Hanne Martine Eckhoff,
Silvia Luraghi and
Marco Passarotti (eds.)

Diachronic Treebanks for Historical Linguistics

BENJAMINS CURRENT TOPICS

113

Diachronic Treebanks for Historical Linguistics

Benjamins Current Topics

ISSN 1874-0081

Special issues of established journals tend to circulate within the orbit of the subscribers of those journals. For the Benjamins Current Topics series a number of special issues of various journals have been selected containing salient topics of research with the aim of finding new audiences for topically interesting material, bringing such material to a wider readership in book format.

For an overview of all books published in this series, please see
benjamins.com/catalog/bct

Volume 113

Diachronic Treebanks for Historical Linguistics
Edited by Hanne Martine Eckhoff, Silvia Luraghi and Marco Passarotti

These materials were previously published in *Diachronica* 35:3 (2018)

Diachronic Treebanks for Historical Linguistics

Edited by

Hanne Martine Eckhoff
University of Oxford

Silvia Luraghi
University of Pavia

Marco Passarotti
Catholic University of the Sacred Heart, Milan

John Benjamins Publishing Company
Amsterdam / Philadelphia

DOI 10.1075/bct.113

Cataloging-in-Publication Data available from Library of Congress.

ISBN 978 90 272 0798 2 (HB)
ISBN 978 90 272 6045 1 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

Table of contents

Introduction

The added value of diachronic treebanks for historical linguistics

Hanne Martine Eckhoff, Silvia Luraghi and Marco Passarotti
University of Oxford / University of Pavia /
Catholic University of the Sacred Heart, Milan

1. Ancient languages and digital sources

Over the last few decades, the widespread diffusion of digital technology and the growing ease of transferring information via the Internet have made an enormous amount of textual data available to scholars. The vastly increased availability of primary sources has radically changed the everyday life of scholars in the humanities, who are now able to access, query and process a wealth of empirical evidence in ways not possible before.

This development also encompasses ancient languages. The first aim in the eighties and the nineties was to digitize textual data and make them available on CD-ROM and online. Later, the need for linguistic annotation gave rise to projects aimed at building corpora enhanced with increasingly complex layers of metalinguistic information, such as part-of-speech (PoS) tagging and syntactic annotation, opening the field to precise queries for particular linguistic phenomena. We are now at a stage where several of these syntactically annotated corpora, or treebanks, have reached a mature state, providing representative selections of texts for several diachronic stages of a given language. These new resources allow for a new approach to diachronic studies of syntactic phenomena where scholars previously had to content themselves with empirical work on a much smaller scale.

This volume brings together a set of papers that report research on various diachronic matters supported by evidence from diachronic treebanks for different languages, i.e., treebanks that provide data for a language across several historical stages. We show that diachronic treebanks can provide considerable methodological advances in terms of greater transparency and better ways of exploiting frequently problematic source material, thus allowing us to shed new light on vexed topics.

https://doi.org/10.1075/bct.113.int

2. What is a treebank?

In linguistics and philology, the term 'corpus' has traditionally been used simply to denote a set of texts used to explore some linguistic phenomena. Many types of digital text resources are now also referred to as 'corpora'. McEnery et al. (2006) simply define a corpus as "a body of naturally occurring language", while Sinclair (2005) gives a much stricter definition: "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research". However, not even the strictest definitions have linguistic annotation of any kind among the criteria. Thus there is a great deal of variation in the amount of work that has gone into building and processing corpora and in the usefulness of the resource for linguists researching particular phenomena in a given corpus. A corpus may be anything from a digitized, machine-readable text collection that only allows queries for text strings, to a sophisticated, multi-layered text resource with several types of linguistic markup, queriable by a dedicated query engine. In this volume, we concern ourselves with one of the most labor-intensive corpus types of all: the treebank.

A treebank is a text corpus with exhaustive syntactic annotation, typically applied on top of lemmatization, PoS tagging and morphological annotation. Each of these annotation layers adds to the precision of queries. Lemmatization allows for queries for all word forms subsumed under a single lemma, eliminating the need to use regular expressions. Part-of-speech and morphological tags allow for queries for specific combinations of linguistic features at the word level, without having to refer to the word form. Syntactic tagging makes it possible to search for groups of words that are syntactically related, regardless of whether they are adjacent to each other or not. Since syntactic queries are mostly multi-word queries, and are typically combined with features from other layers, they can quickly become quite complex and require either a good query engine or that users master a query language. However, given such facilities, a treebank allows queries of great precision: if the annotation is good enough, it is possible to make queries almost entirely free of noise in terms of false positives and false negatives. For example, in a given language one may find all infinitives with preverbal pronominal direct objects in a single query.

Although some treebanks are annotated in accordance with the formalism of a particular syntactic framework, most strive to be relatively theory-neutral. There are two major groups of annotation schemes: phrase-structure-based schemes and dependency-based schemes. The first major treebank to be released, the Penn Treebank (1989–1996; Taylor et al. 2003a), used a (simplified) phrase-structure scheme, which is continued in the many of the Penn daughter treebanks. Many of the numerous dependency treebanks are inspired by the groundbreaking Prague

Dependency Treebank (Bejček et al. 2013). Existing dependency treebanks employ a number of different annotation schemes. As a response to this, the Universal Dependencies[1] initiative has developed a universal consensus-based scheme and works to convert as many treebanks as possible into that scheme.

The two main treebank styles are based on two different syntactic notions, both of which clearly have some psychological reality. Phrase-structure treebanks are based on the idea that words are organized into groups (constituents) with certain properties; for example, that an entire constituent can be substituted by a pro-word and will normally move together. Dependency treebanks, on the other hand, are based on the idea that every word in a sentence has one and only one syntactic head. As a brief illustration of the differences between these two main treebank styles, consider the two syntactic trees below. The tree in Figure 1 is the original Penn-style analysis of the opening of John 11.47 (American Standard Version; see Taylor & Pintzuk this volume). The tree in Figure 2 is the same passage analyzed

