GHENT UNIVERSITY

fwo Opening new horizons

GLIMS

9th International Conference on

GRAMMAR & CORPORA | 2022

Programme & book of abstracts

30/06/2022 – 02/07/2022
Ghent, Belgium

Info and registration: gac2022.ugent.be

Keynote speakers:

Gert De Sutter | Maria del Carmen Parafita Couto | Anke Lüdeling
Marieke Meelen | Florent Perek | Sali Tagliamonte

Grammar & Corpora 2022: provisional programme (2 June 2022)

Author names and presentation titles are hyperlinked to the individual abstracts

Before the conference – Wednesday 29/06/2022 [Location: ???]

???	Warm-up	Location:

Day 1 – Thursday 30/06/2022 [Location: *Het Pand*, Onderbergen]

09:00 – 9:30	Opening session	Location:
9:30 – 10:30	KEYNOTE 1: **Sali Tagliamonte** *Grammar, statistics and 1000 people's stories: Studying language variation and change in natural speech data*	**Location:** **Chair:** Anne-Sophie Ghyselen

	Text & discourse	**Cross-linguistic approaches**	**Learner language**
	Location:	Location:	Location:
10:30 – 11:00	**da Cunha & Abeillé** *Subject or object? A gender bias in function assignment in French*	**Trawiński & Schlotthauer** *Distributional properties of clausal subjects: A cross-linguistic corpus study*	**Shin, Kyung Jung & Mun** *To what extent neural network models reveal L2 constructs? Relationship between text similarity and learner proficiency*
11:00 – 11:30	**Liégeois** *A corpus-based approach to language of immediacy and distance: A theoretical proposal based on French and German weather reports*	TBD	**Wedig & Strobl** *The Belgisches Deutschkorpus (Beldeko) as a resource to investigate cohesion in German learner language: A preliminary analysis of corpus homogeneity*

11:30 – 12:00	Coffee break

	Signed and spoken languages	Alternations: Spanish	Morpho-syntax: German
	Location:	Location:	Location:
12:00 – 12:30	**Otte, Anke, Wähl & Langer** *Numeral Incorporation as Grammaticalization? A Corpus Study on DGS*	**Vázquez Rozas & García-Miguel** *Case marking alternation with psychological verbs in Spanish: Combining different corpus data sources*	**Weber, Bildhauer & Münzberg** *Control nouns: finite vs. non-finite adnominal clauses in German*
12:30 – 13:00	**Lepeut, Vandenitte, Lombart & Meurant** *Comparable corpora of spoken and signed languages: towards a pluri-semiotic perspective on language*	**Granvik** *Alternate ways of explaining: the case of Spanish causal subordinate clauses introduced by dado que, puesto que and ya*	**Husić & Roch** *Mining accordance and information source PPs of German nach*

13:00 – 14:30	Lunch break

	Particles	Theory and method	Multilingual settings
	Location:	Location:	Location:
14:30 – 15:00	**Li, Lorenz & Siemund** *The ages of pragmatic particles in Colloquial Singapore English: A corpus study based on oral history interviews*	**Aplonova, Nikitina, Arkhangelskiy, Hantgan-Sonko, Jordanoska, Sokur, Silué & Paterson** *Multilingual SpeechReporting database: tools, methods and techniques*	**Busso** *"I AGREE TO THE TERMS AND CONDITIONS": legal-lay language comprehensibility in a bilingual English- Italian corpus*
15:00 – 15:30	**Kim** *On the linking adverbial "besides": A corpus-based study*	**Mejri** *Grammatical rule vs. linguistic theory: the case of reflexive pronouns in locative PPs*	**Enrique-Arias** *Testing the Uniformitarian Principle in language contact: variation and change in the Spanish of Mallorca*
15:30 – 16:00	**Esposito** *A Microtext corpus for the analysis of Spanish discourse particles*	**DeVore** *Can network science help explain development of second language complexity?*	**Vanhaverbeke, Enghels & Balam** *Diminutive expressions in bi/multilingual discourse: a cross-community study of Spanish-English codeswitching in Miami and Northern Belize*

16:00 – 16:30	Coffee break

	Romance	**Noun phrases**	**Phraseology & word-formation**
	Location:	Location:	Location:
16:30 – 17:00	**Torres Soler** *Causative constructions with the Spanish motion verbs llevar and traer: a diachronic corpus-based analysis*	**Aarts** *Participles and so-called synthetic compounds as attributive noun modifiers in English*	**Ordines** *No hay boda sin ramo de novia: Capturing the creative potential of snowclones in Spanish*
17:00 – 17:30	**Primerano** *Marrying corpus linguistics, dialectology, and philology: the grammaticalisation of the future and conditional in Central Ibero-Romance (13th-14th century)*	**Basile** *Finnish partitive e-NP constructions in web corpora*	**Lensch & Bloem** *On incoming passers-by and bystanding lookers-on: A quantitative approach to variable particle placement in English particle verbs*

17:30 – 18:30	KEYNOTE 2: **Gert De Sutter** *Understanding grammatical variation from a bilingual perspective: descriptive, methodological and theoretical insights from corpus-based translation research*	Location: Chair: Ludovic De Cuypere
???	Welcome reception	Location: Faculty of Humanities

Day 2 – Friday 01/07/2022 [Location: *Het Pand*, Onderbergen]

09:30 – 10:30	KEYNOTE 3: **Florent Perek** *Constructions and the company they keep*	Location: Chair: Peter Lauwers

	Clause syntax: Catalan	**Productivity panel**	**Discourse**
	Location:	Location:	Location:
10:30 – 11:00	**Torres-Latorre & Sentí** *A corpus study of clitic placement in Old Catalan*	**Panel on "Syntactic Productivity": introduction** *Language Productivity@Work Consortium*	**Orrequia-Barea & Almazán-Ruiz** *Dimensions of modality: lexical modals in the brexit political discourse*
11:00 – 11:30	**Herbeck** *Subject pronoun expression in Valencian Catalan varieties – A quantitative and qualitative study of the corpus Parlars*	**Le Bruyn, Fuchs, van der Klis, Liu, Mo & de Swart** *Measuring productivity through language comparison*	**Dorgeloh** *Corpus-based retrieval and the function of inversion in discourse*

11:30 – 12:00	Coffee break

	Light verb constructions	**Productivity panel**	**Conditionality**
	Location:	Location:	Location:
12:00 – 12:30	**Alvarez-Morera** *Modification in light verb constructions: a corpus study in Germanic and Romance languages*	**Van den Heede, Van Hulle, Colleman, De Cuypere, Enghels, Taverniers & Lauwers** *Productivity (metrics) and semantics: a principal components analysis on minimizing and inchoative data*	**Barrios** *Conditional types, patterns and factuality: a corpus study*
12:30 – 13:00	**Wiskandt & Turus** *Systematic semantic differences between object-experiencer LVCs and corresponding simplex verbs in German*	**Schoonjans** *On the internal and external productivity of IAW phrases in German*	**Vander Haegen** *German 'wh-ever' and 'no matter wh-' as allostructions*

13:00 – 13:45	Lunch break

	Methodological challenges: Spanish	**Productivity panel**	**Theory and method**
	Location:	Location:	Location:
13:45 – 14:30	*Lunch break*	Poster session **Bezinska** *Bulgarian 3- to 6-year-old children's productivity with causatives* **Quentin Feltgen & Georgeta Cislaru** *How Does Productivity Impact Production? Investigating the Ties between Co-Occurrence Frequencies and Cognitive Cost in Real-Time Keylogging Data* **Godts & Taverniers** *Researching {verb, medium} colligations* **Van den Stock, Ghyselen & Colleman** *Measuring inter-individual variation in attitudes towards productivity: from corpus data to acceptability experiment*	*Lunch break*
14:30 – 15:00	**Guajardo** *Using Naive Discriminative Learning to Predict Adverb Placement in Spanish*		**Chan** *Grammar "bores the crap out of me!": A mixed-method study on the "X the Y out of Z" construction and its usage by ESL and ENL speakers*
15:00 – 15:30	**Van Den Driessche** *Recent language change in Spanish teenage talk: the construction with es que*	**Garachana & Sol Sansiñena** *Combinatorial productivity of Spanish verbal periphrases as an indicator of their degree of grammaticalization*	**Maekelberghe & Delaere** *Through the translation glass: How parallel corpora can help us understand fuzzy grammatical categories*
15:30 – 16:00	TBD	**Kiik & Pilvik** *Productivity and functions of neoclassical initial combining forms in Estonian*	**Molochieva, Faghiri & van Lier** *The bi-absolutive construction in Chechen: a comparison of two corpora and elicitation data*

16:00 – 16:30	Coffee break

	Alternations: Germanic	Productivity panel	Usage-based linguistics
	Location:	Location:	Location:
16:30 – 17:00	**Mikkelsen & Glynn** *Behavioural patterns and grammatical constructions: predicting alternation choices between English future constructions*	**Smith** *Measuring the extensibility of a construction relative to the constructional network based on onomasiological domain and discourse situation (pragmeme)*	**Kyung Jung & Shin** *L2-Korean learners' use of constructional components for Korean locative postposition–verb construction: Relationship between L2 textbook and L2 writing*
17:00 – 17:30	**De Pascale & Pijpops** *Modelling meaning differences in syntactic alternations with token-based vectors*	**Baltais, Hartsuiker & Jessen** *Do corpus-derived productivity measures predict language processing? The case of the Spanish inchoative*	**Bernasconi** *Polyfunctional particles in discourse: a corpus-based study of Russian čto li*

17:30 – 18:30	KEYNOTE 4: **Marieke Meelen** *Corpus annotation of the 'conscious self'. From manuscript to egophoric grammars in low-resource historical languages*	Location: Chair: Jóhanna Barðdal
???	Conference dinner	Location:

Day 3 – Saturday 02/07/2022 [Location: Faculty of Humanities, Blandijnberg]

9:30 – 10:30	KEYNOTE 5: **Anke Lüdeling (joint work with Julia Lukassek and Anna Shadrova)** *Variability in grammatical categories and structures: The case of word formation*	Location: Chair: Torsten Leuschner

	Register	**Verb phrases**	**Semantics**
	Location:	Location:	Location:
10:30 – 11:00	**Fernández-Pena & Pérez-Guerra** *Fragments in written and spoken Present-Day English: A corpus-driven constructional account*	**Masini & Busso** *Pleonastic verb particle constructions in Italian: a corpus-based investigation*	**Morei** *A corpus-based study of the semantics of the Pluperfect in spoken Italian*
11:00 – 11:30	**Meyer, Demian, Buchmüller, Szucsich** *Syntactic complexity across registers in Russian*	**Hrbek & Schallert** *Diachronic and diatopic variation of the Middle High German bipartite negation marker ne ... niht*	**Wyrośląk** *Delineating the scope of a benefactive alternation*

11:30 – 12:00	Coffee break

	Morphology	Methodological challenges	Semantics: Mandarin Chinese
	Location:	Location:	Location:
12:00 – 12:30	**Kalnača, Deksne** & **Pakalne** *Latvian deverbal nouns in -ien- and -um- and derivational productivity: a corpus-based analysis*	**Poppek, Masloch & Kiss** *A Corpus-based Perspective on "Split Stimuli" in German*	**Hu** *Semantic coercion of adjectives and numeral classifiers in Mandarin Chinese: A corpus-based study*
12:30 – 13:00	**Górski** *The morphology of Polish imperfective future tense*	**Romain, Milin & Divjak** *Article use in English: construal and constraints*	**Lin** *Aspectual symmetry of correlative coordination in Mandarin Chinese*
13:00 – 13:30	**Danon** *Soft morphophonological constraints on Hebrew genitive choice*	**Denison & Oudesluijs** *Tracking verb changes in a corpus of non-printed manuscript materials*	**Zhang & Liu** *Comparing contextual factors of the eight two-character modal auxiliaries through the lens of modality*

13:30 – 14:30	Lunch break

14:30 – 15:30	KEYNOTE 6: **Maria Carmen Parafita Couto** *The role of multilingual corpora in describing multilingual grammars*	Location: Chair: Renata Enghels
15:30 – 16:00	Closing session	Location:
16:00 - ???	Sightseeing	Location:

Grammar, statistics and 1000 people's stories: Studying language variation and change in natural speech data

Sali Tagliamonte

University of Toronto, Canada

In this presentation I offer an overview of my research program investigating language variation and change. The data come from a large archive of vernacular speech from Ontario, Canada, in which I have been documenting language variation and change among people born from the late 1800's up to the early 2001's. As of 2021, the archive comprises 19 communities with representation from the largest city, Toronto, to many localities in the Near North (e.g. Tagliamonte, 2013; 2014).

Using a selection of well-studied, variable, grammatical features as case studies, I demonstrate how the findings arising from these materials provide important new insights into the nature of language variation and change. In some cases, cross-linguistic regularities expose typological tendencies (negation) (Burnett et al., 2018) or long time assumptions about ongoing grammatical trends are overturned by evidence of stability (Rothlisberger & Tagliamonte, to appear). Simultaneously, smaller pockets of change within variable systems are exposed, such as the lexicalization of certain constructions (*ever)* (Franco & Tagliamonte, to appear) or the splitting off of separate developments (*anyway/anyways)* (Franco & Tagliamonte, 2020). At the same time, social, geographic and cultural influences are also at play (e.g. Tagliamonte et al., 2010; Gardner & Tagliamonte, 2020). Taken together, these findings demonstrate that variation is best understood within a broad, contrastive perspective and that statistical techniques applied to corpus data offer an important means to detect patterns, not only within the variety or dialects under investigation, but also across languages leading to more integrated explanations.

References

Burnett, H., Tagliamonte, S. A. & Koopman, H. (2018). Soft Syntax and the Evolution of Negative and Polarity Indefinites in the History of English. *Language Variation and Change* 30(1): 83-107.

Franco, K. & Tagliamonte, S. A. (2020). New -way(s) with -ward(s): lexicalization, splitting and sociolinguistic patterns. *Language Variation and Change* 32(2): 217-239.

Franco, K. & Tagliamonte, S. A. (to appear). The most stable it's ever been: The preterit/present perfect alternation in spoken Ontario English. *English Language and Linguistics*.

Gardner, M. & Tagliamonte, S. A. (2020). The bike, the back, and the boyfriend: Confronting the "definite article conspiracy" in Canadian and British English. *English World Wide* 41(2): 226-255.

Rothlisberger, M. & Tagliamonte, S. A. (to appear). The social embedding of a syntactic alternation: Variable particle placement in Ontario English. *Language Variation and Change* 32(3): 317-348.

Tagliamonte, S. A. (2013). *Roots of English: Exploring the history of dialects.* Cambridge: Cambridge University Press.

Tagliamonte, S. A. (2014). System and society in the evolution of change: The view from Canada. In Green, E. & Meyer, C. (Eds.), *Variability in Current World Englishes* Berlin and New York: Mouton de Gruyter. 199-238.

Tagliamonte, S. A., D'Arcy, A. & Jankowski, B. (2010). Social work and linguistic systems: Marking possession in Canadian English. *Language Variation and Change* 22(1): 1-25.

Understanding grammatical variation from a bilingual perspective: descriptive, methodological and theoretical insights from corpus-based translation research

Gert De Sutter

Ghent University, Belgium

Grammatical variation has played a key role in understanding how language use in general and variation in particular functions in society, how it is constrained and how it can be represented cognitively. As such, it is a central topic in many usage-based linguistic disciplines, such as sociolinguistics, psycholinguistics, corpus linguistics, probabilistic linguistics and cognitive linguistics, with empirical research into grammatical variation leading to major advances in terms of description, methodology and theory.

It nevertheless seems fair to say that current understanding of grammatical variation is primarily based on studies of monolingual language use, which does not do justice to the ever-increasing amount of multilingual communication, due to increased mobility and global communication. Therefore, it is crucial to investigate to what extent our current understanding of grammatical variation applies to bilingual language production contexts or, alternatively, how it should be adjusted in order to incorporate insights from bilingualism more accurately.

To address this issue, I will present three corpus-based translation studies of grammatical variation phenomena, namely pre- vs postverbal subject placement in Dutch, English *that/zero* alternation in complement clauses and the genitive alternation in Dutch**,** which rely on parallel corpora, i.e. source texts and their translations, and/or monolingual comparable corpora, i.e. translated and non-translated texts in the same language. This will allow us to evaluate (i) to what extent translations, being a prototypical example of bilingual communicative events, exhibit probabilistic patterns of grammatical variation similar to monolingual text production, (ii) to what extent mainstream multifactorial statistical analysis is capable of accurately detecting variation patterns in this type of multilingual data and (iii) to what extent bilingual text production is affected by constraints such as structural priming, structural integration cost, markedness of coding and statistical pre-emption. To conclude, I will discuss the implications of these insights for (probabilistic, cognitive-linguistic) theory of grammatical variation, for statistical analysis and for the use of parallel corpus data in linguistic research.

Constructions and the company they keep

Florent Perek

University of Birmingham, United Kingdom

One major contribution of corpus linguistics to the study of grammar is the realisation that there is a non-trivial relation between words and the syntactic contexts in which they occur (cf. Sinclair 1991, Hunston & Francis 2000, *inter alia*). Many corpus-based studies consistently report that grammatical constructions can be very choosy as to what words they can combine with, sometimes in seemingly unpredictable ways, which has led some scholars to consider that syntactic constructions, just like morphological patterns, display varying degrees of productivity (e.g. Goldberg 2006). In diachrony, lexical fillers of constructions may also vary over time, as speakers come to use language in slightly different ways to their forebearers, gradually expanding (or shrinking) the distribution of constructions (e.g. Rudanko 2011).

In this talk, I will show how distributional semantic models (also known as vector space models) can be used as a powerful tool to explore lexico-grammatical associations of this kind and how they vary (especially over time). In line with the idea that "you shall know a word by the company it keeps" (Firth 1957: 1), distributional semantics aims to capture the meaning of words through their lexical collocates in large corpora, drawing on the intuition that words with a similar meaning are expected to co-occur with a common set of lexical items (Lenci 2008). Distributional semantic methods offer a robust, data-driven way to identify semantic areas in the distribution of a construction, and track changes in it. Through a number of case studies on the productivity of constructions in diachrony (e.g. Perek 2016, 2018), I show that despite the still prevalent emphasis on type frequency (the number of different items) in the literature as a key measure of syntactic productivity, the crucial factor is actually the variability and spread of a construction in semantic space. Indeed, different constructions with a similar increase in type frequency can still display very different degrees of openness, as captured by the distributional semantic methods I will demonstrate.

In the last part of the talk, I offer some reflections on the future of this method and introduce some variations of how it can be applied. I also discuss recent improvements and developments made possible through new technological advances, notably in machine learning (e.g. word2vec, BERT).

References

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in linguistic analysis (Special volume of the Philological Society)*, 1–32. Oxford: Blackwell.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Hunston, S., & Francis, G. (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica* 20(1). 1–31.

Perek, F. (2016). Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics* 54(1): 149–188.

Perek, F. (2018). Recent change in the productivity and schematicity of the *way*-construction: a distributional semantic analysis. *Corpus Linguistic and Linguistic Theory*, 14(1), 65-97.

Rudanko, J. (2011). *Changes in Complementation in British and American English: Corpus-Based Studies on Non-Finite Complements in Recent English*. Basingstoke: Palgrave Macmillan.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Corpus annotation of the 'conscious self'. From manuscript to egophoric grammars in low-resource historical languages

Marieke Meelen

University of Cambridge, United Kingdom

The verbal systems of some languages mark the speaker's personal involvement in an event; for example, in Lhasa Tibetan *nga em-chi* ***yin*** '**I'm** a doctor', *'di nga'i bu-mo* ***yin*** 'This is **my** daughter', and *'di khyed-rang-gi gsol-ja* ***yin*** 'This is your tea [that **I have made** for you]' all end with ***yin***, and all three sentences involve the speaker somehow. In another branch of the Tibeto-Burman language family, Kathmandu Newar uses vowel lengthening to indicate a speaker's involvement (*ji: a:pwa twan-****ā*** '**I** drank too much' vs. *chã/wa a:pwa twan-a*. 'you/(s)he drank too much'), but only if self-conscious . There is a clear distinction, for example, between the egophoric morpheme indicated by the long ***-ā*** *ji: Mānaj nāpalān-****ā*** '**I** met Manoj **as planned**' vs. the non-egophoric short vowel *-a* in *ji: Mānaj nāpalān-a* 'I met Manoj by coincidence'. This phenomenon, whereby the speaker's knowledge, experience or personal involvement is grammatically expressed is called 'egophoricity'. It was first described in Kathmandu Newar (Hale 1970) and Lhasa Tibetan (DeLancey 1980), but today is known in languages of the Himalayas, New Guinea, and equatorial South America. But how do languages develop egophoric marking?

In this talk I will show that Newar and Tibetan offer an excellent starting point to answer this nearly unexplored question. Unlike other languages with egophoric marking, such as Awa Pit (Barbacoan), Kaluli (Trans New Guinea) and Guambiano (Coconucan), Tibetan and Newar varieties have long literary traditions (Tibetan since 650 CE and Newar since 1112 CE). Unlike their present-day descendants, neither Classical Tibetan or Classical Newar exhibit egophoricity (Tournadre & Jiatso 2001). This means that, in theory, we should be able to create annotated diachronic corpora to explore this unanswered research question.

In practice, however, things are not so simple. Since historical Newar and Tibetan are both low-resource and under-researched historical languages, creating well-annotated corpora is not a straightforward task. I will therefore first discuss the creation of deeply-annotated corpora in six different Tibetan and Newar varieties. For some of these varieties, we currently only have photographs of 16-17th c. manuscripts, for others, we need to go to Nepal to collect data in the field and then transcribe it. Each of these therefore present some unique challenges that need to be addressed to arrive at the sophisticated level of annotation we need to understand how egophoric marking emergence and develops over time.

In this talk I'll present some crucial case studies at different stages of the annotation workflow to illustrate how challenges of low-resource historical languages can be overcome and why close collaborations of philologists, NLP experts and linguists in different areas (e.g. those specializing in historical linguistics, phonetics, morphosyntax, semantics & pragmatics) is essential to tackle complex questions of language variation and change, such as the emergence of egophoricity.

Variability in grammatical categories and structures: The case of word formation

Anke Lüdeling, Julia Lukassek, Anna Shadrova

Humboldt-Universität zu Berlin, Germany

How variable are word formation categories across speakers? This question is interesting because (concatenative) word formation is both a grammatical process and intimately tied in with the lexicon. Most of the complex words that we find in any given corpus are highly lexicalized. At the same time, word formation is grammatical in the sense that we can identify categories and patterns that allow for productivity and are subject to language dynamics such as grammaticalization, similarly to syntactic pattern formation. There is evidence that these categories and patterns are accessible even in highly lexicalized words (Smolka/Libben/Dressler 2019). We know from previous research that parts of speech, constituents, and dependencies show highly stable proportions across corpora (otherwise, many applications of natural language processing such as probabilistic tagging or parsing would not be possible). The lexicon, on the other hand, is less easily described in statistical terms (Piantadosi 2014, Williams et al.. 2015).

In our talk, we will discuss the distributions of word-formation patterns of verbs and nouns in two corpora of German. The very high inter- and intra-speaker variability that we find (cf. Shadrova et al.. 2021) has far-reaching methodological and theoretical implications. Specifically, we will address issues with usage-based modeling of grammar and acquisition, such as notions of "constructions all the way down" (Goldberg 2006, 18) or or the construct of a homogeneous native speaker in corpus-based grammar research.

References

Goldberg, Adele E. (2016) *Constructions at Work: The Nature of Generalizations in Language*. Oxford University Press.

Piantadosi, Steven T. (2014) Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21(5), 1112-1130.

Shadrova, Anna, Pia Linscheid, Julia Lukassek, Anke Lüdeling & Sarah Schneider (2021) A challenge for contrastive L1/L2 corpus studies: Large inter-and intra-individual variation across morphological, but not global syntactic categories in task-based corpus data of a homogeneous L1 German group. *Frontiers in Pschology* 12, https://doi.org/10.3389/fpsyg.2021.716485.

Smolka, Eva, Gary Libben & Wolfgang U. Dressler (2019) When morphological structure overrides meaning: Evidence from German prefix and particle verbs. *Language, Corgnition and Neuroscience* 34(5), 599-614.

Williams, Jake Ryland, Paul R. Lessard and Suma Desu and Eric M. Clark and James P. Bagrow and Christopher M. Danforth and Peter Sheridan Dodds (2015) Zipf's law holds for phrases, not words. *Scientific Reports* 5(1), 1-7, arXiv:1406.5181.

The role of multilingual corpora in describing multilingual grammars

Maria Carmen Parafita Couto

Leiden University, the Netherlands

In this talk, I will (i) illustrate how (open access) corpora of naturalistic multilingual speech help us describe distributional patterns that arise and shed light on the grammaticality of these structures, as well as (ii) discuss whether the psycho-/neurolinguistic findings align with the corpora-based findings. For these purposes, I will discuss the case of competing theoretical and methodological tensions in the structural study of code-switching, that is, when multilingual speakers "go back and forth" between the languages they speak within a conversation, or even within a sentence (Deuchar 2012). I will show how corpus analyses of production data can provide a wealth of information about the naturalistic occurrences of code-switches, and enable the predictions of different theoretical models to be assessed in an ecologically valid way. Determining the grammatical constraints that may predict code-switching patterns has been the focus of attention of many recent studies (cf. Backus 2015, Balam et al. 2020, Toribio 2017, López 2020, among many others), some of which also employ psycho-/neurolinguistic measures (Beatty-Martínez et al. 2018, Pablos et al. 2018, Van Hell et al. 2018, Vaughan-Evans et al. 2020, inter alia). I will discuss how processing of code-switched speech often aligns with the code-switching patterns that have previously been reported in naturalistic production in the specific multilingual community, highlighting the importance of studying code-switching from a language ecological perspective. I will finish with a call for rapprochement between domains and argue for open access corpora, which are often collected at public expense. The availability of these data will help us further unravel recent theoretical and empirical questions and criticisms being raised about the description and nature of code-switching grammars (e.g. Toribio 2018, Parafita Couto et al. in press).

References

Backus, A. (2015). A usage-based approach to code-switching: The need for reconciling structure and function. In G. Stell & K. Yakpo (Eds). *Code-switching between Structural and Sociolinguistic Perspectives* (19-37). Berlin/Munich/Boston: De Gruyter.

Balam, O., Parafita Couto, M.C., & Stadthagen-González, H. (2020). Bilingual verbs in three Spanish/English code-switching communities. *International Journal of Bilingualism*, *24*(5–6), 952–967.

Beatty-Martínez, A. L., Valdés Kroff, J. R., & Dussias, P. E. (2018). From the field to the lab: A converging methodsapproach to the study of codeswitching. *Languages*, *3*(2), 1–19.

Deuchar, M. (2012). Code-switching. In Chapelle, C.A. (ed.) *Encyclopedia of Applied Linguistics*. New York: Wiley, 657-664.

López, L. (2020). Bilingual grammar. Toward an integrated model. Cambridge University Press.

Pablos, L., Parafita Couto, M.C., Boutonnet, B., De Jong, A., Perquin, M., De Haan, A., & Schiller, N.O. (2019). Adjective-Noun order in Papiamento-Dutch code-switching. *Linguistic Approaches to Bilingualism*, Volume 9, Issue 4, p. 710 – 735.

Parafita Couto, M.C. Greidanus Romaneli, M. &Bellamy, K. (in press) Code-switching at the interface between language, culture, and cognition. Lapurdum, IKER UMR 5478 CNRS.

Parafita Couto, M. C. Bellamy, K. & Ameka, F. (in press). Theoretical Linguistic Approaches to Multilingual code-switching. Cambridge Handbook of Third Language Acquisition and Processing. Eds. Cabrelli, J. Chaouch-Orozco,A., González Alonso, J., Pereira Soares,, S., Puig-Mayenco, E. & Rothman, J. Cambridge University Press.

Toribio, A. J. (2017). Structural approaches to code-switching: Research then and now. In R.E.V. Lopes, J. Ornelas de Avelar & S. M. L. Cyrino (Eds.) *Romance Languages and Linguistic Theory 12. Selected papers from the 45th Linguistic Symposium on Romance Languages (LSRL), Campinas, Brazil* (pp. 213-233). Amsterdam: John Benjamins Publishing Company.

Toribio, A. J. (2018). The future of code-switching research. In López, L. (Ed.*). Code-Swiching-Experimental Answers to Theoretical Questions: In honor of Kay González Vilbazo.* Issues in Hispanic and Lusohpone Linguistics 19. John Benjamins, pp. 257–267.

van Hell, J. G., Fernandez, C., Kootstra, G. J., Litcofsky, K. A., & Ting, C.Y. (2018). Electrophysiological and experimental-behavioral approaches to the study of intra-sentential code-switching. Linguistic Approaches to Bilingualism, 8(1), 144-171.

Vaughan-Evans, A., Parafita Couto M.C., Boutonnet, B., Hoshino, N., Webb-Davies, P., Deuchar, M. and Thierry, G. (2020). Switchmate! An Electrophysiological Attempt to Adjudicate Between Competing Accounts of Adjective-Noun Code-Switching. *Frontiers in. Psychology.* 11:54976

Subject or object? A gender bias in function assignment in French

Yanis da Cunha & Anne Abeillé

Laboratoire de Linguistique Formelle, Université Paris Cité

The choice of syntactic function for a given argument is known to be sensitive to several factors, such as length, animacy, pronominality in English (Hundt et al., 2021) and French (da Cunha & Abeillé, 2020): subjects tend to be shorter, more often animate and pronominal than objects.

In a sample of 500 sentences from the FrenchTreebank (FTB, Abeillé et al., 2019), (da Cunha & Abeillé, 2020) have also found a bias towards masculine subjects. This gender bias has also been observed in specific corpora, e.g. examples in linguistic papers (Kotek et al., 2021; Pabst et al., 2018). For French, Richy & Burnett (2020), in the examples of the linguistics journal *Langue Française* (1969-1971; 2008-1017), found that male human characters tend to be subjects, pronouns and agents, more often than female ones, without an effect of period or author's gender (Table 1).

Social gender	Pronoun/noun rate	Subject/object rate	Agent/non agent rate
Female	22,6%	76,3%	26,6%
Male	32,2%	91,0%	48,9%

Table 1.: Human NPs in examples from *Langue Française* (1969-1971 and 2008-1017) (Richy & Burnett, 2020).

We extend these results by analyzing 900 sentences from French spoken (CEFC, Benzitoun et al., 2016) and written (FTB, Abeillé et al., 2019) corpora. Our sample contains as many actives as passives. We annotated arguments' grammatical gender and humanness, and speaker/author's gender (when the metadata are available). We were interested in the choice between (passive) subject and (active) object. Using mixed-effects logistic regression models (Table 2), we report three main results. First, we found a significant main effect of argument gender : masculine arguments are more likely to be subject than object (p=0.04). Next, this effect only matters for human arguments (significant interaction, p=0.02, Figure 1), i.e. when grammatical gender is interpreted as social gender. Finally, argument gender interacts significantly with speaker's social gender (p=0.04, Figure 2) : only male authors/speakers show a bias towards masculine subjects, while female authors/speakers do not seem to show this bias. These results are consistent in both written and spoken corpora.

	Estimate	Std. Error	p-value
Intercept	0.07	0.1	0.49
Feminine argument	-0.17	0.08	0.04
Female speaker	-0.20	0.08	0.01
Human argument	0.31	0.09	0.001
Feminine argument: Female speaker	0.16	0.08	0.04
Feminine argument: Human argument	-0.22	0.09	0.02

Table 2. Mixed effects logistic regression modeling passive subject vs. active object alternation (n = 763). Positive estimates indicate that subject coding is more likely, negative ones indicate the contrary.

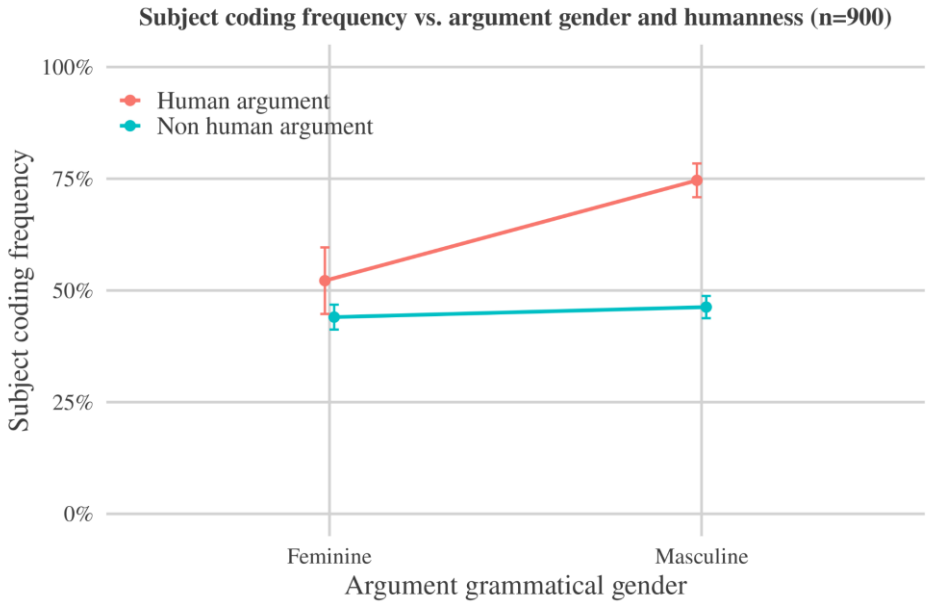


Figure 1. Sample of 900 active/passive sentences from FTB and CEFC.

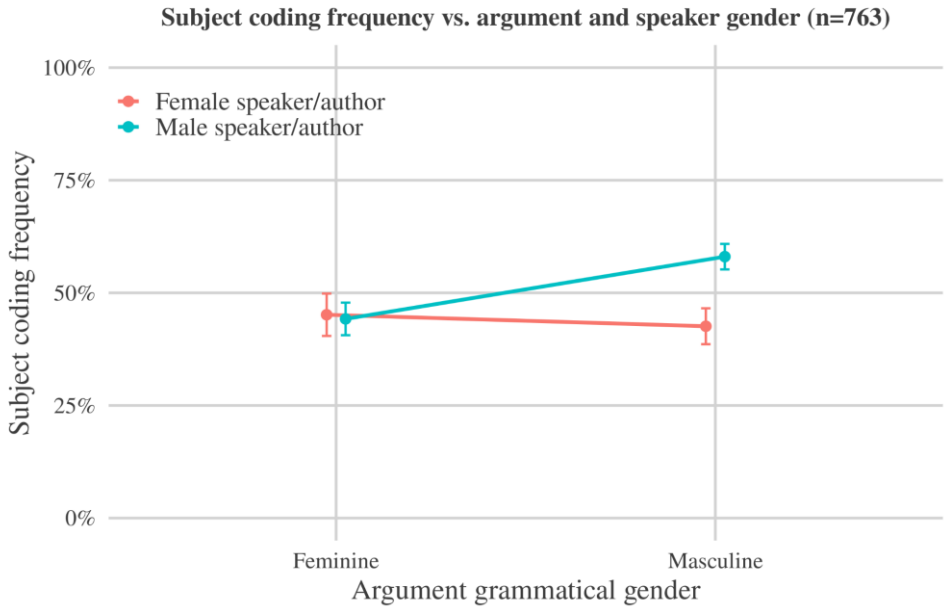


Figure 2. Sample of 763 sentences from FTB and CEFC.

We are extending these results by investigating the rest of the FTB (Candito et al., 2014 version), putting aside expletive *il* "it" and plural arguments (to avoid generic and mixed gender interpretations). We found 53999 maculine NPs and 43648 feminine NPs (ratio 1.24), but masculine arguments are biased towards pronominalization, subject function and active verbs (Table 3). We are currently annoting humanness with Flexique (an animacy-annotated French dictionary) and running statistical analysis on these data. Our results are similar to Richy & Burnett (2020), and suggest that gender biases can be found across genres, here in newspaper texts. They also suggest that gender biases in function assignment may influence syntactic preferences (Esaulova & Von Stockhausen, 2015) and should be taken into account in NLP studies (Costa-jussà, 2019; Sun et al., 2019; Wisniewski et al., 2021).

Grammatical gender	Pronoun/noun rate	Subject/object rate	Active/passive subject rate
Feminine	8,1%	58,0%	88,5%
Masculine	16,7%	68,7%	92,9%

Table 3. Preliminary results for the whole FTB (1990-1993).

We conclude there is evidence for gender effect in syntactic function assignment in French, which holds across different registers and genres (linguists' examples in Richy and Burnett 2020, conversations/interviews in CEFC, newspapers in the FTB).

References

Abeillé, A., Clément, L., & Liégeois, L. (2019). Un corpus annoté pour le français : Le French Treebank. *TAL*, *60*, 19-43.

Benzitoun, C., Debaisieux, J.-M., & Deulofeu, H.-J. (2016). Le projet ORFÉO : Un corpus d'étude pour le français contemporain. *Corpus*, *15*, 91-114.

Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D., & de La Clergerie, É. V. (2014). Deep syntax annotation of the Sequoia French treebank. *International Conference on Language Resources and Evaluation (LREC)*.

Costa-jussà, M. R. (2019). An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, *1*(11), 495-496.

Da Cunha, Y., & Abeillé, A. (2020). L'alternance actif/passif en français : Une étude statistique sur corpus écrit. *Discours*, *27*.

Esaulova, Y., & Von Stockhausen, L. (2015). Cross-linguistic evidence for gender as a prominence feature. *Frontiers in Psychology*, *6*.

Hundt, M., Röthlisberger, M., & Seoane, E. (2021). Predicting voice alternation across academic Englishes. *Corpus Linguistics and Linguistic Theory*, *17*(1), 189-222.

Kotek, H., Dockum, R., Babinski, S., & Geissler, C. (2021). Gender bias and stereotypes in linguistic example sentences. *Language*.

Richy, C., & Burnett, H. (2020). J*ean does the dishes while Marie fixes the car* : A qualitative and quantitative study of social gender in French syntax articles. *Journal of French Language Studies*, *30*(1), 47-72.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating gender bias in natural language processing : Literature review. *arXiv preprint arXiv:1906.08976*.

Wisniewski, G., Zhu, L., Ballier, N., & Yvon, F. (2021). Biais de genre dans un système de traduction automatique neuronale : Une étude préliminaire. *TALN*, 11-25.

A corpus-based approach to language of immediacy and distance: A theoretical proposal based on French and German weather reports

Vince Liégeois

Background

Discourse studies have greatly benefited from corpus linguistic methodology (Partington et al. 2013: 10-14). However, within the discourse-oriented research paradigm of Koch & Oesterreicher (1985, [1990] 2011, 2012) such corpus linguistic inquiries are currently lacking (Ágel & Hennig 2006: XI; Schafroth 2013: 293-294). The Koch & Oesterreicher tradition is particularly interested in (the features of) spoken language and distinguishes between the level of the medium (spoken/written) and the level of conception, i.e., whether an utterance is oriented more towards a spoken (= language of immediacy) or written norm (= language of distance). Yet, it has often been criticised for not providing a clear methodology for either qualitative or quantitative research (Hennig & Feilke 2016: 2).

Theoretical proposition

To this aim, we propose an approach combining a corpus-based linguistic methodology (cf. Gries & Stefanowitsch 2006; Weisser 2016) with the conditions of communication individuated in Koch & Oesterreicher's theory – which characterise either language of immediacy or distance – functioning as the metadata for the qualitative aspect of such a corpus-based analysis (cf. Tab. 1).

Criterium	immediacy	distance
dialogicity	positive	negative
familiarity of the partners	positive	negative
face-to-face-interaction	positive	negative
free thematic development	positive	negative
public	negative	positive
spontaneity	positive	negative
'involvement'/ 'detachment'	'involvement'	'detachment'
context embeddedness	positive	negative

Table 1: Metadata based on Koch & Oesterreicher's conditions of communication (1985: 23; 2012: 450)

In doing so, we hope to (i) provide a qualitative-quantitative methodology to properly study spoken and written language based on Koch & Oesterreicher's discourse theory, and (ii) determine correlations between the conditions of communication enlisted above and communication types like domain-specific discourse traditions and text genres, the study of which has yet to include approaches involving this discourse theory (cf. Ágel & Hennig 2010: 11-12).

Corpus & analysis

To exemplify our methodology, we will conduct a case-study of French and German audiovisual weather reports for which we assembled two corpora containing 300,000 tokens each. As such, we are dealing with a domain-sensitive text genre (from the domain of meteorology) which seeks to transmit domain-specific information to a lay audience (Blondeau & Labeau 2016: 245). More concretely, we will look at differences in the use of so-called "evaluation markers", e.g., fixed strings like *bad weather, stifling hot* and "*no luck today!"* (cf. Partington et al. 2013; Klimczak & Dynel 2018; Liégeois 2021) between different weather reporters.

