New Ways of Analyzing Syntactic Variation 2

Ghent

Belgium

19-20 May 2016

NEW WAYS OF ANALYZING SYNTACTIC VARIATION 2

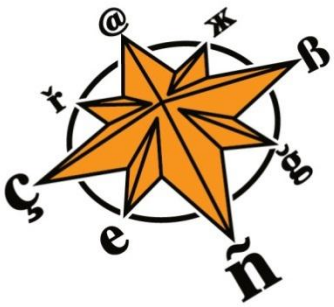
Ghent, 19-20 May 2016

Syntactic variation is a multidimensional concept: it can refer to the existence (in a single language variety) of several syntactic patterns or constructions "competing" for the same functional space (i.e, to **grammatical alternations**), or to any kind of sociolinguistic or "lectal" variation in the formal and/or functional properties of syntactic patterns, along regional, social, diachronic, stylistic, ethnic, gender, etc. dimensions (i.e., to **syntactic patterns or constructions as sociolinguistic variables**), or to a combination of both.

Syntactic variation is a major area of research in different schools of linguistics – including, but not limited to, construction grammar and related usage-based approaches, generative grammar, variationist sociolinguistics (cf. the advent of socio-syntax), psycholinguistics, language acquisition research, and computational linguistics/NLP – and has accordingly been approached from quite divergent theoretical and methodological perspectives. A common trend in all of these approaches is the **increasing use of advanced methods and tools**for the compilation and analysis of empirical data. In addition, there is a growing consensus that linguistic argumentation demands **converging evidence based on an interdisciplinary approach** and a growing body of work hence combines multiple empirical approaches to tackle one and the same linguistic phenomenon (e.g. combining advanced corpus analyses with survey and psycholinguistic experimental designs).

NWASV2 offers a forum for original work on syntactic variation, in any language or (present or historical) language variety – or varieties, for that matter – and from any theoretical perspective, in which an awareness of **recent theoretical insights and advances** is paired with a concern for appropriate **empirical validation, methodological innovation and interdisciplinarity**.

Sponsors



wijs *in* taal

Local Organizing Committee:

Bert Cappelle (University of Lille 3)
Timothy Colleman (Ghent University, GLIMS)
Ludovic De Cuypere (Ghent University, GLIMS)
Gert De Sutter (Ghent University, EQTIS)
Emmeline Gyselinck (Ghent University, GLIMS)
Natalia Levshina (FRS-FNRS, Université Catholique de Louvain)
Clara Vanderschueren (Ghent University, GLIMS)
Freek Van de Velde (University of Leuven)
Annelore Willems (Ghent University, EQTIS)

Scientific Committee:

Antti Arppe (University of Alberta)
Jóhanna Barddal (Ghent University)
Ronny Boogaart (Leiden University)
Anne Breitbarth (Ghent University)
Joan Bresnan (Stanford University)
Bernard De Clerck (Ghent University)
Bart Defrancq (Ghent University)
Renata Enghels (Ghent University)
Olga Fischer (University of Amsterdam)
Adele Goldberg (Princeton University)
Liliane Haegeman (Ghent University)
Torsten Leuschner (Ghent University)
Peter Lauwers (Ghent University)
Koen Plevoets (Ghent University)
Peter Petré (University of Antwerp)
Benedikt Szmrecsanyi (University of Leuven)
Dirk Speelman (University of Leuven)
Geert Stuyckens (Ghent University)
Sali Tagliamonte (University of Toronto)
Miriam Taverniers (Ghent University)
Graeme Trousdale (University of Edinburgh)
Antal van den Bosch (Radboud University Nijmegen)
Sonia Vandepitte (Ghent University)
Johan van der Auwera (University of Antwerp)
Roeland van Hout (Radboud University Nijmegen)
Marleen Van Peteghem (Ghent University)
Arie Verhagen (Leiden University)
Klaas Willems (Ghent University)

Plenary papers

Language variation and change: the case of heritage grammars

Artemis Alexiadou (Humboldt University, Berlin)

artemis.alexiadou@hu-berlin.de

A recent turn within generative linguistics is the investigation of the linguistic properties of the grammars of multilingual speakers. A particularly interesting group thereof are speakers of so-called Heritage languages (HLs). HLs are minority languages that are part of the family's cultural heritage, associated with the country of origin of speakers' parents or grandparents. These HLs are spoken at home or otherwise readily available to young children, but, crucially, they are not one of the dominant languages of the larger (national/majority) society (Rothman 2009). HLs are often referred to as minority languages. HL speakers live in a situation of frequent/permanent language contact between the dominant and the minority language and it has been observed that their grammar differs in significant respects from that of native speakers of the minority language. It is therefore an important task to identify how their grammars differ. In this talk, I will show that HL grammars instantiate patterns that are found in other dialects/registers of the minority language. Thus they exhibit patterns typical of language change and not of language attrition or incomplete acquisition.

The changing functions of competing forms: Attraction and differentiation

Hendrik De Smet (University of Leuven)

hendrik.desmet@arts.kuleuven.be

The relation between functionally similar forms is often described in terms of competition. This leads to the expectation that, on the long term, either one form must survive (substitution) or each form must find its unique niche in functional space (differentiation). Thinking along these lines is backed up by the isomorphic principle, which states that ideally one form corresponds to one meaning and vice versa. However, explanations of change based on competition face several problems. They cannot easily explain what causes synonymy in the first place and, where subsequent differentiation takes place, they often cannot predict how form-function mappings will be reorganized. It is shown here that long-term functional developments in competing forms are subject to forces not usually taken into account in competition-based accounts. First, competing forms often show attraction, becoming functionally more (instead of less) alike. It is argued that this is due to blending and analogy operating between the competing forms. Second, each competing form is anchored to its own constructional network. It is proposed that cases of differentiation typically result from changes in the constructional network of one of the competing forms. The argument is supported by three case studies, addressing attraction and differentiation in the English system of aspectual constructions, the English system of secondary predicate constructions, and in the Dutch pair of expressions *ver van* and *verre van*, both meaning 'far from'. The discussion provides evidence of a phenomenon competition-based accounts could not predict (attraction), and offers a solution for one they could not very well explain (differentiation).

Syntactic alternation research: taking stock and suggestions for the future

Stefan Th. Gries (University of California, Santa Barbara)

stgries@linguistics.ucsb.edu

Over the last 20 or so years, research on syntactic alternations has made great strides in both theoretical and methodological ways. On the theoretical sides, much of the research on syntactic alternations outside of variationist sociolinguistics was restricted to generative linguistics debating how near synonymous constructions differed slightly in meaning and/or how one (and which one) was derived from the other (transformationally). On the methodological side, much research outside of variationist sociolinguistics consisted of monofactorial studies based on relatively simple text counts. One exception to both these characterizations was variationist sociolinguistics, which early on explored performance aspects of alternations and (by now unfortunately) the multifactorial analysis of alternations using Variable Rule Analysis. By now, however, syntactic alternation research has become much more functional (in a broad sense of the term) and much more methodologically sophisticated: much work is now motivated/interpreted psycholinguistically or in a broadly usage-based/cognitive linguistic framework and much work has now adopted a regression-based analytical strategy. These attractive developments notwithstanding, much remains to be done. In this talk, I will sketch some recent developments in (largely) separate alternation studies that I would like the field to adopt more broadly. These developments can be grouped into ones that have to do with (i) the statistical analysis of corpus-based and experimental alternation data, (ii) new predictors that explain typically unexplored aspects of variability in alternations, and (iii), time permitting, aspects of visualization of statistical modeling.

Subtitles, tweets, and syntax. Big data from small input.

Stefan Grondelaers

(with Antal van den Bosch and Dirk Speelman)

s.grondelaers@let.ru.nl

With the advent of ever larger corpora and ever more powerful computers, big data have become the shiniest new toy in a wide number of scientific disciplines, but many theoretical linguists continue to regard the computational handling of such data as "linguistics with some practically useful but theoretically irrelevant and obfuscating nerdy add-ons" (Spärck Jones 2007: 440). In this talk, we present three case studies to demonstrate that big data and computational techniques are not toys, but pivotal tools for the (theoretically inclined) socio-syntactic researcher.

A first case study (with Antal van den Bosch) explores *how much* syntactic variation there is in Dutch. One of the reasons why variation in the syntax of Dutch was largely neglected until recently is the fact that both analysts and users are mostly oblivious to it. In order to go beyond the handful of syntactic variables hitherto identified, and obtain a more aggregate view of syntactic variation in Dutch, we use the bottom-up technology pioneered in Bannard & Callison-Burch (2005) and Van den Bosch (2014), and a parallel corpus of Dutch translations of the English subtitles to 6700 movies. Statistical machine translation software is used to identify plausible mappings between English n-grams and their Dutch translations in order to obtain paraphrases, i.e. stretches of interchangeable text that carry approximately the same meaning. In the pilot study presented here, we investigated which proportion of the set of matching Dutch alternatives which map with a specific English n-gram, represented possible instantiations of syntactic variables (instead of, for instance, idioms or multi-word units). We found evidence for the usual suspects – *er*-variation, word order alternations, complementizer omission, etc. – but also a number of alternations we had not anticipated as interesting variables.

In a second case study (with Dirk Speelman), we use distributional semantic analysis to investigate the prestige triggers which sustain the rapid spread of two syntactic innovations in Netherlandic Dutch, the proliferation of the object pronoun *hun* "them" in subject position (as in *Als je zo speelt krijgen hun natuurlijk altijd kansen* "If you play like that them will always get chances"), and the emergence of *do*-support in Dutch (as in *Doe jij de vaatwasser uitruimen* lit. "do you clean out the dishwasher"). Both excite social controversy, as well as irritation and concern on the part of the cultural and educational establishment. Yet, there must be some prestige trigger for their rapid spread, and the standard experimental access into this trigger – the Speaker Evaluation technique – has offered some evidence for a new kind of (covert) dynamic prestige. In order to tap into qualitatively richer prestige perceptions, we conducted Free Response experiments in which evaluations of the investigated variants were elicited in terms of keywords spontaneously produced by the experimental participants. Distributional semantic analysis was used to cluster the 700 + keywords into 35 evaluative dimensions. The pilot study reported in this talk reveals that whereas the computationally enriched Free Response data and the Speaker Evaluation data access the same perceptual clusters, the former are much more informative.

A third case study (if time permits) investigates the merits of a data source with big data potential, though the analysis reported is the result of manual data coding. In order to study the language-internal conditioning of subject-*hun* (cf. above) on a sufficiently large dataset, we extracted a set of 12.903 tweets containing either standard subject-*zij* or subject-*hun*. While regression analysis confirmed previous findings pertaining to subject-*hun's* linguistic conditioning, the large effect of "intentional misspelling" – which demonstrates that tweeters consciously stylize themselves as dynamic and anti-authoritarian provocateurs – on the distribution of *hun*, forces us to refine the analysis, and question the value of Twitter as a data source.

Shared syntax in multilinguals: learning, translation, and variation

Robert J. Hartsuiker (Ghent University)

rob.hartsuiker@ugent.be

Based on experiments demonstrating structural priming across the languages of multilinguals, I have proposed an account by which multilinguals share syntactic information whenever the structures involved are sufficiently similar (e.g., Hartsuiker et al., 2004). In this talk, I will discuss the implications of such a shared-syntax view for (a) learning syntactic information in a second language; (b) sentence processing during translation (in non-expert translators); (c) syntactic variation at the level of larger groups as a function of language contact. Regarding learning, I will argue that the 2004 shared syntax model should be seen as the endpoint of a developmental trajectory. I will propose a sketch of this trajectory, based on priming experiments with participants varying in second-language proficiency (Hartsuiker & Bernolet, in press). Next, I will turn to a recent study in which German-English and English-German bilinguals (without translator training) translated sentences (Maier, Pickering, & Hartsuiker, submitted). The study provided evidence for an account of translation in which structural priming of the source language sentence strongly affects the choice of syntactic structure in the target language. Finally, I will speculate on the possibility that cross-linguistic priming of shared structures is a mechanism by which language contact can relatively easily change the distribution of structures at a macro level (e.g., in all speakers of a particular geographic area) and discuss a number of predictions of such an account.

References

Hartsuiker, R. J., & Bernolet, S. (in press). The development of shared syntax in second language learning. *Bilingualism: Language and Cognition.*

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish/English bilinguals. *Psychological Science*, *15*, 409-414.

Maier, R. M., Pickering, M. J., & Hartsuiker, R. J. (submitted). Does translation involve structural priming?

Regular session papers

So similar in principle, but so different in practice. Mixing texts, elicitation and experimentation in the study of the Plains Cree independent and conjunct verb constructions

Anti Arppe (University of Alberta)(antti.arppe@iki.fi)

Atticus Harrigan (University of Alberta)

Katie Schmirler (University of Alberta)

The verbal grammar of Plains Cree (ISO code: crk), a polysynthetic Algonquian language spoken across Canada, exhibits a pervasive morpho-syntactic constructional alternation, namely the *independent* vs. *conjunct orders* (e.g. 1-2):

(1) *kikî-nôhtê-wâpamâw*
Ind.2Sg+Pret+PreV/want+*see*.Verb.Transitive.Animate+3SgO
You (Sg.) wanted to see him/her.

(2) *ê-kî-nôhtê-wâpamat*
Cnj.Preverb/ê+Pret+Prev/want+*see*.Verb.Transitive.Animate+2Sg.3SgO
You (Sg.) wanted (were wanting) to see him/her.

In linguistic descriptions and pedagogical materials of Plains Cree, these two alternative constructions are given straightforward English translations: independent verb forms are translated as simple forms ('he does something', 'he did something'), while conjunct verb forms are translated as progressive (*-ing*) forms ('he is doing something', 'he was doing something'). Thus, these two orders are described as being semantically disparate. Another distinction between the two constructions involves where and how they are presumed to be used: independent verbs are claimed to occur in phrases that can stand alone as sentences, while conjunct forms occur alongside an independent phrase: 'he did something (independent) while something else was happening (conjunct)' (Wolfart 1973, 1996; Okimāsis 2004). Wolvengrey (2011) echoes this, indicating that the conjunct may be used in both main and secondary clauses, while the independent is "most closely associated with the main clause" (ibidem, 45). However, conjunct verbs can occur independently outside the context of an independent verb, and many speakers, when asked, will find the conjunct more natural in numerous contexts. Cook (2008) explored these facts and her investigations suggest that the context for a conjunct verb is not necessarily an independent verb form, but that a conjunct verb must simply occur in some pre-existing context within the discourse or conversation.

Nevertheless, any Plains Cree verb can presumably be used in either order, with restrictions based primarily on the overall pragmatics and semantics of the message being conveyed. Moreover, each order can in principle both express the same range of person/number features for their actor (subject) and/or goal (object), but with distinct morphemes, and each can be modified with the same set of preverbs, particle-like prefix morphemes that in part resemble auxiliary verbs and in part adverbs (see e.g. *-nôhtê-* 'want' in examples 1-2 above, or *-âpihtâ-kîsikâwi-* 'at noon').

Furthermore, Wolvengrey (2011) claims that the choice of order has no systematic bearing on a sentence's word order. Though previous sources offer no concrete numbers or indications of which order seems to be most common, Cook (2014) notes that the independent is more syntactically restricted in interpretation, and thus potentially rarer, in comparison to the conjunct.

Nevertheless, prior corpus-based research with other morphologically rich languages such as Finnish has indicated that verbs may substantially differ as to their overall inflectional profiles, and that these differences may be semantically motivated (Karlsson 1985, 1986; Arppe 2006). Therefore, we have conducted a similar empirical study of Plains Cree verbs with a corpus of 94,000 words, where we have observed significant differences in the number of independent vs. conjunct for individual lemmata (Figure 1). For instance, *itwêw* 's/he says thus' occurs almost only in the independent, while *tipiskâw* 'it is night/dark' occurs only in the conjunct), with the rest of the verbs exhibiting a continuum of varying proportions between independent vs. conjunct forms. Moreover, we have also observed significant differences on the whole for the four conjugation classes, in how verb forms altogether per each class (Intransitive Inanimate, Intransitive Animate, Transitive Inanimate, Transitive Animate) are used in either order.

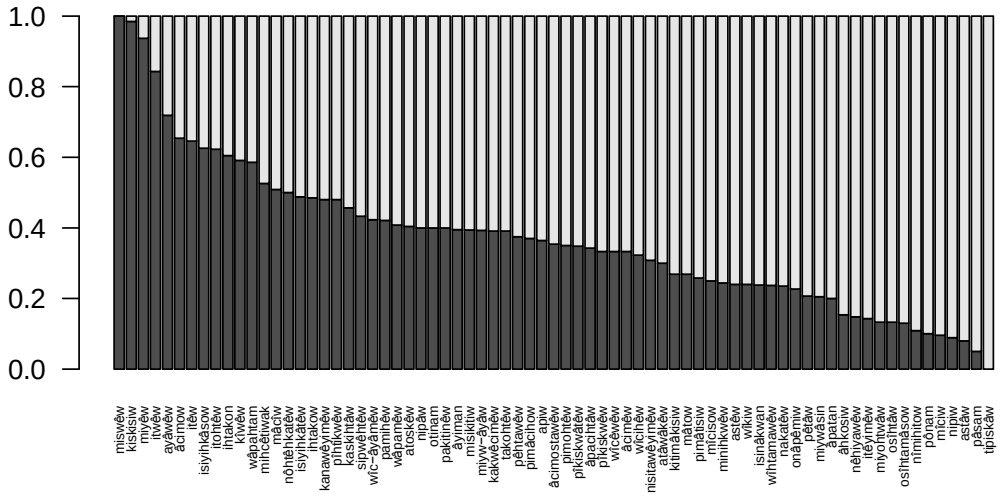


Figure 1. The proportions of independent vs. conjunct forms for a selection of frequent verbs (1.0 ~ only independent forms; 0.0 ~ only conjunct forms observed).

Following up on these empirical corpus-based observations, we will explore possible motivations and explanations for such divergent preferences, starting with the morphological structure of the verbal constructions, e.g. associations between order and actor/goal person, tense, as well as preverbs. Moreover, we will scrutinize whether more fine-grained semantic classifications of the lemmata, than the four general conjugation classes, might explain order preference. Finally, we will expand to the sentential context and evaluate how order choice is determined when multiple verbal constructions are used in the same sentence.

Furthermore, as the texts that are at our disposal for Plains Cree are quite small, and furthermore restricted in terms of their genre, we will contrast these corpus-based results by presenting preliminary qualitative results from field linguistic elicitation that we are conducting in a Cree community. This has allowed us to collect metalinguistic, explicit introspection on when either order is, or should be, used, as well as pursue more indirect probing of native speaker linguistic knowledge through systematically structured experimentation on which of the two alternative orders of a verb is preferred by our consultants, either in isolation, or within example sentences, providing some carefully selected contexts. This allows us to contrast and mix multiple types of linguistic evidence, namely naturally produced language as well as linguistic judgments, as part of language documentation work.

References

Arppe, Antti 2006. Frequency Considerations in Morphology, Revisited - Finnish Verbs Differ, Too. In: *A Man of Measure. Festschrift in Honour of Fred Karlsson in his 60th Birthday*. Suominen, Mickael et al. (Eds). Special Supplement to SKY Journal of Linguistics, Volume 19/2006, pp. 175-189. Linguistic Association of Finland, Turku.

Cook, Clare. 2008. *The Syntax and Semantics of Clause-typing in Plains Cree.* PhD dissertation, University of British Columbia.

Cook, Clare. 2014. *The Clause-Typing System Of Plains Cree.* Oxford: Oxford University Press.

Karlsson, Fred. 1985. Paradigms and word forms. *Studia gramatyczne*, VII: 135–154.

Karlsson, Fred. 1986. Frequency considerations in morphology. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 39.1: 19–28.

Okimāsis, Jean. 2004. *Cree: Language of the Plains / nēhiyawēwin: paskwāwi-pīkiskwēwin*. Regina: Canadian Plains Research Center.

Wolfart, H. Christoph. 1973. *Plains Cree: A Grammatical Study*. Transactions of the American Philosophical Society New Series, vol. 63 (5). Philadelphia: The American Philosophical Society.

Wolfart, H. Christoph. 1996. Sketch of Cree, an Algonquian Language. *Handbook of North American Indians* 17:390-439.

Wolvengrey, Arok. 2011. *Semantic and pragmatic functions in Plains Cree syntax*. Utrecht/Amsterdam: LOT Universiteit van Amsterdam Host.

A crowdsourcing approach to the description of regional variation in French clitic-clusters

Mathieu Avanzi (Universität Zürich)(Mathieu.avanzi@gmail.com)

Elisabeth Stark (Universität Zürich)

Our contribution is dedicated to the empirical testing of alleged stylistic, sociolinguistic and regional variants of clitic-clusters in modern French in France, Belgium and Switzerland. We will show some intriguing new insights into the regional distribution of non-standard variants and discuss two hypotheses to explain their coming into being, i.e. language-contact (with Francoprovençal and Occitan, two Galloromance languages genetically close to, but not identical with French) and/or analogical leveling.

In European French like in other Romance languages (cf. Miller & Monachesi 2003), the internal ordering and the placement of (object) clitics is subject to variation (documented for centuries, cf. De Kok 1985; Gilliéron & Edmond 1902-1910). Besides standard structures exemplified in (a) [type (1), (2) and (3)], one can find variants like the ones exemplified in the examples under (b)-(e):

(1) a. donne -le -moi!
give:IMP:2.Sg it:PRO:ACC:3.SG:cli to me:PRO:1.SG:ton[1]
b. donne -moi -le!
give: IMP:2.Sg to me:PRO:1.SG:ton it:PRO:ACC:3.SG:cli
c. donne -me -le!
give:IMP:2.Sg to me:PRO:1.SG:cli it:PRO:ACC:3.SG:cli
'Give it to me!'

(2) a. ne m' en donne pas!
NEG to me:PRO:1.SG:cli of it:PRO:PART:cli give:IMP:2.Sg not
b. donne m' en pas!
give:IMP:2.Sg to me:PRO:1.SG:cli of it:PRO:PART:cli not
c. m' en donne pas!
to me:PRO:1.SG:cli of it:PRO:PART:cli give:IMP:2.Sg not
d. donne -z[2]-en -moi pas!
give:IMP:2.Sg of it:PRO:PART:cli to me:PRO:ton not
e. donne -moi -z-en pas!
give:IMP:2.Sg to me:PRO:1.SG:ton of it:PRO:PART:cli not
'Don't give some of it to me!'

(3) a. je le lui donne.
I:PRO:NOM:1.SG:cli it:PRO:ACC: 3.SG:cli to him:PRO:DAT:3.SG:cli give:PRES:1.Sg
b. je lui le donne.

[1] *Ton* = *tonique*, i.e. strong pronoun in the classification of Cardinaletti & Starke (1999).

[2] [z] is a subjacent consonant usually realized in front of vocalic onsets of the following word, a *liaison* phenomenon, the analysis of which is highly debated, see Morin (1979). We will abstract away from this phenomenon and consider the form *z-en* as an allomorph of *en*.

I:PRO:NOM:1.SG:cli to him:PRO:DAT:3.SG:cli it:PRO:ACC: 3.SG:cli give:PRES:1.Sg
'I give it to him'

Very little work has been conducted regarding clitic-cluster variation in French. In the literature, examples such as (1b), with inverted order between the strong, non case-marked pronoun *moi* and the accusative clitic *le*[3], (2d) and (2e) with additional enclitic position (as opposed to the standard proclitic one), with similarly varying orders and different patterns of *liaison* (Morin 1979) of postverbal pronouns in a negated imperative, have been, when not singled out as ungrammatical in the standard (Goldbach 2007), essentially considered as features characterizing "français populaire" (Dauzat 1930; Bauché 1951; Riegel et al. 1994; Grévisse & Goosse 2008), i.e. of non-standard colloquial French. Additionally, examples such as (1c), and (3b), displaying consistent dative-accusative order between the clitics, have been claimed to be specific to some regional varieties of French: according to Tuaillon (1983) and Violin-Wigent (2010), dative-accusative order in (3b) as opposed to standard (3a) is typical for Grenoble and Briançon, two cities located in the Francoprovençal and the Eastern Occitan areas, namely. The form (2b), finally, with enclitic pronouns as opposed to the proclitic standard form in (2a), is widely recognized as a possible non-standard variant of (2a), while the examples under (2d) and (2e) are mentioned as even more marked (see above).