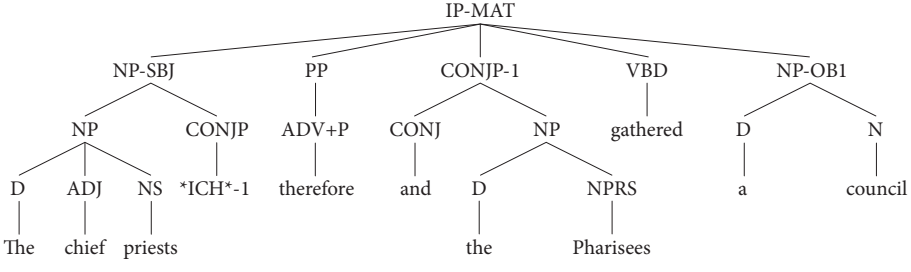


Figure 1. A Penn-style phrase structure tree

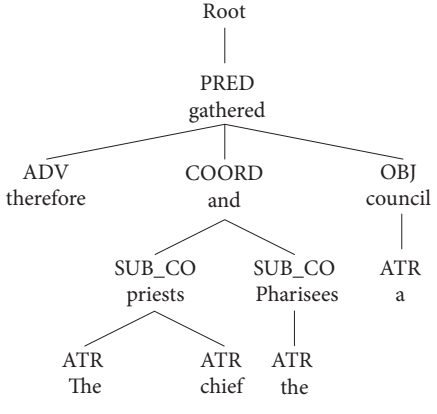


Figure 2. A Prague dependency treebank tree

1. http://universaldependencies.org/

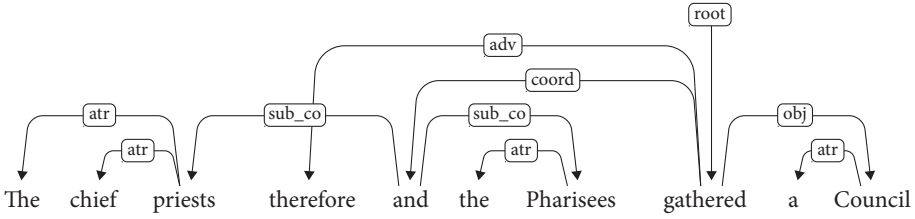


Figure 3. A Prague dependency treebank tree in linear order

in the Prague Dependency Treebank format. Both analyses are given in tree format for ease of comparison. Figure 3 presents the dependency analysis in linear order.

Here we see that the phrase-structure analysis in the Penn-style tree is fairly flat, which brings the two analyses closer than they might have been if the Penn scheme had been binary-branching. The most striking difference in these examples is that the Penn analysis cannot have crossing branches, and therefore it deals with split coordination (the topic of Taylor and Pintzuk's paper) with a trace (the *ICH*-1). The index of the trace is then picked up again in the CONJP-1, the second part of the coordination which is represented in its linear place in the sentence. In the dependency analysis, the fact that the coordination is split is not represented at all and can only be retrieved by combining the dependency analysis with word order information stored in a different layer (visualized in Figure 3). However, this analysis is computationally simpler, since every node in the tree corresponds to a lexical item.

3. Historical corpora and treebanks

Historical linguistics necessarily relies on corpora. This observation is captured by the German term *Korpussprachen* to refer to historical languages. Indeed, with most historical languages, all we have is a more or less extended corpus of written texts. This constitutes a limitation (one cannot check if something that is not in the corpus is not there because it is ungrammatical or through mere accident), but it also enables linguists working on these languages to base their assumptions on all attested forms. Extended corpora, even if finite, often exceed the linguist's ability to check all occurrences: for this reason, the introduction of digitized corpora has been a welcome addition to historical linguistics, as indeed in research on spoken language. Parsed corpora have the further advantage of adding information at various levels of linguistic analysis through the addition of metadata. Among these, treebanks have become an increasingly useful resource for the data-driven study of linguistic structures at various levels (Freddi & Luraghi 2013).

Diachronic treebanks are closely related to synchronic treebanks, and thus generally use the same annotation schemes or schemes inspired by these. For example,

there are several diachronic treebanks using the Penn phrase-structure format, such as the York-Toronto-Helsinki Corpus,[2] and the Penn Corpora of Historical English[3] for English (Taylor et al. 2003a and b; Kroch & Taylor 2000; Kroch et al. 2004); and the French diachronic treebank "Modéliser le changement: les voies du français" (MCVF; Martineau 2008),[4] all represented in this volume. There are also several diachronic dependency treebanks. Some of them directly use the Prague Dependency Treebank format, for example the Perseus Latin and Ancient Greek Dependency Treebanks[5] (Bamman & Crane 2011) and the *Index Thomisticus* Treebank[6] (Passarotti 2011). Other diachronic dependency treebanks, such as the PROIEL family of treebanks (Haug & Jøhndal 2008; Eckhoff & Berdičevskis 2015; Eckhoff et al. 2018) and the Syntactic Reference Corpus of Medieval French (SRCMF; Stein & Prévost 2013), have developed annotation schemes of their own, inspired by more classic schemes, but with modifications intended to increase expressivity.

Diachronic treebanks also take part in standardization initiatives such as the Universal Dependencies initiative. For instance, both PROIEL and Perseus have converted their diachronic Ancient Greek treebanks to Universal Dependencies format and released them.[7] The newly developed Treebank of Vedic Sanskrit (Hellwig et al. 2020), which is partly also available in the Proiel scheme of annotation, is now being expanded based on Universal Dependencies.[8] These treebanks will enable users to run multilingual comparative diachronic analyses and will strengthen the essential role that treebanks will doubtless increasingly play in the near future, in historical linguistics as well as in general linguistics.

Nonetheless, diachronic treebanks, like synchronic treebanks for historical/ancient languages, show a number of peculiar features that make them different from those for modern languages. These features concern both the building and the use of such linguistic resources.

With respect to building, diachronic treebanks raise specific issues concerning the selection, nature and size of the source data. While selectional criteria are a core issue in the building of synchronic treebanks for modern languages, diachronic

2. http://www-users.york.ac.uk/~lang22/YcoeHome1.htm

3. https://www.ling.upenn.edu/hist-corpora/

4. http://www.voies.uottawa.ca/corpus_pg_en.html

5. https://perseusdl.github.io/treebank_data/

6. http://itreebank.marginalia.it

7. http://universaldependencies.org/treebanks/grc_proiel/index.html, http://universaldependencies.org/treebanks/grc_perseus/index.html

8. https://github.com/UniversalDependencies/UD_Sanskrit-Vedic/blob/master/README.md, http://foni.uio.no:3000/sources/110

treebanks are obliged to exploit to the maximum the few texts that remain from the past. This is an obvious consequence of data availability: while the bulk of available data for modern languages increases on a daily basis, corpora for ancient languages are closed, with no (or very few) new additions. Corpora for modern languages strive to be as representative as possible of a language or of one specific variety of it, and they rely on a huge amount of – in principle – open-ended material. Diachronic treebanks, by contrast, tend to include full texts of single authors or, often, all available texts in order to compensate for the limited size of the available dataset.

Diachronic treebanks are usually smaller than treebanks for modern languages, but the limited amount of available texts is typically not the only reason. Most efforts to build diachronic treebanks are quite recent, and the annotation work is typically more difficult and time-consuming than annotation of texts in modern languages. Modern treebanks can often balance shallow annotation with the availability of huge masses of data, as shown by the recent and already widespread use of deep learning[9] techniques such as word embedding. 'Word embedding' is a collective name for techniques that aim to quantify and categorize semantic similarities between linguistic items on the basis of their co-occurrence patterns in large datasets, converting their distributional profiles to vectors of numbers. Such techniques are extremely data-intensive and therefore normally out of reach for historical texts: there is not enough training data, nor is there enough data to compensate for the noise produced by automatic shallow annotation. To be of value, the annotation must be manually corrected.