These evaluation markers were not yet studied within the Koch & Oesterreicher research tradition, yet they play a crucial role in communicating domain-specific knowledge to the lay audience. More specifically, we want to determine whether certain weather reporters use more evaluation markers than others and whether correlations with the use of immediacy- or distance-features can be found. To this aim, we will not only correlate our results (= the expressivity markers found in our corpus-based analysis) with the metadata singled out above, but also with other characteristic immediacy-traits: deixis (personal, spatial and temporal), a freer choice of tense and right-/left-dislocation.

Results

The results of our analysis show that evaluation markers appear more frequently in the speech of those weather reporters that make more use of immediacy features (e.g., those singled out above). Furthermore, from a qualitative perspective, different types of evaluation markers (e.g., with prosodic emphasis: "*nous avons été gâtés aujourd'hui*!") are attested between the different weather reporters, with more lexical variation being found in the speech of weather reporters using more immediacy-features. Consequently, our analysis does not only provide an appropriate methodology for the Koch & Oesterreicher research paradigm, but also for the study of domain-specific discourse and text genres.

References

Ágel, Vilmos and Hennig, Mathilde (2006): Einleitung. In: Vilmos Ágel and Mathilde Hennig (Eds.): *Grammatik aus Nähe und Distanz. Theorie und Praxis am Beispiel von Nähetexten 1650-2000*. Berlin-New York: De Gruyter, IX-XII.

Ágel, Vilmos and Hennig, Mathilde (2010): Einleitung. In: Vilmos Ágel and Mathilde Hennig (Eds.): *Nähe und Distanz im Kontext Variationslinguistischer Forschung* [Linguistik – Impulse & Tendenzen 35]. Berlin-New York: De Gruyter, 1-22.

Blondeau, Hélène and Labeau, Emmanuelle (2016): La référence temporelle au futur dans les bulletins météo en France et au Québec. Regard variationniste sur l'oral préparé. *The Canadian Journal of Linguistics / La revue canadienne de linguistique* 61/3. 240-258.

Gries, Stefan T. and Stefanowitsch, Anatol (2006): *Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis* [Trends in linguistics. Studies and monographs 172]. Berlin-New York: De Gruyter.

Hennig, Mathilde and Feilke, Helmuth (2016): Perspektiven auf ‚Nähe und Distanz'. Zur Einleitung. In: Helmuth Feilke & Mathilde Hennig (Eds.): *Zur Karriere von "Nähe und Distanz". Rezeption und Diskussion des Koch-Oesterreichers-Modells* [Reihe Germanistische Linguistik 306]. Berlin-New York: De Gruyter, 1-10.

Klimczak, Karl M. and Dynel, Marta (2018): Evaluation Markers and Mitigators in Analyst Reports in Light of Market Response to Stock Recommendations. *International Journal of Business Communication* 55/3, 310-337.

Koch, Peter and Oesterreicher, Wulf (1985): Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36, 15-43.

Koch, Peter and Oesterreicher, Wulf ([1990]) 2011: *Gesprochene Sprache in der Romania. Französisch, Italienisch, Spanisch* [Romanistische Arbeitshefte 31]. Berlin-New York: De Gruyter.

Koch, Peter and Oesterreicher, Wulf (2012): Language of Immediacy – Language of Distance. Orality and Literacy from the Perspective of Language Theory and Linguistic History. In: Claudia Lange, Beatrix Weber and Göran Wolf (Eds.): C*ommunicative Spaces. Variation, Contact, and Change. Papers in Honour of Ursula Schaefer*. Frankfurt am Main: Peter Lang, 441-473.

Liégeois, Vince (2021): Zur Diskursivität des Nähe-Distanz-Kontinuums. Theoretische Vorschläge am Beispiel eines Korpus gesprochener und geschriebener französischer Wetterberichte. In: Laurent Gautier, Michael Schreiber & Simon Varga (Eds.): *Fachsprachen kontrastiv* [Kontraste/Contrastes – Studien zum deutsch-französischen Sprach- und Diskursvergleich 7]. Frankfurt am Main: Peter Lang. In press.

Partington, Alan/Duguid, Alison/Taylor, Charlotte (2013): *Patterns and Meanings in Discourse. Theory and practice in corpus-assisted discourse studies (CADS)* [Studies in Corpus Linguistics 55]. Amsterdam: John Benjamins.

Schafroth, Elmar (2013): Diskurstraditionen der Sprachapologetik. In: Elmar Schafroth, Martina Niklaus, Christine Schwarzer and Domenico Conte (Eds.): *Italien, Deutschland, Europa. Kulturelle Identitäten und Interdependenzen* [Beiträge zur Kulturwissenschaft 27]. Oberhausen: Athens, 294-349.

Weisser, Martin (2016): *Practical Corpus Linguistics. An Introduction to Corpus-Based Language Analysis.* Hoboken: John Wiley & Sons.

Distributional properties of clausal subjects: A cross-linguistic corpus study

Beata Trawiński & Susan Schlotthauer

IDS Mannheim, Germany

It is known that there is a cross-linguistic variation in marking and licensing of (clausal) arguments, including subject arguments (cf. Schmidtke-Bode 2014). As far as the European linguistic area is concerned, two classes of languages can be distinguished: case-marking languages and configurational languages (cf. Haspelmath 2001). In case-marking languages, such as German, Polish or Hungarian, case-marking allows to identify arguments. The subject is typically in the nominative case and triggers verb agreement, the object is in the accusative case. In configurational languages, such as English or Italian, word order is much more important for distinguishing arguments. The subject typically precedes the verb and triggers verb agreement, the object follows the verb. Clausal subjects are licensed by specific (lexical) environments such as epistemic, deontic, evaluative or phasal contexts, but there is a cross-linguistic variation related to structural circumstances (cf. Schlotthauer et al. 2014 for the licensing of infinite clausal subjects in German, Hungarian and Rumanian). In this study, the distribution and properties of clausal subjects in five European languages have been investigated: German (1), Polish (2) and Hungarian (3) (case-marking languages) and English (4) and Italian (5) (configurational languages). Five monolingual treebanks annotated with Universal Dependencies (version UD 2.3, de Marneffe et al. 2021) have been used as data source. The analyzed parameter included: the PoS of the syntactic head of the subject clause, the PoS of the matrix head, the finiteness (finite versus infinite) and the position (pre-verbal vs. post-verbal) of the subject clause. The results show that all investigated languages exhibit clear preferences for nominal subjects as opposed to clausal subjects (94%-99% to 6%-1%). The analysis of the distribution of the clausal subjects shows that the two configurational languages (English and Italian) use infinite rather than finite clausal subjects. In Italian, the opposite is the case, whereas German and Polish show no (very strong) preferences for finite versus infinite clausal subjects (Figure 1). As far as the position of clausal subjects is concerned, German and Italian show no preferences, whereas English and Hungarian tend to use clausal subjects in pre-verbal position and Polish in post-verbal position (Figure 2). Finally, in German, clausal subjects are licensed most frequently by copula constructions, whereas in Polish, they are mainly licensed by main verbs. No preferences in this respect can be observed in English, Hungarian and Italian (Figure 3). How these findings correlate with other typological and language-specific properties of the investigated languages (such as the use of expletive pronouns) will be explored and discussed in the full paper.

(1) Zum Zahnarzt zu gehen, war für mich immer eine große Überwindung.
'Going to the dentist has always been a great effort for me.'
(2) Szybko okazało się, że moje obawy były niepotrzebne.
'It quickly turned out that my fears were unnecessary.'
(3) Igazságosnak lenni az objektivitást jelenti.
'To be fair is to be objective.'
(4) That he managed to get on the wagon […] is laudable.
(5) Nei primi giorni che andiamo al mare, è meglio prendere il sole per poco tempo.
'In the first days that we go to the beach, it is better to sunbathe for a short time.'

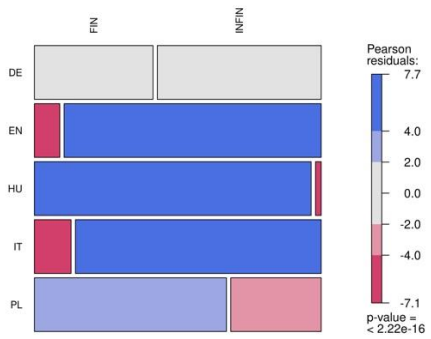


Figure 1 Distribution of finite and infinite clausal clausal subjects across the five languages

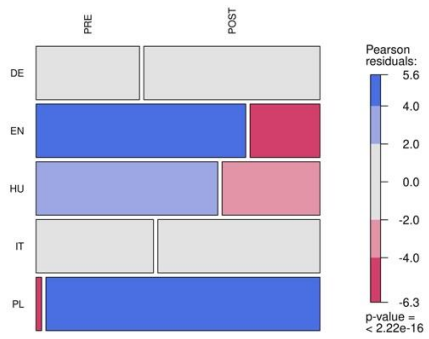


Figure 2 Distribution of pre- and post-verbal subjects across the five languages

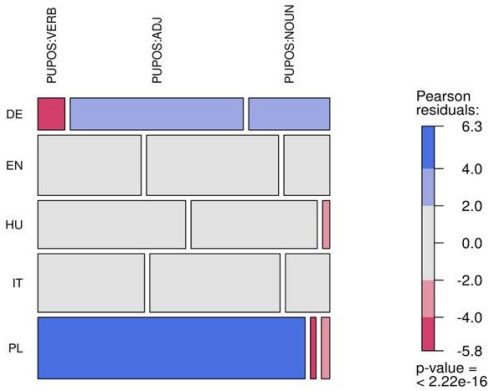


Figure 3 Distribution of the selecting categories across the five languages

References

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre and Daniel Zeman. 2021. Universal Dependencies. In *Computational Linguistics* 47(2), 255–308.

Haspelmath, Martin (2001). Non-canonical marking of core arguments in European languages. In: Aikhenvald, Alexandra Y. & Dixon, R.M.W & Onishi, Masayuki (eds.) *Non-canonical marking of subjects and objects*. (Typological Studies in Language, 46.) Amsterdam: Benjamins, 53-83.

Schlotthauer, Susan/Zifonun, Gisela/Cosma, Ruxandra(2014): Verbale und nominale Infinitive – Strukturelle Eigenschaften und Funktion als Subjekt. In: RuxandraCosma, Stefan Engelberg, Susan Schlotthauer, SperanţaStănescuund Gisela Zifonun(Hg.): Komplexe Argumentstrukturen. Kontrastive Untersuchungen zum Deutschen, Rumänischen und Englischen. (= Konvergenz und Divergenz 3). Berlin/Boston: de Gruyter. S. 253–282.

Schmidtke-Bode, Karsten (2014). Complement Clauses and Complementation Systems: A Cross-Linguistic Study of Grammatical Organization. PhD dissertation, University of Jena.

To what extent neural network models reveal L2 constructs? Relationship between text similarity and learner proficiency

Gyu-Ho Shin, Boo Kyung Jung & Seongmin Mun

Palacký University Olomouc; University of Pittsburgh; Chosun University

With the recent development of NLP techniques, a number of second language (L2) studies utilise these techniques to automatically analyse learner corpora (Meurers, 2015). One area in learner corpus research is text quality, which concerns semantic–pragmatic aspects of language use to influence overall text quality (e.g., Crossley et al., 2019). Despite increasing interests in employing various NLP techniques (e.g., Dascalu et al., 2017), little attention has been paid to how similarly/differently each technique reveals L2 constructs such as learner proficiency. In addition, NLP-based L2 research is heavily biased towards L2-English, which does not ensure the generalisability of its implications. Against this background, we investigate the relationship between learner proficiency and text similarity of L2-Korean learners' written production (relative to native speakers' writing) measured through neural network models.

Method (Table 1). Thirty-six L1-Chinese L2-Korean learners (age: mean = 24.2; *SD* = 3.11) were asked to write argumentative essays on two topics: *preservation vs. exploitation of the nature; competition vs. cooperation*. Learner proficiency was measured separately, using the Korean C-test (Lee-Ellis, 2009; ranging from 0 to 188; mean = 135.98; *SD* = 32.23). Essays from 10 native Korean speakers were collected as a reference text. After extracting content words from the essays, we computed cosine similarity scores between individual learner writing and the reference text by employing two neural network models—Word2Vec (Mikolov et al., 2013; bag-of-words; context-independent) and BERT (Devlin et al., 2018; transformer; context-dependent). The similarity scores (predictor) and proficiency scores (outcome) were then submitted to linear regression models.

Table 1. Information about data by topic (numeric values = number of words)

Topic	L2 learner			Native speaker		
	Mean (SD)	Minimum	Maximum	Mean (SD)	Minimum	Maximum
1	107 (36.36)	62	201	158 (21.27)	131	194
2	113 (38.48)	57	203	166 (33.89)	110	211

Results (Figure 1 & Table 2). The Word2Vec model showed a significant relationship between the two variables for both topics: $F(1, 34) = 3.405$, $p = .074$, $R^2 = .064$, $B = 113.86$ for Topic 1 (albeit marginal); $F(1, 34) = 8.748$, $p = .006$, $R^2 = .181$, $B = 172.59$ for Topic 2. For the BERT model, the slope of the regression line for Topic 1 was nearly horizontal whereas that for Topic 2 was positively oblique. This was reflected in the linear regression analysis, with marginal significance only for Topic 2: $F(1, 34) = 3.79$, $p = .060$, $R^2 = .074$, $B = 32.296$. To further examine how each neural network model classified the participants into the same group uniformly, we created two proficiency groups (highest; lowest) with seven essays by topic. These models demonstrated distinctive classification patterns, yielding weak congruency across the topics/models.

Together, these results indicate that (i) the degree that neural network models explain L2 constructs (learner proficiency in this study) was asymmetric and (ii) these models' performance was sensitive to

essay topics (and particularly to word use such as repetitions of keywords), manifesting some limits on addressing individual variability of L2 writing as well. Given the recent trend that NLP techniques are widely used in learner corpus research, our findings suggest the need for researchers to be aware of NN models' algorithmic characteristics, together with possible influences of topic variations, in conducting automatic L2 text analysis research in pursuit of addressing L2 constructs.

Word2Vec BERT

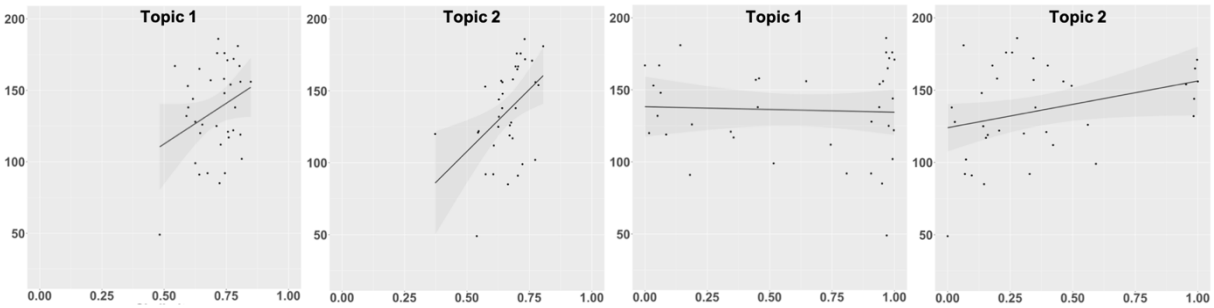


Figure 1. Scatterplot: Similarity scores (X-axis) and proficiency scores (Y-axis).

Table 2. Seven highest/lowest similarity scores and their proficiency scores

	Word2Vec						BERT					
	Topic 1			Topic 2			Topic 1			Topic 2		
	PPT	SIM	PRF	PPT	SIM	PRF	PPT	SIM	PRF	PPT	SIM	PRF
	13	0.848	156	32	0.805	181	19	1.000	171	13	1.000	156
	11	0.812	102	12	0.786	154	2	0.997	122	19	0.997	171
	34	0.807	119	13	0.774	156	35	0.993	144	25	0.989	165
Highest	33	0.807	156	11	0.773	102	11	0.993	102	35	0.985	144
	7	0.803	167	19	0.761	171	26	0.991	176	9	0.983	132
	32	0.796	181	18	0.733	172	18	0.979	172	12	0.953	154
	10	0.785	138	21	0.731	186	20	0.977	125	16	0.593	99
	23	0.625	128	36	0.605	92	34	0.085	119	15	0.096	91
	35	0.616	144	6	0.575	92	28	0.063	148	11	0.073	102
	5	0.597	138	4	0.573	153	14	0.057	167	6	0.069	92
Lowest	4	0.595	153	2	0.547	122	9	0.051	132	32	0.064	181
	9	0.590	132	3	0.544	121	4	0.034	153	23	0.029	128
	14	0.543	167	22	0.539	49	27	0.017	120	5	0.016	138
	22	0.482	49	27	0.373	120	7	0.000	167	22	0.000	49

Note. PPT = participant; PRF = proficiency score; SIM = similarity score

References

Crossley, S., Kyle, K., & Dascalu, M. (2019). *Behavior Research Methods*, 51, 14–27.

Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., & Kurvers, H. (2017). In *Artificial Intelligence in Education 2017 Lecture Notes in Computer Science, 10331* (pp. 52–63).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186).

Lee-Ellis, S. (2009). *Language Testing, 26*(2), 245–274.

Meurers, D. (2015). In *The Cambridge handbook of learner corpus research* (pp. 537–566).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013).

The Belgisches Deutschkorpus (Beldeko) as a resource to investigate cohesion in German learner language: A preliminary analysis of corpus homogeneity

Helena Wedig & Carola Strobl

University of Antwerp

helena.wedig@uantwerpen.be, carola.strobl@uantwerpen.be

We present a new learner corpus for the investigation of German as a foreign language (L2), Beldeko (Belgisches Deutschkorpus). Beldeko contains 301 summaries produced by writers with Dutch as first language (L1) (70,774 tokens). The corpus was created with the aim to investigate novice academic writing in German L2, more specifically, the characteristics of academic writing of writers with Dutch L1, which is a novelty in learner corpora. In this presentation, we will focus on the representativeness of Beldeko as a German learner corpus, especially with regard to selected cohesive devices, which are central to the advanced communicative competence of language learners.

To investigate the representativeness of the Beldeko corpus as a potential resource to investigate cohesion in German learner language, first its level of homogeneity needs to be established. The texts of the corpus were written by 115 students majoring in German (CEF level of B2-C1). They produced summaries of two popular-scientific texts about language variation in contemporary German under test conditions. With regard to corpus homogeneity, two hypotheses were investigated: (1) Based on the writers' common linguistic background and similar overall proficiency level, we hypothesized that the texts show a similar pattern regarding the distributions and the frequencies of layers containing general linguistic information, e.g. part-of-speech (POS)-tags. (2) Since the use of cohesive devices is strongly related to individual writing style and vocabulary size, we hypothesized to find a higher heterogeneity in the corpus regarding these elements. Especially the first condition needs to be met to guarantee that the corpus is balanced enough as a resource for the analysis of the use of cohesive devices.

For the statistical analysis of the corpus, the data were pre-processed and several linguistic annotation layers were added automatically (e.g. POS-tags). Furthermore, we used an online tool for automated text analysis (CTAP: Weiss & Meurers, 2019) to investigate the distribution of cohesive devices, such as connectives. Subsequently, the descriptive statistical analysis was performed via R. The findings reveal a rather homogenous picture of the corpus on the overall grammatical level: the texts show similar frequencies regarding POS-tags. In contrast, the texts show a heterogeneous distribution of connectives. In conclusion, the results confirm that the corpus is suitable for the analysis of German learner language, more specifically, to investigate the use of cohesive devices by advanced learners of German.

References

Strobl, Carola (2020). Beldeko Summary Corpus v1.0.0, Eurac Research CLARIN Centre, http://hdl.handle.net/20.500.12124/15.

Weiss, Z., & Meurers, D. (2019). Broad linguistic modeling is beneficial for German L2 proficiency assessment. In Widening the Scope of Learner Corpus Research, Selected Papers from the Fourth Learner Corpus Research Conference (pp. 419-435). Presses Universitaires de Louvain.

Numeral Incorporation as Grammaticalization? A Corpus Study on DGS

Felicitas Otte, Anke Müller, Sabrina Wähl & Gabriele Langer

Keywords: sign language grammar, numeral incorporation, corpus study

In sign language morphology, one interesting instance of simultaneity present in many sign languages is numeral incorporation, that is, the integration of a numeral handshape into another sign. In German Sign Language (DGS), for example, temporal signs such as week1a can incorporate numeral handshapes to denote the meaning 'quantity of temporal unit'. Figures 1, 2, and 3 show three synonymous constructions meaning 'three weeks': a phrasal construction (figure 1), a fully incorporated form (figure 3), and an intermediate form (figure 2). The phrasal construction consists of two separate signs (three and week1a), each in its full form. In the intermediate form, which we will refer to as cliticization going forward, the numeral is produced for a short time and its handshape perseveres during the production of week1a. In the fully incorporated version, or affixation, only one sign remains. This sign keeps the handshape of the and retains the movement and hand orientation of week1a. We hypothesize that the emergence of the incorporating forms is an instance of language change and will explore it within the framework of grammaticalization theory.

Fig. 1: phrasal construction: three week1a

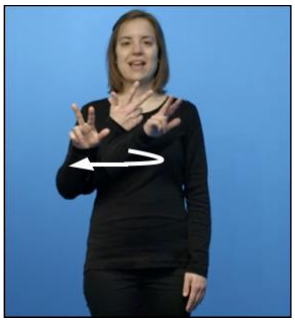


Fig. 2: cliticization: three=week1a

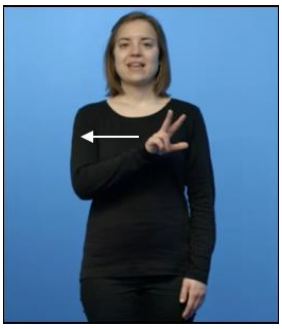


Fig. 3: affixation: three-week1a

Rathmann and Mathur (2010) describe numeral incorporation as numeral morphemes being attached to full signs with the numeral morphemes having originated from the corresponding full number signs. As such, the incorporation of numeral morphemes could be the result of grammaticalization, as its emergence shows several of the typical attributes, as described by Lehmann (1985): progression from free element to bound morpheme, loss of phonological substance, emergence of a closed paradigm, and reduction of the element's scope.

The aim of our study is to analyze the expansion of the usage of numeral morphemes and to estimate trends of grammaticalization using data from the DGS corpus. It contains about 560 h of video data,

signed by 330 signers (Hanke et al. 2020). We used the annotated part of the corpus as the base for this study. The signs we investigate have been evidenced as incorporating numerals in at least five instances. We compare the relative frequency of incorporation vs. phrasal constructions across different groups of signs. Following Bybee (2011: 77f.), the frequency of one type of construction as opposed to the others can be taken as an indicator of whether a grammaticalized item has become entrenched; we thus compare the percentage of tokens found for each construction.

We propose a cline of three different constructions that show the progression towards grammaticalized numeral incorporation: the phrasal construction, cliticization, and affixation. We suggest that the multiples of ten, a hundred, and a thousand, which are realized as a number handshape (from 1-10) combined with a meaningful movement, were reanalysed as a morpheme affixed to a base. In this interpretation, the number sign system becomes a model for a new set of constructions and the numeral morpheme is then affixed to other signs, such as temporal expressions. Himmelmann (2004) describes the base categories as host classes and this process as host class expansion.

From a list of 12 candidates that exhibit numeral incorporation, we compare seven signs with the numerals $num-hundred1 and $num-thousand1 as the presumed original host class. The numerals show a very strong tendency for affixation/full incorporation, as would be expected (see figure 4 and table 1). The existence of cliticized forms in this group could support our hypothesis that the numerals are a base for reanalysis of the number handshape as an affix. The non-numeral signs show diverging tendencies that split them into two groups: the first group, consisting of temporal signs, with over 50% affixation, and the second one with under 50%. The temporals show a strong tendency towards affixation with day2 being the outlier at 53,85% of affixation. The two signs in the last group, euro1 and old8b, appear by far the most often in phrasal constructions. However, particularly old8b does show the beginning of numeral incorporation with a relatively large percentage of cliticized forms. These frequencies might be influenced by other factors such as the phonological form of the base sign.

Overall, our findings are consistent with the hypothesis of a grammaticalization of numeral handshapes, as the frequencies in the different groups do suggest a host class expansion. In our presentation, we will further explore this claim and include an analysis of the use of numeral incorporation across different age groups, following the apparent time framework by Bailey et al. (1991).

Figure 4: Relative frequencies of the different constructions for numerals, temporals and other candidates (as of 10/2021)

Host class	Gloss		**Numeral Incorporation**		
		1. Phrasal Construction	**2. Cliticization**	**3. Affixation**	**Tokens in total**
Numerals	**$num-hundred1**	2	19	651	672
	$num-thousand1	0	52	388	440
	Tokens in total	**2**	**71**	**1039**	**1112**
Temporals	**hour2abc**	3	11	393	407
	day2	3	3	7	13
	week1abc	17	13	166	196
	month1	14	9	154	177
	year1abc	13	26	656	695
	Tokens in total	**50**	**62**	**1376**	**1488**
Other	**euro1**	64	2	3	69
	old8b	38	19	2	59
	Tokens in total	**102**	**21**	**5**	**128**
	Total	**154**	**154**	**2420**	**2728**

Table 1: Absolut frequencies of the different constructions for individual signs (as of 10/2021)

References

Bailey, Guy, Tom Wikle, Jan Tillery and Lori Sand (1991). “The Apparent Time Construct”. In: Language Variation and Change 3.3, pp. 241-264. DOI: 10.1017/S0954394500000569.

Bybee, Joan L. (2011). “Usage-based Theory and Grammaticalization”. In: The Oxford Handbook of Grammaticalization. Ed. by Bernd Heine and Heiko Narrog. Oxford University Press, pp. 69–78. DOI: 10.1093/oxfordhb/9780199586783.013.0006.

Hanke, Thomas, Marc Schulder, Reiner Konrad, Elena Jahn (2020). “Extending the Public DGS Corpus in Size and Depth”. In: Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. Marseille, France: European Language Resources Association (ELRA), pp. 75–82. url: https://www.aclweb.org/anthology/2020.signlang-1.12.

Himmelmann, Nikolaus P. (2004). “Lexicalization and Grammaticalization: Opposite or Orthogonal?” In: What Makes Grammaticalization? Ed. by Walter Bisang, Nikolaus P. Himmelmann, and Björn Wiemer. Vol. 158. Trends in Linguistics. Studies and Monographs [TiLSM]. Berlin, New York: Mouton de Gruyter, pp. 21–42. DOI: 10.1515/9783110197440.1.21.

Lehmann, Christian (1985). “Grammaticalization: Synchronic Variation and Diachronic Change”. In: Lingua e Stile 20, pp. 303–318.

Rathmann, Christian and Gaurav Mathur (2010). “Constraints on Two Types of Nonconcatenative Morphology in Signed Languages”. In: Deaf around the World: The Impact of Language. DOI: 10.1093/acprof:oso/9780199732548.003.0003.

Comparable corpora of spoken and signed languages: towards a pluri-semiotic perspective on language

Alysson Lepeut[1], Sébastien Vandenitte[1], Clara Lombart[1,2] & Laurence Meurant[1]

[1]Université de Namur (NaLTT, LSFB-Lab), [2]Université de Mons

Language use is a phenomenon that exhibits different methods of signaling: description, depiction, and indication (Peirce, 1955; Clark, 1996; Enfield, 2009). It has often been approached with a focus on description only, due partly to linguists relying on non-spontaneous (often written) data (Linell, 2011). However, from the mid 1980s onwards, scholars have shown an intricate connection between speech and gesture (McNeill, 1992; Kendon 2004). More recently, researchers working on gesture as well as on signed languages have highlighted that both spoken and signed interactions combine depictive, indicative, and descriptive strategies (Ferrara & Hodge, 2018). They have also called for a multi-channel and multimodal approach to the comparison of speakers' and signers' languaging (Vermeerbergen & Demey, 2007; Müller, 2018). However, essentially because of methodological constraints related to the availability and comparability of data, empirical work adopting this approach is scarce (Hodge et al., 2019). As a result, much remains to be understood about the semiotic diversity and the use of composite utterances across different human ecologies (Ferrara & Hodge, 2018). Since all communities use these three methods of communication, what is the impact of several factors such as different sensory experiences, the use of different articulators, or cultural differences on hearing speakers' and deaf signers' communicative practices?

Taking the example of LSFB (French Belgian Sign Language) and its ambient spoken language – French, this presentation seeks to illustrate the benefits of a corpus-based, multimodal, pluri-semiotic, and contrastive approach to language and interaction. First, the comparable corpora composed of the LSFB Corpus (cf. Figure 1; Meurant, 2015) and its French counterpart FRAPé (*FRAnçais Parlé, '*spoken French'; cf. Figure 2; Meurant et al., ongoing) will be described. In these corpora, pairs of signers or speakers interact with each other and perform the same set of tasks (i.e., narrations, explanations, descriptions, argumentations, and conversations). The methods used to collect the data are identical in both corpora.

Next, four studies relying on the LSFB-FRAPé datasets will be presented. Three of them mainly focus on one of the semiotic modes: 1) the use of descriptive cues in the marking of information structure; 2) the contribution of indication to the management of interaction; 3) the use of depictive resources in reporting action. The last one concentrates on the combination and alternation of semiotic strategies in the act of reformulation. A common result of those four studies is that speakers and signers use similar semiotic strategies to express, using their hands and/or the rest of their body (i.e., facial expressions, head movements, and body leans), the range of meanings investigated. Differences arise however between signed and spoken productions in terms of i) the frequency of occurrence and the number of bodily articulators they mobilize; ii) their formal properties such as the ways specific articulators are used (e.g., different types of movements or handshapes); and iii) their combination(s) with each other.

Ultimately, these works illustrate how using comparable signed and spoken data contributes to redefining linguistics to accommodate the composite nature of language (Enfield, 2009).

right hand: stay statue ca:man same touch-heart pt:pro3
translation: 'He remained in the same position, holding something in his hand and staring at her, like he'd fallen in love with her'.

Figure 1: Screenshot of a video from the LSFB Corpus (LSFB Corpus, Task 12, S59: 00:04:47:221 – 00:04:50:023

French sentence: Des étoiles, tu vois, pleines, comme ça.
translation: 'Stars, you see, fully colored, like that'.

Figure 2: Screenshot of a video from the FRAPé Corpus (FRAPé Corpus, Task 16, L002: 00:01:44:00 – 00:01:47:00)

References

Clark, H.H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Enfield, N.J. (2009). *The Anatomy of Meaning: Speech, Gesture, and Composite Utterances*. Cambridge: Cambridge University Press.

Ferrara, L., & Hodge, G. (2018). Language as Description, Indication, and Depiction, *Frontiers in Psychology 9*(1). https://doi.org/10.3389/fpsyg.2018.00716.

Hodge, G., Sekine, K., Schembri, A., & Johnston, T. (2019). Comparing signers and speakers: Building a directly comparable corpus of Auslan and Australian English, *Corpora 14*(1),63–76. https://doi.org/10.3366/cor.2019.0161.

Kendon, A. (2004). *Gesture. Visible Action as Utterance*. Cambridge: Cambridge University Press.

Linell, P. (2011). *The written language bias in linguistics: Its nature, origins and transformations*. London: Routledge.

McNeill, D. (1992). *Hand and mind: What gestures reveal about though*. Chicago: University of Chicago Press.

Meurant, L. (2015). *Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB)*, Laboratoire de Langue des signes de Belgique francophone (LSFB-Lab), FRS-F.N.R.S and Université de Namur. http://www.corpus-lsfb.be.

Meurant, L., Lepeut, A., Gabarró-López, S., Tavier, A., Vandenitte, S., Lombart, C., & Sinte, A. (Ongoing). *The Multimodal FRAPé Corpus: Towards building a comparable LSFB and Belgian French Corpus*, Laboratoire de Langue des signes de Belgique francophone (LSFB-Lab), University of Namur.

Müller, C. (2018). Gesture and Sign: Cataclysmic Break or Dynamic Relations?, *Frontiers in Psychology 9*(1651). https://doi.org/10.3389/fpsyg.2018.01651.

Peirce, C.S. (1955). *Philosophical writings of Peirce*. Mineola: Dover.

Vermeerbergen, M., & Demey, E. (2007). Sign + Gesture = Speech + Gesture? In M. Vermeerbergen, L. Leeson, & O. Crasborn (eds.), *Simultaneity in Signed Languages: Form and function* (p. 257–282). Amsterdam: John Benjamins.

Case marking alternation with psychological verbs in Spanish: Combining different corpus data sources

Victoria Vázquez Rozas & José María García-Miguel

Universidade de Santiago de Compostela; Universidade de Vigo

The study deals with the case alternation (accusative or dative) of the pronominal object clitic referring to the experiencer of Spanish psych verbs such as *alegrar* (class II of Belleti & Rizzi 1988). These verbs cast the experiencer as the object (a)-(b) (unlike class I; e.g. *amar* 'love') and exhibit the anticausative alternation (c) (unlike class III; e.g. *gustar* 'like').

a. La ocurrencia **los**=alegró sobremanera [HIS:042.20]
the witticism 3m.pl.**acc**=make_happy.pst.3sg exceedingly
'The witticism made them very happy'

b. Al viejo **le**=alegra ese recuerdo [SON:152.09]
to.the old_man 3sg.**dat**=make_happy.prs.3sg that memory
'That witty remark makes the old man happy'

c. Ella **se**=alegra al ver=le [SON:205.01]
she refl.3=make_happy.prs.3sg to.art see.inf=3sg.dat
'She is glad to see him'

Previous literature has put forward several factors behind this case variation, mainly related to the aspectual and agentive characteristics of the clauses. However, only Miglio et al. (2013) is based on a systematic corpus research and uses statistical tests to assess the significance of the variables under study. These authors manually annotated 1656 instances of 55 verbs from *CdE-hist*, and their analysis indicated that the fixed effects significantly associated with the dependent variable (Acc vs. Dat) were the stimulus animacy, the interactions of genre:tense and genre:clausal stimulus, and the authors' region.

Our research takes Miglio et al.'s findings into account, but we opt for a different strategy in handling the corpus data. We use data from the ADESSE database, which contains detailed lexical, semantic and syntactic information about a 1.45 million word corpus (ARTHUS), as well as data searched in CORPES (312 million words), a lemmatized and POS tagged reference corpus of Spanish, which is not only bigger but also reflects more regional variation (although it is not enriched with syntactic and semantic annotations comparable to those of ADESSE).

ADESSE allows the combination of very specific syntactic and semantic features so that we can identify all the instances of psychological verbs (class: 'sensation') with experiencer objects and anticausative alternation. We found 821 active and 720 anticausative clauses corresponding to 79 verbs, 59 of which were documented in texts from both Spain and America. The statistical analysis of these data proves that the animacy of the stimulus (animate, inanimate, or clause) and the general geographical area (Spain / America) are significant factors in the choice of experiencer case (Acc / Dat). As the corpus is relatively small, there are few data for individual verbs and for more specific geographical zones, but ADESSE nevertheless provides some information on regional variation in Peninsular Spanish.

Next, these same verbs were searched for in CORPES followed by a 3rd person clitic pronoun. After filtering out certain low-frequency verbs, 53 verbs and 25,406 cases were left. The statistical analysis of these frequency data confirms that both verb and geographical region are significant factors in the

choice of experiencer case; but not the interaction between them. Verbs and regions can be ranked from more to less dative preference, but verbs show a surprisingly homogeneous behavior regardless of the region. This suggest that some aspects of verb meaning play a relevant role.

References

ADESSE: Base de datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español. http://adesse.uvigo.es

ARTHUS: Archivo de Textos Hispánicos de la Universidad de Santiago.

CdE-hist: Corpus del Español: Género/histórico. https://www.corpusdelespanol.org/hist-gen/

CORPES: Corpus del Español del siglo XXI. http://web.frl.es/CORPES/

Belletti, Adriana & Luigi Rizzi. 1988. Psych-verbs and θ-theory. *Natural Language & Linguistic Theory* 6(3). 291–352. https://doi.org/10.1007/BF00133902.

Miglio, Viola G., Stefan Th. Gries, Michael J. Harris, Eva M. Wheeler & Raquel Santana-Paixão. 2013. Spanish *lo(s)-le(s)*: Clitic Alternations in Psych Verbs: A Multifactorial Corpus-Based Analysis. In *Selected Proceedings of the 16th Hispanic Linguistics Symposium*, 268–278. Somerville: Cascadilla Proceedings Project. http://www.lingref.com/cpp/hls/16/paper2939.pdf.

Control nouns: finite vs. non-finite adnominal clauses in German

Thilo Weber, Felix Bildhauer & Franziska Münzberg

In German, under certain conditions, finite clauses introduced by *dass* ('that') (1a) alternate with *zu* ('to') infinitive clauses (1b). The latter lack an explicit subject and are considered to be possible only where their implicit subject (often represented by PRO) is "controlled by" (and thus co-referent with) an expression in the surrounding context or has arbitrary/generic reference. The previous literature has focused on cases in which the clauses function as complements of a verb. Cases as in (1), where they are dependent on a 'control **noun**' (highlighted in bold), and where their status as complements or adjuncts is often less clear, have received less attention (but see e. g. Restle 2006 on adnominal infinitives).

(1) Solche Ausstellungen$_i$ hätten den **Vorteil**
Such exhibitions have. 3rd.past.subj the advantage …
'Such exhibitions would have the advantage'
a. dass sie$_i$ Unternehmen zusammenbringen
that they companies together.bring.3rd.pl.pres
'that they bring companies together'
b. Unternehmen zusammenzubringen
companies together.bring.inf
'of bringing companies together'

We investigate the distribution of the two clause types based on samples drawn from the German Reference Corpus (Kupietz et al. 2010) and the German web corpus DECOW16B (Schäfer & Bildhauer 2012), thus covering conceptually written registers (as typically found in newspaper texts) as well as less formal registers (as typically found in internet forums). As a first step, cases were identified in which finite clauses indeed appeared to be interchangeable with an infinitive ('choice contexts' in the sense of Rosenbach 2013; inter-rater agreement $\kappa_{Cohen} = 0.8$). The resulting data set (n = 6268 clauses, of which 651 are finite) was then subjected to further analysis.

We first show that the majority of control nouns represented in our sample only occurs with one of the two clause-types, with only a minority showing variation. Among the most frequently attested nouns only occurring with finite clauses, we find a striking proportion of retrospective nouns (e.g. *Rache* 'revenge', *Verzeihung* 'forgiveness') while the most frequently attested nouns only occurring with infinitives are mostly accounted for by prospective nouns (e.g. *Ehrgeiz* 'ambition', *Bereitschaft* 'willingness'). After discarding nouns that showed no variation at all in the sample, a mixed-effects logistic regression model was used to analyse the remaining data (n = 1928, of which 445 are finite) with respect to a number of factors that arguably influence speakers' choices within those variable contexts.

Among other things, the results indicate that, in line with previous work on verb-dependent infinitives and finite clauses (Brandt 2019), the alternation is influenced by the modality and agentivity of the subordinate-clause-predicate. Moreover, the alternation is guided not only by the presence or absence of a co-referring expression but also by its location relative to the control noun: infinitives are most likely to occur where a co-referring expression occurs within the same NP as the control noun. Finally, we show that infinitives are more likely in newspaper texts compared to internet forums.

In sum, our analysis suggests that the majority of examples from our initial data set represent categorical contexts with only a minority being subject to probabilistic constraints.

References

Brandt, Patrick. 2019. Alternation von *zu*- und *dass*-Komplementen: Kontrolle, Korpus und Grammatik. In Eric Fuß, Marek Konopka & Angelika Wöllstein (eds.), *Grammatik im Korpus*: *Korpuslinguistisch-statistische Analysen morphosyntaktischer Variationsphänomene*, 211–297. Tübingen: Narr.

Kupietz, Marc, Cyril Belica, Holger Keibel & Andreas Witt. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC ´10)*, 1848–1854. Valletta, Malta: European Language Resources Association (ELRA).

Restle, David. (2006). *Kontrollnomina. Eine Untersuchung zum Verhalten attributiver Infinitivkonstruktionen im Deutschen*. Habilitationsschrift. Ludwig-Maximilians-Universität München.

Rosenbach, Anette. (2013). Combining elicitation data with corpus data. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 278–294. Cambridge, MA: Cambridge University Press.