To our knowledge, no systematic study has ever been conducted to test the empirical validity of the variationist descriptions of the different constructions (regional and stylistic distribution) based on data gathered from a large sample of informants. The main reason for this is that the constructions, whether they are standard or non-standard, are very rare in spoken and written corpora, so that a viable new methodology is necessary to rapidly gain as much information and judgments as possible, such as a web-based survey. In our crowdsourcing online experiment on regional variation in French[4], participants were asked to tick a box corresponding to one of the sentences describing a situation illustrated in a picture. They were asked to respond as naturally as possible, as if they were having an informal chat with peers living in the place where they spent the major part of their lives. For clitic-clusters, they were, e.g., presented a picture of a young boy holding a cell phone, with the following text: "you lent your mobile phone to your little brother, but he does not want to give it back to you. A little bit upset, you tell him". They then had three choices: "a/*rends-le-moi*; b/*rends-moi-le*; c/*rends-me-le*" (cf. examples (1a-c)) and had to select the answer corresponding best to the situation presented. Until now, almost 20,000 participants, native speakers of French varieties from France, Belgium and Switzerland, took part in this survey.

Preliminary statistical analyses based on R (R development Core team 2015) conducted on questions built to test clitic-clusters of type (1) and (3) revealed interesting results, pointing to a regional variation. As illustrated on figures 1, 2 and 3 below, the non-standard variant (1b) is almost as widely spread as the standard variant (1a) (though the latter remains rarer). As for variant (1c), it seems to be more used in the Francoprovençal area and in the South-West of France than elsewhere, even if it is attested in other areas of France and Belgium, too.

[3] Variants such as (1c), where a non-case marked clitic (*me*) is used in non-final position, are generally described as typical features of the French spoken in Switzerland according to Singy (1996).
[4] The links to participate to the survey can be found on the website of the project (http://francaisdenosregions.com/).

As for figures 4 and 5 below, they show that the standard variant (3a) is much more used than the variant (3b), which in turn is more often employed in the Francoprovençal area. There seems to be a regional preference for consistent dative-accusative order irrespectively of construction or grammatical person in this area, which has to be explained.

Statistical tests revealed also an interesting effect of the speakers' socio-economic status regarding examples under (3): working-class speakers use the non-standard variant to a greater extent than the other speakers, and this is true regardless of age, origin or sex. Such effects were not found for the sentences under (1), speaking maybe in favor of the 'learnt' character of imperative constructions and the greater normative awareness during their use.

Figures

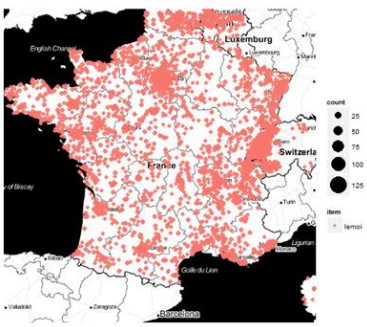


Fig. 1 : *donne-le-moi*

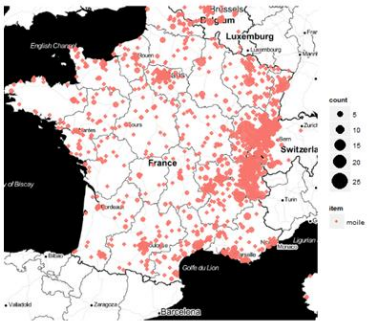


Fig. 2: *donne-moi-le*

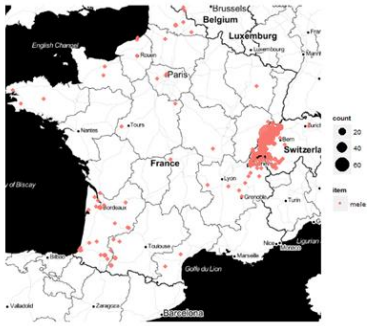


Fig. 3: *donne-me-le*

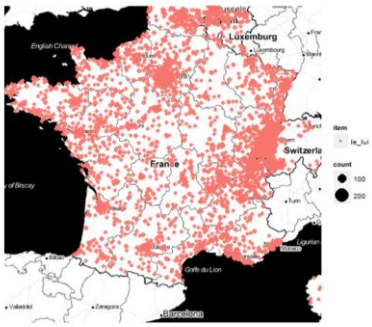


Fig.4: *il le lui donne*

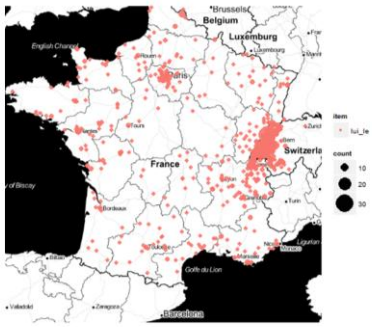


Fig.5: *il lui le donne*

References

Cardinaletti, A. & M. Starke (1999). The typology of structural deficiency: A case study of three classes of pronouns". In Clitics in the languages of Europe, ed. Henk van Riemsdijk, 145-233. Berlin: Mouton.

Dauzat, A. (1930). *Histoire de la langue française*. Paris: Payot.

De Kok, A. (1985). *La place du pronom personnel régime conjoint en français : une étude diachronique*. Amsterdam: Rodopi Editions.

Gilliéron, J. & E. Edmont (1902-1910). *Atlas linguistique de la France*. Paris: Champion.

Goldbach, M. (2007). The Distributed Morphology of Object Clitics in Modern French. *Zeitschrift für Sprachwissenschaft*, 26.

Grévisse, M. & A. Goosse (2008). *Le bon usage*. Bruxelles: Duculot.

Miller, P. & Monachesi (2003). Les pronoms clitiques dans les langues romanes. In *Les langues romanes, problèmes de la phrase simple,* ed. D. Godard, 67-123. Paris: Editions du CNRS.

Morin, Y. (1979). La morphophonologie des pronoms clitiques en français populaire. *Cahiers de linguistique,* 9, 1-36.
Peters, H. (2014). The morpho-syntactic status of ne and its effect on the syntax of imperative sentences. *Journalof French Language Studies*, 24, 49-81.
R Development Core Team (2015). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, available at http://www.R-project.org.
Riegel, M., J.-C. Pellat & R. Rioul (2009). *Nouvelle grammaire méthodique du français*. Paris: PUF.
Rowlett, P. (2014). French imperatives, negative ne, and non-subject clitics. *Journal of French Language Studies*, 24, 29-47.
Singy, P. (1996). *L'image du français en Suisse romande. Une enquête sociolinguistique en Pays de Vaud*. Paris: L'Harmattan.
Tuaillon, G. (1983). Régionalismes grammaticaux. *Recherches sur le français parlé,* 5, 227-240.
Violin-Wigent, A. (2010). *Tu lui les as donnés* ? Variation dans l'ordre des pronoms multiples en phrases déclaratives. In *Actes du 2ème CMLF,* eds. F. Neveu, V. Muni Toke, J. Durand, T. Klingler, L. Mondada & S. Prévost, 2001-2014.

Merging Verb Cluster Variation

Sjef Barbiers (Meertens Institute Amsterdam)(sjef.barbiers@meertens.knaw.nl)

Hans Bennis (Meertens Institute Amsterdam)

Lotte Hendriks (Meertens Institute Amsterdam)

As is clear from the Syntactic Atlas of the Dutch Dialects (SAND I & II, Barbiers et al. 2005/2008) variation in word order in Dutch dialects is a rather infrequent phenomenon. Most variation is found in the domain of morphosyntax, and thus relates to variation in form rather than in order. In those cases the form of a particular word varies across dialects. Well-known examples include subject pronouns, relative pronouns, complementizers and verbal inflection (cf. SAND I).

Compared to morphosyntactic variation, the word order we find within the Dutch language area is remarkably constant. For instance, all 267 dialects that are part of the SAND-research show exactly the same pattern for the placement of the finite verb. There is no variation with respect to Verb Second, although the placement of the finite verb is very variable cross-linguistically, as is clear from the vast literature on Verb Movement in for instance Germanic and Romance languages. Similarly, although the OV-order is cross-linguistically relatively exceptional, all Dutch dialects have the verb following the object in subordinate clauses. Without exception, Dutch dialects have OV-order and move the finite verb to the beginning of the clause in clauses without a complementizer.

However, there is one domain in which word order variation is abundant. This concerns the famous verb raising phenomenon in Dutch (and German). If we find more than one verb at the end of the clause, e.g. a main verb and one or more auxiliary or modal verbs, the order appears to be unstable across dialects. We find four orders in a subordinate clause with a main verb and two modals:

(1)
a.	Ik vind dat iedereen moet kunnen zwemmen.	V1-V2-V3
b.	Ik vind dat iedereen moet zwemmen kunnen.	V1-V3-V2
c.	Ik vind dat iedereen zwemmen kunnen moet.	V3-V2-V1
d.	Ik vind dat iedereen zwemmen moet kunnen.	V3-V1-V2

I find that everyone must can swim
'I think that everybody should be able to swim.'

These sentences are semantically and pragmatically identical. The different orders found in verb clusters appear to be determined by:

(2)
(i) geographical location of the dialect
(ii) type of the auxiliaries in the verbal cluster
(iii) hierarchy of auxiliaries in the verbal cluster

There is a vast literature on verb cluster formation and the variation in order of the verbs in clusters, starting with the seminal publication of Arnold Evers in (1975). Our approach differs from most of the existing literature in at least four respects.

First of all, our analysis takes dialect geography as a starting point. We will concentrate on the variation we find at different locations to see if there are particular co-occurrence patterns that might help us to understand the phenomenon of verb clustering.

Secondly, we approach different order possibilities as the consequence of the structure building process Merge, together with differences in the categorial status of the elements in the cluster. We thus present an analysis without movement rules such as Verb Raising (rightward head movement) or VP-intraposition (leftward XP-movement).

Thirdly, we briefly show that the results of our theoretical analysis of dialect geographical variation converge to a large extent with the results of the reversed dialectometrical, i.e. quantitative-statistical approach to verb clusters recently put forward by Jeroen van Craenenbroeck (2015). Van Craenenbroeck concludes that the variation in verb cluster ordering can be reduced to three grammatical parameters.

Last but not least, we will demonstrate that the intuitions speakers have of the various orders in verbal clusters, even with respect to cluster orders they don't produce themselves, correlate quite nicely to the patterns we find within the Dutch speaking area. We argue that this must be due to their syntactic knowledge and cannot be due to processing preferences or familiarity with the various orders.

We show that the word order variation in verb clusters in the Dutch language area as found in SAND Volume 2 can be reduced to two truly verbal orders: V1-V2-V3 and V3-V2-V1. In two other attested orders, V1-V3-V2, V3-V1-V2, the main verb (V3) is not verbal but adjectival (in the case of a participle) or nominal (in the case of an infinitive). The order V2-V3-V1 is exceptional in that it is only possible if V2 and V3 form an adjectival cluster. The order V2-V1-V3 is unattested.

With the help of the geographic distribution of the various orders we argue that these variation data are best captured in terms of a syntactic analysis that merges verb clusters (without any movement) and includes three syntactic parameters:

(i) A dialect is {descending/ascending} in the ordering of verbs.
(ii) A dialect {does/does not} have verbal participles.
(iii) A dialect {does/does not} have nominalized infinitives in “verb” clusters.

This syntactic analysis is supported by correlations between cluster orderings and so-called cluster interruptions, i.e. the occurrence of non-verbal material inside a cluster. For example, in the Dutch dialects in Belgium the order V1-V3participle-V2 is quite frequent, which follows from the strong preference in that area for participles to be adjectival and the fact that most of these dialects allow cluster interruption with predicative adjectives.
We compare the geographic SAND data with the results of a ranking experiment in which respondents from the whole Dutch language area had to provide a relative ranking of the six logically possible word orders in three-verb clusters. The most important feature of this experiment was that the respondents thus had to give judgments on word orders that do not occur in their own language varieties. Strikingly, the rankings of the respondents show a strong convergence, independently from the dialect area that they live in. This shows that the linguistic environment of the respondents does not influence their rankings and that these rankings therefore cannot be explained in terms of familiarity or frequency of use. Ease of processing is also shown to predict the wrong rankings. These rankings must be due to the grammatical knowledge of the respondents. This strengthens our explanation of verb clustering variation within the Dutch language area.

Handling syntactic variation in idioms and compounds

Leah S. Bauke (University of Wuppertal)(leah.bauke@gmail.com)

It is generally known that certain compound types have an idiomatic reading, i.e. English *red neck, pick pocket, egg head, green back, walk man, sit-in ...* Characteristic of all these is that their meaning cannot be compositionally derived from the meaning of the component parts, nor can the syntactic category of the compound necessarily be derived from the categories of the component parts: e.g. $[sit_V\ in_P]_N$. The same can be found in many other languages (cf. e.g. the examples in (1) in Chinese from Zhang 2007):

(1) a. yi ge hen bao-shou de ren
one CL very keep-defend mod person
'a very conservative person'

b. hen mao-dun
very spear-shield
'very contradictory'

c. zhe zhang zhuozi de da-xiao
this CL table mod big-small
'the size of this table'

d. yi ge kai-guan
one CL open-close
'a switch' (e.g. 'a power switch')

These compounds can be analyzed in terms of a content matching analysis (cf. Borer 2013). According to this analysis meaning assignment in compounds can either be to the individual roots, which leads to a compositional interpretation, or to both roots in a compound frame, where meaning is assigned to the whole compound as a single unit and a non-compositional interpretation results. So for the English compound *fish slice* content matching can either assign meaning to the two roots separately (with the interpretation: *a slice of fish*) or it can be, as a unified *en*-search, to the whole complex (with the interpretation of: *a kitchen utensil used for tossing food in pans*). Both versions are schematically illustrated in table 1:

Table 1:

en-searches (domains boxed)		Content and	Composition
i. en-search 1	$[_{C_1}[_{C_2}\sqrt{\text{FISH}}]\ \sqrt{\text{SLICE}}]$	FISH	$[\sqrt{\text{SLICE}}]$
ii. en-search 2	$[_{C_1}[_{C_2}\sqrt{\text{FISH}}]\ \sqrt{\text{SLICE}}]$	FISH	SLICE
unified en-search	$[_{C_1}[_{C_2}\sqrt{\text{FISH}}]\ \sqrt{\text{SLICE}}]$	FISHSLICE	

In this respect compounds seem to pattern with idioms where we can also observe meaning that cannot be compositionally derived from the meaning of the elements that constitute the idiom and where the category is also not necessarily derivable from the categories of the elements that constitute the idiom. This is, however, not unproblematic, because idioms in contrast to compounds are partially compositional in the sense that they allow internal modification, pronominal reference, contrastive focus, etc. (cf. e.g. Nunberg et al. 1994, Kovecses & Szabó 1996, Borer 2013). All of these properties presuppose the existence of functional structure inside idioms, which makes them unavailable for a content matching along the lines described for compounds, because content matching can only operate on roots and not on functional structure.

In my talk I investigate one type of construction in German that straddles the line between idiom and compound in that it shows non-compositional content and categorial unfaithfulness, while it allows some modification that seems to presuppose functional structure. The construction is exemplified in (2):

(2) a. Mäuse melken
mice milking
'frustrating'

b. Eier legen
eggs laying
'wonderful'

c. Bäume ausreißen
trees tearing
'fit, vigorous'

d. junge Hunde kriegen
young dogs having
'exasperating'

These constructions are problematic for a content matching analysis, because they all involve plural forms - which clearly indicate functional structure. What is striking about these forms though is that their occurrence is strictly limited to nominalized infinitives with *zum* (resembling nominal gerunds in English):

(3) a. Die Situation ist zum Mäuse melken
'The situation is frustrating'

b. Das Wetter ist zum Eier legen
'The weather is wonderful'

c. Er fühlt sich zum Bäume ausreißen
'He feels vigorous'

d. Der Vortrag war zum junge Hunde kriegen
'The talk was very exasperating'

This is very much reminiscent of another rather huge set of *zum* + infinitive constructions, which also have an idiomatic interpretation:

(4) zum Wiehern/Brüllen/Kugeln/Schreien/Kringeln/Quieken,...
neighing/yelling/rolling/screaming/curling/squeaking
'very funny'

These latter forms can be interpreted straightforwardly under a content matching analysis, where content matching must take place when the roots are merged with the nominalizer *zum* at the latest. (Alternatively content matching takes place at the root, leading to the literal interpretations).

For the forms in (2/3) this means that content matching can likewise be delayed until merger of *zum,* where the merger of the roots e.g. *Maus* and *melken* is an instantiation of a compound frame, thus alining the derivation of these idioms to a compound interpretation in a compound frame. Plural assignment can be derived from remerging one of the constituents (*Maus* in this case) under the relevant functional structure, i.e. adjoined to *zum* - a nominalizer!

With this much in place, the system can be easily extended to further idiom types:

(5)	zum Greifen nahe	zum Affen machen	zum Zuge kommen
	to grab close	to mokey make	to move get
	'within easy reach'	'to disgrace oneself'	'become active'

References

Borer, H. 2013. *Taking form*. Oxford: OUP.

Kovecses, Z. & Szabó, P. 1996. 'Idioms a view from cognitive semantics. Applied Linguistics 17: 326-55.

Nunberg, G., et al. 1994. 'Idioms'. *Language* 70: 491-593.

Zhang, N. 2007. Root merger in Chinese compounds. *Studia Linguistica* 6: 170-84.

Verbal cluster order and processing complexity

Jelke Bloem (University of Amsterdam)(j.bloem@uva.nl)

Arjen P. Versloot (University of Amsterdam)

Fred Weerman (University of Amsterdam)

This study aims to provide a linguistic explanation for the observation that many different factors appear to play a role in Dutch verbal cluster order variation. We claim that there is a link between the processing complexity of verbal cluster constructions, and their word order variation. We take a quantitative approach and extract data from the Lassy Large corpus (van Noord, 2009), an automatically annotated corpus of Dutch texts, to look for patterns among a set of 827.709 verb clusters. We then use a multifactorial logistic regression analysis to obtain information about the effect size of each factor, an approach pioneered by Gries (2001).

When producing Dutch two-verb clusters, a speaker can choose between different grammatical word orders or constructions to express essentially the same meaning:

1) Ik denk dat ik het begrepen heb (2-1 order)
   I think that I it understood have
   `I think that I have understood it.'
2) Ik denk dat ik het heb begrepen (1-2 order)
   I think that I it have understood

These orders are called the 2-1 and 1-2 orders, where 1 indicates the position of the head verb and 2 that of the main verb. Corpus studies have shown that a range of factors contribute to this order variation (De Sutter, 2005; Coussé et al., 2008; Bloem et al., 2014), ranging from morphosyntactic and semantic factors to properties such as sentence length and word frequency. While these studies reveal and quantify interesting patterns in the variation and associations with linguistic factors, they do not really explain why all of these different factors are involved when people choose to use one word order or the other. We conducted a large-scale corpus study in order to investigate what these factors have in common.

Some of the variation can be attributed to semantic or information-structural factors, as in other word order phenomena. However, even when these factors are controlled for, variation remains. We argue that any variation that cannot be attributed to semantics or information structure, should be attributed to processing. We observe that a variety of factors that are related to verbal cluster word order, can also be related to the processing complexity of the cluster's context. Speakers generally prefer to minimize their use of cognitive resources, formulating sentences in a way that minimizes processing complexity when they have multiple grammatical ways to express something. Speakers can therefore adapt the word order of their verbal clusters to the complexity of the context to facilitate processing. Similar processing hypotheses have been proposed for other variation phenomena. Gries (2001) already formulated the basic idea when discussing optional particle movement in English, stating "The multitude of variables (...) that seems to be related to Particle Movement can all be related to the processing effort of the utterance". In the case of verb clusters there are factors unrelated to processing difficulty as well, but we have controlled for them in this study.

Various theories suggest that one of the two word orders may be the `default' order, and/or is acquired earlier. On the basis of acquisition evidence, Meyer and Weerman (submitted) argue that the 1-2 order is the evidence children need to learn about verb cluster constructions, which would make it more entrenched in speakers' grammars or lexicons, and therefore easier to access. This leads to a testable explanation of the variation in terms of processing complexity: in a context that is more difficult to process, when there is a higher processing load, more default-order clusters will be produced. On the other hand, De Sutter (2005) suggests that the 1-2 order is stylistically preferred. If this is true, we would expect to find the 1-2 order in contexts that are less difficult to process. If the 1-2 order is the default order, we would expect to find it in contexts that are more difficult to process.

In this work, we conduct a systematic analysis of factors that were linked to Dutch verb cluster order variation in previous work, using the data from the Lassy Large corpus in a process that requires no additional manual annotation. This allows for the analysis of large amounts of data and for the detection of even small probabilistic tendencies in the data. It also enables detailed analysis of more specific forms of the construction.

In our multifactorial regression analysis, we indeed observe that several factors indicating higher sentence complexity, such as longer clauses and more complex verbs, are associated with the same word order in the corpus. This word order is the 1-2 order (Example 2), which provides evidence for a default 1-2 order hypothesis. There is also some conflicting evidence however, such as the observation that 2-1 orders are more frequent in spoken language, which is generally considered to cause a higher processing load. But overall, these findings indicate that Dutch speakers will produce the more entrenched order when the context is more difficult to process, at least when writing, and that this entrenched order is the 1-2 order. More generally, this result shows that it is necessary to go beyond semantic and syntactic factors in search of an explanation for this word order alternation. We conclude that processing complexity is another motivation for variation, in addition to the usual suspects of semantics and information structure.

References

Jelke Bloem, Arjen Versloot, and Fred Weerman. Applying automatically parsed corpora to the study of language variation. In Proceedings of COLING 2014, pages 1974-1984, Dublin, Ireland, August 2014.

Evie Coussé, Mona Arfs, and Gert De Sutter. Variabele werkwoordsvolgorde in de Nederlandse werkwoordelijke eindgroep. Taal aan den lijve, pages 29-47, 2008.

Gert De Sutter. Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen. University of Leuven: PhD thesis, 2005.

Stefan T Gries. A multifactorial analysis of syntactic variation: particle movement revisited. Journal of quantitative linguistics, 8(1):33-50, 2001.

Caitlin Meyer and Fred Weerman. Cracking the cluster: The acquisition of verb raising in Dutch. Manuscript in preparation, submitted.

Gertjan van Noord. Huge parsed corpora in Lassy. Proceedings of TLT7. LOT, Groningen, The Netherlands, 2009.

Agreement Mismatches in Dutch Relatives

Gosse Bouma (Rijksuniversiteit Groningen) (g.bouma@rug.nl)

Dutch relative clauses are normally headed by a pronoun that agrees with the antecedent noun. However, if the antecedent noun is singular neuter, speakers increasingly use the non-agreeing form (b) instead of the agreeing form (a).

a. Ik moet gewoon nu nog beginnen aan een essay dat morgen af moet zijn
(I simply still have to start on an essay that has to be finished tomorrow)
b. Er is een liedje ie ik leuk vind
(There is a song that I like)

Using data obtained from the Corpus of Spoken Dutch, Audring (2006) argues that the Dutch pronominal gender agreement system is undergoing a change from a `grammatical, lexical' gender agreement system to a `semantic' system, especially for nouns that are high in the animacy/concreteness ranking. Language acquisition research has shown that both children and non-native, immigrant, speakers have a tendency to use the common gender definite determiner 'de' with neuter nouns. Cornips (2008) observes that this 'overgeneralization to the common gender' persists in the speech of some immigrant children that otherwise have native-speaker knowledge of Dutch, and that this phenomenon is a stylistic marker of certain forms of ethnically flavored Dutch. Previous research has not considered agreement with the relative pronoun in detail. Also, only corpora of limited size and relatively small samples of elicited speech have been used, making it hard to study the influence of (the semantic class of) the noun, grammatical context, or the speed at which this perceived language change is taking place. Using a large corpus of Twitter messages, we show that we are able to address all of these issues in detail.