There are also complications of historical treebanks not shared by modern treebanks, brought about by editorial issues connected with historical texts. Because of the way they are collected, texts of treebanks for modern languages do not usually raise specific editorial issues, as they normally come in a single version. When dealing with historical texts however, there can be several different versions of the same text, and choosing the critical edition to record in the corpus is a substantial part of the selection work, which also concerns the question of whether or not to include editorial apparatus (such as variants) in diachronic treebanks.

A further difference lies in the nature of the source data. In most cases, texts included in treebanks for modern languages are digital at the outset, while those in diachronic treebanks result from digitization processes. Especially when digitization results from the application of optical character recognition techniques, this can lead to errors, which need to be found and corrected.

With respect to use, treebanks including historical data differ from those for modern languages primarily through a different attitude to the texts themselves.

9. https://en.wikipedia.org/wiki/Deep_learning

Historical treebanks tend to include literary, historical, philosophical or documentary texts. Treebanks for modern languages, on the other hand, are often based on texts from newspapers. Users of data from modern treebanks are interested in exploiting such resources to provide empirical evidence either in support of general tendencies of a language or for different purposes in the area of natural language processing (such as, for instance, inducing grammars and training stochastic parsers). Users of historical treebanks, on the other hand, are in most cases interested in the texts themselves. This makes the results of the analyses that are run on such treebanks more bound to the specific kind of data provided by the resource in question and, thus, less portable in terms of linguistic generalizations. But such an interest in the very empirical evidence provided by diachronic treebanks also increases the value itself of the distinctive features of the texts included there and supports research focused on specific linguistic aspects in a specific period of time, which is a major issue in diachronic studies. In this volume, Korkiakangas' paper exemplifies work that exploits data from quite peculiar Latin texts (medieval charters), whose results cannot be generalized to the entire Latin language but concern one of its specific instantiations.

4. Historical treebanks in use

So far only a small number of scholars in historical linguistics use treebanks in their everyday work, as can be demonstrated, among other things, by the limited number of papers describing treebank-based research published, for instance, in a top-class journal like *Diachronica*. This is partly due to the still limited availability of representative diachronic treebanks, but also partly to the fact that many historical linguists are simply not informed about the very existence of such resources. Treebank providers must therefore increase awareness of their products. As a matter of fact, several papers in this volume are authored by scholars presenting research that makes use of treebank data that they have directly contributed to building, and it is one of the aims of this volume to foster broader use of treebanks in historical linguistics.

Diachronic treebanks allow data extraction aimed to assess the scope and effects of diachronic developments, managing a large amount of data and retrieving information whose relevance can then be evaluated through statistical methods. In this book, we especially emphasize the point made by Haug (2015): the use of treebanks in historical linguistics allows us to publish research that is truly replicable.

Publications in historical linguistics that deal with the same topic and use the same text sources may still reach very different conclusions and report very different statistics for the phenomenon under scrutiny. As an example, Haug (2015) cites

word order statistics reported by a number of researchers for the Gospel of Luke and Acts (which were written by the same author). The statistics differ to an alarming extent, because the researchers have used different selectional criteria and made different theoretical assumptions, without necessarily making these criteria and assumptions explicit. This means that their results are impossible to compare and replicate. Research on several of the topics covered in this volume also suffers from such problems, for instance the literature on the rise of *po* delimitatives discussed by Eckhoff. Indeed, by extracting her data from the Tromsø Old Russian and OCS Treebank, Eckhoff shows that in the OCS and Old East Slavic data sets, the *po* delimitative is not as marginal as is commonly claimed in the literature.

The use of treebanks can help to remedy this situation in two ways. First, a good treebank is annotated consistently, using an annotation scheme founded on broad consensus in the linguistic community, such as the simplified phrase structures in the Penn annotation scheme or the notion of syntactic dependency in the various dependency grammar schemes. Such treebanks are not created for the aims of one specific line of research, and exploiting the empirical evidence provided by such (publicly available) linguistic resources prevents the vicious circle of creating a corpus with the specific aim of studying a single linguistic phenomenon (Sinclair 2005). Thus, simply using a good treebank makes a number of underlying theoretical assumptions explicit.

Second, to retrieve data from a treebank requires a clear and explicit query expression which lists the selectional criteria. Ideally, the query criteria should therefore be published with the research results, as should the full data set. Any additional annotations and classifications that do not come directly from the treebank should be made explicit in the data set. Online data repositories offering persistent identifiers are the best way to publish such data sets. We intend this volume to serve as an example in this respect: all of the papers report the queries that have been run to collect the empirical evidence used to support the authors' conclusions, either in appendices or in online repositories. Several also provide full data sets and the scripts used to process them. Given that the audience of this book does not consist of computer scientists and computational linguists, the authors have also striven to explain the details of their queries (according to the specific query language they used) and analyses, and to provide readers with the opportunity of running them themselves.

5. Aims and scope of this volume

This volume presents a series of studies that demonstrate the potential of a number of mature diachronic treebanks that are now available. For English, we use the York-Toronto-Helsinki corpus[10] and the Penn Corpora of Historical English[11] (Taylor et al. 2003a and b; Kroch & Taylor 2000; Kroch et al. 2004). For French, we use the MCVF corpus[12] (Martineau 2008). For Russian and Old Church Slavonic, we use the Tromsø Old Russian and OCS Treebank[13] (Eckhoff & Berdičevskis 2015), as well as PROIEL[14] (Haug & Jøhndal 2008), which is also used for Latin and Ancient Greek. For Latin, we also use the Late Latin Charter Treebank (Korkiakangas & Passarotti 2011), which was built along the lines of the Perseus Latin and Ancient Greek Dependency Treebanks[15] (Bamman & Crane 2011) and the *Index Thomisticus* Treebank[16] for Latin (Passarotti 2011).

Our aim is to demonstrate the multiple ways in which diachronic treebank data may be used to advance historical linguistics. Treebank data may not only shed new light on vexed topics in the literature, but may also be the foundation of considerable methodological advances. Provided that they are built in a way that takes into consideration both the peculiarities of the text material and the overall standards in the treebank community, they may provide much-needed transparency and replicability to studies in historical linguistics. They may also be used to develop techniques that can to some extent compensate for the inherent problems of the historical text sources, which tend to be archaic, skewed to certain genres and often sparse. This volume contains papers that touch on all of these topics.

5.1 Old topics, new methods

The bulk of the papers in this volume exploit the advent of diachronic treebanks to subject longstanding issues to large-scale data analysis, which was not possible before these resources became available. For example, Taylor & Pintzuk analyze split coordination in English. Crucially, PoS-tagged corpora would not suffice to allow data extraction to the same extent as a treebank does, nor would it collect

10. http://www-users.york.ac.uk/~lang22/YcoeHome1.htm

11. https://www.ling.upenn.edu/hist-corpora/

12. http://www.voies.uottawa.ca/corpus_pg_en.html

13. https://nestor.uit.no, http://torottreebank.github.io/

14. https://proiel.github.io/

15. https://perseusdl.github.io/treebank_data/

16. http://itreebank.marginalia.it

a sufficient number of occurrences of the construction studied in this paper. Split coordination in English is not a construction that can be located automatically in either plain text or even simple PoS-tagged corpora, and both its long period of attestation and its relative rarity preclude manual searching. The treebank-based study by Taylor & Pintzuk reveals an interesting result that could not possibly have been reached without the help of this type of resource: 'split coordination' in fact comprises two different constructions, one of which remained stable over time while the other was lost in the post-Middle English period.