Schäfer, Roland & Felix Bildhauer. (2012). Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 486–493. Istanbul: ELRA.

Mining accordance and information source PPs of German *nach*

Halima Husić & Claudia Roch

Ruhr University Bochum

{halima.husic, claudia.roch}@ruhr-uni-bochum.de

In this abstract we sketch the results of an annotation study to discriminate *nach* PP senses that we aim to integrate into formal analyses. German *nach* (en. *after*) exhibits a range of different senses, most prominently a temporal sense (cf. Kiss et. al 2020). While Schröder (1986) and Helbig & Buscha (2007) subsume accordance and information source under each other, we commit to attributing them to different classes, manner adverbials and reportative evidentials. For illustration, consider the near-minimal pairs in (1)-(2). The accordance PP (1) describes the manner of the implementation event as conforming with Barcelona's model. Differently, the information source PP (2) introduces the reports as the source of information and identifies the contents as the proposition described in the main clause: the reports contain the information that the city implements superblocks.

(1) **Nach dem Modell** von Barcelona richtet die Stadt Superblöcke ein.
after the model of Barcelona implements the city superblocks ptkvz
'The city implements superblocks in accordance with Barcelona's model.'

(2) **Nach den Meldungen** richtet die Stadt Superblöcke ein.
after the reports implements the city superblocks ptkvz
'According to the reports the city implements superblocks.'

Note, (2) is ambiguous: The deverbal nominalization *Meldungen* (en. *reports*) can refer to (i) an event of reporting that occurs prior to a second event, which yields a temporal, or to (ii) the contents of the reports which combines with the information source sense. We conducted two annotation mining studies to grasp distributional differences based on 1600 sentences from NZZ's newswire corpus that varied in determiner realization and restrictions on internal arguments. We annotated the preposition sense, lexical ambiguity and the semantics of P's argument yielding important generalizations:

1. information source PPs require an internal argument denoting informational content, e.g., *Mitteilung* ('announcement'), and an overt realization of the referred content in the remaining sentence, underlined in (2).

2. The nouns are flexible in sortal type and depend on P's selectional restrictions.

3. Full PPs are more often ambiguous between a temporal and an accordance/information source sense than determinerless PPs.

As for semantic analyses of these phenomena, we propose that accordance *nach* is an adverbial manner supplement (for existing analyses cf. e.g. Anderson & Morzycki 2015). On the contrary, information source functions as a modal operator. accordance *nach* in (1) entails that the city implements superblocks, but information source *nach* in (2) does not, such that we assume that information source PPs create an opaque context. Here, we analyze the preposition *nach* as a reportative evidential operator similar to en. *according to (X)* (cf. Kaufmann & Kaufmann 2019, Krifka Forthcoming). Its semantics has to capture a subset relation between the propositional content of the internal argument of the P and the proposition denoted by the remaining sentence, i.e., in (2) the reports include (among others) also the proposition about the city's implementation of superblocks.

We have studied the semantic differences in two senses of German *nach*-PPs: accordance and information source. Based on an annotation mining study with 1600 occurrences of *nach*, we have argued for functional differences among the senses which lays a foundation for formal analyses of these less researched PP senses.

References

[1] Anderson, C. & Morzycki, M (2015). Degrees as kinds. *Natural Language* & *Linguistic Theory, 33*(3), 791–828.

[2] Helbig, G. & Buscha, J. (2007). *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht.* Langenscheidt.

[3] Kaufmann, M. & Kaufmann, S. (2019). Talking about sources. In M. Asatryan, Y. Song & A. Whitmal (Eds.), *Proceedings of NELS, 50*, Vol. 3 (pp. 77-90). GLSA.

[4] Kiss, T., Müller, A., Roch, C., Stadtfeld, T., Börner, K. & Duzy, M. (2020). Ein Handbuch für die Bestimmung und Annotation von Präpositionsbedeutungen im Deutschen. *SLLDS*, Vol. 2.

[5] Krifka, M. (Forthcoming). Layers of assertive clauses: Propositions, judgements, commitments, acts. In J. M. Hartmann & A. Wöllstein (Eds.), *Propositional Arguments in Cross-Linguistic Research: Theoretical and Empirical Issues*. Mouton de Gruyter.

[6] Schröder, J. (1986). *Lexikon deutscher Präpositionen.* VEB Verlag Enzyklopädie.

Multilingual SpeechReporting database: tools, methods and techniques

Ekaterina Aplonova, Tatiana Nikitina, Timofey Arkhangelskiy, Abbie Hantgan-Sonko, Izabela Jordanoska, Elena Sokur, Lacina Silué & Rebecca Paterson

The SpeechReporting database (Nikitina et al. 2021) is a collection of traditional folk stories in 10 languages around the world. It is updated regularly with newly available data, including data from new languages. All of the texts are transcribed, glossed, and translated, as well as being annotated for a number of discourse phenomena. In this talk, we present our workflow and the tools that we used to construct the corpora of the different languages comparable and suitable for cross-linguistic research. We also present two case studies that illustrate how our data can be used for quantitative and qualitative cross-linguistic comparison.

Our toolkit includes the following technologies: ELAN-CorpA (Chanard 2015; 2019) and ELAN Tools (Chanard et al. under development), The SpeechReporting Template (Nikitina et al. 2019), and Tsakorpus (Arkhangelsky 2019). The workflow contains the following steps:

1. For segmenting, transcribing, and glossing new data we use ELAN-CorpA (which has an extension for glossing time-aligned transcription). Existing corpora that are already glossed in FLEx or in Toolbox can be imported into ELAN using ELAN Tools.
2. Once the data are time-aligned and glossed, these are annotated using the SpeechReporting Template for instances of reported speech and thought, their semantic and syntactic type, and the use of pronominal elements used to refer to participants.
3. Annotated ELAN files are transformed into JSON files and uploaded to a server. The Tsakorpus interface was configured for additional searches from SpeechReporting template values.

Based on quantitative data from the two largest corpora of genealogically unrelated and typologically different languages Wan (Mande, West Africa) and Chuvash (Turkic, Central Russia), we specifically observed that interjections are associated with a quotative function; they help signal instances of reported speech. Moreover, the presence of an interjection is negatively correlated with the presence of a grammaticalized quotative element. Constructional typology of reported thought is an example of a qualitative case-study based on our multilingual dataset. We identified three strategies for the expression of thought: (1) reported speech constructions directly recruited for the expression of thought or “inner speech”; (2) structurally diverse reported thought constructions without equivalent among expressions of reported speech; (3) speech-to-thought coercion.

In addition to the two case studies discussed above, we are currently working on other research questions related to discourse reporting that cannot be answered without meticulously annotated and comparable corpora. We believe that our methodology and tools can be also relevant for a broader range of linguistic topics.

References

Arkhangelskiy, Timofey. 2019. Corpora of social media in minority Uralic languages. In Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages, 125–140. Tartu, Estonia. http://volgakama.web-corpora.net/Social_media_corpora_IWCLUL2019_final.pdf (5 October, 2020).

Chanard, Christian. 2015. ELAN-CorpA: Lexicon-aided annotation in ELAN. In Amina Mettouchi, Martine Vanhove & Dominique Caubet (eds.), Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages (Studies in Corpus Linguistics 68), 311–332.

Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.68.10cha. https://benjamins.com/catalog/scl.68.10cha (2 October, 2020).

Nikitina, Tatiana, Aplonova, Ekaterina, Hantgan-Sonko Abbie, Jordanoska, Izabela, and Perekhvalskaya, Elena (eds.) The SpeechReporting Corpus: Discourse Reporting in Storytelling. Villejuif-Paris: CNRS-LACITO.

Nikitina, Tatiana, Hantgan-Sonko Abbie & Chanard Christian. 2019. *Reported speech annotation template for ELAN* (The SpeechReporting Corpus). Villejuif-Paris: LLACAN.

Grammatical rule vs. linguistic theory: the case of reflexive pronouns in locative PPs

Amani Mejri

Background: Reflexive pronouns are locally bound by a coragument antecedent, based on the structural constraints of binding (Dillon, 2014; Ruijendijk & Schmacher, 2020). They can equally look for antecedents that are not structurally defined, hence beyond the grammatical boundaries that local binding defined (Charnavel & Sportiche, 2016; Pollard & Sag, 1992; Reinhart & Reuland, 1993). A recurrent example is the distribution of reflexive pronouns in locative PPs. However, based on the standard rule the use of the reflexive in these constructions is ungrammatical (https://learnenglish.britishcouncil.org/grammar/english-grammar-reference/reflexive-pronouns).

-John$_i$ pulled the blanket over himself$_i$/ him$_i$

-The soldiers$_i$ put their guns in front of themselves$_i$/ them$_i$

-Jane$_i$ saw a snake next to her$_i$/ herself$_i$

In linguistic theory, reflexives in these constructions encode perspective/ point-of-view, or signal a body-oriented reading intended by the speaker and not available via the use of the non-reflexive pronoun (Lederer, 2013). The former is postulated in Reflexivity theory (Reinhart & Reuland, 1993) and in a recent study conducted by Kaiser (2020) on perspective shifts and pronominal choices. The latter claim is underscored in a corpus-based study conducted by Lederer (2013, 2009) that defends the hypothesis of encoding physical closeness via the reflexive in place phrases, resting on Kuno's empathy theory (1987).

Aim of the work: This talk addresses user compliance with the standard grammatical rule of the non-reflexive in place phrases via retrieving corpus-based data and conducting a small-scale follow up experiment. The aim is to gauge the validity of the claims forwarded in linguistic theory on the relevance of the reflexive in encoding distance.

Corpus & Methodology: An initial corpus-based analysis examines the distribution of pronouns in locative PPs (close to/ next to/ in front of/ behind/ around/ behind) in the BNC and the recent Timestamped Corpus. The obtained dataset is used to quantitatively measure the frequency of the reflexive and the non-reflexive in these constructions. Upon this measure, the second phase consists in using the corpus dataset in a follow-up experiment with 20 English native speakers. Test sentences are associated with visuals where the object is once placed close to the referent's body, and once away from the referent's body. The study informants have to choose which sentence (including the reflexive or the non-reflexive) best describes the picture:

-Max$_i$ put the opened bottle down next to him$_i$/ himself$_i$

Tentative Results and Implications The corpus findings in the BNC show that English native speakers tend to use the non-reflexive with the set of the examined place prepositions. Irrespective of object proximity or distance, the non-reflexive pronoun is the default and the most frequent use in the corpus (90%). However, in the recent Timestamped corpus covering the British variety in journalism, reflexives occurred in locative PP (24%). In their turn, the follow-up study results show that speakers preferred the non-reflexive form (78%), although distance and closeness are clearly shown on a grid in the visuals. Notwithstanding their variation, these results suggest that native speakers comply with the grammatical rule and consider using the reflexive has no input regarding physical closeness or distance.

What remains controversial is that some reflexives still occur in locative PPs, and based on linguistic theory the aim is to encode physical distance. Based on these findings, this view is not viable and the semantics of the reflexive pronoun does not contribute in encoding proximity or distance in space.

References

Charnavel, I., & Sportiche, D. (2016). Anaphor binding: What French inanimate anaphors show. *Linguistic Inquiry*, *47*(1), 35-87.

Dillon, B. (2014). Syntactic memory in the comprehension of reflexive dependencies: an overview. *Language and Linguistics Compass*, *8*(5), 171-187.

Kaiser, E. (2020). Shifty behavior: Investigating predicates of personal taste and perspectival anaphors. *Semantics and Linguistic Theory*,*30*, 821-842

Kuno, S. (1987). *Functional syntax: anaphora, discourse and empathy*. Chicago: University of Chicago Press

Lederer, J. S. (2009). *Understanding the self: The distribution of anaphora within prepositional phrases*. University of California, Berkeley.

Lederer, J. (2013). Understanding the Self: How spatial parameters influence the distribution of anaphora within prepositional phrases. *Cognitive Linguistics, 24*(3), 483-529.

Reinhart, T, Reuland E. (1993). Reflexivity. *Linguistic inquiry*, *24*(4), 657-720.

Pollard, C., & Sag, I. A. (1992). Anaphors in English and the scope of binding theory. *Linguistic inquiry*, *23*(2), 261-303.

Ruigendijk, E., & Schumacher, P. B. (2020). Variation in reference assignment processes: psycholinguistic evidence from Germanic languages. *The Journal of Comparative Germanic Linguistics*, *23*(1), 39-76.

Can network science help explain development of second language complexity?

Susanne DeVore

The usage-based theory of language acquisition posits that the frequency and distribution of words and constructions that a learner receives in the input is directly related to cognitive processes of language development (Ellis et al., 2014; P. of P. B. MacWhinney & O'Grady, 2014; Tomasello, 2003). Research undertaken from this theoretical approach has focused on the relationship between verbs and the verb-argument constructions (VACs) in which they appear and to a lesser extent on phraseological constructions (Ellis & Ferreira-Junior, 2009; Paquot, 2019) and have found that in general, learners use words that are more strongly associated at higher levels of language proficiency.

When looking at language development, three things complicate these analyses: First, there are constructions besides clausal and phrasal constructions, such as morphological and lexical; Second, these constructions are embedded within each other – lexical items within phrases and lexical items and phrases within clauses; Third, there is overlap between them – a noun may appear in either a noun phrase or a prepositional phrase. Although current methods used in second language acquisition do not capture this kind of interaction and overlap, Ke & Yao (2008) and Ninio (2006) show that network science can be used to do so in child language acquisition. This study asks:

Can network science be used to analyze the emergence of grammar in adult language learning by capturing the interactions within and across constructions at different levels?

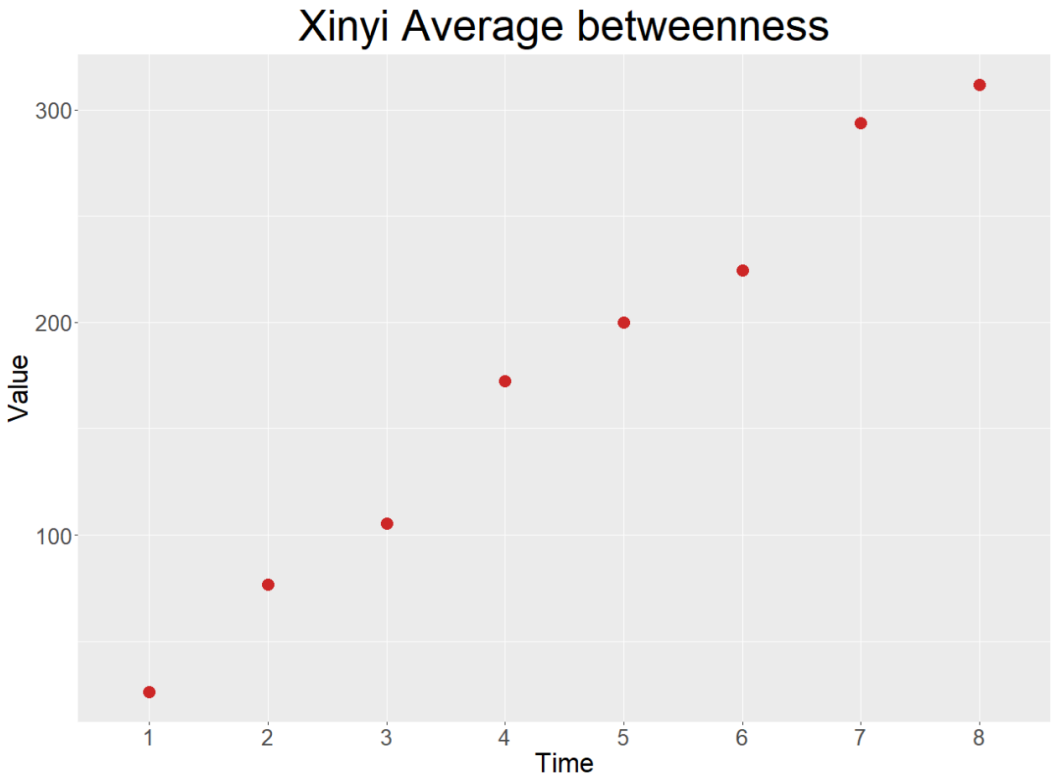
In order to do this, this project developed a longitudinal corpus of early learners of Mandarin. This was dependency parsed using StandfordCoreNLP (Manning et al., 2014) and the dependency relationships were then used to create a network with the words being the nodes and dependency relationships acting as connections. Network science could then be used to quantify the network. Two analyses were initially conducted and indicated that *average betweenness* and *average degrees* showed clear trends in the longitudinal data and were also significant predictors of proficiency in a large cross-sectional corpus. Because *average betweenness* is characterized as the extent to which a word acts as a bridge between different subsystems in the linguistic system, community structure was then used to quantitatively identify the subsystems that emerge and then qualitatively identify which (if any) linguistic structures they represent. Development of these structures is then traced over the 20-week period.

The preliminary analysis shows that

1. Learners increase not just the size of the network over time, but the average number of connections between words.
2. *Average betweenness* appears to be an important factor in early language development (Figure 1) and was also a significant predictor of proficiency across a wide range of proficiency levels (Table 1).
3. The quantitatively identified communities (or subsystems) represent phrasal and clausal structures and these can be used to trace development over time.

These results indicate that network science can be used effectively to quantify the individual trajectories of language learners in a way that reveals new and relevant information about the processes involved in language development.

Figure 1: *Average betweenness score*



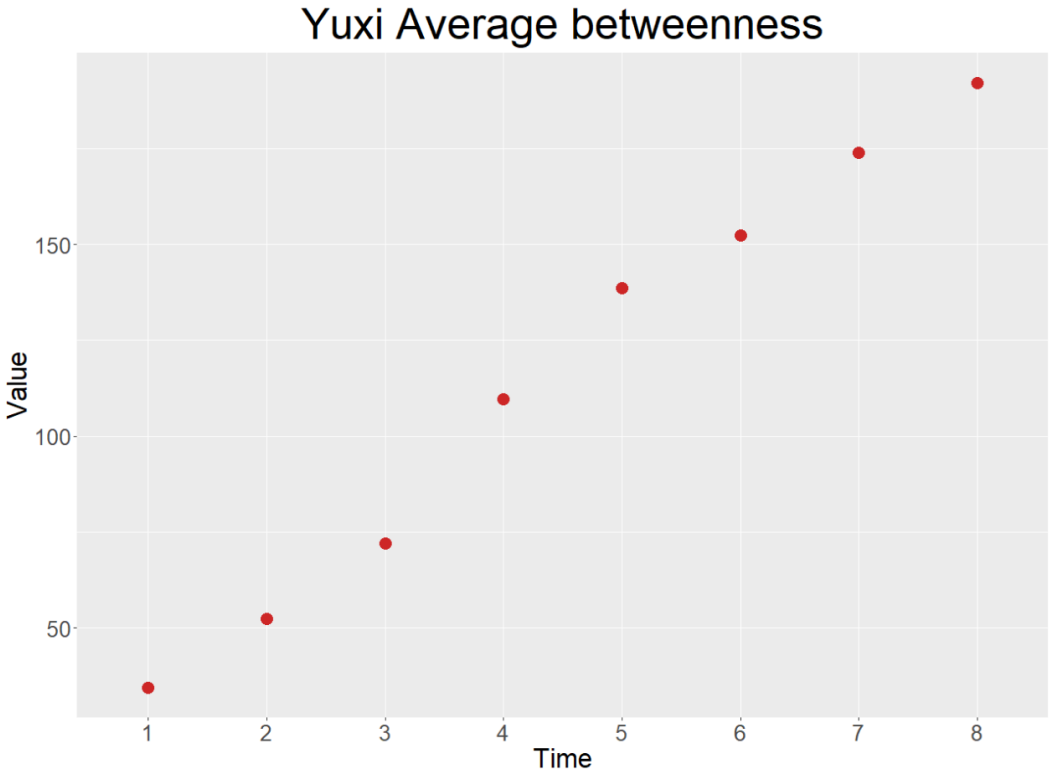


Table 1: *Network Science Indices Model Summary*

	LMG	**Est**	**SE**	**t value**	**Pr(>|t|)**

intercept	NA	30.724	5.774	5.321	p < 0.001
ave_betweenness	0.353	0.124	0.008	14.864	p < 0.001
ave_deg	0.018	7.105	1.887	3.766	p < 0.001

Adjusted R2: .3677; F(2, 388) = 114.4; p < .001

References

DeVore, S., & Kyle, K. (under review). Assessing syntactic and lexicogrammatical use in L2 Mandarin. *Language Learning*.

Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*, 188–221. https://doi.org/10.1075/arcl.7.08ell

Ellis, N. C., O'Donnell, M., & Romer, U. (2014). Usage-based language learning. In B. MacWhinney & W. O'Grady (Eds.), *The Handbook of Language Emergence: Handbook of Language Emergence*. John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/uhm/detail.action?docID=1895429

Ke, J., & Yao, Y. (2008). Analysing Language Development from a Network Approach. *Journal of Quantitative Linguistics*, *15*(1), 70–99. https://doi.org/10.1080/09296170701794286

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*, 513–535. https://doi.org/10.1177/0265532217712554

MacWhinney, P. of P. B., & O'Grady, A. P. D. of S. and A. W. (2014). *The Handbook of Language Emergence: Handbook of Language Emergence* (B. MacWhinney & W. O'Grady, Eds.). John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/uhm/detail.action?docID=1895429

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Ninio, A. (2006). *Language and the Learning Curve: A new theory of syntactic development*. University Press. https://doi.org/10.1093/acprof:oso/9780199299829.001.0001

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

"I AGREE TO THE TERMS AND CONDITIONS": legal-lay language comprehensibility in a bilingual English- Italian corpus

Lucia Busso

From terms of use of websites to consumer contracts, legal-lay language (henceforth: LLL) – i.e., the genre used to communicate legal contents to non-specialised audiences – has become increasingly important in our society and everyday life (Tiersma 1999; Benoliel & Becher 2019). It has variously been noted how LLL's lack of comprehensibility leads to many problems, both legislative and linguistic (Bhatia 1983; Frade 2007; Haapio 2011). Previous research has mainly used computational techniques to analyse legal language and LLL's complexity and readability (Brunato & Venturi 2014; Van Boom et al. 2016). Other works have instead used psycholinguistic tests to explore LLL comprehensibility (Conklin et al. 2019). However, a linguistic study of LLL and of its characteristic linguistic markers still lacks from the literature. The present contribution aims at filling this gap by combining corpus-based evidence with experimental methods.

The study takes the constructionist standpoint that language is formed by *constructions*, holistic pairs of form and function (Goldberg 2006). Construction Grammar is in fact increasingly applied in both synchronic and diachronic corpus-based studies (Hilpert 2013; Perek, 2016), as well as in the analysis of genre (Groom 2019).

Three main exploratory research questions underpin the study:

1. Is LLL more readable than legal jargon in both English and Italian?
2. Are readability scores accurate in measuring LLL's accessibility compared to speaker judgments?
3. What are the principal components of Italian and English LLL's lexical and syntactical complexity with respect to legal jargon and general-domain written prose?

As data, we use *CorIELLS* (CORpus of Italian and English Legal-lay textS), a specialised 1M words corpus of LLL. CorIELLS is the first openly available corpus of legal-lay language, and includes several sub-genres: terms of websites, bank contracts, utilities legal notices, and summaries of European legislation (Busso, forthcoming). We compare *CorIELLS* to two other genres: specialised legal jargon (using the *EURLEX* English corpus [Baisa et al., 2016] and the legal subcorpus of *CORIS* for Italian [Rossini-Favretti, 2000]) and written prose (using the *BNC* Imaginative subcorpus [BNC 2007] and the fiction subcorpus of *CORIS* [Rossini-Favretti et al, 2002]).

We focus on a set of constructions of different levels of schematicity (Barðdal 2008): nominalisations heading PP-attachment chains, modal verbs, reduced participial constructions, and passives. These constructions were selected based on previous research on legal and bureaucratic grammatical features in both Italian and English (Goźdź-Roszkowski & Pontrandolfo 2015; Mori 2019). Firstly, a sample of 120 concordances (30 per each construction) is extracted from the corpora and analysed in terms of readability scores. The same concordance lines are then presented as stimuli in a survey to a pool of participants and results are compared against text-based readability measures. Secondly, using the same concordances, lexico-syntactic complexity metrics are computed with *Profiling-UD* (Brunato et al., 2020) and analysed using Principal Component Analysis.

Results for both languages overall suggest that LLL is at an intermediate level of readability and comprehensibility, and that readability scores and human judgments behave similarly, although the

former overestimate text difficulty. Moreover, the PCA analysis reveals differences between English and Italian: English LLL is lexically similar to legal jargon, while syntactically it shows much more overlap among the other two genres, suggesting a blended nature. Italian appears to be different from both reference genres for lexical complexity, while syntactically it is a 'perfect mix' of the two neighbour genres, with vast overlaps.

References

Adler, M. (2012). *The Plain Language Movement*. Oxford: Oxford University Press.

Baisa, V., Michelfeit, J., Medveď, M., & Jakubíček, M. (2016). European union language resources in sketch engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2799-2803).

Barðdal, J. (2008). *Productivity: Evidence from case and argument structure in Icelandic* (Vol. 8). Amsterdam: John Benjamins.

Benoliel, U., & Becher, S. I. (2019). The Duty to Read the Unreadable. *SSRN Electronic Journal*.

Bhatia, V. K. (1983). Simplification v. Easification— The Case of Legal Texts. *Applied Linguistics*, 4(1), 42– 54.

BNC Consortium (2007). *The British National Corpus, XML Edition*, Oxford Text Archive, http://hdl.handle.net/20.500.12024/2554.

Brunato, D., and Venturi, G. (2014). Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici. *Informatica e diritto*, XL(XXIII), 111-142.

Busso (forthcoming) CorIELLS: a specialised bilingual corpus of English and Italian legal-lay language. *Proceedings of the workshop "Forensic linguistics between scientific research and legal practice", LIV International Conference of the Italian Linguistics Society*.

Conklin, K., Hyde, R., & Parente, F. (2019). Assessing plain and intelligible language in the Consumer Rights Act: a role for reading scores? *Legal Studies*, 39(3), 378-397.

Frade, C. (2007). Power dynamics and legal English. *World Englishes*, *26*(1), 48-61.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goźdź Roszkowski, S., and Pontrandolfo, G. (2015). Legal phraseology today: corpus-based applications across legal languages and genres. *Fachsprache: Internationale Zeitschrift für Fachsprachenforschung-didaktik und Terminologie*, *37*(3), 130-139.

Haapio, H. (2011). Contract Clarity Through Visualization: Preliminary Observations and Experiments. *Proceedings of the 15th International Conference on Visualization*, IEEE Computer society.

Hilpert, M. (2013) *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax (Studies in English Language*). Cambridge: Cambridge University Press.

Mori, L. (2019). Complessità sintattica e leggibilità. Un monitoraggio linguistico per la valutazione dell'accessibilità dei testi legislativi europei e italiani. *Studi Italiani di Linguistica Teorica e Applicata*, (48), 627-657.

Nikiforidou, K. (2018). Genre and constructional analysis. *Pragmatics & Cognition* (25), 543 – 575.

Perek, F. (2016). Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, *54*(1), 149-188.

Rossini-Favretti, R., Tamburini, F., De santis, C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model, in A. Wilson, P. Rayson, T. Mcenery (eds.), A rainbow of corpora: Corpus linguistics and the Languages of the world (pp. 27-38). Munich: Lincom-Europa.

Tiersma, P. M. (1999). *Legal language*. Chicago: University of Chicago Press.

Van Boom, W. H., Desmet, P., and Van Dam, M. (2016). 'If It's Easy to Read, It's Easy to Claim'—The Effect of the Readability of Insurance Contracts on Consumer Expectations and Conflict Behaviour. *Journal of Consumer Policy*, 39(2), 187–197.

Diminutive expressions in bi/multilingual discourse: a cross-community study of Spanish-English codeswitching in Miami and Northern Belize

Margot Vanhaverbeke, Renata Enghels & Osmer Balam

Universiteit Gent; Universiteit Gent; College of Wooster

The current investigation of the diminutive sheds new light on when and how bi/multilingual speakers alternate their languages – here Spanish and English – in two bi/multilingual contexts, namely Miami and Belize. A prominent question regarding the linguistic nature of intrasentential codeswitching remains whether bi-/multilinguals switch according to an underlying, shared grammar, or according to either individual preferences or community norms. In order to contribute to this question, it is essential to compare the bilingual speech of different Spanish-English communities. However, comparative studies of this kind are not widespread (exc. Torres & Potowski 2008; Blokzijl et al. 2017; Parafita-Couto & Gullberg 2019). The primary aim of this presentation is therefore to contribute to this research field and to examine and contrast the diminutive in the multilingual communities of Miami and Belize. The Bangor Miami Corpus and the Northern Belize Corpus (Balam 2016) serve as a basis for this study.

This cross-community corpus analysis focuses on the diminutive construction in bi/multilingual speech. The conceptual category of diminutiveness can be used to convey a dimensional, scalar or temporal reduction (e.g. *a tiny window, a small group, a little while),* but also to express a wide array of positive or negative connotations towards the diminutivized entity (e.g. *my hubby, that nasty little dog*) (i.a. Bagasheva, 2020; Schneider, 2003). With regard to its morphological configuration, various apparatus can be employed to form diminutives, including affixation (e.g. *mini*skirt, kitty), reduplication (e.g. *a goody-goody*), truncation (e.g. *hon < honey*), and periphrastic constructions (e.g*., the tiny castle, a little strange*) (Schneider, 2013). How speakers express diminutiveness may therefore strongly diverge across languages, as is the case in Spanish and English. While Spanish primarily uses diminutive suffixes (*-ito/a, -illo/a, -ico/a, -uelo/a,* etc.; RAE, 2011), English mostly turns to analytic constructions ((a) *little (bit of), small, tiny,* etc.; Hägg, 2016; Schneider, 2003). Accordingly, the diminutive potentially constitutes a conflict site in Spanish-English codeswitching (Enghels & Vanhaverbeke, 2020; 2021).

The present study discusses the formation and the use of diminutive constructions in the conversational Bangor Miami corpus[1] and in a corpus of sociolinguistic interviews carried out in Belize (Balam, 2016)[2]. The results show that multilingual speakers of Miami and Belize form diminutives with distinct tendencies. Both regarding type and token frequency, Miami bilinguals use significantly more analytic markers than Belize multilinguals. As for synthetic diminutives, they employ different types (e.g. *-ico* in Miami and *-ino* in Belize), although *-ito* remains the prevalent suffix in both communities. Remarkably, Belize multilinguals switch significantly more within the diminutive construction (e.g. *un[SP] lee[Kriol] purs[EN]-ito[SP], 'a small purse'*) than Miami bilinguals. From a functional perspective, Miami and Belize multilinguals again use the diminutive in different manners. In Miami, bilinguals differentiate between function and diminutive language, using English markers mainly to express objective meanings (e.g. *un little estante*) and Spanish ones for affective connotations (e.g. *un partimecito*), while Belize speakers do not make that distinction, using both Spanish and English diminutive to communicate either meaning.

[1] See http://bangortalk.org.uk/speakers.php?c=miami for more information on this online corpus.
2 See Balam (2013) for a detailed overview.

References

Bagasheva, M. (2020). *Ways of expressing the category of diminutiveness in English*. Paisievie Cheteniya, Plovdiv University.

Balam, O. (2013). Overt Language Attitudes and Linguistic Identities among Multilingual Speakers in Northern Belize. *Studies in Hispanic and Lusophone Linguistics*, *6*(2).

Balam, O. (2016). Language use, language change and innovation in Northern Belize contact Spanish (Doctoral dissertation). University of Florida, USA.

Blokzijl, J., Deuchar, M., & Parafita Couto, M. C. (2017). Determiner Asymmetry in Mixed Nominal Constructions: The Role of Grammatical Factors in Data from Miami and Nicaragua. *Languages*, *2*(4), 20.

Enghels, R., Vanhaverbeke, M. (2020). La expresión de valores diminutivos en contextos escritos de cambio de código: un análisis comparativo de novelas latinas. *Glosas,* 9(8), 39 – 55.

Hägg, A. T. (2016). *A Contrastive Study of English and Spanish Synthetic Diminutives*. Thesis. Oslo: Oslo University.

Parafita Couto, M. C., & Gullberg, M. (2019). Code-switching within the noun phrase: Evidence from three corpora. *International Journal of Bilingualism*, *23*(2), 695–714.

Real Academia Española. (2010). La derivación apreciativa. *Nueva Gramática de la Lengua Española Manual,* 163–172. Madrid: Espasa Calpe.

Schneider, K. P. (2003). *Diminutives in English*. Tübingen: Max Niemeyer Verlag GmbH.

Schneider, K. P. (2013). The truth about diminutives, and how we can find it: Some theoretical and methodological considerations. *SKASE Journal of Theoretical Linguistics*, 10(1), 137 – 151.

Torres, L., & Potowski, K. (2008). A comparative study of bilingual discourse markers in Chicago Mexican, Puerto Rican, and MexiRican Spanish. *International Journal of Bilingualism*, *12*(4), 263–279.

Vanhaverbeke, M., Enghels, R. (in press). Diminutive Constructions in Bilingual Speech: a case study of Spanish-English Codeswitching. *Belgian Journal of Linguistics, 35.* John Benjamins Publishing company.

On the linking adverbial "besides": A corpus-based study

Sugene Kim

Nagoya University of Business & Commerce

This study investigates the phraseology of the linking adverbial "besides" to unravel why—unlike similar-meaning transitions such as "in addition"—it sounds unnatural in some contexts. To discern patterns underlying its use, concordances were obtained from a set of academic written English corpora—the Michigan Corpus of Upper-level Student Papers (MICUSP) and the British Academic Written English (BAWE) corpus—and academic sections of another set of English corpora—the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). After restricting the search to texts written by native English speakers for MICUSP and the BAWE corpus or to the academic text type for the BNC and COCA, there remained 154 extracts in which "besides" is used as a linking adverbial. Discourse analysis was performed to identify the discourse environment in which it is used to bind the sentences together. Results suggested that the linking adverbial "besides" co-occurs frequently with pragmalinguistic features typical of argumentations. In 91.5% of the data, the use of "besides" was found to be regulated by the negativity-conditioned nature of what precedes it. In most cases, negation was acquired using an explicit negative word (e.g., "not") (Clark, 1976); a negative affix (e.g., "dis-"); adverbs or determiners with negative implication, such as the exclusive focus particle "only," which has a connection with negation (Pullum & Huddleston, 2002); or covertly negative lexical items (e.g., "fail") (Klima, 1964). Negative assertions were also made by means of constructions conveying a negative proposition, such as a rhetorical question, which functions as a negative assertion; the subjunctive mood, indicating uncertainty, hypotheticality of the propositional content, or counter-factuality (Bennett, 2003; Celce-Murcia & Larsen-Freeman, 1999); or the comparative construction including comparisons with the degree adverb "too" (Meier, 2003; Stassen, 1984). When negativity in the preceding clause is not expressed syntactically or semantically, the clause introduced by "besides" was shown to act as a rhetorical cue for treating previously stated arguments as premises for an inference of a de facto proposition, which takes a negative form without fail.

Keywords: *besides*, linking adverbial, corpus, discourse analysis

References

Bennett, Jonathan. 2003. *A philosophical guide to conditionals*. Oxford: Clarendon Press.

Celce-Murcia, Marianne & Larsen-Freeman, Diane. 1999. *The grammar book: An ESL/EFL teacher's course* (2d ed.). Boston: Heinle & Heinle.

Clark, Herbert H. 1976. *Semantics and comprehension*. The Hague: Mouton.

Klima, Edward S. 1964. Negation in English. In Jerry A. Fodor & Jerrold J. Katz (eds.), *The structure of language: Readings in the philosophy of language,* 246–323. Eaglewood Cliffs, NJ: Prentice-Hall.

Meier, C. 2003. The meaning of *too, enough,* and *so . . . that*. *Natural Language Semantics* 11(1). 69–107.

Pullum, Geoffrey K. & Huddleston, Rodney. 2002. Negation. In Rodney Huddleston & Geoffrey K. Pullum (eds.), *The Cambridge grammar of the English language,* 785–849. Cambridge: Cambridge University Press.

Stassen, L. 1984. The comparative compared. *Journal of Semantics* 3(1–2). 143–182.

Causative constructions with the Spanish motion verbs *llevar* and *traer*: a diachronic corpus-based analysis

Julio Torres Soler

University of Alicante

In the last decade, the use of Construction Grammar to the study of diachronic syntax has experienced a strong success (Hilpert 2013; Traugott & Trousdale, 2013; Barðdal et al., 2015). One area in which Diachronic Construction Grammar (DCxG) has turned out to be particularly fruitful is the evolution of verbal periphrases, since this model of grammar allows to highlight several aspects of grammatical change, which are not well enough explained under the only perspective of the grammaticalization theory (Garachana, 2020). In the field of the Spanish language, Enghels and Comer (2020a, 2020b) employed the (D)CxG framework to the analysis of the causative constructional network, which has as a prototype the causative construction with *hacer* (e. g. *ella hizo llorar a su hermana* 'she made her sister cry'). The authors argue that the causative constructions with *poner* and *meter* ('to put') constitute a particular subschema inside the causative constructional network, which they call *locative*, which has some common features but also significant differences that are observable in their diachronic evolution.

In this work, we attempt to extend the diachronic analysis of the Spanish causative constructional network to the causative constructions with the caused-motion verbs verbs *llevar* ('to take') and *traer* ('to bring') (e. g. *El frío llevó a los pobladores a construir refugios* 'the cold made the settlers buid shelters' and *el panadero trajo a sus clientes a gozar de nuevos sabores* 'the baker made his clients enjoy new flavours'). Therefore, we elaborate a corpus for the study of those pair of constructions, made up of more than 300 occurrences from different periods of the history of Spanish. In the following months, we will classify the occurrences according to several syntactic and semantic variables, such as the presence or not of intercalated lexical elements, the animacity of the main subject and the subordinate subject, and the dynamicity of the infinitive. Then, we will carry out a quantitative analysis of the data, which will highlight similarities and differences in the evolution of the causative constructions with *llevar* and *traer* and will facilitate the placement of the two micro-constructions under study in the causative constructional network of Spanish.

Our preliminary results show that *traer* was employed in the Middle Age more frequently than *llevar*, whereas in later periods this situation was reversed. This might be partially explained by the semantic input of the caused-motion verbs, which progressively acquired the strong deictic character that they show nowadays. We expect that the rise of the centripetal deictic meaning of *traer* as a motion verb ('bring') discouraged its use as an auxiliary verb in causative periphrases, in favour of *llevar*, whose meaning as motion verb adopted a centrifugal deictic value ('take'). On the other hand, as it is typical of the last steps of a constructionalization process, we expect the analyzed constructions to have experienced a gradual increase of productivity and incorporation (Traugott & Trousdale, 2013).

References

Barðdal, J., E. Smirnova, L. Sommerer & S. Gildea (Eds.) (2015). *Diachronic Construction Grammar*. Amsterdam/Philadelphia, John Benjamins Publishing Company.

Enghels, R. & M. Comer (2020a). La red construccional de perífrasis causativas: definición, comparación sincrónica y evolución diacrónica. In M. Garachana (Ed.), *La evolución de las perífrasis verbales en*

español. Una aproximación desde la gramática de construcciones diacrónica y la gramaticalización, (pp. 161-196). Berlin: Peter Lang.

Enghels, R. & M. Comer (2020b). Causative and inchoative constructions with *poner* and *meter* ('to put') in Spanish: A diachronic constructional approach. In J. Fernández Jaén & H. Provencio Garrigós (Eds.), *Changes in Meaning and Function. Studies in historical linguistics with a focus on Spanish* (pp. 21-45). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Garachana, M. (2020). ¿Es necesaria una gramática de construcciones diacrónica? In M. Garachana (Ed.), *La evolución de las perífrasis verbales en español. Una aproximación desde la gramática de construcciones diacrónica y la gramaticalización*, (pp. 45-72). Berlin: Peter Lang.

Hilpert, M. (2013). *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press.

Traugott & Trousdale (2013). *Constructionalization and constructional changes*. Oxford: Oxford University Press.

Marrying corpus linguistics, dialectology, and philology: the grammaticalisation of the future and conditional in Central Ibero-Romance (13th-14th century)

Antoine Primerano

Universiteit Gent / Humboldt-Universität zu Berlin

As a result of the yet incomplete grammaticalisation of the Late Latin periphrasis [infinitive + habere], the medieval Ibero-Romance future and conditional verb forms (fc) show synchronic variation between synthetic, univerbated variants (sfc, e.g. *cantaré*, *cantaría*), and analytic ones (afc, e.g. *cantar lo é*, *cantar lo ía*), where an unstressed personal pronoun interrupts the [infinitive-like form + ending] sequence (Company 2006, *inter alia*). This variation is said to be influenced by clitic placement, as the sfc can occur either with no unstressed pronoun (1a) or in the syntactic-pragmatic environments where it has to be preverbal (1b), while the afc always occur with a mesoclitic pronoun (1c), in the environments where it is postverbal with other tenses (Bouzouita 2011, Castillo Lluch 2002, *inter alia*).