Method

We use a large, representative, sample of Dutch tweets as a corpus of present day informal Dutch. The noisy character of the data, and the lack of linguistic (PoS, dependency) annotation, makes it challening to collect relevant cases automatically. We use the regular expression '(een|het) NEUTER-NOUN (die|dat)' to obtain relevant examples. The nouns that can take the position of NEUTER-NOUN have to be on a list of frequent neuter nouns. As diminutives always have neuter gender, these are included in the list as well.We manually filtered cases that give rise to large numbers of false hits, such as nouns that occur in fixed expressions (i.e. 'een beetje', somewhat), temporal expressions ('uur', hour), nouns that take a subordinate clause complement introduced by 'dat', and matches that are part of exclamatives ('wat een popje, die Matthijs', what a sweetheart, that Mathijs). We collected over 700,000 matching tweets, of which 37.5% contains the non-agreeing form 'die'. Manual verification of a sample of 500 'dat' and 500 'die' cases suggests that the error rate in both sets is between 14 and 15%. These are mostly cases where the antecedent of the pronoun is not the preceding noun, or where 'dat' is used as complementizer. It is re-assuring that the error rate is not biased towards the non-agreeing case, but we should be cautious when making claims w.r.t. specific nouns.

The semantic class of the 300 most frequent nouns in the data was determined manually, using the classification scheme suggested by Audring. We also considered the concreteness ratings for Dutch nouns of Brysbaert et al. (2014).

For a subset of the data, we were able to estimate demographic properties. Usernames that end in 19[0-9][0-9] are assigned the corresponding birthyear. When the prefix of a username matches with a member of a list of common Dutch gender-specific male and female first names, the user is assigned the corresponding gender. Where multiple matches are possible, we use the longest match. A Turkish or Maroccan ethnic background is assigned to those users whose screenname or profile matches with 'turk' or 'marok' (i.e. A0zturk, abdelhakimmarok).

Results

* Grammatical context plays a role. If the determiner preceding the antecedent noun is the neuter definite determiner 'het' non-agreement is 20.5%, but this goes up to 46.8% for cases where the indefinite determiner 'een' (which can be used with both common and neuter nouns) is used.

* The percentage of non-agreement varies strongly per antecedent noun, but in general non-agreement occurs more frequently with animate nouns, thus confirming the role of semantic class. Concreteness ratings show a slight positive correlation with the percentage of non-agreement.

* Morphology plays a role. The macro average for non-agreement for diminutives is 47%. This is significantly higher than the macro average for non-diminutive nouns, even if we take into account the fact that diminutives often refer to a human being (i.e. 'meisje', girl).

* The most important demographic factor is age. Young users use 'die' significantly more often than older users. Women use 'die' more often than men (39.3 vs 32.4%). Contrary to what Cornips (2008) suggests, with 31.9% non-agreement, we do not find that users with a Moroccan or Turkish background use 'die' more than other users.

* Average tweet and word length, type-token ratio, out-of-vocabulary rate, and keywords analysis suggest that non-agreement occurs especially in tweets that are highly informal.

Discussion

Using a large sample of current, informal, Dutch, we were able to show that for relative pronouns, 'generalising to the common gender' is a pervasive phenomenon, with some neuter nouns occurring with 'die' more than 50% of the time. Grammatical context, morphology, semantic class, and level of formality, influence the likelyhood that a neuter noun occurs with a non-agreeing relative pronoun. Ethnic background does not (or no longer) seem to play a significant role. In future work, we would like to investigate to what extent this trend is visible in other recent text corpora (i.e. web pages) as well.

References

Jenny Audring. Pronominal gender in spoken Dutch. Journal of Germanic Linguistics, 18(02):85-116, 2006.

Marc Brysbaert, Michael Stevens, Simon De Deyne, Wouter Voorspoels, and Gert Storms. Norms of age of acquisition and concreteness for 30,000 Dutch words. Acta psychologica, 150:80-84, 2014.

Leonie Cornips. Loosing grammatical gender in Dutch: The result of bilingual acquisition and/or an act of identity? International Journal of Bilingualism, 12(1-2):105-124, 2008.

A corpus-based analysis of pronoun choice in German relative clauses

Patrick Brandt (Institut für Deutsche Sprache)(brandt@ids-mannheim.de)

Eric Fuss (Institut für Deutsche Sprache)

Present-day (standard) German exhibits a range of competing strategies for introducing attributive relative clauses. Possible candidates for the use as relative pronoun include d-pronouns (der, die, das, etc.), which are similar in shape to demonstratives and constitute by far the most common option, wh-determiners (welcher, welche, welches, etc. 'which'), and pure wh-pronouns (*was* 'what', *wie* 'how', *wo* 'where' etc.). The distribution of the alternants is governed by a complex of interacting conditions concerning e.g. the type of head noun or properties of the gap in the relative clause, as well as extra-linguistic factors such as region, register choice, and considerations of style. Of particular interest is the case of neuter singular relativization, where all three options co-occur. While it is commonly agreed that the wh-determiner welch- 'which' is merely a stylistically marked alternative to the series of d-pronouns that is primarily found in the written language (cf. e.g. Duden 2009: 303), the variation between the d-pronoun das and the pure wh-relativizer was seems to be governed by grammatical factors that are poorly understood so far. According to traditional grammars (cf. e.g. Duden 2009), was replaces das in connection with indefinites, demonstratives, and nominalized adjectives:

(1) das Buch, das/*was er gelesen hat (note that *was* is acceptable is colloquial varieties)
the book that/what he read has

(2) a. alles/nichts/vieles, was/??das er sagt
all/nothing/much what/that he says
b. das/von dem, was/??das er sagt
that/of that-DAT what/that he says
c. das Gute, was/das er getan hat
the good what/that he done has
'the good things he did'

However, as indicated in (2), the respective contexts that give rise to relativization by means of was do not do so to the same extent. For instance, while was is almost compulsory after indefinites and demonstratives (2a,b), it co-varies with das in connection with nominalized adjectives (2c). This paper investigates the conditions under which the regular neuter singular relative pronoun das can (or must) be replaced by the wh-pronoun was. Based on a corpus study conducted in the Deutsches Referenzkorpus (DEREKO) with more than 18.000 relevant sentence tokens, and by using statistical methods including cluster analysis and the calculation of Pearson residuals, we present a more detailed picture of the distribution of relative was in present-day (standard) German. Major empirical results include:

(i) The most important factor governing the choice between d- and wh-relatives is the type of head element: Lexical head nouns require das, while in (all) other cases was is usually the preferred option.

(ii) Certain (individuating) indefinites (in particular *jedes* 'everything' and *keines* 'nothing') trigger das.

(iii) Another set of indefinites (in particular *etwas* 'something') allows both d- and wh-relatives; here the choice seems to be linked to semantic differences (*das* is used if *etwas* refers to an entity, while *was* signals an interpretation where *etwas* refers to a proposition or to a property).

(iv) Nominalized substantives do not behave uniformly: While positives allow both options (with d-forms being more frequent), superlatives regularly trigger was.

(v) In connection with nominalized positives, the presence of a quantifiying modifier (e.g. *alles Gute* 'all good') significantly enhances the share of was-relatives. A similar effect can be observed in connection with mass nouns.

We then argue that the distribution of das and was is most successfully captured by an approach that does not treat was as an exception but analyzes it as the elsewhere case instead that applies when identification with the antecedent cannot be established by syntactic means via agreement in phi-features. Our empirical results point to the conclusion that this is the case whenever there is no lexical nominal in the antecedent that, following Geach (1962) and Baker (2003), supplies a criterion of identity needed to establish sameness of reference between the antecedent and the relativizer. This approach facilitates a uniform treatment of wh-relative constructions that lack a nominal antecedent, including free relatives and attributive relatives that refer to predicates and propositions. Our analysis is couched within the framework of Distributive Morphology (Halle & Marantz 1993). We argue that was corresponds to a heavily underspecified phonological exponent that can only be chosen for insertion when the licensing requirements of the more specified candidate das are not met (cf. for a similar proposal regarding Dutch *wat* Boef 2013). Specifically, we propose that das signals the presence of a referential index that has been valued in the course of the syntactic derivation. We furthermore show how apparent cases of free variation between das and was in connection with nominalized adjectives relate to subtle semantic differences that are rooted in syntactic structure; these differences are the effect of a nominalizing syntactic head being combined with a minimal lexical root or with a derived syntactic structure respectively.

References

Baker, Mark. 2003. Lexical Categories. Cambridge: Cambridge University Press.

Boef, Eefje. 2013. Doubling in relative clauses: Aspects of morphosyntactic microvariation in Dutch. Utrecht: LOT.

Geach, Peter. 1962. Reference and Generality. Ithaca, NY: Cornell University Press.

Deutsches Referenzkorpus (DeReKo): http://www.ids-mannheim.de/cosmas2/.

Dudenredaktion (ed.). 2009. Duden: Die Grammatik. 8. Auflage. Mannheim/Leipzig/Wien/ Zürich: Dudenverlag.

Halle, Morris & Alec Marantz. 1993. Distributed Morphology and the pieces of inflection. In Kenneth Hale & Samuel J. Keyser (eds.), The View from Building 20, 111-176. Cambridge: The MIT Press.

Cognitive constraints on language variation: Agreement with English presentational 'there be' and Spanish presentation 'haber'

Jeroen Claes (K.U. Leuven, Quantitative Lexicology and Variational Linguistics)
(jeroenclaes@gmail.com)

Daniel Ezra Johnson (University of Southern California)

The most fundamental question of variational linguistics may well be: "What constrains variation?" In our presentation, we will explore the hypothesis that morphosyntactic variation is constrained by three general cognitive constraints on the spreading activation of constructions: markedness of coding, statistical preemption, and structural priming. Roughly, the first constraint consists in that constructions which provide a better match with the conceptual import speakers wish to express will reach a higher level of activation than their competitors (Langacker, 1991: 298). The second constraint implies that a highly accessible entrenched instance of a construction will reach a higher level of activation than a competing, more abstract construction, provided the entrenched construction and its more abstract competitor match the conceptualization equally well (Goldberg, 2006: 94). The third constraint indicates that once a construction has been activated, it remains more active than others for some period of time, favoring its use in subsequent variable contexts (Pickering & Ferreira, 2008).

To test whether and how these constraints impact linguistic alternations, we will present a case study of the agreement variation that is observed with plural NPs in the English presentational There be construction (e.g., There was/were problems) and the Spanish presentational haber construction (e.g., Había/habían problemas). We will argue that this agreement variation is symptomatic of a competition between, respectively, <There be Subj AdvP> and <There.Imp.Subj be Complement AdvP> and <AdvP haber Subj> and <AdvP haber Obj>.

Subsequently, we will use mixed-effects logistic regression and data from the spoken sections of the British National Corpus and a corpus of recording sessions with speakers of Puerto Rican Spanish to trace patterns of variability. The results show that relatively more prominent NP arguments (in terms of their definiteness/specificity and typical action-chain position) trigger agreement more often. Since subjecthood is claimed to be the grammatical expression of exceptional cognitive prominence (Langacker, 1991: 306), these results support that markedness of coding constrains the variation. For English, the present tense favors agreement. For Spanish, agreement is preferred in expressions that involve tenses other than synthetic present or preterit.

These results are contextualized and interpreted as evidence in favor of the hypothesis that statistical preemption constrains the variation. Speakers of both English and Spanish are also more likely to establish agreement with the NP when the previous token represents agreement, suggesting that (i) tokens involving different verb forms (e.g., is and was) activate the same constructions (Goldberg, 2006: 125) and (ii) that structural priming constrains the variation.

Furthermore, the conditional permutation of these variables in a random forest model reveals that the relative contributions of the predictors to explaining the variation are nearly identical for English and Spanish.

The combined facts that (i) the same predictors condition the variation in English and Spanish, (ii) that these predictors have the same directionalities of effects for both languages and (iii) that the predictors contribute in the same proportions to the fit of a random forest model make a strong case for considering the effects of these variables as reflexes of general cognitive constraints on spreading activation (cf. Tagliamonte, 2013).

Determiners and feature hierarchy in Old French

Monique Dufresne (Queen's University)(dufresne@queensu.ca)

Mireille Tremblay (Université de Montréal)

Rose-Marie Dechaine (University of British Columbia)

Research context

In the diachrony of French, we observe a change in the determiner (D) system such that in earlier stages of French, D was not obligatory, while in Modern French D is obligatory. Based on parsed corpus data from two OF texts — *Brendan* (*circa* 1106-1121) and *Marie de France* (*MdeF)* (*circa* 1154-1189) — we undertake a quantitative analysis of the distribution of D. The findings of previous qualitative research indicate that the following factors are relevant in tracking the distribution of D: (i) predicate/argument contrast; (ii) grammatical function, in particular subject/object contrast; (iii) semantic class, namely count/non-count contrast; (iv) definite/indefinite contrast; (v) number, namely the singular/plural contrast.

Findings

A quantitative analysis using Goldvarb confirms the following: (i) nominal predicate favour determinerless nouns (i.e. bare Ns); (ii) object position favours bare Ns; (iii) non-count (abstract and mass) favours bare Ns; (iv) indefiniteness contexts favour bare Ns; (v) plural marking favours bare Ns. In addition, the quantitative analyses reveal a conditioning factor that has not been previously noticed the qualitative literature, namely (vi) feminine gender favours bare Ns, but only in the earlier text. Not only does the quantitative analysis confirm the activity of these factors, it also allows us to measure their relative weight to describe the interplay between them, and thus track stable and changing properties of the grammar.

Stable factors Across the two texts, the relative weight of the factors relating to argument and grammatical function are stable, i.e. they do not change over time. We take this to indicate that these are stable properties of the grammar, and moreover that they reflect properties of Universal Grammar.

Changing factors We observe an increase in range of the factor group semantic class between the two texts. This increase is partly due to the surprising ***increase of bare non-count nouns*** (mass and abstract) in *MdeF*. As for definiteness and number, because there are no definiteness nor number contrasts with abstract and mass Ns, the crucial class to track is count Ns. Within the count N domain, definite and indefinite count nouns pattern differently: while the change appears nearly completed with definites, it is in progress with indefinites, showing a steady decrease of bare indefinites from *Brendan* to *MdeF*. Moreover, number is significant in both texts with indefinites, but significant only in *MdeF* with definites. As for gender, it is a significant factor only with *Brendan*'s indefinites.

Analysis We argue that the change in significance and relative weight of semantic class, number, and gender reflects a reorganization of the feature system, triggered by the reanalysis of *un* from a cardinality numeral ('one') to a marker of indefiniteness. The use of *un* as an indefinite marker forces a paradigmatic re-organization of the D-paradigm. With the introduction of the plural indefinite determiner *des* attested in *MdeF*, the paradigm of overt determiners is complete and number contrast is now formally encoded in D. Specifically, the D paradigm of *Brendan* is organized around a case contrast, and within this paradigm, number is part of a portmanteau morphemes that combine features for case, number and gender. In contrast, the D paradigm of *MdeF* is organized around a number contrast, to which case is subordinated. With plural count Ns, the suffixal case marker *-s* which codes either (1a) or (1b) in *Brendan*, is moving towards being re-analyzed as plural marking (2) in *MdeF*, where it is no longer subordinated to case.

(1) a. ***-s*** [CASE: NOMINATIVE; PERSON: 3; NUMBER: SG; GENDER: MASC]
b. ***-s*** [CASE: ACCUSATIVE; PERSON: 3; NUMBER: PLURAL]

(2) ***-s*** [PERSON: 3; NUMBER: PLURAL]

Consequences This change in feature hierarchy explains the surprising increase of bare mass and abstract nouns. We attribute this increase to the re-organization of the definite D paradigm, and more specifically to the neutralization of case inflection in favour of number marking. While the selection of a definite D in Brendan is conditioned only by grammatical function, in MdeF it is also conditioned by number. Mass and abstract Ns differ from count Ns in that only the latter show a singular/plural contrast. In *Brendan*, overt Ds are still number-neutral, and so can co-occur with mass and abstract Ns. In *MdeF,* the D paradigm encodes number contrasts, with overt Ds morphologically marked for singular and plural. Therefore overt definite Ds, which are no longer number-neutral, mark a singular/plural contrast, and so are no longer compatible with mass and abstract Ns which, because of their semantic properties resist individuation.

The change in feature hierarchy also accounts the novel observation that GENDER is no longer a significant factor in *MdeF*'s indefinites. In Old French, only masculine nouns display a case contrast, that is, the feature hierarchy in (1) only applies to masculine nouns, creating a gender asymmetry in the system. The reanalysis of the suffixal case marker *–s* as plural marking yields a uniform case-neutral system in both genders (2), thereby accounting for the disappearance of the gender effect.

Conclusion

The gradual disappearance of bare Ns in French reflects the emergence of paradigmatically conditioned D. Confirming this requires the integration of two research methods: qualitative methods to identify which linguistic factors participate in the change; quantitative methods to measure the relative significance of each factor group.

References

Boucher, P. (2005). *Definite Reference* in *Old* and *Modern French:* The *Rise* and *Fall* of *DP,* in Batllori et al. (eds), *Grammaticalization and Parametric Variation*. New York: OUP.

Carlier, A. (2013) Grammaticalization in Progress in Old French: Indefinite Articles, in Deborah L. Arteaga (ed.), *Research on Old French: The State of the Art*. Dordrecht: Springer.

Carlier, A. (2007) From preposition to article: The grammaticalization of the French partitive, *Studies in Language* 31, 1-49.

Foulet (1928/1974) *Petite syntaxe de l'ancien français*. Paris: H. Champion.

Mathieu, E. (2009). From local blocking to cyclic Agree: the role and meaning of determiners in the history of French, dans Ghomeshi et al. (eds.), *Determiners: variation and universals*, John Benjamins.
Moignet, G. (1976) *Grammaire de l'ancien français.* Paris : Librairie Klincksieck.

Persian Phrase Structure:
Lessons from Studies on Word Order Variations

Pegah Faghiri (Université Sorbonne Nouvelle - Paris 3)(pegah.faghiri@gmail.com)

Pollet Samvelian (Universit de Paris 3)

In this paper, we present a usage-based study to constituent ordering in Persian in line with studies on word order variations that take into account functional factors. The results of our empirical study, combining corpus-based and experimental data, goes against the widespread theoretical view of Persian's phrase structure. More precisely, our results undermine the backbone argument put forward to support this view which consist of a broadly admitted claim regarding the relative order between objects in a ditransitive sentence.

Persian is an SOV language with DOM triggered by definiteness and/or specificity and marked by the enclitic *=rā*, (1). This enclitic is also used as a topicalizer, placed in the initial position, beyond DO, (2). It should be noted that Persian do not formally mark nouns for definiteness, a single noun without any formal determination, *ketāb*, can receive two reading (except in the DO position): 1) bare noun, that is, a noun lacking any determination or quantification, nonspecific and under-specified for number, 'a book, some books'; 2) definite noun 'the book'. (Existential) indefiniteness is, however, formally marked by the determiner *ye(k)*, *ye ketāb*, the enclitic *=i*, *ketāb=i*, or the combination of both, *ye(k) ketāb=i*, 'a book'. Furthermore, disposing only of around 250 simplex verbs, verbal concepts are mainly expressed in Persian by complex predicates (cf. Samvelian 2012), that is the combination of a simplex verb and a non-verbal element, prototypically involving bare nouns (*dars xāndan* 'to study'), (2).

(1) a. Maryam ketāb xarid
Marya book bought-3SG
'Maryam bought a book/some books.'
b. Maryam ketāb=rā xarid
Maryam book=DOM bought-3SG
'Maryam bought the book.'

(2) tābestān=rā dars mi-xān-am
summer=DOM lesson IPFV-read.PRS-3SG
'Summer, I will study.'

Most theoretical studies, namely in the framework of the generative grammar, postulate a phrase structure for Persian reflecting a structural asymmetry between rā-marked and non-rā-marked DOs (e.g. Brown & E. Karimi 1994, Ghomeshi 1997, Karimi 2003, 2005). Despite their substantial difference, these studies posit two different positions, either base-generated or resulting from a movement, for rā-marked vs. non-rā-marked DOs, roughly represented in (3). Furthermore, they share the same methodological approach, based mainly, if not exclusively, on the authors (categorical) grammaticality judgments.

(3) a. [VP DP[+rā] [V' PP V]]
b. [VP [V' PP [V' DP[-rā] V]]]

Karimi (2005) provides the most thorough argumentation in favor of this analysis. Considering rā-marking a matter of specificity, the author assimilates non-rā-marked DOs, to bare nouns, minimizing their significant differences, and puts forward a number of observations suggesting syntactic and semantic asymmetries between rā-marked and bare DOs to support this analysis. Namely, the semantic incorporation of non-rā-DOs into the verb, similar to the combination involved in CPs, and their lack of (semantic and syntactic) autonomy contrary to rā-marked DOs, and the unmarked word order between the DO and the IO, which is, broadly assumed to follow the schema in (4).

(4) a. DO[+rā] IO V
b. IO DO[-rā] V

Adopting a truly empirical approach, we have studied the relative order between the DO and IO in Persian in line with studies on word order preferences (e.g. Hawkins 1994, Arnold et al 2000, Yamashita & Chang 2001, Wasow 2002, Bresnan et al 2007). Including a preliminary multifactorial corpus study and follow up (off-line) psycholinguistic experiments conducted via web-based questionnaires, our study takes into account functional factors shown to play significant role in word order preferences cross-linguistically such as length, givenness, humanness, verbal lemma, via mixed-effect regression modeling.

Most importantly, our results show that the oversimplifying dual classification of DOs based on markedness is flawed: in comparable context, while rā-marked DOs show a strong preference for the DO-IO order and bare DOs for the inverse, the picture is less clear-cut for in-between DOs. In one hand, indefinite/quantified (non-rā-marked DOs) - contrary to what is expected by the theory - show a clear preference, although moderate, for the DO-IO order, grouping with rā-¬marked DOs. In the other, bare DOs carrying adjuncts, show a much less strong preference for the IO-DO order. More generally, extra-syntactic factors such as relative length and humanness show up to play a significant role. Indeed, ordering preferences between the DO and the IO, rather than being dichotomous, are best reflected by a continuum based on the degree of determination of the DO - closely related to its discourse accessibility (cf. Gundel et al. 1993) - as well as other (functional) factors related to the conceptual accessibility, such as relative length (cf. Yamashita & Chang 2001) and humanness.

Furthermore, examining all other arguments put forward, we conclude that there is no empirical ground for positing two syntactic positions for the DO. Accordingly, the dual phrase structure analysis for Persian that provides wrong predictions with respect to the unmarked word order is refuted.

This study thus highlights the importance of using empirically solid methods in theoretical syntax.

References

Arnold, J. E. et al. 2000. Heaviness vs. newness: The effects of complexity and information structure on constituent ordering. Language 76(1). 28–55.

Bresnan, J. et al. 2007. Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts (ed.), Cognitive foundations of interpretation, 69–94.

Browning, M. & E. Karimi.1994. Scrambling to object position in Persian. In N. Cover, H. van Riemsdijk (eds), Studies on scrambling, 61-100.

Ghomeshi, J. 1997. Topics in Persian VPs. Lingua 102(2). 133–167.

Gundel, J. et al. 1993. Cognitive status and the form of referring expressions in discourse. Language 69 (2): 274 – 307.
Hawkins, J.A. 1994. A performance theory of order and constituency. Cambridge University Press
Karimi, S. 2003. Object positions, specificity and scrambling. In Karimi, S. (ed.) Word Order and Scrambling, 91-125.
Karimi, S. 2005. On object positions, specificity, and scrambling in Persian. In Simin Karimi (ed.), Word order and scrambling, 91–124. Blackwell Publishing.
Samvelian, Pollet. 2012. Grammaire des prédicats complexes: les constructions nom-verbe, Hermès–Lavoisier.
Wasow, T. 2002. Postverbal behavior. CSLI lecture notes.
Yamashita, H., & Chang, F. 2001. 'Long before short' preferences in the production of a head final language, Cognition 81(2): B45-B55.