The volume is not restricted to studies of syntactic change. Eckhoff demonstrates how enriched treebank data can be employed in an analysis of an important semantic-morphological change in Russian: the rise of the *po* delimitative and its consequences for the Russian aspect system. The main focus is on the association between derivational morphology and semantics. To explore this, verbs in the treebank have been enriched with tags indicating their internal structure (prefixation, stem, suffixation), as well as semantic tags. Thus, information about word-internal structure and semantics may be combined with the morphological and syntactic information that is already present in the treebank. The syntactic level makes it possible to identify potential delimitative contexts. Eckhoff employs all of this data to reevaluate substantially the chronology of the change.

Diachronic treebanks with good coverage also make large-scale statistical modeling possible. Ponti & Luraghi use treebank data to model a major syntactic shift: the rise of configurationality in Greek and Latin. Notably, data extraction from treebanks allows examination of issues that could not easily be addressed based on PoS-tagged corpora, such as the decline of null anaphora for referential null objects. This feature was approximated by the absolute frequency of the part-of-speech tags of the nodes adjacent to verbs. The loss of zero anaphora could be expected to be related to the skyrocketing rate of (personal) pronouns in that position for late varieties, which turns out to be the case. This paper is also a good example of another common feature of the papers mentioned above: the co-existence of existing historical linguistic questions (and related literature) with methods for data analysis and evaluation borrowed from other disciplines. For example, to run network analysis on data in Ancient Greek and Latin, Ponti & Luraghi apply a tool developed in the field of computational biology, both to build the networks and to calculate the topological indices used to evaluate their physical properties. The results are then interpreted linguistically in order to understand the rise of configurationality in these languages.

Overall, addressing core linguistic questions, exploiting empirical evidence enhanced with metalinguistic information, applying multi-disciplinary techniques of data analysis and providing replicable results are all distinctive characteristics of the papers included in this volume.

5.2 Treebanks, text attestations and methodology

In historical linguistics, the available text sources are a perpetual problem. We are, as already noted, restricted to what written sources have come down to us, and the text inventory we are left with is normally skewed, to some extent random, and often very sparse. The fact that we are restricted to only written sources is itself a limitation, since even poorly standardized written material tends to be conservative and does not necessarily reflect ongoing linguistic change. Some genres, such as religious texts, legal documents and metric poetry, tend to be more archaic than more narrative genres, often even formulaic. The advent of richly annotated diachronic treebanks provides new ways to handle the challenges posed by the state of the sources. Two of the papers in this volume make a direct methodological contribution to this problem by applying treebank data and statistical modeling to assess the relationship between attested texts and the vernacular.

Simonenko, Crabbé & Prévost address the issue of genre differences, by looking at the relationship between prose and verse in historical French. There is a long-held intuition that prose is more progressive than verse in reflecting linguistic change. Statistically modeling two major syntactic changes in the history of French, the loss of null subjects and the loss of OV word order, their initial result is that there appears to be a significant difference between the rates of change in prose and verse. However, casting their analysis in terms of grammar competition and operationalizing criteria to identify the competing abstract grammars, the authors are able to show that the pace of change is the same in prose and verse if one corrects for metalinguistic factors. Their paper thus demonstrates both the analytic and methodological advances a large-scale treebank study can offer in combination with abstract grammatical representations.

Korkiakangas' study of late Latin charters complements Simonenko, Crabbé & Prévost's work by using treebank data to assess the relationship between written text and formulaicity within single texts, rather than between genres. Charters are legal documents and always contain a considerable amount of formulaic language, but they also always contain a 'free' part in which the case in question is described. Korkiakangas exploits this difference to measure the likely visibility of various types of change in the formulaic part of the charters. He selects a number of phenomena that are commonly considered to be undergoing change in late Latin, measuring the extent to which they are reflected in the formulaic and free parts. He demonstrates that a number of innovations are in fact found in both text types and concludes that only phenomena which are particularly salient, either perceptually or syntactically, are preserved in their conservative form in the formulaic parts of the charters. Studies of this kind make it possible to make better use of conservative text genres in linguistic studies, rather than discarding them altogether.

6. Conclusions

The papers in this volume demonstrate the current use of diachronic treebanks in historical linguistics. They offer considerable advances, not only in providing structured data that allow innovative interpretations of longstanding issues, but also by opening new methodological avenues. A good treebank annotated according to the existing standards enables researchers to be explicit about their theoretical assumptions and selectional criteria, and opens the way for replication. Large-scale treebank data also allow filtering of problematic data in ways that were not available to us before. Diachronic and historical treebanks also open considerable opportunities for the creation of other, treebank-based resources. One example is the Homeric Dependency Lexicon (HoDeL) available at https://studiumanistici.unipv.it/?pagina=p&titolo=ling-larl-hodel.

HoDeL is a resource that was developed at the University of Pavia, which allows to extract all verbs from the Homeric poems along with their dependents, adding other morphosyntactic information and providing the link to the English translation. The dependents have been extracted using the Perseus treebank (see Zanchi & Luraghi 2020). Similarly, the syntactic subcategorization lexicon IT-VaLex (accessible at: https://itreebank.marginalia.it/itvalex; downloadable from: https://github.com/CIRCSE/ITVALEX) was induced from the Index Thomisticus Treebank (McGillivray & Passarotti, 2015). The Syntacticus treebank browsing facility (https://syntacticus.org) exploits data from the PROIEL family of treebank in similar ways: it provides generated dictionaries for each language, generated paradigms of attested forms for each lemma and attested valency frames for every verb. Such facilities thus give easy access to a wealth of information that could not be included in traditional dictionaries.

There is still much work to do in treebanking for ancient languages, and it looks as though syntactic annotation is just a step on a longer path. Together with building new treebanks for still under-resourced languages, and enlarging the already available treebanks, the research community dealing with linguistic resources for ancient languages is now in the process of enhancing textual corpora with different layers of semantic and textual information, including semantic role labeling, ellipsis resolution and coreference analysis. This trend is already being pursued by some of our authors (e.g. in Eckhoff's paper), and we hope such efforts will soon impact the world of diachronic linguistics, supporting research across all linguistic levels.

References

Bamman, David & Gregory Crane. 2011. The ancient Greek and Latin dependency treebanks. In Caroline Sporleder, Antel van den Bosch & Kalliopi Zervanou (eds.), *Language technology for cultural heritage: Selected papers from the LaTeCh workshop series*, 79–98. Berlin: Springer. https://doi.org/10.1007/978-3-642-20227-8_5

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová et al. 2013. *Prague dependency treebank 3.0*. Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czech Republic. http://ufal.mff.cuni.cz/pdt3.0/ (last accessed 5 July 2018.)