(1) 13th-century Castilian:
- a. *Si non, **dexar-emos** Burgos*
  'and they did it so, as we **will tell** later' (cited in Company 2006: 351)
- b. *todo **lo contar-emos***
  'we **will tell** everything' (cited in Bouzouita 2011: 116)
- c. ***Enprenar-te as** e avras fijo*
  'you **will get pregnant** and have a son' (cited in Bouzouita 2011: 113)
- d. *e **tornaré los** a sus logares*
  'and I **will return them** to their places' (cited in Bouzouita 2016: 270)

So far, little attention has been paid to the sfc with postverbal pronouns (1d), which appear in exactly the same grammatical environments as the afc (except for Bouzouita 2013, 2016; Eberenz 1991; Graham 2018). This can be interpreted as a progression in the grammaticalisation process, since these structures show complete univerbation of the two components of the original periphrasis and thus seem to be crucial for our understanding of the change in question. The reason for this lack of attention is the scarce documentation of this structural possibility in 13th-century Castilian, as it only gains frequency from the 15th century onwards in this variety (Eberenz 1991). Consequently, no large-scale corpus studies on the alternation between afc and sfc with postverbal pronouns have been carried out yet. This has also hidden the fact that these structures seem to be synchronically more frequent in the Eastern Ibero-Romance varieties (i.e. Navarro-Aragonese, Catalan, including Occitan) than in the more Western ones (Castilian, Asturian-Leonese, Galician-Portuguese), hinting at the possibility that the change might have started in the East and spread gradually towards the West of the Iberian Peninsula, as proposed initially by Bouzouita (2016) and worked out in more detail by Bouzouita & Sentí (in press [2022]) and Primerano & Bouzouita (subm.).

The objective of the present talk is twofold. One the one hand, it aims to examine whether there actually exist diatopic differences in the grammaticalisation of the fc by means of a broad corpus study comparing three Central Ibero-Romance varieties (Navarro-Aragonese, Castilian, Asturian-Leonese) in the 13th and 14th centuries. In other words, the hypothesis of an East-to-West spread of the change will be partially verified. Quantitative analyses will demonstrate that sfc with postverbal pronouns remain extremely infrequent in Castilian and Asturian-Leonese in that period, while the Navarro-Aragonese data show an increase in their frequency of use, thus corroborating the starting hypothesis.

Additionally, this talk will present both the advantages and shortcomings of corpus methods in the study of Ibero-Romance historical morphosyntax from a dialectological perspective. More concretely, it will provide a critical assessment of the large-scale databases available up to this day and address general issues, such as the reduced number of available sources, the varying philological reliability of extant texts, and the difficulty of composing comparable, representative, and dialectally contrastive historical corpora.

References

Bouzouita, M. (2011). Future constructions in medieval Spanish: Mesoclisis uncovered. In R. Kempson, E. Gregoromichelaki & C. Howes (Eds.), *The dynamics of lexical interfaces* (pp. 107-123). Stanford, UK: CSLI.

Bouzouita, M. (2013). La influencia latinizante en el uso del futuro en la traducción bíblica del códice Escorial I.i.6. In E. Casanova & C. Calvo (Eds.), *Actes del 26é Congrés de Lingüística i Filologia Romàniques* (pp. 353-364). Berlin: W. de Gruyter.

Bouzouita, M. (2016). La posposición pronominal con futuros y condicionales en el códice escurialense I.i.6: un examen de varias hipótesis morfosintácticas. In J. Kabatek (Ed.), *Lingüística de corpus y lingüística histórica iberorrománica* (pp. 271-301). Berlin: W. de Gruyter.

Bouzouita, M., & Sentí, A. (in press [2022]). La gramaticalización del futuro y el condicional en el iberorromance del siglo xiv a partir de traducciones bíblicas paralelas: el caso del castellano y el catalán antiguos. In A. Enrique-Arias (Ed.), *Traducción bíblica e historia de las lenguas iberorrománicas*. Berlin: W. De Gruyter.

Castillo Lluch, M. (2002). Distribución de las formas sintéticas y analíticas de futuro y condicional en español medieval. In M. T. Echenique & J. P. Sánchez (Eds.), *Actas del V Congreso Internacional de Historia de la Lengua Española* (pp. 541-550). Madrid: Gredos.

Company Company, C. (2006). Tiempos de formación romance ii: Los futuros y condicionales. In C. Company Company (Ed.), *Sintaxis histórica de la lengua española. Primera parte: La frase verbal. Vol. 1.* (pp. 349-422). Mexico: Fondo de Cultura Económica/UNAM.

Eberenz, R. (1991). Futuro analítico y futuro sintético en tres obras con rasgos coloquiales: El 'Corbacho', 'La Celestina' y 'La Lozana Andaluza'. In R. Lapesa (Ed.), *Homenaje a Hans Flasche: Festschrift zum 80. Geburtstag 25. November 1991* (pp. 499-508). Stuttgart: Franz Steiner Verlag.

Graham, L. (2018): An analysis of morphosyntactic variation in the Old Spanish future and conditional. *Journal of Historical Linguistics*, *8*, 192-229.

Primerano A., & Bouzouita, M. (submitted). La gramaticalización de los futuros y condicionales en el navarroaragonés de los siglos xiii y xiv: una comparación con el castellano medieval.

Participles and so-called synthetic compounds as attributive noun modifiers in English

Bas Aarts

University College London

English allows participial forms of verbs to modify nouns, as in the following example:

(i) The Rapids in 1834 was *a straggling village* whose 44 residents clustered mainly along the river on the east side of a single dirt path – the future Front Street. (*iWeb Corpus*)

Setting aside deverbal adjectives such as *interesting*, *satisfying*, etc., which can be preceded by intensifying adverbs such as *very* and *extremely*, there is no agreement in the linguistic literature about whether attributive V-*ing* premodifiers in noun phrases are verbs or adjectives. In this paper, after a brief discussion of earlier analyses (e.g. Roeper and Siegel 1978, Fabb 1984, Brekke 1988, Meltzer 2007/2010, Borer 2013, amongst others) I will discuss the evidence for regarding them as verbs, using corpus data. This includes the fact that V-*ing* modifiers have verbal roots and semantics; that they cannot occur as the complement of *seem*, *appear*, etc.; and the fact that they can take a full range of verbal dependents in so-called 'synthetic compounds', such as *cake-eating, beer-swilling* and *wall-straggling*. Corpus evidence shows that subjects are also possible as dependents in such compounds (Bauer and Renouf 2001; Bauer et al. 2013 and Lieber 2016), though many linguists have ruled this out ever since Roeper and Siegel (1978: 208) formulated their First Sister Principle ("All verbal *-ing* compounds are formed by the incorporation of a word in first sister position of the verb"). My research, using various corpora, including the *iWeb Corpus*, as databases, shows that it is even possible to have multiple dependents, as in (ii) and (iii):

(ii) But it is left to [$_{NP}$ *the then still living Robin Cook*] to enter the epitaph on the war on terror, of which Iraq is argued to be so necessary an ingredient. (*The Guardian*)

(iii) If you'd like to request aid for an upcoming non profit fundraising event from [$_{NP}$ *The Backyard Axe Throwing League*], please contact BATL London for eligibility requirements. (*iWeb Corpus*)

Example (ii) shows that two adjuncts are possible, whereas (iii) shows that an adjunct and a complement are possible. Examples like these have not been noticed in the literature before. The cumulative evidence argues strongly in favour of analysing attributive V-*ing* premodifiers as verbs.

References

Bauer, Laurie and Antoinette Renouf (2001) A corpus-based study of compounding in English. *Journal of English Linguistics* 29.2. 101-123.

Bauer, Laurie, Rochelle Lieber and Ingo Plag (2013)(eds.) *The Oxford reference guide to English morphology*. Oxford: Oxford University Press.

Borer, Hagit (2013) *Structuring sense: taking form*. Oxford University Press.

Brekke, Magnar (1988) The experiencer constraint. *Linguistic Inquiry* 19.2, 169-180.

Fabb, Nigel (1984) *Syntactic affixation*. Cambridge MA: MIT dissertation.

Lieber, Rochelle (2016) On the interplay of facts and theory: revisiting synthetic compounds in English. In: Daniel Siddiqi and Heidi Harley (2016)(eds.), *Morphological Metatheory*, Amsterdam/Philadelphia: John Benjamins Publishing Company. 513-536.

Meltzer, Aya (2007) The experiencer constraint revisited. In: N. Hilton, R .Arscott, K. Barden, A. Krishna, S. Shah and M. Zellen (2007)(eds.) *CamLing 2007*. Cambridge: Cambridge Institute of Language Research. 177-184.

Meltzer-Asscher, Aya (2010) Present participles: categorial classification and derivation. *Lingua* 120, 221-2239.

Roeper, Thomas and Muffy Siegel (1978). A lexical transformation for verbal compounds, *Linguistic Inquiry* 9. 199–60.

Finnish partitive e-NP constructions in web corpora

Rodolfo Basile

In Finnish, canonical subjects in nominative may alternate with partitive ones, which change clause-level semantics on many levels. They usually introduce a new referent and are featured in existential constructions (Hakanen 1972). Existential subjects have also been defined as *e-NP*s (existential NP, Huumo & Helasvuo 2015), in that they are not subjects in a canonical way: in Finnish, these include clause final nominative NPs or partitive NPs both in clause-initial and clause-final position. Existential constructions are also characterized by the absence of verb agreement. Unlike in most other languages (Creissels 2014, 2019), Finnish existential constructions can be characterized by verbs different than 'to be', which are also called *lexical existentials* (Larjavaara 2019), in that they add lexical information on how the referent is existing or being located (e.g. walking, swimming, etc.).

This paper investigates the frequency of usage of lexical existentials with partitive-marked e-NPs in Finnish web corpora, namely Suomi24:2017, a collection of forum posts in the year 2017, and proposes an alternative sampling method to apply to collostructional analysis (Stefanowitsch & Gries 2003, Gries & Stefanowitsch 2004). The main research question is: which verbs occur the most with partitive e-NPs? The reason why we need alternative sampling is the complexity of this construction, whose partitive e-NP is often not recognized by the machine, which mistakes it for other partitive-marked elements (e.g. time adverbials).

The new sampling is based on what I call *Expected Sample size*, which allows for observation of large-scale effects while having a rather small sample. It goes the following way: a small sample (e.g. 1000 sentences) for each verb is manually checked, obtaining an observed occurrence frequency of the construction studied. This frequency is then compared to the overall occurrence of the verb in the whole corpus and the occurrence of the same verb in a corpus search constrained by grammatical features that trigger the appearance of the partitive e-NP (e.g. absence of verb agreement), obtaining a new, smaller sample which is proportional to all these occurrence frequencies. When collostructional analysis is carried out with the new numbers, the new sampling eliminates the bias originally provided by the first sampling of 1000 sentences per verb, which ignored their overall occurrence in the corpus. The results show an estimate of how large the new sample is supposed to be, in order to observe the frequency obtained in the original biased sample of 1000 sentences.

A statistically ordered set of p-values describes the association between verbs (collexemes) and the partitive e-NP collostruction better than the first sampling, providing an estimate of their behavior on the whole-corpus level. It points toward interesting qualitative findings, which help classifying the different verbs used in the Finnish partitive e-NP construction.

This paper contributes to the development of a new way of sampling in corpus linguistics, which can be used not only with the construction I presented, but potentially also with other constructions that may face quantitative researchers with the problem of having to manually polish excessive amounts of data.

References

Creissels, Denis. 2014. Existential predication in typological perspective. Ms. Université Lyon.

Creissels, Denis. 2019. Inverse-Locational Predication in Typological Perspective. *Italian Journal of Linguistics* 31 (2): 38–106.

Gries, Stefan Th., and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on `alternations'. *International Journal of Corpus Linguistics* 9 (1): 97–129. https://doi.org/10.1075/ijcl.9.1.06gri.

Hakanen, Aimo. 1972. Normaalilause ja eksistentiaalilause. *Sananjalka* 14 (1): 36–76. https://doi.org/10.30673/sja.86366.

Huumo, Tuomas, and Marja-Liisa Helasvuo. 2015. On the subject of subject in Finnish. *Subjects in Constructions–Canonical and Non-Canonical*, 13–41.

Larjavaara, Matti. 2019. *Partitiivin Valinta*. Suomalaisen Kirjallisuuden Seura.

Stefanowitsch, Anatol, and Stefan Th Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8 (2): 209–43.

No hay boda sin ramo de novia: Capturing the creative potential of snowclones in Spanish

Pedro Ivorra Ordines

Key words: snowclones, lexically filled constructions, creativity, productivity

Construction-based studies to the analysis of phraseological phenomena – with a special attention to idioms with empty slots (Dobrovol'skij 2016: 81) – has covered a great variety of approaches ranging from the study of idiom modification (Mellado Blanco 2020) to the analysis of the interaction of constructional idioms within the constructicon (Mollica & Stumpf 2021) or of the productivity and creativity of comparative constructional idioms (Ivorra Ordines 2022). But it is not only in the field of syntactic frames that CxG offers promising results. In line with Mellado Blanco (2022), I advocate for the study of substantive or lexically filled constructions (e.g., idioms, proverbs), reinforcing the idea that phraseological phenomena are flexible units endowed with a great creative potential (see Traugott & Trousdale 2014 on the process of constructionalization).

Against this background, this study aims at scrutinizing snowclones, defined by Pullum (2003) as "reusable customizable easily-recognized twisted variant of a familiar but not literary quoted or misquoted saying". Drawing on data extracted from the esTenTen18 corpus (Sketch Engine), our investigation highlights the importance of analogical extensions in the extension of a grammatical pattern based on only one model item (Barðdal 2008), as in the examples below:

(1) Busquen e indaguen cuanto quieran, **no hay boda sin ramo de novia**. (SkE 1036325935)
'Search and investigate as much as you want, there is no wedding without a bridal bouquet.'
[*No hay* X *sin* Y] (Y is essential for X to take place)

(2) Las más de 50 entidades acreedoras, con las que refinanció más de 5.000 millones de euros in extremis en 2010 han dicho basta […], **más vale piso en mano que dinero volando**. (SkE 6077667071)
'The more than 50 creditor institutions, with which more than 5,000 million euros were refinanced in extremis in 2010, have said enough […], better a flat in hand than money flying.'
[*Más vale* X *que* Y *volando*] (applied to those who leave uncertain situations, waiting for better but uncertain things)

(3) **Limpieza llama a limpieza**, si estamos ante una ciudad limpia, […] probable que nos lo pensemos dos veces de no agacharnos a recoger el papel que se nos ha caído. (SkE 425840396)
'Cleanliness calls for cleanliness – if we are dealing with a clean city, […] we are likely two to think twice about not bending down to pick up the paper we have dropped.'
[X *llama a* X] (something positive triggers something positive in the inmediate future)

It is argued that formula of this type can be best studied as a type of constructional idiom placed in the middle of the lexicon-grammar continuum "with varying degrees and kinds of freedom as to what can fill the slots in the pattern and with varying degrees of semantic and pragmatic specialization" (Zwicky 2006). In addition, our data not only confirms speakers' ability in chopping snowclones up to individual lexical items but also shows that creative formations by analogical extensions follow certain patterns of creativity – i.e., the regularity in the irregularity can be captured.

References

Barðdal, J. 2008. *Productivity: Evidence from Case and Argument Structure in Icelandic*. Benjamins.

Dobrovol'skij, D. 2016. Fraseología y Gramática de Construcciones. *Language Design* 18, 71-106.

Ivorra Ordines, P. 2022. Comparative constructional idioms A corpus-based study of the creativity of the [*más feo que* X] construction. In C. Mellado Blanco (ed.), *Productive Patterns in Phraseology and Construction Grammar. A Multilingual Approach*. De Gruyter, 29-52.

Mellado Blanco, C. 2020. La desautomatización desde el prisma de la Gramática de Construcciones: un nuevo paradigma de la variabilidad fraseológica. *Nasledje* 45, 17-34.

Mellado Blanco, C. 2022. Phraseology, Patterns and Construction Grammar: An introduction. In C. Mellado Blanco (ed.), *Productive Patterns in Phraseology and Construction Grammar. A Multilingual Approach*. De Gruyter, 1-25.

Mollica, F. & Stumpf, S. 2021. Families of constructions in German A corpus-based study of constructional phrasemes with the pattern [X_{NP} attribute]. In C. Mellado Blanco (ed.), *Productive Patterns in Phraseology and Construction Grammar. A Multilingual Approach*. De Gruyter, 79-106.

Pullum, G. 2003. Phrases for lazy writers in kit form. *Language Loc*. http://itre.cis.upenn.edu/~myl/languagelog/archives/000061.html

Traugott, E. C. & Trousdale, G. 2014. Contentful constructionalization. *Journal of Historical Linguistics* 4/2, 256-283.

Zwicky, A. 2006. Snowclone mountain? *Language Log*. http://itre.cis.upenn.edu/~myl/languagelog/archives/002924.html

On *incoming passers-by* and *bystanding lookers*-on: A quantitative approach to variable particle placement in English particle verbs

Anke Lensch & Jelke Bloem

University of Koblenz-Landau & University of Amsterdam

Some highly frequent Present-day English *-er* nominalizations of particle verbs appear to violate Goldberg's (1995: 67) Principle of No Synonymy as they are attested in two different forms that appear to have equivalent descriptive meaning, compare (1) and (2):

1. others were *passers-by* out for a drive or a stroll (*The Guardian* 2002)
2. a *bypasser* randomly kicked him in the head. *The Guardian* 1992)

Both *passerby* and *bypasser* denote a 'someone walking past' and thus they constitute a doublet. When comparing the bases of the two derivatives, we observe that the placement of *by* in relation to *pass* can encompass a small difference in meaning, compare (3) and (4).

3. I just happened to be *passing by* the biscuit tin (*The Guardian* 2000)
   'moving past another object or event'
4. all the while *bypassing* the time-consuming, tedious and costly business of divorce.
   'the voluntary passing or avoidance of something unpleasant' (*The Guardian* 2001)

According to Los et al. (2012: 139f.), variation in the placement of the particle in relation to the verbal base was much more frequent until the Middle English period and has since all but fallen out of use. Today, in most particle verbs the particle cannot be placed in front of the verb, consider *chat up* vs **upchat* (cf. Rodríguez-Puente 2019:120f.). While the bases of some verb particle combinations (e.g., s*tand by* vs. *bystand*) are hardly ever attested in Present-day English data, the alternation in particle placement can still be conserved in their derivatives, consider (5) to (8).

5. a bit of recreational *bystanding* (*The Guardian* 2005)
6. He did not try to munch a *bystander's* apple (*The Guardian* 1994)
7. He has the presence of a *stander-by* (*The Guardian* 2003)

By undertaking a large-scale quantitative and qualitative corpus analysis, we identify the factors that have allowed a small subset of particle verbs and some of their derivatives to preserve variable particle placement. We gauge the frequencies of particles, prepositions, verbal bases, and particle verb derivatives in British English and American English corpora comprised of data from the 17th century to the present day (Chadwyck-Healey Collection, Mainz Corpus Collection) in order to establish to what extent frequency could influence the preservation of particle placement alternation. Using this data, we find particle verb derivatives with particle alternation, and perform a collostructional analysis (Stefanowitsch & Gries, 2003) to reveal the degree to which particle verb types are associated with one of the two patterns, as well as to reveal strongly lexicalized combinations.

Close qualitative analysis of the data confirms that the alternation in particle placement is preserved in those particle verb derivatives where the particular combination of the particle and the verbal base still allow for literal readings, whereas more lexicalized combinations do not allow for alternation. Moreover, the data reveals that particles that are attested with a wider range of verbal bases are less likely to allow for alternating particle placement. These results show that Goldberg's Principle of No Synonymy needs to be understood in a broader sense in that various factors, such as frequency and the degree of semantic transparency can have an effect on long-term variation.

References

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

Los, Bettelou; Blom, Corrien; Booij, Geert; Elenbaas, Marian and van Kemenade, Ans (2012) *Morphosyntactic Change: A Comparative Study of Particles and Prefixes*. Cambridge: Cambridge University Press.

Rodríguez-Puente, Paula (2019) *The English Phrasal Verb, 1650-Present. History, Stylistic Drifts, and Lexicalization.* Cambridge, Cambridge University Press.

Stefanowitsch, Anatol, & Gries, Stefan (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209-243.

A corpus study of clitic placement in Old Catalan

Aina Torres-Latorre & Andreu Sentí

Weak pronouns or clitic pronouns are easily identified by going with a verb. In Catalan, the distribution of these pronouns around the verb has evolved since the first stages of the language. In fact, the contemporary Catalan clitic position is characterised by morphological restrictions: the so-called proclitic pronouns appear before finite verbs except imperative while the enclitic ones appear after non-finite verbs and imperative (Bonet, 2002: 937). However, in the medieval language, according to Batllori *et al.* (2005), the pronouns were distributed according to syntactic restrictions, that is, their position was determined by the syntactical context. The distribution of the pronouns following syntactic restrictions has also been observed in other Romance languages, such as Spanish (see Granberg, 1988; Bouzouita, 2008).

The present work aims to study this distribution diachronically using a selected corpus. We will first try to confirm the syntactic hypothesis from the observation of empirical data. In particular, we have selected the chronicle *Llibre dels fets*, also known as *La crònica de Jaume I*, because of its cultural relevance, its narrative nature and its chronological location (the book was written during the second half of the 13th century and the oldest manuscripts that have been conserved date from the first half of the 14th century). In *Llibre dels fets*, the occurrences of weak pronouns with every verbal tense have been analysed and classified keeping in mind the different syntactical contexts.

The second main aim of this work is the description of two verbal tenses: future and conditional. The romance future and conditional tenses are the result of the grammaticalization of the Latin periphrasis cantāre habeō. In some medieval Romance languages, such as Occitan, Catalan, Aragonese, Spanish and Portuguese, these tenses had two types of forms: the synthetic forms, with proclitic (*el faré*) or enclitic pronouns (*faré-lo*), and the analytical forms (*fer-lo he*), characterised by the presence of the weak pronoun between the infinitive and the auxiliar. Analytical forms do no present univerbation (cf. Lehmann, 1985) and are thus less grammaticalized than the synthetic forms.

To study these two tenses, a diachronic corpus has been made up of several texts from the 13th century to the 15th century. The purpose of this diachronic and quantitative view is to observe the evolution of the grammaticalization of future and conditional. Analytical forms and synthetic forms with enclitic pronouns coexisted in the same syntactical contexts and, therefore, the advancement of the grammaticalization can be observed in these contexts. Moreover, the results are compared with Spanish (cf. Bouzouita, 2016a) and Navarroaragonese (cf. Primerano, 2019) because it has been observed that the grammaticalization is more advanced in the eastern languages of the Iberian Peninsula than in western ones (Bouzouita, 2016b; Bouzouita & Sentí, in press), so we expect an advance in Catalan in relation to these two other peninsular languages.

References

Batllori, Montserrat / Iglésias, Narcís / Martins, Ana Maria (2005): «Sintaxi dels clítics pronominals en català medieval», *Caplletra* 38, p. 137-177.

Bonet, Eulàlia (2002), «Cliticització», in Solà, Joan; Lloret, Maria Rosa; Mascaró, Joan; Pérez Saldanya, Manuel, *Gramàtica del català contemporani*, vol. 1, p. 933-989.

Bouzouita, Miriam (2008): *The Diachronic Development of Spanish Clitic Placement*, Londres: King's College London, PhD thesis.

— (2016): «La posposición pronominal con futuros y condicionales en el códice escurialense I.i.6: un examen de varias hipótesis morfosintácticas», in: Kabatek, Johannes: *Lingüística de corpus y lingüística iberorománica*, Berlín: De Gruyter.

— (2016b): «La posposición pronominal con futuros y condicionales en el castellano medieval: ¿un caso de contacto de lenguas?», presented in the seminar *El orden de palabras en las lenguas iberorrománicas medievales*, 29th of July of 2016, University of Girona.

Bouzouita, Miriam / Sentí, Andreu (in press): «La gramaticalización del futuro y el condicional en el iberorromance del siglo xiv a partir de traducciones bíblicas paralelas: el caso del castellano y el catalán antigues», in: Enrique-Arias, Andrés (ed.): *Traducción bíblica e historia de las lenguas iberorrománicas*, Berlín: De Gruyter. Beihefte zur Zeitschrift für romanische Philologie.

Granberg, Robert Arthur (1988): *Object Pronoun Position in Medieval and Early Modern Spanish*, Los Angeles: University of California.

Lehmann, Christian (1985): «Grammaticalization: Synchronic variation and diachronic change», *Lingua e stile*, vol. 20, p. 303-318.

Primerano, Antoine (2019): *La gramaticalización de los futuros y condicionales en el navarroaragonés de los siglos xiii y xiv: análisis morfosintáctico-pragmático*, Ghent University, master thesis.

Panel on "Syntactic Productivity": introduction

Language Productivity@Work Consortium

When speakers produce or interpret language structures, they rely on a structured inventory of grammatical rules or patterns. Some of these are highly productive: they have a broad domain of application and are readily available to coin new expressions.

This phenomenon has long been observed in morphology. For instance, speakers of Dutch can readily apply the morphological rule Verb+*baar* to create new adjectives meaning 'that can be Verb-ed', such as in *een twitter·baar stuk tekst* 'a twitterable text chunk'. By contrast, other rules such as Verb+*(e)lijk*, as in *ondraag·lijk* 'unbearable' are not productive (Booij 2002). As a consequence, **twitter·lijk* is completely out. But also syntactic rules and constructions can be productive to varying degrees, since they can be applied to a range of words (which fill one or more slots), including neologisms. For instance, the transitive construction X+Verb+Y (e.g. *to eat an apple*), which accepts many verbs, is far more productive than the nominative-genitive construction in Modern German, which is restricted to only a handful of verbs (Barðdal 2008: 150), one of which is *gedenken* (*Wir gedenken der Opfer* 'We remember the victims').

Productivity is an abstract property of linguistic structures that forms part of the implicit knowledge speakers have about a language. Not only does it play a fundamental role in synchronic language description, it is also a crucial concept in language change (e.g. Hilpert 2013, Traugott & Trousdale 2013, Perek 2016) and language acquisition (Tomasello 2003; Hartsuiker & Bernolet 2017). Up until the present day, however, the phenomenon of productivity is poorly understood, especially regarding syntactic constructions.

This panel is hosted by the Language Productivity@Work Consortium (Ghent University; https://www.languageproductivity.ugent.be/).

In the introduction, its organizers will address the following topics:

- Recall the key concepts in productivity research on syntactic constructions, including the most important productivity measures, such as type/token ratio, as discussed in the work of Baayen (1992, 2001, 2009) and Zeldes (2012), for instance;
- outline some research challenges in productivity research and show how the *Language productivity@work* consortium wants to tackle some of these interdisciplinary issues, uniting corpus linguistics (cf. Goldberg 2019), psycholinguistic experiments and sociolinguistic surveys;
- introduce the programme

References

Baayen, R. H. 1992. Quantitative aspects of morphological productivity. In *Yearbook of Morphology 1991*, ed. Booij, G. E., & J. Marle. Dordrecht: Kluwer. 109–149.

Baayen, R. H. 2001. *Word frequency distributions*. (Text, Speech and Language Technologies 18). Dordrecht: Kluwer.

Baayen, R. H. 2009. Corpus linguistics in morphology: morphological productivity. In *Corpus Linguistics. An international handbook*, ed. Lüdeling, A.& M. Kyto. Berlin: De Gruyter. 900–919.

Barðdal, J. 2008. *Productivity: Evidence from Case and Argument Structure in Icelandic*. (Constructional Approaches to Language 8). Amsterdam: Benjamins.

Booij, G. 2002. *The Morphology of Dutch*. Oxford: Oxford UP.

Firth, J.R. 1957. *A synopsis of linguistic theory 1930-1955*. *Studies in Linguistic Analysis*. Oxford : Philological Society, 1-32.

Goldberg, A. 2019. *Explain me this. Creativity, Competition, and the Partial Productivity of Constructions.* Princeton: Princeton University Press.

Hilpert, M. 2013. *Constructional Change in English .Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge UP.

Perek, F. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics* 54(1): 149–188.

Tomasello, M. 2003. *Constructing A Language: A Usage Based Theory of Language Acquisition*. Harvard: UP.

Traugott, E. & G. Trousdale. 2013. *Constructionalization and constructional change*. Oxford: Oxford UP.

Zeldes, A. 2012. Productivity *in Argument Selection: From Morphology to Syntax.* Berlin: De Gruyter.

Measuring productivity through language comparison

Bert Le Bruyn, Martín Fuchs, Martijn van der Klis, Jianan Liu, Chou Mo & Henriëtte de Swart

We identify a semantic dimension of productivity that frequency-based measures cannot capture (§1) and propose a translation-corpus-based operationalization of this dimension (§2).

§1. Productivity in syntax/semantics is not only about the frequency of a construction but also about its felicity in semantic context types (SCTs):

(1) [...] zodat ik de huiskamertafel naar mijn eigen kamer **heb moeten** brengen. [DUTCH]
'and I**'ve had** to move the living room table to my own room.'

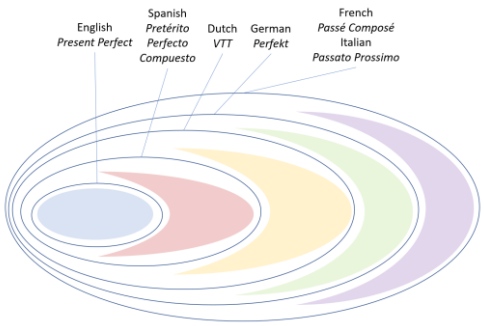
(2) Er had zich een stapel vrachtbrieven op mijn tafel opgehoopt, [DUTCH]
die ik allemaal **moest** afdoen.
'There was a pile of shipping manifests on my desk that I **had** to work through.'
(Morriën, translating *L'Étranger*)

(1) represents a resultative SCT in which a past obligation leads to a result with current relevance. The SCT represented by (2) is different: there's no relevant result at the moment of writing and the narrator just refers to an obligation at some past-time point. With the SCTs thus defined, *moeten* obligatorily appears in the perfect (VTT) in (1) and in the past (OVT) in (2).

Had Dutch had one tense marker for past reference, *moeten* in (1) and (2) would have appeared with the same tense marker. The difference in tense marker shows that the OVT/VTT limit each other's productivity in these SCTs. Frequency-based measures cannot directly capture this SCT-dependency: without coding for SCTs, it is impossible to tease apart the frequencies of constructions and those of the different SCTs they appear in.

§2. The role of semantics in productivity has been noted before: [1] argues that a verb's lexical semantics influences the productivity of morpho-syntactic processes and proposes word embeddings as a corpus-based operationalization of lexical semantics. However, word embeddings do not suffice to operationalize SCTs: (1) and (2) represent different SCTs but the (lexical) verb in them is the same. To remedy, we build on insights from [2] and resort to translation corpora to build a counterpart of word embeddings for constructions.

In [2], we studied the *have*-perfect (present tense of *have/be* + past participle) on the basis of *L'Étranger* and its translations. We established that the contexts the English perfect appears in are a subset of the contexts the Spanish perfect appears in and so forth for Dutch/German/French/Italian:

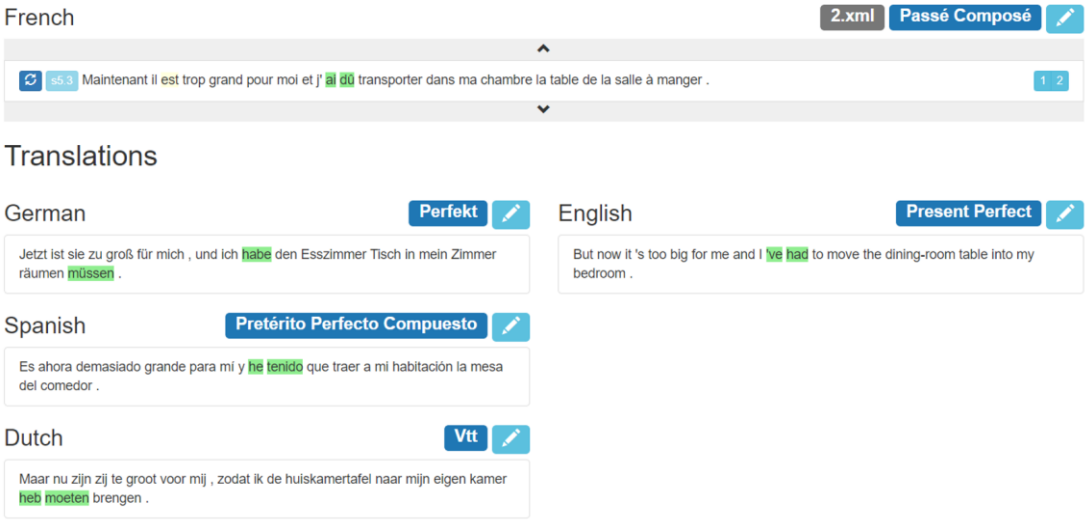


In [2], we further showed that the contexts that appear in the English perfect as well as each set of contexts that is added from one language to the next – the colored sets – can be characterized in a

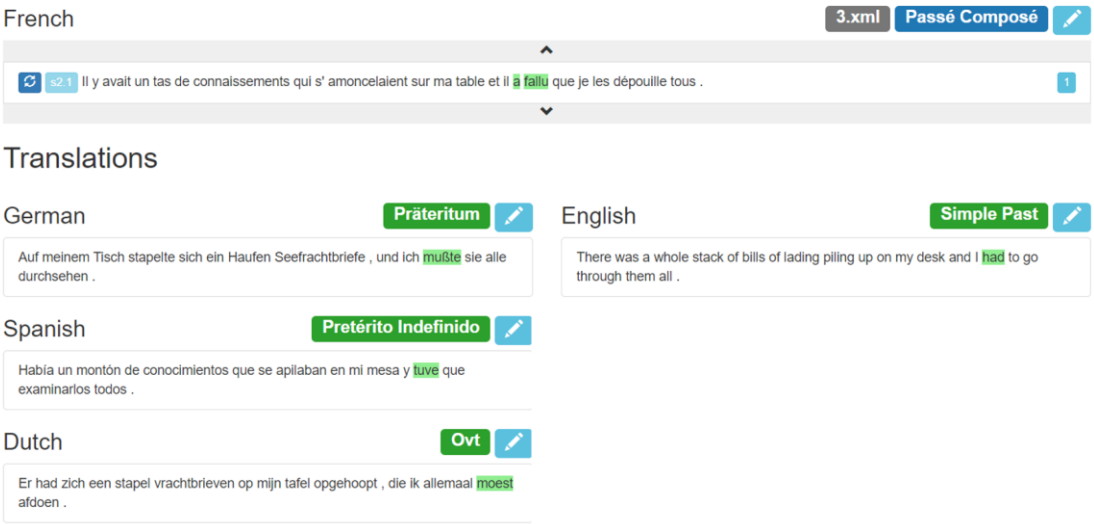
semantically homogenous way and can insightfully be related to restrictions on the *have*-perfect noted in the literature.

We argue that the colored sets are a corpus-based operationalization of SCTs. In our example, the blue set brings together resultative contexts like (1) and the purple set brings together 'state in story' contexts like (2):

(1') Source



(2') Source



Each colored set corresponds to a specific configuration of language-specific tense markers. In our example, the blue set corresponds to <*Passé Composé*, *Perfekt*, *PPC*, *VTT*, *Present Perfect*>. Based on our insights in [2], these tuples suffice to operationalize SCTs, and we thus propose them as the construction counterpart of the vectors that are standard in corpus-based operationalizations of lexical semantics.

We conclude that the move from monolingual to translation corpora allows us to operationalize SCTs without manually annotating for meaning. This, in turn, allows us to cross frequencies of constructions with those of the SCTs they appear in, enriching classical productivity measures (i.a., [3]) and capturing the impact of SCTs on productivity.

References

[1] Suttle, L. & A. Goldberg. 2011. The partial productivity of constructions as induction, *Linguistics* 49/6: 1237-1269.

[2] Van der Klis, M., Le Bruyn, B., & De Swart, H. (2022). A multilingual corpus study of the competition between past and perfect in narrative discourse. *Journal of Linguistics*, 58(2), 423-457.

[3] Baayen, R. H. 2009. Corpus linguistics in morphology: morphological productivity. In *Corpus Linguistics. An international handbook*, ed. Lüdeling, A.& M. Kyto. Berlin: De Gruyter. 900–919.

Dimensions of modality: lexical modals in the brexit political discourse

Aroa Orrequia-Barea & Encarnación Almazán-Ruiz

Keywords: modality, lexical modals, corpus linguistics, political discourse, Brexit

The semantic category of modality is mainly associated with the speaker's attitude towards the situation or facts expressed in the clause. Modality can be conveyed through different linguistic procedures in the discourse: lexical, grammatical and prosodic. Apart from including a (semi)auxiliary modal verb in the verb phrase, there are other linguistic devices that the speaker can use to express modality in the utterance. According to Huddleston and Pullum (2002, p. 173), "lexical modals" are other word classes (i.e., adjectives, adverbs, nouns or lexical verbs) that can also convey the same meaning as (semi)auxiliary modal verbs. Considering the meaning expressed in the utterance, a division can be established between two main types of modality, epistemic and deontic (Huddleston, 1988, p. 78-80).

The meaning potential of language becomes apparent based on the lexical choices that the speaker makes. This potential is especially relevant in political discourse, where word choice can be crucial in influencing voters' decisions. Due to Brexit, there were eloquent political clashes between two of the most important candidates of the UK at that time: Boris Johnson and Jeremy Corbyn.

This paper compares and analyses the lexical modals used in their political speeches during the last months of the Brexit process, focusing on whether lexical choice reveals politicians' real perspectives. Our primary research questions will determine which *lexical modals* each politician uses and which semantic implication(s) they convey. This study is descriptive-interpretative using the methodology of Corpus-assisted Discourse Analysis (Baker, 2020). We aim to unveil the connections between the use of language, particularly the expression of modality through *lexical modals*, and the political context in which it occurs, namely the Brexit process (Paltridge, p. 186). Thus, we will explore whether the politicians' involvement and attitudes are reflected in discourse.

The corpus consists of 51,491 words composed from a collection of speeches delivered by Boris Johnson, Prime Minister, and Jeremy Corbyn, leader of the opposition at that moment. The speeches were delivered from the 24th of June, 2019, when Boris Johnson was appointed Prime Minister, until the 31st of January, 2020, the so-called Brexit day. Using as a base Huddleston and Pullum's classification of lexical modals, both corpora have been manually annotated by the authors, considering the lexical modals used and their meaning. Afterwards, the most frequent meaning of the lexical modals has been statistically determined. Finally, the Sketch Engine corpus tool has been used to explore our own corpus (Kilgraff et al. 2010). The function 'concordances' has been really useful to study the lexical modals in context and address their semantic implications.

As expected, results indicate that the reading of epistemic modality is more frequent among lexical modals than that of deontic. Thus, some of the most frequent lexical modals in the corpora are *possible, sure* or *perhaps.* The comparison of lexical modals between both politicians reveals interesting differences in their perspectives towards Brexit. Whereas Johnson presented Brexit as a plausible process, Corbyn presented it as a remote and tentative possibility. In conclusion, the prominence of the epistemic modality is presented as a linguistic resource that politicians use not to manifest their commitment, at least when talking about Brexit.

References

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum.

Baker, P. (2020). Corpus-assisted discourse analysis. In C. Hart (Ed.), *Researching Discourse: A Student Guide*, 124-142. Routledge. https://doi.org/10.4324/9780367815042-8

Chilton, P. (2004). *Analysing Political Discourse: Theory and Practice*. Routledge.

Huddleston, R. D. (1988). *English Grammar: An Outline*. Cambridge University Press.

Huddleston, R. D. & Pullum G. K. (2002) *The Cambridge Grammar of the English Language*. Cambridge University Press.

Kilgarriff, A., Baisa, V., Bušta, J., Jakub.ček, M., Kov.ř, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography* 1(1), 7-36. https://doi.org/10.1007/s40607-014-0009-9

Palmer, F. R. (1988). *Mood and Modality.* Cambridge University Press.

Paltridge, B. (2012). *Discourse Analysis: An Introduction.* Bloomsbury Publishing.