Syntactic variation and probabilistic indigenization in World Englishes

Jason Grafmiller (KU Leuven)(jason.grafmiller@kuleuven.be)

Benedikt Heller (KU Leuven)

Melanie Röthlisberger (KU Leuven)

Benedikt Szmrecsányi (KU Leuven)

We report on an ongoing project combining variationist methods (e.g. Tagliamonte 2001) with the Probabilistic Grammar framework (e.g. Bresnan et al. 2007) to study syntactic variation within and across nine varieties of English from around the world: British, New Zealand, Canadian, Irish, Indian, Singaporean, Hong Kong, Philippines, and Jamaican English. Our main interest lies in the extent to which language users' grammatical knowledge differs across these speech communities. In particular, we seek to shed light on questions regarding the extent to which different English varieties share a core grammar, the degree to which individual probabilistic constraints are cross-lectally malleable, and the degree to which certain syntactic alternations exhibit cross-constructional parallelisms. In treating variation as a "core explanandum" (Adger and Trousdale 2007: 274), the project is thus situated at the crossroads of variationist linguistics, usage-based theoretical linguistics, cognitive linguistics, and dialectology.

Our case study explores three well-known patterns of syntactic variation in English, drawing on data from the International Corpus of English (ICE): the genitive alternation (*Mary's speech* versus *the speech of Mary*), the dative alternation (*Tom sent Mary a letter* versus *Tom sent a letter to Mary*), and particle placement (*Tom looked the word up* versus *Tom looked up the word*). In total, our data comprise 8-10k observations of each construction, which were semi-automatically extracted from the ICE corpora: N = 10594 genitives; N = 8549 datives; N = 8072 particle verbs. We carefully circumscribe the variable contexts to identify variants in the corpora, annotating for numerous factors conditioning the choice of constructional variant. Such factors include the semantics of verbs (and particles), as well as the animacy, frequency, definiteness, givenness, and length (in characters) of all construction constituents. Throughout the extraction and annotation process, both automatic (via Perl/Python scripts) and manual coding techniques are employed to maximize efficiency and accuracy. We then use multivariate techniques such as generalized linear mixed-models (Pinheiro and Bates 2000) to investigate variable effects in the conditioning factors that constrain syntactic choices. The effects of such conditioning factors, e.g. the tendency for longer constituents to follow shorter ones, can be seen as stochastic generalizations about language usage, which – according to experimental evidence (Bresnan and Ford 2010) – language users implicitly know about. Thus, we aim to illuminate the variability in the linguistic knowledge that language users with differing English backgrounds implicitly command.

Overall, mixed-effects logistic regression models fit the data quite well for each construction: genitives model, C = .98; datives model, C = .97; particle verbs model, C = .92. We find that the direction of the influence of individual factors is largely consistent across all nine varieties, but that a few cross-varietal differences in the strength of their influence nevertheless emerge in small corners of the data. Such differences are more likely to emerge in outer circle (non-native) varieties, e.g. Hong Kong or Indian English, than inner circle varieties, e.g. Canadian English, but there are few consistent patterns with respect to which factors differ in specific varieties. Interestingly, we also find that some alternations, e.g. particle placement, involve a greater number of cross-lectally varying factors than others, suggesting that different alternation phenomena are not equally amenable to cross-varietal innovation. We speculate that variation in expectation-based processing effects, which are grounded in usage probabilities, could explain the variation exhibited across different varieties. This is supported by recent work showing that the more a constructional alternate is lexically entrenched (i.e. occurs with the same lexical items, Hoffmann 2014) the more likely it is to exhibit cross-varietal variation (AUTHORS forthcoming). Furthermore, an alternation's degree of entrenchment in a given variety correlates inversely with that variety's placement along the path of development in Schneider's (2007) Dynamic Model of nativization. This model predicts that innovations in developing varieties of English typically occur at the syntax-lexicon interface, where new patterns emerge as differences in the habitual associations of constructions with specific lexical items. Results of our models thus reflect the reshaping of stochastic patterns of internal linguistic variation due to shifting usage frequencies among speakers of post-colonial varieties, a process we refer to as `probabilistic indigenization'. We argue these findings support a probabilistic model of linguistic knowledge which is shaped both by general, higher-level cognitive factors as well as by surface level, community-specific usage norms.

(Complete list of datasets by variety, conditioning factors in our models, and model summary statistics are provided online at http://tinyurl.com/oct94j8)

References

Adger, D. and G. Trousdale (2007). Variation in English syntax: Theoretical implications. English Language and Linguistics 11, 261–278.

Bresnan, J., A. Cueni, T. Nikitina, and B. Harald (2007). Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts (Eds.), Cognitive Foundations of Interpretation, pp. 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, J. and M. Ford (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. Language 86(1), 168–213.

Hoffmann, T. (2014). The cognitive evolution of Englishes: The role of constructions in the dynamic model. In S. Buschfeld, T. Hoffmann, M. Huber, and A. Kautzsch (Eds.), Varieties of English Around the World, Volume G49, pp. 160–180. Amsterdam: John Benjamins Publishing Company.

Pinheiro, J. C. and D. M. Bates (2000). Mixed Effects Models in S and S-PLUS. New York: Springer.

Schneider, E. (2007). Postcolonial English: Varieties Around the World. New York: Cambridge University Press.

Tagliamonte, S. (2001). Comparative sociolinguistics. In J. Chambers, P. Trudgill, and N. Schilling-Estes (Eds.), Handbook of Language Variation and Change, pp. 729–763. Malden and Oxford: Blackwell.

Jason Grafmiller, Benedikt Heller, Melanie Röthlisberger and Benedikt Szmrecsányi (forthcoming). Syntactic alternations, schematization, and collostructional diversity in World Englishes.

Syntactic variation and change in the rise of the shell noun construction in 16th and 17th century Spanish: presence vs. absence of the preposition de in nominal complement clauses (N que vs. N de que)

Anton Granvik (University of Gothenburg)(anton.granvik@sprak.gu.se)

In present day Spanish, the so-called shell content construction (Schmid 2000) comes in three versions, or subconstructions, as indicated by examples 1a-1c including one of its most salient nouns, *idea* 'idea':

(1) a. La idea de ir me parece estupenda.
'The idea of going seems fabulous (to me).'
b. La idea de que vengas me parece estupenda.
'The idea that you'll come seems fabulous.'
c. La idea es que vayamos todos juntos.
'The idea is that we'll all go together.'

The focus of this paper is the b version of the construction, which has undergone a peculiar formal change in the history of Spanish, i.e. the preposition *de* has been introduced before the complementizer *que*. When this subconstruction first appeared in Spanish, from the mid 13th century onwards, it did so in the schematic format N *que*. However, as the use of the shell noun construction increased in the 16th century this process was accompanied by the introduction of the preposition *de* before the complementizer (Girón 2004, Serradilla 2010).

The expansion of the N *(de) que* subconstruction in Spanish thus takes place in the middle of a period of syntactic variation in the formal expression of clausal complements. This variation is shown in examples 2a and 2b including the shell noun *señal* 'sign, signal'. Both examples are found in the same work by late 16th century writer Alonso de Villegas:

(2) a. *y ofreció su cuerpo y alma a los demonios, y* ***en señal que*** *se les entregava, arrojó la moneda que le quedava.* (Villegas, Alonso de, Fructus sanctorum y quinta parte del Flossanctorum, 1594)
'and he offered his body and soul to the demons, and as a sign that he offered the, he threw to coin he had left.'

b. *Por el amor del mismo Jesucristo, yo te perdono. -Y* ***en señal de que*** *le perdonava, le levantó de tierra y le dio beso de paz.* (Villegas, Alonso de, Fructus sanctorum y quinta parte del Flossanctorum, 1594)
'For the love of Jesus Christ himself, I forgive you. –And as a sign [of] that the forgave him, he raised him from the ground and gave him a kiss of peace.'

The aim of this paper is to address this specific case of syntactic variation leading to change, with the goal of finding a motivation for how the new variant, N *de que* came to surpass the earlier one, N *que*. Building on data from the Corpus del español (Davies 2002-) and Corpus Diacrónico del Español (CORDE), which are the most extensive diachronic corpora of Spanish available today, and range from the medieval period up to the end of the 20th century, and by running parallel collostructional analyses (Stefanowitsch & Gries 2003), I identified the 78 most important shell nouns in Spanish. Retrieving all cases of 31 of these nouns, which occur among the top 30 nouns in at least two centuries, from CORDE, nine of them were identified by means of a Distinctive Collexeme Analysis (Gries & Stefanowitsch 2004) as being of particular importance for the early alternation between the two subconstructions, namely *causa, duda, esperanza, fe, opinion, recelo, señal, sospecha*, and *temor*.

The main focus of this study, then, is the use of the nine nouns mentioned above in the N *que* and N *de que* variants from 1510 to 1699. This period was identified as the first of three salient groupings in a VNC analysis (Gries & Hilpert 2008, Hilpert 2012) performed on the data including the 31 nouns.

The hypothesis for explaining the presence vs. absence of *de* in the sentential complements of these nine nouns, is that the N *que* variant correlates with what I have called complex predicates, i.e. semi-lexicalized verb+noun combinations. In contrast, I hypothesize that the N *de que* variant is more frequent when the noun is used independently, as indicated e.g. by the presence of a determiner.

A sample consisting of 1123 cases of the nine nouns in both constructional variants was annotated for contextual features aiming to capture this hypothesized difference. In addition to characterizing the grammatical usage context, I also took into account extra-linguistic factors such as the different authors and textual genres. The annotated sample was the subjected to a (binomial) logistic regression analysis.

The results indicate that the N *que* variant is significantly associated with the absence of determiners and the noun functioning as direct object. These findings clearly seem to support the hypothesis. Second, looking at the extra-linguistic factors, the two constructional variants were associated to a varying degree with different textual genres: the prepositional variant is preferred in Didactic and Newspaper texts, whereas the N *que* variant predominates in Scientific texts. Third, different authors also show distinct preferences for the two variants, some relying exclusively or predominantly on either N *de que* or N *que* (innovators vs. conservatives), and a (smaller) group using both variants (undecided).

References

CORDE = REAL ACADEMIA ESPAÑOLA: Banco de datos (CORDE) [en línea]. Corpus diacrónico del español. <http://www.rae.es>

Davies, Mark (2002-): Corpus del español: 100 million words, 1200s-1900s. Online: <http://www.corpusdelespanol.org>.

Girón Alconchel, José Luis (2004): «Cambios gramaticales en los Siglos de Oro», in Rafael Cano Aguilar (coord.): Historia de la lengua española. Barcelona: Ariel, 859-893.

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

Gries, Stefan Th./Martin Hilpert (2008): "The identification of stages in diachronic data: variability-based neighbour clustering", Corpora Vol. 3 (1): 59-81.

Hilpert, Martin (2012): “Diachronic collostructional analysis: How to use it and how to deal with confounding factors”, in K. Allan (ed.) Topics in English Linguistics: Current Methods in Historical Semantics. Berlin: de Gruyter.
Schmid, Hans-Jörg (2000): English Abstract Nouns as Conceptual Shells. From Corpus to Cognition. Berlín/Nueva York: Mouton de Gruyter.
Serradilla Castaño, Ana María (2010): “La subordinación completiva en español antiguo: continuidad y ruptura de los modelos latinos», in Mónica Castillo Lluch and Marta López Izquierdo (eds.): Modelos latinos en la Castilla medieval. Madrid-Frankfurt am Main: Iberoamericana-Vervuert, 145-158.
Stefanowitsch, Anatol/Stefan Gries (2003): “Collostructions: Investigating the interactions of words and constructions”, International Journal of Corpus Linguistics 8(2), 209–243.

Barking up the right tree: Idiomatic constructions and syntactic domains in English and Dutch

William Harwood (KU Leuven)(himbol@gmail.com)

Tanja Temmerman (KU Leuven)

1 Summary

This paper aims to show that the structural size of syntactic domains is a source of cross-linguistic variation. This is primarily evidenced by idiomatic constructions in English and (dialectal) Dutch: it is shown that the syntactic material upon which the idiomatic interpretation is dependent varies cross-linguistically.

2 Background: VP-domain

It is commonly agreed that the lexical verb, together with its internal arguments, forms a syntactic domain, typically referred to as the "VP-domain", which exhibits a certain degree of syntactic, semantic, and phonological independence.

This is, for instance, evidenced by idioms. Idioms are expressions with a non-compositional interpretation: their meanings are not simply predictable from the literal meanings of their parts. A canonical example is *kick the bucket* ('to die'), the meaning of which has nothing to do with kicking or buckets. Crucially, if any of the parts are altered, the figurative interpretation is lost: neither *kick the can* nor *knock the bucket* means 'to die'. It has long been noted (e.g. Chomsky 1980) that verbal idioms are typically comprised of the verbal predicate and its internal arguments: *bite the dust* (DP-theme), *come to the point* (PP-goal), *all hell breaks loose* (DP-theme subject). Svenonius (2005) observes that, although idioms can co-occur with other syntactic material, such as aspect/voice/tense, the idiomatic interpretation is never dependent on the presence of these items. There thus seems to be a strict separation between the VP-domain and the domain of tense/aspect: a verb forms an idiom with its internal arguments, i.e. material internal to the VP-domain, but not with material generated outside of this domain.

This VP-domain can be defined through various means, such as the clause-internal phase (Chomsky 2000), the predicational layer (Grohmann 2003), the event layer (Ramchand & Svenonius 2014), to name but a few. This domain is often considered to be universal: in all languages, the lexical verb forms a domain together with its internal arguments, to the exclusion of higher syntactic material.

3 Cross-linguistic variation: verbal idioms

However, we argue that the size of the "VP-domain" should be considered a point of cross-linguistic variation: in English, this domain actually extends as far as the progressive aspectual layer, so as to include both passive voice and progressive aspect, while in Dutch, this domain extends as far as the perfect aspectual layer.

3.1 English This claim is evidenced in English by the fact that certain verbal idioms *are* reliant on additional syntactic material. A number of idioms depend on passive voice (Bowers 2010), i.e. when passive voice is absent, the figurative interpretation is lost:

(1) a. He was hoisted by his own petard. = He was foiled by his own scheming.
b. His own petard hoisted him. ≠ He was foiled by his own scheming.

Moreover, many idioms exist in English that depend on progressive aspect:

(2) a. Bob is pushing up daisies. = Bob is dead.
b. Bob pushed up daisies. ≠ Bob is dead.

Despite extensive research, there appear to be no real idioms in English that are dependent on perfect aspect or tense. Thus, in English, it seems that the progressive aspect layer constitutes the upper boundary of the traditional VP-domain.

3.2 Dutch In order to determine the size of Dutch idioms, we systematically study data gathered from a survey of 10 dialect dictionaries, informal judgment tasks and formal interviews with native speakers of 13 dialects (in Flanders and The Netherlands).

In (dialectal) Dutch, verbal idioms reliant on passive voice and progressive aspect are attested as well:

(3) a. *van 'n éizel op kèiremis beschéite* ***wèrre***
of the donkey on fair shat become
'become pregnant' (lit. 'be shat on by a donkey at the fair')

b. *# 'n éizel beschéit éir op kèiremis*
the donkey shat her on fair
≠ 'she became pregnant'
[Dendermonde Dutch]

(4) a. *Hij is aan het poepgaain.*
he is on it poo-jay-ing
'He's doing nothing.'

b. # *Hij poepgaait altijd.*
he poo-jays always
≠ 'He's doing nothing'
[Brugge Dutch]

(Dialectal) Dutch idiomatic expressions differ from English idioms, however, in that they can also be dependent on perfect aspect:

(5) a. *Z'**heit** teigen den hoek van een ronne taufel geloeipen.*
she-has against the corner of a round table run
'She's pregnant and she doesn't know who the father is.'
(lit. 'she has run against the corner of a round table')

b. # *Ze liep teigen den hoek van een ronne taufel.*
she ran againstthe corner of a round table
≠ 'she was/is pregnant'
[Aalst Dutch]

Despite extensive research, there appear to be no idioms in Dutch that are dependent on particular tense forms. Thus it can be concluded that, in Dutch, the perfect layer constitutes the upper boundary of the traditional "VP-domain".

3.3 Theoretical implications If verbal idioms are indeed constrained by the "VP-domain" (e.g. Chomsky 1980, Svenonius 2005), then the data above imply that this domain in English extends as far as the progressive layer, and as far as the perfect layer in Dutch. This means that syntactic domains are not rigid and absolute, as Chomsky (2000) claims, but are instead a point of cross-linguistic variation.

4 Consequences and supporting evidence

The advantage of this claim is that it provides us with a new tool to analyse many forms of cross-linguistic syntactic variation. We will show, for instance, how this understanding can be used to explain the differences in (size of the) ellipsis site between VP ellipsis in English, and modal complement ellipsis in Dutch.

References

Bowers, J. 2010. *Arguments as Relations*. Cambridge: MIT Press.

Chomsky, N. 1980. *Rules and Representations*. New York: Columbia University Press.

Chomsky, N. 2000. Minimalist inquiries: The framework. In R. Martin, D. Michaels & J. Uriagereka (eds.), *Step by Step*. Cambridge: MIT Press. 89-155.

Grohmann, K. 2003. *Prolific domains: On the anti-locality of movement dependencies.* Amsterdam: John Benjamins.

Ramchand, G. & P. Svenonius. 2014. Deriving the functional hierarchy. *Language Sciences* 46:152-174.

Svenonius, P. 2005. Extending the extension condition to discontinuous idioms. *Linguistic Variation Yearbook* 5:227-263.

Does word order affect attention to changes in complement clauses? Testing a semantic hypothesis experimentally.

Tanya Karoli Christensen (University of Copenhagen)(tkaroli@hum.ku.dk)

Torben Juel Jensen (University of Copenhagen)

Marie Herget Christensen (University of Copenhagen)

In modern spoken Danish, we find a high degree of variation between so-called "main clause" and "subordinate clause" word order in complement clauses as determined by the relative ordering of finite verb and sentence adverbials: In short, main clause word order is Verb > Adverb (1), and subordinate clause word order is Adverb > Verb (2) (cf. Diderichsen 1946 and later versions).

1. så vidste jeg jo bare at der **var ikke** noget mellem ham og hende
   then knew I JO just that there was not anything between him and her
   'then of course I just knew that there was nothing between him and her'
2. så vidste jeg jo bare at der **ikke var** noget mellem ham og hende
   then knew I JO just that there not was anything between him and her
   'then of course I just knew that there was nothing between him and her'

The use of Main Clause Phenomena (MCP) in complement clauses has been studied for a range of languages and from diverse theoretical perspectives, and there is wide agreement that the distribution of MCP to a large extent is conditioned by semantic or discourse-pragmatic factors (e.g. Hooper and Thompson 1973; Meinunger 2006; Wiklund, Bentzen et al. 2009; Aelbrecht, Haegeman et al. 2012). One line of research hypothesizes that main clause word order in subordinate clauses signals that the more important information, sometimes called the 'main point of the utterance' or the 'foreground' is to be found in the subclause, rather than in the main clause (e.g. Boye and Harder 2007; Simons 2007; Wiklund, Bentzen et al. 2009; Jensen and Christensen 2013).

Previous observational studies of spoken language data in Danish have shown a range of statistically significant, linguistic predictors for the use of Verb > Adverb word order in complement clauses (Jensen and Christensen 2013). Two of the most important predictors are presence or absence of a subordinator (typically at 'that') and the type of matrix clause predicate. SUBORDINATOR ABSENCE clearly favours Verb > Adverb (i.e. main clause) word order (cf. Thompson and Mulac 1991); and while COMMUNICATIVE and COGNITIVE predicates also favour the use of Verb > Adverb word order, FACTIVE (i.e., presupposition-generating) predicates disfavour them (Kiparsky and Kiparsky 1971). Semi-factive predicates allows for two readings (cf. Simons 2007) and occupy the middle ground regarding the distribution of the two word orders (Jensen and Christensen 2013).

A reasonable extension of the foregrounding hypothesis is that a foregrounded clause will attract more attention from listeners or readers (in psycholinguistic terms: comprehenders) than a backgrounded clause. In order to test whether comprehenders are in fact more attentive to complement clauses with main clause word order than to those with regular subordinate clause word order, we have performed an experiment under the Text Change-paradigm. This paradigm assumes that comprehenders are disinclined to notice change made to already processed discourse, so-called change blindness. However, the degree to which change blindness occurs is affected by a number of linguistic and cognitive factors, including the age of the comprehender (Price 2008) and well as sentential complexity (Sanford, Sanford et al. 2005). Linguistic factors that capture the attention of the comprehender are found to reduce change blindness. This suggests that discourse under the scope of attention capturing devices is processed fuller and more deeply than discourse which is not (Sanford, Sanford et al. 2006), which makes the testing paradigm relevant for the hypothesis of the semantico-pragmatic function of word order in subordinate clauses. If Verb > Adverb is a foregrounding signal, we would expect foregrounded clauses to cause less change blindness than non-foregrunded clauses.

In a first round of experiments using a text change-design, 38 students each read 24 complex sentence constructions twice. Each construction contained a complement clause with either Verb > Adverb or Adverb > Verb order. Half of the complement clauses were governed by a semifactive predicate (open to both word orders) and the other half by what we call a semantically secondary predicate (favouring Verb > Adverb word order). All complement clauses were initiated by the subordinator at 'that' to ensure that comprehenders read the relevant clause as a complement rather than as a main clause.

Attention to the complement clause was tested by measuring how disinclined they were to notice change of words in the complement clause when re-reading the same construction after a minimal break (see an English example in 3).

3. First presentation: then she realised that I did not have any sort of plan
   Second presentation: then she realized that I did not have any kind of plan

A mixed-models statistical analysis showed significantly less change blindness with Verb > Adverb complements than with Adverb > Verb complements. We consider this a preliminary confirmation of the semantico-pragmatic hypothesis that main clause word order functions as a foregrounding signal for comprehenders.

We will discuss possible objections to the design, among them the fact that the stimuli were presented in written form, whereas the foregrounding hypothesis relates to spoken language.

References

Aelbrecht, et al. (2012). Main Clause Phenomena. John Benjamins.

Boye and Harder (2007). Complement-taking predicates. Studies in Language 31(3).

Diderichsen (1946). Elementær Dansk Grammatik. Gyldendal.

Hooper and Thompson (1973). On the Applicability of Root Transformations. Linguistic Inquiry 4(4).

Jensen and Christensen (2013). Promoting the demoted. Lingua 137.

Kiparsky and Kiparsky (1971). Fact. Steinberg and Jakobovits (eds.) Semantics. CUP.

Meinunger (2006). On the discourse impact of subordinate clauses. Molnár and Winkler (eds.) The Architecture of Focus. Mouton de Gruyter.

Price (2008). The use of Focus Cues in Healthy Ageing. Univ. Glasgow.

Sanford et al. (2005). Depth of lexical-semantic processing and sentential load. Journal of Memory and Language 53(3).
Sanford et al. (2006). Shallow processing and attention capture in written and spoken discourse. Discourse Processes 42(2).
Simons (2007). Observations on embedding verbs, evidentiality, and presupposition. Lingua 117(6).
Thompson and Mulac (1991). The discourse conditions for the use of the complementizer that in conversational English. Journal of Pragmatics 15.
Wiklund, Bentzen et al. (2009). On the distribution and illocution of V2 in Scandinavian that-clauses. Lingua(119).

A production bias model of the Constant Rate Effect

Henri Kauhanen (The University of Manchester)

(henri.kauhanen@manchester.ac.uk)

George Walkden (The University of Manchester)

Kroch (1989) advanced the hypothesis that when two grammatical options compete across a number of linguistic contexts and one replaces the other over time, the rate of change will be the same in all contexts. To date, this hypothesis has been studied in a number of languages and data sets (Kroch 1989; Santorini 1993; Taylor 1994; Pintzuk 1995; Postma 2010; Wallage 2013) and has accumulated enough support for it to be referred to as the Constant Rate Effect, or CRE (e.g. Pintzuk 2003).

