005.ber

Bylund, Emanuel, Antfolk, Jan, Abrahamsson, Niclas, Olstad, Anne Marte Haug, Norrman, Gunnar, Lehtonen, Minna. 2023. Does bilingualism come with linguistic costs? A meta-analytic review of the bilingual lexical deficit. *Psychonomic Bulletin & Review*. 30(3), 897–913. https://doi.org/10.3758/s13423-022-02136-7

Engemann, Helen. 2022. How (not) to cross a boundary: Crosslinguistic influence in simultaneous bilingual children's event construal. *Bilingualism: Language and cognition*. 25(1), 42–54. https://doi.org/10.1017/S1366728921000298

Gardani, Francesco. 2021. On how morphology spreads. *Word Structure*. 14(2), 129–147. https://doi.org/10.3366/word.2021.018

Gardani, Francesco. 2022. Contact and borrowing. *The Cambridge handbook of Romance linguistics*, 845–869. https://doi.org/10.1017/9781108580410.034

Gardani, Francesco. 2024. L'italoromanzo in contatto: osservazioni sulla permeabilità differenziale della grammatica al prestito. *Revue de Linguistique Romane*. 88, 325–358. https://doi.org/10.5167/uzh-264583

Gunnerud, Hilde Lowell, Ten Braak, Dieuwer, Reikerås, Elin Kirsti Lie, Donolato, Enrica, Melby-Lervåg, Monica. 2020. Is bilingualism related to a cognitive advantage in children? A systematic review and meta-analysis. *Psychological Bulletin*. 146(12), 1059–1083. https://doi.org/10.1037/bul0000301

Pavlenko, Aneta, Jarvis, Scott 2002. Bidirectional transfer. *Applied linguistics*. 23(2), 190–214. https://doi.org/10.1093/applin/23.2.190

Sebastian-Galles, Nuria, Santolin, Chiara. 2020. Bilingual acquisition: The early steps. *Annual Review of Developmental Psychology*. 2(1), 47–68. https://doi.org/10.1146/annurev-devpsych-013119-023724

Van Dijk, Chantal, Van Wonderen, Elise, Koutamanis, Elly, Kootstra, Gerrit Jan, Dijkstra, Ton, Unsworth, Sharon. 2022. Cross-linguistic influence in simultaneous and early sequential bilingual children: a meta-analysis. *Journal of Child Language*. 49(5), 897–929. https://doi.org/10.1017/S0305000921000337

IMPLICIT EVALUATION IN CoWITE: A SYSTEMIC FUNCTIONAL PERSPECTIVE

Francisco Alonso-Almeida, Francisco J. Álvarez-Gil, Ivalla Ortega-Barrera & Elena Quintana-Toledo

University of Las Palmas de Gran Canaria

Keywords: implicit evaluation, instructional texts, gendered discourse, Systemic Functional Linguistics, Late Modern English

This study uses the *Corpus of Women's Instructive Texts in English* (CoWITE) to explore implicit evaluation in women's instructional writing from the standpoint of Systemic Functional Linguistics (SFL). A crucial component of meaning-making in instructional discourse is implicit evaluation, in which writers purposefully communicate attitudes, convictions, and authority without passing judgment outright (Martin, White 2005). The following inquiries are addressed by this study: (1) What language techniques do women use in instructional texts to express implicit evaluation? (2) How do these devices stack up against those found in writings written by men during the same time period? (3) What can be learned about the rhetorical positioning of female writers in Late Modern English from these strategies?

In terms of methodology, the study makes use of SFL's appraisal framework, paying special attention to *graduation* (force and focus) and *attitude* (emotion, judgment, and appreciation) (Martin, White 2005). CoWITE (Alonso-Almeida et al. 2025) is a morphosyntactically annotated corpus comprising over 400,000 words of instructional texts, including medical advice, cooking instructions, and therapeutic guidance, written by English-speaking women between 1700 and 1899. The texts are drawn from both printed and manuscript sources preserved in UK and US libraries. To assist comparative research, the study also incorporates a control corpus of approximately 300 000 words of instructional texts written by men from the same period and covering the same domains. Linguistic indicators of implicit evaluation, such as intensification, hedging (Hyland 1998), and vocabulary choices that covertly convey stance (Biber, Finegan 1989), are excerpted using corpus methods.

According to preliminary research, female authors of instructional writing frequently use intensification and directive language to establish credibility while also using epistemic devices to get over social restrictions on their authority. Male-authored writings, on the other hand, make direct references to expertise and display more overt forms of judgment. As a contribution to historical discourse analysis research, we hope that this study will demonstrate how gendered linguistic choices impact authorial positioning and evaluative meaning in instructional discourse in Late Modern English. This study seeks to shed fresh light on the relationship between language, authority, and gender in Late Modern English by unveiling the implicit evaluative devices used by women.

Acknowledgments

This research was supported by the Agencia Estatal de Investigación through the Plan Estatal de Investigación Científica, Técnica y de Innovación 2021–2023, European Union "Una manera de hacer Europa," under grant PID2021-125928NB-I00. The authors gratefully acknowledge this support.

References

Alonso-Almeida, Francisco, Álvarez-Gil, Francisco, Ortega-Barrera, Ivalla, Quintana-Toledo, Elena, Bator, Magdalena, de la Cruz-Cabanillas, Isabel, Sánchez-Cuervo, Margarita Esther, & Gómez-Calderón, Maria José. 2025. *Corpus of Women's Instructive Texts in English (1800–1899)* (CoWITE19). Las Palmas de Gran Canaria: University of Las Palmas de Gran Canaria. https://doi.org/10.5281/zenodo.15097949

Alonso-Almeida, Francisco, Álvarez-Gil, Francisco, Ortega-Barrera, Ivalla. 2025. *Corpus of Women's* Instructive *Texts in English (1700–1799) (CoWITE18).* Las Palmas de Gran Canaria: University of Las Palmas de Gran Canaria. https://doi.org/10.5281/zenodo.15151249

Biber, Douglas, Finegan, Edward. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text – Interdisciplinary Journal for the Study of Discourse.* 9(1), 93–124. https://doi.org/10.1515/text.1.1989.9.1.93

Hyland, Ken. 1998. *Hedging in Scientific Research Articles.* Amsterdam: John Benjamins.

Martin, J. R., White, P. R. R. 2005. *The Language of Evaluation: Appraisal in English.* Basingstoke: Palgrave Macmillan.

Salager-Meyer, Françoise. 1994. Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes.* 13(2), 149–170.

THE DUPLICATION OF THE LATVIAN PREPOSITIONS *CAUR* 'THROUGH', *PĀR* 'OVER, ABOVE' AND THE ADVERBS *CAURI* THROUGH', *PĀRI* 'BEYOND': A CORPUS-BASED ANALYSIS

Ieva Auziņa

University of Latvia

Keywords: prepositions, adverbs, prefixes, verbs, duplications

In the Latvian language, prepositions, adverbs, and prefixes with similar or identical meanings are frequently used in connection with verbs (Nītiņa, Grigorjevs 2013; Kalnača, Lokmane 2021). This study examines the usage of the prepositions *caur, pār* and the adverbs *cauri, pāri* in Latvian, particularly in colloquial and literary texts, where they often appear in one construction (Dimiņš 2017), as in examples (1) and (2):

(1) *Izvilku kabatlakatiņu, ar kāju sagriezu atpakaļ somu,*
izvēru *kabatlakatiņu* ***cauri*** ***caur***
thread.PST.1SG handkerchief.ACC.SG through through
nesamām *cilpām* *un* *sasēju.*
carrying.DAT.PL loop.DAT.PL and tie.PST.1SG
'I pulled out a handkerchief, turned the bag back upright with my foot, threaded the handkerchief through into the carrying loops and tied.' (LVK2022)

(2) *Radošā* *enerģija* *rāpsies* ***pāri*** ***pār***
creative.NOM.SG energy.NOM.SG climb.FUT.3 beyond over
vāroša *katliņa* *malām.*
boiling.PTCP.GEN.M pot.GEN.SG edge.DAT.PL
'Creative energy will climb over the edges of a boiling pot.' (LVK2022)

The use of these constructions has been selected for research because, as demonstrated by the examples, there would be no change in the sentence's meaning if only a preposition or only an adverb were used. However, such duplications are common in both Latvian colloquial speech and literary texts – 335 instances were found in the corpora where the preposition *pār* is duplicated with *pāri*, and *caur* is duplicated with *cauri*. This paper explores constructions in which verbs associated with the preposition and the adverb are employed in both literal and figurative senses, to determine whether the doubling of *caur* and *cauri*, *pār* and *pāri* is related to whether the verb is used figuratively or literally.

The paper concludes that the doubling of *caur* and *caur, pār* and *pāri* in Latvian language serves a stylistic function, heightening the visual and figurative aspects of the language in both direct and metaphorical contexts. The findings suggest that this doubling emphasizes the process of action, making it more vivid to the reader. In most cases, the preposition and the adverb are doubled when the verb, used figuratively, forms an idiom with the adverb.

The empirical data is sourced from various corpora, specifically "Balanced Corpus of Modern Latvian Texts" (LVK2018 and LVK2022), "LATE Media Speech Corpus" (LATE-mediji) etc. Additionally, the *NoSketch Engine* is utilized for data processing and analysis. The use of corpora and their tools allows this research to be reliable and systematic, firstly, by identifying the most frequently used verbs with the selected prepositions and adverbs to conduct an objective study, it also provides the opportunity to perform a statistical analysis.

This study was carried out in the framework of the Latvian Council of Science research project "Database of Latvian Morphemes and Derivational Models" (No. lzp-2022/1-0013).

References

Auziņa, Ilze, Darģis, Roberts, Levāne-Petrova, Kristīne, Auziņa, Anna, Saulīte Baiba, Ļaksa-Timinska, Ilze, Gailīte, Elīna, Nešpore-Bērzkalne, Gunta, Rābante-Buša, Guna, Pokratniece, Kristīne, Klints, Agute. 2024. *LATE Media Speech Corpus (LATE-mediji).* CLARIN-LV digital library at IMCS, University of Latvia. Available at: http://hdl.handle.net/20.500.12574/114

Dimiņš, Dens. 2017. Dažas latviešu valodas verbu konstrukcijas ar priedēkļa dubultojumu ar prievārdu vai apstākļa vārdu. *Semantika, sintakse, valodas kultūra. Res Latvienses.* 4, 63–73. https://doi.org/10.22364/RL.4.6

Kalnača, Andra, Lokmane, Ilze. 2021. *Latvian Grammar.* Rīga: University of Latvia Press. https://doi.org/10.22364/latgram.2021

Levāne-Petrova, Kristīne, Darģis, Roberts. 2018. *Balanced Corpus of Modern Latvian (LVK2018),* CLARIN-LV digital library at IMCS, University of Latvia. Available at: http://hdl.handle.net/20.500.12574/11

Levāne-Petrova, Kristīne, Darģis, Roberts, Pokratniece, Kristīne & Viesturs Jūlijs Lasmanis. 2023. *Balanced Corpus of Modern Latvian (LVK2022),* CLARIN-LV digital library at IMCS, University of Latvia. Available at: http://hdl.handle.net/20.500.12574/84

Nītiņa, Daina, Grigorjevs, Juris (eds.). 2013. *Latviešu valodas gramatika.* Rīga: Latvijas Universitātes Latviešu valodas institūts.

SOME ASPECTS OF CONNECTED SPEECH: LATE SPEECH CORPORA ANALYSIS

Ilze Auziņa & Guna Rābante-Buša

Institute of Mathematics and Computer Science, University of Latvia

Keywords: speech corpus, connected speech, modification, deletion, linking

Speech corpora play a crucial role in advancing not only the language technology development – automatic speech recognition (ASR) and text-to-speech synthesis (TTS) in particular – but also the understanding and insights of phonetics and prosody, morphology and syntax, semantics, and pragmatics of a language. The aim of this paper is to present an overview of recent Latvian speech corpora – open-access corpora for linguistic and connected speech analysis.

Connected speech refers to spoken language when analysed as a continuous sequence, as in normal utterances and conversations (Crystal 2008, 101). When a word is uttered alongside other words, its pronunciation can undergo numerous processes, that is, important changes happen when words or phrases are used in connected speech. Connected speech could be affected by multiple factors, among which are speech speed, semantics, word frequency, and phonological awareness. Studies on connected speech focus mostly on the English language, whereas studies on other languages are comparatively rare (Bi et al. 2022). A classification for connected speech processes differs, but we use the one proposed by Alameen and Levis (2018) to describe processes in Latvian connected speech: (1) linking; (2) deletion; (3) insertion; (4) modification; (5) reduction; and (6) multiple processes.

Newly developed open-access Latvian speech corpora (*LATE-mediji* (2024a), *LATE-sarunas* (2024b), *LATE Phonetically Annotated Speech Corpus* (*fonLATE*) (2024c)) include spontaneous or read speech data (over 120 hours of orthographically transcribed and morphologically annotated speech data) that allows systematic analysis of connected speech.

The first results of the corpora data analysis show that widespread processes in Latvian connected speech are:

(1) modification (e.g., assimilation of voicing: *jū**s** **b**ūsiet* 'you will be' [j̯uː**z** **b**uːsi͜et]; *krēsls u**z** **k**ā* [u**s** **k**ɑː] *sēdēt* 'chair to sit on'; assimilation by place of articulation: *lī**dz** **š**im* [liː**t͡ʃ** ʃim] 'until now),

(2) reduction (e.g., vowel reduction: *ārpus Rīg**a**s šīs kultūr**a**s iespēj**a**s baudīt daudz mazāk**a**s* [ɑːrpŭs riːgʃ ʃiːs kultuːrs i͜espeːi̯z bɑ͜udiːt dɑ͜ud͡z mɑzɑːks] ‘outside of Riga, there are much fewer opportunities to enjoy this culture’).

In a certain position in connected speech linking is also found – the structure of syllables and the location of syllable boundaries change due to the fusion of words, e.g., *jau ir* [j̯ɑ.vir] ‘already is’, *cits cilvēks* [t͡si.t͡sil.væːks] ‘another person’. So far, deletion and sound insertion have been detected very rarely, and depend on the individual speaker.

This research is funded by the Latvian Council of Science project “Standardization of Modern Latvian Pronunciation” (No. lzp-2024/1-0613).

References

Alameen, Ghinwa, Levis, John M. 2018. Connected speech. In Reed, Marnie, Levis, John M. (eds.). *The Handbook of English Pronunciation.* Malden, MA: Wiley Blackwell. https://doi.org/10.1002/9781118346952.ch9

Auziņa, Ilze, Darģis, Roberts, Levāne-Petrova, Kristīne, Auziņa, Arta, Saulīte, Baiba, Ļaksa-Timinska, Ilze, Gailīte, Elīna, Nešpore-Bērzkalne, Gunta, Rābante-Buša, Guna, Pokratniece, Kristīne, Klints, Agute. 2024a. LATE Media Speech Corpus V1 (LATE-mediji). *CLARIN-LV digital library at IMCS*, University of Latvia. Available at: http://hdl.handle.net/20.500.12574/114

Auziņa, Ilze, Darģis, Roberts, Rābante-Buša, Guna, Ļaksa-Timinska, Ilze, Gailīte, Elīna, Auziņa, Arta. 2024b. LATE Conversational Speech Corpus V1 (LATE-sarunas). *CLARIN-LV digital library at IMCS*, University of Latvia. Available at: http://hdl.handle.net/20.500.12574/113

Auziņa, Ilze, Rābante-Buša, Guna, Darģis, Roberts. 2024c. LATE Phonetically Annotated Speech Corpus V1 (fonLATE). *CLARIN-LV digital library at IMCS,* University of Latvia. Available at: http://hdl.handle.net/20.500.12574/115

Bi, Huichao, Zae, Samed, Kania, Ursula, Yan, Rong. 2022. A systematic review of studies on connected speech processing: Trends, key findings, and implications. *Frontiers in Psychology.* 13. https://doi.org/10.3389/fpsyg.2022.1056827

Crystal, David. 2008. *A Dictionary of Linguistics and Phonetics.* 6th edition. Oxford: Blackwell Publishing.

MULTIPLE *WH*-QUESTIONS IN ROMANIAN: A CORPUS-BASED APPROACH

Gabriela Bîlbîie

University of Bucharest & LLF

Keywords: *wh*-questions, coordination, Romanian, corpus, construction

Romanian, like other languages (Bulgarian, Hungarian, Serbo-Croatian, Russian, etc., see Citko, Gračanin-Yuksek 2013), allows the alternation between coordinated *wh*-questions (CWQs) (1a) and 'paratactic' multiple *wh*-questions (MWQs) (1b), where two (or more) *wh*-phrases are fronted with (1a) or without (1b) a conjunction, regardless of their syntactic function (in particular, in both constructions, the fronted *wh*-phrases can be arguments of a verbal head, like in (1)).

(1) a. ***Cine*** ***și*** ***ce*** *a* *mâncat?* (CWQs)
who and what has eaten
b. ***Cine*** ***ce*** *a* *mâncat?* (MWQs)
who what has eaten
'Who ate what?

Previous literature (Comorovski 1996, Raţiu 2011, Citko, Gračanin-Yüksek 2013) assumes that the two patterns (CWQs and MWQs) are distinct constructions, with different semantic and syntactic properties, as schematized in Table 1.

However, previous works are based solely on introspection data, many examples being artificial and lacking appropriate context. Our main goal is to confront previous research with corpus data to account for the behaviour of CWQs in actual usage. Here we present the preliminary results of a corpus investigation based on the authentic uses from the *CoRoLa* (*The Reference Corpus of the Contemporary Romanian Language*, Barbu Mititelu et al. 2018) that comprises both a written and an oral part.

Table 1. Properties of CWQs vs. MWQs

	Coordinated *wh*-questions (CWQs)	**Multiple *wh*-questions (MWQs)**
Semantic interpretation	single-pair reading	pair-list reading
Ordering constraints (superiority effects)	no ordering constraints	strict ordering constraints

One of the most striking properties of these two constructions in corpus is their preference for embedded contexts. In previous research, the prototypical pattern is in a main clause configuration. Our corpus data show, however, a main/subordinate clause asymmetry, as noted by Gazdik (2011) for Hungarian.

Moreover, there is no clear-cut correlation between the construction type (CWQs vs. MWQs) and semantic interpretation: both single-pair and pair-list readings are available with both constructions. In particular, CWQs are compatible not only with single-pair reading, but also with pair-list reading (*contra* Citko, Gračanin-Yüksek 2013).

Concerning so-called superiority effects, although CWQs do not impose strict ordering constraints as MWQ, there are still preferences for maintaining the same ordering as in MWQ. We observe a preference for 'animate-first' (and in particular 'human-first') if one of the *wh*-phrases is an animate subject, in line with the animacy hierarchy (Silverstein 1976).

Crucially, our corpus data show that CWQs in particular do not form an homogeneous class in Romanian, but they give rise to three potential syntactic analyses: (i) mono-clausal CWQs with a subclausal coordination; (ii) bi-clausal CWQs with ellipsis in the first clause (i.e., the first *wh*-phrase is a fragmentary 'short question') and (iii) bi-clausal CWQs with ellipsis in the second clause (i.e. the second *wh*-phrase is a fragmentary 'short question' and has a parenthetical status, while the first *wh*-phrase belongs to the full clause). The bi-clausal analyses with 'sharing' proposed by Citko and Gračanin-Yüksek (2013) are challenged by those attested data that involve mismatch, where the verbal head is not compatible with the first *wh*-clause. This kind of data is problematic for any syntactic approach that appeals to a syntactic reconstruction mechanism, but not for approaches assuming semantic reconstruction and a fragmentary syntax (Ginzburg, Sag 2000).

References

Barbu Mititelu, Verginica, Tufiş, Dan, Irimia, Elena. 2018. The reference corpus of the contemporary Romanian language (CoRoLa). In Calzolari, Nicoletta, Choukri, Khalid, Cieri, Cristopher, Declerck, Thierry et al. (eds.). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), European Language Resources Association (ELRA), 1178–1185.

Citko, Barbara, Gračanin-Yuksek, Martina. 2013. Towards a typology of coordinated wh-questions. *Journal of Linguistics*. 49(1), 1–32. https://doi.org/10.1017/S0022226712000175

Comorovski, Ileana. 1996. *Interrogative phrases and the syntax-semantics interface*. Dordrecht: Kluwer.

Gazdik, Anna. 2011. *Multiple questions in French and in Hungarian: A Lexical-Functional analysis with special emphasis on the syntax-discourse interface*. PhD dissertation. Paris: Université Paris Diderot Paris 7.

Ginzburg, Jonathan, Sag, Ivan A. 2000. *Interrogative investigations: The form, meaning and use of English interrogatives*. Stanford: CSLI Publications.

Raţiu, Dafina. 2011. A multidominance account for conjoined questions in Romanian. In Herschensohn, Julia (ed.). *Romance linguistics 2010*. Seattle, Washington: John Benjamins, 257–270. https://doi.org/10.1075/cilt.318.16rat

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In Dixon, R. M. W. (ed.), *Grammatical categories in Australian languages*. Canberra: Australian Institute of Aboriginal Studies, 112–171.

EXPLORING NUMERICAL CONCEPTS THROUGH LANGUAGE: A COMPARATIVE CORPUS ANALYSIS

Adriano Cerri

University of Pisa

Keywords: numerals, numerical concepts, frequency, corpora, Baltic languages

Laboratory experiments have demonstrated that the cognitive salience of numerical concepts is not homogeneous; rather, it varies significantly depending on distinctions between high and low cardinalities, as well as between simple units and round numbers (Rosch 1975; Dehaene 1997). Furthermore, within this variability, certain cross-linguistically recurring patterns can be identified, which can be partially explained by cognitive factors (Dehaene, Mehler 1992).

In this presentation, I aim to explore how linguistic data can shed light on the cognitive architecture of speakers with regard to numerical concepts. The examined data pertain to the absolute frequency of lexical items directly related to number concepts (i.e., numerals) in contemporary Latvian and Lithuanian. These data are sourced from the two largest available corpora of these languages, namely, the Balanced Corpus of Contemporary Latvian Texts (LVK2022) and the Corpus of Contemporary Lithuanian (DLKT). Previous studies on earlier versions of these corpora (Cerri 2019) will be expanded with new data extracted from their latest releases, as well as from two corpora specifically devoted to spoken language, i.e., the Latvian Subtitles of Public Broadcasting (SUBTITRI) and the Corpus of Spoken Lithuanian Language (SLKT). The findings will also be compared with data collected by other scholars for genealogically and typologically diverse languages to identify common distribution patterns.

This research aims to achieve the following objectives: (i) to assess whether the frequency of linguistic forms in a corpus can provide a tangible reflection of cognitive architecture in the domain of numeracy; (ii) to examine elements of consistency and discrepancy in numeral usage across written and spoken corpora, with particular attention to evaluating the persistence of recessive numeral forms, such as cardinals for *pluralia tantum* nouns; and (iii) to highlight the main cross-linguistic regularities by comparing Baltic language data with data from other languages.

By integrating linguistic and cognitive perspectives, this study seeks to contribute to understand the relationship between language use and numerical cognition.

References

Cerri, Adriano. 2019. The frequency of the use of Baltic numerals: Cognitive, linguistic, and cultural factors. *Baltistica*. 54(2), 257–286. https://doi.org/10.15388/baltistica.54.2.2393

Dehaene, Stanislas. 1997. *The Number Sense: How the Mind Creates Mathematics*. New York: Oxford University Press.