Zhang, J. (2019). A Semantic Approach to the English Modality. *Journal of Language Teaching and Research*, vol. 10, no. 4, 879–85. https://doi:10.17507/jltr.1004.28.

Corpus-based retrieval and the function of inversion in discourse

Heidrun Dorgeloh

Heinrich Heine Universität Düsseldorf, dorgeloh@hhu.de

Subject-verb inversion in English is both commonly described as an information-structuring construction and a topic-management device in discourse (Birner 1994; Dorgeloh 1997; Ward, Birner & Huddleston 2002; Kreyer 2006; Prado-Alonso 2011; Dorgeloh & Wanner forthc.). However, the corpus-based retrieval of such a marked constituent order, exemplified in (1), is far from trivial. For this reason, it is almost impossible to treat the occurrence of inversion in discourse neither as being subject to text-linguistic variation (i.e., studying rates of occurrence within different types of discourse or registers) nor as a proper case of syntactic variation (i.e., contrasting all inverted as opposed to non-inverted sentences in a given corpus) (Biber 2012). The paper therefore raises the question which corpus-based procedure can be reasonably applied for an analysis of the function of inversion with large-scale, usage-based data.

(1) Given the vagaries of the presidential selection process and President Bush's recent drop in the polls, those eager to see Bush's stamp on the high court are increasingly anxious. *Equally nervous are the president's liberal foes*, who may be anticipating their own worst-case scenario for 2005 […]. (COCA Corpus, *News*, 2003)

The paper will show that inversions can be retrieved sufficiently well using some lexical shortcuts, i.e. by retrieving a set of inversions from a corpus based on selected lexemes: The so-called 'locative' type of inversion typically occurs with locative and directional prepositions, while 'non-locative' inversion, notably AdjP-inversion, like in (1), is known to follow a limited set of subjective or connective adverbs (e.g., *totally*, *also, more/most*), also showing a clear bias for a limited group of adjectives (e.g., *great*, *strong*, *important*) (Dorgeloh & Kunter 2015). Using data from a procedure which applies lexical shortcuts based on these insights, the paper will present an analysis of inversion with respect to its function of marking an upcoming topic in discourse. Applying an operationalization of topic persistence that has previously been applied to topicalization and left-dislocation (Gregory & Michaelis 2001), I will discuss outcomes of a study based on a set of 530 inversions retrieved from the COCA Corpus (Davies 2008 -). The attestations were analyzed for the recurrence of the NP referent (or a referent related to it) within a fixed amount of subsequent utterances. The analysis provides evidence for a discourse function of inversion related to topicality and highlights a solution for dealing with discourse-functional categories, such as topic or information structure, in a corpus.

References

Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37.

Birner, Betty. 1994. "Information status and word order: An analysis of English inversion." *Language* 70(2), 233–259.

Davies, Mark. 2008-. The Corpus of Contemporary American English (COCA): One billion words, 1990–2009. www.english-corpora.org/coca/.

Dorgeloh, Heidrun & Anja Wanner. Forthc. *Discourse Syntax: English Grammar Beyond the Sentence*. Cambridge: Cambridge University Press.

Dorgeloh, Heidrun & Gero Kunter. 2015. “Modelling adjective phrase inversion as an instance of functional specialization in non-locative inversion.” In: Sanchez-Stockhammer, Christina (ed.), *Building bridges into the future: Can we predict linguistic change*? Special issue of VARIENG.

Dorgeloh, Heidrun. 1997. *Inversion in Modern English: Form and Function*. Amsterdam/Philadelpia: Benjamins.

Gregory, Michelle L. and Laura A. Michaelis. 2001. “Topicalization and left-dislocation: a functional opposition revisited.” *Journal of Pragmatics*, 33(11), 1665–706.

Kreyer, Rolf. 2006. *Inversion in Modern Written English: Syntactic Complexity, Information Status and the Creative Writer*. Tübingen: Narr.

Prado-Alonso, Carlos. 2011. “Text structuring in written English: The role of inversion”. *English Studies* 92(4), 449–463.

Ward, Gregory, Betty Birner & Rodney D. Huddleston. 2002. “Information packaging”. *The Cambridge Grammar of the English Language*, ed. by Rodney D. Huddleston & Geoffrey K. Pullum, 1363–1449. Cambridge: Cambridge University Press.

Modification in light verb constructions: a corpus study in Germanic and Romance languages

Georgina Alvarez-Morera

Universitat Rovira i Virgili

Introduction

Light verb constructions (LVCs) are structures with a verb followed by a non-verbal element (NVE), usually a determiner phrase, in which the meaning is derived from the noun (i.e. *take a shower*). Although traditional approaches to light verbs dismiss their contribution to the whole construction because their meaning is faded (Jespersen 1954, Kearns 1988), more recent studies maintain that light verbs are not completely semantically empty since they do contribute to the semantics of the LVC (Butt 2010).

Moreover, most LVCs have a correspondent full verb with the same meaning. A question that arises is if speakers are conveying the same with the synthetic verb (*to answer*) and the LVC (*to give an answer*). Since Wierzbicka (1982), it has been proposed that LVCs have the aspectual function of being the telic counterpart *(have a thought)* of atelic verbs (*to think*). However, Bonial & Pollard (2020) show that this aspectual function is not the main motivation for the choice of LVCs over synthetic verbs: their corpus study reveals that the modification possibilities of the nominal element is the main factor.

In fact, there is a consensus that LVCs are frequent cross-linguistically because the noun can take flexible modification (Leech et al. 2009, Huddleston & Pullum 2002, Bonial & Pollard 2020, among others), but most studies have focused only on Germanic languages (Levin & Ström Herold 2015). This corpus-based study consists of a contrastive investigation on LVCs in Germanic languages (English and German) and Romance languages (Catalan and Spanish) with focus on the modification of the NVE.

Corpus & methodology

From the basic repertoire of light verbs that is shared cross-linguistically (Butt 2010), our proposal analyses the verb *give.* The study is based on a random sample of 9.000 tokens of LVCs from online annotated corpora for the four languages: English, COCA; German, DWDS; Catalan, CTILC; Spanish, CORPES XXI. The sample was obtained by searching for the collocates of the LV and the NVE. After manually excluding instances of non-light structures, for every LVC a sample of up to 200 occurrences were analysed. All examples were classified according to the following grammatical factors: (i) determination, (ii) modification and (iii) number of the NVE.

Selected results

The study shows that the majority of NVE in all four languages are introduced by a determiner. However, results related to adjectival modification of LVCs (which is the most frequent kind of modification, Levin & Ström Herold 2015) show that ca. 50% of English *give*-LVCs are modified, which contrasts with ca. 30% in German *geben*-LVCs. For Romance languages, results are more consistent: ca. 25% of Catalan *donar*-LVCs and ca. 20% of Spanish *dar*-LVCs are modified by an adjective.

Our results show that adjectival modification of LVCs is not very frequent with *give*-LVCs. Thus, modification cannot be considered a reason for the high frequency of use of LVCs cross-linguistically.

Implications

The results confirm that the majority of LVCs in all four languages appear unmodified, even if they can have flexible modification patterns. Thus, the reason for their frequency must be found elsewhere. These results are also compatible with the idea that LVCs and their verbal counterpart are not interchangeable in all contexts (Sanromán Vilas 2009). Further research should focus on the differences between LVCs and their synthetic counterparts on the basis of corpora.

References

Bonial, C. & K. A. Pollard. 2020. Choosing an event description: What a PropBank study reveals about the contrast between light verb constructions and counterpart synthetic verbs. *Journal of Linguistics*, 1–24.

Butt, M. 2010. The light verb jungle: Still hacking away. In M. Amberber, B. Baker, M. Harvey (eds.). *Complex Predicates: Cross-linguistic Perspectives on Event Structure*, 48–78. Cambridge: Cambridge University Press.

Huddleston, Ro. & G. K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: CUP.

Jespersen, O. 1954. *A Modern English Grammar*, vol. 4. New York: Barnes and Noble.

Kearns, K. 1988. *Light Verbs in English*. Ms., MIT.

Leech, G., M. Hundt, C. Mair & N. Smith. 2009. *Change in Contemporary English. A Grammatical Study*. Cambridge: Cambridge University Press.

Levin, M., Ström Herold, J. 2015. Give and Take : A contrastive study of light verb constructions in English, German and Swedish. In S. Oksefjell Ebeling and H. Hasselgård (eds.). *Cross-Linguistic Perspectives on Verb Constructions*. Cambridge Scholars Publishing, 144-168.

Sanromán Vilas, B. 2009. Diferencias semánticas entre construcciones con verbo de apoyo y sus correlatos verbales simples. *ELUA,* 23, 289-314.

Wierzbicka, A. 1982. Why Can You Have a Drink When You Can't *Have an Eat?. *Language,* 58(4), 753-799.

Systematic semantic differences between object-experiencer LVCs and corresponding simplex verbs in German

Niklas Wiskandt & Dila Turus

Heinrich-Heine-Universität Düsseldorf • Department of General Linguistics

In the literature on German light verb constructions (LVCs), it is often claimed that they typically have corresponding simplex verbs (SVs), e.g. *zur Aufführung bringen*, lit. 'transfer into performance'– *aufführen* ‚perform' (e.g. Polenz 1987). However, there is no consensus on how far LVCs and their corresponding simplex verbs are used as synonyms (e.g. Glatz 2006, Polenz 2018). The debate raises the question: Why should German have two different syntactic types of predicates that denote the very same event?

We address this question in an intensively discussed lexical domain: Object-experiencer predicates such as *frighten* or *worry* are known to show several peculiarities at the syntax-semantics interface since the study of Belletti & Rizzi (1988). Recently it has been investigated to what extent object-experiencer verbs can be classified as agentive; it seems that some of them frequently occur with human, and thus potentially agentive, subjects, while others strongly prefer inanimate entities or situations as subjects.

In German, there are frequent object-experiencer predicates both in the shape of SVs, e.g. *ängstigen* (‚frighten'), and of LVCs, e.g. *in Angst bringen* (literally ‚bring into fear') or *in Angst versetzen* (lit. ‚transfer into fear'). The *in N versetzen* (lit. ‚transfer, shift into N') pattern is particularly productive and works with numerous nouns denoting emotions, many of which are derived from object-experiencer SVs. We argue that there is a systematic difference in meaning between object-experiencer SVs and corresponding LVCs. This difference is visible in the argument selection of the predicates in corpus data.

We selected ten pairs of *in N versetzen* object-experiencer LVCs and corresponding SVs, and searched the W archive of the German reference corpus *DeReKo* (Leibniz-Institut, 2021), a sub-corpus, comprising a variety of newspapers and belles-lettres, using the COSMAS II interface (Leibniz-Institut, 2020). The first search string "in n /s0 &versetzen" targeted the preposition, the noun and the inflected forms of the light verb *versetzen* within one sentence. The second string "&v" targeted all inflected forms of the SV. The sample consists of 1,000 randomly collected sentences for each SV, and a maximum of 1,000 sentences for each LVC. In the cases where less than 1,000 tokens of an LVC were found by the search string, all instances were included in the sample. All sentences were analyzed manually. As the main annotation step, we analyzed the syntactic and semantic type of the arguments of the predicates and categorized subjects (I.) and objects (II.) into the classes listed in Table 1.

Parameter		**I. Type of the subject referent (non-experiencer argument)**	**II. Type of the object referent (experiencer argument)**
Value	a	[human] NP/pronoun	[animate, individuated] NP/pronoun
	b	[non-human animate] NP/pronoun	[animate, collective] NP/pronoun
	c	[inanimate, eventive] NP/pronoun	[animate, generic] NP/pronoun
	d	[inanimate, non-eventive] NP/pronoun	[inanimate] NP/pronoun
	e	infinitive construction	Ø (no overt argument)
	f	subject sentence	

	g	Ø (no overt argument)

Table 1: Categorization of argument referents

The quantitative analysis of our annotation results shows that differences between the LVC and SV patterns are visible in both subject referent and object referent type frequencies.

Two of the effects we carve out are the following, illustrated on the pair *in Begeisterung versetzen* vs. *begeistern*: Collective object NPs are particularly frequent with LVCs, exemplified in (1), while generic object NPs tend to appear with SVs (2). Inanimate, non-eventive subjects strongly prefer the SV as in (2) over the LVC, while inanimate eventive subjects favour the latter.

(1) *Der österreichische Saxophonist Benny Horatschek alias Mr. Soulsax **versetzt** das Publikum **in Begeisterung**.*
'The Austrian saxophonist Benny Horatschek alias Mr. Soulsax fills the audience with enthusiasm.'
(A07/OKT.08550 St. Galler Tagblatt, 18.10.2007, S. 61; Stimmung pur mit Mr. Soulsax)

(2) *Andere Fahrzeuge wiederum **begeistern** Technik-Interessierte und sportliche Lenker mit dem, was sie ausserhalb der Fahrgastzelle zu bieten haben.*
'Then again, other vehicles delight people interested in technology and racy drivers with the features they offer outside the passenger cabin.'
(A97/JUN.07770 St. Galler Tagblatt, 06.06.1997, Ressort: TB-AUT (Abk.); Ein rollendes Kraftwerk der leisen und sanften Art)

Our study provides an answer to the much-disputed question about the nature of German LVC-SV pairs, advances the knowledge on argument selection of object-experiencer predicates, and advocates the importance of empirical studies for answering theoretical questions.

References

Belletti, Adriana & Luigi Rizzi. 1988. Psych-Verbs and θ-Theory. *Natural Language & Linguistic Theory* 6(3), 291-352.

Glatz, Daniel. 2006. Funktionsverbgefüge - semantische Doubletten von einfachen Verben oder mehr? In Kristel Proost, Gisela Harras & Daniel Glatz (eds.)*, Domänen der Lexikalisierung kommunikativer Konzepte*. 129-178 Tübingen: Narr.

Leibniz-Institut für Deutsche Sprache. 2020. *COSMAS II (Corpus Search, Management and Analysis System)*. Leibniz-Institut für Deutsche Sprache, Mannheim.

Leibniz-Institut für Deutsche Sprache. 2021. *Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2021-I (Release vom 02.02.2021)*. Leibniz- Institut für Deutsche Sprache, Mannheim.

von Polenz, Peter. 2008. *Deutsche Satzsemantik. Grundbegriffe des Zwischen den-Zeilen-Lesens*. Berlin, Boston: De Gruyter.

Productivity (metrics) and semantics: a principal components analysis on minimizing and inchoative data

Margot Van den Heede, Sven Van Hulle, Timothy Colleman, Ludovic De Cuypere, Renata Enghels, Miriam Taverniers & Peter Lauwers

Ghent University

Productivity in its application to syntactic structure, especially in the framework of usage-based Construction Grammar (Goldberg 2019, Barðdal et al. 2015), refers to the domain of application of a grammatical pattern. More specifically, it concerns "the range of lexical items that may fill the slots of constructions" (Perek 2016: 66), hence their *lexical openness*.

The study has two goals. First, we wish to compare different productivity measurements, including both well-established and more innovative metrics. The well-known productivity measures that are examined are type/token, hapax/token and hapax/type ratios (Barðdal 2008, Zeldes 2012, Perek 2016). The 'anti-productivity' measures (Van Wettere 2021) include the token frequency of the most token frequent filler and the mean and standard deviation of the frequencies of the three most frequent fillers. In addition, the slope of the fitted Zipfian distribution (van Egmond 2013) of token frequencies is also taken into account.

Secondly, we additionally aim to disentangle the relation between lexical and semantic openness, as captured by distributional semantic analyses (Perek 2016). We understand semantic openness in terms of semantic range and semantic density. Semantic range is defined as the proportion of semantic clusters covered by a *given* micro-construction (i.e. an auxiliary or a minimizer construction, cf. below), within the onomasiological space delineated by the fillers of the *whole set* of micro-constructions (as such it is close to semantic variability, cf. Goldberg 2019). Semantic density, here applied to the whole micro-construction, captures the average semantic diversity of its types, computed on the basis of the average cosine distance between filler pairs (Perek 2016; Lenci 2018). In order to calculate the correlations between all these metrics, a Principal Components Analysis (= PCA) is conducted (Jolliffe & Cadima 2016, Van Wettere 2021).

As a testing ground, we compare the verb slot of two constructions, the minimizing construction in Netherlandic Dutch and the inchoative auxiliary construction in Peninsular Spanish. Minimizers are nouns that are used to reinforce sentential negation (Hoeksema 2002). They are recruited from different semantic categories, such as taboo terms (*geen reet uitmaken*, lit. 'to not matter an asshole'), units referring to distance (*voor geen meter vertrouwen*, lit. 'to not trust for a metre'), or weight (*geen gram aankomen,* lit. 'to not gain a gram'). The inchoative construction expresses the onset of an event, and incorporates auxiliaries from various semantic domains, for instance: change of state verbs (*rompió a llorar*, lit. 'he broke to cry'), motion verbs (*se echó a reír*, lit. 'she threw herself to laugh') or put verbs (*se mete a escribir*, lit. 'she puts herself to write') (Garachana 2017; Enghels & Van Hulle 2018; Fernández Martín 2019).

A first inspection of the PCA shows that (a) productivity and anti-productivity measures strongly correlate and that (b) semantic range correlates with lexical openness. In addition, the behaviour of inchoative constructions is more extreme, with on the one hand semantically very open auxiliaries and on the other hand idiomatic expressions. Based on this data, semantic density shows a moderate inverse correlation: the more types, the less dense (or sparser) the spectrum is on average, which might point to the effect of semantic coherence (Barðdal 2008). As for the minimizing data, a second dimension of productivity needs to be added, which correlates with hapax/type ratio and semantic

density. This corresponds to minimizers with a high number of (semantically diverse) hapaxes and types, despite the presence of a high token frequent predicate.

References

Barðdal, J. 2008. *Productivity: Evidence from case and argument structure in Icelandic* (Vol. 8). Amsterdam: John Benjamins Publishing.

Barðdal, J. et al. 2015. *Diachronic construction grammar*. Amsterdam: John Benjamins Publishing.

Enghels, R., & Van Hulle, S. 2018. El desarrollo de perífrasis incoativas cuasi-sinónimas: entre construccionalización y lexicalización. *Elua 32*. 91–110.

Fernández Martín, P. 2019. *Las perífrasis verbales del español: una perspectiva histórica*. Madrid: Arco/Libros.

Garachana, M. ed. 2017. *La gramática en la diacronía: la evolución de las perífrasis verbales modales en español*. Madrid: Iberoamericana Vervuert.

Goldberg, A. E. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton, NJ: Princeton University Press.

Hoeksema, J. 2002. Minimaliseerders in het Standaardnederlands. *Tabu 32*(3-4). 105-174.

Jolliffe, I. T., & Cadima, J. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374*(2065).

Lenci, A. 2018. Distributional models of word meaning. *Annual review of Linguistics 4*, 151-171.

Perek, F. 2016. Recent change in the productivity and schematicity of the way-construction: a distributional semantic analysis. *Corpus Linguistics and Linguistic Theory 14*(1). 65-97.

van Egmond, M. 2013. Calculating Zipf's law (and building growth curves) – Tutorial.

Van Wettere, N. 2021. Productivity of French and Dutch (semi-) copular constructions and the adverse impact of high token frequency. *International Journal of Corpus Linguistics*, *26*(3). 396-428.

Zeldes, A. 2012. *Productivity in argument selection: From morphology to syntax.* Berlin: De Gruyter Mouton.

On the internal and external productivity of IAW phrases in German

Steven Schoonjans

Alpen-Adria-Universität Klagenfurt & KU Leuven

IAW phrases, also called aggressively non-D-linked phrases, are a group of expressions that can be used especially in wh-questions to express "incomprehension from the side of the speaker with regard to the sentence proposition" (Stefanowitsch 2011:190, my translation). Typical examples for German include phrases such as *in aller Welt* 'in all world', *zur Hölle* 'the hell' and *um Himmels willen* 'for heaven's sake', as illustrated in (1–2):

(1) Warum *in aller Welt* sollte man ein Rockkonzert besuchen? (Stefanowitsch 2011:190)

'Why IAW should you visit a rock concert?'

(2) Wieso *zum Teufel* habt ihr überhaupt gewettet? (Catasso 2021:141)

'Why IAW did you bet in the first place?'

While Bayer & Trotzke (2015:21) claim that these phrases "are idiomatic and cannot be changed ad libitum", a search in different corpora and on the internet has yielded nearly 2000 different IAW phrases that represent seven different patterns, five of which can be considered as to some extent productive in present-day German: [*zu* X] 'the X', [*bei* X] 'by X', [*beim Barte* Xs] 'by X's beard', [*in* Xs *Namen*] 'in X's name', and [*um* Xs *willen*] 'for X's sake' ('Xs' in German standing for the genitive form of X or the paraphrase '*von* X').

The aim of this presentation is to get a better view of the productivity of these five patterns, both from an internal and from an external point of view. To this end, I present an analysis of the IAW phrases in four German corpora: the German reference corpus *DeReKo* (in COSMAS-II), the Reddit corpus *GeRedE* (Blombach et al. 2020), the e-mail corpus *CodE Alltag* (Eder et al. 2020), and a self-compiled set of texts from the Harry Potter and SpongeBob SquarePants communities.

External productivity is understood as the contexts in which these patterns can be used. More precisely, I will look at the question words they can combine with, showing that not all patterns are equally productive and that they show collocational preferences for different question words. As regards internal productivity, I will look at the elements (mainly nouns and noun phrases) that can occur in the X slot of the five patterns. Here as well, differences between the patterns can be found, [*bei/zu* X] being clearly more productive and showing both entrenched, highly token-frequent fillers and more occasional ones, while the other patterns may be somewhat more old-fashioned and seem to be more restricted to a small number of highly entrenched elements. Nevertheless, similar tendencies can be found for all five patterns, relating among other things to their semantics and pragmatics (e.g. dominance of taboo words (mainly, but not only, religious ones) and elements with a negative connotation) or to the usage context (e.g. community-specific slot fillers). Finally, the question will be addressed if there is a correlation between internal and external productivity, in the sense that the patterns that are more productive internally are also more productive from an external point of view.

References

Bayer, Josef & Andreas Trotzke. 2015. "The derivation and interpretation of left peripheral discourse particles." in: Bayer, Josef et al. (eds.). *Discourse-oriented Syntax*. Amsterdam: Benjamins. 13-40.

Blombach, Andreas et al. 2020. “A corpus of German Reddit exchanges (GeRedE).” in: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6310-6316.

Catasso, Nicholas. 2021. “Is German *warum* so special after all?” in: Soare, Gabriela (ed.). *Why is ‘Why’ Unique? Its syntactic and semantic properties*. Berlin: De Gruyter. 115-150.

Eder, Elisabeth et al. 2020. “CodE Alltag 2.0 — A pseudonymized German-language Email corpus.” in: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 4466-4477.

Stefanowitsch, Anatol. 2011. “Keine Grammatik ohne Konstruktionen: Ein logisch-ökonomisches Argument für die Konstruktionsgrammatik.” in: Engelberg, Stefan et al. (eds.). *Sprachliches Wissen zwischen Lexikon und Grammatik*. Berlin: De Gruyter. 181-210.

German 'wh-ever' and 'no matter wh-' as allostructions

Flor Vander Haegen

Ghent University

In a number of recent corpus studies within usage-based Construction Grammar, regression models are used to detect functional and distributional differences between the variants of an alternation (cf. e.g. De Vaere et al. 2021 and Pijpops et al. 2021). Whereas such studies are usually concerned with argument-structure alternations, my own contribution will use the same methodology to investigate a hitherto undiscussed alternation between two subtypes of conditionals, as illustrated in (1) and (2) below:

(1) Was *immer* wir probieren, das Auto macht nicht mit.

(2) *Egal* was wir probieren, das Auto macht nicht mit.
('Whatever/No matter what we try, the car is no longer working.')

Conditionals like (1) and (2) are known as 'universal concessive conditionals', or UCCs for short (Haspelmath & König, 1998). They differ from prototypical conditionals in that the protasis receives a quantificational reading thanks to the combination of a wh-word with either a clause-internal particle like *immer* '-ever' in (1) or a clause-external expression like *egal* 'no matter' in (2). Historically, the two variants represent a case of layering (Hopper, 1991): whereas internally marked UCCs are very old, sharing an ancestor with free relatives, their externally marked counterparts have been emerging from routinised discourse patterns involving 'no matter'-type predicates with embedded interrogatives over the past two centuries (Leuschner, 2006; Vander Haegen, 2019). This makes their synchronic relationship all the more intriguing, as any functional or distributional differences between them in contemporary German remain to be investigated.

In my talk, I will present preliminary results from my ongoing doctoral project on the horizontal relationships between the subtypes of concessive conditionals in the German constructicon. A logistic regression model based on data from the German Reference Corpus DeReKo (*N* = 3,000) reveals different preferences between the subtypes in (1) and (2) with regard to lexical specification and register, resulting in part from their distinct diachronic origins. Given their shared quantificational semantics, I argue that internally and externally marked UCCs constitute allostructions, i.e. alternating realisations of a formally underspecified UCC constructeme (cf. Perek, 2015, 153f. on allostructions and constructemes). Not only do my findings support a probabilistic view of syntactic variation, they also suggest interesting prospects for crosslinguistic comparison and identify methodological challenges specific to the investigation of non-argument structure alternations.

References

De Vaere, H., De Cuypere, L., & Willems, K. (2021). Alternating constructions with ditransitive *geben* in present-day German. *Corpus Linguistics and Linguistic Theory*, *17*(1), 73–107.

Haspelmath, M. & König, E. (1998). Concessive conditionals in the languages of Europe. In J. van der Auwera (ed.), *Adverbial constructions in the languages of Europe* (pp. 563–640). De Gruyter.

Hopper, P.J. (1991). On some principles of grammaticization. In E.C. Traugott & B. Heine (eds.), *Approaches to grammaticalization. Volume 1: Focus on theoretical and methodological issues* (pp. 17–35). Benjamins.

Leuschner, T. (2006). *Hypotaxis as building site: The emergence and grammaticalisation of concessive conditionals in English, German and Dutch.* LINCOM Europa.

Perek, F. (2015). *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives.* Benjamins.

Pijpops, D., Speelman, D., Van de Velde, F., & Grondelaers, S. (2021). Incorporating the multi-level nature of the constructicon into hypothesis testing. *Cognitive Linguistics*, *32*(3), 487-528.

Vander Haegen, F. (2019). Die Emergenz irrelevanzkonditionaler Subjunktoren des Typs *egal was*: Variation und Grammatikalisierung anhand des Deutschen Referenzkorpus. *Germanistische Mitteilungen*, *45*, 113–138.

Recent language change in Spanish teenage talk: the construction with *es que*

Nele Van Den Driessche

Ghent University

Recently, it has repeatedly been shown that language can change in the course of just a few decades. The 21st century, characterized by important sociocultural changes, such as the expansion of social media (Jenkins 2009), that have profoundly changed language, including the Spanish language, forms a highly interesting time period to investigate. However, the speed and nature of these "current changes" (Aarts *et al.* 2013: 1) have not yet been monitored systematically.

For this study, the construction with *es que* (e.g. *No viene porque es que se ha puesto enfermo,* Lit *'He doesn't come because it is that he is sick;* Fuentes Rodríguez 1997: 241) will be analyzed in teenage talk, since adolescents are said to play an important role in current changes as catalysts of language change. Teenage talk is indeed considered as one of the language variants that changes the fastest due to the creative language use of adolescents. Some of these linguistic innovations can end up being adopted by other generations with the aim of rejuvenating their language (Zimmerman 2002). Although *es que* is a highly typical feature of teenage talk, its (socio)linguistic behavior and evolution in the past three decades have not received many attention (Van Den Driessche & Enghels in press).

In view of this, the research objectives are twofold. First, the analysis wants to describe how the frequency and use of *es que* in Spanish teenage talk has changed over the past three decades. Secondly, I aim to contribute to the discussion on the use of corpora to examine processes of language change. Concretely, language change can be analyzed by applying different methodologies, namely by conducting a Real-Time Analysis and/or an Apparent-Time Analysis (i.a. Blas Arroyo 2005; Meyerhoff 2006). A Real-Time Analysis consists in the comparison of the speech of a constant group of speakers over different time periods, resulting in a longitudinal study (Bailey *et al.* 2001). However, scholars are frequently confronted with a lack of such comparable data, especially with regard to oral corpora. This presentation aims to demonstrate such Real-Time Analysis by using two highly comparable corpora.

In order to accomplish these goals, data of the COLAm corpus (*Corpus Oral de Lenguaje Adolescente de Madrid*, Jørgensen 2007, 2013), compiled at the beginning of the 21st century, will be compared with data from the CORMA corpus (*Corpus Oral De Madrid*, Enghels *et al.* 2020), that collects data from present-day Spanish. The data have then been subjected to a detailed functional, formal and sociolinguistic analysis, taking into account both internal (morphosyntactic and semantic-pragmatic) and external (sociolinguistic) parameters as well as measures of productivity (e.g. token frequency).

The first results show that *es que* has undergone some interesting changes in the past three decades. First, the data indicate that teenagers now use more often *es que* than at the beginning of the 21st century. Together with this increased token frequency, the analysis suggests an extension of the uses and contexts in which the construction with *es que* can be used.

References

Aarts, B., Close, J., Leech, G. N., & Wallis, S. (2013). *The verb phrase in English. Investigating recent language change with corpora*. Cambridge: Cambridge University Press.

Bailey, A., Ailey, G., Wikle, T., Tillery, J., & Sand, L. (1991). The apparent time construct. *Language Variation and Change, 3,* 241-264.

Blas Arroyo, J. L. (2005). *Sociolingüística del español: Desarrollos y perspectivas en el estudio de la lengua española en contexto social*. Madrid: Cátedra.

Enghels, R., De Latte, F., & Roels, L. (2020). El Corpus Oral de Madrid (CORMA): Materiales Para El Estudio (Socio)Lingüístico Del Español Coloquial Actual. *Zeitschrift fur Katalanistik, 33,* 45-76.

Fuentes Rodríguez, C. (1997). Los conectores en la lengua oral: *Es que* como introductor de enunciado. *Verba, 24,* 237-263.

Jenkins, H. (2009). *Confronting the Challenges of Participatory Culture: Media Education for the 21st Century*. Cambridge: MIT Press.

Jørgensen, A. M. (2007). COLA: Un corpus oral de lenguaje adolescente. *Oralia, 3,* 225-34.

Jørgensen, A. M. (2013). Spanish teenage language and the COLAm corpus. *Bergen Language and Linguistics Studies, 3*, 151-166.

Meyerhoff, M. (2006). *Introducing Sociolinguistics*. London: Routledge.

Zimmermann, K. (2002). La variedad juvenil y la interacción verbal entre jóvenes. In F. Rodríguez González (Ed.), *El lenguaje de los jóvenes* (pp. 137-164). Barcelona: Ariel.

Van Den Driessche, N. & Enghels, R. (In press). *La construcción con es que en el lenguaje juvenil madrileño: entre marcador discursivo, insubordinación y estrategia de dislocación*. Revue Romane.

Bulgarian 3- to 6-year-old children's productivity with causatives

Yanka Bezinska

Grenoble Alpes University

Keywords: causatives, causative situation, morphosyntactic productivity, linguistic complexity, language acquisition, Bulgarian

Productivity is a crucial concept in synchronic language description (Baayen 2009; Goldberg 2019; Zeldes 2012; Van Wettere 2021) and language change (Traugott & Trousdale 2013; Perek 2016). Surprisingly, this phenomenon has not been closely examined with regard to morphosyntax and language acquisition (Tomasello 2003; Hartsuiker & Bernolet 2017).

This proposal focuses on morphological and syntactic productivity from a developmental perspective. Our study is based on Bulgarian causatives, displaying various morphosyntactic and semantic complexity. This language has three causative mechanisms: lexical in (1), morphological in (2) and syntactic in (3):

(1) Majka-ta **hrani** bebe-to.
mother.F.SG-DEF feed.PRS.3SG (CAUS V) baby.N.SG-DEF
'The mother feeds the baby'

(2) Žaba-ta **raz-**plaka Tarzan.
frog.F.SG-DEF CAUS PREF-cry.PST.3SG Tarzan
'The frog made Tarzan cry'

(3) Momiče-to **kara** bebe-to **da** **se smee**.
girl.N.SG-DEF make.PRS.3SG (CAUS V) baby.N.SG-DEF conj laugh.PRS.3SG
'The girl makes the baby laugh'

Lexical devices are formally simple and tend to express direct (contact) causation where an agentive *causer* physically manipulates a patientive *causee* (Shibatani 1976). They serve to encode a causative situation that is cognitively considered as simple, because it is conceptualized as a single event, involving the spatio-temporal overlap of the causing and the caused event (Shibatani & Pardeshi 2002).

By the adding of the causative prefix 'raz-' or the combination of one causative and one lexical verb in a periphrastic construction, morphological and syntactic devices present a higher degree of formal complexity. Both mechanisms typically express indirect (distant) causation where an agentive *causer* gives instructions to an agentive *causee* (id.). This causative situation is cognitively more complex; it is conceptualized as the sum of two distinct causing and caused event, having their own agents and spatio-temporal profiles (id.).

In this study, we focus only on productive causatives, morphological and periphrastic (Shibatani 1976). We try to answer three main questions. The first one is related to the factors which influence the productivity with causatives. *Are these mechanisms produced spontaneously and if not, which strategies Bulgarian speakers could use if the language offers them several alternative constructions?* The second question concerns the impact of syntactic priming on causatives productivity. *To what extent providing structural model of morphological and periphrastic causatives could encourage participants to choose these devices?* The final question is linked to the availability of causative mechanisms in production and comprehension. *Are Bulgarian children able to correctly interpret causative events, even if the spontaneous use of productive causatives remains a challenge between 3 and 6 years of age?*

96 Bulgarian speakers (56 children and a control group of 40 adults) took part in the current study. Children were divided into three age groups (18 children aged 3 to 4, 17 children aged 4 to 5 and 21 children aged 5 to 6). They participated in three experimental tasks: production, comprehension and imitation. The first task was *production*, consisting of watching animated cartoons with various causative actions. Each video was visualized three times requiring participants to answer three gradual questions: *What X did? What Y did? What X did to Y?* The experiment continued with the *comprehension* task, consisting of simulating causative actions with plastic figurines (e.g. *The mother **makes** the baby **dance**. It's your turn now, do like the mom!*). Finally, *imitation* was designed as an elicited production task, by priming causative structures (e.g. *The mother **makes** the baby **dance**. Look carefully and tell me what is the daddy doing?* expected answer: *The daddy is **making** the girl **dance***.). As a control group, adults took part only in the production task.

Results reveal that participants' production strategies with causatives depend on a package of parameters such as the morphosyntactic complexity (or *cue cost*), the input frequency (or *cue availability*) and the specialization of each form in causation coding (or *cue reliability*) (MacWhinney 2005). Despite of their structural simplicity, morphological causatives show low averages in production (3% - 5% - 9% for children and 30% for adults, with statistically significant difference between groups: F (3;92) = 7,29; $p < .001$). Being polysemic and limited to only fifty Bulgarian verbs, these synthetic devices have not full linguistic availability and reliability, which explains their late emergence in children's language (beyond five years of age). Periphrastic causatives also display a limited use in the production task (11% - 7% - 15% for children and 18% for adults, with significant difference between groups: F (3;92) = 4,46; $p = .01$). These analytic devices describing strongly coercive causative macro-situations are formally and cognitively more complex, which explains speakers' preferences to some alternative constructions (lexical causatives or focalization on the causing or the caused event only).

Our results also indicate that between 3 and 6 years of age, Bulgarian children have a sufficiently precise cognitive representation of causation, because they can understand all causative devices available in their language (average scores over 56%, with no statistically significant difference between groups).

Finally, the data have revealed increasing averages in children's use with productive causatives during the imitation task (from 31% to 67% for the morphological devices and from 31% to 39% for the periphrastic mechanisms). Syntactic priming has clearly a significant positive impact on the availability of Bulgarian productive causatives, encouraging children to choose these constructions in case of linguistic competition.

References

Baayen, R. H. 2009. Corpus linguistics in morphology: morphological productivity. In *Corpus Linguistics. An international handbook*, eds. Lüdeling, A. & M. Kyto. Berlin: De Gruyter. 900–919.

Goldberg, A. 2019. *Explain me this. Creativity, Competition, and the Partial Productivity of Constructions.* Princeton: Princeton University Press.

Hartsuiker, R. J. & S. Bernolet. 2017. The development of shared syntax in second language learning. *Bilingualism: Language and Cognition* 20(2): 219–234.

MacWhinney, B. 2005. A unified model of language acquisition. In *Handbook of Bilingualism: Psycholinguistic Approaches*, eds. Kroll, J. F. & A. M. B. DeGroot. Oxford/New York: Oxford University Press. 49-67.

Perek, F. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics* 54(1): 149–188.

Shibatani, M. 1976. The grammar of causative constructions: a conspectus. In *Syntax and Semantics: The Grammar of Causative Constructions*, Vol. 6, ed. Shibatani, M. New York/San Francisco/London: Academic Press. 1-39.

Shibatani, M., & P. Pardeshi. 2002. The causative continuum. In *Typological Studies in Language*, Vol. 48, ed. Shibatani, M. Amsterdam: John Benjamins Publishing. 85–126.

Tomasello, M. 2003. *Constructing A Language: A Usage Based Theory of Language Acquisition*. Harvard: UP.

Traugott, E. & G. Trousdale. 2013. *Constructionalization and constructional change*. Oxford: Oxford UP.

Van Wettere, N. 2021. Productivity of French and Dutch (semi-)copular constructions and the adverse impact of high token frequency. *International Journal of Corpus linguistics* 26: 396-428.

Zeldes, A. 2012. *Productivity in Argument Selection: From Morphology to Syntax.* Berlin: De Gruyter.

Researching {verb, medium} colligations

Martin Godts & Miriam Taverniers

Many English verbs can function in transitive/intransitive alternations with the same constituent as (i) either direct object in a transitive clause or (ii) subject in an intransitive clause. In this alternation, the verb expresses an event as (i) either instigated by an agent with a patientlike object undergoing the process or (ii) a patientlike subject performing the act as if self-instigated – as initially defined by Davidse (1992). The constituent shared in both structures refers to the same medium (Halliday 1985: 144). The imagery linked to such an alternation is very similar and focuses on a semantic verb-noun embracement in which verbs and mediums license, specify and restrict each other's meanings, evoking an event equally embedded in both verb and medium, with verbs and mediums thus colligating.

This linguistic phenomenon can be considered lexically construed 'ergativity' as implemented in accusative languages, analogous to the typical morphological alignment as it appears in ergative languages – cf. McGregor (2009). At first, verbs relating some kind of physical change have been involved in this alternation, but through time application has expanded to all kinds of phasal, procedural, emotional and metaphorical changes involving patientlike mediums with an active twist. Many of the verbs participating in the alternation have inchoative and anticausative meanings in the intransitive alternant and causative aspects in the transitive counterparts. Additional features include resultative, perfective aspects and lexical aspects restricting the types of mediums fit to function in the alternation. Specifically in English, many of the verbs involved use the same form in both alternant constructions, and are referred to as 'labile' (Haspelmath 1993)). In contrast, the other Germanic and all Romance languages show more morphological and phrasal markers distinguishing the transitive and intransitive versions of the verbs.

This research is inquisitive and typological in nature in its attempt, first, to sketch a large synchronic picture of labile verbs in English and second, to map the emergence and occurrence of labile verbs in historical corpora. It starts from collections of English verbs attested as participating in causative alternations – i.e. Levin (1993), Francis et al. (1996) and McMillion (2006), the latter also including a diachronic account.

By extracting present-day instances of labile verbs in relation to colligate mediums from a huge web corpus, a data collection is obtained allowing to draw up an instance report for all the attested verbs, aiming at studying all potential occurrences of lexical ergativity. This includes co-occurring verbal particles, expanding the labile verbs with phrasal variants, and co-occurring prepositional objects and other adjuncts, which apparently play a role in licensing the acceptability of clauses as ergative alternants within a broader discourse context. Moreover, these data grant to cluster mediums and verbs in verb clusters with common mediums, in medium clusters with common verbs and in prototypical colligate clusters. On the basis of these reports, research results are visualized in a semantic vector space. Furthermore, as a syncretism may exist between an active ergative perfective and a passive construction – see Abraham (2003: 4-5), and compound nominalization signals a further step in {verb, medium} colligation, a delineation of the grammatical structures at hand in a multilayered semantic map including conceptual and more finegrained substantiated submaps concludes this part of the research.