CREs provide a fresh perspective on causation in syntactic change: evidence of a CRE is evidence against the view (e.g. Bailey 1973) that linguistic innovations adapt to linguistic contexts based on their functionality; instead, patterns of use observed in historical data are to be thought of as reflexes of more abstract, underlying grammatical changes. Despite the wealth of empirical studies that over the years have sought to establish CREs in historical data, this central intuition of Kroch (1989) has, however, never been explicated formally in a detailed model of change that takes both grammatical competition and contextual effects into account.

What is more, certain doubts have recently been raised concerning the standard way of detecting CREs in corpus data, which is to fit a number of independent logistic curves, one per each context of interest. Firstly, (1) Wallenberg (2015) and Willis (2015) show that, using this method, CREs can be empirically demonstrated in situations where they cannot be taken to support underlying grammatical unity: across languages and across geographical areas, respectively. On the other hand, (2) customary research practice in diachronic syntax has long acknowledged that fitting a number of independent logistic curves to a set of contexts leaves variation in the time dimension entirely unexplained: it would, in principle, be possible to establish a CRE across two contexts where the change goes to completion in one before it even takes off in another (see Figure 1 in the supplementary document). Together, problems (1) and (2) imply that the standard operationalization of CREs is neither necessary nor sufficient for assuming that a single underlying change has occurred.

In this talk, we aim to overcome these problems and to shed light on the nature of causation in language change by introducing a model of the CRE that is more tightly constrained, and therefore makes stronger (more restricted) empirical predictions than the traditional formulation. Starting with Yang's (2000) mathematical model of grammar competition, we augment the model with production biases across an arbitrary number of linguistic contexts. We show that this extension of Yang's framework naturally gives rise to the CRE in computer simulations (Figure 2). Crucially, however, it is a theorem of the model that the time separation possible between any two contexts of one underlying grammatical change has a finite upper bound which is inversely proportional to the rate of the underlying change. This time separation theorem overcomes problem (2) identified above, and invites us to reconsider a number of data sets in which CREs have previously been studied using the independent logistics operationalization.

For this purpose, we introduce a novel curve-fitting algorithm based on nonlinear least squares regression (Bates & Watts 1988).

We investigate the model in the light of historical data by focussing on two changes for which a CRE has previously been established using the method of independent logistics: the rise of periphrastic do in English (Kroch 1989) and, to take a phonological example to illustrate the generality of the procedure, the loss of final fortition in Early New High German (Fruehwald, Gress-Wright & Wallenberg 2009). We show that the fit of our model to these data is no worse than a fit made using the traditional method (Figure 3). Crucially, however, our model implies a maximal time separation for each change, which we also test, finding that the empirically observed time separations fall within the range prescribed by our model (Figure 4).

We therefore show that a more constrained model of the CRE can fit historical data no worse than a less constrained one, and that it also generates new empirical predictions, also in line with the data, in the form of the time separation theorem. To complement these results, we investigate a number of pseudo-CREs - data sets that appear to exhibit a CRE if probed using the traditional method of independent logistics but that plausibly cannot due to unassailable a priori grounds (see problem (1), above). We show that here, when quantified by the residual error of the regressions, our model gives consistently worse fits than the traditional method, as desired (Figure 3).

Finally, we discuss a number of additional predictions the model makes about change in the presence of contextual biases. In brief, we show that in this extended model Yang's (2000: 239) Fundamental Theorem of Language Change ceases to hold, so that a distributional difference in the proportion of sentences parsed by two competing grammars is neither a sufficient nor a necessary condition of change on its own: the production biases induce a bifurcation in the parameter space of the model, and whether an innovatory grammatical option overtakes a conventional one comes to depend on a nonlinear interaction of grammar advantages (as defined in Yang 2000) and the magnitude and direction of the production biases. Conducting a full bifurcation analysis of the two-grammar case of the extended model, we work out the exact mathematical form of this dependence, and discuss its implications for population-level modelling of language change (Figure 5).

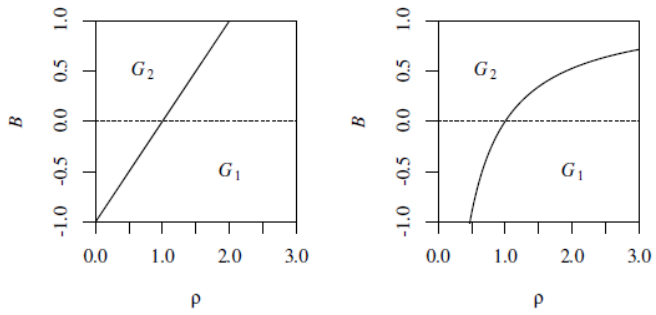


Figure 1

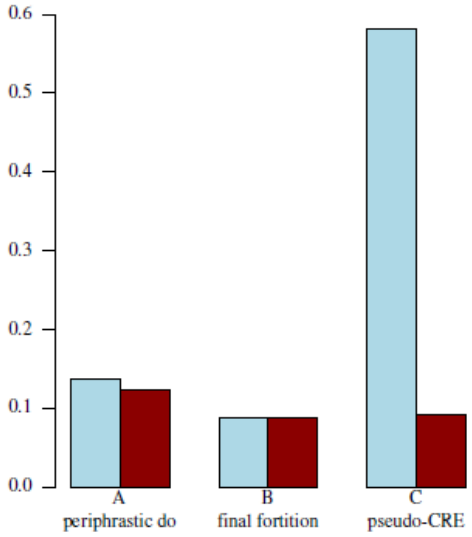


Figure 2

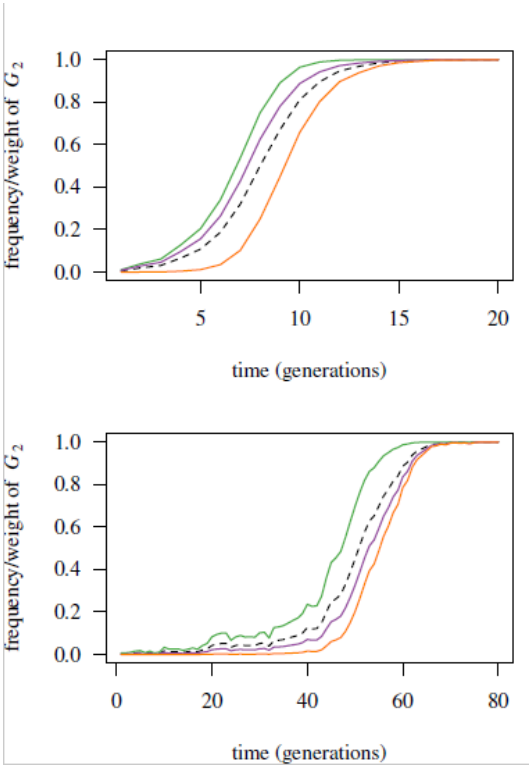


Figure 3

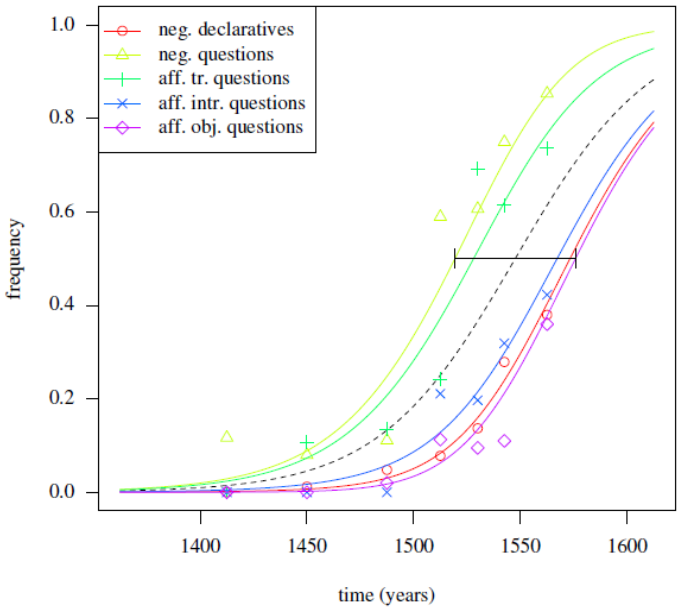


Figure 4

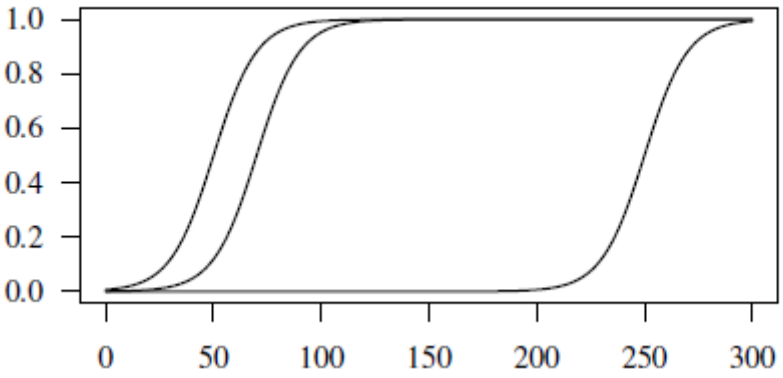


Figure 5

Dative subject construction and their syntactic variants in Eastern Slavic: Archaism or early innovation?

Leonid Kulikov (Ghent University)(kulikovli@googlemail.com)

General background and subject

This paper focuses on the historical sources and evolution of constructions with non-canonical (dative) subject in the history of Slavic languages, foremost in Eastern Slavic, paying special attention to syntactic variation attested for a particular class of constructions, such as possessive constructions and constructions with experiencer verbs. I will argue that the existence of two (or, in some languages, even more) competing variants must be due to the interaction of two opposing tendencies in the history of the Slavic (and, more generally, Eastern Indo-European) syntax.

Traditionally, constructions such as (1), attested in some Eastern Slavic languages, are regarded as a feature of the literary language resulting from early borrowing or from the influence of Southern Slavic languages (e.g. Gallis 1956):

(1) Old Russian
bě emu synŭ
was he-DAT son-NOM
'He had a son'

This opinion leads to inadequate understanding of the origins of constructions with non-canonical subject in the history of Slavic languages, and in the history of Indo-European, in general. This traditional view will be subjected to critical analysis in the present paper.

Main research questions

In this paper I will concentrate on evidence from the history of (mainly Eastern) Slavic languages that unambiguously points to the antiquity of constructions with the dative subject in this Indo-European branch, most likely inherited from the proto-language. In particular, I will try to corroborate the idea expressed in an early paper by V. Toporov (1960) (unfortunately largely disregarded in later scholarship) about the syntactic influence of Middle Iranian languages on Eastern Slavic dialects in the contact zone (see Ščecova 1996). While Old Indo-Aryan (Sanskrit) has few traces of Differential Subject Marking preserved after the split of Proto-Indo-Aryan and Proto-Iranian, the closely related Iranian languages proved to be much more conservative in this part of the syntactic system. In particular, they furnish valuable evidence for the Proto-Indo-European dative subject constructions (especially in possessive constructions). This paper will focus on possible impact of Iranian syntax that could have supported the preservation of a number of syntactic archaisms in early Eastern Slavic dialects which were lost in many other Slavic languages of both Western and Southern groups.

Methodology

The methods applied in the study include modern typological analysis as well as diachronic analysis of the evidence available from the history of language systems, foremost in the domain of grammatical relations (subject/object) and transitivity oppositions.

Integrating data from Slavic (mostly Old and early Modern Russian), Old and Middle Iranian and some other ancient Indo-European branches with evidence from non-Indo-European languages which attest similar phenomena, I will investigate the main mechanisms and diachronic scenarios documented for the history of oblique subjects in this part of the Indo-European area.

Main results

I argue that the diachronic scenarios documented for the history of non-nominative subjects depend on several parameters, such as, above all, the semantic class of verbs. In particular, both verbs of possession and experiential verbs prove to be most conservative as far as the preservation of the oblique subjects inherited from Proto-Indo-European (see especially Barðdal et al. 2012; Barðdal & Smitherman 2013) is concerned.

In particular, there are good reasons to assume the supporting influence of Iranian dative constructions that helped to preserve this and some other archaic features of the Slavic syntax arguably inherited from Proto-Indo-European.

Later on, when the Iranian influence decreases and eventually ceases, constructions with the dative subject are slowly ousted by other syntactic patterns, such as constructions with the preposition u 'at' + genitive, as in (2):

(2) Old Russian (Pov.vrem.let)

i	ne	be	u	nix	carja
and	not	was	at	they:GEN	king:GEN

'And they did not have a king'.

The syntactic variation attested for these two competing patterns may reflect two basic tendencies than can be formulated for the history of (Eastern) Slavic syntax. The preservation of the original Differential Subject Marking with the persistent syntactic variation, particularly common for certain semantic classes of verbs (such as verbs of experience, happenstance, ontological states, attitudes and some others) and the tendency to suppress the oblique (Dative) subject constructions in favor of the canonical nominative subject (especially in the Western and South-Western parts of the Slavic area). While religious texts and, to some extent, chronicles reveal the increasing amount of competing constructions (with *u* 'at' + genitive or with the verb *iměti* 'have'), most probably due to the so-called 'South Slavic influence' (see e.g. Lunt 1987), secular texts preserve more traces of the archaic patterns, going back to Proto-Indo-European oblique subject constructions.

In addition, evidence from the history of Slavic syntax furnishes valuable material for a diachronic typology of constructions with dative (and, more generally, oblique) subjects, instantiating a number of basic mechanisms responsible for the rise of Differential Subject Marking, and its subsequent decline in some branches of Indo-European. Furthermore, from the historical point of view, this pattern can be considered as an indirect trace of the archaic type of alignment (ergative/active) that can be reconstructed for early Proto-Indo-European.

References

Barðdal, Jóhanna et al. 2012. Reconstructing constructional semantics: The dative subject construction in Old Norse-Icelandic, Latin, Ancient Greek, Old Russian and Old Lithuanian. Studies in Language 36.3: 511-547.

Barðdal, Jóhanna & Thomas Smitherman. 2013. The quest for cognates: A reconstruction of oblique subject constructions in Proto-Indo-European. Language Dynamics and Change 3.1: 28-67.
Gallis, Arne. 1956. The syntax of relative clauses in Serbo-Croatian viewed on a historical basis. Oslo: Aschehoug.
Lunt, Horace G. 1987. On the relationship of Old Church Slavonic to the written language of early Rus'. Russian linguistics 11.2: 133-162.
Ščecova, Tat'jana G. 1996. Dativ v drevnerusskom pis'mennom jazyke XI—XIV vv. (na materiale datel'nogo posessivnogo, gradacionnogo i priimennogo sub''ektno-ob''ektnogo) [The dative case in Old Russian literary language of the 11-14 centuries: Possessive, gradational and adnominal subject-object constructions]. Moscow, PhD diss.
Toporov, Vladimir N. 1960. Ob odnoj irano-slavjanskoj paralleli iz oblasti sintaksisa [An Irano-Slavic syntactic parallel]. Institut slavjanovedenija, Kratkie soobščenija 28: 3-11.

Subordination in language contact situations: Complement clauses in Caucasian Urum

Johanna Lorenz (Bielefeld University)(johanna.lorenz@uni-bielefeld.de)

Background and research questions

In line with the debate on the question to what extend grammar or syntax can be borrowed, it is particularly interesting to study the languages of speech communities characterised by bi- or multilingualism including both genealogically and typologically distinct languages. The study presented in this talk will do so by focusing on complement clauses in Caucasian Urum (henceforth Urum). Urum is a variety of Anatolian Turkish spoken by ethnic Greeks in the Small Caucasus in Georgia. Almost every speaker is fluent in Russian and most are competent in Georgian as well. At present, the language is considered to be highly endangered.

Several studies report changes of subordination patterns in Turkish or Turkic languages in settings with dominant Indo-European languages (e.g. Valk and Backus 2013 on Turkish; Menz 2001, 2006 on the Turkic language Gagauz; Matras 2009 on Macedonian Turkish; Johanson 2006 on West Rumelian Turkish). They claim that a growing realisation of finite complement clauses introduced by subordinating conjunctions substitutes the primarily synthetic and non-canonical morphological marking in Turkic complement clauses. In a pilot corpus study, I identified four basic complementation patterns in Urum: complement clauses introduced with a complementizer and with a canonical subordinate verb, complement clauses without a complementizer but with a non-canonical verb, either only with an infinitive morpheme or with an additional dative morpheme, and clauses without any specific morphosyntactic subordination encoding.

The purpose of this study is twofold. Firstly, it aims to provide a description of the possibilities to express complement clauses in Urum in comparison to Turkic languages and to Russian. It will concentrate on the identification of language change phenomena. At the moment, only data from Tbilisi are available and therefore the age of the speakers is analysed as a decisive factor influencing the choice of alternative patterns. In the near future, data from speakers living in the original settlements in Georgia (K'vemo K'art'li) and from speakers who have migrated to Greece are expected to be available for further analysis.

Secondly, it attempts to identify critical factors on different linguistic levels that determine the distribution of variations and tendencies. The research issues include the question, whether complement-taking predicate classes or single verbs pose selectional constraints on complementation strategies. Many studies within the field of complementation primarily discuss functional verb classes. Further questions arise when licensing conditions on the semantic level are taken into account. In several (crosslinguistic) studies, dependencies on the veridicality status of complement clauses, the implicativity or assertiveness of the superordinate verb or the predetermination of semantic characteristics in the complement relation have predominantly been reported separately. With regard to the syntactic level, I will also examine how the different patterns in Urum can be accounted for by looking at functional categories within a minimalistic approach (e.g. CP, TP, vP and VP).

Methodology

The study is largely corpus-based. A parallel dataset was designed to allow the calculation of variations between speakers. Within a repeated observation design, the speakers were instructed to tell short stories on given themes.

This communicative setting is not as artificial as with experiments, where less variation with respect to linguistic properties can be expected, but it is also not as free as with naturalistic data containing more variation of discourse settings and topics. The analysis to date is based on five narratives of each 32 speakers. Recently, another eight narratives from 13 speakers have become available.

Complement clauses were classified as clauses with or without subordination marking. Within the first group, a distinction between clauses with a complementizer or with a special verb was drawn. Finally, special verbs were subdivided according to the presence or absence of a dative suffix.

Potential explanatory factors will be annotated and their influence will be estimated by means of Mixed effects logistic regression models. Probabilistic models will be compared as depending upon a variety of factors and it will be tested, which factors can be removed from a model without a significant loss of information.

For contexts where striking syntactic variations can be found, experiments with a 2x2 factorial design concentrating on potentially decisive factors will be carried out in February 2016.

Results

The data examined so far indicate that the language fits quite well into crosslinguistic observable tendencies with regard to the influence of different classes of complement-taking predicates on the choice of alternative patterns. The selectional constraints are likely to be more related to verb classes than to single verbs. Furthermore, the Urum speakers replace primarily left-branching and synthetic Turkic patterns in favour of right-branching and analytic constructions. Younger speakers seem to avoid the usage of additionally dative marked subordinate verbs. This is interesting because the usage of dative suffixes as a subordinating device is not attested in Russian. The analysis of potential explanatory factors will be carried out in the next weeks.

References

Johanson, Lars. 2006. Turkic language contacts in a typology of code interaction. In Turkic languages in contact, ed. Hendrik Boeschoten, and Lars Johanson, 4-26. Wiesbaden: Harrassowitz.

Matras, Yaron. 2009. Language contact. Cambridge: Cambridge University Press.

Menz, Astrid. 2001. Gagauz right-branching propositions introduced by the element ani. Turkic languages 5:234-244.

Menz, Astrid. 2006. On complex sentences in Gagauz. In Turkic languages in contact, ed. Hendrik Boeschoten, and Lars Johanson, 139-151. Wiesbaden: Harrassowitz.

Valk, Pelin Onar, and Ad Backus. 2013. Syntactic change in an immigrant language: from non-finite to finite subordinate clauses in Turkish. Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics 4:7-29.

Alternative constructions for Romance deadjectival verbs: relevant correlations between eventive and argument structure

Eugenia Mangialavori Rasia (CONICET)(eukenia@gmail.com)

Ana Laura Marusich (UNR)

Countless proposals aimed at elucidating the way in which meaning and morphosyntax of event structure are related (e.g. Jackendoff 1990, Grimshaw1990, Travis 2000, Borer 2005, Hale&Keyser 2002); the general wisdom being that the syntactic projection of arguments correlates non-trivially with event structure. In particular, the causative–inchoative alternation has been the subject of much debate and the source of theoretical innovation (cf. Levin&Rapapport 2005; Chierchia 2004, Reinhart 2002; Ramchand 2008 i.a.). It might also be a case where new theoretical insights can be combined with the use of new technologies providing them with empirical validation.

Here, by applying novel theoretical insights on core verbal meaning, we set out to account for certain variability and systematicity of argument structure realizations in relation to the alternatives open by the way a change-of-state[COS] predication can be construed in Romance.

Problem

Apart from the alternation between the well-known inchoative/unaccusative and causative/transitive COS structures, Romance renders an additional layout, featuring the Initiator/Causer as sole argument of the deadjectival verb (1). This draws an interesting contrast with languages like English in at least two respects: (i)either participant can be realized in the monoargumental version in Romance; (ii)there is no mechanism guiding the interpretation of the unique DP as Undergoer/Causee.

(1) where A=[Causer/Initiator] and B=[Causee/Undergoer]

a. El chocolate engorda a los niños — Type1:X1-V-X2
'Chocolate makes kids fat'

b. Los niños engordan — Type2:X2-V
'Kids fatten [get fat/gain weight]'

c. El chocolate engorda — Type3:X1-V
'Chocolate fattens [is fattening]'

Data & Analysis

A basic rule of event composition states that the event structure of COS verbs combines two basic components: cause and process; the former, when present, causally implicating the latter (cf. Ramchand 2008 i.a. for overview). These subevents are (allegedly) ordered in a hierarchical embedding relation —in a constructivist approach, participant relations being built up recursively from successively embedded (sub)event descriptions in first-phase/lexical-syntax.

However, a closer look on (1)-(3) suggests that the causational and the process components may be equally optional. Hence, we must define whether we are facing a complex structure with unrealized projections/arguments; or if either subevent (with the corresponding projection filled by the DP denoting the subject of the particular subevent) can occur independently, assuming that there is a general combinatorial semantics that interprets the (lexical/first-phase) syntactic structure in a regular and predictable way, and that the semantics of event structure and event participants is read directly off the structure (Ramchand 2008:42), and not directly off information encoded by lexical items.

As for the first option, absence of the Theme/Causee/Undergoer could be taken as indication of this eventive portion not being encoded in the verb, but optionally construed as one of many alternatives for deriving a property-denoting root into a verb. Otherwise, cases like ((1)c) would instantiate an argument related to a projection absent in the lexical entry (InitP/VCAUSE), and a process projection lexically specified taking no argument (thus violating the requirement that all subevental projections must have a filled specifier, Ramchand, 2003: 27) .

Composite role hypothesis discarded by impossibility of reflexive morphology (2)—and since Romance does not provide us with any overt morphological indicator of such derivational process or eventive complexity—another option would be to evaluate if relevant evidence can be gleaned from the aspectual properties of these constructions, assuming that the absence of an Undergoer/Causee shall correlate with the absence of its corresponding eventive portion (i.e., process/change). Indeed, whereas (deadjectival) COS verbs in English are prototypically argued to be [proc] verbs (Ramchand 2008:91), data indicates that Romance Causer-only alternatives do not comprise a process portion. Corpus and analytic data—comprising nearly 30 deadjectival Spanish verbs and its corresponding Italian/Portuguese/Catalan equivalents—show that Type3 is consistently stative: unlike Type1/2, it only occurs in present tense (3) and fails to accommodate duration(framing)/endpoint adverbials (4) and maximality/culminative modifiers (5). Interestingly enough, proportional modifiers (mucho/bastante) are allowed only if interpreted as a degree scope on the capacity of the causer to determine the corresponding change.