Dehaene, Stanislas, Mehler, Jacques. 1992. Cross-linguistic regularities in the frequency of number words. *Cognition*. 43(1), 1–29. https://doi.org/10.1016/0010-0277(92)90030-L

DLKT – *Dabartinės lietuvių kalbos tekstynas / Corpus of the contemporary Lithuanian language*. Kaunas: Vytautas Magnus University. Available at: http://corpus.vdu.lt/lt/

LVK2022 – *Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss / Balanced Corpus of Contemporary Latvian Texts*. Rīga: University of Latvia. https://korpuss.lv/id/LVK2022

Rosch, Eleanor. 1975. Cognitive Reference Points. *Cognitive Psychology*. 7(4), 532–547. https://doi.org/10.1016/0010-0285(75)90021-3

SLKT – *Sakytinės lietuvių kalbos tekstynas / Corpus of Spoken Lithuanian*. Kaunas: Vytautas Magnus University. Available at: http://sakytinistekstynas.vdu.lt/

SUBTITRI – *Latvijas sabiedrisko mediju subtitru korpuss / Latvian Subtitles of Public Broadcasting*. Rīga: University of Latvia. Available at: https://korpuss.lv/id/Subtitri

ANALYSIS AND DETECTION OF PHRASEOLOGICAL UNITS (PUS) IN THE FRENCH CORPUS: A LEXICO-GRAMMATICAL APPROACH IN NLP

Chen Lian & Dao Huy-Linh

LLL-CNRS-University of Orléans & CRLAO–CNRS–INALCO

Keywords: PUs, identification of PUs, lexico-grammatical, NLP

Phraseology, also known as "fixation" (Gross 1982, 1984, 1988, Nunberg et al. 1994, Klein, Lamiroy 1994, 2005, Gross 1996, Polguère 2002, Lamiroy 2003, Mejri 1997, 2003, Chen 2021, etc.), is still a relatively young discipline within the language sciences. Its status remains debated: some researchers consider it a branch of lexicology, while others view it as closely related to syntax. In our project, we investigate phraseological units (PUs) in French through a lexico-grammatical approach, combining lexical and syntactic analysis in line with Maurice Gross's (1975) work on lexicon-grammar.

We focus on the most frequent "verbal idiomatic expressions" (Gonzalez Rey 2002) in the corpus. Our objective is to develop automatic processing tools to detect these units in a large text corpus. For this purpose, we use a French corpus extracted from Wikipedia (OPUS Moses version v1.0, 2018: http://opus.nlpl.eu/Wikipedia-v1.0.php), which contains approximately 15 791 884 words. The corpus was divided into 200 segments ("chunks") of around 79 000 words each. The experiments were conducted on the first 10 chunks.

Lexical Approach

We extracted a lexicon of 2,248 phraseological units (PUs) from XML files in the ATILF laboratory database and converted it into CSV format. Using SpaCy (https://spacy.io/), we detected approximately 118 occurrences of idiomatic expressions in the corpus. However, this detection remains partial, as the morphosyntactic variability of PUs poses significant challenges. E.g., the expression *casser sa pipe* 'to kick the bucket' can appear in different forms: with pronominalization (*il l'a cassée* 'he broke it'), variation of the possessive (*leur pipe* 'their pipe'), addition of adverbs, or even insertion of intervening elements.

Syntactic Approach

To address these limitations, we use Stanza (https://github.com/stanfordnlp/stanza) and define several syntactic patterns that correspond to typical structures of verbal phraseological units (PUs):

N + V (e.g., *un ange passe* ‘an angel passes’)

V + direct object (e.g., *chercher noise* ‘to pick a fight’)

V + complements (e.g., *faire flèche de tout bois* ‘to make use of every means’)

V + adjective/adverb (e.g., *voir rouge* ‘to see red’)

V + comparison (e.g., *pleuvoir comme vache qui pisse* ‘to rain heavily’)

These patterns are essential for identifying discontinuous and structurally complex expressions. We also designed a script to identify the 30 most frequent verbs in the corpus segments, assuming that these verbs might indicate the presence of phraseological units (PUs).

However, this simple frequency-based method quickly reveals its limitations: many of the detected expressions are not idiomatic (e.g., ***mettre** en conformité* ‘to bring into compliance’, or ***faites** par un humain* ‘made by a human’).

These false positives highlight the importance of combining frequency data with syntactic and contextual analysis.

Annotation

These examples show that, although the verbs are correctly extracted, the identified expressions are not always idiomatic in the linguistic sense. This highlights the need to explore more robust methods, such as syntactic pattern detection (based on chunking or parsing), to better capture the typical structures of phraseological units (PUs).

Thus, this method is complemented using the Arborator-Grew tool (https://arborator.grew.fr), which enables fine-grained syntactic annotation of PUs, allowing for a better understanding of their complex structures. Annotated example:

Expression: “Pleuvoir des hallebardes, tomber des cordes”

pleuvoir	VERB	_	root	_	_
des	DET	_	det	_	hallebardes
hallebardes	NOUN	_	obj	_	pleuvoir
tomber	VERB	_	conj	_	pleuvoir
des	DET	_	det	_	cordes
cordes	NOUN	_	obj	_	tomber

Automatic identification of PUs requires both reliable linguistic resources (e.g., lexicons) and robust syntactic models. Combining lexical and syntactic analyses allows for a better understanding of the diversity and variability of idiomatic expressions in text corpora.

References

Caseli, Helena de Madeiros, Ramisch, Carlos, das Graças Volpe Nunes, Maria, Villavicencio, Aline. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*. 44(1), 59–77.

Constant, Matthieu, Nivre, Joakim. 2016. A transition-based system for joint lexical and syntactic analysis. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2016).* 1, 161–171. https://doi.org/10.18653/v1/P16-1016

Constant, Mathieu, Eryiğit, Gülşen, Monti, Johanna, van der Plas, Lonneke, Ramisch, Carlos, Rosner, Michael, Todirascu, Amalia. 2017. Multiword expression processing: A survey. *Computational Linguistics*. 43(4), 837–892. https://doi.org/10.1162/COLI_a_00302

Constant, Matthieu, Sigogne, Anthony, Watrin, Patric. 2012. La reconnaissance des mots composés à l'épreuve de l'analyse syntaxique et vice-versa: évaluation de deux stratégies discriminantes. *Conférence sur le Traitement Automatique des Langues Naturelles*, 57–70.

Constant, M. et Tellier, I. 2012. Evaluating the impact of external lexical resources into a CRF-based multiword segmenter and part-of-speech tagger. *The 8th International Conference on Language Resources and Evaluation (LREC'12)*, 646–650.

Chen, Lian. 2021. *Analyse comparative des expressions idiomatiques en chinois et en français (relatives au corps humain et aux animaux)*. Thèse en linguistique. Paris: Cergy Paris Université.

Gross, Maurice. 1982. Une classification des phrases "figées" du français. *Érudit, Revue québécoise de linguistique*. 11(2), 151–185. https://doi.org/10.7202/602492

Gross, Maurice. 1984. Une classification des phrases 'figées' du français. *Lingvisticæ Investigationes Supplementa*, Pierre ATTAL et Claude MULLER (éds.) De la syntaxe à la pragmatique. 8, 141–180. https://doi.org/10.1075/lis.8.08gro

Gross, Maurice. 1988a. Les limites de la phrase figée. *Langage, Fait partie d'un numéro thématique: Les expressions figées*. 90, 7–22.

Gross, Maurice. 1988b. Sur les phrases figées complexes du français. *Langue française*. 77, 47–70.

Gross, Gaston. 1996. *Les expressions figées en français: noms composés et autres locutions*. Paris: Ophrys.

Klein, Jean René, Lamiroy, Béatrice. 1994. Lexique-Grammaire du Français de Belgique: Les Expressions Figées. *Lingvisticæ Investigationes*. 18(2), 289–320. https://doi.org/10.1075/li.18.2.04kle

Klein, Jean René, Lamiroy, Béatrice. 2005. Relations systématiques entre expressions verbales figées à travers quatre variétés du français. *Cahiers de l'Institut de Linguistique de Louvain: Peeters*, 31 (2–4), 77–92. https://hal.science/hal-01618782v1

Mejri, Salah. 1997. *Le figement lexical: descriptions linguistiques et structuration sémantique, série linguistique*. X. Publications de la Faculté des lettres de la Manouba.

Mejri, Salah. 2003. Le figement lexical. In *Le figement lexical, Cahiers de lexicologie*. Salah MEJRI (dir.). 1(82), 23–39.

Nunberg, Geoffrey, Sag, Ivan A., Wasow, Thomas. 1994. Idioms. *Language*. 70(3), 491–538.

Polguère, Alain. 2002. *Notions de base en lexicologie*. (Version préliminaire septembre 2002, pour LNG 1080), Observatoire de Linguistique Sens-Texte.

THE NATURE OF MOUTH ACTIONS IN POLISH SIGN LANGUAGE (PJM): A CORPUS-BASED STUDY

Rafał Darasz

University of Warsaw

Keywords: sign linguistics, Polish Sign Language, mouth actions

The topic of mouth actions has been researched throughout numerous sign languages since the 1980s. Despite extensive research, the exact nature and functions of mouth actions have not been specifically defined to a satisfactory extent. Two main types of mouth actions occurring alongside manual signs can be distinguished: mouthings related to words of a spoken language and mouth gestures autonomous from spoken languages (Boyes-Braem, Sutton-Spence 2001). Mouth actions in Polish Sign Language (PJM) have not been analysed extensively so far.

The aim of this study is to analyse the variety and grammatical functions of mouth actions in PJM, their obligatoriness, as well as elaborate on the mouth actions variation in terms of gender and age. The research material comes from the Corpus of Polish Sign Language (Wójcicka et al. 2020) created by the Section for Sign Linguistics at the University of Warsaw, which is currently one of the world's largest signed language corpora (about 700 000 annotated units). It contains recordings of deaf PJM users aged 18–90 performing a series of tasks designed to elicit linguistic data (Mostowski 2014). Using ELAN software, additional tiers were added to the existing .eaf-file annotations with gloss and translations tiers. The tiers were partly based on the Annotation Conventions for the NGT Corpus (Crasborn et al. 2015) and the Auslan Corpus Annotation Guidelines (Johnston 2019). The preliminary data analysed so far consists of recordings of 16 PJM signers from the open access Open Repository of Polish Sign Language (https://www.korpuspjm.uw.edu.pl/en).

The PJM data are comparable with similar research on other SLs. Out of total 1990 sign tokens, 1455 (~73%) of them were accompanied by mouth gestures (Auslan [77% manual signs with mouth actions; 17,002 manual sign tokens; (Johnston et al. 2016)] and BSL [71%; 1099 manual sign tokens; (Crasborn et al., 2008)]. 58% of mouth gestures were classified as mouthing (results comparable with BSL (51%) and STS (Swedish SL – 59%). The preliminary results also show that there is variation in terms of gender and age (e.g., men and younger participants use fewer informants than women and older informants).

References

Boyes-Braem, Penny, Sutton-Spence, Rachel (eds.). 2001. *The hands are the head of the mouth: The mouth as articulator in sign languages*. Hamburg: Signum-Verl.

Crasborn, Onno A., Bank, Richard, Zwitserlood, Inge, van der Kooij, Els, de Meijer, Anna, Sáfár, Anne. 2015. *Annotation Guidelines for Corpus NGT*.

Crasborn, Onno A., van der Kooij, Els, Waters, Dafydd, Woll, Bence, Mesch, Johanna. 2008. Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language & Linguistics*. 11(1), 45–67. https://doi.org/10.1075/sll.11.1.04cra

Johnston, Trevor. 2019. *Auslan Corpus Annotation Guidelines*.

Mohr, Susanne. 2014. *Mouth Actions in Sign Languages: An Empirical Study of Irish Sign Language*. Boston: De Gruyter Mouton/Ishara Press. https://doi.org/10.1515/9781614514978

Wójcicka, Joanna, Kuder, Anna, Mostowski, Piotr, Rutkowski, Paweł (eds.). 2020. *Otwarte Repozytorium Korpusu Polskiego Języka Migowego*. Warszawa: Wydział Polonistyki Uniwersytetu Warszawskiego. Available at: https://www.korpuspjm.uw.edu.pl

THE BALTIC PERFECT AND EVIDENTIAL: A SEARCH IN A PARALLEL CORPUS OF LITERARY PROSE

Anna Daugavet

Vilnius University

Keywords: parallel corpus, Baltic, grammaticalization, participles, perfect, evidential, narrative

In the Baltic languages Latvian and Lithuanian the predicative use of active past participles (glossed PPA), either accompanied by the copula/auxiliary 'be' (inflected for tense) or without it, underlies the grammatical categories of perfect and evidential. The two categories are historically related and cannot be always easily distinguished even when the presence of the copula/auxiliary is considered (Nau 2005; Wiemer 2006). To illustrate the point, (1) from LiLa is ambiguous between an experiential perfect and a quotative evidential.

(1)	Latvian	*Tādas*	*sētas*	*viņa*
	Lithuanian	*Tokias*	*sodybas*	*ji*
		such.ACC.PL	homestead.ACC.PL	3SG.NOM.F
	Latvian	***redzēj-us-i***	*vienīgi*	*Krievijā.*
	Lithuanian	***mači-us-i***	*tik*	*Rusijoje.*
		see-PPA-NOM.SG.f	only	Russia.LOC.SG

a) 'She has only seen such homesteads in Russia.'
b) 'She claims to have only seen such homesteads in Russia.'

As a parallel corpus containing translations of Lithuanian literary prose into Latvian and vice versa, LiLa offers a relatively convenient tool for comparing the use of perfect and evidential forms in the two Baltic languages, although the methods employed in the study, as well as its results, are clearly affected by the choice of data.

Narratives, common in LiLa texts, favour the use of evidentials (2), that are more frequent in literary prose, at least in Lithuanian, in comparison to other types of data, such as *Facebook* comments, treated as a more natural source of Lithuanian perfects in Kapkan (2024).

(2) Lithuanian (original) *Jame* ***gyven-us-i*** *kraugerė*
there live-PPA-NOM.SG.F bloodthirsty.NOM.SG
Lithuanian (original) *demonė* *Kasmali.*
demon PN
'There (in the graveyard) lived a bloodthirsty demon called Kasmale.'

Narratives also interfere with perfect forms that have to be differentiated between deictic use, as in dialogues (3), and non-deictic use in narrative sequences, e.g., in the historical present (4), revealing differences between the two Baltic languages. (See Fleischman (1990) on the function of tense and aspect in narratives, as well as Ritz & Engel (2008) on the narrative use of perfect.) The original Latvian present perfect is replaced with simple past and present forms in the Lithuanian translation. Also note the addition of *jau* 'already' to the Lithuanian translation, absent from the Latvian original, see Dahl (2022) on iamitives.

(3) Latvian (original) *Vai jūs* ***esat*** ***aizmirs-us-i****,* *ka* <...>
Q 2PL.NOM be.PRS.2PL forget-PPA-NOM.SG.F COMP
Lithuanian (translation) *Ar jūs* *jau* ***pamiršote****,* *kad* <...>
Q 2PL.NOM already FORGET.PST.2PL COMP
'Have you (already) forgotten that <...>'

(4) Latvian (original) <*Gabriēla kodī pirkstus un*>
nez kāpēc ***ir*** ***apklus-us-i****.*
for.some.reason be.PRS.3 fall.silent-PPA-NOM.SG.F
Lithuanian (translation) <*Gabriela kramto pirštus ir*>
nežinia kodėl ***nutyla****.*
for.some.reason fall.silent.PRS.3
'<Gabriela is biting her fingers and> has fallen silent for some reason.'

The study confirms the intuitive proposition that functions associated with perfect in Western and Northern European languages (Dahl 2022), such as the meaning of current relevance in (3), are more developed in Latvian than in Lithuanian. Meanwhile, the lesser degree of grammaticalization does not preclude Lithuanian participles from acquiring evidential extensions of resultative uses, as described for other languages in Aikhenvald (2004).

Sources

LiLa = Lithuanian-Latvian-Lithuanian Parallel Text Corpus at https://korpuss.lv/en/id/LiLa

References

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford University Press.

Dahl, Östen. 2022. Perfects Across Languages. *Annual Review of Linguistics*. 8(1), 279–297. https://doi.org/10.1146/annurev-linguistics-031120-123428

Fleischman, Suzanne. 1990. *Tense and Narrativity: From Medieval Performance to Modern Fiction*. London: Routledge.

Kapkan, Danguolė Kotryna. 2024. *The Grammaticalization of BE Perfects and Beyond: Case Studies in Lithuanian, Bulgarian and Barese*. Doctoral dissertation. Vilnius University.

Nau, Nicole. 2005. Perfekts un saliktā tagadne latviešu valodā [Perfect and Compound Present in Latvian]. *Baltu filoloģija*. 14(2), 137–154.

Ritz, Marie-Eve E. and Dulcie M. Engel. 2008. "Vivid narrative use" and the meaning of the present perfect in spoken Australian English. *Linguistics*. 46(1), 131–160. https://doi.org/10.1515/LING.2008.005

Wiemer, Björn. 2006. Grammatical Evidentiality in Lithuanian (A Typological Assessment). *Baltistica*. XLI(1), 33–49.

A CORPUS-BASED REASSESSMENT OF OLD CHURCH SLAVONIC WORD ORDER

Andrea Di Manno

University of Naples "L'Orientale"

Keywords: Old Church Slavonic, word order, Shannon's entropy, typology, corpora

Descriptions of word order in Old Church Slavonic (OCS) are relatively scarce, and there is little consensus among scholars. Excluding generativist and functionalist studies, most accounts consist of brief annotations in grammars or typological works. Within this typological tradition, some scholars propose a basic syntactic order for OCS (e.g., SVO in Siewierska, Rijkhoff, Bakker 1998, Östling 2015, Choi et al. 2021, see also Biagetti et al. 2023), while others (e.g., Levshina 2019) argue that its word order cannot be adequately captured by syntactic relations.

This study provides a comprehensive characterization of OCS word order by analysing data from treebanks annotated using the *Universal Dependencies* framework (Zeman et al. 2024). Data from the *Codex Marianus*, *Codex Suprasliensis*, and *Psalterium Sinaiticum* (Haug, Jøhndal 2008; Eckhoff, Berdicevskis 2015) are compared to both ancient and modern (esp. Slavic and Baltic) Indo-European languages and to typologically well-characterized reference languages, such as Hungarian (free word order) and English (rigid word order).

Shannon's (1948) entropy will serve as the primary metric, following its successful application in studies of grammaticalization (Rovai 2008), syncretism (Milizia 2013), and word order (Futrell et al. 2015, Kuboň et al. 2016, Levshina 2019, Merlo, Samo 2022). Derived from observed frequencies in a dataset, entropy captures how predictable or variable a linguistic feature – such as word order – is in each language: higher entropy reflects greater flexibility and unpredictability, while lower entropy indicates fixed patterns.

Our findings will corroborate Meillet's (1924) claim that word order in OCS lacks grammatical significance. Like other ancient Indo-European languages, and notably Old Russian, OCS exhibits high entropy across most analysed parameters, demonstrating substantial syntactic flexibility. In contrast, modern Slavic languages display diachronic trends toward greater rigidity in word order (*pace* Bakker 1998), with shifts ranging from moderate (e.g., Czech, Russian) to pronounced (e.g., Bulgarian).

The extreme syntactic flexibility of OCS, which enables it to replicate Greek word order without producing ungrammatical structures, is further supported by variability in individual texts across specific parameters. This aligns with observations by Turner (2006) and McAnallen (2009) regarding the influence of textual genres. However, the findings also raise concerns about the gradient word order approach proposed by Levshina et al. (2023), as applying such models to free word order languages risks capturing corpus-specific traits rather than characteristics of the language itself.

References

Biagetti, Erica, Inglese, Guglielmo, Zanchi, Chiara, Luraghi, Silvia. 2023. Reconstructing Variation in Indo-European Word Order: A Treebank-Based Quantitative Study. *Language Dynamics and Change*. 13(2), 198–231. https://doi.org/10.1163/22105832-bja10025

Bakker, Dik. 1998. Flexibility and Consistency in Word Order Patterns in the Languages of Europe. In Siewierska, Anna (ed.). *Constituent Order in the Languages of Europe*, Berlin-New York: De Gruyter Mouton, 383–420.

Choi, Hee-Soo, Guillaume, Bruno, Fort, Karën, Perrier, Guy. 2021. Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting. In Mitkov, Ruslan, Angelova, Galia (eds.). *RANLP 2021 – Recent Advances in Natural Language Processing*. INCOMA Ltd, 181–290. https://doi.org/10.26615/978-954-452-072-4_033

Eckhoff, Hanne M., Berdicevskis, Aleksandrs. 2015. Linguistics vs. Digital Editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta*. 14–15, 9–25.

Futrell, Richard, Mahowald, Kyle, Gibson, Edward. 2015. Quantifying Word Order Freedom in Dependency Corpora. In Nivre, Joakim, Hajičová, Eva (eds.). *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala: Uppsala University, 91–100.

Haug, Dag T. T., Jøhndal, Marius L. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In Sporleder, Caroline, Ribarov, Kiril (eds.). *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27–34.

Kuboň, Vladislav, Lopatková, Markéta, Hercig, Tomáš. 2016. Searching for a Measure of Word Order Freedom. In *Proceedings of the 16th ITAT Conference Information Technologies – Applications and Theory*. Bratislava: CreateSpace Independent Publishing Platform, 11–17.

Levshina, Natalia. 2019. Token-based typology and word order entropy. *Linguistic Typology*. 23(3), 533–572. https://doi.org/10.1515/lingty-2019-0025

Levshina, Natalia, Namboodiripad, Savithry, Allassonnière-Tang, Marc, Kramer, Mathew, Talamo, Luigi, Verkerk, Annemarie, Wilmoth, Sasha, Garrido Rodriguez, Gabriela, Gupton, Timothy M., Kidd, Evan, Liu, Zoey, Naccarato, Chiara, Nordlinger, Rachel, Panova, Anastasia, Stoynova, Natalia. 2023. Why we need a gradient approach to word order. *Linguistics*. 61(4), 825–883. https://doi.org/10.1515/ling-2021-0098

McAnallen, Julia. 2009. The competing roles of SV(O) and VS(O) word orders in *Xoždenie igumena Daniila*. *Russian Linguistics*. 33, 211–228. https://doi.org/10.1007/s11185-009-9039-6

Meillet, Antoine. 1924. *Le slave commun*. Paris: Champion.

Merlo, Paola, Samo, Giuseppe. 2022. Exploring T3 languages with quantitative computational syntax. *Theoretical Linguistics*. 48(1–2), 73–83. https://doi.org/10.1515/tl-2022-2032

Milizia, Paolo. 2013. *L'equilibrio nella codifica morfologica*. Roma: Carocci.

Östling, Robert. 2015. Word Order Typology through Multilingual Word Alignment. Zong, Chengqing, Strube, Michael (eds.). *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 2: Short Papers). Beijing: Association for Computational Linguistics, 205–211. https://doi.org/10.3115/v1/P15-2034

Rovai, Francesco. 2008. Mutamento ed Entropia. Un approccio informazionale ai processi di grammaticalizzazione. *Studi e Saggi Linguistici*. 46, 93–114.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*. 27(3), 379–423.

Siewierska, Anna, Rijkhoff, Jan, Bakker, Dik. 1998. Appendix – 12 word order variables in the languages of Europe. Siewierska, Anna (ed.). *Constituent Order in the Languages of Europe*. Berlin-New York: De Gruyter, 783–812.