A second, diachronic line of inquiry tracks the usage of {verb, medium} colligation in a series of historical corpora including biblical texts, prose, poetry, news, letters, etc. This part of the research wants to deepen the understanding of what triggers the emergence, expansion and productivity of the {verb, medium} colligation. A neural network setup describes and applies a procedure to track and

delineate particular grammatical features of the anticausative alternation. I track these features across the historical corpora referred to in the references. Moreover, I investigate whether a transformer model can be developed that can generate (learn) acceptable instances of the {verb, medium} colligation.

References

Abraham, Werner. 2003. Ergative diagnostics: temptation redux. Linguistik Online. 13. 10.13092/lo.13.868.

Davidse, Kristin. 1992. Transitivity/ergativity: the Janus-headed grammar of actions and events. In Martin Davies & Louise Ravelli (eds.) Advances in Systemic Linguistics: Recent Theory and Practice, 105-135. London/New York: Printer.

Francis, Gill; Susan Hunston & Elizabeth Manning et al. 1996-2022. Collins COBUILD Grammar Patterns – Ergative verbs, chapter 7. As retrieved online *https://grammar.collinsdictionary.com/grammar-pattern/chapter-7-ergative-verbs*, May 2022.

Halliday, M.A.K. 1985,1994. 4th Ed, revised by Christian M.I.M. Matthiessen. Halliday's introduction to Functional Grammar London. London, New York: Routledge.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. Bernard Comrie, and Maria Polinsky (eds.) *Causatives and Transitivity*. Amsterdam: Benjamins. 87-120.

Levin, Beth. 1993. Verb classes in English and alternations: A Preliminary Investigation. Chicago: University of Chicago Press.

McGregor, William. 2009. Typology of Ergativity. In Language and Linguistics Compass 3/1, Wiley.

McMillion, Allan. 2006. Labile Verbs in English. Their Meaning, Behavior and Structure. PhD dissertation, Stockholm University.

Historical Corpora

Kytö, Merja and Culpeper, Jonathan, 2006, A Corpus of English Dialogues 1560-1760 (CED), Oxford Text Archive, http://hdl.handle.net/20.500.12024/2507.

University of Oxford, 2003, Corpus of Early English Correspondence Sampler (CEECS), Oxford Text Archive, http://hdl.handle.net/20.500.12024/2461.

Amberley, John Russell, viscount, 1842-1876; Bell, Gertrude Lowthian, 1868-1926; Dowson, Ernest Christopher, 1867-1900; et al., 1994, Corpus of Late Modern English prose / David Denison, Oxford Text Archive, http://hdl.handle.net/20.500.12024/2077.

Nevalainen, Terttu; Raumolin-Brunberg, Helena; Keränen, Jukka; et al., 2006, Parsed Corpus of Early English Correspondence (PCEEC), Oxford Text Archive, http://hdl.handle.net/20.500.12024/2510.

University of Oxford, 2003, The English language of the north-west in the late Modern English period: a Corpus of late 18c Prose, Oxford Text Archive, http://hdl.handle.net/20.500.12024/2468.

The Corpus of Late Modern English Texts, version 3.1 (CLMET3.1) has been created by Hendrik De Smet, Susanne Flach, Hans-Jürgen Diller and Jukka Tyrkkö, as an offshoot of a bigger project developing a database of text descriptors (Diller, De Smet & Tyrkkö 2011). CLMET3.1 is a principled collection of public domain texts drawn from various online archiving projects.

University of Oxford, 1991, Helsinki corpus of English texts, Oxford Text Archive, http://hdl.handle.net/20.500.12024/1477.

University of Oxford, 2001, The York-Helsinki parsed corpus of Old English poetry (YCOEP), Oxford Text Archive, http://hdl.handle.net/20.500.12024/2425.

Measuring inter-individual variation in attitudes towards productivity: from corpus data to acceptability experiment

Anouk Van den Stock, Anne-Sophie Ghyselen & Timothy Colleman

Ghent University

Corpus-based investigations of productivity tend to abstract away from user-related variables. This is problematic, since productivity is also often referred to as a constrained form of creativity (Goldberg 2019). Creativity, however, is a property of the speaker, and not of the language, and is therefore likely to be influenced by individual, user-related variables (Hofmann 2018). Taking corpus data as its starting point, this study combines insights and methods of cognitive linguistics and sociolinguistics in an attempt to (i) establish whether there is individual variation in attitudes towards grammatical productivity/creativity, and (ii) explore which user-related variables affect language users' evaluation of productive instantiations of grammatical constructions.

This is done on the basis of a large-scale, internet-based acceptability rating experiment in which over 700 native speakers of Dutch rated both conventional and unconventional/productive/creative instantiations of two selected Dutch argument structure patterns, namely the *weg*-pattern in (1) and the *krijgen*-passive in (2), on a 7-point Likert scale.

(1) Hij maakte/baande/danste/hopte zich een weg door de menigte.

(2) Els kreeg een kaartje toegestuurd/afgeleverd/overhandigd.

Materials were selected on the basis of preliminary corpus investigations of the types of verbs frequently and less frequently encountered in the patterns at stake. As for the creative instantiations of these constructions – operationalized as being one-offs (i.e. hapaxes) of a construction – we also considered the verb's lemma frequency and its degree of semantic compatibility to the construction, modelled by Vector Space Semantics. A first glimpse at the results of the acceptability rating experiment reveals considerable variation in how speakers judge the less conventional instantiations of the above-mentioned constructions. In a next step, mixed ordinal regressions will be employed to map the correlation between attitudes towards grammatical productivity and the following three types of user-related variables: (i) 'prototypical' sociolinguistic variables such as age, gender and education, (ii) personality traits such as Extraversion, Openness and Conscientiousness – measured through the BFI-2 questionnaire (Soto & John 2017), and (iii) cognition-related measures such as receptive vocabulary and print exposure.

This poster presentation will present a detailed description of the methodological approach employed the study outlined above. We will tackle, for instance, which criteria we put forward to select critical items based on corpus data, the procedure of the acceptability rating experiment, and the general design of the internet-based survey.

References

Goldberg, A. (2019). *Explain me this. Creativity, Competition, and the Partial Productivity of Constructions*. Princeton: Princeton University Press.

Hoffmann, T. (2018). Creativity and Construction Grammar: Cognitive and Psychological Issues. *Zeitschrift für Anglistik und Amerikanistik*, *66*(3), 259-276.

Soto, C. J., & John, O. P. (2017). The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1), 117–143.

Combinatorial productivity of Spanish verbal periphrases as an indicator of their degree of grammaticalization

Mar Garachana & María Sol Sansiñena

University of Barcelona; KU Leuven

One of the most productive syntactic patterns for the expression of modal, temporal and aspectual values in Spanish is the combination VERB + VERB (see, a.o., Dietrich 1983, Gómez Torrego 1988, Fernández-Montraveta et al. in press). The studies on the constitution of the periphrastic system in Spanish show that, to a large extent, this productivity is based on the ease with which these combinations admit different verbal pieces in the second verb slot (Rodríguez Molina 2004, Garachana 2016). The changes in the type and token productivity over time have a direct impact on the type of subjects that a certain construction admits (e.g. meteorological and existential verbs admit zero subjects) and on the tense and mood in which the periphrasis is conjugated. The more grammaticalized a periphrasis gets, the more tenses and moods it will admit (Garachana 2016, 2017).

However, it has not been sufficiently explored whether the evolution of combinatorial patterns in near-synonymous periphrases follow similar grammaticalization paths. Adopting a constructionist, usage-based approach, we investigate the evolution of the Spanish near-synonymous periphrases *dejar de* + inf y *parar de* + inf, as in *Deja de/Para de gritar* (Lit. Stop of to shout 'Stop shouting') and we pose the following research questions:

1. Does the evolution of these two near-synonymous periphrases follow parallel collostructional patterns of host-class expansion? Given the observed lower productivity of *parar de* + inf — with respect to *dejar de* + inf— over time, how does productivity relate to changing token frequency of each of the types of each of the periphrases?

2. Is the expansion that *parar de* + inf experiences in the XXI century —compared to a much smaller use in previous centuries— due to a modification of the combinatorial patterns of the periphrasis or to the attraction that *dejar de* + inf may have exerted on it? If *dejar de* + inf acts as a *supporting construction*, the extension of tokens in the inf slot should not have a direct correlation with the types.

Corpus data are drawn from the CORDE, as well as from not exhaustive searches on CORPES XXI that allow access to data from more recent decades and less formal registers. Each token is analysed in terms of morphosyntactic and semantic-pragmatic parameters, as well as contextual elements. We conduct a collostructional analysis to investigate which lexemes are strongly attracted or repelled by the non-finite verb form slot in the construction. Specifically, we test the attraction between the two periphrastic constructions and verbs extracted from the CORDE through a distinctive collexeme analysis (Gries & Stefanowitsch 2004), which is known to be well suited for the study of related constructions.

We hypothesize that the relatively low productivity of *parar de* + inf may be affected by the high productivity of *dejar de* + inf. Our preliminary results show that the greater productivity of *dejar de* + inf can be mostly explained by its polysemic nature and not so much by the type of verbs that may appear in the inf slot.

References

CORDE = *Corpus diacrónico del español.* http://www.rae.es.

CREA = *Corpus del español actual.* http://www.rae.es.

CORPES XXI = *Corpus del español del siglo xxi.* http://www.rae.es.

Dietrich, W. (1983). *El aspecto verbal perifrástico en las lenguas románicas: estudios sobre el actual sistema verbal de las lenguas románicas*. Gredos.

Feltgen, Q. (2020). Diachronic Emergence of Zipf-like Patterns in Construction-Specific Frequency Distributions: A Quantitative Study of the *Way Too* Constructions. *Lexis* [Online], 16. https://journals.openedition.org/lexis/4968

Fernández-Montraveta, A., Vázquez, G., & Topor, M. (in press). A contrastive study of the degree of grammaticalization of verbal periphrases in Catalan, Spanish and Romanian. In M. Garachana, S. Montserrat, & C. D. Pusch (Eds.), *From composite predicates to verbal periphrases in Romance languages*. John Benjamins.

Garachana Camarero, M. (2016). Restricciones léxicas en la gramaticalización de las perífrasis verbales. *Rilce*, 32(1), 136–158.

Garachana Camarero, M. (2017). Los límites de una categoría híbrida. Las perífrasis verbales. In M. Garachana Camarero (Ed.), *La gramática en la diacronía. La evolución de las perífrasis verbales modales en español* (pp. 35–80). Iberoamericana-Vervuert.

Gómez Torrego, L. (1988). *Perífrasis verbales. Sintaxis, semántica y estilística*. Arco Libros.

Gries, S. & Stefanowitsch, A. (2004). Extending collostructional analysis A corpus-based perspective on 'alternations'. *lnternational Journal of Corpus Linguistics* 9(1), 97-129.

Rodríguez Molina, J. (2004). "Difusión léxica, cambio semántico y gramaticalización: el caso de *haber*+ participio en español antiguo". *Revista de filología española* 84(1), 169–209.

Productivity and functions of neoclassical initial combining forms in Estonian

Johanna Kiik & Maarja-Liisa Pilvik

University of Tartu, Estonia

Neoclassical constructions, prototypically complex words comprising morphemes of Latin and Greek origin, are widely used in many Indo-European as well as Finno-Ugric languages. They usually consist of combining forms that are derived from Latin and Greek stems. Often, the corresponding free lexemes have not been borrowed, and thus, neoclassical formations can be composed of bound morphemes linked without a stem (e.g. Estonian *mega-* + *-liit = megaliit* 'megalith' < kr *μέγας* 'big'' + *λίθος* 'stone') unlike native compounds and derivatives which have to consist of at least one free morpheme. In fact, neoclassical constructions are not uniform, their syntactic freedom, semantic transparency, functions and origin can vary greatly (Bauer 1998), which makes the grammatical status of neoclassical morphemes a question of debate.

Some neoclassical morphemes have started to combine with native stems over time (e.g. *antikangelane* 'antihero') and have become thus a productive mechanism of vocabulary expansion in Estonian as well as in many other languages. Some neoclassical elements have become even more independent, functioning as adverbial or adjectival modifiers (*megamõnus / mega mõnus* 'mega comfortable', *super õhtu* 'super evening'), and even appearing in predicative positions (*pidu oli mega* 'the party was mega', *probleem on pseudo* 'the problem is pseudo').

The aim of our study is to analyze the productivity and syntactic behaviour of more frequent neoclassical initial combining forms (ICFs) and their constructional patterns in Estonian. More specifically, the study will address the following research questions:

> Q1. Which functions do neoclassical ICFs perform in Estonian?
>
> Q2. Does morphological productivity correlate with syntactic freedom and flexibility of neoclassical ICFs?

We manually collected neoclassical ICFs from the Estonian "Dictionary of Foreign Words" (Vääri et al. 2012) and extracted the corresponding data from the Estonian National Corpus 2017 (approx. 1.3 bln tokens). In the paper, we analyze the neoclassical constructions using different productivity measures (e.g. those suggested by Baayen 1992) and examine the relationship between morphological and syntactic productivity as well as the semantic motivations behind different constructions.

Our initial results indicate that many neoclassical morphemes are productive in Estonian (to a different degree) and have a wide range of syntactic and semantic functions, although only some of them are used productively and spontaneously with native stems. The more popular ICFs of degree and evaluation like *super-* and *mega-*, in particular, are syntactically flexible and can appear as adjectival and adverbial modifiers as well as in predicative constructions. The positive correlation between frequency, productivity, and syntactic freedom of the ICFs is likely to result from a) the clearer perception of morpheme boundaries and b) the tendency to use the neoclassical morphemes as abbreviations of or stand-ins for longer formations.

References

Baayen, R. H. 1992. Quantitative aspects of morphological productivity. In Yearbook of Morphology 1991, ed. Booij, G. E., & J. Marle. Dordrecht: Kluwer Academic Publishers. 109–149.

Bauer, L. 1998. Is there a class of neoclassical compounds, and if so is it productive? – Linguistics 36(3): 403–422.

Estonian National Corpus 2017. Center of Estonian Language Resources. Vääri, E., R. Kleis, J. Silvet. 2012. Võõrsõnade leksikon. Tallinn: Valgus. https://www.eki.ee/dict/vsl/

Grammar "bores the crap out of me!": A mixed-method study on the "X *the* Y *out of* Z" construction and its usage by ESL and ENL speakers

Nok Chin Lydia Chan

Different from Generative Grammar which sees grammar as a formal system of how words are put together to form sentences, Construction Grammar (CxG) suggests that grammar is more than that as it includes many form-and-meaning pairings which are called constructions. For years, Construction Grammarians have been investigating constructions with various approaches, including corpus-linguistics, pedagogical, second language acquisition and so on, yet there is still room for exploration.

The present paper aims to further investigate the [V *the* $N_{taboo\text{-}word}$ *out of*]-construction (Hoeksema & Napoli, 2008; Haïk, 2012; Perek, 2016; Hoffmann, 2020) (e.g., *I kick the hell out of him.*) and propose a new umbrella construction, "X *the* Y *out of* Z" (XTYOFZ) construction, for it. Another aim is to examine the usage and comprehension of the XTYOFZ construction by English as a Second Language (ESL) and English as Native Language (ENL) speakers. To obtain these objectives, this paper attempts to answer the following research questions (RQs):

1. What are the syntactic and semantic properties, and the usage contexts of the XTYOFZ construction?
2. Why should XTYOFZ construction be considered as a construction?
3. How do ESL speakers comprehend and use the XTYOFZ construction compared to ENL speakers?

To answer RQs 1 and 2 (i.e., the first research aim), a corpus-based study with the Corpus of Contemporary American English (COCA) was performed. Two searches were done in total, one focusing on the frequencies of the words used as X, Y and Z in the construction, and another one looking at the frequency of occurrences of the construction across different genres. As to answer RQ 3 (i.e., the second research aim), a timed Lexical Decision Task (LDT) and a follow-up survey were conducted. The follow-up survey consisted of questions about participants' English acquisition and usage, as well as a short task testing participants' production and comprehension of the construction.

Corpus data shows that the combination of non-motion action verbs (e.g., *scare*, *beat*) as X and taboo terms (e.g., *shit*, *hell*) as Y was the most common. Also, it was found that the construction occurs mostly in non-academic contexts such as websites and TV/movies. Furthermore, the corpus data supports that XTYOFZ construction should be considered as an umbrella construction which embeds various sub-constructions that share the same syntactic form, non-compositional meaning, and idiosyncratic constraints. On the other hand, results from the LDT show that ESL speakers access constructional meaning slightly more slowly than ENL speakers. It was found that ENL speakers had slightly faster reaction times and much lower error rate than the ESL speakers. The follow-up survey also reflects that ESL speakers had a harder time to produce and comprehend the construction compared to ENL speakers. It was rather obvious that ENL participants were much more familiar with the construction than the ESL participants.

By investigating the features of a relatively less-discussed construction and its usage by ESL speakers, this study hopes to increase the knowledge base of CxG and ESL construction comprehension and usage, particularly on the constructions that are mainly used in more casual settings.

References

Haïk, I. (2012). *The hell* in English grammar. In N. Le Querler, F. Neveu & E. Roussel (Eds.), *Relations, connexions, dépendances: Hommage au professeur Claude Guimier* (pp. 101-126). Rennes: Presses Universitaires de Rennes.

Hoeksema, J. & Napoli, D. J. (2008). Just for the hell of it: A comparison of two taboo-term constructions. *Journal of linguistics*, 44(2), 347-378.

Hoffmann, T. (2020). Marginal argument structure constructions: the [V the $N_{taboo\text{-}word}$ out of]-construction in post-colonial Englishes. *Linguistics vanguard*, 6(1).

Perek, F. (2016). Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, 54(1), 149-188.

Through the translation glass: How parallel corpora can help us understand fuzzy grammatical categories

Charlotte Maekelberghe & Isabelle Delaere

KU Leuven

Various frameworks, most notably those in the tradition of cognitive-functional linguistics, recognize the existence of "fuzziness" in grammar, i.e., that boundaries between categories may be non-discrete (Aarts et al. 2004). The English gerund, with its combination of nominal and verbal properties, is a typical illustration of such a fuzzy category, blurring the distinction between the classes of nouns and verbs. an

While the structural hybridity of fuzzy categories can be resolved on the basis of formal diagnostics, functional hybridity is much more difficult to objectively describe. In this study, we claim that examining fuzzy categories through the lens of translation can help us better understand their functioning. Because languages such as German and Dutch lack a hybrid gerund construction, translations of English gerunds into these languages undergo an obligatory grammatical shift. More specifically, English gerunds can be translated into German and Dutch by either fully nominal constructions, such as *-ung* and *-ing* nominalizations or nominalized infinitives (1a-b), or by fully clausal structures, including finite and non-finite verb phrases (2a-b). We argue, then, that the choice for a particular translation strategy may render explicit the (nominal or clausal) construal imposed on a situation by the original English gerund.

(1) a. ***Building*** *a solid foundation for sustainable development* is a responsibility shared by developed and developing countries.
Die Schaffung *einer soliden Grundlage für die nachhaltige Entwicklung* ist eine Verantwortung, die von Industrie- und Entwicklungsländern geteilt wird. (CroCo)

b. We know that one of the tasks in ***increasing*** *energy efficiency* is to improve public awareness.
We weten dat het bewustmaken van het publiek een belangrijke taak is bij ***het bevorderen van*** *een efficiënter gebruik van energie*. (DPC)

(2) a. President Bush has been reestablishing American trade leadership by ***moving*** *on multiple fronts*.
Präsident Bush stellte die amerikanische Führungsrolle im Handel wieder her, indem *er Maßnahmen an mehreren Fronten* ***ergriff***. (CroCo)

b. Of course, we must find ways of ***ensuring*** *that vessels are safe.*
Uiteraard moeten wij enerzijds manieren vinden om *de veiligheid van vaartuigen* ***te waarborgen***. (DPC)

On the basis of data from the English-German parallel corpus CroCo (Hansen-Schirra et al. 2012) and the Dutch Parallel Corpus (Macken et al. 2011), we manually annotated the strategies that were used to translate a set of 3,000 English gerunds (viz. nominal, clausal, omission of the gerund, or another strategy). We then conducted a conditional inference tree and random forest analysis to examine the factors that predict the choice for a particular translation strategy, incorporating language-internal (coreferentiality relation, clausal function, length of the gerund phrase) and language-external (target language, genre) variables.

Findings show that coreferentiality is the strongest predictor of translation strategy, with clausal strategies being significantly preferred over nominal ones when the gerund is controlled by the subject

of its matrix clause. A second variable acting as a strong predictor is genre, where a distinction is made between genres which prefer clausal strategies (e.g. fiction, popular-scientific texts) and those which prefer nominal strategies (e.g. essays, manuals). Differences between German and Dutch only emerge within certain genres, viz. essays and official speeches, which can be partly attributed to the distinct composition of these subcorpora in CroCo and DPC.

Our study not only confirms the functional hybridity of the English gerund, it also reveals the specific contexts which foreground either its nominal or clausal profile. On a more general level, we highlight the benefits that parallel corpora can offer to the study of grammatical variation.

References

Aarts, Bas, David Denison, Evelien Keizer & Gergana Popova (Eds.). 2004. *Fuzzy Grammar: A reader*. Oxford: Oxford University Press.

Hansen-Schirra, S., Neumann, S. & Steiner, E. (2012). *Cross-Linguistic Corpora for the Study of Translations. Insights from the language pair English-German*. Berlin: Mouton de Gruyter.

Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus. *Meta*, 56(2), 374–390.

The bi-absolutive construction in Chechen: a comparison of two corpora and elicitation data

Zarina Molochieva, Pegah Faghiri & Eva van Lier

University of Amsterdam

This talk discusses the so-called bi-absolutive construction (henceforth BC) in Chechen, a Nakh-Daghestanian language spoken in the Chechen Republic, Russia. In particular, we will compare the (restrictions on) usage of this construction in two different corpora and in elicited speech.

Chechen is an ergative language; the basic transitive construction is illustrated in (1a). However, under specific circumstances the agent can take absolutive instead of ergative case marking. This amounts to the BC, shown in (1b). The same alternation is possible with verbs that take a dative-absolutive case frame in the basic construction, as in example (2a). In the BC in (2b), the dative experiencer takes absolutive case. In addition, (1b) and (2b) show that the BC involves a combination of an auxiliary ('be'), which agrees in gender with the agent/experiencer, and a lexical verb in the simultaneous converb form, which agrees with the gender of the patient/stimulus (which also, by default, takes absolutive case marking). The letters J, V, D, B indicate gender, which is morphologically unmarked on the noun.

(1) a. beerasha kiertash kegjo
child.PL.ERG fences-PL(J) break-J-make.PRS
'The children break the fences.' (elicited)

b. *oj beerash d-u gondahw kiertash keg-j-ie-sh*
INTERJ child.PL(D) D-be.PRS around fences-PL(J) break-J-make-CVBsim
'Oh, the children are breaking the fences there,' (witch_522)

(2) a. *k'ant-ana i jo? j-ieza.*
boy(v)-DAT dem girl(J) J-love.PRS
'The boy loves the girl.'

b. *k'ant I jo? j-ieza-sh v-u.*
boy(V) girl(J) (J)-love-SIM.CVB v-BE.PRES
'The boy loves the girl.'

Previous research shows that the bi-absolutive construction is restricted to progressive aspect contexts, to human or animate agents/experiencers, and to verb classes with either an ergative-absolutive or a dative absolutive basic case frame, as opposed to other frames, such as absolutive-lative (Molochieva 2011:178, Forker 2012).

Our study compares the use of the BC across three data sources: a corpus of spoken narratives consisting of approximately 1000 clauses (Molochieva & Walker in prep.), a corpus of written newspaper articles consisting of approximately 1880 clauses (Komen, to appear), and native-speaker production data elicited with visual stimuli. The corpora were annotated for humanness/(in)animacy of the agent/experiencer argument and for verb class. In addition, we systematically manipulated these factors when using visual stimuli (pictures and short videos) to elicit the BC with native speakers.

Our results show, firstly, that BCs are rare in the narrative spoken corpus: only 4 clear instances are found. In contrast, in the written corpus, we identified 17 instances by using the available syntactic annotation – via the corpus web-based research application (https://cesar.science.ru.nl/). We explain this frequency difference in terms of the mismatch between the function of the BC and the narrative

genre of the spoken corpus: BCs are used to describe situations that are ongoing at the moment of speech, while narratives usually consist of chains of subsequent events. In Chechen, such chains are typically expressed by means of series of non-finite verb forms, and often involve a high proportion of unexpressed arguments. Secondly, we will discuss constraints on the BC that are not genre-dependent, but bear on referential properties of the agent and on predicate class. The elicitation data show that BCs are strongly preferred with human and animate agents, and dispreferred with inanimate agents. Experiencer verbs also participate in the BC, but not verbs with other case frames.

Overall, by combining two distinct corpora and controlled elicitation data, our study sheds light on the different types of constraints – lexical, semantic, and discourse-pragmatic – that affect the use of alternating constructions like the BC.

References

Forker, Diana. 2012. The bi-absolutive construction in Nakh-Daghestanian. Folia Linguistica 46(1), 75 – 108.

Komen, Erwin (to appear). Constructing a corpus of modern Chechen. Available online: http://erwinkomen.ruhosting.nl/che/crp/

Molochieva, Zarina. 2011. Tense, aspect, and mood in Chechen. PhD Thesis University of Leipzig

Molochieva, Zarina & Katherine Walker (in prep.). Multi-CAST Chechen. In Haig, Geoffrey & Stefan Schnell (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts.*

Behavioural patterns and grammatical constructions: predicting alternation choices between English future constructions

Olaf Mikkelsen[a,b] & Dylan Glynn[a,b]

[a]Paris 8 University, [b]Adam Mickiewicz University

Keywords: multifactorial feature analysis, construction grammar, future construction, alternations, English

Expressions of futurity such as *will* and *BE going to* have received a lot of attention in the literature on English tense usage (see Wekker 1976, Fleischman 1982, Bergs 2010 *inter alia*), and a number of conceptual-functional motivations for the choice of construction have been proposed. The present study shows that the alternation can be adequately described in terms of the following dimensions of meaning: (1) intentionality, (2) temporal proximity, (3) epistemic certainty and (4) present relevance. These are all predicted to correlate positively with *BE going to*. While some recent studies have tackled this subject quantitatively (Hilpert 2008, Denis & Tagliamonte 2018, Flach 2021) by looking at collocational patterns, this study goes further in the testing of previous introspection-based research employing multifactorial feature analysis (Gries 2006, Glynn 2009), the results of which are interpreted in line with the theoretical assumptions of Construction Grammar (Goldberg 2006).

Data is drawn from the British English part of the LiveJournal corpus (Speelman & Glynn 2012) and a subsample of 200 occurrences of each construction is controlled for stylistic, interpersonal and syntactic variation, but not for lexical slots. The annotation schema consists of 20 semantic variables, and the data is manually annotated by two annotators (inter-rater agreement obtained using Cohen's Kappa). Mixed-effects binomial logistic regression with Markov Chain Monte Carlo cross validation is then used to model the results of the feature analysis. Initial results suggest that the alternation between *will* and *BE going to* can be explained in terms of the four abovementioned meaning dimensions, but that one of them does not behave according to previous findings: while *BE going to* significantly correlates with temporal proximity, epistemic certainty and present relevance, intentionality is a significant predictor of *will*. The explanation for this unexpected finding could be found in the interaction with grammatical person: while previous research has focused on first person uses (where *BE going to* is the more frequent form), the speaker's intention can also be expressed with the passive construction, which overwhelmingly favours *will*. We will discuss possible reasons for why speakers may want to construe their intentions in this way, and how well the findings fit the semantic description of the two constructions under scrutiny.

References

Bergs, Alexander. (2010). Expressions of futurity in contemporary English: A Construction Grammar perspective. *English Language and Linguistics, 14*(2), 217-238. doi:10.1017/S1360674310000067

Denis, Derek & Tagliamonte, Sali A. (2018). The changing future: Competition, specialization and reorganization in the contemporary English future temporal reference system. *English Language and Linguistics, 22*(3), 403-430. doi:10.1017/S1360674316000551

Flach, Susanne. (2021). Beyond modal idioms and modal harmony: A corpus-based analysis of gradient idiomaticity in mod adv collocations. *English Language and Linguistics, 25*(4), 743-765. doi:10.1017/S1360674320000301

Fleischman, Suzanne. (1982). *The Future in Thought and Language*. New York: Cambridge University Press.

Glynn, Dylan. (2009). A usage-based method for Cognitive Semantics. V. Evans & S. Pourcel (eds.), *New Directions in Cognitive Linguistics*, 77–106. Amsterdam: John Benjamins.

Goldberg, Adele. (2006). *Constructions at work. The Nature of Generalization in Language*. Oxford: Oxford University Press.

Gries, Stefan Th. (2006). Corpus-based methods and Cognitive Semantics: The many senses of to run. St. Th. Gries, & A. Stefanowitsch (eds.), *Corpora in Cognitive Linguistics*, 57–99. Berlin: Mouton de Gruyter.

Hilpert, Martin. (2008). *Germanic Future Constructions*. Amsterdam: John Benjamins.

Speelman, Dirk & Dylan Glynn. (2012). LiveJournal Corpus of British and American online personal diaries. *University of Leuven*.

Wekker, Herman Christian. (1976). *The Expression of Future Time in Contemporary British English*. Amsterdam: North Holland.

Modelling meaning differences in syntactic alternations with token-based vectors

Stefano De Pascale & Dirk Pijpops

Researchers in usage-based construction grammar are increasingly using distributional vector modelling, e.g. to study changes in the productivity of syntactic constructions (Perek 2016) or lexical biases in the choice between syntactic alternants (Pijpops et al. 2018). This technique models the meaning of a word or a construction by representing the textual contexts of its occurrences in a representative corpus as a mathematical vector (see Lenci 2018 for an accessible introduction, as well as further references). Until now, this research has primarily used type-based vectors, whereby each vector represents the entire polysemy of a lemma or a constructional variant. We aim to add to this research by employing token-based vectors (Heylen et al. 2015; De Pascale & Zhang 2021). These vectors overcome the problematic conflation of senses in type-based vectors, by representing the meaning of individual occurrences of a word or a construction, as the weighted average of the type-based vectors of the context words of that occurrence.

The present study uses token-based vectors to investigate the choice between two seemingly interchangeable constructions (cf. Gries and Stefanowitsch 2004; Dosedlová and Lu 2019). As a case study, we look at the alternation between the Dutch transitive construction and the so-called *naar*-construction, as in (1)-(2). We calculate token-based vectors for all occurrences of a verb in either construction and plot these vectors using Multidimensional Scaling. This allows us the delineate the semantic space where both constructions overlap, as well where they diverge.

(1) *Ik verlang echt (naar) betere resultaten.*
I desire really (to) better results
'I really desire the better results.'

(2) *De man greep (naar) een mes en stak een van zijn kameraden twee keer.*
The man grabbed (to) a knife and stabbed one of his comrades two times
'The man grabbed a knife and stabbed one of his comrades twice.'

We first focus on the variation of the verb *verlangen* 'desire', which has already been taken under scrutiny in Pijpops et al. (2021). This allows us to confirm the validity of our technique. Next, we turn to the variation of the verb *grijpen* 'grab', which has so far not been studied in any real depth. Our results indicate that token-based distributional vectors can be a useful addition to the methodological toolbox of researchers in usage-based construction grammar.

References

De Pascale, Stefano & Weiwei Zhang. 2021. Scoring with Token-based Models. A Distributional Semantic Replication of Sociolectometric Analyses in Geeraerts, Grondelaers, and Speelman (1999). In G. Kristiansen, K. Franco, S. De Pascale, L. Rosseel & W. Zhang (eds.), *Cognitive Sociolinguistics Revisited*, 186–199. Berlin, Boston: De Gruyter Mouton.

Dosedlová, Aneta and Wei-lun Lu. 2019. The near-synonymy of classifiers and construal operation A corpus-based study of ke and zhu in Chinese. *Review of cognitive linguistics* 17(1). 113–130.

Gries, Stefan Thomas and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on "alternations." *International journal of corpus linguistics* 9(1). 97–130.

Heylen, Kris, Thomas Wielfaert, Dirk Speelman and Dirk Geeraerts. 2015. Monitoring Polysemy. Word Space Models as a Tool for Large-Scale Lexical Semantic Analysis. *Lingua* 157. 153–172.

Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics* 4. 151–171.

Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics* 54(1). 14–188.

Pijpops, Dirk, Dirk Speelman, Stefan Grondelaers and Freek Van de Velde. 2021. Incorporating the multi-level nature of the constructicon into hypothesis testing. *Cognitive Linguistics* 32(3). 487–528.

Pijpops, Dirk, Dirk Speelman, Stefan Grondelaers and Freek Van de Velde. 2018. Comparing explanations for the Complexity Principle. Evidence from argument realization. *Language and Cognition* 10(3). 514–543.

Measuring the extensibility of a construction relative to the constructional network based on onomasiological domain and discourse situation (pragmeme)

Chris A. Smith

CRISCO EA4255; Université de Caen

This paper seeks to address the relation between semantics, pragmatics and the productivity of a low level lexico-grammatical construction Have the N (body part/) to ↔attitude. The question posed is how semantics affects productivity, in the generative sense of extensibility of a construction (a form meaning pairing). I take an onomasiological approach to a lexicogrammatical construction, *have X to V*: Have the N (body part/) to ↔attitude". *Have the N to* is a construction that takes a wide selection of nouns relating to body parts or emotions (see data below from the COHA). The construction falls under a general semantic abstract schema of metaphor and metonymy relating to body parts and emotions and attitudes (confidence, arrogance). I argue this construction belongs to what can be called the "pragmeme of politeness (approval/disapproval)". The expressions falling under this pragmeme are expressions of reproach or admiration in the face of a perceived attitude (such as insolence, assurance/ confidence, or courage): *have the heart to, have the nerve to, have the cheek to, have the guts to.* My position is that these constructions can be subsumed under a larger constructional network which can be identified as a **pragmeme** (Mey 2010: 2884) or "general situational prototypes of acts that are capable of being executed in a particular situation or cluster of situations".

The pragmeme is instantiated by discourse formulae, which can represent the locus of change in a constructional network (Torres Cacoullos and Walker 2009) and so should be taken into account in measuring extensibility. If the productivity of a schema is measured purely as generational productivity i.e the extension capacity of the schema, or the frequency of use (token), this doesn't take into account the pool of available forms fitting the pattern. I argue that **that productivity levels** of the construction are relative to the constructional architecture and have to be considered relatively within the network. This means that relative productivity (extensibility) gives information regarding the salience and attractivity of the construction. In addition, the extension to new "filler" words (via analogy) is dependent on the availability of alternatives in the onomasiological paradigm. Empty slot "fillers" are far more than lexical fillers and cannot be reduced to a quantitative approach. "Fillers" represent exemplar connections within the lexicogrammatical continuum and therefore are themselves low level constructions than can affect other constructions in the network (and thus affect semantic change in lexical items). A bottom-up approach to constructions tends to consider the importance of low-level constructions and how they affect higher level constructions, following Gyselinck 2020, Budts and Petré 2020 on the importance of considering the horizontal links between constructions.

The paper is organised in three parts:

1) We first identify the specificity and variations of the Have the N (body part/) to ↔attitude construction within the pragmeme of politeness using the COHA

2) We attempt to track the constructional architecture based on the instantiations of the construction, taking into account horizontal links (synonymy, polysemy)

3) We measure the attractivity of the construction based on the extensibility within the onomasiological frame (available pool of forms expressing an attitude/emotion).

Data

HAVE THE N (concept)	Tokens	HAVE THE N (body part)	Tokens
HAVE THE RIGHT TO	1561	HAVE THE **HEART** TO	234
HAS THE RIGHT TO	854	HAD THE HEART TO	82
HAD THE RIGHT TO	603		
HAVING THE RIGHT TO	44		
HAVE THE POWER TO	720	HAD THE **NERVE** TO	199
HAS THE POWER TO	469	HAVE THE NERVE TO	169
HAD THE POWER TO	518	HAS THE NERVE TO	33
HAVING THE POWER TO	39		
HAD THE COURAGE	561	HAVE THE **FACE** TO	35
TO HAVE THE COURAGE TO	449		
HAS THE COURAGE TO	89		
HAVING THE COURAGE TO	49		
HAS THE AUTHORITY TO	78	HAD THE **GUTS** TO	77
HAVE THE AUTHORITY TO	123	HAVE THE GUTS TO	119
HAD THE AUTHORITY TO	46		
HAVE THE OPPORTUNITY	334	HAD THE **GALL** TO	40
TO	247		
HAD THE OPPORTUNITY TO			
HAVE THE HONOR TO	333		
HAVE THE HONOUR TO	85		
HAD THE HONOR TO	84		
HAS THE HONOR TO	32		
HAVE THE ABILITY TO	254		
HAS THE ABILITY TO	150		
HAD THE ABILITY TO	106		
HAD THE CHANCE TO	239		
HAVE THE CHANCE TO	213		
HAS THE CHANCE TO	31		
HAD THE MISFORTUNE TO	233		
HAVE THE MISFORTUNE TO	50		
HAVE THE MONEY TO	186		
HAVE THE TIME TO	183		
HAD THE TIME TO	84		
HAVE THE STRENGTH TO	181		
HAD THE STRENGTH TO	72		
HAVE THE GOODNESS TO	150		
HAD THE AUDACITY TO	131		
HAS THE AUDACITY TO	30		
HAD THE SENSE TO	112		
HAVE THE SENSE TO	50		
HAVE THE MEANS TO	101		
HAD THE GRACE TO	87		
HAVE THE RESOURCES TO	87		
HAVE THE CAPACITY TO	87		
HAS THE CAPACITY TO	59		

HAD THE CAPACITY TO	42		
HAD THE FORESIGHT TO	73		
HAD THE TEMERITY TO	72		
HAVE THE KEY TO	71		
HAD THE MONEY TO	66		
HAS THE OPPORTUNITY TO	65		
HAVE THE ENERGY TO	65		
HAVE THE DECENCY TO	55		
HAVE THE KINDNESS TO	47		
HAD THE IMPUDENCE TO	46		
HAD THE EFFECT TO	45		
HAD THE LUCK TO	44		
HAD THE MEANS TO	40		
HAD THE WIT TO	40		
HAD THE PLEASURE TO	36		
HAD THE CURIOSITY TO	35		
HAD THE SATISFACTION TO	34		
HAVE THE PATIENCE TO	33		
HAD THE URGE TO	33		
HAD THE HAPPINESS TO HAVE THE HAPPINESS TO	32 30		
HAVE THE WILL TO	31		

Table: Variations of HAVE THE N TO in the COHA

References

Budts Sara and Petré Peter. (2020). Putting connections centre stage in diachronic construction grammar. In Lotte Sommerer & Elena Smirmova (eds), *Nodes and Networks in Diachronic Construction grammar*, 317-351. Amsterdam/ Philadephia: John Benjamins.

Goldberg Adele (2016). Partial productivity of linguistic constructions: Dynamic categorization and statistical preemption. *Language and Cognition*, 8, pp 369-390 doi:10.1017/langcog.2016.17

Goldberg, Adele E. (2019). *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton; Princeton University Press.

Gries Steven. Th. & Anatol. Stefanowitsch. (2004). Extending collostructional analysis: A corpus-based perspectives on 'alternations'. *International Journal of Corpus Linguistics*, 9(1), 97–129.

Gyselinck Emmeline. (2020). (Re) Shaping the constructional network. In Lotte Sommerer & Elena Smirmova (eds), *Nodes and Networks in Diachronic Construction grammar*, 107- 140. Amsterdam/ Philadephia: John Benjamins.

Lorenz David (2020). Converging variations and the emergence of horizontal links. In Lotte Sommerer & Elena Smirmova (eds), *Nodes and Networks in Diachronic Construction grammar*, 243-276. Amsterdam/ Philadephia: John Benjamins.

Mey Jacob. (2010). Reference and the pragmeme. *Journal of Pragmatics,* 2883-2888.

Perek Florent. (2020). Productivity and schematicity in constructional change. In Lotte Sommerer & Elena Smirmova (eds), *Nodes and Networks in Diachronic Construction grammar*, 141- 166. Amsterdam/ Philadephia: John Benjamins.

Petré, Peter. (2019). How constructions are born. The role of patterns in the constructionalization of be going to INF. In B. Busse & R. Möhlig-Falke (Eds.), *Patterns in language and linguistics: New Perspectives on a Ubiquitous Concept* (Topics in English 104), 157–192. Berlin: Mouton de Gruyter.

Torres Cacoullos, Rena & Walker, James A. (2009). The present of the English future: Grammatical variation and collocations in discourse. *Language* 85: 321–354.