The situation holds a relevant correlation between eventive and argument structure: as soon as a causee/undergoer is added (even a generic/defective one), a process (COS) event obtains. In this sense, the additional construction contributed by Romance provides clear evidence in favor of Ramchand's (2007) intuition about InitP(V_{CAUSE}) being a state. Further tests in languages with more reliable ergativity diagnostics (e.g., auxiliary selection/ne-clitization/passivization) confirm that, unlike Type2 (consistently unaccusative), Type3 show consistent unergative patterns.

(2) Las luces (*se) enceguecen. cf. La gente (se) enceguece.
the lights SE.REFL go blind the people SE.REFL go blind

(3) El chocolate {está engordando/engordó} *(a todo el mundo/a la población).
'Chocolate {is making/made} (everybody/the population) fat'

(4) El chocolate *(te) engorda {por un tiempo/en pocos días}
*'Chocolate makes *(you) fat {for a while/in few days}'*

(5) El chocolate engorda {*totalmente/*completamente/mucho/bastante/seriamente}
'Chocolate can make somebody {totally/completely/a lot/enough/seriously} fat'

(6) a. #El ruido ensordece
'Noise causes deafness'
b. # El aire engorda
'Air causes fatness'

Results & Conclusions

Two possible derivations are distinguished: incorporation of the root into (unaccusative V^0) en route to the causative verbal head (V^0)—i.e., those DVs delivering a COS predicate with a necessary undergoer/causee—and direct conflation of the property-denoting root into the unergative phonologically null verb—i.e., those DVs that only involve a causer and are consistently stative.

The operation is relatively free, and restrictions are essentially encyclopedic (the only argument needs to feature some property allowing them to instigate the corresponding change (6)).

The alternative syntactic structures may be freely built up by conflation/incorporation instead of requiring lexical tagging—thus, lexical items may not necessarily provided of minimal syntactic information constraining insertion in syntactic structures. The proposed syntax avoids the requirement of finer specification on the roots forcing the nonprojection of a specifier. Finally, the systematicity of the phenomena under discussion consistently points towards structural composition rather than lexical encoding. In this sense, theoretical and methodological innovation can be seen to converge in the detection and analysis of underexplored patterns.

Selected references

Jackendoff, R. 1990. *Semantic structures*. Cambridge: MIT Press.

Hale, K. & Keyser, S.J. 2002. *Prolegomenon to a Theory of Argument Structure*, Cambridge: MITPress.

Ramchand, G.C. 2008. *Verb meaning and the lexicon*. Cambridge: CUP.

Plurality in Vietnamese, Chinese and Korean

Trang Phan (Ghent University)(chengnn85@gmail.com)

Eric Lander (Ghent University)

The paper presents a comparative analysis of plural markers in Vietnamese, Chinese and Korean, with the main empirical focus on the two Vietnamese plural markers *những* and *các*, which have not drawn much attention in the literature in comparison with its Chinese and Korean counterparts (*men* and *tul*, respectively). The goal of the paper is to address the three following questions, as part of a broader goal of determining what the fundamental nature of plurality in classifier languages is: (i) Are these plural markers truly optional as commonly assumed? That is to say, what is the point of adding the plural marker when a bare noun can by itself have a plural reading in classifier languages? (ii) In what way are these markers different from plural morphemes in non-classifier languages? (iii) How syntactically similar/different are these plural markers within classifier languages?

In order to answer the first question, we compare the behaviour of the nominal phrase containing the plural markers and that of bare nouns in four environments: (i) the internal argument of relational possessive constructions, (ii) predicative nominals, (iii) the generic reading, and (iv) scope with respect to negation. The sharp contrast between the two kinds of nominals exhibited in all four of these environments clearly shows that the presence of plural markers does make a significant difference to the meaning of the sentence, suggesting that they are not really optional. Furthermore, there are some contexts in which *những* and *các* are used obligatorily. Optionality might be an illusion arising from other factors, for instance, the existence of the *Num*-neutral general number category (Nomoto 2013).

With regard to the second question, assuming the Split DP hypothesis (Alexiadou et al. 2007) and the Split Plural hypothesis (Mathieu 2014), we argue that these plural markers are not only semantically but also structurally different from English-type plural morphemes. In particular, plural morphemes in English-type languages are classifying plurals which only encode the plurality of referents of nouns and occupy the head position of ClP, whereas plural morphemes in classifier languages are counting plurals which are base-generated in a higher position as the head of NumP, and also encode other semantic information in addition to plurality. Therefore, the presence of plural markers in classifier languages does not constitute a counter-example to Chierchia's (1998) nominal mapping parameter.

At the micro-parametric level, adopting the so-called Spanning approach (Ramchand 2008, Dekany 2009), we argue that Vietnamese, Chinese and Korean share the same underlying functional structure (UniqueP > SpecificityP > CountingP > ClassifyingP > N), but that they differ in the syntactic 'size' of their plural markers. Specifically, the size of the plural markers in each language is determined on the basis of three diagnostics: (i) Do the plural markers exhibit a specificity effect? (Nomoto 2013's tests) (ii) Do they show up in indefinite specific contexts? (iii) Can they co-occur with classifiers? We thus put forth an explanation for the parametric difference in terms of spanning. For Vietnamese, *những* is the spellout of the span [SpecificityP [CountingP]], whereas *các* is the spellout of [UniqueP [SpecificityP [CountingP]]]. Korean tul has a similar size as Vietnamese *những*, which only spells out [SpecificityP [CountingP]]. Chinese men is the largest, spelling out the entire span of [DefiniteP [SpecificityP [CountingP [ClassifyingP]]]].

Finally, we discuss the implications of our proposal for the structure of bare argument nominals, the presence of covert D in article-less languages and the fine-grained distinctions between two closely related concepts of specificity and definiteness (Enc 1991, Abbott 1999, 2006, Gillon 2006).

References

Abbott, B. (1999). "Support for a unique theory of definiteness." In: Matthews, T. & Strolovitch, D. (Eds.), Proceedings from SALT IX, pp. 1– 15. Ithaca, NY: Cornell University.

Abbott, B. (2006). Definite and indefinite. In: Brown, K. (ed.), The Encyclopedia of Language and Linguistics, 2nd ed., vol. 3, pp. 392-399. Oxford: Elsevier.

Alexiadou, A. Haegeman, L. & Stavrou, M. (2007). Noun Phrase in the Generative Perspective. Berlin: Mouton de Gruyter.

Chierchia, G. (1998). "Reference to kinds across languages". Natural Language Semantics 6 4:339–405.

Dékány, E. (2009). "The nanosyntax of Hungarian postpositions". In Svenonius, P. et al (Eds.) Nordlyd: Special issue on Nanosyntax 36, pp. 41-76.

Emeneau, M. (1951). Studies in Vietnamese (Annamese) Grammar. Berkeley / Los Angeles: University of California Press.

Enç, M. (1991). "The semantics of specificity". Linguistic Inquiry22(1):1–25.

Gillon, C. (2006). "The Semantics of Determiners: Domain Restriction in Skwxwú7mesh". PhD, University of British Columbia.

Mathieu, E. (2014). "Many a plural". In: Aguilar-Guevara, A. et al (Eds.), Weak Referentiality, pp. 157-181. Amsterdam: John Benjamins.

Nomoto, H. (2013). Definite-like plurals. Ms, Tokyo University of Foreign Studies.

An alternation study of Dutch psych verbs

Dirk Pijpops (University of Leuven; Research Foundation Flanders)

(dirk.pijpops@kuleuven.be)

Dirk Speelman (University of Leuven)

I. INTRODUCTION

In Dutch, a number of psych verbs exhibit an alternation between a reflexive (1) and transitive argument construction (2). The present corpus study investigates what factors drive the choice of the language user between these constructions, for the verbs ergeren ('to annoy'), interesseren ('to interest'), storen ('to disturb') and verbazen ('to amaze').

(1) Reflexive construction (experiencer-subject)
Daar erger ik me groen en geel aan. (CGN)
There annoy I myself green and yellow to
'That greatly annoys me.'

(2) Transitive construction (stimulus-subject)
Dit [...] ergerde de Romeinen mateloos. (ConDiv)
This [...] annoyed the Romans excessively
'This [...] excessively annoyed the Romans.'

II. HYPOTHESES

A. Agentivity hypothesis

The agentivity hypothesis is put forward, be it in varying forms, in quite different theoretical frameworks (a.o. Dowty 1991; Langacker 1995; Pesetsky 1995). It may be summarized as follows.

For mental states or events, it is not always clear which of the participants, i.e. the stimulus or experiencer, is more agentive. This causes variation in argument realization. The more agentive participant is assigned subject position.

This hypothesis may operate at two levels. At the type level, the agentivity hypothesis states that verbs whose lexical meaning attributes a more agentive role to the experiencer, will be more compatible with experiencer-subject constructions. The operationalization of the agentivity hypothesis at this level is taken over from Van de Velde (2004: 53–55) and embodied by the variable Verb. This operationalization leads us to consider interesseren ('to interest') to entail the most agentive experiencer, followed by either ergeren ('to annoy') or storen ('to disturb'), and finally verbazen ('to amaze'). Preference for the transitive construction is therefore expected to rise from interesseren to either ergeren or storen and finally to verbazen.

The second level is the token level. Here, the agentivity hypothesis predicts that given a particular utterance, the language user will put the currently most agentive participant in subject position. The operationalization at the token level is taken over from Levin and Grafmiller (2012) and embodied by the variable Stimulus-Animacy. It predicts that utterances with animate stimuli will prefer the transitive construction, while inanimate objects, especially abstract entities, will prefer the reflexive construction.

B. Etymology hypothesis

The etymology hypothesis is inspired on Klein and Kutscher (2005), who posit that it's not the psychological meaning of psych verbs that determines their argument construction, but rather their (ties with a former) physical meaning. Etymological inquiry led us to suspect that storen most strongly favors the transitive construction, followed by either ergeren ('to annoy') or verbazen ('to amaze'), and finally interesseren ('to interest').

C. Topicality hypothesis

The topicality hypothesis is operationalized through the variables Stimulus- and Experiencer-Topicality. These variables present a scale ranging from the first and second persons, to the third person pronouns, the definite nouns and the indefinite nouns. It is expected that preference for object position rises as we go to the end of this scale.

III. RESULTS

All instances of the four verbs were extracted from the Corpus of Spoken Dutch (CGN, Oostdijk et al. 2002) and the ConDiv corpus (Grondelaers et al. 2000). These instances were manually checked, and a number of them had to be excluded. The resulting dataset contained 1810 occurrences, which were tagged for the hypothesis-driven variables presented above, and number of nuisance variables. Next, a logistic regression model was composed using a stepwise variable selection procedure. The hypothesis-driven variables turned out to be the most important predictors in the model. Their effect plots can be found in Figure 1.

The variable Verb does not confirm the agentivity hypothesis at the type-level, nor the etymology hypothesis. Conversely, the variable Stimulus-Animacy does more or less confirm the animacy hypothesis at the token level. The topicality hypothesis is confirmed by Stimulus-Topicality, but Experiencer-Topicality behaves exactly opposite to what was predicted. However, we will show that in retrospect, such behavior might not be as aberrant as it appears on first sight.

IV. CONCLUSIONS

To end with, we shortly summarize the relevance of this study for theories of argument realization. First, the study has shown that inter- and intralingual generalizations such as the agentivity and topicality hypothesis definitely seem possible (cf. Levin and Rappaport Hovav 2005). Second, our failure to confirm the type level agentivity hypothesis means that caution may be in order when applying the agentivity hypothesis too rigidly at the type level. Finally, the confirmation of the token level agentivity hypothesis seems to indicate that argument constructions do seem to add meaning to utterances, separately from the meaning of the verb (Goldberg 1995; Colleman and De Clerck 2009).

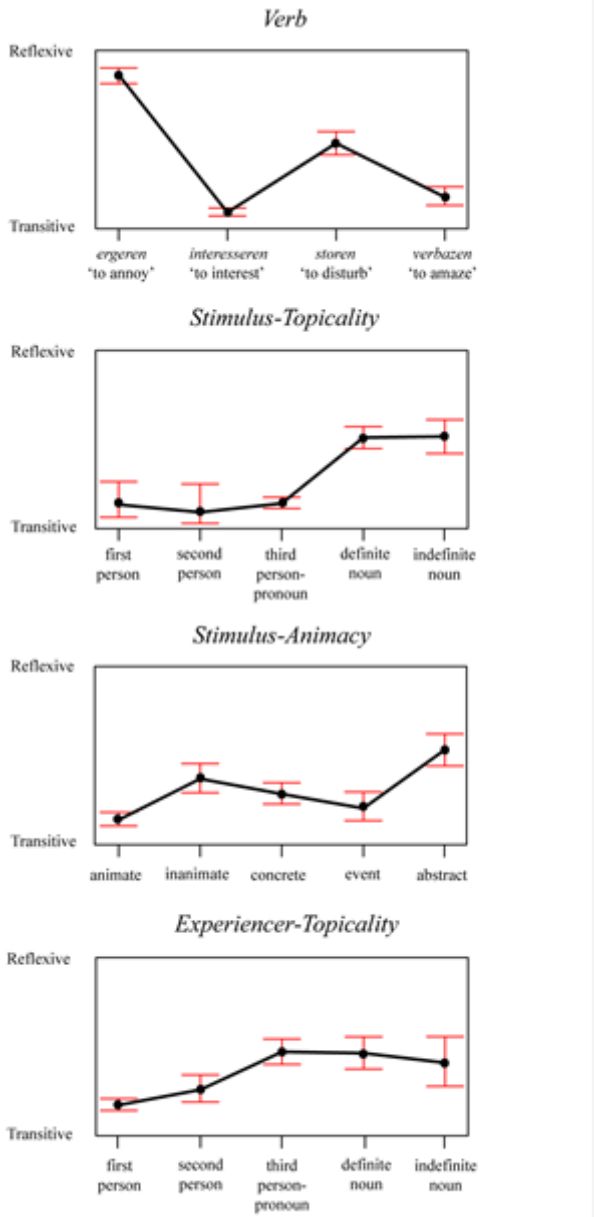


Fig. 1. Effect plots of the hypothesis-driven variables

References

Colleman, Timothy and Bernard De Clerck. 2009. "Caused motion"? The semantics of the English to-dative and the Dutch aan-dative. Cognitive Linguistics 20(1). 5–42.

Dowty, David. 1991. Thematic proto-roles and argument selection. Language 67(3). 547–619.

Goldberg, Adele Eva. 1995. Constructions: a construction grammar approach to argument structure. Chicago: University of Chicago press.

Grondelaers, Stefan, Katrien Deygers, Hilde Van Aken, Vicky Van den Heede and Dirk Speelman. 2000. Het CONDIV-corpus geschreven Nederlands. Nederlandse Taalkunde 5(4). 356–363.

Klein, Katarina and Silvia Kutscher. 2005. Lexical Economy and Case Selection of Psych-Verbs in German. Working papers of the SFB 282, Theory of the Lexicon [Arbeitspapiere des SFB 282 Theories des Lexicons] Nr. 122.

Langacker, Ronald Wayne. 1995. Foundations of cognitive grammar. Stanford: Stanford University Press.
Levin, Beth and Jason Grafmiller. 2012. Do you always fear what frightens you? In Tracy Holloway King & Valeria de Paiva (eds.), From Quirky Case to Representing Space: Papers in Honor of Annie Zaenen, 21–32. Stanford: CSLI Publications.
Levin, Beth and Malka Rappaport Hovav. 2005. Argument realization. Cambridge: Cambridge university press.
Oostdijk, Nelleke, Wim Goedertier, Frank Van Eynde, Louis Boves, Jean-Pierre Martens, Michael Moortgat and Harald Baayen. 2002. Experiences from the Spoken Dutch corpus project.
Pesetsky, David. 1995. Zero Syntax. Cambridge: MIT Press.
Van de Velde, Freek. 2004. De Middelnederlandse onpersoonlijke constructie en haar grammaticale concurrenten: semantische motivering van de argumentstructuur. Nederlandse Taalkunde 9(1). 48–76.

The alternation strength of causative verbs: a quantitative and qualitative analysis of the interaction between the verb, theme and construction

Laurence Romain (Université Lille)(laurence.romain1@gmail.com)

This paper presents a corpus-based study of the causative alternation in English, where a transitive (causative) construction (e.g., John opened the door, henceforth Cx2) alternates with an intransitive (non-causative) construction (e.g., The door opened; Cx1). This alternation covers a wide variety of verbs (cf. Levin 1993 for a good overview). A first step in our analysis of this alternation has been to measure the distribution of verbs in each construction, using a collostructional analysis. This already shows considerable differences between verbs, some of which are significantly attracted to one construction or the other; for example 'thicken' is significantly attracted to Cx1 while 'open' is significantly attracted to Cx2. This distributional variation can be, at least partially, explained by the meaning of each verb that may fit better with one construction rather than the other. Lemmens (2006) argues that high frequency of a theme may also influence the frequency of the verb in one construction. A theme with a high frequency would therefore increase the number of occurrences in one or the other construction and an analysis focusing solely on such frequency would be less representative of the actual distribution of the verb. Lemmens (ibid.) suggests, following Stefanowitsch & Gries's collexeme analysis & distinctive collostructional analysis (2003, 2004, 2005), that measuring the degree of Theme overlap may yield more precise results, give a better idea of the distribution of verbs between the two constructions, and thus more information on the construction itself, since, as Perek points out "the strongest collexemes of a construction, as the most semantically compatible lexemes are a potential source of information about the meaning of the construction" (2014: 66).

The second step of our analysis has thus been to calculate the degree of Theme overlap, which is likely to give a more representative idea not only of the meaning of each construction but also of the alternation strength of a verb, that is, its degree of alternation. Lemmens' case studies (ibid.) focus on individual themes and find a very low alternation index; however, he suggests a semantic grouping of themes. Taking up on his suggestion, we have set up such a grouping following Perek's (2014) method of using WordNet as a basis for semantic grouping so as to avoid the issue of arbitrariness or subjectivity of groupings. A WordNet search for each verb under study is complemented by a calculation of the semantic proximity of the different Themes found in collocation with each verb. Such a proximity is measured in terms of shared collocates via a semantic vector space model (Turney & Pantel 2010). While a query on WordNet provides only one synonym for 'thicken': 'inspissate', which is extremely rare, a look at the shared collocates of the different Themes associated with 'thicken' provides a reliable basis for objective semantic groupings.

As we will show in our discussion, such grouping of Themes into semantic categories (in relation to the verb(s) they appear with) indeed yields a more accurate view of a verb's alternation strength and by extension, a better insight into the meaning of each verb and the construction it occurs in.

Priming and Frequency in Language Change: the Spanish Past Subjunctive

Malte Rosemeyer (Albert-Ludwigs-Universität Freiburg)

(malte.rosemeyer@rocketmail.com)

Scott Schwenter (Ohio State University)

Usage-based approaches to language change demonstrate the relevance of frequency for language change in that the repeated use of a linguistic element leads to the conservation of that element (Bybee 2006). Recently, it has also been suggested that priming, i.e. short-term activation triggered by the previous mention of an element, might have an influence on language change, particularly grammaticalization (Eckardt 2008). We demonstrate that like frequency, priming can have a conserving effect in language change, and that there is a predictable relationship between priming and frequency in processes of language change.

To substantiate this claim, we employ conditional inference trees and random forests, available in R (cf. Breiman et al. 1984, Tagliamonte and Baayen 2012), to analyze the alternation between the Spanish past subjunctive forms ending in *-ra* and *-se* (as in *comiera* and *comiese* 'had eaten') in a corpus of nearly 4000 past subjunctive forms taken from the Corpus del Español (Davies 2002). The older form *-se* has been gradually replaced by the *-ra* form since the 13th century, leading to a low relative frequency of *-se* in contemporary Spanish. This makes the alternation a perfect example for the analysis of the influence of priming on language change.

Previous studies of this variation (Asratián 2007; Day 2011) claim that the two forms are governed by semantic differences such as (non-)realization, emphasis or (im)probability, but our analysis suggests that these factors play at best a marginal role. Rather, due to the ongoing replacement of *-se* by *-ra*, priming and frequency effects are the best predictors of the alternation. First, the probability for a prior *-se* to lead to the choice of *-se* over *-ra* in the following context is significantly greater than the probability of a prior *-ra* to lead to *-ra* over *-se*. Second, although *-se* is basically restricted to 3rd person singular morphology in contexts without priming, thus demonstrating paradigmatic atrophy (Leech et al. 2009: 80), when primed by *-se* this restriction is drastically reduced. Third, *-se* is significantly more likely to occur with high-frequency verbs (e.g. *ser* 'to be') than low-frequency verbs (e.g. *entrar* 'to enter').

These results suggest that priming has a strong conserving effect in language change, reflecting previous results from studies such as Rosemeyer (2014) or Tamminga and Ecay (2014). Following Jäger and Snider (2013), we assume that the strength of priming effects is dependent on surprisal: the less expected a linguistic element, the stronger its priming effect. Given that, in language change, decreasing frequency will necessarily cause the obsolescing element to be less expected, obsolescing elements will nearly always have a particularly strong priming effect.

Some researchers, such as Tamminga and Ecay (2014), have interpreted this effect in terms of a mere corollary of language change. In other words, one could assume that priming does not actually conserve the disappearing variant, but rather has a temporal effect that vanishes after the priming effect has expired. Our results however suggest that this is not the case and that priming does indeed conserve obsolescing variants. In particular, the fact that priming to some degree counteracts the paradigmatic atrophy experienced by obsolescing forms suggests that the priming effect does not only activate a certain instantiation of the construction (such as comiese), but rather activates the entire representation of the construction (in this case, the V + *-se* construction). Since priming thus leads to an increase in abstract constructional knowledge, we can maintain the stronger hypothesis that priming has a conserving effect in language change.

Our results also shed light on the relationship between frequency and priming as competing factors in language change. Although both can result in conservation, the conserving effect of frequency will always cause irregularity (such as the paradigmatic atrophy of Spanish *-se* forms). Our study demonstrates that priming can serve to temporarily reestablish regularity with the result that, in priming contexts, *-se* forms no longer display paradigmatic atrophy.

References

Asratián, Arucia (2007): "Variación *-ra/-se* en el español hablado en Caracas", in: Boletín de lingüística 19 (5), 5-41.

Breiman, Leo et al. (1984): Classification and Regression Trees. New York: Chapman and Hall.

Bybee, Joan L. (2006): "From usage to grammar: the mind's response to repetition", in: Language 82 (4), 711-733.

Davies, Mark (2002): Corpus del español (100 million words, 1200s-1900s). Available online at http://www.corpusdelespanol.org. Last access: 23 March 2015.

Day, Meagan (2011): "Variation in the use of the –ra and –se forms of the imperfect subjunctive in Modern Spoken Peninsular Spanish". Talk given at NWAV 40, Georgetown University.

Eckardt, Regine (2008): "Concept priming in language change", in: Theoretical Linguistics 34 (2), 123-133.

Jäger, Florian and Neal E. Snider (2013): "Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience", in: Cognition 127, 57-83.

Leech, Geoffrey et al. (2009): Change in Contemporary English. A Grammatical Study. Cambridge: Cambridge University Press.

Rosemeyer, Malte (2014): Auxiliary Selection in Spanish. Gradience, Gradualness, and Conservation. Amsterdam, Philadelphia: Benjamins.

Tagliamonte, Sali and Harald Baayen (2012): "Models, forests and trees of York English: Was/were variation as a case study for statistical practice", in: Language Variation and Change 24 (2), 135-178.

Tamminga, Meredith and Aaron Ecay (2014): "Priming as a diagnostic of grammatical status: two case studies". Talk given at Penn Linguistics Conference 38.