Turner, Sarah. 2006. Post-verbal subjects in Early East Slavonic. *Transactions of the Philological Society*. 104(1), 85–117. https://doi.org/10.1111/j.1467-968X.2006.00163.x

Zeman, Daniel et al. 2024. *Universal Dependencies 2.15*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. https://doi.org/10.1162/coli_a_00402

ATTRIBUTIVE ADJECTIVE PLACEMENT IN FRENCH, ITALIAN AND SPANISH NEWS ARTICLES: IS THERE A PREPOSABILITY CLINE IN ROMANCE?

Eric Engel

University of Cologne

Keywords: adjective position, syntactic variation, French, Italian, Spanish

Background and research questions. Attributive adjectives in most Romance languages can occur in prenominal or in postnominal positions. Despite the general availability of both syntactic positions, researchers comparing both orders across languages have reported differences as to how much each language makes use of the prenominal position (Delbecque 1990; Radatz 2001; Gulordava, Merlo 2015). Focusing on French, Italian, and Spanish, their results are summarized in (1).

(1) Expected relative frequency of prenominal (*vs.* postnominal) attributive adjectives
Italian > Spanish > French

However, the methodology employed in these studies does not allow us to distinguish between two possible interpretations, which are the focus of this study: Does the order in (1) reflect differences in size of the variable context (Tagliamonte 2012), i.e. the number of adjective-noun combinations that show variable adjective placement? Or are the variable contexts comparable across the three languages, but within the group of potentially variable adjective-noun combinations, the linguistic constraints on the prenominal position differ?

Method. To answer these questions, we extracted all concordances of immediately adjacent adjectives and nouns (using POS sequences) occurring in the news feeds shown in Table 1 in the year 2016 from the Timestamped JSI web corpora 2014–2016 (Bušta et al. 2017) for French, Italian, and Spanish. Those adjective-noun combinations meeting a frequency threshold of 100 tokens (combining prenominal and postnominal position of the adjective) were matched with the closest translation equivalents in the other languages, e.g., French *absence totale/totale absence*, Spanish *ausencia total/total ausencia* and Italian *assenza totale/totale assenza* were grouped under the label TOTAL ABSENCE.

The final sample comprises 347 matched adjective-noun combinations, each occurring at least 100 times in the French, Italian and Spanish data.

Preliminary results and outlook. As a first step, we wanted to know whether the size of the variable context, defined as those adjective-noun combinations that occur at least once in each order, varies as predicted by (1). First results suggest that there is no clear difference between the Italian and the Spanish subsample; however, the same adjective-noun combinations show much more categorical behavior in the French data.

We will next code those adjective-noun combinations that show variable behavior for factors that proved relevant in language-specific analyses of adjective position: the relative length of adjective and noun (Abeillé, Godard 1999; File-Muriel 2006), the type of determiner (Thuilier 2014), and syntactic dependents of the adjective. We plan to analyse the effects of these factors on variable adjective position using mixed-effects modelling, to assess whether cross-linguistic differences in adjective preposing are to be located in the syntax (as language-specific constraint hierarchies) or in the lexicon (as language-specific delimitations of the variable context).

Table 1. News feeds included in the sample

	French	Italian	Spanish
General	liberation.fr	repubblica.it	publico.es
	lemonde.fr	corrieree.it	elpais.com
	lefigaro.fr	ilgiornale.it	elmundo.es
Finance & Economy	lesechos.fr	ilsole24ore.com	eleconomista.es
Sports	lequipe.fr	corrieredellosport.it	marca.com

References

Abeillé, Anne, Godard, Danièle. 1999. French word order and lexical weight. In Borsley, Robert (ed.). *The Nature and Function of Syntactic Categories*. Brill, 325–360. https://doi.org/10.1163/9781849500098_012

Bušta, Jan, Herman, Ondřej, Jakubíček, Miloš, Krek, Simon, Novak, Blaž. 2017. JSI Newsfeed Corpus. In *9th International Corpus Linguistics Conference*. University of Birmingham, 382.

Delbecque, Nicole. 1990. Word order as a reflection of alternate conceptual construals in French and Spanish: Similarities and divergences in adjective position. *Cognitive Linguistics*. 1(4), 349– 416. https://doi.org/10.1515/cogl.1990.1.4.349

File-Muriel, Richard J. 2006. Spanish adjective position: Differences between written and spoken discourse. In Clements, J. Clancy, Yoon, Jiyoung (eds.). *Functional Approaches to Spanish Syntax: Lexical Semantics, Discourse and Transitivity*. Palgrave Macmillan, 203–218.

Gulordava, Kristina, Merlo, Paola. 2015. Structural and lexical factors in adjective placement across Romance languages. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 247–257. https://doi.org/10.18653/v1/K15-1025

Radatz, Hans-Ingo. 2001. *Die Semantik der Adjektivstellung: Eine kognitive Studie zur Konstruktion <Adjektiv + Substantiv> im Spanischen, Französischen und Italienischen*. Berlin: Max Niemeyer Verlag. https://doi.org/10.1515/9783110944396

Tagliamonte, Sali A. 2012. *Variationist Sociolinguistics: Change, Observation, Interpretation*. Wiley-Blackwell. https://doi.org/10.7764/onomazein.28.6

Thuilier, Juliette. 2014. An experimental approach to French attributive adjective syntax. *Empirical Issues in Syntax and Semantics*. 10, 287–304.

DERIVATIONAL PRODUCTIVITY IN A DIACHRONIC PERSPECTIVE: A CORPUS-BASED STUDY OF ENGLISH SUFFIXES

Tamás Fekete

University of Pécs

Keywords: productivity, derivation, suffixation, diachronic corpora, language change

It is a generally held view corroborated by different treatments of morphological productivity that type frequency, especially the frequency of hapax legomena, is a reliable indicator of how productive a given word-formation method is (e.g., Baayen, Lieber 1991). My aim in this paper is to analyse the productivity of derivational suffixes from a diachronic perspective, utilizing four different corpora, corresponding to four different historical periods of English. With this paper, I seek to answer the following research questions:

1) How did the productivity of English derivational suffixes change from Old to Middle to Early Modern and finally to Present-Day English?
2) What patterns are observable in the frequencies of derivational suffixes between each historical period and how do these frequencies correlate with productivity?

In this exploratory study, on the one hand, each period is treated as one entity to be compared with another, while on the other hand, analysis is also split into 50-year segments for a more fine-grained assay (cf. Barðdal et al 2024) to mitigate the potential pitfalls of a monolithic approach. The analysis is expected to yield results that provide a more nuanced interpretation of diachronic productivity.

The analysis is split into the following historical periods of English and relies on the following corpora.

Old English period:

- York-Toronto-Helsinki Parsed Corpus of Old English Prose (1.5 million words)
- Corpus of poetry collected by the author from the *Old English Poetry in Facsimile* (OEPF 3.0, Foys et al 2019) website (160 thousand words)

Middle English period:

- Middle English corpus (5 million words) comprising works of prose and poetry, collected by the author from the *Middle English Text Series* website

Early Modern English period:

- Corpus of Early English Correspondence (5 million words)

Present-Day English period:

- The Baby version of the British National Corpus (4 million words)

Each period is represented by a separate corpus of about 5 million words, except for the Old English period, where text availability is the limiting factor of corpus size. The Baby version of the BNC (instead of the whole 100-million-word edition) was chosen specifically for reasons of comparability. Frequency values are extracted with the AntConc concordancing software (version 4.3.1) and are normalized to 1 million words. Statistical analysis is carried out in SPSS version 27.

Frequency data is analysed quantitatively, focusing on three factors: concentration, frequency and diversity. For concentration and frequency, the Pareto and Zipf distributions are analysed, respectively, hypothesizing that suffixes with greater productivity follow the Pareto and Zipf distributions more closely. Diversity is measured via Shannon's entropy which accounts for the probability with which a given suffix is likely to occur. Higher entropy values are hypothesized to correspond to higher productivity. Formal and semantic transparency, as well as preference for native or foreign bases is also factored into the analysis (cf. Palmer 2014).

References

Barðdal, Jóhanna, Renata Enghels, Quentin Feltgen, Sven Van Hulle & Peter Lauwers. 2024. Productivity in Diachrony. In Ledgeway, Adam, Aldridge, Edith, Breitbarth, Anne, Kiss, Katalin É, Salmons, Joseph, and Simonenko, Alexandra (eds.). *Wiley-Blackwell Companion to Diachronic Linguistics*, edited by. Oxford: Wiley-Blackwell.

Baayen, Harald, & Rochelle Lieber. 1991. Productivity and English derivation: a corpus-based study. *Linguistics*. 29(5), 801–843.

Corpus of Early English Correspondence. 1998. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin at the Department of Modern Languages. Helsinki: University of Helsinki.

Foys, Martin et al. (eds.). 2019. *Old English Poetry in Facsimile 3.0* (Center for the History of Print and Digital Culture, University of Wisconsin-Madison). https://doi.org/10.21231/t6a2-jt11

Palmer, Chris C. 2014. Measuring productivity diachronically: nominal suffixes in English letters, 1400–1600. *English Language and Linguistics*. 19(1), 107–129. https://doi.org/10.1017/S1360674314000264

The BNC Baby, version 2. 2005. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at: http://www.natcorp.ox.ac.uk/

The Middle English Texts Series (website). 2024. The Rossell Hope Robbins Library at the University of Rochester. Available at: https://metseditions.org

The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE). 2003. Oxford Text Archive. Available at: http://hdl.handle.net/20.500.12024/2462

CORPUS-BASED ANALYSIS OF LEXICO-GRAMMATICAL PATTERNS AND ENGAGEMENT MARKERS IN ELF FOR MEDICAL PROFESSIONALS: BRIDGING THEORY AND PRACTICE

Olga Freimane

University of Latvia

Keywords: corpus-based approach, lexico-grammatical patterns, English as a Lingua Franca, medical professionals, interactional competence

This study examines the use of lexico-grammatical patterns and engagement markers in English as a Lingua Franca (ELF) among Latvian medical professionals during monologic interactions in formal spoken discourse. The research adopts a corpus-based approach (McEnery, Hardie 2012) to identify recurring linguistic features across the discourse. By focusing on lexico-grammatical commonalities and engagement markers, the study bridges theoretical understanding with practical applications for improving interaction in multilingual, high-stakes environments. The analysis draws on a self-compiled specialised corpus created for this study, consisting of 60 transcriptions (200 099 words) across three professional genres: scientific conference presentations, public discussions, and professional interviews. These were sourced from freely accessible institutional platforms and transcribed using the AI-powered tool *Notta*. To ensure relevance, only recent samples (2019–2025) were included. The speakers include practicing and future medical professionals, postgraduate residents, and specialists in related health fields. Access to the specialised corpus is restricted due to ethical and data protection considerations. The study is guided by Seidlhofer's (2004) framework on ELF lexico-grammatical commonalities and Hyland's (2005) model of engagement markers, enabling a systematic evaluation of language use in diverse professional contexts. Although limited to Latvian medical professionals, the findings provide a foundation for broader cross-contextual research in ELF medical communication. The findings reveal key patterns in the use of lexico-grammatical features and engagement markers that shape interaction in ELF medical discourse. Common lexico-grammatical features identified among Latvian medical professionals include bulky or unclear phrasing, invariant word order, non-standard noun phrase formation, and

frequent use of fillers. These features may compromise the precision and professionalism expected in formal medical settings. At the same time, the study highlights specific engagement markers, such as listener pronouns, directives, questions, shared knowledge references, and personal asides, which are critical tools for enhancing interaction in monologic interactions. These markers facilitate audience engagement, promote understanding and bridge linguistic gaps, especially in ELF contexts where listeners may have varying levels of proficiency. Within the Latvian context, engagement markers are unevenly distributed across the genres, which can limit or distort opportunities for effective interaction in ELF medical discourse.

The practical implications of this research are significant for both educators and practitioners. The findings provide a foundation for developing targeted English language training programmes, focusing on lexical and grammatical precision and audience engagement. By addressing the specific needs of medical professionals, these programmes can enhance monologic interactional competence, boost confidence and improve the overall effectiveness of multilingual interaction. This study contributes to a broader understanding of ELF's role in global professional discourse, demonstrating how corpus-based approaches can inform practical solutions for enhancing communication in diverse, interdisciplinary fields such as medicine.

The participation in the conference is funded within the project "Internal and External Consolidation of the University of Latvia" (No. 5.2.1.1.i.0/2/24/I/CFLA/007).

References

Hyland, Ken. 2005. *Metadiscourse: Exploring Interaction in Writing.* London: Continuum.

McEnery, Tony, Andrew, Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Seidlhofer, Barbara. 2004. Research perspectives on teaching English as a Lingua Franca. *Annual Review of Applied Linguistics.* 24, 209–239. https://doi.org/10.1017/S0267190504000145

CORPUS-BASED TRACKS TO THE PAST: DIACHRONIC INSIGHTS INTO THAT/ZERO COMPLEMENTATION OF THE VERB 'THINK'

Vassiliki Geka

National & Kapodistrian University of Athens

Keywords: zero/that complementation, 'think', diachrony, CxG

Drawing on diachronic corpus evidence (COHA) spanning Late Modern English (ca. 1820) to Present-Day English (ca. 2010), the paper makes a case for the contribution of corpora to the study of syntactic variation through time and through the lens of Construction Grammar (Traugott 2005; Traugott, Trousdale 2013; Barðdal et al. 2015). In particular, it examines the variation that the verb 'think' exhibits between *that* and *zero* complementation, as in 1–2.

(1) *"The thunder has fallen, at last, and I think that I shall be in heaven or hell, before that clock strikes...'*
(COHA, FIC: Logan: A Family History, Volume 2, 1822)

(2) *"I think she's an intense young woman. Smart and aloof..."*
(COHA, FIC: Analog Science Fiction & Fact, 2010)

The two variants and their constructional differences are empirically examined both qualitatively and quantitatively based on 7200 random tokens collected across the 190-year spectrum specified above. Following stratified sampling per decade (1820–2010) and genre, the data, mined exclusively from fiction texts, are first annotated with respect to different contextual, matrix and complement clause related parameters. These include – *inter alia* – verbal morphology, subject correreferentiality, pronominality, pre- and post-subject intervening linguistic material, harmony of polarity and contemporality. The annotated data are then subjected to statistical analysis through Pearson's test, linear regression and stepwise logistic regression with a double aim: (a) to identify statistically significant correlations holding between the data and the annotation parameters and (b) to examine whether – as entertained in the literature (e.g., Thompson, Mulac 1991; Rissanen 1991; Finegan, Biber 1995; Diessel, Tomasello 2001; Shank et al. 2016a, 2016b; Shank, Plevoets 2018) – the annotation parameters investigated could possibly have a foretelling/predictive (or possibly conditioning) function for the emergence of each variant.

In light of the evidence collected, the study ultimately argues for the benefits of applying a corpus-based, constructionist approach to the analysis of syntactic variation phenomena while it addresses the following research questions: (a) in what ways can corpora, as principled collections of texts, provide important insights into the developmental trajectory of language patterns, (b) what meaning differences can be detected between the constructional variants despite their partial overlap (cf. minimal constructional synonymy (Goldberg 1995)), (c) whether and to what extent diachronic trends favouring each alternative can be empirically established; and (d) what morpho-syntactic, contextual, and/or other factors may jointly or independently motivate each variant and to which degree their presence in discourse may be considered predictive.

References

Barðdal, Jóhanna, Smirnova, Elena, Sommerer, Lotte, & Gildea, Spike (eds.). 2015. *Diachronic Construction Grammar*. Amsterdam: John Benjamins. https://doi.org/10.1075/cal.18

Diessel, Holger, Tomasello, Michael. 2001. The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics*. 12(2), 97–141. https://doi.org/10.1515/cogl.12.2.97

Finegan, Edward, Biber, Douglas. 1995. That and zero complementizers in late Modern English: Exploring ARCHER from 1650–1990. In Aarts, Bas, Meyer, Charles F. (eds.). *The Verb in Contemporary English*. Cambridge: Cambridge University Press, 241–257.

Goldberg, Adele. 1995. *A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Rissanen, Matti. 1991. On the history of that/zero in object clause links in English. In Aijmer, Karin, Altenberg, Bengt (eds.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 272–289. https://doi.org/10.4324/9781315845890

Shank, Carl, Van Bogaert, Joost, Plevoets, Koen. 2016a. The diachronic development of zero complementation: A multifactorial analysis of the that/zero alternation with *think*, *suppose*, and *believe*. *Corpus Linguistics and Linguistic Theory*. 12(1), 31–72. https://doi.org/10.1515/cllt-2015-0074

Shank, Carl, Plevoets, Koen, Van Bogaert, Joost. 2016b. A multifactorial analysis of that/zero alternation: The diachronic development of the zero complementiser with *think*, *guess*, and *understand*. In Yoon, Jae Jung, Gries, Stefan Th. (eds.). *Corpus-Based Approaches to Construction Grammar*. Amsterdam: John Benjamins, 201–240. https://doi.org/10.1075/cal.19.08sha

Shank, Carl, Plevoets, Koen. 2018. Investigating the impact of structural factors upon that/zero complementizer alternation patterns in verbs of cognition: A diachronic corpus-based multifactorial analysis. *Research in Corpus Linguistics*. 6, 83–112. https://doi.org/10.32714/ricl.06.07

Thompson, Sandra, Mulac, Anthony. 1991. The discourse condition for the use of complementizer *that* in conversational English. *Journal of Pragmatics*. 15, 237–251. https://doi.org/10.1016/0378-2166(91)90012-M

Traugott, Elizabeth, Closs. 2005. Lexicalization and grammaticalization. In Cruse, D. Alan, Hundsnurscher, Franz, Job, Michael, & Lutzeier, Peter R. (eds.). *Lexikologie/ Lexicology*. 2, 1702–1712. Berlin: Walter de Gruyter.

Traugott, Elizabeth C., Trousdale, Graeme. 2013. *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.

ON POLISH GRADATION

Rafał L. Górski

Jagiellonian University; Institute of Polish Language PAS

Keywords: morphology, Polish language, gradation

Polish – as many Indo-European languages – has two means of marking the degree: the morphological comparative with suffix (*-szy/-ejszy*) attached to the adjective stem (e.g., *bogatszy* 'richer') and the periphrastic comparative with *bardziej* in front of adjectives (e.g., *bardziej bogaty* 'more rich', also labelled as "the analytic comparative"). Similarly superlative is derived from comparative by a prefix *naj-* attached either to the adjective itself or (in case of periphrastic comparative) to *bardziej*, resulting with the forms *najbogatszy*, *najbardziej bogaty*. However, the morphological gradation is limited to a small number of adjectives, for all the other lexemes the periphrastic form is the only option.

If a particular adjective admits both markers, the periphrastic one is rarely chosen. Still, the number of occurrences of such forms in texts prompts normative linguists to deplore their spread (Buttler et al. 1986, Jadacka 2005). The factors which block morphological degree are discussed in a number of papers (e.g., Dick 1976). In contrast, the only attempt to find out what shapes the choice between the two forms when both are available is found in Gębka-Wolak (2017), nevertheless, it seems that the paper overlooked some factors.

The aim of this presentation is to explore the factors which increase the chances that the speaker would use the periphrastic form rather than the (available) morphological one. These factors are:

1. Overall frequency of the lemma in the balanced subcorpus of NCP (regardless of its degree).
2. Its frequency in comparative and superlative.
3. The proportion between the positive and comparative+superlative.
4. Word forming morpheme; simplicia (or non-derived adjectives) a separate class among this category.
5. Number of syllables of the positive nominative singular.
6. The allomorph, which marks the degree: either *-sz-* or *-ejsz-*.
7. The degree (comparative vs. superlative).
8. The grammatical case.
9. Modifications of the stem in comparative and superlative.

10. Apart from (9), full suppletivism was marked in a separate column.

11. Text-type in which the particular hit occurs.

The data were drawn from the balanced subcorpus of the National Corpus of Polish. Only these adjectives were considered, which allow for both morphological and periphrastic gradation. The full dataset contains 1 053 306 occurrences of comparative and superlative degree, of which only 5% represents the periphrastic variant. Since the data were so heavily biased, the sample with morphological degree was downsampled to the same size as the periphrastic one. This dataset was then analysed via logistic regression. The obtained results show, *inter alia*, that with the increase of the frequency of adjectives the probability of occurring with periphrastic gradation drops, or in other words the latter is avoided with frequent adjectives. In turn, factors which favour the periphrastic gradation are nominative case (in opposite to all oblique cases), the length of the adjective, finally comparative (in contrast to superlative), the *-ejszy* allomorph and the modifications of the stem caused by the morpheme of the degree. The word forming morphemes affect the choice to various degrees, the same can be said about text-types.

The Nagelkerke pseudo-R^2 for this model is 0.85. The model predicts the choice with high accuracy (>0.9) and the agreement between the predicted and actual values is substantial (Cohen's kappa >0.8).

References

Buttler, Danuta, Kurkowska, Halina, Satkiewicz, H. 1986. *Kultura języka polskiego. Zagadnienia poprawności gramatycznej.* Warszawa: Państwowe Wydawnictwo Naukowe.

Dick, J. H. 1976. Stopniowanie przymiotników © przysłówków we współczesnej polszczyźnie kulturalnej. In Orzechowska, H. (ed.). *Zagadnienia kategorii stopnia w językach słowiańskich*. Warszawa, 43–52.

Gębka-Wolak M. 2017. Syntetyczny model stopniowania przymiotników we współczesnej polszczyźnie (uwagi na podstawie badania ilościowego), *Язык Сознание Коммуникация* 55, 82–97.

Jadacka, Hanna. 2005. *Kultura języka polskiego. Fleksja, słowotwórstwo, składnia*. Warszawa: Wydawnictwo Naukowe PWN.

NON-CANONICAL GRAMMAR AND VERY LARGE CORPORA: A CASE STUDY OF GERMAN ADPOSITIONS

Stefan Hartmann

Heinrich Heine University Düsseldorf

This paper investigates non-canonical uses of the German circumposition um ... willen as well as the postposition wegen. Both adpositions usually govern the genitive case, although there is quite some variation especially in the case of wegen. (1) illustrates typical uses of both adpositions. However, in non-standard data, we also find numerous cases like (2), where the genitive morpheme -s is "relocated" to the postposition.

(1) um des Frieden-s willen / des Frieden-s wegen
for the peace-GEN sake for peace-GEN sake

(2) um des Frieden willen-s / des Frieden wegen-s
for the peace sake for peace sake

The present study investigates these non-canonical uses of um...willen(s) and wegen(s) based on data from the 20-billion-word webcorpus DECOW16B (Schäfer, & Bildhauer 2012; Schäfer 2015), testing the hypothesis that the drive towards cleft-formation, which Nübling et al. (2017, 117) see as the most important syntax-typological feature of German, is among the main motivations for the displacement of the s-morpheme, along with analogical inferences and frequency effects. If the hypothesis is correct, we would expect (a) that we find significantly more (masculine and neuter) nouns from inflection classes that show a genitive-s in the wegens- and um ... willens data than in comparison datasets with the canonical variants; (b) that we find a significantly higher proportion of non-canonical s-less genitives in the wegens- and um ... willens-data than in the comparison datasets: while s-omission is quite common especially in the case of low-frequency words, loan words, and proper nouns (Zimmer 2018), we can predict that it occurs more frequently in combination with the non-canonically s-suffixed postpositions. Drawing on an exhaustive search for wegens and um ... willens and samples of 5000 attestations each for the canonical forms, binomial regression models as well as CART trees and random forests (Tagliamonte & Baayen 2012) were used to test the above-mentioned predictions. In line with prediction (a), the proportion of feminine nouns is much lower in the case of the non-canonical forms. In simple binomial mixed regression models with "Gender" as binary response variable, "Variant" as predictor variable, and "lemma" as random variable, "Variant" emerges as a highly significant predictor both for um...willen(s) and for wegen(s). Turning to prediction (b), the random forest models show that for both constructions, the variant makes a clear difference for the presence or absence of s-less genitives. In the case of um ... willen(s), the variables that proved most influential in Zimmer (2018) emerge as significant predictors of s-lessness in the canonical variant. For the non-canonical variant, by contrast, only frequency makes a difference. This is reflected in the measure of conditional permutation variable importance (Strobl et al. 2008): here, "Variant" emerges as the most significant predictor by far. In sum, then, the results lend support to the hypothesis that the principle of cleft formation plays a major role in the relocation of the genitive *-s*.