Eckhoff, Hanne Martine & Aleksandrs Berdičevskis. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS treebank. *Scripta & e-Scripta* 14–15. 9–25.

Eckhoff, Hanne Martine, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Trygve Truslew Haug, Odd Einar Haugen & Marius Larsen Jøhndal. 2018. The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52(1). 29–65. https://doi.org/10.1007/s10579-017-9388-5

Freddi, Maria & Silvia Luraghi. 2013. Appendix – Resources in syntax. In Silvia Luraghi & Claudia Parodi (eds.), *The Bloomsbury companion to syntax*, 463–468. London: Bloomsbury.

Haug, Dag Trygve Truslew. 2015. Treebanks in historical linguistic research. In Carlotta Viti (ed.), *Perspectives on historical syntax*, 185–202. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.169.07hau

Haug, Dag Trygve Truslew & Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In Caroline Sporleder & Kiril Ribarov (eds.), *Proceedings of the language technology for cultural heritage data workshop (LaTeCH 2008)*, 27–34. Marrakech, Morocco.

Hellwig, Oliver, Salvatore Scarlata, Elia Ackermann & Paul Widmer. 2020. The Treebank of Vedic Sanskrit. In *Proceedings of LREC*.

Korkiakangas, Timo & Marco Passarotti. 2011. Challenges in annotating medieval Latin charters. *Journal of Language Technology and Computational Linguistics* 26. 103–114.

Kroch, Anthony, Beatrice Santorini & Lauren Delfs. 2004. *The Penn-Helsinki parsed corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania.

Kroch, Anthony & Ann Taylor. 2000. *The Penn-Helsinki parsed corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania.

Martineau, France. 2008. Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus* 7. http://corpus.revues.org/1508 (last accessed 5 July 2018.)

McEnery, Tony, Richard Xiao & Yuko Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.

MvGillivray, Barbara & Passarotti, Marco. 2015. Accessing and using a corpus-driven Latin Valency Lexicon. In Gerd V. M. Haverling (ed.), *Latin Linguistics in the Early 21st Century. Acts of the 16th International Colloquium on Latin Linguistics, Uppsala, June 6th–11th , 2011*, 289–300. Uppsala: Uppsala Universitet.

Passarotti, Marco. 2011. Language resources: The state of the art of Latin and the index Thomisticus treebank project. In Marie-Sol Ortola (ed.), *Corpus anciens et Bases de données*, « ALIENTO. Échanges sapientiels en Méditerranée », N°2, Nancy, Presses universitaires de Nancy, 301–320.

Sinclair, John. 2005. Corpus and text: Basic principles. In Martin Wynne (ed.), *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books: 1–16. http://ota.ox.ac.uk/documents/creating/dlc/ (last accessed 5 July 2018.)

Stein, Achim & Sophie Prévost. 2013. Syntactic annotation of medieval texts: The syntactic reference corpus of Medieval French (SRCMF). In Paul Bennett, Martin Durrell, Silke Scheible & Richard Whitt (eds.), *New methods in historical corpora*, 75–82. Tübingen: Narr.

Taylor, Ann, Anthony Warner, Susan Pintzuk & Frank Beths. 2003a. *The York–Toronto–Helsinki parsed corpus of Old English prose*. University of York.

Taylor, Ann, Marcus Mitchell & Beatrice Santorini. 2003b. The Penn treebank: An overview. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, 5–22. Dordrecht: Kluwer Academic Publishers.

Zanchi, Chiara & Silvia Luraghi. 2020. Presenting HoDeL – A new resource for research on Homeric Greek verbs. In *Papers from the Annual International Conference "Dialogue" (2020)*, additional volume, http://www.dialog-21.ru/dopmat/2020/

Split coordination in English

Why we need parsed corpora

Ann Taylor and Susan Pintzuk
University of York

In this article we provide a practical demonstration of how syntactically annotated corpora (treebanks), particularly the English Historical Parsed Corpora Series, can be used to investigate research questions with a diachronic depth and synchronic breadth that would not otherwise be possible. The phenomenon under investigation is split coordination, in which two parts of a conjoined constituent appear separated in the clause (e.g., *and this is where **my aunt** lives **and my uncle***). It affects every type of coordinated constituent (subject/object DPs, predicate and attributive ADJPs, ADVPs, PPs and DP objects of P) in Old English (OE); and it, or a superficially similar construction, occurs continuously throughout the attested period from approximately 800 to the present day. Despite its synchronic range and diachronic persistence, split coordination has received surprisingly little attention in the diachronic literature, with the exception of Perez Lorido's (2009) limited study of split subjects in eight OE texts. Its modern counterpart is most frequently analysed as Bare Argument Ellipsis (BAE). Although the OE and Present-Day English constructions appear superficially similar, we show that not all of the OE data is amenable to a BAE analysis. We bring to bear different types of evidence (structural, discourse/performance effects, rate of change, etc.) to argue that split coordination in fact represents two different constructions, one of which remains stable over time while the other is lost in the post-Middle English period.

Keywords: coordination, ellipsis, annotated corpora, treebanks, syntactic change, history of English

1. Introduction

In this article we provide a practical demonstration of how treebanks, i.e. morpho-syntactically annotated (parsed) corpora, in particular the English Historical Parsed Corpora Series, can be used to investigate research questions with a diachronic depth and synchronic breadth that would not otherwise be possible. The English

https://doi.org/10.1075/bct.113.01tay

Historical Parsed Corpora Series is a collection of historical treebanks created at the University of Pennsylvania and the University of York, which provides continuous coverage of the English language from the earliest attested Old English (OE) texts through to Present-Day English (PDE). The corpora are all annotated using the same guidelines, so that syntactic variation and change can be tracked through the entire history of English.

The phenomenon under investigation, which we refer to descriptively as 'split coordination', is illustrated in (1), with examples from OE taken from the York-Toronto-Helsinki Corpus of Old English prose (YCOE). Every type of coordinated constituent (subject and object DPs, predicate and attributive ADJPs, ADVPs, PPs and DP objects of P) can be split. Furthermore, this is not just an OE phenomenon; split coordination, or a construction that is superficially similar, occurs continuously throughout the attested period from approximately 800 to the present day, as the parallel examples in (2) from PDE, taken from the Switchboard Corpus (Marcus et al. 1999), demonstrate.