A verb-centered approach to the ACC/DAT alternation of German two-way prepositions: Integrating qualitative and quantitative methods

Jonah Rys (Ghent University)(jonah.rys@ugent.be)

Background

A little-known alternation in German morphosyntax is the 'intransparent' accusative-dative alternation with two-way prepositions, i.e. those occasions where both cases can be used in reference to the same event (e.g. Das Flugzeug prallte auf dieACC/derDAT Wasseroberfläche auf 'The plane crashed down on the water surface'). The phenomenon has already been tackled from different perspectives (e.g., Paul 1920, Leys 1989, Smith 1995, Willems 2011), and as is common in alternation research, all of these approaches have been concerned with defining a general abstract function for each construction. Despite theoretical disagreements, there is consensus that the alternation serves as a subjective perspectivation device, allowing the speaker to present a dynamic event as 'completed' (DAT) or 'not completed' (ACC) (Leys 1989), 'established' (DAT) or 'being established' (ACC) (Paul 1920, Willems 2011), or with focus on the path (ACC) or endpoint (DAT) of the movement (Smith 1995, cf. also Duden 2006). These proposals, however, rely mainly on introspective reasoning and selective, small samples of corpus data, which leaves the question to what extent they can stand the test of a detailed, large-scale study of the alternation in actual language use.

Research questions and hypotheses

That corpus-based research can offer a more detailed insight in the nature of the alternation, has already been shown in recent studies (Rys et al. 2014, Willems et al. to appear). In particular, the functionally motivated use of the alternation can only partially account for the observed variation and it proves difficult to delimit a general, overarching function that explains all of the variation. We propose that the alternation is better described on a verb-by-verb basis, taking into account that i) the alternation may serve many different 'local' functions, i.e. relevant for only one or a small set of related verbs, that cannot be generalized to a single, overarching function and ii) case marking is not necessarily functional, but may be guided by conventional usage norms. The presence of both non-generalizable functions and non-functional conventionalized case marking tendencies call for a differentiated empirical approach that enables one to go beyond defining the alternation in general semantic terms.

Methodology

This paper reports the findings of a case study of the ACC/DAT alternation with 19 German contact verbs (including aufprallen, landen, einschlagen, anstoßen etc.). A three-step procedure is proposed that takes individual verbs as the focus of the analysis: First, for each verb, a sample of 300 sentences (extracted from the Deutsches Referenzkorpus Mannheim) is analyzed qualitatively to determine potential morphosyntactic, semantic and lexical factors that exhibit a preference for either construction.

Second, the effect of these factors is tested quantitatively using bivariate (chi square, fisher's exact) and multivariate (classification tree analysis) analyses. This leads to a detailed profile of the constructional variation for each verb, revealing both overarching tendencies and between-verb contrasts.

Finally, the corpus data are compared with acceptability judgments from native speakers. This step is often absent in alternation studies but indispensable to uncover contrasting individual usage norms that are masked by generalized corpus data. In this study, case marking in 23 corpus sentences were judged by native speakers using a 5 point Likert scale.

Results and conclusions

i) Case marking is motivated by a wide variety of semantic and lexical factors, making case marking variation with certain intransparent verbs much more predictable than has hitherto been acknowledged. ii) Despite some overarching tendencies (e.g., the DAT preference for the 'X lands on Y' sense of verbs such as *aufsetzen, niedergehen, landen*), case marking is often governed by verb-specific, non-generalizable factors (e.g., the quasi-obligatory use of ACC for the 'X crashes down on Y' sense of *niedergehen*, which is not found for semantically related verbs). iii) For several verbs (e.g., *landen*, *niederkommen, auffahren*), no case variation occurs, although such variation is to be expected based on observed variation with near-synonymous verbs (e.g. *aufsetzen, niedergehen, auflaufen*). iv) Although in general, judgment rates correlate positively with corpus tendencies, the acceptability test bears witness to contrasting intuitions among individual speakers, with speakers regularly dismissing highly frequent occurrences as incorrect. In summary, the study confirms that syntactic alternation research benefits from a lager focus on verb-specific particularities (alongside generalized regularities) and individual (alongside collective) preference norms.

References

Duden. 2006. Die Grammatik (7th ed.). Mannheim: Dudenverlag.

Leys, O. 1989. „Aspekt und Rektion räumlicher Präpositionen". Deutsche Sprache 17: 97–113.

Paul, Hermann. 1920. Deutsche Grammatik, Vol. IV, Teil IV: Syntax (Zweite Hälfte). Halle: Max Niemeyer.

Rys, J., Willems, K., De Cuypere, L.. "Akkusativ und Dativ nach Wechselpräpositionen im Deutschen. Eine Korpusanalyse von versinken, versenken, einsinken und einsenken in". In I. Doval and B. Lübke (eds.) Raumlinguistik und Sprachkontrast. Neue Beiträge zu spatialen Relationen im Deutschen, Englischen und Spanischen. München: Iudicium Verlag.

Smith, M. B. 1995. "Semantic motivation vs. arbitrariness in grammar: toward a more general account of the dative/accusative contrast with German two-way prepositions." In: I. Rauch and G. Carr (eds.) Insights in Germanic linguistics I: Methodology in transition. Berlin/New York: Mouton de Gruyter.

Willems, K. 2011. "The semantics of variable case marking (Accusative/Dative) after two-way prepositions in German locative constructions. Towards a constructionist approach". Indogermanische Forschungen 116: 324–366.

Willems, K., Rys, J., De Cuypere, L. "Case alternation in argument structure constructions with prepositional verbs. A case study in corpus-based constructional analysis". To appear in: H. C. Boas and A. Ziem (eds.): Constructional approaches to argument structure in German. Berlin/New York: Mouton de Gruyter.

Quantificational dimension of Taraldsen's Generalisation: The loss of pro-drop and rich verbal inflection in French

Alexandra Simonenko (Labex EFL)(alexandra.simonenko@mail.mcgill.ca)

Benoit Crabbé (Alpage)

Sophie Prévost (LaTTiCe)

1. This paper examines the relation between the availability of null referential subjects and the richness of verbal inflection, known as Taraldsen's Generalisation ([12], [4], [1]), from the point of view of historical variation. The original generalisation is stated as a categorical implication, namely, that rich verbal agreement implies the possibility of non-expression of subjects in a given language, where richness is often taken to equal a morphologically distinct ending for each person ([11]). On the diachronic dimension, [11] has to reject a direct connection between the loss of richness and the loss of pro-drop in French because of an apparent temporal lag between the two (XI-XII century vs. XVI century estimates respectively). In this paper we show that there was no temporal lag, rather, there exists an inverse dependency between the likelihood of pronominal subject omission and the richness of verbal inflection. We first establish the temporal profile of the pro-drop variation, then propose an operationalization of the notion of inflectional richness as a measure of ambiguity for verbal endings and finally relate the two phenomena using the same measures.

2. French has been widely considered to be a pro-drop language in the early and high Medieval period and to have lost that property by the end XVII c. ([3], [13], [11]). However, it has also been noted that even in the earliest attested documents French appears to differ from the "classic" version of a pro-drop grammar exemplified, for instance, by Italian, in allowing for overt impersonal pronouns and demonstrating a high rate of overt subjects in subordinate clauses ([14], [10]).

We calculated the rate of pro-drop (as a proportion of clauses with overt subject pronouns among clauses with either pronominal or null subjects, excluding cases of coordination ellipsis) in a sample of finite clauses which excludes imperatives and wh-clauses targeting subjects in MCVF and Penn Supplement (a treebank of tagged, parsed and functionally annotated French texts from X to XVIII cc. with Penn treebank style annotation scheme of approx. 1 mln words). As Fig. 1 shows, the limited pro-drop character of Old French is especially striking in prose. Compare >50% of overt subjects in the XII c. prose with ≈30% of overt subjects in the speech of adult Italian and Spanish speakers ([2] and [8] respectively). We conclude that even in the earliest documented periods French was close to losing pro-drop.

3. In this paper we present the first quantitative evaluation of the diachronic syncretisation of (a part of) French verbal paradigm which can help adjudicate the ongoing debate about a potential link between the loss of rich verbal inflection and that of pro-drop (e.g. [6], [13], [11]).

Part of the impoverishment of verbal inflection was the spread of the ending *-e* from 3rd to 1st person singular in verbs of the traditional 1st group (*-er* infinitives) (e.g. *aim* '(I) love' becomes *aime*, as in *il aime* 'he loves') and the spread of the ending *-s* from the 2nd to the 1st person singular in verbs of the traditional 2nd group (other types of infinitives): e.g. *di* '(I) say' becomes *dis*, as in *tu dis* 'you say') ([5]).

Most likely syncretisation extended beyond *-e* and *-s* in oral language affecting all final stops and fricatives and making all endings phonologically indistinguishable except for 1st and 2nd person plural. However, due to the unavailability of oral data, we have to approximate this process by focusing on the fate of *-e* and *-s*, which can be quantified. We also had to exclude the X-XI period because the data are too sparse to estimate the relevant measures.

In order to estimate the goodness of an ending for identification of subject's person we estimate conditional entropy of subject person given a certain ending. To that effect, we defined a binary variable *Person* with sample space {1st, 3rd} or {1st, 2rd} and estimated its conditional entropy given the ending *-e* or *-s*, respectively. We coded all utterances with verbs of the 1st and 2nd groups with *-e*/*-s* endings and an overt subject for the person of the subject. We thus propose conditional entropy as a measure of the capacity of verbal inflection to identify the person of the subject, thus operationalizing the notion of "richness". On this view a paradigm can be more or less rich, depending on how unambiguous endings are.

4. Finally, we compare the loss of the richness to that of pro-drop. On Fig. 3 we added entropy of the variable *Subject* with sample space {overt, null}. We see that as the entropy of *Person* given ending *-e* goes up, that is, this ending becomes a weaker signal of person, the entropy of *Subject* goes down, that is, the likelihood of having a pronominal subject expressed becomes progressively greater than not having it expressed (there were not enough relevant data for X-XI cc.). On Fig. 4 we plotted the entropy of *Person* for *-e* and *-s* combined, the data on *-s* alone being too scarce for it to be considered in isolation. These results strongly speak against a temporal lag between the two changes.

Figure 2: Pronominal subject expression in verse

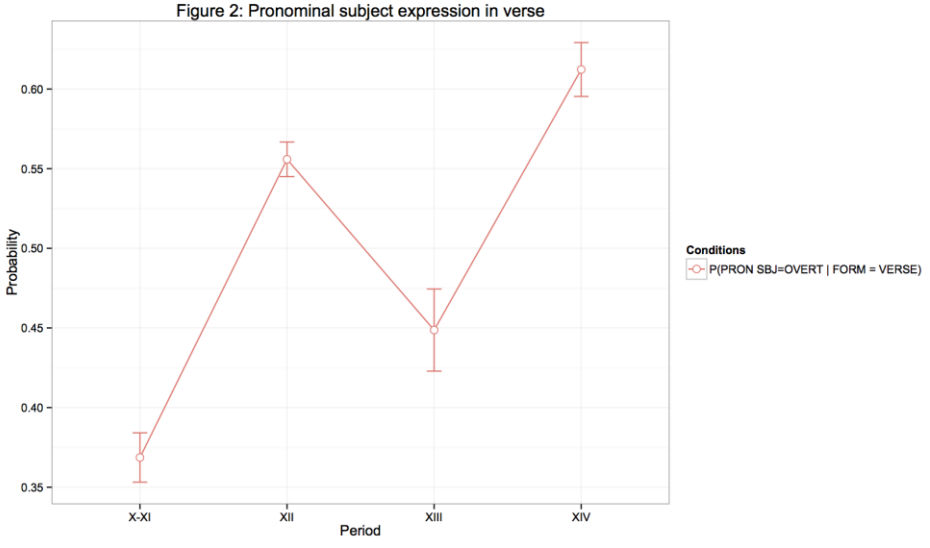


Figure 1: Pronominal subject expression in prose

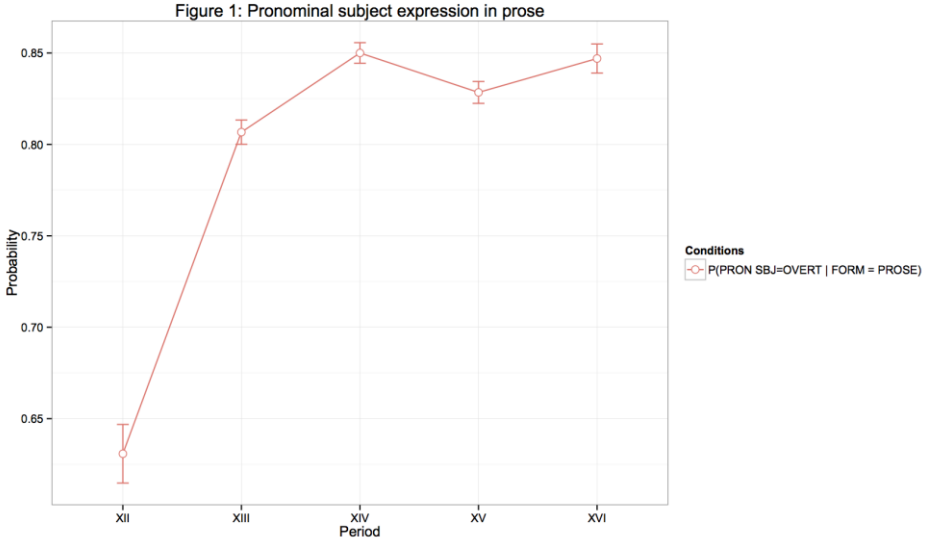


Figure 3: Entropy of person given ending e & entropy of pronominal subject

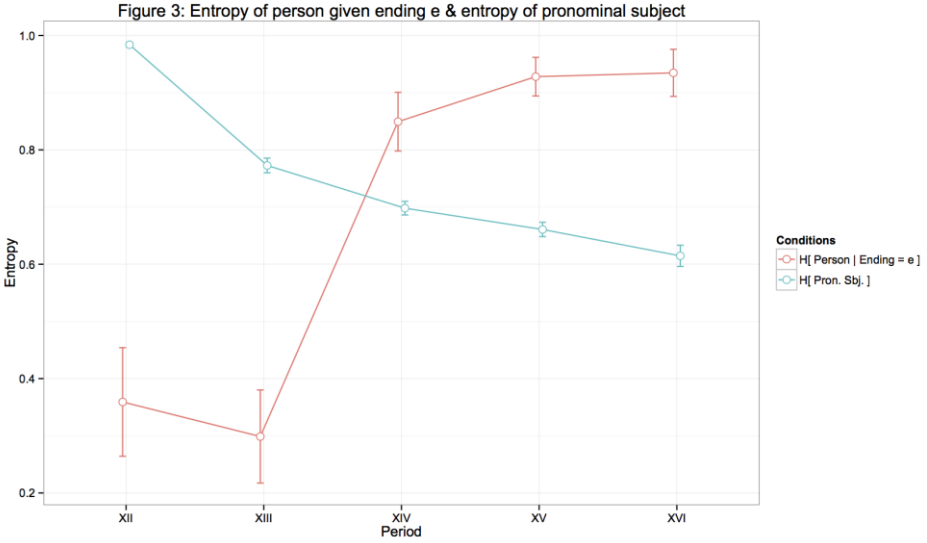
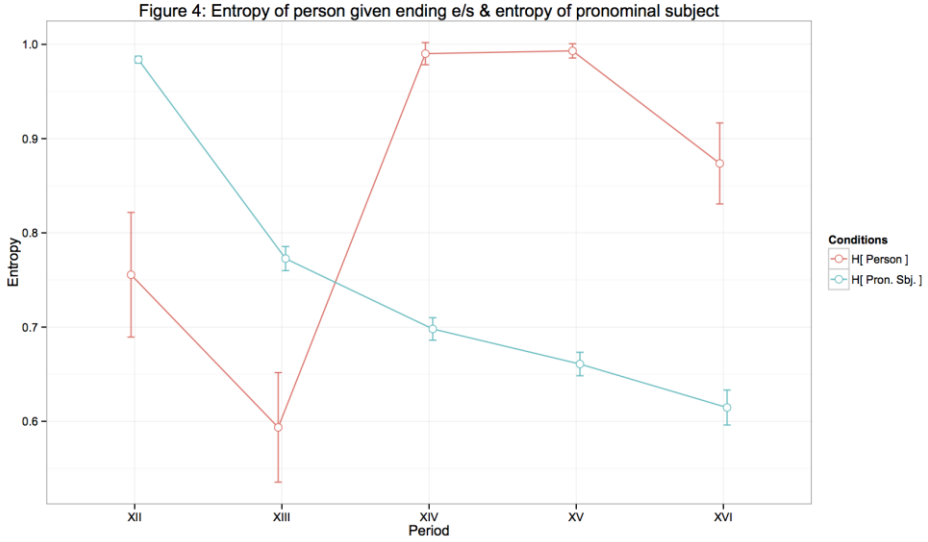


Figure 4: Entropy of person given ending e/s & entropy of pronominal subject



References

[1] Adams, Marianne. 1987. From Old French to the theory of pro-drop. *Natural Language & Linguistic Theory* 5:1–32.

[2] Bates, Elizabeth. 1976. Language and context: The acquisition of pragmatics, volume 13. New York: Academic Press.

[3] Foulet, Lucien. 1928. Petite syntaxe de l'ancien fran¸cais. Paris: Champion, troisième édition revue. Rééedition 1982.

[4] Luigi. 1986. Null objects in italian and the theory of pro. *Linguistic Inquiry* 17:501–557.

[5] Marchello-Nizia, Christiane. 1992. Histoire de la langue française aux XIVe et XVe siècles. Paris: Dunod.

[6] Mathieu, Eric. 2006. Stylistic fronting in old french. *Probus* 18:219–266.

[7] MCVF. 2010. Corpus MCVF annoté syntaxiquement, sous la direction de France Martineau, avec Paul Hirschbühler, Anthony Kroch et Yves Charles Morin. http://www.voies.uottawa.ca/corpus pg en.html.

[8] Otheguy, Ricardo, Ana Celia Zentella, and David Livert. 2007. Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language* 770–802.

[9] Penn Supplement to MCVF. 2010. Penn Supplement to the MCVF Corpus by Anthony Kroch and Beatrice Santorini.

[10] Prévost, Sophie. To appear. Emergence and development of personal pronoun subjects: study of a bilingual Latin-old French corpus.

[11] Roberts, Ian. 2014. Taraldsen's Generalization and Language Change: Two Ways to Lose Null Subjects. In Functional Structure from Top to Toe: The Cartography of Syntactic Structures, ed. Peter Svenonius, volume 9. Oxford University Press.

[12] Taraldsen, Tarald. 1980. On the NIC, vacuous application and the that-trace filter. Indiana University Linguistics Club.

[13] Vance, Barbara. 1997. Syntactic change in medieval French. Studies in Natural Language and Linguistic Theory. Dordrecht/Boston/London: Kluwer.

[14] Zimmermann, Michael. 2009. On the evolution of expletive subject pronouns in Old French. In Proceedings of the workshop “null-subjects, expletives, and locatives in romance”. arbeitspapier 123, ed. E.-M. Remberger G.A. Kaiser, 63–92. Fachbereich Sprachwissenschaft, Universität Konstanz.

The noemen/heten alteration: on the rapid emergence of a new variant

Dirk Speelman (KULeuven, QLVL)(dirk.speelman@arts.kuleuven.be)

Stefan Grondelaers (Radboud University Nijmegen)

1. The linguistic variable

In this paper, we discuss the alternation exemplified in (1) and (2):

(1) Hij 'heet' Tom. (He 'is called' Tom.)
(2) Hij 'noemt' Tom. (He 'is called' Tom.)

In Colloquial Belgian Dutch (CBD), the verb 'noemen', which in standard Dutch means TO CALL, can also have the meaning TO BE CALLED (2), so that it becomes a competitor for the (standard Dutch) verb 'heten' (1).

Frequency counts reported in an early study on this use of 'noemen', De Schutter (1973), suggest that in the early seventies of the previous century the 'noemen' variant was not (yet) all that frequent, and that moreover it was mostly restricted to the provinces of Oost-Vlaanderen (OV, 14.6% 'noemen' versus 85.4% 'heten') and West-Vlaanderen (WV, 12.8% 'noemen' versus 87.2% 'heten') [See map in FIG. 1]. Meanwhile, however, the 'noemen' variant has gained ground rapidly, especially among younger speakers. This change, which is a recent, rapid and dramatic evolution, offers an exceptional opportunity to monitor 'language change in action'.

2. Research questions

In the present study, our general research question can be formulated as:

- RQ1 Which internal and external predictors (see next section) motivate the synchronic and diachronic status of the the 'noemen'/'heten' alternation?

More specific research questions are:

- RQ2 How has the geographic distribution of the variant 'noemen' evolved and which social groups lead the change?
- RQ3 Do we find internal predictors that indicate that 'noemen' emerged from a lexically and functionally specific expression/pattern?

The second research question, RQ2, which zooms in on external predictors, aims to position the emergence of the 'noemen' variant in the broader picture of CBD. More specifically, the question is to which extent 'noemen', starting out as a local phenomenon, has spread across the whole of Flanders, thus contributing to the homogeneity of CBD.

The third research question, RQ3, which zooms in on internal predictors, addresses the question whether the 'noemen' variant evolves from more lexically specific high frequency usages (especially 'hoe noemt dat') to less lexically fixed usages (cf. Bybee, 2006) and from specific functional contexts (e.g. questions, or still more specifically, expressions of WHAT'S IT CALLED AGAIN?, when trying to remember a word) to other usages. This hypothesis is based on anecdotal evidence that seems to suggests that 'noemen' is typically used in questions, and moreover in lexically specific questions (e.g. 'hoe noemt dat?'). Somewhat more technically, RQ3 can be phrased as "is restriction to lexically specific context increasing or decreasing from one generation to the next"?

3. Materials

Data come from the face-to-face and telephone conversations in the Spoken Dutch Corpus (n=394, with 184 cases of 'heten' and 210 cases of 'noemen'), which were recorded between 2000 and 2004.

4. Variables in model

The variables for the regression model are:

- variant: levels 'heten', 'noemen' [response variable]
- age: speaker age, in years [apparent-time construct]
- sex: speaker sex, 'F' and 'M'
- region: speaker birth region, levels 'WV', 'OV', 'AB', 'LI' (see FIG. 1)
- occup.type: which education level does the speaker's occupation require?
- (levels 'high', 'neutr', 'no.high')
- s.type: sentence type, levels 'D', 'Q' (resp. declarative and question)
- collo': collocate present in observation? level 'H', '-','N' (resp. presence in observation of collocate that is generally attracted to 'heten', neither, or 'noemen').
- speaker: speaker id, with 188 levels [random factor in mixed model]
- conver.type: conversation type, with 2 levels [(limited) random factor in mixed model]

5. Mixed-effects regression analysis

The glmer regression model that was selected, was variant ~ age*region + occup.type + age*collo + (1|speaker) + (1|conver.type). Its effects are shown in FIG 2, FIG 3 and FIG 4. Disregarding random effects, the model has C=0.83. Although, admittedly, with 13 fixed effect regressors this is a rather complex model, relative to the size of the dataset, and therefore there could be some risk of overfitting the data, the effects and the relative importance of the predictors are largely confirmed by a conditional inference tree analysis (FIG 5; C=0.81) and a random forest analysis (FIG 6; C=91).

6. Conclusions

Regarding RQ1, the most important conclusion of the study is that external predictors ('age', 'region', and 'occup.type') have a much stronger effect than internal predictors, but that there is at least one internal predictor at work ('collo'), which moreover interacts with the external predictors (especially 'age').

Regarding RQ2, it can be said that the distribution of 'noemen' now covers the whole of Flanders and that the linguistic change is mostly led by the social group without higher education.

Regarding RQ3, the conclusion is that there are (weak) lexically/functionally specific usages of 'noemen', but that they are restricted to the younger users. More specifically, there is weak evidence that collocates of 'noemen' to some extent 'trigger' the use of 'noemen' in the case of younger speakers, but have no such effect in the case of older speakers. Conversely, there are (stronger) indications of lexically specific contexts where older speakers avoid 'noemen'. More specifically, collocates of 'heten' tend to 'trigger' the avoidance of 'noemen' in the case of older speakers, but have no such effect in the case of younger speakers. This seems to be the effect of prescriptive language purification the older, but not the younger generations, were subjected to in Flanders.