References

Nübling, Damaris, Dammel, Antje, Duke, Janet & Renata Szczepaniak. 2017. *Historische Sprachwissenschaft des Deutschen: eine Einführung in die Prinzipien des Sprachwandels.* 5th ed. Tübingen: Narr.

Schäfer, Roland. 2015. Processing and querying large corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen & Andreas Witt (eds.). *Challenges in the Management of Large Corpora (CMLC-3)*, 28–34. Available at: http://corpora.ids-mannheim.de/cmlc.html

Schäfer, Roland & Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Terry, Doğan, Mehmet Uğur, Maegaard, Bente, Mariani, Joseph, Moreno, Asuncion, Odijk, Jan & Piperidis, Stelios (eds.). *Proceedings of LREC 2012*, 486–493.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics*. 9(307). https://doi.org/10.1186/1471-2105-9-307

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*. 24(2), 135–178. https://doi.org/10.1017/S0954394512000129

Zimmer, Christian. 2018. *Die Markierung des Genitiv(s) im Deutschen: Empirie und theoretische Implikationen von morphologischer Variation.* Berlin, Boston: De Gruyter. https://doi.org/10.1515/zrs-2020-2032

THE HISTORY OF NEGATION IN FRISIAN AND DUTCH: A CORPUS-BASED COMPARISON OF JESPERSEN'S CYCLE AT THE NORTH SEA COAST

Daniel Hrbek

Osnabrück University

Keywords: Old Frisian, Dutch, Jespersen's Cycle, historical linguistics, comparative Germanic corpus linguistics

Since Proto-Germanic times, there have been many syntactic developments that characterise modern West Germanic dialects – including (sentential) negation. As the examples from Old Frisian (1) and Old/Middle Dutch (2) show, the original Germanic negative element was a preverbal particle *ni*/*ne* (a), which later – due to phonological weakening – had to be reinforced (b) with a postverbal particle OFri. *nāwet*/ODt. *niewiht*, Late OFri. *naet*/MDt. *niet* (< PGmc. **ni io uuiht* 'not a thing'). Ultimately, *ne* was dropped completely (c), so that all modern West Germanic languages now only have a single, postverbal negative particle. This phenomenon is known as *Jespersen's Cycle* (Jespersen 1917), with stage II, the so-called bipartite negation, being particularly prominent. Although both negative elements occur simultaneously, they do not cancel each other out; rather, they are only able to reverse the proposition in combination.

In recent years, urgently needed research has been done for the history of negation in High (e.g., Jäger 2008) and Low German (Breitbarth 2014). However, even if parts of the Continental West Germanic dialect continuum now can be considered well-studied, there is still a lack of a comprehensive overview that analyses this highly conspicuous pattern in a cross-lingual and comparative way. This is aggravated by the fact that only minor studies exist for the other two languages, Dutch (Vosters & Vandenbussche 2012; Zeijlstra 2002) and Frisian; in the case of the latter, there is only a single overview (Bor 1990) that that does not meet modern requirements. Hereby, I would like to bridge this gap and take a closer look at the change of negation in the closely related North Sea Germanic languages Frisian and Dutch.

For this purpose, the development of sentential negation is analysed not only diachronically, but also (as far as possible) diatopically, to gain an overview of the spatio-temporal spread of this phenomenon. Other factors associated with this cycle (e.g., prefixiation and position of the finite verb) will also be included, so that a comparison with the well-researched dialects of High and Low German can be made, leading to a better insight into this process and its areal/dialectal realisation. In the case of (Old) Frisian, this will also provide the first (simultaneously large-scale) empirical evidence ever that Jespersen's Cycle occurred there at all.

Modern corpora such as *Brieven als buit* and *Corpus Oudnederlands* (Dutch), as well as *Corpus Oudfries* (Frisian), which are balanced according to various criteria, serve as the source of data for this, enabling a comparable analysis of the two languages, their historical stages and dialects. Therefore, I will also present what a corpus linguistic methodology that allows a comparison across languages and language stages might look like during my talk. In doing so, I not only want to break a lance for the (chronically) under-researched languages Old Dutch and in particular Old Frisian (e.g., Hrbek in press), but also share my experience in the field of comparative historical corpus linguistics, which is also committed to historical dialectology (cf. Wiesinger 2017).

(1) a. *and nammermar* ***ne*** *mot hi anda godis*
and nevermore NEG may he in God's
huse wesa mith ore kerstene lioden
house be with other Christian people
First Rüstring Manuscript; XVII.6 *(On Killing a Relative)*

b. *Ief hi dan* ***naet*** *komma* ***ne*** *welle*
if he then NEG come.INF NEG wants
Jus Municipale Frisonum; III.57,6 (*Elder Skeltariucht*)

c. *Jsrahel, dines Godes nama scheltu* ***naet*** *wrswerra*
Israel yours God's name shall=you NEG take in vain
Jus Municipale Frisonum; II.8d (*Haet is riucht? What is law?*)

(2) a. *minon eygenen wingardon* ***ne****mochte ich behoodan*
my own vineyard NEG=could I cultivate
Old Dutch (*Leiden Willeram*; f. 16v)

b. *Want ic* ***ne*** *wille* ***niet****, broeder, dat ghi onwetende sijt*
because I NEG want NEG brother that you unknowing are
Early Middle Dutch (14th c.) (*Lectionarium van Amsterdam*; 40)

c. *want menne mach Gode* ***niet*** *deylen*
because drove may God NEG divide
Late Middle Dutch (16th c.) (*Gheestelike brulocht*; 1,206)

References

Bor, Arie. 1990. The use of the negative adverbs *ne* and *nawet* in Old Frisian. In Bremmer, Rolf Hendrik, van der Meer, Geart, Vries, Oebele (eds.). *Aspects of Old Frisian Philology*. Amsterdam: Brill, 26–41.

Breitbarth, Anne. 2014. *The History of Low German Negation*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199687282.001.0001

Hrbek, Daniel. Forthcoming. Negationspartikeln auf Abwegen: *ne* zwischen Exzeption, Koordination und Negation. Ein Versuch der Disambiguierung am Beispiel des Altfriesischen. In Schnee, Lea, Huber, Henriette, Hartmann, Stefan (eds.). *Historische Grammatik(en): Sprachwandelforschung zwischen Norm und System*. Berlin, Boston: De Gruyter.

Jäger, Agnes. 2008. *History of German Negation*. Amsterdam, Philadelphia: Benjamins. https://doi.org/10.1075/la.118

Jespersen, Otto. 1917. *Negation in English and Other Languages*. Kopenhagen: Høst & Søn.

Vosters, Rik, Vandenbussche, Wim. 2012. Bipartite Negation in 18th and Early 19th Century Southern Dutch: Sociolinguistic Aspects of Norms and Variation. *Neuphilologische Mitteilungen*. 113(3), 343–364.

Wiesinger, Peter. 2017. *Strukturelle historische Dialektologie des Deutschen. Strukturhistorische und strukturgeographische Studien zur Vokalentwicklung deutscher Dialekte*. Hildesheim: Georg Olms Verlag.

Zeijlstra, Hedde. 2002. What the Dutch Jespersen Cycle may reveal about Negative Concord. In Alexiadou, Artemis, Fischer, Susann, Stravrou, Melita (eds.). *Papers from the Workshop "Languages Change from a Generative Perspective"*. Potsdam: Universitätsbibl https://doi.org/iothek Potsdam, 183–206.

DERIVATIONAL RELATIONSHIPS OF BORROWED DENOMINAL PERSONAL NOUNS IN CONTEMPORARY LITHUANIAN

Lina Inčiuraitė-Noreikienė & Erika Rimkutė

Vytautas Magnus University

Keywords: derivational relationship, denominal personal nouns, simplex loanword, complex loanword, contemporary Lithuanian

Over the past decade, a growing interest in the interplay between word formation and borrowing has been observed (cf. Panocová 2015; ten Hacken, Panocová 2020). The adaptation of material borrowings across languages has been extensively studied (cf. Matras, Sakel 2007; Haspelmath, Tadmor 2009; Wohlgemuth 2009; Gardani et al. 2014; Ralli 2016). However, these studies have predominantly focused on how loanwords adapt to the phonological and morphological system of the recipient language rather than on their derivational relationships.

The following two research questions are posed: 1) How are derivational relationships established among suffixed borrowed denominal personal nouns in contemporary Lithuanian? 2) Do complex denominal personal nouns have a lower token frequency than their corresponding simplex counterparts in the Corpus of Contemporary Lithuanian Language (CCLL)? Borrowed nouns belong to different categories of word formation, such as personal names, instruments, place nouns, etc. Denominal personal nouns have been chosen because they represent one of the major categories in derivational morphology, naming people according to their roles, professions, affiliations or characteristics. We use Seifart's (2015, 513) three criteria to determine derivational relations between loanwords and examine their frequency in the CCLL:

1. A group of complex loanwords exists, each containing a borrowed suffix and sharing a common, identifiable meaning component, e.g., a set of words with the same suffix denotes personal nouns, e.g., *afer-ist-as, -ė* 'fraudster', *masaž-ist-as, -ė* 'masseur', *skandal-ist-as, -ė* 'scandalist', etc.
2. A set of loanword pairs can be identified, one with the affix and one without, showing regular and clearly discernible changes in meaning. These pairs typically consist of a simplex loanword and its complex counterpart. The former refers to actions, processes, events, etc., whereas

the latter denotes personal nouns, e.g., *afer-ist-as, -ė* 'fraudster' ⇠ *afer-a* 'fraud', *masaž-ist-as, -ė* 'masseur' ⇠ *masaž-as* 'massage', *skandal-ist-as, -ė* 'scandalist' ⇠ *skandal-as* 'scandal', etc.

3. In pairs of complex and corresponding simplex loanwords, complex loanwords tend to have a lower frequency of use than their simplex counterparts. For example, the token frequency of *aferistas* 'fraudster' in the CCLL is 929, while the token frequency of *afera* 'fraud' is 1,105. The former occurs less frequently than the latter.

The research relies on data from the electronic Dictionary of Internationalisms *Interleksis* (DI), the Dictionary of Contemporary Lithuanian (DCL), the Dictionary of Standard Lithuanian (DSL), the Dictionary of Lithuanian (DL), the Database of Lithuanian Neologisms (DLN) and the Corpus of Contemporary Lithuanian Language (CCLL). The synchronic approach to the derivation of borrowed suffixed personal nouns has been adopted in the research. It is planned to analyse approximately 200 to 300 borrowed denominal personal nouns.

The expected results of the study are that derivational relationships between borrowed personal nouns can be established. Complex personal nouns contain borrowed suffixes, identifiable by sharing the same root with the corresponding simplex nouns (cf. Stundžia, Inčiuraitė-Noreikienė 2023, 58). More specifically, complex personal nouns exhibit both formal and semantic motivation, which is why they are included in the description of the Lithuanian word-formation system (see Stundžia 2016, 3094; Urbutis 1965, 409; Urbutis 2005, 138). Additionally, the study anticipates that complex denominal personal nouns will have lower token frequencies than their corresponding simplex counterparts in the CCLL.

Acknowledgements

This project has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-PD-24-170.

References

Gardani, Francesco, Arkadiev, Peter, Amiridze, Nino (eds.). 2015. *Borrowed Morphology*. Berlin, Boston, Munich: De Gruyter Mouton. https://doi.org/10.1515/9781614513209

Haspelmath, Martin, Tadmor, Uri (eds.). 2009. *Loanwords in the World's Languages: A Comparative Handbook*. Berlin, New York: Mouton de Gruyter.

Matras, Jaron, Sakel, Janete. 2007. *Grammatical Borrowing in Cross-linguistic Perspective*. Berlin: Mouton de Gruyter.

Panocová, Renáta. 2015. *Categories of Word Formation and Borrowing: An Onomasiological Account of Neoclassical Formations*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Pius ten Hacken, Panocová Renáta (eds.). 2020. *The Interaction of Borrowing and Word Formation*. Edinburgh: Edinburgh University Press. https://doi.org/10.1515/9781474448215

Ralli, Angela. 2016. Strategies and patterns of loan verb integration in Modern Greek varieties, in Ralli, Angela (ed.). *Contact Morphology in Modern Greek dialects*, Cambridge: Cambridge Scholars Press, 73–108.

Seifart, Frank. 2015. Direct and indirect affix borrowing. *Language*. (91)3, 511–533.

Stundžia, Bonifacas. 2016. Lithuanian. In Müller, Peter O., Ohnheiser, Ingeborg, Olsen, Susan, Rainer, Franz (eds.). *Word-Formation. An International Handbook of the Languages of Europe*. 5, Berlin: De Gruyter Mouton, 3089–3106.

Stundžia, Bonifacas, Inčiuraitė-Noreikienė, Lina. 2023. Exploring competing patterns in morphological derivation: The case of personal and agent nouns with borrowed roots in contemporary Lithuanian, *Skase Journal of Theoretical Linguistics*. 20(2), 56–78.

Urbutis, Vincas. 1965. Daiktavardžių daryba, in Ulvydas, Kazys (ed.). *Lietuvių kalbos gramatika I: Fonetika ir morfologija*. Vilnius: Mintis, 251–473.

Urbutis, Vincas 2005. Daiktavardžių daryba, in Ambrazas, Vytautas (ed.). *Dabartinės lietuvių kalbos gramatika*, 4-oji pataisyta laida, Vilnius: Mokslo ir enciklopedijų leidybos institutas, 86–167.

Wohlgemuth, Jan. 2009. *A typology of verbal borrowings* (Trends in Linguistics. Studies and Monographs 211). Berlin, New York: Mouton de Gruyter.

Sources

CCLL – Corpus of Contemporary Lithuanian Language, Kaunas: Vytauto Didžiojo universitetas, 2021. Available at: corpus.vdu.lt (accessed 2024-12-30).

DCL – Keinys, Stasys (ed.). Dabartinės lietuvių kalbos žodynas [Dictionary of Contemporary Lithuanian], 8th edition (revised and updated) Vilnius: Lietuvių kalbos institutas, 2021, electronic version, 2021. Available at: https://ekalba.lt/dabartines-lietuviu-kalbos-zodynas (accessed 2024-12-30).

DI – Kinderys, Algimantas (ed.). Kompiuterinis tarptautinių žodžių žodynas [Computerised Dictionary of Internationalisms] *Interleksis*, Vilnius: Alma littera, Fotonija, 2003.

DL – Lietuvių kalbos žodynas [Dictionary of Lithuanian], 1–20 (1941–2002): electronic version, Naktinienė, Gertrūda (ed.). Vilnius: Lietuvių kalbos institutas, 2005 (updated 2018). Available at: www.lkz.lt (accessed 2024-12-30).

DNL – Miliūnaitė, Rita, Aleksaitė, Agnė. Lietuvių kalbos naujažodžių duomenynas [the Database of Lithuanian Neologisms] [Continuous Online Reference since 2011 m.], compiled by Rita Miliūnaitė, Vilnius: Lietuvių kalbos institutas. available at: https://ekalba.lt/naujazodziai/ (accessed 2024-12-30).

DSL – Liutkevičienė, Danutė (ed.). Bendrinės lietuvių kalbos žodynas [Dictionary of Standard Lithuanian]. Available at: https://ekalba.lt/bendrines-lietuviu-kalbos-zodynas (accessed 2024-12-30).

IDENTIFICATION OF ADJECTIVIZED AND ADJECTIVAL LITHUANIAN PARTICIPLES BASED ON A CORPUS LINGUISTICS METHODOLOGY

Laima Jancaitė-Skarbalė

Vytautas Magnus University

Keywords: adjectivized participles, adjectival participles, corpus linguistics, grammar, Lithuanian language

The participles of the Lithuanian language are declined forms of the verb, but their usage differs significantly not only from conjugated and other verb forms but also from each other. Based on their functions and meanings, participles can be divided into three groups: adjectivized, adjectival, and verbal.

Adjectivized participles are participles that retain the participial form but function as adjectives and carry the meaning of an adjective. These have been discussed by Lithuanian linguists such as Paulauskienė (1994), Gaivenis and Keinys (1990, 40), and Ambrazas (1979), as well as by foreign researchers, including Petrunina (2021). Examples include *nepamirštamas* 'unforgettable', *prieinamas* 'accessible', and *tinkamas* 'suitable'. Adjectivized participles often shift their lexical meaning from the corresponding verbs, lose the voice and tense categories typical of participles, and acquire other adjectival features.

Adjectival participles, while functioning as adjectives and conveying adjectival meaning, remain semantically close to their base verbs. E.g., the participle *miręs* 'dead' is the antonym of the adjective *gyvas* 'alive', but it is also a form of the verb *mirti* 'to die' (*miręs žmogus* 'a person who has died').

Verbal participles are those that do not have an adjectival meaning (e.g., they are not synonymous or antonymous with adjectives). Examples include *naudojamas* 'used', *ieškantis* 'searching', and *pamatęs* 'seen'.

It is important to identify adjectivized and adjectival participles both theoretically and practically, for tasks such as writing dictionaries and grammars, teaching Lithuanian as a foreign language, annotating corpora morphologically and syntactically, and conducting various linguistic studies.

In this study, adjectivized and adjectival participles were identified using the following criteria: 1) lexical-grammatical criteria (change in lexical meaning from the corresponding verb; synonyms and antonyms with adjectives; frequent attributive function and absence of verbal arguments; compatibility with adverbs of measure/degree and gradation (e.g., *mylimas* 'beloved' – *mylimiausias* 'most beloved'); frequent occurrence in the definite form); 2) derivational criteria (formation of adverbs with the suffix *-ai* or abstract nouns with the suffixes *-umas* and *-ybė* from participles); 3) quantitative criteria (frequent use of participles in corpora compared to other verb forms).

These criteria were identified based on the work of other linguists and the analysis conducted by the author of this presentation. They were applied using corpus data. Criteria that more reliably help identify adjectivized and adjectival participles are considered more important, while others are considered less important. The study assumes that the more essential criteria a participle meets, the more likely it is to be adjectivized.

The study analysed 288 participles derived from the 200 most frequent verbs in the Lexical Database of Lithuanian Language Usage (https://kalbu.vdu.lt/mokymosi-priemones/leksikonas/), e.g., *mylėti* 'to love', *žinoti* 'to know', *eiti* 'to go'. These participles were analysed in the Pedagogic Corpus of Lithuanian (https://kalbu.vdu.lt/mokymosi-priemones/mokomasis-tekstynas/) and the Corpus of the Contemporary Lithuanian Language (http://tekstynas.vdu.lt/tekstynas/). After conducting the study, 53 (18.4%) adjectivized and adjectival participles were identified.

References

Ambrazas, Vytautas. 1979. *Lietuvių kalbos dalyvių istorinė sintaksė*, Vilnius: Mokslas.

Gaivenis, Kazimieras & Stasys Keinys. 1990. *Kalbotyros terminų žodynas*, Kaunas: Šviesa.

Paulauskienė Aldona. 1994. *Lietuvių kalbos morfologija: paskaitos lituanistams*. Vilnius: Mokslo ir enciklopedijų leidykla.

Petrunina, Uliana. 2021. *Adjectivization in Russian. Analyzing participles by means of lexical frequency and constraint grammar.* Doctoral thesis. Tromsø: The Arctic University of Norway. Available at: https://munin.uit.no/bitstream/handle/10037/20757/thesis.pdf?sequence=2&isAllowed=y

ON THE EMERGENCE OF LITHUANIAN PARTICLE CLUSTERS

Erika Jasionytė-Mikučionienė

University of Vilnius

Keywords: Lithuanian particles, particle clusters, syntactic scope, degrees of integration

In the last decades, new cross-linguistic research has addressed the question of particle clusters (Lohmann & Koops 2016, Crible et al. 2017, Crible 2018, Haselow 2019). The main theoretical question is raised whether the resulting particle combination functions as a single – semantically and syntactically indivisible – unit, or whether it consists of independent and easily separable units or components (cf. Josep Cuenca, Crible 2019). In Lithuanian linguistics, the existing descriptions focus more on individual particles (Petit 2010, Sawicki 2012, Panov 2019, Ruskan 2019, among others), while a more systematic account of particle clusters based on synchronic as well as diachronic data is still lacking. Thus, the present paper aims at investigating the combinations of Lithuanian particles considering their development, possible degrees of integration and order in a combination.

The data for the study was obtained from several corpora: the sub-corpora of fiction and spoken language of the Corpus of the Contemporary Lithuanian Language, the Corpus of Spoken Lithuanian, and 16th and 17th-century Old Lithuanian texts compiled by the Institute of the Lithuanian Language. Moreover, a corpus of 19th-century fiction texts was compiled and used in the analysis.

Traditional Lithuanian grammars define particle combinations as collocations of particles with other independent and dependent words. However, the criteria proposed for determining whether a combination can be considered a morphological unit or combination *per se* are insufficient. The research presented in this paper reveals that an important distinguishing criterion, which is not mentioned in the works of Lithuanian linguists, is syntactic scope. The analysis of the empirical data from several corpora of the Lithuanian language confirms that, based on the criterion of syntactic scope, particle combinations can be categorized according to the three degrees of their integration: juxtaposition (a), addition (b), and composition (c):

(a) juxtaposed markers take scope on different units, e.g.:
**A: Ar aš gal ne taip supratau…*
B:* ***Nu tai gal* *irgi suvėlė, suvėlė, ane?* (CSL)
'A: Maybe I misunderstood… B: Well, maybe he said it vaguely, vaguely, didn't he?'

(b) added markers take scope over the same unit; they combine but keep their individual meaning, e.g.:
**A: Tai bet ką? Čia yra viskas, kas C?*
B:* ***Ne kad**.* (CCLL-Sp)
'A: But what? Is it the same as C? B. But no.'

(c) combined markers take scope over the same unit; however, the combined particles function as a single marker and their individual meaning cannot be disentangled anymore, e.g.:
**A: – Sunkiai gyvenat.*
B: – Taip, močiutė padėdavo, o dar pati tai uogaudavau, tai grybaudavau,* ***taip kad**…* (CCLL-Fic)
'A: Life is not easy for you. B. Yes, my grandmother used to help me. Besides, I myself used to pick berries and mushrooms, so…'

However, the results show that the boundaries between the degrees of integration in particle combinations are not always clearly defined. Therefore, this study complements the discussions in general linguistics regarding the boundaries and degree of integration of particle combinations.