(1) a. DP subject
oðþæt ***þæt ad*** *wæs forburnen,* ***and ealle þa tunnan***
until the pile was burned and all the casks
"until the pile and all the casks were burned up"
(coaelive,+ALS_[Julian_and_Basilissa]:332.1143)[1]

b. DP object
God sende ða ***fyr*** *on merigen* ***and fulne swefel*** *him to*
God sent then fire in morning and foul brimstone him to
"God then sent fire and foul brimstone to him in the morning"
(coaelive,+ALS[Pr_Moses]:211.2976)

c. PP
& ***on sorhge*** *leofodon &* ***on geswincum*** *siþþan*
and in grief lived and in torment afterwards
"and [they] lived afterwards in grief and torment"
(colsigewZ,+ALet_4_[SigeweardZ]:117.49)

(2) a. *and this is where* ***my aunt*** *lives* ***and my uncle,***
b. *you put, um,* ***really good vanilla flavoring*** *in it* ***and some butter***
c. *my only experience with it, I was* ***in Central America*** *for a while,* ***and, uh, in San Salvador, in El Salvador,***

Despite its synchronic range and diachronic persistence, split coordination has received surprisingly little attention in the diachronic literature. Its occurrence in OE is often mentioned (Kohonen 1978; Mitchell 1985: §§1464–1472; Reszkiewicz

1. Example references are to the corpus.

1966; Sielanko 1994; Traugott 1972); but beyond Perez Lorido's (2009) suggestive, but rather limited, study of split subjects in eight OE texts, it hasn't been seriously investigated. Its modern counterpart, which is most frequently analysed as a type of Gapping known as Stripping or Bare Argument Ellipsis (BAE), has been discussed in the literature since at least the sixties (Hankamer & Sag 1976; Johnson 2006; Reinhart 1991; Ross 1967), but no empirical corpus-based studies of its use exist.

One reason for the lack of quantitative, empirical investigations of this construction is, perhaps, that while the number of split coordinations is by no means negligible, neither is it high enough that sufficient numbers of examples can easily be collected without computational aids. Perez Lorido's study, based on manually collected data, is a case in point. He limits his study to coordinated subjects only, and the total number he collects from his eight texts is 731, of which 142 (19.4%) are split. By contrast, a search of the YCOE uncovers 3,391 coordinated subjects – more than four times as many as Perez Lorido – out of a total of 139,775 nominative subjects (2.4%), with 629 of these split (18.5%), not to mention over 2,000 cases of conjoined objects as well as smaller but still healthy numbers of the other categories. The situation in PDE is even more difficult, as here the construction occurs at frequencies well below 10%.

In addition to the problem of low frequency, this construction, whether split or not, is not uniquely marked by any lexical item; the only item common to these constructions is the conjunction itself. Therefore, without a parsed corpus, we could search only for *and* and other conjunctions and their variant spellings, or, in a part-of-speech (POS) tagged corpus, for the POS 'conjunction'. However these retrievals will suffer badly from low precision (too much unwanted data), since the search will retrieve every token containing a conjunction, and there is no way to disregard conjunctions that are irrelevant, e.g., those conjoining clauses rather than smaller constituents, and no automatic way to separate split coordination from non-split coordination.

Given the low frequency and non-uniqueness of this construction, retrieving it manually from printed texts or even from a text/POS-tagged corpus will be at best limited in scope and inefficient and at worst error-prone and unrepresentative.[2] In the English parsed corpora we use in this investigation, by contrast, (split) coordinations are explicitly marked, making it possible to quickly and accurately retrieve the relevant data.

2. This is not to say smaller manual studies are impossible, as Perez Lorido (2009) shows.

2. The case study

As noted above, all coordinated categories can and do split, but to keep things manageable, we focus here only on subject coordination, the most common type. Furthermore, we limit the study to the following three research questions, out of the many we could pursue:

1. What is the frequency of split subject coordination over time? Is it a stable construction? Is it changing? In which direction?
2. Is the construction in the historical corpora the same in all respects as that found in PDE?
3. What factors (weight, information structure, etc.) affect splitting/non-splitting?

2.1 Extracting the data

The data for this study are taken from the following corpora (all corpora with sources are listed in Appendix 1):

Corpus	Size in words	Date range
The York-Toronto-Helsinki Corpus of Old English Prose (YCOE)	1,450,376	800–1150
Penn-Helsinki Parsed Corpus of Middle English 2 (PPCME2)	1,155,965	1150–1500
Parsed Corpus of Early English Correspondence (PCEEC)	2,159,132	1400–1710
Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)	1,794,010	1500–1700
Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE)	948,895	1700–1915
The Brown Corpus (Fiction and Imaginative Prose) (BROWN)	432,879	1960s
Wall Street Journal (WSJ)	851,496	1980s
CallHome (CH)	166,619	1990s
Switchboard (SWBD)	1,253,960	1990s

Although all the corpora are in ascii format and thus can be used on any platform and viewed and searched with any word processing program, in order to fully utilize the annotations, a search program sensitive to structure is required. We use CorpusSearch (http://corpussearch.sourceforge.net/), a program conceived and designed by Ann Taylor and Anthony Kroch and implemented by Beth Randall in Java. CorpusSearch is not corpus specific but will search any corpus in the correct format, including all the corpora in the English Parsed Corpora Series and related corpora in other languages. Queries that can be used to extract the data for this study are included in Appendix 2.

While all the corpora in our dataset are parsed in the Penn Treebank style, there are some differences between the historical and the present-day corpora with regard to how particular constructions, including coordination, are handled. For this reason, some of the searches differ in detail although the material retrieved is the same.

Example (3) illustrates the structure of split coordination in the historical corpora. The 2nd conjunct is linked to the rest of the subject by means of a co-indexed trace (*ICH*).[3] Example (3a) in tree form is given in (3c).

(3) The structure of split coordination in the historical corpora

a.

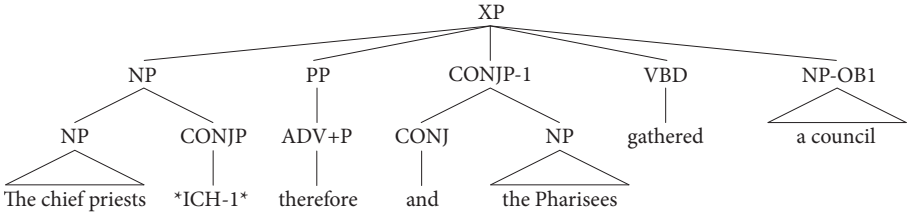
```
/~*
The chief priests therefore and the Pharisees gathered a
council,
(ERV-NEW-1881,11,40J.1030)
*~/
((IP-MAT (NP-SBJ (NP (D The) (ADJ chief) (NS priests)) <-- 1st
                                                          conjunct
                      (CONJP *ICH*-1)) <-- trace of 2nd conjunct
         (PP (ADV+P therefore))
         (CONJP-1 (CONJ and)          <-- 2nd conjunct
                  (NP (D the) (NPRS Pharisees)))
         (VBD gathered)
         (NP-OB1 (D a) (N council))
         (. ,))
(ID ERV-NEW-1881,11,40J.1030))
```

b.

```
/~*
Besides both Jesus was invited, and his Disciples to the
Marriage.
(PURVER-NEW-1764,2,1J.98)
*~/
((IP-MAT (ADVP (ADV Besides))
         (NP-SBJ (CONJ both)       <-- 1st conjunct
                 (NP (NPR Jesus))
                 (CONJP *ICH*-1))  <-- trace of 2nd conjunct
         (BED was)
         (VAN invited)
         (, ,)
         (CONJP-1 (CONJ and)       <-- 2nd conjunct
                  (NP (PRO$ his) (NS Disciples)))
         (PP (P to)
             (NP (D the) (N Marriage)))
         (. .))
(ID PURVER-NEW-1764,2,1J.98))
```

3. ICH is an inherited label from the Penn Treebank. It is not a theoretical construct but just stands for 'Interpret Constituent Here'.

c. Example (a) in tree form



In the PDE corpora the label dominating the 2nd conjunct is different (NAC rather than CONJP),[4] but the structure is essentially the same, as can be seen in (4), and thus retrieval is equally easy and accurate.