A methodological conclusion is that sociosyntactic analysis should go beyond the traditional researcher-defined higher-order predictors and include lexical effects – for instance, in the forms of collocations.

7. Future steps

Other types of data will be needed to make a sharper distinction between apparent-time effects and age-grading. Also larger datasets will be needed to further study the weaker effects of the internal predictors (and their interactions with the external predictors). Therefore, our plan is to replicate the analysis using (more recent) data from a large Twitter corpus.

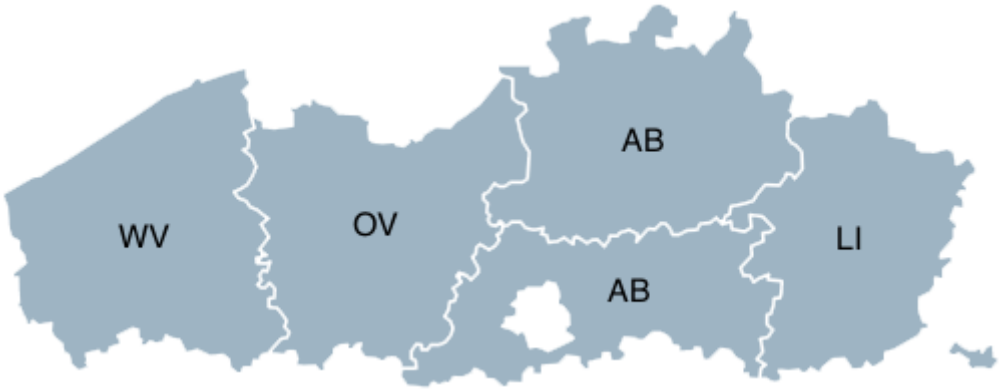


Figure 1: Regions distinguished in the variable REGION

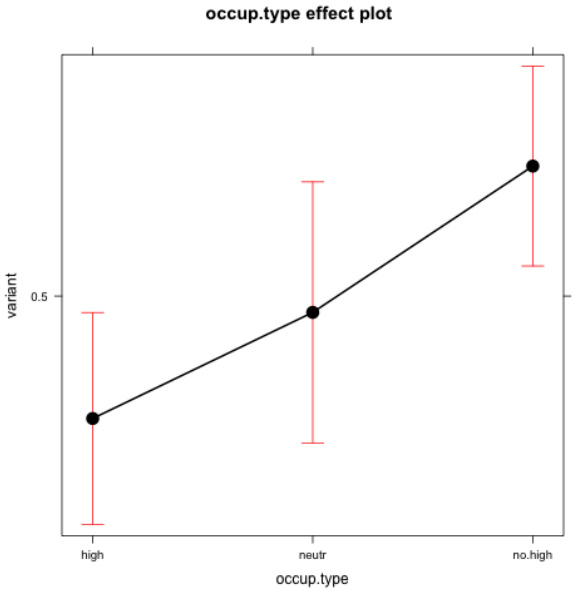


Figure 2: Effect of OCCUP.TYPE in mixed model

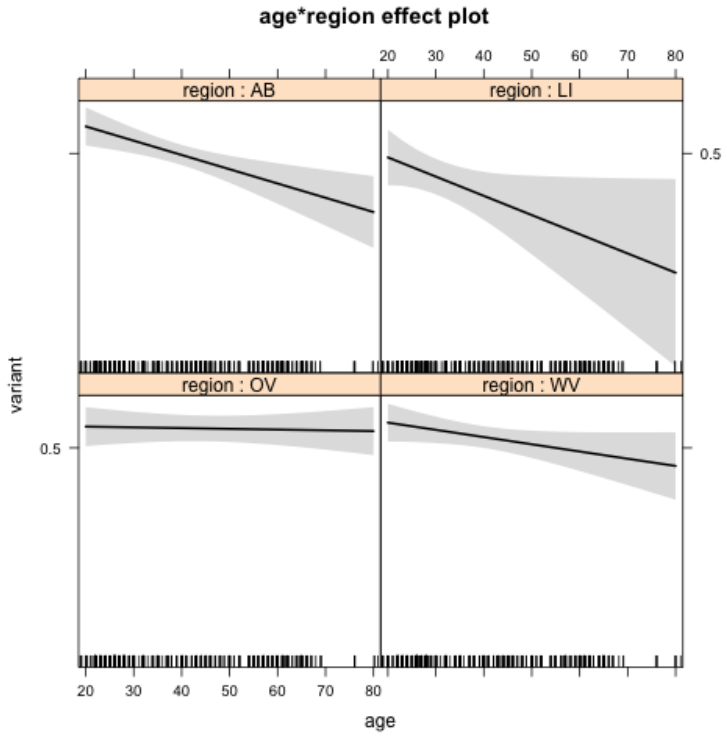


Figure 3: Interaction AGE*REGION in mixed model

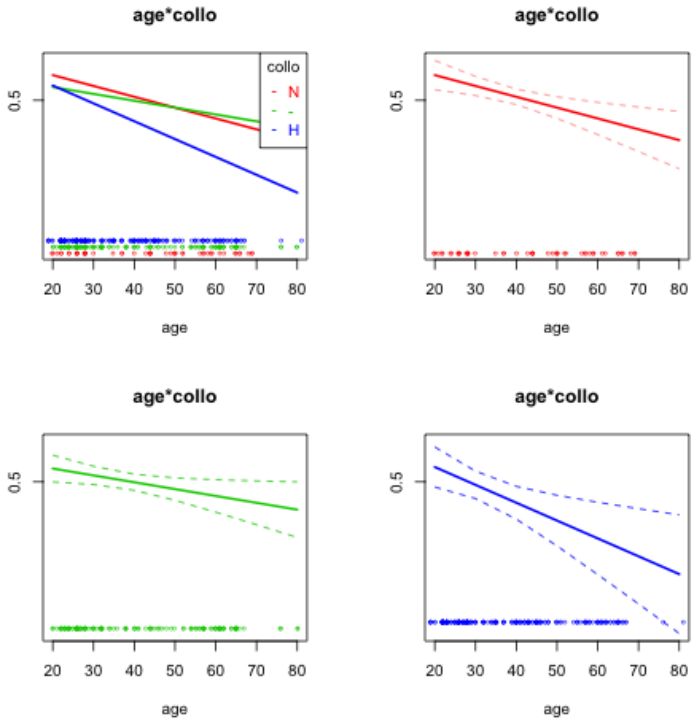


Figure 4: Interaction AGE*COLLO in mixed model

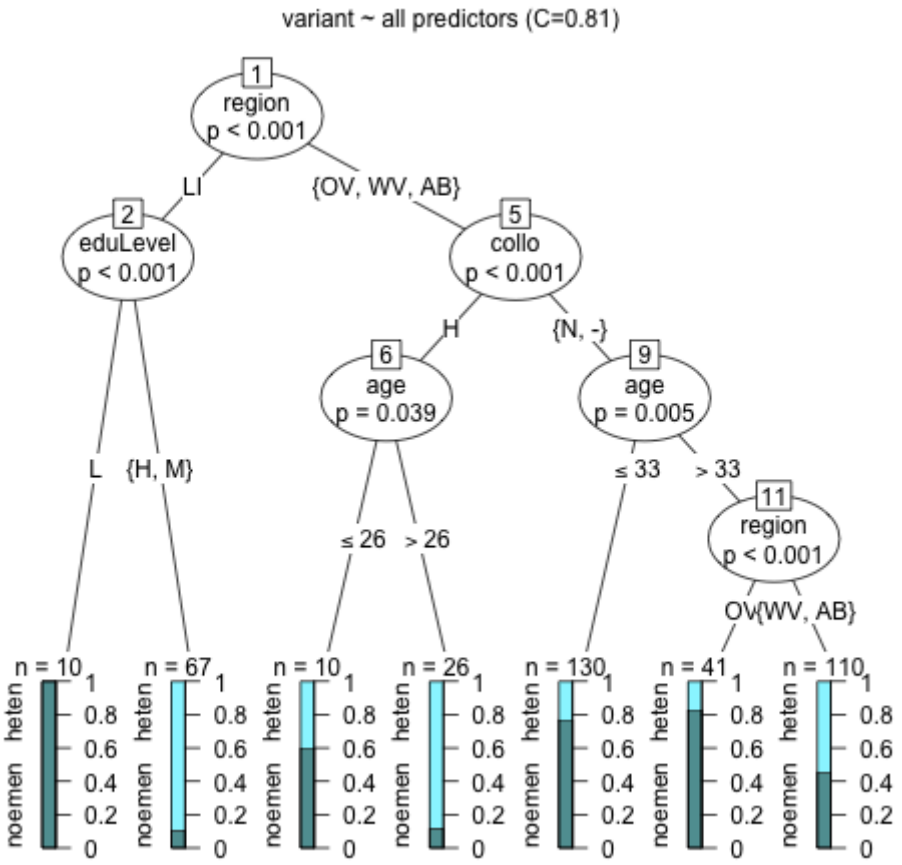


Figure 5: Conditional inference tree

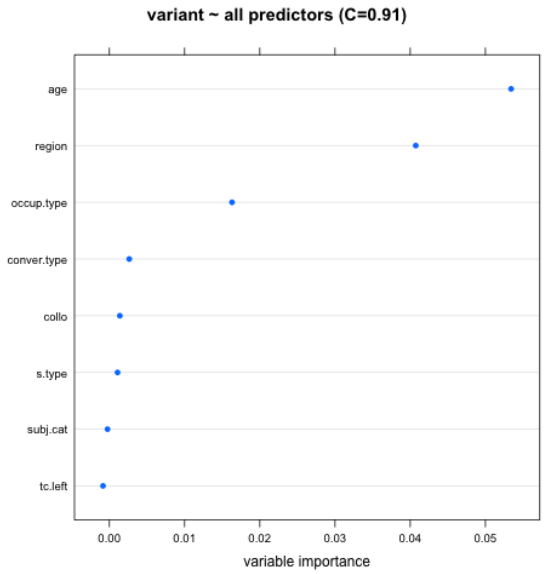


Figure 6: Random forest

References

Bybee, J. (2001). 'From usage to grammar: the mind's response to repetition'. Language 82, 711-733.

De Schutter, G. (1973). ‘Eksogeen taalgebruik in Zuid-Nederland’. In: W. Pée. Album Willem Pée. De jubilaris aangeboden bij zijn zeventigste verjaardag. p. 117-122. Tongeren: George Michiels.

Case variation in German PPs: Influence of the regional distribution

Simone Ueberwasser (University of Zurich)(simone.ueberwasser@ds.uzh.ch)

In German, many prepositions demand the argument of the PP to be in a specific case (eg. mit einem_dat Akzent_dat ('with an accent')). Some prepositions, however, show relatively free variation in the use of the case. The prepositions discussed in this presentation, trotz, wegen and während ('in spite of', 'because of', 'during'), can take a genitive or a dative. While the preference for one or the other case is most often said to depend on style or register (with the genitive being the prestigious variant), some studies have shown the regional origin of the investigated texts to be another influencing factor (e.g. Petig, 1997).

In my presentation, I will look at ~200,000 PPs from a corpus that contains data from Austria, Germany and Switzerland. For the investigation, only PPs that can be distinctly assigned to the genitive or to the dative based on the morphology of the argument's NP are taken into account. The research confirms that the dative is more frequently used in the south of the area in question in this corpus as is the case in other corpora. In a second step, I will look at other factors that promote the use of the dative in all regions in the corpus and show that these features are more frequent in the south. The increased use of the dative in the south is thus not always an independent factor but often the result of specific constructions that are more often used in the south and thus promote the use of the dative.

The corpus used contains data taken from online newspapers and is stratified by countries. Additionally, each country and its data are subdivided into linguistically relevant subregions. Using Austria and Switzerland as independent units and comparing them against the six regions defined for Germany allows to show a continuously increasing use of the dative towards the south.

On a sociolinguistic level, the influencing factors that promote the dative in the south include audience design as introduced by Bell (1984), which can be shown by taking the data from individual newspapers and comparing them against other data from the same regions or countries. This factor is in itself influenced by semantic factors, because NPs which can be marked as colloquial, dialectal or pejorative show a preference for the dative and are more frequent in the newspaper taken as an example for audience design.

Intralinguistic factors that influence the use of the dative are constructions that can be identified by applying methods such as multiple correspondence analyses. As a general summary, a NP of the form Determiner-Adjective-Noun can be seen as an unmarked form, the further the NP deviates from this form, the more likely it is to take a dative.

References

Bell, Allan (June 1984). Language style as audience design. Language in Society 13: 02, 145–204.

Petig, William E. (1997). Genitive Prepositions uses with the Dative in Spoken German. Die Unterrichtspraxis - Teaching German 30: 1, 36–39.

Syntactic variation and syntactic uniformity across languages: A crosslinguistic corpus study on linearization devices

Elisabeth Verhoeven (Humboldt University Berlin)(verhoeve@cms.hu-berlin.de)

Summary

An empirical challenge for the understanding of syntactic variation is to distinguish between the syntactic properties that vary and those that remain constant across languages. This talk presents an interesting case at issue in a large-scale corpus study in four typologically different languages (Chinese, German, Greek, Turkish). The study examined the factors determining the choice of linearization of arguments of transitive roots (actors and undergoers). Two strategies were examined within this functional space: (a) the choice of word order; (b) the choice of subject (actor/undergoer) and the concomitant choice of voice that is used as a linearization device (among else) (Branigan et al. 2008; Lamers & de Hoop forthc.). In a nutshell, the results show that the choice of a marked word order strongly depends on the syntax of the language at issue. That is, orders in which the object precedes the subject do not appear under identical conditions in the four examined languages. In contrast to word order, the choice of voice does not substantially vary across languages, i.e., if a voice alternation is available and can be used without substantial influence on the propositional content, its occurrence in discourse is not constrained by language-specific rules.

Method

We selected 20 transitive verbs per language (10 causative verbs; 10 experiencer-object verbs) and extracted a corpus of 250 sentences per verb (total = 5000 sentences per language) from written corpora (Chinese: *CCL* Corpus, Beijing University; German: DeReKo, IDS-Mannheim; Greek: *HNC* from ILSP, Athens; Turkish: *TS corpus*, Mersin University). After restricting the sample to declarative main clauses, the data was annotated for four annotation categories. The categories to serve as dependent variables are: (a) WORD ORDER (SOV|SVO|OSV|OVS|VSO|VOS), and (b) VOICE (active|non-active). The categories to serve as fixed factors are prominence scales that are known to affect the choice of linearization in discourse (Aissen 1999, Bresnan et al. 2001): (a) ANIMACY (animate|inanimate), (b) DP-type of arguments (zero|pronoun|definite|indefinite).

Results

The findings in the choice of subject (actor/undergoer) reveal similar patterns in the four languages: the proportions of undergoer-subjects (i.e., the choice of non-active voice) increases with experiencer-object verbs and is sensitive to animacy (more frequent if the undergoer outranks the actor in the animacy hierarchy). The corresponding findings in the choice of word order show that the conditions determining the choice of word order differ across languages. While an asymmetry in animacy (such that actor $<_{\text{animacy}}$ undergoer) explains a part of the occurrence of OS orders in German the corresponding linearization in Chinese does not occur under these conditions. Both languages have syntactic constructions in which objects precede subjects in the linearization. Crucially, these constructions are different. OS orders (e.g., OVS in main clauses) are an instance of scrambling in German, while OSV orders are an instance of left-dislocation in Chinese. I.e., the object constituent is outside the core clause in the latter construction, a pattern that is highly marked and contextually restricted in comparison to German scrambling. (Turkish and Greek results are similar to German).

Conclusions

The relevance of the summarized results for linguistic theory is that a part of the observed variation across languages can be explained if we take into account the syntactic properties of the object languages. We assume that the role of prominence scales such as animacy or referential asymmetries holds true at a level that is not language-specific (namely, at the discourse level). Their manifestation in particular languages interacts with grammatical systems and depends on the properties of language-specific constructions. The comparison between subject choice and order choice is an interesting minimal pair for understanding the locus of variation: syntactic constructions that are dedicated to particular discourse functions (e.g., left-dislocation in Chinese) are contextually restricted, while other linearization options may be chosen in a wide array of contexts – among else under the influence of the prominence scales examined in this study.

References

Aissen, J. 1999. Markedness and subject choice in optimality theory. *NNLT*, 17:673-711.

Branigan, H.P., Pickering, M.J., Tanaka, M. 2008. Contributions of animacy to grammatical function assignment and word order during production. *Lingua* 118, 72-189.

Bresnan, J. , Dingare, S., Manning, C.D. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG01 Conference*, M. Butt & T. H. King (eds.). Stanford: CSLI Publications.

Lamers M., de Hoop, H., forthc. Object fronting in Dutch. In: Moravcsik, E., Malchukov, A. (Eds.), Competing Motivations.

Understanding PP placement in written Dutch. A corpus-based multifactorial investigation of the principal syntactic, semantic and discursive determinants

Annelore Willems (Ghent University)(annelore.willems@ugent.be)

Gert De Sutter (Ghent University)

The present paper describes the driving forces behind the constituent ordering of non-predicate prepositional phrases (PPs). In Dutch subordinate clauses, language users fundamentally have the choice to put PPs either before the final verb cluster (the so-called *middle field position*; example 1) or after the final verb cluster (the *postfield position*; example 2)([dpc-cam-002333-nl]):

(1) ... [dat]$_{\text{1st pole}}$ de trainer door een laptop naast het veld $_{\text{middle field}}$ [vervangen wordt]$_{\text{final verb cluster}}$
... [that]$_{\text{1st pole}}$ the coach by a laptop next to the field $_{\text{middle field}}$ [replaced will be]$_{\text{final verb cluster}}$

(2) ... [dat]$_{\text{1st pole}}$ de trainer $_{\text{middle field}}$ [vervangen wordt]$_{\text{final verb cluster}}$ door een laptop naast het veld $_{\text{postfield}}$
... [that]$_{\text{1st pole}}$ the coach $_{\text{middle field}}$ [replaced will be]$_{\text{final verb cluster}}$ by a laptop next to the field $_{\text{postfield}}$
'that the coach will be replaced by a laptop next to the field.'

Although this variable syntactic structure is well known in Dutch syntactic literature, remarkably little is known about what governs PP placement. On the basis of our corpus-based investigation, we are able to confirm the multifactorial nature of PP placement in Dutch, as is already often shown for other types of syntactic variation (e.g., Grondelaers 2000, De Sutter 2005), as well as propose a more fine-grained version of the theoretical framework in which PP placement in Dutch is traditionally described and understood.

Building on journalistic data in the Dutch Parallel Corpus (Macken et al. 2011), which yielded 1,718 relevant observations, it is first shown that non-predicate PPs are significantly more often placed in postfield position than in middle field position. The theoretical implication of this result is that the structuralist idea of the middle field as the only standard slot (especially in the ANS, the standard grammar of Dutch, but also in generative-linguistic literature (e.g., Koster 1974, 1978, Hoekstra 1984 and Zwart 1990) and in a few corpus studies (e.g., Jansen 1978, 1979, 1990 and Braecke 1990) has to be abandoned in favor of a theoretical model that considers middle field and postfield as equivalent positions, at least for written communication.

Second, a binary logistic regression model was fitted with PP placement as response variable and 12 manually annotated predictor variables (partially verified via inter annotator agreement).

	Odds ratio	p-value
Length of the midfield	1.077	0.012 *
Length of the postfield	0.406	2.27e-07 ***
Length of the PP	1.361	< 2e-16 ***

Discourse accessibility	1.174	5.44e-08 ***
Particle verbs	1.627	0.004 **
Verbs in idioms	2.142	0.000 ***
Copulative verbs	2.435	4.57e-05 ***
Semantic category 1: e.g. location,...	0.330	1.22e-08 ***
Semantic category 2: time	0.102	5.91e-16 ***

TABLE 1

It should be first noted that the overall quality of the model is very good, with a c-index of 0.82, and that the model does not suffer from multicollinearity (only the significant predictors are presented in table 1). This model furthermore confirms the multifactorial nature of PP placement in Dutch, which adds further proof to the existence of a so-called probabilistic grammar. Word order is consequently not determined in a clear-cut all-or-nothing manner, but on the basis of fine-grained linguistic and contextual constraints, which language users seem to internalize through exposure and use (cf. Bybee & Beckner 2011). Third, the model shows that both syntactic complexity and discourse-related aspects play a unique, decisive role in PP placement, something which was suggested in previous literature (e.g., ANS), but never empirically tested. More particularly, the multivariate analysis points out that the length of the PP and the length of the middle field are positively correlated with postfield position, which is in line with what was previously found (Jansen 1978, 1979). More remarkably, the length of the postfield, which is a new predictor, also influences the word order. In fact, it works as an inhibitory variable: the heavier the postfield, the fewer PPs in postfield position. As for the discourse-related variables, the results show that less accessible PPs are placed significantly more often in the postfield, which is in line with the general literature on the Given-before-New principle (see among others the ANS, Gundel et al. 1993): the given (and thus accessible) information is placed before the new (less accessible) information. As for the semantic class of the PP, our results show that locative and temporal PPs prefer middle field position. Finally, verb type affects PP placement significantly in that postfield position is preferred in sentences with a semantically 'empty' verb, such as copulative verbs and verbs in idioms. Although it is not completely clear how to interpret this result, one could imagine that these verbs can be processed more quickly (cf. Segalowitz & Lane 2000, Tremblay et al. 2011), as a result of which they are preferably placed before a less accessible clausal element (the PP).

References

ANS – Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and van den Toorn, M.C. *Algemene Nederlandse Spraakkunst*. Groningen/Deurne: Martinus Nijhoff – Wolters Plantyn, 1997.

Braecke, C. ''Uit de tang' of [+ or -] prominent?". *Taal en Tongval*, 3, 125-134, 1990.

Bybee, J. and Beckner, C. "Usage-based Theory". In: B. Heine and H. Narrog (eds.), The Oxford Handbook of Linguistic Analysis. Oxford: Oxford University Press, 826-855, 2011.

De Sutter, G. *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen*. Unpublished PhD University of Leuven, 2005.

Grondelaers, S. *De distributie van niet-anaforisch er buiten de eerste zinsplaats. Sociolexicologische, functionele en psycholinguïstische aspecten van er's status als presentatief signaal*. Unpublished PhD University of Leuven, 2000.

Gundel, J.K., Hedberg, N. and Zacharski, R. "Cognitive status and the form of referring expressions". Language, 69, 274-307, 1993.

Hoekstra, T. *Transitivity: grammatical relations in Government-Binding theory*. Dordrecht: Foris, 1984.

Jansen, F. "Hoe krijgt een spreker zijn woorden op een rijtje? Taalgebruiksaspekten van de 'PP over V' konstruktie". In: J. G. Kooij (Ed.), *Aspekten van de woordvolgorde in het Nederlands*. Leiden: Vakgroep Nederlandse Taal- and Letterkunde, 70-104, 1978.

Jansen, F. "On tracing conditioning factors of movement rules: extraposition of PP in spoken Dutch". In: M. Van de Velde and W. Vandeweghe (Ed.), *Sprachstruktur, Individuum und Gesellschaft*. Tübingen: Niemeyer, 83-93, 1979.

Jansen, F. "Tangconstructies en geen reacties daarop". *Taal en Tongval*, 3, 135-151, 1990.

Koster, J. "Het werkwoord als spiegelcentrum". *Spektator*, 3, 601-618, 1974.

Koster, J. *Locality Principles in Syntax*. Dordrecht: Foris, 1978.

Macken, L., Declercq, O. and Paulussen, H. "Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus". META, 56 (2), 374-390, 2011.

Segalowitz, S.J. and Lane, K.C. ‚Lexical access of function versus content words'. Brain and Language, 75(3): 376-389, 2000.

Tremblay, A., Derwing, B., Libben, G. and Westbury, C. „Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks". *Language Learning*, 61(2): 569-613, 2011.

Zwart, J.W. *PP Extraposition from NP in Dutch*. Ms. U. of Groningen, 1990.