References

Crible, Ludivine, Degand, Liesbeth & Gaëtanelle Gilquin. 2017. The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis) fluency. *Languages in Contrast.* 17(1), 69–95. https://doi.org/10.1075/lic.17.1.04cri

Crible, Ludivine. 2018. *Discourse Markers and (Dis)fluency: Forms and Functions across Languages and Registers*, Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/pbns.286

Haselow, Alexander. 2019 Discourse marker sequences: Insights into the serial order of communicative tasks in real-time turn production. *Journal of Pragmatics.* 146, 1–18. https://doi.org/10.1016/j.pragma.2019.04.003

Josep Cuenca, Maria & Ludivine Crible. 2019. Co-Occurrence of Discourse Markers in English: From Juxtaposition to Composition. *Journal of Pragmatics.* 140, 171–84. https://doi.org/10.1016/j.pragma.2018.12.001

Lohmann, Arne & Christian Koops. 2016. Aspects of discourse marker sequencing: Empirical challenges and theoretical implications. In Keizer, Evelien, Kaltenböck, Gunther, Lohmann, Arne (eds.). *Outside the Clause: Form and Function of Extra-clausal Constituents*, Amsterdam/Philadelphia: John Benjamins, 417–446. https://doi.org/10.1075/slcs.178.14loh

Panov, Vladimir. 2019. *Juk* and *gi*, and "particles" in contemporary Lithuanian: Explaining language-particular elements in a cross-linguistic context. *Kalbotyra.* 72, 58–86. https://doi.org/10.15388/Kalbotyra.2019.3

Petit, Daniel. 2010. On presentative particles in the Baltic languages. In Nicole, Nau, Ostrowski, Norbert (eds.). *Particles and Connectives in Baltic*, Vilnius: Vilniaus Universitetas, Asociacija "Academia Salensis", 151–170.

Ruskan, Anna. 2019. Functional variation of discourse particles in Lithuanian: A look at clause Peripheries. *Corpus Pragmatics.* 3, 303–325. https://doi.org/10.1007/s41701-019-00061-0

Sawicki, Lea. 2012. Responsive discourse particles in Lithuanian dialog. *Baltic Linguistics.* 3, 151–175. https://doi.org/10.32798/bl.422

WORD-MEDIAL ELEMENTS IN LATVIAN COMPOUNDS: A CORPUS-BASED STUDY

Andra Kalnača & Tatjana Pakalne

University of Latvia

Keywords: compounds, linking elements, left-hand components

Elements occurring at the boundary between two parts of a compound in many Indo-European languages, e.g., Slavic, Romance, Germanic, Baltic languages, Greek, have been classified in literature as a type of morpheme, i.e. interfix (e.g., Kalnača 2004, Haspelmath, Sims 2010), and also as formatives that are not themselves morphs (Bauer 2018), generally termed 'linking elements' (LE). The characteristic features of LEs across languages include lack of semantics, erratic or complex distribution and constraints on productivity, functional non-uniformity, as well as interchangeability with one another or with a null element (Szczepaniak 2020). LEs should be distinguished from inflectional suffixes or endings in the same position, e.g., inflection of the left-hand component, to which they are sometimes homonymous, (Lieber, Štekauer 2009, Kürschner, Szczepaniak 2013).

Latvian has at least three kinds of elements occurring at the boundary between compound components (word-medial elements or WMEs), all categorized as interfixes in literature (Haspelmath, Sims 2010, Nītiņa, Grigorjevs 2013): elements synchronically corresponding to an inflectional word-final morpheme of the left-hand (LH) component (1), elements unlike any relevant inflectional morpheme, occurring in compounds with borrowed parts (2), and remnants of an old stem thematic vowel *-a-* (3) (Endzelīns 1951).

(1) a. NOM.SG
*vec-**ā**-māt-e* 'grandmother'

b. GEN.PL
*ac-**u**-mirkl-is* 'instant (N)'

c. INS.PL
*kāj-**ām**-gājējs* 'pedestrian'
*aus-**īm**-dzirdams* 'audible'

d. ACC.SG
*pirm-**o**-reiz* 'the first time'

e. LOC.SG
*gal-**ā**-tikšana* 'coping'

f. *palīdz-ē-**t**-tieksme* ‘willingness to help’, *strād-ā-**t**-prieks* ‘joy in working’,
*konserv-ē-**t**-aizsākts* ‘started to get preserved’ (INF)

(2) a. *imun-**o**-krāsošana* ‘immunostaining’
b. *kart-**o**-shēma* ‘schematic map’
c. *Austr-**o**-ungārija* ‘Austria-Hungary’

(3) *plik-**a**-dīda* ‘pauper’, *niek-**a**-bīlis* ‘idle talker’, *abr-**a**-kasis* ‘a small loaf of dough leavings’, *kaz-**a**-kuņģis* ‘goat stomach’

Differentiation between inflection of the LH-component, on the one hand, and ‘empty’ elements inserted between two compound components, i.e. LEs, on the other hand, is not always straightforward. E.g., while (2) and (3) above are clear LEs, the WMEs in (1) are more inflection-like. This poses a question as to whether these are actually inflectional endings or homonymous LEs inserted between compound components, e.g., by analogy to already existing compounds. These seemingly minor particulars are significant, because they potentially touch upon deeper issues: 1) word-formation models underlying compounding in Latvian, esp. using whole words vs. stems, 2) available compounding strategies, i.e. combining two separate source words vs. transforming a single syntactic construction, e.g., an noun phrase, a prepositional phrase, a coordinative word group into a compound (the so-called syntactic mode of word-formation), 3) the semantic representation of Latvian compounds, incl. the role of grammatical semantics expressed by the word-final morpheme of the LH-component, such as case, gender, number.

The proposed study is intended as groundwork for further systematic discussion of these questions. It includes creating a comprehensive dataset on contemporary Latvian compounding from the data of the *Balanced Corpus of Modern Latvian (LVK2018)* via the *Database of Latvian Morphemes and Derivational Models (DLMDM)* (Project No. lzp-2022/1-0013) and providing a quantitative and qualitative analysis of properties related to the presence or absence of WMEs in compounds. By DLMDM project mid-term, this amounted to 21638 multi-root lemmas grouped according to 1) the type of the LH-component (an indeclinable whole word; a non-segmentable whole word; a stem + WME formally corresponding to a declinable whole word; a stem without a WME; a stem with a non-inflectional WME, i.e. LE); 2) grammatical semantics expressed by specific WMEs: case, number, part of speech (POS), as in (1); 3) word stems occurring in lemmas with and without WMEs, as ‘kaln’ in (4).

(4) *kaln-**ā**-kāpējs* ‘mountain climber’ (LOC.SG), *kaln-**u**-mētra* ‘savory’ (GEN.PL), *kaln-**ø**-raktuve* ‘mine’ (NULL)

The picture that emerges from the data is that, on the one hand, among LH-components of declinable POSs, stems without WMEs corresponding to inflectional endings are much more widespread than stems with such WMEs formally corresponding to whole words. On the other hand, whole words as LH constituents are not altogether alien to Latvian. There are long-established formations with morphemically non-segmentable pronouns: ***šīs***-*dienas* 'today's', ***šo***-*pavasar* 'this spring', ***šai***-*pus* 'on this side', ***tā***-*dēļ* 'therefore', numerals: *trī*-***s***-*simt* 'three hundred', *simt*-***s***-*procentīgs* 'hundred-percent', prepositions, older adverbs, and nouns with old endings as WMEs: *gad*-***s***-*kārta* 'season', *rāt*-***s***-*nams* 'town hall', *likten*-***s***-*bērns* 'destiny's child', *vien*-***is***-*prātis* 'of the same opinion', as well as an open class of LH-components with WMEs synchronically corresponding to a broad range of grammatical form markers: the infinitive for verbs, all case forms for nouns and, to a lesser extent, adjectives, pronouns and numerals. This richness of grammatical forms among LH-components boils down to certain characteristic pairings of POSs and grammatical forms (compounding models). Some are much more productive than others. E.g., among 1983 compounds whose LH-component formally corresponds to a declinable whole word, 73% LH-components are N GEN, out of which 83% are PL. By contrast, ADJ NOM SG, as in (1a), or N INS PL, as in (1c), are limited to a handful of lexemes probably giving rise, from time to time, to analogical formations. Apart from quantitative data and definition of compounding models underlying compounds with WMEs in Latvian, the study also offers an overview of factors restricting and enhancing the productivity of WMEs. Future work might include analysis in terms of compounding strategies and the role of WMEs in the semantic representation of compounds.

This study was carried out in the framework of the Latvian Council of Science research project "Database of Latvian Morphemes and Derivational Models" (No. lzp-2022/1-0013).

References

Bauer, Laurie. 2018. Morphological Entities: Overview and General Issues. *Oxford Research Encyclopedia of Linguistics.* Retrieved 11 Feb. 2025, from https://oxfordre-com.datubazes.lanet.lv/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-540

Endzelīns, Jānis. 1951. *Latviešu valodas gramatika.* Rīga: Latvijas Valsts izdevniecība.

Haspelmath, Martin, Sims, Andrea D. 2010. *Understanding Morphology.* London: Hodder Education.

Kalnača, Andra. 2004. *Morfēmika un morfonoloģija*. Rīga: Latvijas Universitātes Akadēmiskais apgāds.

Kürschner, Sebastian, Szczepaniak, Renata. 2013. Linking elements – origin, change, and functionalization. *Morphology*. 23(1), 1–6. https://doi.org/10.1007/s11525-013-9215-7

Levāne-Petrova, Kristīne, Darģis, Roberts. *Balanced Corpus of Modern Latvian (LVK2018)*. CLARIN-LV digital library at IMCS, University of Latvia. 2018. Available at: http://hdl.handle.net/20.500.12574/11

Lieber, Rochelle, Štekauer, Pavol. 2009. Introduction: status and definition of compounding. In *Oxford Handbook of Compounding*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199695720.013.0001

Nītiņa, Daina, Grigorjevs, Juris (eds.). 2013. *Latviešu valodas gramatika*. Rīga: Latvijas Universitātes Latviešu valodas institūts.

Szczepaniak, Renata. 020. Linking Elements in Morphology. In *Oxford Research Encyclopedia of Linguistics*. Retrieved 11 Feb. 2025, from https://oxfordre-com.datubazes.lanet.lv/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-571

VARIATION IN THE CASE FORMS OF THE INDEFINITE PRONOUN *KEEGI* 'SOMEONE': A COMPARATIVE CORPUS STUDY OF WRITTEN AND SPOKEN ESTONIAN

Annika Kängsepp

University of Tartu

Keywords: corpus linguistics, language variation, indefinite pronouns, morphology, Estonian

Variation in the case forms of the Estonian indefinite pronouns *keegi* 'someone', *miski* 'something', *kumbki* 'either', and *ükski* 'none' has been observed in written language (Rull 1917) and dialects (Saareste 1955) for over a century. This variation concerns the placement of the *-gi/-ki*, which is typically regarded as an emphatic clitic in other contexts. Within indefinite pronouns, however, it can occur after the case ending (e.g., *kelle-le=gi* someone-ALL=CLIT 'to someone'), before the case ending (e.g., *kellegi-le* someone.INDF-ALL), between two case endings (e.g., *kelle-le-gi-le* someone-ALL-INDF-ALL) or before and after the case ending (e.g., *kellegi-le=gi* someone.INDF-ALL=CLIT) (Saareste 1923, 1936). In indefinite pronouns, *-gi/-ki* is perceived more as part of the stem. The variation has a strong dialectal background, as forms where *-gi/-ki* is placed after the case ending have historically been common only in Southern and Northeast Estonia (Saareste 1955). Nevertheless, in standard Estonian, the normative placement occurs only after the case ending.

As this variation has been systematically studied very little and only in written contemporary Estonian (Pant 2018, 2020), this presentation aims to provide an overview of its extent and to identify the factors influencing the variation in the case forms of the pronoun *keegi* through a comparative analysis of written and spoken Estonian. The analysis of written language is based on data from the Estonian National Corpus (2.4 billion words; Koppel, Kallas 2022). For the analysis of spoken language, data were drawn from two corpora: the Estonian Public Broadcasting's Radio Corpus (109 million words; Lippus et al., 2023a), and the Estonian Podcast Corpus (85 million words; Lippus et al., 2023b). To examine the variation, the proportion of occurrences was calculated, while statistical analysis was applied to identify the factors influencing the placement of *-gi/-ki*.

The findings reveal that in written Estonian, *-gi/-ki* is predominantly placed after the case ending, accounting for 78.6%, while forms where *-gi/-ki* precedes the case ending constitute 20.6%. Instances where *-gi/-ki* appears between case endings or both before and after it represent the remaining 0.8%. In spoken Estonian, the distribution is more varied, with *-gi/-ki* appearing after the case ending in 54.2% of occurrences, before the case ending in 43.4%, and in other positions in 2.4%. Univariate analysis of written Estonian identified genre, the occurrence of the pronoun as an attribute, and the function of the pronoun in a clause as significant factors influencing variation. In spoken Estonian, the corpus, the position of the pronoun in the clause, the gender of the speaker, case, and speech rate were found to significantly affect the placement of *-gi/-ki*. Multivariate analysis further indicated that genre, and whether the text had been edited, was the strongest factor influencing written language variation. In spoken language, speech rate emerged as the most influential factor, with faster speech favouring the placement of *-gi/-ki* before or between two case endings. Additionally, male speakers were more likely to produce forms where *-gi/-ki* precedes or occurs between case endings (Kängsepp 2024, 2025).

This presentation argues that corpus-based analysis provides valuable insight into grammatical variation across registers, emphasizing how using both corpora offers a more comprehensive view by incorporating linguistic, sociolinguistic, and prosodic factors.

This work was supported by the National Program: Estonian Language and Culture in the Digital Age project “Morphosyntactic variation in Estonian” (EKKD-TA2).

References

Koppel, Kristina, Kallas, Jelena. 2022. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eesti keele kogu [Estonian National Corpus 2013–2021: The largest collection of Estonian language data]. *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics.* 18, 207–228. https://doi.org/10.5128/ERYa18.12

Kängsepp, Annika. 2024. Käändevormide varieerumine ja seda mõjutavad tegurid kirjalikus keeles indefiniitpronoomeni keegi näitel [The variation in the case forms of the indefinite pronoun keegi ‘someone’ in written Estonian]. *Keel ja Kirjandus.* 11, 1016–1037. https://doi.org/10.54013/kk803a3

Kängsepp, Annika. 2025. Indefiniitpronoomenite keegi ja miski käändevormide varieerumine suulises keeles [Variation in the case forms of the indefinite pronouns keegi ‘someone’ and miski ‘something’ in spoken Estonian]. *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics.* 21, 121–139. https://doi.org/10.5128/ERYa21.07

Lippus, Pärtel, Alumäe, Tanel, Orasmaa, Siim, Tsepelina, Katrin, Lindström, Liina. 2023a. *Eesti Rahvusringhäälingu raadiosaadete korpus* [Estonian Public Broadcasting's Radio Corpus]. https://doi.org/10.23673/re-441

Lippus, Pärtel, Alumäe, Tanel, Orasmaa, Siim, Pilvik, Maarja-Liisa, Lindström, Liina. 2023b. *Eesti taskuhäälingukorpus* [Estonian Podcast Corpus]. https://doi.org/10.23673/re-445

Pant, Annika. 2018. *Asesõnade keegi, miski, kumbki käändevormide varieerumine eesti kirjakeeles* [*The variation of case forms in pronouns keegi* 'someone', *miski* 'something', *kumbki* 'either' *in written Estonian*]. Bakalaureusetöö. Tartu: Tartu Ülikool. Available at: http://dspace.ut.ee/bitstream/handle/10062/60560/Pant_Annika_BA2018.pdf?sequence=1&isAllowed=y

Pant, Annika. 2020. *Pronoomenite keegi, miski, kumbki, ükski käändevormide kasutus tänapäeva eesti keeles* [*The variation of case forms usage of pronouns keegi* 'someone', *miski* 'something', *kumbki* 'either', *ükski* 'none' *in modern Estonian*]. Magistritöö. Tartu: Tartu Ülikool. Available at: http://dspace.ut.ee/bitstream/handle/10062/68143/Pant_Annika_MA2020.pdf?sequence=1&isAllowed=y

Rull, Aado. 1917. Kaaslõpp -gi ja ta kidurad kaimud [The clitic -gi and its lesser relatives]. *Eesti Kirjandus*. 2, 82–88.

Saareste, Albert. 1923. Kellegile, mingit, kuskil [To someone, some, somewhere]. *Eesti Keel*. 4, 116–117.

Saareste, Andrus. 1936. Eesti õigekeelsuse päevaküsimustest [On Current Issues in Estonian Orthography]. *Eesti Kirjandus*. 12, 548–556.

Saareste, Andrus. 1955. *Petit atlas des parlers estoniens. Väike eesti murdeatlas* [*A Small Estonian Dialect Atlas*]. Uppsala: Almqvist & Wiksells.

PERFECT VARIATIONS ACROSS SLAVIC AND BALTIC: TRANSLATION MINING APPROACH TO CORPUS DATA

Dorota Klimek-Jankowska[1], Alberto Frasson[1], Antonina Mocniak[2], Andrzej Żak[3], Elena Vaiksnoraite[4], Tanya Ivanova-Sullivan[5], Daria Seres[6], Vladimir Cvetkoski[7], Diana Androva[8] & Patrick Mihaylov[8]

University of Wrocław[1], Jagiellonian University[2], Institute of Slavic Studies at the Polish Academy of Sciences[3], Ohio State University[4], University of California Los Angeles[5], University of Graz[6], Ss. Cyril and Methodius University in Skopje[7] & Sofia University[8]

Keywords: Translation Mining, Perfect Tense, Past Tense, True Perfect, Aoristic Perfect, Evidential Perfect, Slavic, Baltic

Studies on languages encoding Perfect Tense have provided evidence of its cross-linguistic variation (Migdalski 2006, Arkadiev & Wiemer 2020, Dahl 2021, de Swart 2022, Fuchs, González 2022, Bertrand et al. 2022, Le Bruyn 2022, Mulder et al. 2022, Corre et al. 2025). We use a Translation Mining corpus-based approach (Van der Klis et al. 2017) to systematize this variation in Baltic and Slavic by comparing the English_(En) version of *Harry_Potter_and_the Philosopher's_Stone* with its translations in Bulgarian_(Bg), Croatian_(Cr), Latvian_(Lv), Lithuanian_(Lt), Macedonian_(Mc), and Serbian_(Sr). We extracted all the present perfect contexts in the English original and we counted the frequencies of tenses used to translate them in the target languages. A quantitative analysis employing a chi-square independence test was performed on language and tense categories used for translations of the En present perfect, and we report a statistically significant association (p < .001) between language and tense choice. Two-sample proportion tests were used as post hoc comparisons to follow up on the chi-square result. These tests revealed that Mc significantly differs from Bg (p = .01), Cr (p < .001) and Sr (p < .001) in the frequency of present perfect use. No significant differences were found between Bg and Sr (p = 0.02) or Cr and Sr (p = 0.051). Lv differs significantly from Cr (p = .0002) and Sr (p =.003), but does not differ significantly from Bg (p = .08) and Mc (p =.44). Lt differs significantly from all the lgs in our sample. To reach a deeper understanding of the statistically significant contrasts we used our Translation Mining tools to carefully analyse the linguistic data and

arrive at microtypological and formal generalizations. In this talk we plan (i) to revisit Klein 1992's Present Perfect puzzle to account for the compatibility of Serbian and Croatian present perfect with definite time adverbials, lifetime effect contexts and sequence of events contexts; (ii) extend Klein 1992's Present Perfect puzzle to evidential perfect contexts which can also be modified by definite past time adverbials in Bulgarian, Macedonian and Latvian; (iv) to show that Macedonian *be*-perfect is more advanced in transitioning to evidential perfect as compared to Bulgarian and Latvian. This explains why *be*-perfect is significantly less frequent in Macedonian than in Bulgarian (and also Latvian) which license more instances of true perfect contexts; (v) to discuss present tense translations of English universal present perfect contexts.

This research was supported by the Polish National Science Center (NCN) grant SONATA BIS-11 HS2 (No. 2021/42/E/HS2/00143).

References

Arkadiev, Peter, Wiemer, Björn. 2020. Perfects in Baltic and Slavic. In *Perfects in Indo-European Languages and Beyond*. John Benjamins, 123–214. https://doi.org/10.1075/cilt.352.05ark

Bertrand, Anne et al. 2022. Nobody's perfect. *Languages*. 7(2), 148. https://doi.org/10.3390/languages7020148

Corre, Eric et al. 2025. Intermediate perfects: A comparison of Dutch, Catalan and Breton. *Languages in Contrast*. 25(1), 1–22. https://doi.org/10.1075/lic.22008.cor

Dahl, Östen. 2021. "Universal" readings of perfects and iamitives in typological perspective. *The Perfect Volume*: *Papers on the Perfect*. Amsterdam, Philadelphia: John Benjamins, 43–63. https://doi.org/10.1075/slcs.217.02dah

De Swart, Henriëtte et al. 2022. Perfect variations in Romance. *Isogloss*. 8(5), 1–31. https://doi.org/10.5565/rev/isogloss.213

Fuchs, Martín, González, Paz. 2022. Perfect-Perfective variation across Spanish dialects: a parallel-corpus study. *Languages*. 7(3), 166. https://doi.org/10.3390/languages7030166

Klein, Wolfgang. 1992. The present perfect puzzle. *Language* , 525–552.

Le Bruyn, Bert et al. 2022. Parallel corpus research and target language representativeness: The contrastive, typological, and translation mining traditions. *Languages*. 7(3), 176. https://doi.org/10.3390/languages7030176

Migdalski, Krzysztof. 2006. *The syntax of Compound Tenses in Slavic*. PhD thesis. Tilburg: University of Tilburg.

Migdalski, Krzysztof. 2015. On the loss of Tense and verb-adjacent clitics in Slavic. In Biberauer Theresa, Walkden, George (eds.). *Syntax over Time: Lexical, Morphological, and Information-Structural Interactions*. Oxford University Press, 179–196. https://doi.org/10.1093/acprof:oso/9780199687923.003.0011

Mulder, Gijs et al. 2022. Tense and aspect in a Spanish literary work and its translations. *Languages*. 7(3), 217. https://doi.org/10.3390/languages7030217

Van der Klis, Martijn et al. 2017. Mapping the perfect via translation mining. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 2, 497–502.

Van der Klis, Martijn et al. 2022. A multilingual corpus study of the competition between past and perfect in narrative discourse. *Journal of Linguistics*. 58(2), 423–457. https://doi.org/10.1017/S0022226721000244

EXTRACTING GRAMMATICAL AND MORPHOLOGICAL ELEMENTS FOR SUBJECTIVITY DETECTION IN BACHTRACK REVIEWS OF THE LATVIAN ARTISTS' PERFORMANCES

Kristīna Korneliusa

University of Latvia

Keywords: subjectivity, corpus linguistics, derivational suffixes, part-of-speech tagging, Sketch Engine

Being a subject refers to the capability "of being in conscious states and of being the subject of conscious events" (Peacocke 2012, 90). Since consciousness is fully internal, so is a human being's subjective perception. To communicate it to others, verbal language is used as one of the possible semiotic systems – though complex and developed, it still has its limitations, and part of the intended meaning can be lost. However, this problem ensures the topicality of finding ways of "self-expression in language" (Fina 2009, 171), as subjectivity is defined in linguistics. The importance of detecting subjectivity especially concerns the texts which deal with the communication of one's aesthetic experience, e.g., performing arts reviews. To make the findings generalisable, a large dataset is necessary. Hence, a corpus-based approach is preferable for data extraction and processing. It requires a list of specific linguistic markers expressing subjectivity. The goal of the current research is to detect linguistic subjectivity markers applicable for the corpus-based analysis of performing arts reviews. The following research questions are asked: (1) which grammatical and morphological elements can be identified as linguistic markers of subjectivity and (2) what are the advantages and drawbacks of detecting subjectivity using grammatical and morphological elements compared to semantic annotation?

Based on the theoretical considerations of linguists and philosophers (see Wiebe et al. 2004; Solomon 2005, Baumgarten et al. 2012, Pho 2012, Peng 2024), the concept of subjectivity is broken down into the following components: opinion, judgement, beliefs, attitude, evaluation and assessment, wishes and desires, feelings and emotions, and linguistic markers – semantic and grammatical – are attributed to them.