(4) The structure of split coordination in the PDE corpora (SWBD)

```
/~*
but then today the wind has dropped off, and also, the
temperature,
*~/
((S (CONJP (CC but) (RB then))
    (NP-TMP (NN today))
    (NP-SBJ (NP (DT the) (NN wind))   <-- 1st conjunct
            (NAC (-NONE- *ICH*-1)))   <-- trace of 2nd conjunct
    (VP (VBZ has)
        (VP (VBN dropped)
            (PRT (RP off))
            (, ,)
            (NAC-1 (CC and)           <-- 2nd conjunct
                   (ADVP (RB also))
                   (, ,)
                   (NP (DT the) (NN temperature)))))
```

In addition to the split coordinations, we need to collect the non-split cases, not only in order to generate frequencies, but also to act as a control on any potential explanation for why splitting occurs. Here the difference between the parsing of the historical and PDE corpora is a bit larger, but the relevant examples can, nevertheless, be retrieved with great accuracy in both cases. Examples of non-split coordinations in the historical corpora and the PDE corpora are given in (5) and (6).

4. NAC stands for 'Not A Constituent' and results from the lack of a Conjunction Phrase encompassing the conjunction and 2nd conjunct in the Penn Treebank parsing scheme; see Example (4). Note that no label (-NONE-) is given to the trace in (4); in (3), the trace is a CONJP. Split coordination does not occur in the Wall Street Journal, the first Penn Treebank corpus that was parsed and the one that much of the parsing scheme was developed to handle. The NAC label, originally used for other constructions, was co-opted to handle this construction during the parsing of the spoken Switchboard corpus.

(5) The structure of non-split coordinations in the historical corpora

```
/~*
for the bagpipes and the musicke went to wracke – (ARMIN,-
E2-H:11.98)
*~/
((IP-MAT (CONJ for)
         (NP-SBJ (NP (D the) (N+NS bagpipes))      <-- 1st
                                                   conjunct
                 (CONJP (CONJ and)
                        (NP (D the) (N musicke)))) <-- 2nd
                                                   conjunct
         (VBD went)
         (PP (P to)
             (NP (N wracke)))
         (. -))
(ID ARMIN,-E2-H:11.98))
```

(6) The structure of non-split coordination in PDE corpora (SWBD)

```
/~*
Both my mother's parents and my father's parents were
immigrants E_S
*~/
((S (NP-SBJ (DT Both)
            (NP (NP (PRP$ my) (NN mother) (POS's))
                (NNS parents))              <-- 1st conjunct
            (CC and)
            (NP (NP (PRP$ my) (NN father) (POS's))
                (NNS parents)))             <-- 2nd conjunct
    (VP (VBD were)
        (NP-PRD (NNS immigrants)))
    (-DFL- E_S))
```

2.2 The distribution of split subject coordination over time

The first step necessary for an empirical investigation of split subject coordination is to retrieve all the relevant tokens from each corpus, i.e., all coordinated subjects, whether they are split or not (see Appendix 2, A.1). We can then separate the split from the non-split tokens (Appendix 2, A.2). From these data, we can get an overall picture of the distribution of split subject coordination from the OE to PDE periods in all the corpora we have available, ordered approximately by date, as shown in Table 1.

Table 1 reveals a basic downward trend in the frequency of split subject coordination over time, with a strong rise at the end of the 1990s. This rise, however, is somewhat deceptive, as it comes from two speech corpora. Clearly this construction, at least in PDE, is restricted to more oral registers. The best modern comparators to our earlier corpora, all of which are necessarily written, are thus the written

Table 1. The frequency of split coordination in English historical and present-day corpora

Corpus	Non-split	Split	Total	% split
YCOE (800–1150)	2762	629	3391	18.55%
PPCME (1150–1500)	2212	277	2489	11.13%
PCEEC (1400–1710)	4049	292	4341	6.73%
PPCEME (1500–1700)	3948	223	4171	5.35%
PPCMBE (1700–1915)	1922	14	1936	0.72%
BROWN (1960s)	696	7	703	1.00%
WSJ (1980s)	1633	0	1633	0.00%
CALLHOME (1990s)	60	5	65	7.69%
SWBD (1990s)	398	38	436	8.72%

BROWN/WSJ corpora.[5] Given the oral aspect of this construction, we also need to be careful with the PCEEC corpus, which, while obviously not representing speech, is made up solely of personal letters and thus is designed to be as vernacular as possible. The PCEEC also overlaps partially in time with the PPCME and PPCEME, further complicating matters. We will thus exclude the PCEEC as well as the speech corpora from the main investigation. Removing the PCEEC, CALLHOME and SWBD from Table 1, and collapsing BROWN and WSJ, gives Table 2.

Table 2. The frequency of split coordination in English historical and present-day corpora, excluding and collapsing corpora

Corpus	Non-split	Split	Total	% split
YCOE (800–1150)	2762	629	3391	18.55%
PPCME (1150–1500)	2212	277	2489	11.13%
PPCEME (1500–1700)	3948	223	4171	5.35%
PPCMBE (1700–1915)	1922	14	1936	0.72%
WSJ/BROWN (late 20th c.)	2329	7	2336	0.30%

We have now answered our first question. Split subject coordination has always been a low frequency construction (in written texts), but shows a clear and fairly steady decrease over time. In addition, we have also identified one issue relevant to question 3, that split coordination is apparently sensitive to register.

5. This is the best available, but clearly not perfect, as the range of registers represented is much more limited than in the earlier corpora. A much better comparator would be the written part of the British National Corpus, but as it is not parsed, it is impossible to extract the relevant data, as discussed above.

2.3 A comparison of PDE with earlier stages of the language

We focus in this section on the cause of the decline in frequency of the split coordination construction. The sensitivity to register raises the possibility that the decline evident in Table 2 is simply an external effect, perhaps the result of standardization and/or a prescription against this construction in writing, and doesn't represent a change in the syntax of the language but only a change in register norms, or something similar. This might explain the fact that the major decline in split coordination post-dates the Middle English period and that in the modern language it can be found in speech and fiction; in contrast, the Wall Street Journal corpus furnishes no examples despite its large size. A closer look at the data, however, raises the possibility that, although the OE and PDE constructions appear superficially similar, we are actually looking at two different constructions. The relevant difference can be seen by comparing the examples in (1) and (2) and in (7) and (8); while the 2nd conjunct in PDE is overwhelmingly found in clause-final position,[6] the same is not true in the earlier stages of the language. In fact, final position for the 2nd conjunct in Old and Middle English is actually less common than non-final position: only about 30% of the subject 2nd conjuncts occur in final position in these early stages.