Smith, Chris A. 2021 (in press). Approche constructionnelle de l'émergence d'expressions de l'insolence en anglais. *Cahiers de lexicologie*, 119: 119-148.

Do corpus-derived productivity measures predict language processing? The case of the Spanish inchoative

Mariia Baltais, R. Hartsuiker & Anna Jessen (alphabetical order)

Syntactic productivity has mainly been looked at so far from a corpus linguistic perspective (e.g., Zeldes 2012; Perek 2015). Measures such as type frequencies, or the ratio between types and the total number of instances (tokens) (= type/token ratio), are indicative of *realized* productivity, whereas the number of one-offs (hapaxes), or the hapax/token ratio can be viewed as a proxy for *potential* productivity (Baayen 2009). But it is debateable whether corpus measures actually suffice to define a construction's productivity, since the construction could be *extensible* beyond the scope of close-ended corpora (Barðdal 2008). The aim of this study is thus to examine productivity on the "speaker-level", and in two different modalities: comprehension and production. We ask whether the productivity measures gathered from the corpus – which reflects the level of the community - are predictive of native speakers' individual use of language. We conducted two experiments on the Spanish inchoative construction (see below): one acceptability rating and one sentence completion task in order to investigate the extensibility/productivity of a construction.

We chose the Spanish inchoative, which expresses the onset of an event, because it is strikingly productive (e.g., García Fernández 2012). A range of verbs can fill the inchoative verb slot, such as change-of-state verbs (*rompió a llorar*, lit. 'she broke to cry'), motion verbs (*se echó a reír*, lit. 'she threw herself to laugh'), and others (Enghels & Van Hulle 2018). Another source of productivity is the infinitive slot because many different events can be described as 'being started'. A corpus-based dataset has been created by extracting and manually cleaning data from the Spanish Web corpus (Van Hulle & Enghels, in press). Based on this dataset, materials for both experiments were developed.

Experiment 1: Comprehension

The acceptability rating survey included 6 inchoative verbs that varied in the productivity of the infinitive slot. Each inchoative was combined with 10 infinitives that varied in their token frequency of co-occurrence with the inchoative in the dataset (including hapaxes and non-attested infinitives). Acceptability ratings were given on a 7-point Likert scale by 96 native speakers of European Spanish. Both higher token frequency of the infinitive with an inchoative and higher hapax/token ratio of the inchoative (indicative of its productivity) corresponded with higher acceptability ratings (Table 1). Furthermore, for infinitives with low token frequency (0-10 tokens in a sample of 500 tokens) there was a significant token frequency by hapax/token ratio interaction: the more productive the inchoative, the smaller the effect of infinitive token frequency (β = -.45, *SE* = .16, *t* = -2.87, $p < .01$). In sum, corpus measures of productivity were predictive of participants' acceptability judgments. The interaction found in the low token frequency bands demonstrates that hapax/token ratio reflects a construction's extensibility towards new items, i.e., its productivity "at work" in speakers' minds.

Inchoative	Type/token ratio corpus	Hapax/token ratio corpus	Mean rating Exp. 1	N token Exp. 2	N type Exp. 2	Type/token ratio Exp.2	Hapax/token ratio Exp.2
empezar	0.56	0.39	6.33	100	43	0.43	**0.24**
lanzarse	0.43	0.27	5.57	62	32	0.52	**0.34**
meterse	0.42	0.28	5.66	**31**	**14**	0.45	0.26

ponerse	0.36	0.24	**6.02**	100	**36**	**0.36**	0.18
romper	0.06	0.03	5.01	99	8	0.08	0.04
echarse	0.03	0.01	**5.50**	**91**	**9**	**0.1**	0.01

Table 1: Corpus measures, ratings Exp. 1, results Exp. 2 – bold print = deviates from corpus data

Experiment 2: Production

Experiment 2 tested 25 inchoative verbs. Participants (n = 100) completed sentences like *Enrique empieza a* …. ('Enrique starts to … '). Sentence completions without an inchoative meaning were excluded, so that token frequencies varied (i.e., often below the maximum of 100). We calculated type/token ratios similarly to the corpus data (Table 1). Here we present the subset of the data corresponding to the 6 inchoative verbs from Experiment 1.

Significant correlations between type/token and hapax/token ratio (Table 1) showed that corpus measures predicted language production. But note that the large differences in sample size complicate direct comparison, therefore we conducted a more reliable analysis with reduced token frequencies, which we will discuss in more detail. Additionally, there was a correlation between mean ratings of Experiment 1 and the number of types per inchoative ("N type Exp. 2", Table 1) in Experiment 2, showing that inchoatives with higher ratings in Experiment 1 also attracted a higher number of different infinitives in Exp. 2. There was no other between-experiment correlation.

In conclusion, corpus data are predictive of both comprehension and production in individual speakers. Interestingly, zooming into single inchoative verbs reveals some differences (e.g., see *echarse* in Table 1). We will discuss these differences further in more detail.

References

Baayen, R. H. (2009). 43. Corpus linguistics in morphology: morphological productivity. *Corpus linguistics. An international handbook*, 900-919.

Barðdal, J. (2008). *Productivity: Evidence from case and argument structure in Icelandic* (Vol. 8). John Benjamins Publishing.

Enghels, R., & Van Hulle, S. (2018). El desarrollo de perífrasis incoativas cuasi-sinónimas: entre construccionalización y lexicalización. *ESTUDIOS DE LINGUISTICA-UNIVERSIDAD DE ALICANTE-ELUA*, 32, 91–110.

Garcia Fernandez, L. (2012). *Las perífrasis verbales en español.* Castalia

Perek, F. (2015). *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives* (Vol. 17). John Benjamins Publishing Company.

Van Hulle, S., & Enghels, R. (in press). De Spaanse inchoatiefconstructie in beeld. Clusteranalyse als antwoord op het quasi-synonymie vraagstuk. *Handelingen – Koninklijke Zuid-Nederlandse maatschappij voor taal-en letterkunde en geschiedenis.*

Zeldes, A. (2012). Productivity in argument selection. In *Productivity in Argument Selection*. De Gruyter Mouton.

L2-Korean learners' use of constructional components for Korean locative postposition–verb construction: Relationship between L2 textbook and L2 writing

Boo Kyung Jung & Gyu-Ho Shin

University of Pittsburgh; Palacký University Olomouc

Usage-based constructionist approaches highlight the role of language use in shaping linguistic knowledge.[1,2,3] For L2 acquisition under foreign language-learning contexts, L2 textbooks seem to provide L2 learners with an essential/primary source of L2 input.[4,5] With this background, the present study investigates the characteristics of L2 written production in Korean, a popular L2 target and yet understudied language in this regard, with a focus on a locative postposition–verb construction (LPVC). This construction (Figure.1) consists of a fixed slot for postposition (i.e., only one of the three particular options [*-ey, -eyse, -(u)lo*] is allowed), a moderately restricted slot for verb (i.e., its selection is contingent upon the semantics of the postposition and the intended event), and a rather open slot for noun unless its meaning is incompatible with the frame semantics (i.e., a location-related event).[6] We specifically ask how L2-Korean learners with different L1 backgrounds (Czech; English) utilize postposition–verb pairs in employing the LPVC, considering properties of L2 textbook input.

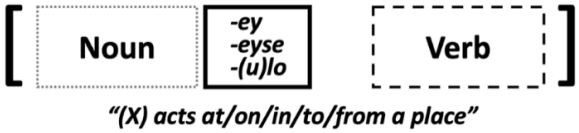
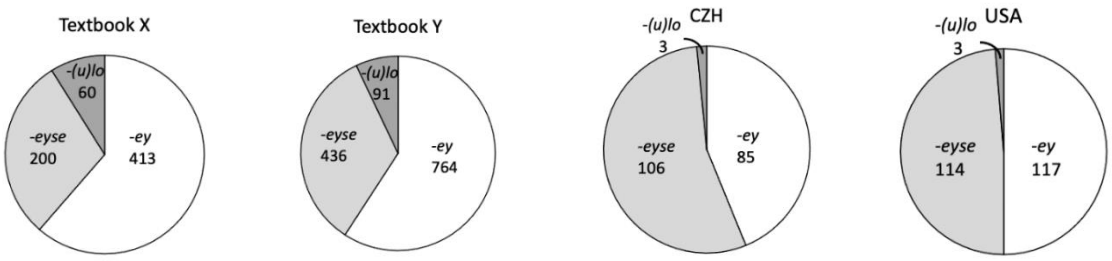


Figure 1. Schema of LPVC

L2-textbook analysis

We analyzed two textbook types, each of which is used widely in the Czech Republic (Textbook.X) and the United States (Textbook.Y) for tertiary-level instruction of Korean. Each type consists of multiple proficiency levels, and we selected the first four volumes of each. We electronically compiled all the sentences in the textbooks, organizing them by textbook type, and isolated LPVC instances in a semi-automatic manner.

Results (Figure.2a & Table.1): Among postpositions, *-ey* occurred most frequently, followed by *-eyse* and *-(u)lo*. The overall Type–Token Ratio (TTR) for the postposition–verb pairs showed that *-ey* was the highest, indicating its intensive use with certain verbs relative to the other two postpositions. Another TTR without *ka-* 'to go' (because of its excessive use in LPVC) showed gradual increase of the ratio from *-(u)lo* to *-eyse*/*-ey*, with *-ey* still the highest.



(a) L2 textbook (b) L2 writing

Figure 2. Overall occurrences of three locative postpositions

Table 1. Type–token ratio: verb use in the textbooks (token: type)

	Overall		Without *ka*- 'to go'	
	Textbook X	Textbook Y	Textbook X	Textbook Y
-ey	413:48 (8.60:1)	764:76 (10.05:1)	257:47 (5.45:1)	480:75 (6.4:1)
-eyse	200:73 (2.74:1)	436:96 (4.54:1)	195:72 (2.71:1)	419:95 (4.41:1)
-(u)lo	60:17 (3.53:1)	91:25 (3.64:1)	29:16 (1.81:1)	63:24 (2.63:1)

L2-writing analysis

34 L1-Czech learners (CZH, M_{age}=23.97, *SD*=2.69), 34 L1-English learners (USA, M_{age}=26.15, *SD*=4.43), and 25 native speakers (NSK; M_{age}=23.6, *SD*=4.10) of Korean wrote three argumentative essays for 20 minutes each. L2ers' proficiency in Korean was measured separately;[7] there was no statistical by-group difference in the proficiency scores. We extracted LPVC instances manually from the essays (spelling/spacing errors uncorrected) and identified the postposition–verb pairs.

Results (Figure.2b & Table2): Overall, L2ers rarely used of *-(u)lo*, compared to their use of *-ey* and *-eyse*, echoing the general textbook composition regarding LPVC. However, they commonly produced *-ey* and *-eyse* at a similar rate and demonstrated compelling TTR for these two postpositions. These trends were inconsistent with those found in the textbooks. This discrepancy may be due to language-specific properties: *-eyse* has two major functions but *-ey* has eight major functions,[8] and the different nature of one-to-many mapping under competition[9] involving the two postpositions may have affected L2ers' postposition choices in producing LPVC.

Table 2. Type–token ratio: verb use in writing (token: type)

	Overall			Without *ka*- 'to go'		
	NSK	CZH	ENG	NSK	CZH	ENG
-ey	158:75 (2.11:1)	85:33 (2.58:1)	117:58 (2.02:1)	158:75 (2.11:1)	79:32 (2.47:1)	111:57 (1.95:1)
-eyse	157:99 (1.59:1)	106:50 (2.12:1)	114:60 (1.9:1)	157:99 (1.59:1)	106:50 (2.12:1)	114:60 (1.9:1)
-(u)lo	7:7 (1:1)	3:2 (1.5:1)	3:3 (1:1)	7:7 (1:1)	3:2 (1.5:1)	2:2 (1:1)

Together, our findings suggest that L2 written production can be best explained by the interaction between L2-input distributions and target-knowledge particularities. This points to the noisier nature of L2 knowledge,[10,11] resulting in developmental trajectories of L2 distinctive from L1 acquisition.[12,13]

References

[1] Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). *Journal of Child Language, 42*(2), 239–273.

[2] Ellis, N. C. (2002). *Studies in Second Language Acquisition*, *24*(2), 143–188.

[3] Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*.

[4] Alsaif, A., & Milton, J. (2012). *Language Learning Journal*, *40*(1), 21–33.

[5] Römer, U. (2004). In G. Aston, S. Bernardini, & D. Steward (Eds.), *Corpora and language learners* (pp. 151–168).

[6] Jung, B., & Shin, G.-H. (accepted). *Corpora*.

[7] Lee-Ellis, S. (2009). *Language Testing, 26*(2), 245–274.

[8] Sohn, H. M. (1999). *The Korean language*.

[9] MacWhinney, B. (2008). In P. Robinson & N. C. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 341–371).

[10] Futrell, R., & Gibson, E. (2017). *Bilingualism: Language and Cognition, 20*(4), 683–684.

[11] Tachihara, K., & Goldberg, A. E. (2020). *Language Learning, 70*(1), 219–265.

[12] Jiang, N., Novokshanova, E., Masuda, K., & Wang, X. (2011). *Language Learning*, *61*(3), 940–967.

[13] Slabakova, R. (2014). *Foreign Language Teaching and Research*, *46*(4), 543–559.

Polyfunctional particles in discourse: a corpus-based study of Russian *čto li*

Beatrice Bernasconi

Roma Tre University – Sapienza University of Rome

The present work focuses on the analysis of the Russian polyfunctional particle *čto li* as a strategy of meaning construction and negotiation between interlocutors in conversation (Auer, 1984; Bazzanella, 2011; Jucker *et al*., 2003; Prince *et al.*, 1982; Wilson & Sperber, 2002). Several studies on Russian pragmatic particles as well as on approximation and other strategies for meaning (non)-definition in discourse have been conducted in recent years (Bodganova-Beglarian *et al*., 2019; Benigni, 2014; Podlesskaja & Starodubceva, 2013; Račeva, 2018; Žukov & Žukov, 2003). However, only little attention, to our knowledge, has been devoted to the complex particle *čto li* and its behaviour in Russian speech (Bogdanova-Beglarian, 2016). This study aims at analysing authentic corpus data in order to provide an overview of the diverse functions that *čto li* carries out in spoken Russian and an explanation that could account for all of them. The research is couched within the theoretical framework of Cognitive Linguistics and adopts a usage-based approach.

Čto li is a discourse complex particle that fulfils various semantic and pragmatic functions during referential work in conversation. It is composed by two elements: *čto*, 'what' or conjuction 'that' and *li*, a particle used to form polar questions or conditional clauses. It mainly occurs in questions (1), but is also attested in declaratives (2) and exhortatives (3):

(1) *Ob"javili, čto li?*
'Did they make an announcement, or what?'

(2) *— Da ešče soobščenie iz Moskvy — o likvidacii Kominterna, čto li.*
'And there was a report from Moscow – something about the liquidation of the Comintern.'

(3) *I davajte čaj pit', čto li.*
'Let's have some tea, shall we?'

Data were collected from the Multimodal Corpus of the Russian National Corpus. A dataset of 300 occurrences was randomly extracted and manually annotated for several factors, such as the sentence type in which *čto li* occurs, the pragmatic function of the question – when it occurs in an interrogative –, whether it approximates the meaning, and its position within the sentence. A quanti-qualitative analysis was conducted on the dataset. Three main functions of the particle (i. approximation, ii. reduction of commitment, iii. interlocutor engagement in the referential process) and four main linguistic contexts in which they are exploited (a. requests for confirmation, b. rhetorical questions, c. declaratives, d. exhortatives) were identified. An explanation that accounts for the behaviour of the particle will be put forward. It will be claimed that the main and basic function of *čto li* is that of a marker of a specific type of irreality, namely non-exclusion of factuality (NEF, Pietrandrea, 2012). As such, *čto li* presents the State of Affairs as just one of a set of alternatives, whose factuality is not to be excluded (Pietrandrea, 2012: 186). It will be shown how the three manifest functions of the particle, as they emerged from the corpus analysis, directly derive from this basic semantic property. Studies on particles with similar behaviour in other languages, e.g., Italian (Masini & Pietrandrea, 2010) will provide further support to our claim.

References

Auer, J. C. P. (1984). Referential problems in conversation. *Journal of Pragmatics*, *8*(5–6), 627–648.

Bazzanella, C. (2011). Indeterminacy in dialogue. *Language and Dialogue*, *1*(1), 21–43.

Bogdanova-Beglarian, N. V. (2016). Čto li v russkoj razgovornoj reči: funkcional'no-semantičeskie vozmožnosti pragmatemy. *Art-sanat*, 183–189.

Bogdanova-Beglarian, N. V., Blinova, O. V., Šerstinova, T. Y., Troščenkova, E. V., Gorbunova, D., & Zaides, K. D. (2019). Pragmatic markers of Russian everyday speech: the revised typology and corpus-based study. In *2019 25th Conference of Open Innovations Association (FRUCT)* (pp. 57-63). IEEE.

Benigni, V. (2014). Strategie di approssimazione lessicale in russo e in italiano. In O. Inkova, M. Di Filippo, & F. Esvan (Eds.), *L'architettura del testo. Studi contrastivi slavo-romanzi* (pp. 165–181). Edizioni dell'Orso.

Jucker, A. H., Smith, S. W., & Lüdge, T. (2003). Interactive aspects of vagueness in conversation. *Journal of Pragmatics*, *35*(12), 1737–1769.

Masini, F., & Pietrandrea, P. (2010). Magari. *Cognitive Linguistics*, *21*(1), 75–121.

Pietrandrea, P. (2012). The conceptual structure of irreality: A focus on non-exclusion-of-factuality as a conceptual and a linguistic category. *Language Sciences*, *34*(2), 184–199.

Podlesskaja, V. I., & Starodubceva, A. V. (2013). O grammatike sredstv vyraženija nečetkoj nominacii v živoj reči. *Voprosy Jazkosnanija*, *3*, 25–41.

Prince, E. F., Frader, J., & Bosk, C. (1982). On Hedging in Physician-Physician Discourse. In R. J. Di Pietro (Ed.), *Linguistics and the professions* (pp. 83–97). Ablex Press.

Račeva, A. A. (2018). The deictic units «zdes'» (here), «tut» (here) and «tam» (there) in oral speech: Marking of discursive processes. *Sibirskiy Filologicheskiy Zhurnal*, *65*, 257–272.

Wilson, D., & Sperber, D. (2002). Relevance Theory. In L. Horn & G. Ward (Eds.), *Handbook of Pragmatics* (pp. 607–632). Blackwell.

Žukov, A. V., & Žukov, K. A. (2003). Semantika neopredelennosti (o slovach i frazeologismach s razmytym i širokim značeniem). *Vestnkik Novgorodskogo gosudarstvennogo universiteta im. Jaroslava* Mudrogo, *25*, 106–111.

Fragments in written and spoken Present-Day English: A corpus-driven constructional account

Yolanda Fernández-Pena & Javier Pérez-Guerra

University of Santiago de Compostela; University of Vigo

Keywords: fragment; corpus-driven; Construction Grammar; parsing

The English language shows a wide variety of stand-alone constructions which, despite their reduced, non-canonical, fragmentary structure, are still semantically, discursively and pragmatically comparable to a complete clause construction, as the sentence fragments in italics in (1)-(3):

(1) Hope the summer's good / – *well done to Giles!* [= 'Say well done to Giles'] (W1B-011 #116:3)

(2) Well that's all my news. / *Regards to Simon*. [= 'Give my regards to Simon'] (W1B-006 #149:4)

(3) A: She might have been that kind of teenager anyway. / B: *Quite likely I think* [= 'That is quite likely, I think'] (S1A-031 #097:1:B)

Prior research on fragments has been mainly framed within the Generativist framework, where scholars have focused on the derivational or non-derivational mechanisms that explain their use and interpretation as full propositional sentences (e.g. Morgan, 1973; Barton, 1990; Ginzburg & Sag, 2000; Merchant, 2004; Stainton, 2006). Corpus-based/-driven analysis of these structures are limited and mainly based on spoken data (cf. Greenbaum & Nelson, 1999; Fernández & Ginzburg, 2002; Fernández et al. 2007; Bowie & Aarts, 2016), with constructional accounts being even scarcer (cf. Goldberg & Perek, 2019; Cappelle, 2021). The research reported in this paper contributes to the empirical and theoretical characterisation of fragmentary expressions in Present-Day English.

On theoretical grounds, this study advocates for a Construction-Grammar compliant account of fragments, according to which a cognitive mechanism identifies a cue, either in the construction itself or in the linguistic context, and paves the way for the conventionalised pairing of the fragment's (reduced) expression and specific interpretation not necessarily conveyed by an 'augmented' or sententialised version. As regards the empirical description of fragments, this study reports the results of a corpus analysis of sentence fragments in written and spoken discourse, based on data retrieved from the parsed version of the British component of the *International Corpus of English* (ICE-GB) (Nelson, Wallis & Aarts, 2002). Approximately 1,000 valid instances of fragments were analysed for variables such as 'category', 'structure', 'missing constituents' or 'augmentation type'. The statistical models have revealed that, although proving more pervasive in speech, fragments are not uncommon in written registers. They are more frequently found in informal written text types (e.g. letters and novels). The most frequent types in the data are verbless fragments (*Back to Cambridge tomorrow*) in written texts, clausal-finite (*If only she would admit it*) in speech, and nominal phrases (*No, no more of this conjecture*) in both registers. Most fragments show a high rate of subject and/or verb omission, most of them being functional elements or latent items that can be left unexpressed (e.g. in the latent ditransitive construction in (1) and (2) above, and in (3), whose sententialised version would require the addition of a semantically bleached subject and verb operator).

References

Barton, E. L. (1990). *Nonsentential Constituents: A Theory of Grammatical Structure and Pragmatic Interpretation*. Amsterdam and Philadelphia: John Benjamins.

Bowie, J. and Aarts, B. (2016). Clause fragments in English dialogue. In M. J. López-Couso, B. Méndez-Naya, P. Núñez-Pertejo and I. M. Palacios-Martínez (Eds.), *Corpus Linguistics on the Move: Exploring and Understanding English through Corpora* (pp. 259-288). Leiden and Boston: Brill.

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

Fernández, R. and Ginzburg, J. (2002). Non-sentential utterances in dialogue: A corpus-based study. *Traitement Automatique des Languages*, 43(2), 13-42.

Fernández, R., Ginzburg, J. and Lappin, S. (2007). Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3), 397-427. doi: 10.1162/coli.2007.33.3.397

Ginzburg, J. and Sag, I. A. (2000). *Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives.* Stanford, CA: Center for the Study of Language and Information.

Goldberg, A. E. and Perek, F. (2019). Ellipsis in Construction Grammar. In J. van Craenenbroeck and T. Temmerman (Eds.), *The Oxford Handbook of Ellipsis* (pp. 188- 204). Oxford: Oxford University Press.

Greenbaum, S. and Nelson, G. (1999). Elliptical clauses in spoken and written English. In P. Collins and D. A. Lee (Eds.), *The Clause in English: In Honour of Rodney Huddleston* (pp. 111-125). Amsterdam and Philadelphia: John Benjamins.

Merchant, J. (2004). Fragments and ellipsis. *Linguistics and Philosophy*, 27(6), 661-738. doi: 10.1007/s10988-005-7378-3

Morgan, J. L. (1973). Sentence fragments and the notion 'sentence'. In B. B. Kachru, R. B. Lees, Y. Malkiel, A. Pietrangeli and S. Saporta (Eds.), *Issues in Linguistics: Papers in Honor of Henry and Renée Kahane* (pp. 719-751). Chicago, IL: University of Illinois Press.

Nelson, G., Wallis, S. and Aarts, B. (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam and Philadelphia: John Benjamins.

Stainton, R. J. (2006). *Words and Thoughts: Subsentences, Ellipsis, and the Philosophy of Language.* Oxford: Oxford University Press.

Pleonastic verb particle constructions in Italian: a corpus-based investigation

Francesca Masini & Lucia Busso

University of Bologna; Aston University

Italo-Romance varieties are traditionally classified as verb-framed languages accordying to Talmy (2000)'s typology of motion events. That is, they typically encode the direction of motion (Path) in the verb root and not in a satellite.

However it has long been noted that Italo-Romance varieties display a complex system of phrasal verbs, which encode the Path into satellite particles (Hijazo-Gascón & Ibarretxe-Antuñano, 2013; Iacobini, 2012; Iacobini & Masini, 2009; Simone, 2008). These verbs and the constructions (in Goldberg, 2019 sense) in which they typically occur in are labelled Verb-Particle Constructions (Masini, 2005). Although VPCs are present in many Romance languages (Iacobini & Fagard, 2011), Italian presents by far the most wide use of VPCs in the lexicalization of spatial events (Iacobini, 2015).

The present paper focuses on an under-researched subset of spatial VPCs, namely constructions in which the satellite particle can be considered pleonastic with the main verb of the construction:

> *Loro non **scesero giù** dalla pianta, noi non **salimmo su***
> 'They didn't **climb down** (lit. **descend down**) from the tree, we didn't **climb up** (lit. **ascend up**)'

These presumably pleonastic constructions (henceforth: PVCP) are often mentioned as instances of double (or multiple) framing in phrasal verbs (Croft et al. 2010). The characters of PVCPs though remain vastly understudied in Romance varieties, with the exception of some studies on Spanish (González Fernández 1997; Mateu y Rigau, 2010; Arroyo Hernandez 2020).

The present study aims at providing the first corpus-based description of PVCPs in Italian, by answering the following research questions:

- Are Italian VPCPs truly redundant, and hence synonymic to their synthetic counterpart? Or are they different, independent constructions?
- And, if they qualify as different constructions, what are their formal, functional and distributional properties (also with respect to their synthetic counterparts)?

For our analysis, we selected a set of widely used intransitive VPCs which qualify as pleonastic, i.e. where the semantic content of particle and verb is redundant:

1. *uscire fuori* 'exit outside'
2. *entrare dentro* 'enter inside'
3. *salire su* 'ascend up'
4. *scendere giú* 'descend down'
5. *cadere giú* 'fall down'
6. *fuggire via* 'escape away'
7. *scappare via* 'escape away'

A corpus-based investigation is conducted using CORIS, the reference corpus for written Italian (Rossini-Favretti et al., 2002). Specifically, we extracted all concordance lines for each of the PVCP analysed, and manually annotated them according to various aspects: their meaning in context (literal

or figurative), agentivity of the subject (±agentive), presence of an explicit/accessible Ground, presence of an additional locative prepositional phrase, syntactic frame (following frames used by *LexIt*, cf. Lenci et al., 2012). We further annotate the concordances for replaceability and classify them in three conditions: completely replaceable by their synthetic counterpart (c1), replaceable with semantic change (c2), irreplaceable by the synthetic verb (c3).

The annotation process is close to completion; we expect to find ample overlap but also some degree of differentiation between PVPCs and their synthetic counterparts in terms of both meaning (possible senses being developed) and distribution (syntactic environments in which they are typical found), where the salience of the Ground seems to play a role.

References

Arroyo Hernández, I. (2020). *Subir arriba*: redundancia e interpretación de construcciones direccionales con partes axiales en español. *Artifara* 20(2), 173-187.

Croft,W., Barðdal, J., Hollman, W., Sotirova, V., Taoka, A. (2010). Revisiting Talmy's typological classification of complex events. In Boas, H. (ed.) *Contrastive construction grammar* (pp. 201-23). Amsterdam: John Benjamins.

Goldberg, A. E. (2019). *Explain Me This*. Princeton: Princeton University Press.

González Fernández, M.J. (1997) Sobre la motivación semántica de las expresiones pleonásticas de movimiento: *subir arriba, bajar abajo, entrar adentro y salir fuera*. In Company, C. (ed.) *Cambios diacrónicos en el español* (pp. 123-141). México, Universidad Nacional Autónoma.

Hijazo-Gascón, A. & Ibarretxe-Antuñano, I. (2013). Las Lenguas Románicas y la Tipología de los Eventos de Movimiento. *Romanische Forschungen* 125,4. 467-494.

Hilpert, M. (2014). *Construction Grammar and its Application to English*. Edinburgh: Edinburgh University Press.

Iacobini, C. (2009). Phrasal verbs between syntax and lexicon. *Italian Journal of Linguistics* 21. 97-118.

Iacobini C. (2015). Particle-Verbs in Romance. In Müller P.O., Ohnheiser I., Olsen S., Rainer F. (eds.) *Word-Formation* (pp. 627-659). Berlin: De Gruyter Mouton.

Iacobini, C. & Masini, F. (2009). I verbi sintagmatici dell'italiano fra innovazione e persistenza: il ruolo dei dialetti. In Cardinaletti, Anna & Munaro, Nicola (eds.), *Italiano, italiani regionali e dialetti*. Milano: Franco Angeli. 115-136.

Iacobini, C., & Fagard, B. (2011). A diachronic approach to variation and change in the typology of motion event expression. *Cahiers de Faits de langue*, *3*, 151-172.

Lenci, A., Lapesa, G., & Bonansinga, G. (2012). LexIt: A computational resource on Italian argument structure. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12).

Masini, F. (2005). Multi-word expressions between syntax and the lexicon: The case of Italian verb-particle constructions. *SKY Journal of Linguistics*, *18*(2005), 145-173.

Mateu, J. and Rigau, G. (2010). Verb-particle constructions in Romance: A lexical-syntactic account. *International Journal of Romance Linguistics,* 22(2), 241-269.

Rossini-Favretti, R., Tamburini, F., De santis, C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model, in A. Wilson, P. Rayson, T. Mcenery (eds.), *A rainbow of corpora: Corpus linguistics and the Languages of the world* (pp. 27-38). Munich: Lincom-Europa.

Simone, Raffaele 2008. Verbi sintagmatici come categoria e come costruzione. In Cini, Monica (ed.), *I verbi sintagmatici in italiano e nelle varietà dialettali. Stato dell'arte e prospettive di ricerca* (pp. 11-30). Frankfurt-am-Main: Peter Lang.

Talmy, L. (2000). *Toward a cognitive semantics*. Vol. 2. Cambridge, MA: MIT Press.

Diachronic and diatopic variation of the Middle High German bipartite negation marker *ne … niht*

Daniel Hrbek[1] & Oliver Schallert[2]

[1]Universität Osnabrück

[2]Ludwig-Maximilians-Universität Munich

In the history of German, there have been drastic changes regarding the expression of sentential negation: The clitic preverbal negation marker OHG *ni* (< PG **ni*) first underwent phonological weakening to MHG *ne/en* (1a) and then became reinforced by the postverbal negation marker MHG *niht* (< OHG *niouuiht* 'nothing') (1b). The preverbal part of the so-called discontinuous negation subsequently fell victim to complete loss (1c) so that the postverbal negation marker *niht* was sufficient to express negation on its own. This traditional scenario is known as the Jespersen's cycle (Jespersen 1917) and can be observed in many languages.

(1) a. die nemugin iz uernemen
they neg=could it hear
(Wiener Physiologus 149va,5)

b. erne wil dich nit lazen
he=neg wants you neg let
(Rolandslied 0a,8969)

c. Sí sprachen neg an dem hilígen tag
they talked NEG on the sacred day
(Oberaltaicher Evangelistar 25ba,37)

Recent studies (Jäger 2008; Pickl 2017), however, offer a more nuanced picture of this scenario: Jäger's data show that phase II – the reinforcement of the phonologically weakened form *ne* by an additional *niht* – must have been very brief, especially in Upper German. Therefore, it never was the dominant negation pattern. Pickl (2017) comes to the conclusion that phase II and III even occurred simultaneously, which contradicts a strictly progressive cyclic development. On the other hand, Schüler's (2016, 2017) data, stemming from charters from the 13th century, point to a stable phase II in Western Central German, notably Ripuarian (with Cologne as its center). Thus, the question emerges why MHG only showed such a brief phase II, contrary to other West Germanic languages such as English, Dutch or Low German where it was longer and much more stable (Breitbarth 2014).

With the aid of the *Referenzkorpus Mittelhochdeutsch* (ReM) (Klein et al. 2016), a modern balanced corpus of MHG, we conducted an in-depth analysis of negation structures in their diatopic and diachronic dimension. For comparative purposes, we rely on mappings based on the *Corpus der altdeutschen Originalurkunden* 'Corpus of Old German Original Charters' (CAO) (Wilhelm et al. 1932–2004). Unlike most other historical documents, deeds can usually be located exactly on the basis of textual evidence and thus offer a more-finely grained picture of the areal profile of the different negation strategies. In combination, data from these two corpora contribute to the still unsolved question why the bipartite MHG negation marker was lost so quickly.

Our most important findings are:

- Contrary to what has been assumed in previous research (e.g. Szczepaniak 2010), there is no change in the direction of clisis: With the exception from Bavarian sources from the 12th century, enclitic forms of *en* only rarely occur.
- During the course of the MHG period, the use of clitic negation in verb last clauses (including so-called 'Späterstellungen') is gaining ground compared to verb second clauses, thus confirming Behaghel's (1924: 84) assessment, based on a much smaller sample. In verb first clauses, clitic negation remains rare, which can be explained by prosodic factors (Hertel 2022).
- In areal terms, the bipartite negation appears quite robustly in West Central German; in all Upper German varieties, by contrast, *niht* is the predominant negation strategy by the end of the 13th century.

References

Breitbarth, A. 2014. The History of Low German Negation (= Oxford Studies in Diachronic and Historical Linguistics 13). Oxford: Oxford University Press.

Behaghel, O. 1924. Deutsche Syntax. Vol. 2: Die Wortklassen und Wortformen. B: Adverbium. C: Verbum. Heidelberg: Winter.

Jäger, A. 2008. History of German negation (Linguistik Aktuell; 118). Amsterdam: Benjamins.

Jespersen, O. 1917. Negation in English and other languages. Kopenhagen: Andr. Fred. Høst & Søn.

Klein, T., K.-P. Wegera, S. Dipper & C. Wich-Reif. 2016. Referenzkorpus Mittelhochdeutsch

(1050–1350), Version 1.0. Tech. rep. https://www.linguistics.ruhr-uni-bochum.de/rem/index.html [last accessed: 30 April 2022].

Pickl, S. 2017. Neues zur Entwicklung der Negation im Mittelhochdeutschen. Grammatikalisierung und Variation in oberdeutschen Predigten. Beiträge zur Geschichte der deutschen Sprache und Literatur 139. 1–46.

Schüler, J. 2016. Alte und neue Fragen zur mittelhochdeutschen Negationssyntax. In Augustin Speyer and Philipp Rauth (ed.), Syntax aus Saarbrücker Sicht 1. Beiträge der SaRDiS-Tagung zur Dialektsyntax, 91–107. (Zeitschrift für Dialektologie und Linguisitk – Beihefte; 165). Stuttgart: Steiner.

Schüler, J. 2017. Negationsstrukturen in den Kölner Urkunden des 13. Jahrhunderts im Vergleich. Rheinische Vierteljahresblätter 81. 1–23.

Hertel, J. [née Schüler] (2022). Zur Negationssyntax im Mittelhochdeutschen. Disseration, University of Saarbrücken.

Szczepaniak, R. 2010. Jespersen's Cycle in German from the phonological perspective of syllable and word languages. In: A. Breitbarth et al. (eds.), Continuity and change in grammar, 321–334. (Linguistics Today; 159). Amsterdam: Benjamins.

Wilhelm, F. (ed.). 1932–2004. Corpus der altdeutschen Originalurkunden bis zum Jahr 1300. Vol. 1: 1200–1282; Vol. 2: 1283–1292; Vol. 3: 1293–1296; Vol. 4: 1297–; Vol. 5: Nachtragsurkunden 1261–1297. Lahr: Schauenburg [Vol. 1–4]; Berlin: Erich Schmidt [Vol. 5].

A corpus-based study of the semantics of the Pluperfect in spoken Italian

Eleonora Morei

In this paper, I present the results of a corpus-based study on the Pluperfect tense in spoken Italian.

The Pluperfect is often described as an anaphoric tense whose function is that of locating an event prior to a reference point in the past. In fact, the Pluperfect's semantics proves to be way more complex than that: as a matter of fact, not less than three distinct anaphoric functions (*perfect-in-the-past*, *reversed result*, *past-in-the-past*) and one deictic function (*past temporal frame*) are with fair confidence associated with the Italian Pluperfect in the existing readership.

Perfect-in-the-past and *past-in-the-past* functions are described by Comrie (1976), while Italian examples of each can be found in Bertinetto (1986):

a. *Alle cinque, Clara **aveva** già **fatto** il bagno //non si sa esattamente quando, ma lo aveva fatto//.* (Perfect-in-the-past)
   at 5 o' clock Clara had already taken a bath //we don't know when, but she had done it//.
b. *Clara **aveva** già **fatto** il bagno alle cinque //esattamente a quell'ora//.* (Past-in-the-past)
   Clara had already taken a bath at 5 //exactly at that time//.

Reversed result and *past temporal frame* functions are described by Squartini (1999), who also provides Italian examples:

c. *Me lo **aveva promesso**, ma adesso fa finta di non ricordarsene.* (Reversed result)
   She promised me, but now she acts as if she didn't.
d. *Su questo argomento tanti anni fa N. ci **aveva scritto** un libro.* (Past temporal frame)
   N. wrote a book on this many years ago.

Literature also suggests that the Pluperfect could be used in spoken Italian as a deictic, generally perfective tense (Bertinetto 2003, 2014; Scarpel 2017), but this still awaits empirical confirmation. On the other hand, typological studies (Dahl 1985; Plungian and van der Auwera 2006) associate the Pluperfect with the marking of remoteness, this also being a subject to further scrutiny.

The studies cited so far are either based on a qualitative observation of the Italian language, questionnaires (Dahl 1985) or literary corpora (Bertinetto 2003, 2004). The present study attempted to investigate the forementioned functions with a corpus-base methodology.

My research moved from the following questions: to which extent do the functions *perfect-in-the-past*, *reversed result*, *past-in-the-past* and *past temporal frame* occur in authentic language samples? and, is it possible to identify any distinguishing features of the context or cotext of each? Is it possible to find instances of a Pluperfect marking remoteness or used as a generally perfective tense?

To answer these questions, I analyzed the context and cotext of 245 occurrences of Pluperfect in ParlaTO (Cerruti and Ballarè 2021), which is a corpus of spontaneous speech collected in Turin between 2018 and 2020. It is a module of the larger KIParla corpus (Mauri et al. 2019). The aim of my analysis was to classify the Pluperfect's occurrences within the four previously mentioned functions, as well as to address the cases that fell outside their scope. These four functions proved to be relevant for a description of spoken Italian: indeed, they managed to account for 239 out of 245 occurrences. Nevertheless, the 6 left out occurrences suggest the need to identify unaccounted functions of the

Pluperfect, in which the marking of shared information (evidentiality or a more general notion of knowledge management?) might be involved.

References

Bertinetto 1986: Bertinetto P.M., *Tempo, aspetto e azione nel verbo italiano. Il sistema dell'indicativo*, Firenze, Accademia della Crusca, 1986

Bertinetto 2003: Bertinetto P.M., *Tempi verbali e narrativa italiana dell'Otto/Novecento. Quattro esercizi di stilistica della lingua*, Alessandria, Edizioni dell'Orso, 2003

Bertinetto 2014: Bertinetto P.M., *Non-conventional uses of the Pluperfect in the Italian (and German) literary prose*, in Bres J., Labeau E. (eds.), *Evolution in Romance Verbal Systems*, Berne, Peter Lang, 2014, pp. 145-170

Cerruti and Ballarè 2021: Cerruti M. and Ballarè S., *ParlaTO: corpus del parlato di Torino*, «Bollettino dell'Atlante Linguistico Italiano (BALI)», 44, 2020, pp. 171-196; www.corpusparlato.com

Comrie 1976: Comrie B., *Aspect*, Cambridge, Cambridge University Press, 1976

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

Mauri et al. 2019: Mauri C., Ballarè S., Goria E., Cerruti M. and Suriano F., *KIParla corpus: a new resource for spoken Italian,* in Bernardi R., Navigli R. and Semeraro G. (eds.), *Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it*, 2019; www.kiparla.it

Plungian and van der Auwera 2006: Plungian V.A. and van der Auwera J., *Towards a typology of discontinuous past marking*, «STUF - Language Typology and Universals», 59 (4), 2006, pp. 317-349

Scarpel 2017: Scarpel S., *Considerazioni sull'uso aoristico del trapassato prossimo*, «Studia de Cultura», 9 (1), 2017, pp. 121-131

Squartini 1999: Squartini M., *On the semantics of the Pluperfect: Evidence from Germanic and Romance*, «Linguistic Typology», 3, 1999, pp. 51-89

Delineating the scope of a benefactive alternation

Piotr Wyrośląk

Paris 8 University, Adam Mickiewicz University in Poznań

pi.wyroslak@gmail.com

The present study concentrates on the problem of *delineating the scope of a constructional alternation*, understood as the process of identifying the kinds of contexts in which language users can be assumed to have an actual choice between a set of constructions viewed as near-synonymous. The lack of clarity on how to determine these contexts constitutes a major methodological problem for constructional alternation research. Arguably, a chance for advances in this area may be sought in the growing body of research indicating the importance of constructions at a low-level of generalisation for the accounts of the alternations defined at more schematic levels (cf. e.g. Glynn 2004; Pijpops et al. 2021; Romain 2021).