The primary tool selected for the analysis is *Sketch Engine*. The detected subjectivity markers are tested in comparison to the ones extracted with the help of the semantic tagger *Wmatrix 5*. The semantic tags denoting subjectivity in *Wmatrix 5* interface are based on components of subjectivity derived from theory. Their list has been compiled by the author in previous research.

The research corpus Bachtrack Reviews Latvia (BARRE) contains 116 699 tokens across 129 reviews of performances by Latvian artists published on Bachtrack.com (Online 1), an international classical music website containing concert, opera and dance reviews.

The preliminary findings suggest that *Sketch Engine* is superior to *Wmatrix 5* in terms of the extraction of first- and second-person pronouns, as semantic tagging recognizes all pronouns as a single category. The extraction of other parts of speech requires the combination of part-of-speech tagging and incorporation of morphological elements into *Sketch Engine* queries. For instance, the adverbs and adjectives containing derivational suffixes are more likely to be among subjectivity markers.

References

Baumgarten, Nicole, Du Bois, Inke, House, Juliane (eds.). 2012. Introduction. In N. Baumgarten, I. Du Bois and J. House (eds.). *Subjectivity in Language and in Discourse*. Bingley: Emerald Group Publishing Limited, 1–14.

Fina, Anna. 2009. Language and subjectivity. *Estudios de Lingüística Aplicada*. 27(50), 117–176.

Peacocke, Christopher. 2012. Subjects and consciousness. In A. Coliva (ed.) *The Self and Self-Knowledge*. Oxford: Oxford University Press, 74–101.

Peng, Bingzhuan. 2024. Subjectivity of Discourse Constructions in News Discourse by Integrating Construction Grammar and Critical Discourse Analysis. *Applied Mathematics and Nonlinear Sciences*. 9(1), 1–16. https://doi.org/10.2478/amns-2024-1050

Pho, Phuong Dzung. 2012. Authorial stance in research article abstracts and introductions from two disciplines. In N. Baumgarten, I. Du Bois & Juliane House (eds.). *Subjectivity in Language and in Discourse*. Bingley: Emerald Group Publishing Limited, 97–114. https://doi.org/10.1163/9789004261921_006

Solomon, Robert. 2005. Subjectivity. In T. Honderich (ed.). *The Oxford Companion to Philosophy*. Oxford: Oxford University Press, 900.

Wiebe, Janyse, Bruce, Rebecca, Martin, Melanie, Wilson, Theresa, Bell, Matthew. 2004. Learning subjective language. *Computational Linguistics*. 30(3), 277–308. https://doi.org/10.1162/0891201041850885

Internet sources

Available at: https://bachtrack.com/

Tools

Sketch Engine. Available at: https://www.sketchengine.eu/

Wmatrix 5. Available at: https://ucrel-wmatrix5.lancaster.ac.uk/

PASSIVE CONSTRUCTIONS IN NORWEGIAN *THEN* AND *NOW*

Maria Klejnowska

Adam Mickiewicz University in Poznań

Keywords: Norwegian, Middle Norwegian, passive voice, corpus, diachronic linguistic

Passive constructions are a common phenomenon across the world's languages, and it is without doubt broadly studied. This paper will examine the use of passive constructions in Norwegian, both in the past and in the present. While significant research has been conducted on the passives in Scandinavian languages, the existing literature predominantly adopts a synchronic perspective, with comparatively limited attention to the diachronic aspect. Among the languages of Mainland Scandinavia, Norwegian is arguably the least studied regarding passive constructions.

The passive constructions could be expressed in both Modern and Medieval Norwegian in two different ways: firstly, with an auxiliary verb and past participle (a) which constitutes a periphrastic construction; and secondly, with an addition of an -s ending to the verb (b) which constitutes a morphological construction.

a) *Maten blir spist.*
The food become.AUX eat.PTCP
'The food is being eaten.'

b) *Maten spise-s.*
The food eat-PASS
'The food is being eaten'

Nevertheless, a considerable number of linguistic changes have emerged between the historical and modern forms of passive constructions. The Middle Norwegian period, which spans approximately 1350–1550 CE, is distinguished by a multitude of linguistic transformations. The passive voice was also subject to change during this period. In periphrastic constructions, the auxiliaries and their usage underwent the most significant shift. This was due to the introduction of a new verb, blífa 'remain', which was brought by the Hanseatic merchants during the Late Middle Ages. This verb quickly became a new auxiliary, and the native verða 'become' dropped in frequency. In addition to the shift in auxiliaries, the frequency of morphological passive seems to be the biggest change that is clearly visible in the data.

This study aims to compare passive constructions found in Norwegian at two distinct historical periods: the present day and the Middle Ages. In order to achieve this, the study will examine and comment on the differences between the two periods in the following aspects:

- Usage of morphological and periphrastic passive
- Tempus of passive constructions
- Modality of passive constructions.
- Type of subject used
- Agentivity of the subject

This study draws data from newspapers and literature written between 1977 and 2005, that constitute the corpus of the modern Norwegian. The data has been compiled by Anu Laanemets (2012). Her data will be compared with the data from my own corpus that comprises the diplomas written between the 13th and 16th century. Diplomas are documents from Middle Ages of legal nature and constitute the most important source for linguistics research on the language during the Middle Norwegian period (ca. 1350–1550 CE). They are collected in Diplomatarium Norvegicum, which is available online.

References

Diplomatarium Norvegicum. Available at: https://www.dokpro.uio.no/

Laanemets, Anu. 2012. *Passiv i moderne dansk, norsk og svensk: Et korpusbaseret studie af tale- og skriftsprog*. Dissertationes Philologiae Scandinavicae Universitatis Tartuensis 2. Tartu: Tartu University Press.

MORPHOLOGICAL FORM AS AN IMPORTANT FEATURE OF WORD PATTERNING

Jolanta Kovalevskaitė

Vytautas Magnus University of Kaunas

Keywords: grammar–semantics interface, corpus pattern, corpus-driven lexical database, Lithuanian, Lexical Database of Lithuanian Language Usage

Corpora are widely recognized as representative linguistic sources where the contextual information of a word (collocations, corpus patterns, valencies, etc.) can be observed. From the lexical perspective, corpus analyses can show a strong interconnection of lexis and grammar (Sinclair 2000). In such morphologically rich language as Lithuanian, corpus-driven evidence about the interconnection of lexis and grammar is needed to achieve more accurate lexicographic description. Recent research based on the new resources for teaching Lithuanian as a second language (Kovalevskaitė et al. 2020) revealed a strong interconnection of word meaning and lexical, grammatical, and semantic environment of the word. This usage information, provided in the form of corpus patterns, was collected from the *Pedagogic Corpus* with the aim to build the *Lexical Database of Lithuanian Language Usage* (https://kalbu.vdu.lt/en/resources/lexical-database-of-lithuanian-language-usage/).

As Bielinskienė et al. (2021) have shown, detected corpus patterns of nouns and adjectives reflect different meanings, and pattern variants show that the grammatical characteristics of a specific word usage are rather individual (the same observation is confirmed by the corpus pattern analysis of verbs, see Kovalevskaitė et al. (2024)). It has been observed that different senses are often related to different patternings. E.g., the entry of the noun *tiesa* 'truth' describes the three senses of the word. 7 patterns describe the 1st sense, whereas the 2nd and the 3rd senses have only one pattern each. However, the form of the word itself can be related to concrete sense of the particular word, e.g., when used in its 1st and 2nd senses, the noun has a singular form, whereas in the 3rd sense, the noun is used in plural: *amžinos tiesos* ('eternal truths'). As to verbs, some infinite forms can be relevant only for some verbs and/or particular senses of these verbs (Kovalevskaitė, Rimkutė et al. 2023).

In other inflected languages, the interconnection of word meaning and its grammatical form has been examined by the automated research of grammatical profiles. E.g., by using the *GramatiKat* application (Kováříková, Kovářík 2023) one can assess typical behavior in individual word by observing the relative frequency distribution of the inflected forms of a word, with a particular focus on missing forms and forms with higher-than-expected frequencies. From this data, one can build clusters based on morphological (and consequently semantic) similarities of the analysed words. However, in case of polysemous words, the automated research of inflectional profiles of lexeme may have limitations if the studied words were not analysed and classified according to their different senses (Arppe 2006).

The dataset used for this study is taken from the *Lexical Database of Lithuanian Language Usage* and consists of about 700 frequently used Lithuanian headwords (verbs, nouns, adjectives, and adverbs) and their patterns. Each pattern, associated with a specific meaning of a headword, presents information on grammatical, semantic, and lexical levels. The presentation aims to show which inflected forms are usually included on the grammatical level of the patterns and, accordingly, how these forms interrelate with separate meanings of the analysed words.

References

Arppe, Antti. 2006. Frequency considerations in morphology, revisited – Finnish verbs differ, too. *Finnish Journal of Linguistics*, 175–189. Available at: https://journal.fi/finjol/article/view/153145

Bielinskienė, Agnė, Kovalevskaitė, Jolanta, Rimkutė, Erika. 2021. Grammatical patterns in the corpus-driven "Lexical Database of Lithuanian". *Language: Meaning and Form*. 12, 7–30. https://doi.org/10.22364/vnf.12.01

Kovalevskaitė, Jolanta, Bielinskienė, Agnė, Rimkutė, Erika. 2024. *Tekstynais paremti kalbos vartosenos tyrimai leksikografijoje: „Mokomojo lietuvių kalbos vartosenos leksikono" atvejis*. Kaunas: Vytauto Didžiojo universitetas. Available at: https://hdl.handle.net/20.500.12259/271925

Kovalevskaitė, Jolanta, Rimkutė, Erika. 2023. Kodėl svarbios neasmenuojamosios formos: Mokomojo lietuvių kalbos vartosenos leksikono veiksmažodžių tyrimas. *Taikomoji kalbotyra*. 19, 57–77. https://doi.org/10.15388/Taikalbot.2023.19.5

Kovalevskaitė, Jolanta, Boizou, Loic, Bielinskienė, Agnė, Jancaitė, Laima, Rimkutė, Erika. 2020, The First Corpus-driven Lexical Database of Lithuanian as L2. In *Proceedings of the Ninth International Conference Human Language Technologies – The Baltic Perspective*. IOS Press, 245–252. https://doi.org/10.3233/FAIA200630

Kováříková, Dominika, Kovářík, Oleg. 2023. *GramatiKat (version 2): A tool for grammatical categories research and grammatical profiles*. FFUK. Available at: https://www.korpus.cz/gramatikat/

Sinclair, John. 2000. Lexical Grammar. *Darbai ir Dienos*. 24, 191-203.

DISCOURSE PRAGMATICS OF PARTICLE-INITIAL CONSTRUCTIONS IN ENGLISH

Anastasiia Protopopova

Université Paris-Cité

Keywords: adverbial particle, preposing, subject-dependent inversion, observer-effect, deixis

Particle-initial constructions (PI-constructions) in English almost exclusively belong to one of two types. Subject-dependent Inversion (SDI) is preferred if the subject is a full NP and Particle Preposing (PartPrep) if the subject is a pronoun.

(1) a. *In came Kim.* (SDI)
b. *In she came.* (PartPrep)

Bolinger (1977) notes that PI-constructions have an effect of bringing the speaker 'on stage', where the event is treated as if directly perceptible. With respect to discourse pragmatics, Cappelle (2002) distinguishes two discourse functions for PI-constructions, Focus Preposing (FocPrep, focus on the particle) and Presentative Preposing (PresPrep, focus the subject). He claims that (i) when the subject is a pronoun, PI is always FocPrep; (ii) certain particles are more frequent with FocPrep while others are more frequent with PresPrep.

The aim of the current study is to determine the mechanism of particle selection for each discourse function of PI-constructions.

I have collected a corpus of 138 PI-constructions on the COCA FICTION (1993–2001). The FICTION subcorpus was accessed through TXM. The queries retrieved sentence-initial uses of PI-constructions. This allowed me to limit the number of instances without constraining the structure searched for, thus making the corpus representative.

The corpus corroborates Cappelle's finding (i) and shows that PresPrep dominates in SDI. The corpus was coded for the discourse status of the three constituents involved (Subject, Verb, Particle) on a scale of 1–4 (discourse-old, easily inferrable, inferrable, discourse-new).

It was found that the verbs are always at least easily inferrable (Birner 1996). The verbs come and go dominate in the corpus (41 and 39 occurrences respectively). The preference for deictic verbs links up to Bolinger's observation that PI-constructions are used deictically. This also suggests that the discourse status of the subject-referent might be linked to the deictic properties of the verb: with *go*, it should tend to be discourse-old (old referent exiting the scene) while the opposite holds for *come* (new referent entering the scene).

This new hypothesis is corroborated by corpus data. Since the 'on stage' effect treats the event as if directly perceptible, the deictic verbs select particles with respect to the location of the deictic centre, which is itself linked with a discourse function of PI. For instance, *In* and *out* can both mark a new referent entering the scene or an old referent exiting the scene. The directional meaning of *away* seems to be at odds with PresPrep. *On*, used to highlight continuousness, does not introduce or remove the subject from the scene.

Therefore, the corpus corroborates Cappelle's observation (ii). Since discourse-old referents tend to appear in FocPrep, and discourse-new referents, in PresPrep, the choice of verb is correlated with the choice of construction. Consequently, since the choice of verb also influences the choice of the particle, it establishes the connection between the particle and the preferred discourse function, thus explaining Cappelle's observation (ii).

References

Birner, Betty J. 1996. *The Discourse Function of Inversion in English*. New York, London: Routledge.

Bolinger, Dwight. 1977. *Meaning and Form*. New York: Longman.

Cappelle, Bert. 2002. And up it rises: Particle preposing in English. *Verb-particle explorations*. Berlin, Boston: De Gruyter Mouton, 43–66. https://doi.org/10.1515/9783110902341.43

ENRICHING LITHUANIAN CORPUS COLLECTION WITH SYNTACTICALLY ANNOTATED DATA

Erika Rimkutė, Agnė Bielinskienė, Jolanta Kovalevskaitė & Jurgita Vaičenonienė

Vytautas Magnus University of Kaunas

Keywords: Lithuanian language, syntactically annotated corpus, Universal Dependencies

There are several corpora for Lithuanian (https://sitti.vdu.lt/en/resources/) which have morphological layers. The development of morphologically annotated corpora allowed the researchers to study the distribution of parts of speech and grammatical categories for Lithuanian in written discourse (Brinkutė 2018), in written and spoken language (Dabašinskienė 2009, Kamandulytė-Merfeldienė 2018). Corpus data have also been used for morphotactics, a little-studied area in Lithuanian (Rimkutė et al. 2016).

The corpus-based analysis is very important to study the syntax of the Lithuanian language, as it helps to achieve a more descriptive approach to syntactic phenomena in grammars; to identify the relationship of different grammatical patterns with styles and varieties; to assess the use of constructions and functions in terms of frequency, synchronicity, and diachrony. Moreover, automatic syntactic analysis contributes to the development of natural processing tools for Lithuanian. Based on the material from *The Pedagogic Corpus of Lithuanian*, attempts have been made to identify lexical and grammatical corpus patterns in texts for learners of Lithuanian as a foreign language (Bielinskienė et al. 2021, Kovalevskaitė et al. 2024).

To study grammatical patterns in large general-type corpora, it is necessary to annotate them not only morphologically, but also syntactically. The syntactically annotated corpus ALKSNIS is too small for comprehensive syntactic research: it consists of 3643 syntactically annotated sentences (~60 000 words) in the PML (Prague Mark-up Language) format (Bielinskienė et al. 2016, Rimkutė et al. 2019). Other typologically similar languages have considerably more syntactically annotated data: e.g., a balanced subset of the Corpus of Modern Latvian (100 million words) has syntactic (Universal Dependencies) and semantic annotation layers (12–17 thousand sentences) (Saulīte et al. 2022); the written Czech corpora (SYN2015 and

SYN2020, 100 million words each) are available with syntactic annotation (Křen 2020).

The project “Morphologically and syntactically annotated text models for training (gold standards)” is developing resources that will help to fill the gap of a larger syntactically annotated corpus in the Lithuanian language and allow syntactically annotating large amounts of data. In this presentation, we will focus on the new syntactically annotated corpus of the Lithuanian language, which will be annotated in the international Universal Dependencies (UD) format and consist of 10 million words.

First, the structure of the corpus (i.e., fiction, nonfiction, newspapers, magazines, and legal documents) will be discussed. Legal issues are particularly important here as full texts will have to be included, not just excerpts. Second, the tokenization strategy (boundaries of typical and atypical lexical units) will be presented. Third, the syntactic annotation techniques will be reviewed. Adaptation of the widely used UD standard to the Lithuanian language poses a number of challenges for its theoretical and practical applicability to Lithuanian. This international standard is quite different from the PML annotation of the ALKSNIS corpus, which basically followed the traditional notion of grammar. Fourth, other tools applied for syntactic analysis and editing of syntactic trees will be presented.

Acknowledgements

Project “Morphologically and syntactically annotated text models for training (gold standards)” (No. 02-098-K-0001). The project is financed by the Recovery and Resilience Facility (RRF) plan “New Generation Lithuania” and the State Budget of the Republic of Lithuania; see https://sitti.vdu.lt/morfologiskai-ir-sintaksiskai-anotuotu-tekstynu-modeliai-dirbtinio-intelekto-apmokymui/.

References

Bielinskienė, Agnė, Boizou, Loic, Kovalevskaitė, Jolanta, Rimkutė, Erika. 2016. Lithuanian Dependency Treebank ALKSNIS. In *Proceedings of the Seventh International Conference Baltic HLT 2016*. Amsterdam: IOS Press, 107–114. Available at: http://ebooks.iospress.nl/volumearticle/45523

Bielinskienė, Agnė, Kovalevskaitė, Jolanta, Rimkutė, Erika. 2021. Grammatical patterns in the corpus-driven Lexical Database of Lithuanian. *Language: Meaning and Form (Valoda: nozīme un forma).* 12, 7–30. https://doi.org/10.22364/vnf.12.01

Brinkutė, Rūta. 2018. *Gramatinių formų pasiskirstymas morfologiškai anotuotame lietuvių kalbos tekstyne*. Kaunas: Vytauto Didžiojo universitetas. Available at: https://portalcris.vdu.lt/server/api/core/bitstreams/db82855c-3f1d-4ca4-8df8-589953211ec5/content

Dabašinskienė, Ineta. 2009. Šnekamosios lietuvių kalbos morfologinės ypatybės. *Acta linguistica Lithuanica*. 60, 1–15.

Kamandulytė-Merfeldienė, Laura. 2018. Nuo buitinės kalbos iki viešojo kalbėjimo: kiekybinis kai kurių leksikos ir gramatikos ypatybių tyrimas tekstynų lingvistikos metodu. *Lituanistica*. 64(4), 255–270. Available at: https://etalpykla.lituanistika.lt/fedora/objects/LT-LDB-0001:J.04~2018~1552548039495/datastreams/DS.002.0.01.ARTIC/content

Kovalevskaitė, Jolanta, Bielinskienė, Agnė, Rimkutė, Erika. 2024. *Tekstynais paremti kalbos vartosenos tyrimai leksikografijoje („Mokomojo lietuvių kalbos vartosenos leksikono" atvejis)*. Kaunas: Vytauto Didžiojo universitetas. https://doi.org/10.7220/9786094676154

Křen, Michal. 2020. Czech National Corpus in 2020: Recent Developments and Future Outlook. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, Marseille: ELRA, 52–57. https://aclanthology.org/2020.cmlc-1.8

Rimkutė, Erika, Kazlauskienė, Asta, Utka, Andrius. 2016. Morphemic Structure of Lithuanian Words. *Open Linguistics*. 2(1), 160–179. https://doi.org/10.1515/opli-2016-0008

Rimkutė, Erika, Bielinskienė, Agnė, Kovalevskaitė, Jolanta, Boizou, Loic, Aleksandravičiūtė, Gabrielė, Brokaitė, Kristina, Utka, Andrius. 2019. *Lithuanian Treebank ALKSNIS (2019-10-24)*, CLARIN-LT digital library in the Republic of Lithuania: http://hdl.handle.net/20.500.11821/21

Saulīte, Baiba, Darģis, Roberts, Grūzītis, Normunds, Auziņa, Ilze, Levāne-Petrova, Kristīne, Pretkalniņa, Lauma, Rituma, Laura, Paikens, Pēteris, Znotiņš, Artūrs, Strankale, Laine, Pokratniece, Kristīne, Poikāns, Ilmārs, Bārzdiņš, Guntis, Skadiņa, Inguna, Baklāne, Anda, Saulespurēns, Valdis, Ziediņš, Jānis. 2022. Latvian National Corpora Collection – Korpuss.lv. In *Proceedings of the 13th International Conference on Language Resources and Evaluation*, Marseille: ELRA, 5123–5129. Available at: http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.548.pdf

THE USE OF LATVIAN CORPORA FOR REVISION AND DEVELOPMENT OF ELECTRONIC DICTIONARY "TĒZAURS"

Laura Rituma, Gunta Nešpore-Bērzkalne, Ilze Lokmane, Lauma Pretkalniņa, Agute Klints & Madara Stāde

Institute of Mathematics and Computer Science,
University of Latvia

Keywords: electronic dictionary "Tēzaurs", Latvian corpora, grammatical information in dictionary, plural-only nouns

"Tēzaurs" is an electronic explanatory Latvian dictionary with ~407 000 entries (Grasmanis et al. 2023, available at: https://tezaurs.lv/). ~106 000 of them contain grammatical information, e.g., part of speech, declension or conjugation class, etc. For declinable words, full inflectional paradigms are given. Also, grammatical information can be added to individual word senses that function only in certain grammatical forms; the grammatical information can indicate the conversion of a word or sense into a different part of speech.

The grammatical information in "Tēzaurs" is obtained both from the original dictionary entry sources and text corpora – mainly the Latvian National Corpora Collection (Saulīte et al. 2022), comprising 39 Latvian corpora of various genres and periods.

When revising "Tēzaurs", an important component is the grammatical information extracted from the corpora, e.g., whether a word is used in its entire paradigm with equal frequency or which of the possible form variants are used, e.g., adverbs are traditionally divided into two groups depending on whether they are gradable or not (Nītiņa, Grigorjevs 2013, 602–603; Kalnača, Lokmane 2021, 322–324). However, corpus data indicates that boundaries between these groups are fuzzy: there are gradation forms that should not be formed due to the semantics of the adverb but are nevertheless used in practice (*optimāli* 'optimally', *mūžīgi* 'forever'). Another noteworthy phenomenon is compound genitives – a special group of Latvian compound nouns that only exist in singular or plural genitive (Kalnača, Lokmane 2016). However, the corpus data shows that some compound genitives are used in other cases (e.g., locative *masveidā* 'massively') and some have even acquired a full paradigm (*ilgtermiņš* 'long term'); furthermore, some compound genitives can be used in both singular and plural forms (*beznosacījuma/beznosacījumu* 'unconditional', *pirmslaulības/pirmslaulību* 'premarital').

The grammatical information in "Tēzaurs" influences the morphological tagging of corpora, as the morphological tagger uses the data from the dictionary (Paikens, Pretkalniņa, Rituma 2024). The tagging, in turn, influences the search query results in the corpora when a specific lemma or grammatical feature is searched.

To demonstrate the use of corpora for revising grammatical information in the "Tēzaurs" entries, we offer an analysis of the **tagging of plural-only nouns**. Only a part of them – the so-called true plurals – are never used in the singular; a larger part is occasionally used in singular, mostly to create a certain stylistic impression or even a different lexical meaning (Nītiņa, Grigorjevs 2013, 341–343). Therefore, plurals in "Tēzaurs" are annotated and displayed differently depending on the corpus data (see Table 1).

Table 1. Words with entry noun in plural in "Tēzaurs"

	Group 1	Group 2	Group 3	Group 4
Description	True plurals – words that are used only in the plural	Plurals with rarely occurring singular forms (usually in colloquial texts; non-standard)	Words usually used in the plural, but singular forms are possible (mostly standard language)	Full paradigm words – the dictionary entry in plural due to the tradition, since the base meaning is plural
Examples	*Limbaži, Ādaži, bēres* 'funeral', *drupas* 'ruins', *ļaudis* 'people', *kāzas* 'wedding', *mokas* 'anguish'	*milti* 'flour', *tauki* 'fat', *bikses* 'trousers', *bailes* 'fear', *slāpes* 'thirst', *brilles* 'glasses', *nepatikšanas* 'trouble'	*sāpes* 'pain', *rūpes* 'care', *rītasvārki* 'dressing gown', *bažas* 'concern', *bēdas* 'sorrow',	*latvieši* 'Latvians', *baptisti* 'Baptists', *kurpes* 'shoes', *roņveidīgie* 'seals', *plaušas* 'lungs'
dictionary entry form	plural	plural	plural	plural
base form (lemma)	plural	plural	singular	singular
inflectional forms shown in entry	plural	plural	singular and plural	singular and plural
forms recognized by tagger	plural	singular and plural	singular and plural	singular and plural
lemma in corpus	plural	plural	singular	singular

To conclude, the addition of grammatical information to "Tēzaurs" is closely connected to actual evidence of existence of certain word forms in the corpus data. Although word usage does not always match standard language norms, our goal is to identify and include all word forms so they can be recognized in search queries and corpora.

Acknowledgments

This study was conducted within the project "Advancing Latvian computational lexical resources for natural language understanding and generation" (No. lzp-2022/1-0443) funded by the Latvian Council of Science.

References

Grasmanis, Mikus, Paikens, Pēteris, Pretkalniņa, Lauma, Rituma, Laura, Strankale, Laine, Znotiņš, Artūrs, Grūzītis, Normunds. 2023. Tēzaurs.lv – the experience of building a multifunctional lexical resource. *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference.*

Paikens, Pēteris, Pretkalniņa, Lauma, Rituma, Laura. 2024. A computational model of Latvian morphology. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).* Torino: ELRA and ICCL, 221–232.

Saulīte, Baiba, Darģis, Roberts, Grūzītis Normunds, Auziņa, Ilze, Levāne-Petrova, Kristīne, Pretkalniņa, Lauma, Rituma, Laura, Paikens, Pēteris, Znotiņš, Artūrs, Strankale, Laine, Pokratniece, Kristīne, Poikāns, Ilmārs, Bārzdiņš, Guntis, Skadiņa, Inguna, Baklāne, Anda, Saulespurēns, Valdis, Ziediņš, Jānis. 2022. Latvian National Corpora Collection – Korpuss.lv. *Proceedings of the 13th Language Resources and Evaluation Conference* (LREC), 5123–5129

Kalnača, Andra, Lokmane, Ilze. 2016. Compound genitives in Latvian. In Lívia Körtvélyessy, Pavol & Salvador Valera (eds.). *Word-Formation across languages.* Newcastle upon Tyne: Cambridge Scholars Publishing, 170–196.

Kalnača, Andra, Lokmane, Ilze. 2021. *Latvian Grammar.* Riga: University of Latvia Press. https://doi.org/10.22364/latgram.2021

Nītiņa, Daina, Grigorjevs, Juris (red.). 2013. *Latviešu valodas gramatika*. Rīga: LU Latviešu valodas institūts.

SEMI-INSUBORDINATION IN BRITISH ENGLISH: DIACHRONIC TRENDS AND FUNCTIONAL INSIGHTS

Xulia Sánchez-Rodríguez

University of Vigo

Keywords: semi-insubordination, diachronic variation, BNC1994, BNC2014

After Evans (2007), the conventionalised use as main clauses of structures which formally resemble subordinate clauses but stand alone with full semantic content is known as 'insubordination', illustrated by the *if*-, *that*- and infinitive clauses in (1)–(3), respectively:

(1) *If you could get that across to her.* (Lastres-López 2020, 51)

(2) ***That** he should have left without asking me!* (Evans 2007, 403)

(3) ***To** think that I was once a millionaire!* (Evans 2007, 404)

Whereas insubordination has been widely investigated (Van Linden, Van de Velde 2014; Sansiñena et al. 2015; Beijering et al. 2019; Lastres-López 2020; Kaltenböck, Keizer 2022; among others), the related phenomenon of so-called 'semi-insubordination' (SIS), illustrated in (4), has received less attention in the literature. Firstly introduced by Van Linden and Van de Velde (2014), insubordination is employed with complement (i.e. dependent) clauses introduced by a head element, as in (4), which is initiated by the adjectival head *lucky*:

(4) *Well, **lucky I'm going to a good news story**.* (Kaltenböck 2021, 133)

Other examples of SIS, according to Kaltenböck's (2021) (only) study in (American) English, consist of a nominal (5), verbal (6) or prepositional (7) head element followed by a (non-)finite subordinate clause:

(5) ***Pleasure** to have you here.* (Kaltenböck 2021, 127)

(6) ***Seems** he chomped down on a sandwich wrap he says contained a, quote "dangerous" substance, specifically an olive pit.* (Kaltenböck 2021, 129)

(7) *Yes, **about time** we got out of this show.* (Kaltenböck 2021, 129)

G. Kaltenböck (2021) also investigates the discourse functions of SIS constructions and identifies two main uses: commenting (8), where the speaker expresses a subjective evaluation of the proposition, and presentative (9), where a proposition is introduced into the discourse:

(8) Unidentified Woman 1: *You're time is up.* (Laughter)
Limbaugh: *Well, my friends, **sorry you ran out of time**, but I'll tell you what it is.* (Kaltenböck 2021, 147)

(9) *Mm, mm, **funny there was a lad here** [...] this chap was here last week he* (Kaltenböck 2021, 148)

Research on SIS remains limited, with questions still unanswered regarding how this phenomenon manifests itself in other varieties of English and its diachronic evolution. This study thus aims to address this gap by exploring SIS in the recent history of British English. Using data from the *British National Corpus*, both BNC1994 (BNC Consortium 2007) and BNC2014 (Love et al. 2017; Brezina et al. 2021), I coded 687 manually-pruned instances for variables such as type of matrix element (adjective, noun, verb, prepositional phrase), lexical head in the matrix element, type of subordinate clause (*how*-, *that*-, *ing*-, *to*-infinitive clauses), time period (BNC1994, BNC2014), medium (written, spoken) and discourse function (anaphoric, exophoric). The findings reveal an ongoing process of functional and lexical productivity of SIS constructions in recent British English. Functionally, the data show a tendency towards lower distributional differentiation among complement-clause types, with growing prevalence of the commenting function. Lexically, the study observes a notable diachronic increase in the range of adjectives used as matrix elements, highlighting the prominence of adjectival matrix elements in SIS constructions. The attested narrowing of choices, as evidenced by the SIS data in terms of both form (fixation of major syntactic slots and decrease of productive lexical types of the head element) and function (pragmatic uses), indicates the emergence of an incipient yet significant process of constructionalisation of semi-insubordination in recent diachrony.

This research is supported by the FPI grant PREP2023-000303 funded by MCIN/AEI/10.13039/5011000110337/FEDER,UE by means of a contract at the University of Vigo.

References

Beijering, Karin, Kaltenböck, Gunther, Sansiñena, María Sol. 2019. Insubordination: Central issues and open questions. In Beijering, Karin, Kaltenböck, Gunther, Sansiñena, María Sol (eds.). *Insubordination. Theoretical and Empirical Issues.* Berlin/Boston: De Gruyter Mouton, 1–28. https://doi.org/10.1515/9783110638288-001

Brezina, Vaclav, Hawtin, Abi, McEnery, Tony. 2021. The Written British National Corpus 2014 – design and comparability. *Text & Talk*. 41(5–6), 595–615. https://doi.org/10.1515/text-2020-0052

British National Corpus Consortium. 2007. *British National Corpus: XML edition*. Oxford: Oxford Text Archive. https://doi.org/10.1515/text-2020-0052

Evans, Nicholas. 2007. Insubordination and its uses. In Nicolaeva, Irina (ed.). *Finiteness: Theoretical and Empirical Foundations*. Oxford: Oxford University Press, 366–431. https://doi.org/10.1093/oso/9780199213733.003.0011

Kaltenböck, Gunther. 2021. *Funny you should say that*. On the use of semi-insubordination in English. *Constructions and Frames*. 13(1), 126–159. https://doi.org/10.1075/cf.00049.kal

Kaltenböck, Gunther, Keizer, Evelien. 2022. Insubordinate *if*-clauses in FDG: Degrees of independence. *Open Linguistics*. 1, 675–698. https://doi.org/10.1515/opli-2022-0212

Lastres-López, Cristina. 2020. Subordination and insubordination in contemporary spoken English. *If*-clauses as a case in point. *English Today*. 36(2), 48–52. https://doi.org/10.1017/S026607841900021X

Love, Robbie, Dembry, Claire, Hardie, Andrew, Brezina, Vaclav, McEnery, Tony. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*. 22(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Sansiñena, María Sol, De Smet, Hendrik, Cornillie, Bert. 2015. Between subordinate and insubordinate. Paths toward complementizer-initial main clauses. *Journal of Pragmatics*. 77, 3–19. https://doi.org/10.1016/j.pragma.2014.12.004

Van Linden, An, Van de Velde, Freek. 2014. (Semi-)autonomous subordination in Dutch: Structures and semantic-pragmatic values. *Journal of Pragmatics*. 60, 226–250. https://doi.org/10.1016/j.pragma.2013.08.022

SYNTAX ERRORS IN THE LATVIAN GRAMMATICAL ERROR CORRECTION AND FLUENCY CORPUS "NORMA"

Baiba Saulīte, Agute Klints & Roberts Darģis

Institute of Mathematics and Computer Science, University of Latvia

Keywords: syntax errors, Latvian language, error-annotated corpus, grammatical error correction (GEC), language mastery

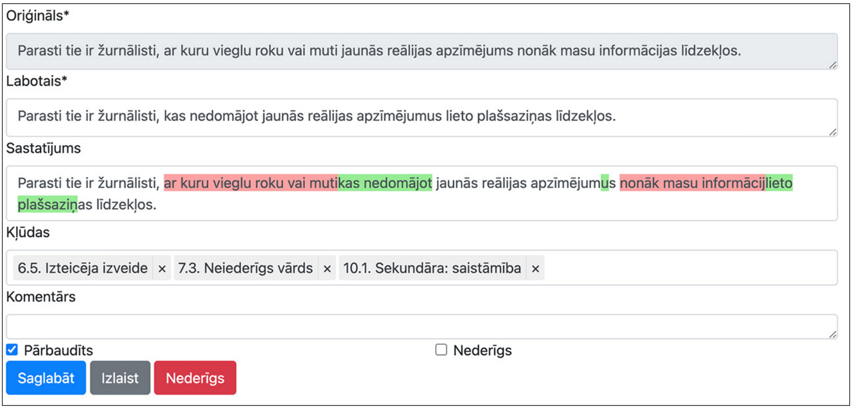
This abstract presents the creation of the Latvian grammatical error correction (GEC) and fluency corpus "Norma" and describes in detail the syntactic errors in this corpus.

"Norma" is a semi-automatically error-annotated corpus of texts produced by native speakers of Latvian, in which the most common errors in the Latvian language are documented, corrected, annotated and explained. The corpus data will be used to analyse how language errors affect the grammatical system of the Latvian language and to develop state-of-the-art corpus-based guidelines for improving the quality of written language. An error-annotated corpus is also needed for high-level grammar checkers, which could detect complex structural errors in addition to low-level spell-checkers.

A special corpus platform is being developed to facilitate data collection, annotation and analysis (see Fig. 1). The corpus and further information about this project can be found on the website https://norma.korpuss.lv/.

Error types used in other GEC or learner corpora are usually limited to GEC errors (Bryant et al. 2017; Náplava et al. 2022), covering only errors such as spelling, punctuation, some verb, noun, adjective forms and word order. However, edits of texts produced by native speakers are rich in fluency errors related to the inaccurate use of lexical or structural units. Specifically, these edits relate to correcting miscollocations and calques, stylistically inappropriate words, and rewriting syntactic structures that contain dysfluencies or that sound awkward to a native speaker (Syvokon et al. 2023).

The initial set of Latvian error types includes 10 error types with several subtypes: (1) Typographical formatting; (2) Spelling; (3) Derivation; (4) Word formation; (5) Punctuation; (6) Syntax; (7) Inaccurate use of lexical or structural units; (8) Text structure; (9) Other errors; (10) Secondary errors.



Oriģināls*

Parasti tie ir žurnālisti, ar kuru vieglu roku vai muti jaunās reālijas apzīmējums nonāk masu informācijas līdzekļos.

Labotais*

Parasti tie ir žurnālisti, kas nedomājot jaunās reālijas apzīmējumus lieto plašsaziņas līdzekļos.

Sastatījums

Parasti tie ir žurnālisti, ar kuru vieglu roku vai mutikas nedomājot jaunās reālijas apzīmējumus nonāk masu informācijlieto plašsaziņas līdzekļos.

Kļūdas

6.5. Izteicēja izveide × 7.3. Neiederīgs vārds × 10.1. Sekundāra: saistāmība ×

Komentārs

☑ Pārbaudīts ☐ Nederīgs

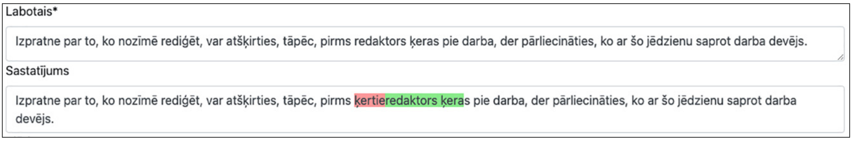
Figure 1. Corpus annotation tool and editing interface:
(1) original, (2) edited, (3) alignment, (4) error types

Table 1. Most common error types in "Norma"

Main error category	**Sentences**	**Percentage**
7 Use of lexical or structural units	2283	28%
1 Typographical formatting	1431	18%
6 Syntax	1410	17%
5 Punctuation	909	11%
2 Spelling	679	8%
10 Secondary errors	574	7%
8 Text structure	250	3%
4 Word formation	216	3%
3 Derivation	210	3%
9 Other errors	179	2%

Over 4000 sentences have been annotated, each with one or more error types. The most common types of errors are (1) inappropriate use of lexical or structural units; (2) typographical formatting; and (3) syntactic errors (see Table 1).

There are 1410 syntax errors annotated in this version of the Norma corpus. These errors are grouped into seven subtypes: 6.1 agreement; 6.2 secondary predicates; 6.3 negation; 6.4 word order; 6.5 predicate; 6.6 clause division; 6.7 coordination; and 6.8 type of subordinate clause. The most popular and varied subtype is predicate formation (600), e.g., an infinitive instead of a finite verb (see Fig. 2).



Labotais*

Izpratne par to, ko nozīmē rediģēt, var atšķirties, tāpēc, pirms redaktors ķeras pie darba, der pārliecināties, ko ar šo jēdzienu saprot darba devējs.

Sastatījums

Izpratne par to, ko nozīmē rediģēt, var atšķirties, tāpēc, pirms ķertieredaktors ķeras pie darba, der pārliecināties, ko ar šo jēdzienu saprot darba devējs.

Figure 2. Sentence *The understanding of what it means to edit can vary, so before* ***the editor gets to work****, it is worth clarifying what the employer means by this concept.* Correction: verb *ķerties* 'to get' in the infinitive form was replaced by the 3rd person verb *ķeras* 'gets', and the subject *redaktors* 'editor' was added.

Syntax errors, such as inaccuracies in agreement or negation formation, are closely related to a language's grammatical system. Corpus "Norma" and error analysis can help proofreaders, writers and others to master language fluency.

Acknowledgement

This work was supported by the Latvian Council of Science grant "Common Writing Errors in Latvian" (No. lzp-2023/1-0481).

References

Bryant, Christopher, Felice, Mariano, Briscoe, Ted. 2017. Automatic annotation and evaluation of error types for grammatical error correction. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (1). Vancouver, Canada: Association for Computational Linguistics, 793–805.

Náplava, Jakub, Straka, Milan, Strakov'a, Jana, Rosen, Alexandr. 2022. Czech grammar error correction with a large and diverse corpus. *Transactions of the Association for Computational Linguistics*. 10, 452–467.

Syvokon, Oleksiy, Nahorna, Olena, Kuchmiichuk, Pavlo, Osidach, Nastasiia. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, 96–102. Dubrovnik, Croatia. Association for Computational Linguistics.

CASE CHOICE IN UKRAINIAN VOCATIVE EXPRESSIONS: A STUDY OF PARLIAMENTARY TRANSCRIPTS (1990–2024) ANNOTATED WITH UNIVERSAL DEPENDENCIES

Maria Shvedova & Arsenii Lukashevskyi

National University "Kharkiv Polytechnic Institute"; & University of Jena and National University "Kharkiv Polytechnic Institute"

Keywords: Ukrainian, direct address, vocative case, variation, Universal Dependencies

In Ukrainian, as in Bulgarian, Polish, Czech, and Slovak, the dedicated Slavic vocative case has been preserved. It is regularly used in direct address forms for singular animate masculine and feminine nouns. Less commonly, vocative forms are used by inanimate nouns and neuter nouns, which are typically used in the nominative case. Although the modern standard requires the obligatory use of the dedicated form for declined masculine and feminine animate nouns (Ukrainian Orthography 2019), in actual practice, they are also often used in the nominative (Ponomariv 1999). This variation has not yet been properly studied across a large corpus, as no Ukrainian reference corpus has syntactic annotation required to distinguish between the nominative in the address and the subject nominative.

To research the case choice in direct address expressions, a Universal Dependencies corpus of Ukrainian was used as source. UD_Ukrainian_ParlaMint contains over 509 instances of direct address, providing enough data for the model trained on it to more accurately identify the vocative as dependency relation. This model was published in UD at the end of 2024 (UD Ukrainian ParlaMint 2024). Using it, we annotated the corpus of Ukrainian parliamentary transcripts from 1990 to 2024, totalling 88 million tokens, from which we obtained over 128 thousand contexts with the vocative relation. The accuracy of the data was manually verified. We did not include in these data vocative expressions consisting of a single surname, as the model often fails to distinguish between masculine and homonymous feminine forms that do not decline. We also did not include examples with indeclinable nouns (e.g., *pani* 'madam', *Jerry*, *Geo*), neuter nouns, nouns that decline according to the adjectival paradigm (e.g., *včenyj* 'scholar'), and plurals, because they do not have a special vocative form.

Our data show an approximately equal distribution of nominative and vocative forms in addresses, except for periods with 100% vocative, which were likely edited and thus excluded from the analysis (see Fig. 1).

For the study, 152 lemmas were selected based on two criteria: (1) a minimum of 10 occurrences in address positions, and (2) a 95% confidence interval for the vocative share with a width not exceeding 0.5.

The material reveals the following patterns: (1) the share of vocatives is higher in more frequent lemmas and decreases with lower frequency; (2) vocatives are more frequent in male names than in female names; (3) stem endings influence vocative usage—male names ending in *-ij* (e.g., *Jurij*, *Andrij*) show a lower share of vocatives compared to male names overall; (4) pragmatic factors also play a role—vocatives are more frequent in solemn addresses and after words expressing respect (e.g., *šanovnyj*).

We also examined cases with variable vocative endings (*Ihore/Ihorju, Oleže/Olehu*), as well as instances where, in multi-noun addresses, only one noun appears in the vocative while the other remains in the nominative, e.g., *pane* (Voc) *ministr* (Nom).

The Universal Dependencies annotation opens opportunities for exploring other homonymous grammatical forms in Ukrainian, as well.

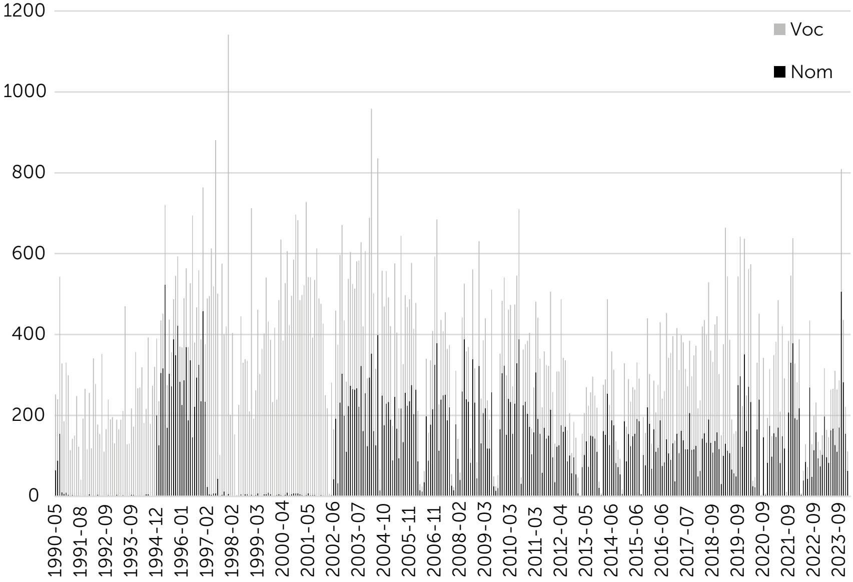


Figure 1. Distribution of the vocative and nominative cases in sentences with direct address by month.

Acknowledgments

We are grateful to UniDive COST Action CA21167 for the shared resources and support in creating the UD treebank and for the inspiring scientific events.
The research was partially funded by the Humboldt Foundation.

References

Ukraïns'kyj pravopys [Ukrainian Orthography]. 2019. Kyïv: Naukova dumka.

Ponomariv, O. 1999. *Kul'tura slova: movnostylistyčni porady* [The Culture of Words: Linguistic and Stylistic Advice]. Available at: http://ponomariv-kultura-slova.wikidot.com/

UD Ukrainian ParlaMint. 2024. Available at: https://universaldependencies.org/treebanks/uk_parlamint/index.html

PLUPERFECT IN LARGE SLAVIC CORPORA: PRODUCTIVITY AND LEXICAL RESTRICTIONS

Dmitri Sitchinava

University of Potsdam

Keywords: pluperfect, Slavic languages, lexical restrictions, productivity

The verbal category of pluperfect is construed both as a combination of two simpler categories (Past-in-the-Past and/or Perfect-in-the-Past) and as a non-compositional category in its own right (cf. Dahl 1985). Pluperfect is polyfunctional, featuring such uses as cancelled result, "discontinuous past" (Plungian, Auwera 2006), irrealis (as *If you had come, we would not lose* in English), experientials or evidentials.

Research questions. We research the variation of productivity vs. lexicalization of pluperfect among Slavic languages (see Sitchinava 2019 and bibliography referred therein on pluperfect in Slavic). In which languages is the pluperfect a productive tense, occurring with the majority of the lexicon, and where is it intertwined with specific verb classes or even prefers certain concrete lexemes? Additionally, we address the question of which verb classes are used reluctantly with the pluperfect, despite their overall frequency in the language, and what this reveals about the semantics of the form.

Data and methods. The non-anterior uses of pluperfect are attested in many Slavic languages and varieties. Pluperfect constructions have low text frequencies except for Bulgarian, Macedonian, and Sorbian, where a strong, albeit not absolute tendency towards English-like consequence of tenses is found. Beyond these uses, the pluperfect, as researched in smaller parallel corpora of fiction, subtitles, and other sources, is not well represented, with particularly scarce data for West Slavic. Large comparable corpora thus offer a more reliable resource for capturing these forms than parallel corpora.