Building upon the insights of this research line, the present study looks at the scope of the benefactive alternation in Polish, observed between the synthetic dative case and the prepositional construction [*dla* + NP_{gen}] (ex. 1-2; cf. e.g. Buttler 1976; Dąbrowska 1997; Lesz-Duk 1988; Rudzka-Ostyn 2000).

(1) *Obiecał,* *że* ***zbuduje*** ***nam*** ***dom*** *taki,*
promise(PFV).PST.M.3SG COMPL build(PFV).NPST.3SG 1PL.DAT house.ACC.SG such.M.ACC.SG
jakiego *nikt* *jeszcze* *nie* *widział* (NKJP)
REL,M.GEN.SG nobody.NOM yet NEG see(IPFV).PST.M.3SG
'He promised that **he would build us a house** like nobody had ever seen'

(2) *Nad* *morzem* ***zbudował*** ***dla*** ***niej*** ***piękny*** ***zamek*** (NKJP)
over sea.INS.SG build(PFV).PST.M.3SG for 3SG.F.GEN beautiful.M.ACC.SG castle.ACC.SG
'On the coast, **he built a beautiful castle for her'**

The objective of the study is to refine the scope delineation for this alternation, based on the low-level, collostructional properties of *dla*-PP and dative constructions. In the course of the adopted alternation scope delineation procedure, one will try to identify the kind of use that is as uncharacteristic for either constructional variant as possible – while, at the same time, as frequent as possible for both of them.

The study uses data from a syntactically annotated corpus, a manually annotated, gold-standard component of a general corpus of Polish (National Corpus of Polish, Przepiórkowski et al. 2012). Instances of ad-verbal uses of *dla* and the dative will be retrieved, together with the data on verb forms, verb lemmas and argument heads. Obtained datasets will provide the frequency data for a battery of association analyses including common collostruction-analytic setups (Stefanowitsch 2013), as well as multidimentional exploratory techniques. The analyses will focus on identifying the collostructional environments that are attested for both alternating constructions – and with smallest frequency disproportion. The last stage will involve the critical evaluation of the obtained ranking-format results in terms of the prevalence of the semantic shifts undetected by the analyses. This evaluation stage will be facilitated with Walenty, (Przepiórkowski et al. 2014), a valency database for Polish, used as the reference resource.

References

Buttler, Danuta. 1976. *Innowacje Składniowe Współczesnej Polszczyzny*. Warszawa: Państwowe Wydawnictwo Naukowe.

Dąbrowska, Ewa. 1997. *Cognitive Semantics and the Polish Dative*. Vol. 9. Cognitive Linguistics Research. Berlin-New York: De Gruyter.

Glynn, Dylan. 2004. 'Constructions at the Crossroads: The Place of Construction Grammar between Field and Frame'. *Annual Review of Cognitive Linguistics*. John Benjamins.

Lesz-Duk, Maria. 1988. *Czasowniki o Składni Przyimkowej w Języku Polskim*. Częstochowa: Wydawnictwo Wyższej Szkoły Pedagogicznej w Częstochowie.

Pijpops, Dirk, Dirk Speelman, Freek Van de Velde, and Stefan Grondelaers. 2021. 'Incorporating the Multi-Level Nature of the Constructicon into Hypothesis Testing'. *Cognitive Linguistics*.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, eds. 2012. *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo naukowe PWN.

Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. 2014. 'Walenty: Towards a Comprehensive Valence Dictionary of Polish'. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2785–92. Reykjavík, Iceland.

Romain, Laurence. 2021. 'Putting the Argument Back into Argument Structure Constructions'. *Cognitive Linguistics*. https://doi.org/10.1515/cog-2021-0021.

Rudzka-Ostyn, Brygida. 2000. 'Celownik w języku polskim'. In *Z rozważań nad kategorią przypadka*, 97–132. Kraków: Universitas.

Stefanowitsch, Anatol. 2013. 'Collostructional Analysis'. In *The Oxford Handbook of Construction Grammar*, edited by Thomas Hoffmann and Graeme Trousdale. Oxford: Oxford University Press.

Latvian deverbal nouns in *-ien-* and *-um-* and derivational productivity: a corpus-based analysis

Andra Kalnača, Daiga Deksne & Tatjana Pakalne

University of Latvia, Rīga

Keywords: verbs, deverbal nouns, suffixes, productivity

Latvian has a rich system of productive word formation featuring, primarily, suffixation (e.g., Nītiņa, Grigorjevs 2013). Suffixal nominalizations are one of the most productive category-changing non-transpositional derivational patterns in contemporary Latvian (e.g., Nau 2016; on nominalizations from a typological perspective see Alexiadou 2017). However, not all derivational patterns are equally productive (e.g., Aronoff, Lindsay 2017; Lieber, Štekauer 2017).

The current study aims at exploring the factors which influence the productivity of two deverbal derivational patterns:

(1) V_{PST} + *-ien-* → N
lēkt 'to jump' – *lēca* (PST.3) – *lēc-**ien**-s* 'jump'
(2) V_{PST} + *-um-* → N
lem-t 'to decide' – *lēma* (PST.3) – *lēm-**um**-s* 'decision'

In both patterns, nominal suffixes are added to the past-tense stem (or, for most -ien- derivatives derived from 3rd conjugation verbs, to the present-tense stem) of the verb to derive *nomina actionis* (masculine). Derivatives with the suffix *-ien-* express instantaneous single actions (1), single actions lasting for some time (3a), and abstract (3b) or concrete (3c) things. Derivatives in *-um-*, on the contrary, are mainly result nominalizations (2), abstract (4a) or concrete (4b) things, and sometimes also event/process nominalizations (4c). Thus, on the whole, these derivational patterns demonstrate the opposition between event and result nominalizations (typologically, see Roy, Soare 2013; Alexiadou 2017). Nevertheless, the *-um-* pattern is much more productive than the *-ien-* pattern.

(3) a. *nāc-ien-s* 'coming'
b. *cēl-ien-s* 'act (in theatre)'
c. *dzēr-ien-s* 'drink'
(4) a. *lik-um-s* 'law'
b. *aud-um-s* 'fabric'
c. *lasīj-um-s* 'reading'

To identify the factors constraining and/or enhancing the productivity of these derivational patterns, we excerpted 1) the infinitive forms of all verbs, 2) the base forms of all deverbal derivatives with the suffixes *-ien-* and *-um-* from "The Balanced Corpus of Modern Latvian". The derivatives were then matched with their base verbs. This resulted in three separate lists: 1) verbs with only *-ien-* derivatives (185 items), 2) verbs with only *-um-* derivatives (1164 items), 3) verbs with both types of derivatives, e.g., *dzēr-ien-s* 'drink' and *dzēr-um-s* 'drunkenness' (52 items). Further analysis revealed the following:

1. ***-ien-* derivatives only**

37% of base verbs are verbs of sound, and the derived deverbal nouns express a short action (1). 80% of base verbs are primary/simple (with no suffix in the infinitive), 20% are suffixal; 39% of the total number of verbs are transitive.

2. ***-um-* derivatives only**

34.2% of base verbs are primary/simple, while the remaining 65.8% are suffixal. Besides that, 76% of verbs are transitive.

3. ***-ien-* and *-um-* derivatives**

92.16% of base verbs are primary/simple, 7.84% are suffixal; 71% are transitive. This group contains verbs of physical action, motion and speaking. Deverbal nouns with *-um-* are more frequent as their average frequency is 399, which is three times higher than the average frequency of *-ien-* derivatives (124).

Thus, at least three clusters of factors that could affect the productivity of deverbal derivational patterns with *-ien-* and *-um-* have been identified: 1) morphological structure of base verbs, 2) semantics of base verbs, 3) possibly, to an extent, argument structure.

References

Alexiadou, A. 2017. Nominal Derivation. *The Oxford Handbook of Derivational Morphology*. Lieber, R., Štekauer, P. (eds.). Oxford: Oxford University Press, 235–256.

Aronoff, M., Lindsay, M. 2017. Productivity, Blocking, and Lexicalization. *The Oxford Handbook of Derivational Morphology*. Lieber, R., Štekauer, P. (eds.). Oxford: Oxford University Press, 67–83.

Lieber, R., Štekauer, P. 2017. Universals in Derivation. *The Oxford Handbook of Derivational Morphology*. Lieber, R., Štekauer, P. (eds.). Oxford: Oxford University Press, 777–786.

Nau, N. 2016. Argument realization in Latvian action nominal constructions: a corpus and text based investigation. Holvoet, A., Nau, N. (eds.). Argument realization in Baltic. John Benjamins, 461–522.

Nītiņa, D., Grigorjevs, J. (eds.). 2013. *Latviešu valodas gramatika.* [A Grammar of Latvian.] Rīga: Latvijas Universitātes Latviešu valodas institūts.

Roy, I., Soare, E. 2013. Event related nominalizations. *Categorization and Category Change*. Iordăchioaia, G., Roy, I. & K. Takamine (eds.). Cambridge Scholars Publishing, 123–152.

Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss 2018 (LVK2018) [The Balanced Corpus of Modern Latvian 2018 (LVK2018)]. Available at: http://www.korpuss.lv/id/LVK2018

The morphology of Polish imperfective future tense

Rafał L. Górski

Jagiellonian University & Institute of Polish Language, PAS

Polish imperfective future tense has two different forms: it consists of an auxiliary with either a participle or an infinitive. It is widely assumed that the two forms are perfectly synonymous, however, the choice of one of the forms is not fully random (cf. Dunaj 1987, Górecka & Śmiech 1972, Łaziński 2006 for some partial observations). There are several – often mutually exclusive – principles, which govern it. Still, we cannot speak of rules but rather strong tendencies. Sentences which violate them are still acceptable, also the corpus shows a number of exceptions. The aim of the presentation is to uncover the pattern of the usage of the both morphological variants of the imperfective future in spoken Polish. The data are drawn from the Spokes corpus (Pęzik 2015). Since the corpus contains no POS-tagging, the concordance was lemmatised and tagged with a morphological analyser Morfeusz.

The principles can be ordered from most general to most specific. The principles read as follows:

1. Do not lose information.

The participle, contrary to infinitive, bears the gender marker (this is also true for number, which is however also coded by the auxiliary). The grammatical gender helps to identify the subject (note that Polish is a Pro-drop language), especially in the third person. In fact the participles are much more frequent than infinitives

participle	infinitive
3985	1864

2. Be sparse, avoid longer forms.

The choice of a participle is not without a cost – the masculine is marked by a null morpheme and is exactly as long as the infinitive, but for feminine and neuter the participle is one syllable longer than the infinitive. The same holds for plural, regardless of the gender. Thus, if the form with a participle is longer than its infinitival counterpart, the infinitive is frequent.

overall number of participial futures and the number of syllables in infinitive and participle

infinitive = participle	infinitive < participle
1888	2093

The above principle has three exceptions:

3a. If the infinitive consists of a no more than two syllables, avoid infinitives.

	infinitive	participle
≤2 syllables	1164	3244
>2 syllables	700	741

3b. Control verbs should not be used in infinitive.

This principle seems to have a double motivation. It can be explained by euphony, namely it bans a sequence of two infinitives, where both verbs bare the same marker *-ć* ([t͡ɕ]). Meanwhile, in such a sequence the contrast between the control and the embedded verb is blurred.

	infinitive	participle
Control	204	1817
Non-control	1660	2168

Now, most control verbs are short, so one can ask whether the principle 3b is an epiphenomenon or it is the syntactic status of the verb rather than the number of syllables which plays a role here. The answer is that both phenomena decide about the form. Bi- and trisyllabic control verbs tend to take participles but the same is true for monosyllabic and bi-syllabic non-control verbs. The differences in 3a and 3b are statistically significant ($p > 0.05$)

3c. Also, frequent verbs tend to take the participle rather than infinitive.

Finally, one can ask whether the sex of the speaker plays a role here. Note, that in first person singular males use almost exclusively the participial future (principle 2 does not apply to this form), while women use both forms, therefore both sexes are exposed to the two patterns with different frequency. The answer however is unclear. Whereas in the plural men tend to use the participle more frequent than women ($p < 0.05$), in 3 person singular the difference between the two sexes shows no statistical significance.

This interplay of the above outlined tendencies towards making exceptions (and exceptions from exceptions) are be modelled with decision trees. The discriminators are: the number of syllables, frequency of the verb, its status in terms of control, as well as the gender of the subject. The accuracy based on tenfold cross-validation shows performance above the baseline.

References

Dunaj, B. 1987. Czas przyszły czasowników niedokonanych w polszczyźnie–uzus i norma. *Język Polski*, *67*, 9-19.

Górecka, J., Śmiech, W. 1972. Czas przyszły złożony w języku polskim. *Rozprawy Komisji Językowej Łódzkiego Towarzystwa Naukowego*, *18*, 11-38.

Kieraś, W., & Woliński, M. (2017). Morfeusz 2–analizator i generator fleksyjny dla języka polskiego. *Język Polski, 97*, 75-83.

Łaziński, M. 2006. *O panach i paniach. Polskie rzeczowniki tytularne i ich asymetria rodzajowo-płciowa.* Warszawa: Wydawnictwo Naukowe PWN

Nitsch, K. (1956). Tajemnice polskiego czasu przyszłego złożonego. *Język polski*, *36*, 190-196.

Pęzik, P. 2015. Spokes – a Search and Exploration Service for Conversational Corpus Data. In *Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands*,. Linköping Electronic Conference Proceedings. Linköping University Electronic Press, Linköpings universitet, 99–109.

Soft morphophonological constraints on Hebrew genitive choice

Gabi Danon

Bar-Ilan University

Modern Hebrew is known to have 3 kinds of genitive constructions. Many aspects of the alternation between these constructions remain poorly understood - including both 'hard' constraints on the grammaticality of a given construction and 'soft' constraints which may disfavor a certain option without rendering it ungrammatical.

This talk focuses on the role of morphophonological factors. Siloni (2003) argues that genitive case assignment in the preposition-less genitive known as the Construct State (CS) is sensitive to prosodic constraints (see also Ouhalla 2009 for similar cases in Spanish Arabic). While that work focuses on categorical consequences of the prosody-based analysis, it also notes in passing that CS-formation when headed by nouns whose surface form in the CS is identical to their free form, as in (1), sometimes leads to reduced acceptability compared to CS headed by alternating nouns as in (2). Perhaps due to their subtle nature and the ease of coming up with counterexamples like (3), these facts have never been addressed directly; this study aims to look specifically at such cases. We aim to show that usage patterns display clear sensitivity to morphophonological properties that are unexpected on syntactic or semantic grounds.

This study uses a sample from the dependency parsed Hebrew Wikipedia treebank (Goldberg 2014). A sample of over 100,000 genitives was extracted and automatically annotated for a variety of lexical, grammatical, and morphological factors. Using this sample we checked for association between different factors and the type of genitive.

To test the hypothesis that CS with overt marking is sometimes more acceptable than CS with no such marking, we exploited the fact that nouns ending with *-a*, the most common feminine suffix, usually alternate between *-a* in the free form and *-at* in the CS (4), while their plural is most commonly the non-alternating *-ot* (5). We thus compared the distribution of the 3 genitive types headed by such nouns, in both singular and plural, to that of other nouns, many of which alternate only in the plural (6)-(7). The results seem to confirm Siloni's observation: The proportion of CS in the alternating cases is higher than in the non-alternating ones. Crucially, while this effect is statistically highly significant, the CS is still unquestionably the most frequent genitive type even in non-alternating cases.

Similar patterns are seen with other comparisons performed, such as between derived nominals where the head noun belongs to an alternating morphological template to ones belonging to a non-alternating template.

Finally, we also checked another morphophonological factor which might affect prosodic structure: word length. Once again, we found a significant association, with longer head nouns being less frequent in the CS than shorter ones.

Overall, the corpus findings provide robust evidence for the role of weak, violable, non-blocking morphophonological constraints which are often barely noticeable in introspective judgments. Whether these follow from the grammar of genitive formation or are factors that merely affect usage preferences remains open. Either way, identifying these patterns has important consequences for teasing apart the various factors involved in genitive choice and avoiding confounds when examining the role of other factors.

Examples

(1) ? me'il ha-yeled
coat(m.s.cs) the-boy
'the boy's coat' (non-alternating head noun)

(2) xulcat ha-yeled
shirt(f.s.cs) the-boy
'the boy's shirt' (alternating head noun)

(3) merkaz ha-xeder
center(m.s.cs) the-room
'the center of the room' (non-alternating head noun)

(4) a. xulca
shirt(f.s.free)
b. xulcat
shirt(f.s.cs) (alternating in the singular)

(5) a. xulcot
shirts(f.p.free)
b. xulcot
shirts(f.p.cs) (non-alternating in the plural)

(6) a. me'il
coat(m.s.free)
b. me'il
coat(m.s.cs) (non-alternating in the singular)

(7) a. me'ilim
coats(m.p.free)
b. me'iley
coats(m.p.cs) (alternating in the plural)

References

Goldberg, Yoav. 2014. *The Hebrew Wikipedia Dependency Parsed Corpus* ver. 1.0.

Ouhalla, Jamal. 2009. Variation and Change in Possessive noun Phrases: The Evolution of the Analytic Type and Loss of the Synthetic Type. *Brill's Journal of Afroasiatic Languages and Linguistics* 1(1). 311–337.

Siloni, Tal. 2003. Prosodic case checking domain: The case of constructs. In Jacqueline Lecarme (ed.), *Research in Afroasiatic Grammar* II, 481–510. John Benjamins.

A Corpus-based Perspective on "Split Stimuli" in German

Johanna M. Poppek, Simon Masloch & Tibor Kiss

Linguistic Data Science Lab, Ruhr-Universität Bochum

{johanna.poppek, simon.masloch, tibor.kiss}@rub.de

Introduction and Database

Stimulus subject (STM-SUBJ) verbs are psychological predicates that realise the stimulus argument as their subject on the canonical transitive pattern. In some sentences though, a part of the semantic stimulus is expressed by a phrase external to the subject, often a PP: In (1) it is not only Oscar, but more specifically his riding style that causes the fans' emotion. Such cases are referred to as "split stimuli" in the literature (e.g. Engelberg, 2015; Hirsch, 2018; Klimek & Rozwadowska, 2004; Temme, 2018).

(1) **[$_{PP}$Mit einer angriffigen Fahrt] hat Oscar […] die Fans […] entzückt.**
with a attacking ride has Oscar the fans delighted
'Oscar delighted the fans with an attacking ride.' (NZZ_1997_06_26_a187_seg5_s1)

While supposed properties of the construction have been used as tests in theoretical literature (e.g. Hirsch, 2018), the only corpus study on this phenomenon in German we are aware of is (Engelberg, 2015), which is based on a comparatively small number of verbs. We use data originating from a larger body of annotations regarding the general distributional, lexicogrammatical, and semantic properties of STM-SUBJ verbs, *GerEO* (Masloch et al., 2021; Poppek et al., 2022). GerEO contains samples for 64 STM-SUBJ verbs (predominantly emotion verbs, more than 10,000 annotated sentences) from a corpus created from the 1993–1999 volumes of the *Neue Zürcher Zeitung*.[3] Examples are annotated by – among others – syntactic pattern, stimulus type, and a presumed stimulus PP, i.e. a PP that can be considered to express (part of) a stimulus.

Results

We find that additional stimulus PPs are by no means rare; however, verbs differ largely in how frequently they cooccur with an additional stimulus PP (and which PP) in the relevant constructions, see Figure 1. The data also revealed that some prepositions possess a reading that allows them to act as a PP expressing part of a stimulus, but which are not routinely listed as such in the literature, including *als* 'as' (as in (2)). Some claims from the literature have to be dismissed based on our data: E.g., Temme (2018, pp. 130–132) proposes that a split is not possible with factive stimuli. While prepositions differ with respect to the kinds of stimulus subject they cooccur with (inanimate NP, animate NP, clausal, …) and "split stimuli" with clausal stimulus subjects are rare, there are counterexamples to Temme's claim like (3), (4). It strikes us as particularly interesting that verbs selecting a dative experiencer object like *imponieren* 'to impress' and *gefallen* 'to like' frequently license stimulus PPs with *mit* 'with' and *durch* 'through': This contradicts Hirsch's (2018) claim that stimulus PPs headed by these prepositions are compatible with non-stative verbs only since STM-SUB verbs taking a dative experiencer are usually considered stative in the literature (see Landau, 2010).

[3] Since this is a written-language corpus in Standard German only, it may reflect some domain-specificissues, however, regional variation is annotated separately.

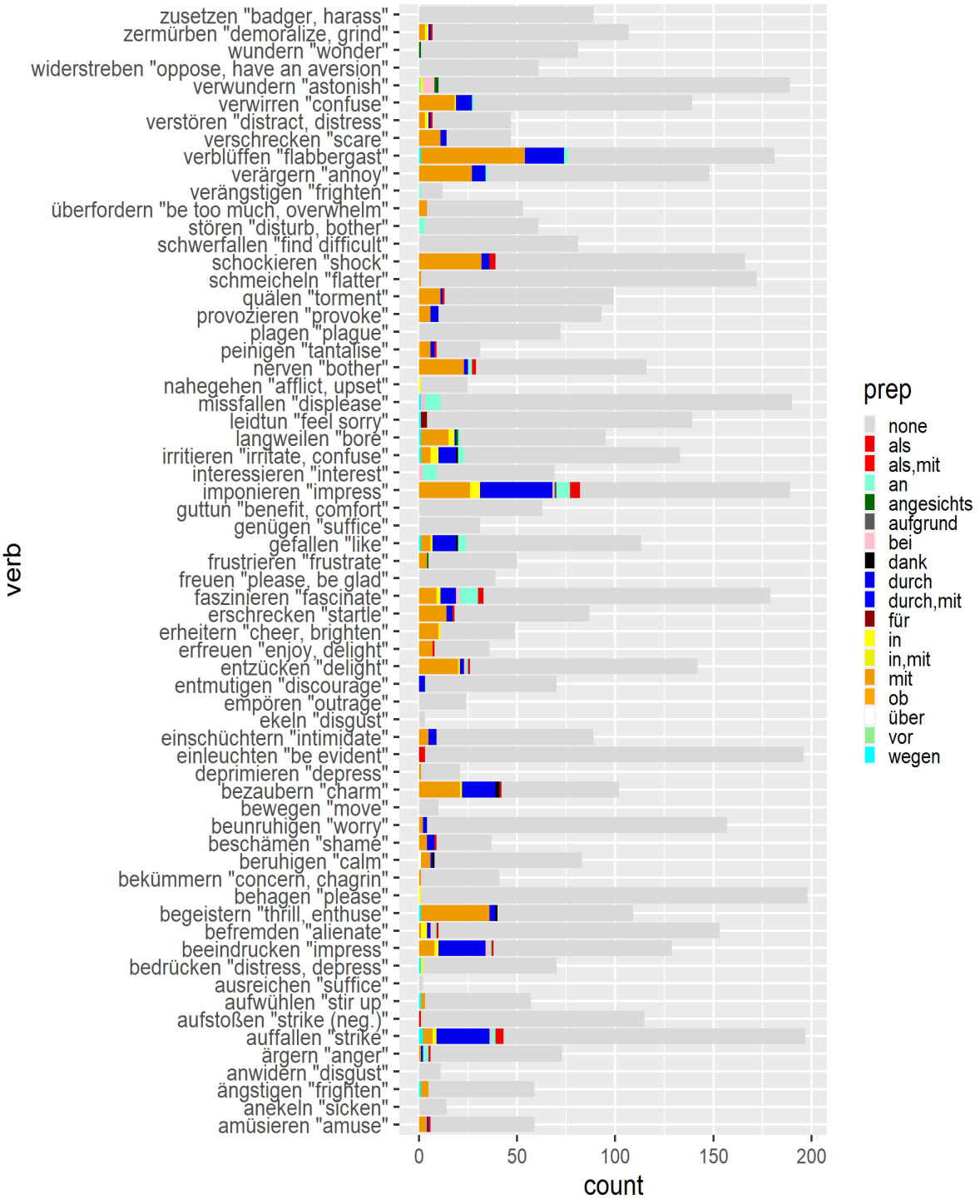
(2) **Die Verteidiger […] imponierten als «Abräumer» an der Bande […]**
the defenders impressed as enforcers at the boards
'The defenders impressed as "enforcers" on the boards.' (NZZ_1997_06_26_a187_seg5_s1)

(3) **Was mich am Lötschberg […] ärgert, ist, dass man sich der Gefahr**
what me at.the Lötschberg annoys is that one Refl the.gen danger.gen
des noch grösser werdenden Röstigraben bedient, um uns
the.gen even bigger becoming Röstigraben.gen employs in.order.to us.dat
den Lötschberg schmackhaft zu machen […]
the.acc Lötschberg.acc tasty to make
'What annoys me about the Lötschberg (tunnel) is that they are using the danger of an ever-growing Röstigraben to make the Lötschberg palatable to us.'
(NZZ_1997_06_26_a187_seg5_s1)

(4) **Dass diese schöne Geschichte geradlinig und einfach erzählt sei,**
That this beautiful story linear and simple told is.sbj
fasziniere ihn an der Urfassung, sagt der Regisseur.
fascinates.sbj him at the original.version says the director
'That this beautiful story is told without frills and in a simple way is what fascinates him about the original version, the director says.' (NZZ_1995_02_07_a138_seg7_s1

From a larger corpus-based perspective, the phenomenon of "split stimuli" appears more complex than previously assumed and it expands beyond claims previously made in the literature, e.g. regarding assumptions like factiveness of the stimulus as well as prepositions licensing it.

Figure 1: Count of stimulus prepositions on transitive (stimulus subject and experiencer object present) or intransitive (no overt experiencer object) pattern and psych reading for each verb



References

Engelberg, S. (2015). Gespaltene Stimuli bei Psych-Verben: Kombinatorische Mustersuche in Korpora zur Ermittlung von Argumentstrukturverteilungen. In S. Engelberg, M. Meliss, K. Proost, & E. Winkler (Eds.), *Argumentstruktur zwischen Valenz und Konstruktion* (pp. 469–492). Narr Francke Attempto.

Hirsch, N. (2018). *German psych verbs – insights from a decompositional perspective* (PhD Thesis). Humboldt-Universität zu Berlin.

Klimek, D., & Rozwadowska, B. (2004). From psych adjectives to psych verbs. *Poznań Studies in Contemporary Linguistics*, *39*, 59–72.

Landau, I. (2010). *The Locative Syntax of Experiencers*. MIT Press.

Masloch, S., Poppek, J. M., Robrecht, A., & Kiss, T. (2021). Syntactic pattern distribution analysis of experiencer-object psych verbs: An Annotation Manual [https://ldsl.rub.de/media/pages/research/resources/sllds/a3914a6ac2- 1640260430/sllds_4.pdf]. *Studies in Linguistics and Linguistic Data Science*, *4*.

Poppek, J. M., Masloch, S., & Kiss, T. (2022). *GerEO: A Large-Scale Resource on the Syntactic Distribution of German Experiencer Object Verbs* [to appear in LREC 2022 Proceedings].

Temme, A. (2018). *The peculiar nature of psych verbs and experiencer object structures* (PhD Thesis). Humboldt-Universität zu Berlin.

Article use in English: construal and constraints

Laurence Romain, Petar Milin & Dagmar Divjak

Keywords: articles, learning, corpus, computational modelling, experimental data

The English article system, while seemingly simple, remains difficult to grasp for learners, especially for those whose language does not use articles. Grammars for learners tend to offer an account that uses two main features: Hearer Knowledge (HK) and Specificity of the Referent (SR) as guiding principles for the choice of article (cf. Huebner's semantic wheel 1983, 1985) but many blur the line between the two features. In this paper, we show that these semantic features partly guide the choice of grammatical structure (i.e., articles) and we explore the use of articles in English in more depth through an analysis of both corpus and experimental data. We will see that while Hearer Knowledge comes out as the most crucial cue for article choice in corpus data, Specificity of the Referent plays a pivotal role in delimiting contexts and potentially constraining article choice in experimental data, with HK being more open to construal.

For the corpus study, we extracted 2000 contexts from the BNC and annotated them for Hearer Knowledge, Specificity of the Referent but also countability, number and elaboration. We then fed this data to a decision tree analysis and a simple learning algorithm (Widrow-Hoff, 1960). Both analyses revealed the predominance of HK as a crucial variable for the choice between the definite article *the* and the indefinite *a/an* but turned out to be less informative for the zero article.

Then, for the experimental study, we ran an online survey in a 3 alternative forced-choice format (3AFC) where participants had to choose between *the, a/an* or *Ø*. To obtain the 3AFC data, we presented 180 native speakers of English with 4 out of 12 journalistic texts from which we removed articles. Despite the fact that most participants retrieved the original article in about 80% of contexts, a number of contexts seemed more prone to variation, judging by the divergence in our participants' responses. We therefore investigated which types of contexts yielded more variation by quantifying construal (i.e., how open to interpretation a context is) via the measure of Entropy. This measure allowed us to identify certain features that tend to constrain the choice of article and others that make it more open to interpretation. We find, for example, that contexts in which the referent was originally annotated as non-specific (SR-) are more open to construal and thus allow the use of different articles. On the other hand, SR+ contexts (i.e., with a specific referent) appear to be more constrained. HK is not particularly prone to high or low Entropy and as such seems more open to interpretation.

Overall, through the combination of corpus and experimental data, we find that the article system relies mostly on a feature that is both crucial and open to interpretation. While this does not seem to pose a problem for native speakers, L2 learners might find this construal-based variation more difficult to grasp.

References

Huebner, T. (1983). *A longitudinal analysis of the acquisition of English*. Ann Arbor: Karoma.

Huebner, T. (1985). System and variability in interlanguage syntax. *Language Learning*, 35(2), 141-163.

Widrow, B. & Hoff, M. E. (1960). *Adaptive switching circuits*. Paper presented at the WESON Convention Record Part IV.

Tracking verb changes in a corpus of non-printed manuscript materials

David Denison & Tino Oudesluijs

University of Manchester

This paper explores (1) the methodology of constructing a manuscript-based corpus of historical materials, and (2) developments in the English auxiliary system in the late 18th and early 19th centuries, focusing on the verb *be* and those other verbs (principally *have* and full modals) that it commutes with in its various uses.

The first topic arises from the fact that many historical corpora containing printed, copy-edited materials (e.g. EEBO, CEEC, PPCMBE2, COHA) are prone to deliberate or unconscious standardisation (see e.g. Kytö & Pahta 2012: 126). Our main corpus of 800k words consists of carefully edited transcriptions of a manuscript collection of personal correspondence and diaries running from c.1760 to c.1820, with detailed levels of TEI mark-up, and so is more faithful to original language use than one taken from copy-edited publications. Even in this modest-sized corpus there are more than 28k examples of the be lexeme. The corpus has both a normalised transcription, used for POS-tagging, and a diplomatic one, from which hits from the CQPweb query engine are displayed for analysis. In addition to overall chronological change, the corpus can show apparent-time change using birth dates of participants, and real-time change, even including change across the lifespan. For comparison across geographical region, social class and topic, we will use a control corpus of 300k words of letters by mainly untutored writers from the same period, also edited carefully from unpublished letters.

The linguistic exploration concerns the half-century or so from 1760, a period which covers grammaticalisation of the progressive, first appearances of the progressive passive and progressive of *be* itself, further encroachment of *have* as auxiliary of the perfect with mutative and motion verbs, loss of non-finite *be to* semi-modal, changes in negative imperatives, and morphological developments in 2nd-person and in contracted forms (see e.g. Warner 1993 etc., Denison 1998, Nevalainen, Palander-Collin & Säily 2018, Sag et al. 2020). The aim is to gain a better understanding of diachronic grammar from detailed micro-histories of verb change, as found in a number of intersecting social networks that operated at a crucial period in the history of English. The paper will present two case studies of such micro-histories, namely pronoun + (un)contracted verb (*I have* vs *I've*), and *you was* vs *you were*.

References

Denison, David. 1998. Syntax. In Suzanne Romaine (ed.), *The Cambridge history of the English language*, vol. 4, *1776-1997*, 92-329. Cambridge: CUP.

Kytö, Merja & Päivi Pahta. 2012. Evidence from historical corpora up to the twentieth century. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford handbook of the history of English*, 123-33. New York: OUP.

Nevalainen, Terttu, Minna Palander-Collin & Tanja Säily (eds.). 2018. *Patterns of change in 18th-century English: A sociolinguistic approach*. Amsterdam: John Benjamins.

Sag, Ivan A., Rui P. Chaves, Anne Abeillé, Bruno Estigarribia, Dan Flickinger, Paul Kay, . . . Thomas Wasow. 2020. Lessons from the English auxiliary system. *Journal of Linguistics* 56, 87-155, 227-8.

Warner, Anthony R. 1993. *English auxiliaries: Structure and history*. Cambridge, etc: CUP.

Aspectual symmetry of correlative coordination in Mandarin Chinese

Ting-Shiu Lin

Université Lumière Lyon 2

It is generally held that conjuncts in an accidental coordination must resemble each other in their semantic type for the sentence to be acceptable (e.g., Munn, 1993; Zhang, 2010). However, this requirement is not sufficient to account for the (un)grammaticality of sentences (1-a) and (1-b). In these sentences, a correlative coordinator *yòu*…*yòu*… or *yě*…*yě*… links two verbal phrases, both of which depict a student's activity in class. Nevertheless, (1-a) is grammatical while (1-b) is not.

(1)	a.	*Xuésheng*	*zài*	*jiàoshì-lǐ*	*yòu/yě*	*tīng kè*	*yòu/yě*	*xiě bǐjì*.
		student	at	classroom-inside	YOU/YE	listen course	YOU/YE	write note
		'Students both listen to the lecture and take notes in the classroom.'						
	b.	??*Xuésheng*	*zài*	*jiàoshì-lǐ*	*yòu/yě*	*tīng kè*	*yòu/yě*	*xiě-xià bǐjì*.
		student	at	classroom-inside	YOU/YE	listen course	YOU/YE	write-down note

This paper hypothesizes that not only semantic resemblance but also aspectual symmetry of conjuncts affects the degree of acceptability of a coordinate structure formed by *yòu*…*yòu*… or *yě*…*yě*… In sentence (1-a), both conjuncts are activity predicates, while in (1-b), an activity predicate is linked to an achievement one (cf., Xiao & McEnery, 2004). Data collected from the corpus of Center for Chinese Linguistics of Peking University (CCL corpus) are in support of this hypothesis. Among 11219 sentences containing *yòu*…*yòu*… and 2123 sentences containing *yě*…*yě*… found in this corpus, more than 99% of them consist of conjuncts that have identical aspectual features.

Furthermore, it appears that whether coordinated predicates should have the same aspectual features is related to the semantic relationship that a coordinator builds between its conjuncts. In Mandarin, coordinators can be classified into at least two types: (i) those that mark a temporal progression or an increase in contextual importance between their conjuncts, such as *bìng*, *bìngqiě* and *érqiě*; (ii) those whose conjuncts are considered equally important in the context and are atemporal or depict events that happen in the same time frame, such as *hé*, *yòu*…*yòu*… and *yě*…*yě*… Only the coordinators of the second type must have conjuncts with identical aspectual features to improve the acceptability of the sentence (2-a vs. 2-b).

(2)	a.	*Xuésheng*	*zài*	*jiàoshì-lǐ*	*tīng kè*	*bìng/ bìngqiě/ érqiě*	*xiě*(*-xià*)	*bǐjì*.
		student	at	classroom-inside	listen course	BING/BINGQIE/ERQIE	write(-down)	note
		'Students listen to the lecture and take notes/write down some notes in the classroom.'						
	b.	*Xuésheng*	*zài*	*jiàoshì-lǐ*	*tīng kè*	*hé*	*xiě*(??*-xià*)	*bǐjì*.
		student	at	classroom-inside	listen course	HE	write-(??down)	note
		'Students listen to the lecture and take notes in the classroom.'						

This paper proposes that aspectual symmetry between conjuncts should be considered as a component of Zhang's (2010) Relativized Parallelism Requirement (RPR). According to Zhang (2010), RPR is a filter on representations of syntactic complexes which rules out sentences that are too difficult to process. The coexistence of temporal parallelism and aspectual symmetry between conjuncts may facilitate information processing and thus makes the sentence more acceptable. It is worth noting that in the data collected from the CCL corpus, the conjuncts having different aspectual features are all

similar in both morphological construction and number of syllables. This suggests that aspectual asymmetry may be rescued by strict symmetry in other linguistic aspects, which is compatible with the hypothesis that the requirement of aspectual agreement between conjuncts is to facilitate information processing.

Selected references

Munn, A. B. (1993). *Topics in the syntax and semantics of coordinate structures* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Xiao, R., & McEnery, T. (2004). *Aspect in Mandarin Chinese: A corpus-based study*. Amsterdam, the Netherlands: John Benjamins Publishing Company.

Zhang, N. Ning (2010). *Coordination in Syntax*. Cambridge: Cambridge University Press.

Comparing contextual factors of the eight two-character modal auxiliaries through the lens of modality

Zhuo Zhang & Meichun Liu

Department of Linguistics and Translation, City University of Hong Kong

Keywords: Chinese modality, two-character modal auxiliaries, Behavioral Profiling approach, 3-grams, cluster analysis

This study focuses on eight two-character modal auxiliaries, including kěnéng(**可能**, *keneng*, 'possible'), kěyǐ(**可以**, *keyi*, 'can'), yídìng(**一定**, *yiding*, 'must'), yīngdāng (应当, *yingdang*, 'should'), yīnggāi (应该, *yinggai*, 'should') bìxū (**必**须, *bixu*, 'must'), nénggòu (**能**够, *nennggou*, 'can') and yuànyì (**必然**, *biran*, 'must be'), and aims to study their distributional (dis)similarities in the web corpus.

Research question(s)

1) How are these modal auxiliaries (dis)similar to each other in sentential contexts?
2) Does modality contribute to the formation of clusters? If yes, what are the more significant features? If not, why?

Hypothesis

Although most modal auxiliaries are polysemous, they mostly have a dominant core sense (Yang, 2017: 21). Based on the form-meaning mapping principle, the contextual features of the eight auxiliaries may form several clusters with a respective focus on modality.

Method

The eight modals display strong preferences in collocations, so the top 20 3-grams of each modal are analyzed to unveil their similarities and differences. Next, we adopted the Behavioral Profiling approach (Gries & Divjak, 2009: 57-75), and the auxiliaries are considered as near-synonyms. The idea is to find and annotate the sentential features of modal verbs, convert the categorical data into vector tables, and evaluate the table by statistical analysis. The potential set of features can include part-of-speech tagging, subject types, and collocational features associated with specific modifiers. In this study, we tried to use as many contextual features as possible, reduce the features with low variance and then conduct hierarchical clustering and report the clustering solution with the highest Silhouette width.

Data

We build a corpus of 8*9,000 sentences with about 4 million characters. We first downloaded eight 10,000 random sentences containing the modal auxiliary from the *Chinese Simplified Web 2017 sample* via the corpus tool Sketch Engine (Kilgarriff et al., 2014). After data cleaning, 9,000 sentences were randomly extracted to ensure a similar number of sentences for each modal. The annotation is fully rule-based with the help of NLP toolkits (Manning et al., 2014).

Results & Discussion

Through analysis of 3-grams, *keneng*, *biran*, and *yiding* are more likely to indicate epistemic meaning. In contrast, *yinggai*, *yingdang*, and *bixu* tend to be used to express the deontic sense, and *keyi* and

nenggou are inclined to dynamic modality. The vector table forms three clusters with AU (Approximately Unbiased) values > 0.95 and Silhouette width=0.55 > 0.2, with a respective focus on epistemic, deontic, and dynamic modality, which indicates the core semantics of the modal auxiliaries in the corpus. Based on the clustering solution of the eight modals, the three-way modality can be applied to analyze two-character modal auxiliaries in a web corpus with a general field.

Conclusion

As the first BP study on Chinese near-synonyms, the study draws implications on quantitative explorations of modal auxiliaries through 3-grams and cluster analysis and finds that modality type is a crucial factor contributing to the (dis)similarities of the two-character modal auxiliaries.