We have drawn and analysed frequency lists of verbs used in pluperfect in the available large web corpora (two languages per subgroup; for Russian, the residual *bylo* construction is counted). The following corpora were used: Ukrainian – GRAC; Modern Russian – RNC; Polish – PlTenTen2019; Slovak – Araneum Slovacum Maius; Serbian – MaCoCu; Bulgarian – BgTenTen. On a graph showing the frequency of the top 100 verbs, the steepest curve corresponds to languages where a smaller number of verbs account for the largest share of pluperfect uses. It can also be quantified by the entropy of the distribution of the construction among the most frequent lexemes: higher entropy scores indicate greater productivity. These data align with subgroups (e.g., distribution in Serbian clusters with Bulgarian despite different absolute frequencies).

Results. Comparing the data from contemporary corpora with the historical data (Old East Slavic and Middle Russian – both RNC), we suggest that pluperfect further increased its productivity in South Slavic, as compared to the oldest texts, but became lexicalized in East Slavic, choosing mainly modals and verbs of attempt.

A separate thread of research concerned lexical restrictions. Cross-linguistically, the less frequent in the context of pluperfect are accomplishment verbs with semantics of success (like Ukrainian *poščastyty* (impers.) 'be lucky', Old East Slavic *poběditi, odolěti* 'win a battle', *vŭzmošči* 'manage to do', Slovak *vyhrat'* 'win a game', *stihnút'* 'manage, catch', Polish *dokonać* 'accomplish'), the situations construed as a result that cannot be reverted or challenged. Imperfectives/statives and factive verbs also tend to be restricted cross-linguistically. These data show pragmatic tendency to use pluperfect as a marker of unsuccessful or cancelled result.

References

Dahl, Östen. 1985. *Tense and Aspect Systems*. Oxford: Basil Blackwell.

Plungian, Vladimir, Auwera, Johan van der. 2006. Towards a typology of discontinuous past marking. *Language Typology and Universals*. 59(4), 317–349.

Sitchinava, Dmitri. 2019. Slavic Pluperfect: loci of variation [in Russ.]. *Voprosy jazykoznanija*. 1, 30–57.

INDEFINITE VERB FORMS AS SECONDARY PREDICATES IN FINNISH AND LITHUANIAN

Lucia Šalagová

Masaryk University, Brno

Keywords: secondary predicate, subordinate action, indefinite verb form, construction, corpus study

Although Finnish and Lithuanian are two different languages genealogically and typologically, some of their features have been discussed and compared in the past few years. One of them, i.e. the notion of partitivity and grammatical cases used in object position, was discussed by Maija Tervola (2015), Marja Leinonen (2016) and Asta Laugalienė (2021). On the other hand, the topic related to indefinite verb forms and their usage in these languages deserves more attention. A systematic overview of all Finnish and Lithuanian indefinite verb forms and their syntactic functions were provided by Lase Bergroth (2009). However, no other researchers have discussed this matter in greater depth.

Indefinite verb forms have a wide range of syntactic uses in both languages. The syntactic function I focus on in my study is as secondary predicates within a sentence. Indefinite verb forms appear in various secondary predication constructions used in Finnish as well as in Lithuanian. At least to some extent, many of these constructions tend to work in a similar manner. However, the similarities and differences of secondary predicates in Finnish and Lithuanian were not discussed in further detail yet.

In my study I focused on explaining in what ways and to what extent particular secondary predication constructions of these languages correspond to each other, and what differences can be observed in their meaning and usage. I did so by focusing on the following aspects of the sentences in which the secondary predication constructions occur: *the type of subject of the main and subordinate action; the type of a sentence* (there are special types of sentences with specific syntactic and grammatical features in both Finnish and Lithuanian that are often used while talking about weather or mood and feelings.); *passive voice; temporal relation of main and subordinate action; syntactic function of the construction; and negation*. The constructions I researched in my study were Lithuanian constructions occurring in so called participial sentences (term used by Ambrazas 2001), some cases of indirect speech and constructions denoting feelings and sensations; and Finnish temporal, modal, final and referative constructions. I answered the following questions:

1. To what extent do certain types of Finnish and Lithuanian secondary predication constructions correspond to each other considering all the individual aspects mentioned above?
2. In what ways (aspects) do the constructions work in the same or similar manner? Why?
3. Are some of the constructions even comparable in some of the aspects? Why? Why not?
4. Which Finnish and Lithuanian constructions are closest to each other in meaning and usage? Why?

I have chosen an empirical approach for my research. The data was acquired from two separate corpora: the annotated corpus of the contemporary Lithuanian language (DLKT – *Dabartinės lietuvių kalbos tekstynas*) and the Finnish parallel corpus, which is part of the multilingual parallel corpora collection (*InterCorp*) approachable via the *KonText* interface on the *Czech national corpus* webpage.

The differences among these constructions arise mainly from the typological characteristics of these two languages. Finnish temporal construction and Lithuanian constructions in participial sentences evince the most similarities despite the typology of these languages.

References

Ambrazas, Vytautas (ed.). 2001. *Lithuanian Grammar*. Vilnius: Baltos lankos.

Bergroth, Lase. 2009. *Suomen ja Liettuan kielen nominaalimuodoista: vertaileva näkökulma* [On the nominal FORMS of Finnish and Lithuanian verbs: a comparative perspective]. Helsinki: University of Helsinki.

Laugalienė, Asta. 2021. PARTITIVITY and object marking in Finnish and Lithuanian. *Philologia Estonica Tallinnensis*. 5, 235–267. https://doi.org/10.22601/PET.2020.05.08

Leinonen, Marja. 2016. Partitives and GENITIVES in negated sentences in Finnish, Latvian and Lithuanian. *Valoda: nozīme un forma*. 7, 89–103. https://doi.org/10.22364/VNF.7.9

Tervola, Maija. 2015. Comparing object case alternation in Finnish and Lithuanian. In Junttila, Santeri (ed.). *Contacts Between the Baltic and Finnic Languages*. Helsinki: Suomalais-Ugrilainen Seura, 205–245.

EXPLICITNESS/IMPLICITNESS OF THE CZECH PRONOUN *JÁ* AND ITS LATVIAN EQUIVALENT *ES* IN TEXTS: A CONTRASTIVE CORPUS-BASED STUDY

Michal Škrabal, Pavel Štoll & Aiga Veckalne

Charles University, Prague; Charles University, Prague; Ventspils University of Applied Sciences, Ventspils

Keywords: 1st person singular pronoun, pro-drop languages, Latvian, Czech, grammar

Our paper is inspired by Svetla Čmejrková's study "Personal pronoun já ('I') in Czech texts" (Čmejrková 2007), presented at the first Grammar & Corpora conference held in Prague in 2005. We have returned to this topic after twenty years – not only to compare the results of earlier and current research (thus tracing potential shifts and/or trends in Czech) but also to examine cross-linguistic comparison. We chose Latvian for several reasons, not only because of this year's conference venue. Latvian and Czech show a considerable number of grammatical parallels, including a higher degree of implicitness of some personal pronouns in texts compared to other languages where their explicit use is more or less obligatory. This is not the case for so-called *pro-drop languages* such as Czech or Latvian: the personal endings of verb forms usually make it clear which person is being referred to, making the use of the personal pronoun redundant. As a result, its explicit presence in a text is then perceived as marked and specific. Yet, it is one of three different means for expressing verbal person (besides person endings and finite forms of auxiliary verbs), "in contexts where verb forms lack person endings, usually, in the conditional, oblique, or debitive moods" (Kalnača, Lokmane 2021, 227).

In our paper, we want to compare the degree of absence/presence of the 1st person singular pronoun já/es in the Czech and Latvian corpora (SYN2020, ORAL v1, ONLINE; LVK2022, LRK2013/BalsuTalka), including the parallel one (InterCorp v16-en), and describe the aforementioned specific cases of this pronoun's explicit use in both languages (cf. Kalnača 1999, 2010). Furthermore, we contrast the situation in the corpora with the description in Czech (Štícha 2013; Cvrček 2015) and Latvian (Endzelīns 1951; MLLVG II 1962; Kalnača, Lokmane 2021; Nītiņa, Grigorjevs 2013; LVG 2015) grammars.

Methodologically, we proceed as follows: to be able to compare our results with those of Čmejrková (2007), we rely on her set of CQL queries, which we also adapt for Latvian. As for the parallel data, we work with Czech originals translated into Latvian and vice versa, noting cases where the personal pronoun a) is equally present (or absent) in both languages, b) is present in the original but not in the translation, c) is absent in the original but occurs in the translation.

References

Cvrček, Václav et al. 2015. *Mluvnice současné češtiny. 1/ Jak se píše a jak se mluví.* Praha: Karolinum.

Čmejrková, Světla. 2007. Osobní zájmeno 1. osoby v českém textu.In Štícha, František, Šimandl, Josef (eds.). *Gramatika a korpus / Grammar & Corpus 2005.* Praha: ÚJČ AV ČR, 31–41.

Endzelīns, Jānis. 1951. *Latviešu valodas gramatika.* Rīga: Latviešu valodas institūts.

Kalnača, Andra. 1999. Verba personas kategorija un darbības visparinājums. *Linguistica Lettica*, 5. sēj. Rīga: Latvijas Universitāte, 60–71.

Kalnača, Andra. 2010. Personas formu semantika latviešu valodā. *Valoda – 2010. Valoda dažādu kultūru kontekstā*. 20, 199–206.

Kalnača, Andra, Lokmane, Ilze. 2021. *Latvian Grammar.* Riga: University of Latvia Press.

LVG. 2015. *Latviešu valodas gramatika*. Rīga: LU Latviešu valodas institūts.

MLLVG II. 1962. *Mūsdienu latviešu literārās valodas gramatika II. Sintakse*. Rīga: LPSR Zinātņu akadēmijas izdevniecība.

Nītiņa, Daina, Grigorjevs, Juris (eds.). 2013. *Latviešu valodas gramatika.* Rīga: LU Latviešu valodas institūts.

Štícha, František et al. 2013. *Akademická gramatika spisovné češtiny.* Praha: Academia.

ON THE SEMANTIC FEATURES OF THE REFERENT AND THEIR INFLUENCE ON THE CHOICE OF THE DEMONSTRATIVES IN THE LITHUANIAN SUBDIALECT OF KRETINGIŠKIAI

Aušrinė Tverskytė & Gintarė Judžentytė-Šinkūnienė

Vilnius University

Keywords: deictic words, semantic features, referent, Kretingiškiai subdialect, Lithuanian

Semantic features are usually discussed in two different domains: 1) deictic features, which indicate the location of the referent in the speech situation; 2) qualitative features, which classify the referent (Lyons 1977, Rauh 1983, Diesel 1999). The deictic features are known as the distance-based approach and have been the subject of extensive research (Lyons 1977, Fillmore 1997, Diessel 1999, 2005, Coventry et al. 2008, Tóth et al. 2014, Reile 2015, etc.). The qualitative features are also given attention by the researchers (Hanks 1990, 2009, Jarbou 2010, Rocca, Tylén, Wallentin 2019), but to a much lesser extent than the quantitative features. Specifically in the case of Lithuanian deictics such studies do not exist at all.

This paper aims to investigate the effect of semantic referent features on the choice of deictic words in one of Northern Samogitian subdialects, i.e., Kretingiškiai subdialect. Based on the study by Rocca, Tylén and Wallentin (2019), the following semantic features were chosen: 'animate'/'inanimate', 'harmful'/'harmless', and 'big'/'small'. The experiment to investigate the influence of the chosen semantic referent properties was carried out in 2024. People from the study area were interviewed using an electronic questionnaire published on the Qualtrics Experience Management Platform. The questionnaire included 40 words taken from the study by Rocca, Tylén, Wallentin (2019) and translated from English into Lithuanian. Participants had to assign one of four proximal or distal demonstrative pronouns *šitas* (-a) 'this', *šitai tas* (-a) 'this', *tas* (-a) 'that', and *ten tas* (-a) 'that', or the adverbs of place *čia* 'here', *šitai* 'here', *ten* 'there', and *tenai* 'there' to each given referent name.

A mini corpus was created from the survey answers provided by the 783 participants representing the chosen subdialect. This mini corpus, compiled from responses across nearly the entire Kretingiškiai area, enabled a detailed linguistic analysis of deictic word choice among different generations and locations. The analysis revealed that, in addition to physical or psychological distance, the semantic characteristics of the referent – such as being 'harmful' or 'big' – influence the use of distal deictics more than referents described as 'harmless' or 'small'. These findings align with the results reported by Rocca, Tylén, and Wallentin (2019) for English, Danish, and Italian speakers.

References

Coventry, Kenny R., Valdés, Berenice, Castillo, Alejandro, and Guijarro-Fuentes, Pedro. 2008. Language within your Reach: Near–Far Perceptual Space and Spatial Demonstratives. *Cognition*. 108(3), 889–95. https://doi.org/10.1016/j.cognition.2008.06.010

Diessel, Holger. 1999. *Demonstratives. Form, Function and Grammaticalization*. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.42

Diessel, Holger. 2005. Demonstrative Pronouns – Demonstrative Determiners. In Dryer, Matthew, Haspelmath, Martin, Gil, David, and Comrie, Bernard (eds.). *World Atlas of Language Structures*,. Oxford: Oxford University Press, 170–173.

Fillmore, Charles J. 1997. *Lectures on Deixis*. Stanford: CSLI Publications.

Hanks, William F. 2005. Explorations in the Deictic Field. *Current Anthropology*. 46(2), 191–220. https://doi.org/10.1086/427120

Jarbou, Samir Omar 2010. Accessibility vs. physical proximity: An analysis of exophoric demonstrative practice in spoken Jordanian Arabic. *Journal of Pragmatics*. 42(11), 3078–3097. https://doi.org/10.1016/j.pragma.2010.04.014

Lyons, John. 1977. *Semantics*. 2. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511620614

Rauh, Gisa. 1983. *Essays on Deixis*. Tübingen: Narr.

Reile, Maria. 2015, Space and Demonstratives: An experiment with Estonian exophoric demonstratives. *Journal of Estonian and Finno-Ugric Linguistics*. 6(2), 137–165. https://doi.org/10.12697/jeful.2015.6.2.06

Rocca, Roberta, Tylén, Kristian, Wallentin, Mikkel. 2019. *This* shoe, *that* tiger: Semantic properties reflecting manual affordances of the referent modulate demonstrative use. *Plos One*. https://doi.org/10.1371/journal.pone.0210333

Tóth, Enikő, Csatár, Péter Banga, Arina. 2014. Exploring Hungarian and Dutch Gestural Demonstratives. In Veselovská, Ludmila, Janebová, Markéta (eds.). *Complex Visibles out there. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*. Olomouc: Palacký University, 607–627.

‘MID-ACTION REPORTS’ WITH *-NE/-TE* IMPERSONALS: COMPLEX CONSTRUCTIONS IN ONLINE, INTERACTIVE REGISTERS OF POLISH

Piotr Wyroślak

Université Paris 8 Vincennes – Saint-Denis

Keywords: Polish, online registers, complex constructions, impersonals, lists

In the presentation, the term *mid-action report* is used to describe constructional patterns consisting of three subsequent components: 1. “checklist”: a participial construction expressing a successfully completed task – or a number of coordinated participles; 2. “linking expression” based on a marker of logical or temporal consequence; 3. “next task” expressing the action to be undertaken in the nearest future. The pattern can be summarised with the formula ‘X:done, now time for Y’:

(1) (X/Twitter: @maciekkot7, Maciej Kot, 1:59 PM · Nov 4, 2015)
Konferencja PZN przed sezonem 😊 Pogadane, teraz czas na działanie i realizacje słów 😎 #kraków #pzn…[link]

‘Pre-season conference of PZN [Polish Ski Association] 😊 Chatting done, now time for action and making the words become reality 😎 #kraków #pzn…’ [link]

Pogadane,	*teraz*	*czas*	*na*	*działanie*
chat.PPART.NEUT	now	time.NOM.SG	on	ACT.GER.ACC.SG

i	*realizacje*	*słów*
CONJ	realisation.ACC	word.GEN.PL

‘Chatting:done, now time for action and making the words become reality 😎’

(2) (X/Twitter: @gdansk, Miasto Gdańsk, 10:00 AM · Apr 17, 2022)
Pojedzone, poleżakowane, więc pora na... poszukiwania! 😅🔍 Dacie radę odnaleźć wszystkie jajka, które ukryliśmy na zdjęciu?🥚Napiszcie w komentarzu, ile pisanek udało Wam się odszukać 👇 [Graphic]

‘**Done eating and lounging, so it’s time for... a hunt**! 😅🔍 Can you spot all the eggs hidden in the picture?🥚Let us know how many painted eggs you’ve found 👇 #gdansk #ilovegdn #ciekawostkigdn [‘gdnfunfacts’]’

The aim of the study is to analyse the substantial formal variation that this pattern exhibits in Polish within its component parts. A 'checklist' may be composed of a different number and different kinds of participles, including *-ne/-te* participle forms that, as in the examples above, can be found in impersonal, non-agreeing uses (cf. e.g., Brajerski 1979, 96, Kibort 2011, Niekrewicz 2024, Słoń 2002). Furthermore, multiple elements can serve as 'linking expressions' – including *to teraz* 'now', *czas na* 'time for', *pora na* 'time for', *można* '[it is] possible', and the combinatorial possibilities between them, such as *to teraz czas na, więc można*, *więc czas na*, *więc pora na* 'so [it is] time for'.

These patterns of variation will be described based on the data from a corpus of Twitter/X posts including *po*-prefixed *-ne/-te* forms, collected using Twitter API for the period up to October 2023. I will discuss the distributional differences between the subsets of tweets discerned based on the presence of specific linking expressions. I will consider the length of the chains of coordinated participles in the "checklists", the participle types, their attraction to a given linking marker, as well as the forms heading the "next task" component. The discussion of the latter problem will be aided by the analysis of the skip-grams generated for the subsets of tweets, thereby providing an example of the application of skip-gram analysis to a highly schematic construction found in online registers.

References

Brajerski, Tadeusz. 1979. Geneza orzeczeń typu (z)jedzono i (wy)pito. *Język polski.* 59(2), 84–98.

Kibort, Anna. 2011. The elephant in the room: the impersonal *-ne/-te* construction in Polish. In Andrej L. Malchukov & Anna Siewierska (eds.). *Impersonal Constructions: A Cross-Linguistic Perspective* (Studies in Language Companion Series 124), 357–394. Amsterdam: John Benjamins.

Niekrewicz, Agnieszka Anna, 2024. Pobadane, postudiowane..., czyli o ekspansji pozornych imiesłowów biernych we współczesnej polszczyźnie. *Język polski.* (3), 74–87.

Słoń, Anna. 2002. Pragmatic aspects of translating constructions with a defocused instigator. In B. Lewandowska-Tomaszczyk & M. Thelen (eds.). *Translation and Meaning.* Part 6, 297–309. Maastricht: Hogeschool Zuyd

TAD 'THEN' IN LATVIAN CORPORA

Evelīna Zilgalve

University of Latvia

Keywords: particles, adverbs, discourse markers, discourse structure, pragmatic functions

Depending on its functions in the utterance, the word *tad* 'then' is able to belong to two parts of speech: particles and adverbs. As an adverb, it indicates (1) a period of time that is mentioned previously; (2) a moment of time that occurs after another period of time; or (3) a situation or conditions that contribute to the implementation of something (Tēzaurs):

(1) *Tas notika janvārī. Tad man bija eksāmeni.*
'It happened in January. Then I had exams.'

(2) *Vispirms mēs devāmies uz kino. Tad uz restorānu.*
'First, we went to the cinema. Then we went to the restaurant.'

(3) *Ja līs, tad es paņemšu lietussargu.*
'If it rains, then I will take the umbrella.'

As an emphasizing particle (Kalnača, Lokmane 2021), *tad* appears both in (4) general and (5) specific questions or (6) declarative, exclamatory or imperative utterances showing some kind of prior assumption or knowledge of the speaker (see Zilgalve 2023, Theiler 2020; Haselow 2011):

(4) *Vai tad tu to nezināji?*
'Did you not know that?
>> The speaker thought the addressee knew that.

(5) *Ko tad tu tur darīji?*
'What did you do there then?'
>> The speaker assumes that the addressee did not have anything to do there.

(6) *– A. Man ir vajadzīga palīdzība.*
– B. Tad runā!
' – A. I need help.
– B. Speak then!'
>> B encourages A to speak without further hesitation.

In the Latvian corpora LVK2022 (Levāne-Petrova et al. 2022), there are 155 023 hits using a simple search of *tad*, and 12 613 hits of *tad* as a particle, using CQL. However, after observing these examples, it becomes obvious that *tad* in many of these utterances functions as an adverb or partly desemanticised adverb (Zilgalve 2013, 2014), and the usage of *tad* is broader than the meanings described in grammars and dictionaries.

This paper deals with the functions of *tad*: the analysis of the data of Latvian corpora shows a broader usage of *tad* in various contexts, like marking the beginning of the main clause after consecutive clause (*lai arī* – *tad* 'although – then') or expressing comparison with *ja* – *tad* 'if – then' constructions.

References

Haselow, Alexander. 2011. Discourse marker and modal particle: The functions of utterance-final *then* in spoken English. *Journal of Pragmatics*. 43, 3603–3623. https://doi.org/10.1016/j.pragma.2011.09.002

Kalnača, Andra, Lokmane, Ilze. 2021. *Latvian Grammar*. Riga: University of Latvia Press. https://doi.org/10.22364/latgram.2021

Levāne-Petrova, Kristīne, Darģis, Roberts, Pokratniece, Kristīne, Lasmanis, Viesturs Jūlijs. 2023. Balanced Corpus of Modern Latvian (LVK2022). *CLARIN-LV digital library at IMCS, University of Latvia*. Available at: http://hdl.handle.net/20.500.12574/84

Spektors, Andrejs et al. 2024. Tēzaurs.lv 2024 (Autumn edition). *CLARIN-LV digital library at IMCS, University of Latvia*. Available at: http://hdl.handle.net/20.500.12574/110

Theiler, Nadine. 2020. DENN as a highlighting-sensitive particle. *Linguistics and Philosophy*. 44, 323–362. https://doi.org/10.1007/s10988-019-09290-7

Zilgalve, Evelīna. 2013. Partikulas gramatizēšanās aspektā. *Valoda: nozīme un forma*. 3, 142–158. https://doi.org/10.22364/vnf.3

Zilgalve, Evelīna. 2014. Vārdu pāra *ja – tad* funkcionālā analīze. *Valoda: nozīme un forma*. 4, 160–176. https://doi.org/10.22364/vnf.4

Zilgalve, Evelīna. 2023. Partikulas *tad* pragmatiskie aspekti. *58. prof. Artura Ozola dienas starptautiskās zinātniskās konferences „Gramatika un vārddarināšana" referātu tēzes.* Rīga: LU Akadēmiskais apgāds, 78–79. https://doi.org/10.22364/aoszk.58.tk

10th International Conference
“Grammar and Corpora”, Book of Abstracts

University of Latvia Press
Aspazijas bulvāris 5–132, Riga, LV-1050, Latvia
www.apgads.lu.lv
Online bookstore: gramatas.lu.lv

G&C

ISBN 978-9934-36-400-6

9 789934 364006 >

FLPP
FUNDAMENTAL AND
APPLIED RESEARCH
PROJECTS

UNIVERSITAS LATVIENSIS 19 19 SCIENTIAE ET PATRIAE

UNIVERSITY OF
LATVIA