(7) a. *but then today* ***the wind*** *has dropped off, and also, the* ***temperature,***
(SWBD)

b. ***a new carrier*** *was coming in* ***and, uh, the, uh, attendant, uh, support vessels.***
(SWBD)

c. ***A cold supper*** *was ordered* ***and a bottle of port.*** (BROWN)

d. *"****Fear*** *possessed me,* ***and the certainty of war****", he has related.* (BROWN)

(8) a. ***Hys apostoli*** *arærdon* ***and heora æftergengan*** *manega men*
his apostles raised and their followers many men
of deaðe
from death
"His apostles and their followers raised many men from death"
(YCOE: coaelhom,+AHom_6:324.1027)

---

6. Two non-final cases occur in SWBD, given in (i) and (ii). The first case has an indirect question following the 2nd conjunct; and the second case, a relative clause modifying both conjuncts.

(i) *on publicity and letting* ***realtors*** *know* ***and key people*** *how wonderful the schools are*

(ii) *I hate to see* ***a car*** *going down the street,* ***or even a truck or bus for that matter****, that's putting out a lot of dark smoke,*

b. *But so it bifel þat* ***Rudak*** *was slayn,* ***and Skater also****, in pleyn bataile*
but so it befell that Rudak was slain and Skater also in open combat
"so it befell that Rudak and Skater also were slain in open combat"
(PPCME: CMBRUT3,23.691)

c. *And* ***both Iesus*** *was called,* ***and his disciples****, to the mariage*
(PPCEME: AUTHNEW,-E2-H:II,1J.166)

Modern syntactic accounts have analysed split coordination as BAE because of the clause-final position of the 2nd conjunct in PDE. Thus, Example (7d) is derived as in (9) from two full conjoined clauses with deletion under identity of everything in the second clause except the subject.[7]

(9) **Fear** possessed me, **and the certainty of war** ~~possessed me~~ (BROWN)

If the derivation involves movement of the remnant (i.e., the XP of the 2nd conjunct) to a left peripheral clause position prior to deletion under identity of the remainder of the second clause (Johnson 2006: 425; cf. also Busquets 2006; Konietzko & Winkler 2010), then it accounts as well for cases where objects and other coordinated constituents are split, as shown in (10):

(10) you put, um, really good vanilla flavoring in it and you put some butter in it
you put, um, **really good vanilla flavoring** in it **and some butter**$_i$ you put t_i in it
you put, um, **really good vanilla flavoring** in it **and some butter**$_i$ ~~you put t_i in it~~

This analysis necessarily produces a clause-final 2nd conjunct and thus works well for PDE, but it is not so clear how the examples in (8) could be derived by the same mechanism, since the 2nd conjunct is clause-internal, followed by a direct object in (8a), a locative PP in (8b) and a PP complement of the verb in (8c). Clearly, earlier stages of English have another way of deriving split coordinations that can be used instead of, or in addition to, BAE. Given this, the proportion of final to non-final 2nd conjuncts over time is clearly of interest. Our next set of searches, therefore, takes all the cases of split subject coordinations and divides them into cases with a final or non-final 2nd conjunct. Of course, we could simply write a query to extract each type from each corpus individually; there is, however, a more efficient way, one that in addition prepares for subsequent searches of the data. CorpusSearch includes a facility to code tokens for any feature for which it is possible to search. We can therefore take our set of split coordinated subjects and code them for the

7. Another possible way to derive these examples is by extraposition of the 2nd conjunct (Munn 1993), which may be more or less attractive depending on one's theory and the structure it assigns to coordinated phrases. We will not pursue this alternative here.

position of the 2nd conjunct (see Appendix 2, A.3). The data can then be exported and analysed in a spreadsheet or statistical analysis program, such as R. Examples are given in (11) for two tokens, the first with a final 2nd conjunct, indicated by the CODING node, the second with a non-final 2nd conjunct.

(11) Using coding strings for easier calculation of statistics

a.

```
((IP-SUB (CODING final)
         (CS1-NP-NOM^1 (NP-NOM (D^N +t+at) (N^N ad))
                       (CONJP *ICH*-1))
         (BEDI w+as)
         (VBN forburnen)
         (, ,)
         (CONJP-1 (CONJ and)
                  (NP-NOM (Q^N ealle) (D^N +ta) (N^N tunnan))))
 (ID coaelive,+ALS_[Julian_and_Basilissa]:332.1143))
```

b.

```
((IP-MAT-SPE (CODING non.final)
             (NEG+CONJ ne)
             (ADVP-LOC (ADV^L +t+ar))
             (CS1-NP-NOM^1 (NP-NOM (N^N w+adla))
                           (CONJP *ICH*-1))
             (NEG ne)
             (BEPI bi+d)
             (, ,)
             (CONJP-1 (NEG+CONJ ne)
                      (NP-NOM (ADJ^N wanhal)))
             (VBN gemet)
             (. .))
 (ID coaelive,+ALS_[Thomas]:80.7594))
```

Table 3 shows the results of separating 2nd conjuncts by position: there is a steady decline in non-final 2nd conjuncts, with an unexpected peak in the PPCMBE.

Table 3. Split coordinated subjects: Position of 2nd conjunct

Corpus	Final	Non-final	Total	% non-final
YCOE (800–1150)	419	210	629	33.4%
PPCME (1150–1500)	202	75	277	27.1%
PPCEME (1500–1700)	179	44	223	19.7%
PPCMBE (1700–1915)	6	8	14	57.1%
WSJ/BROWN (late 20th c.)	6	0	6	0.0%

While it is not possible to investigate this spike in detail, due to space restrictions, a quick look at the non-final examples shows that four out of eight are from the Bible and repeat the word order of an earlier version, as illustrated in (12). Five out of the eight (three from the Bible) have a subject split only by a discourse particle (*therefore, then, too*, etc.), which calls into question their evidence for syntactically split

constituents. The remaining two examples[8] of this type in the PPCMBE give scant evidence for the continuation of this construction post-1700, and we can thus safely date the loss of this type of splitting to the end of the Early Modern English period.

(12) a. Late Modern English (1764)
*Besides **both Jesus** was invited, **and his Disciples** to the Marriage.*
(PPCMBE: PURVER-NEW-1764,2,1J.98)

b. Early Modern English (1611)
*And **both Iesus** was called, **and his disciples**, to the mariage.*
(PPCEME: AUTHNEW,-E2-H:II,1J.166)

(13) a. Late Modern English (1764)
***The real benefits** then **which have been conferred on us by the Resurrection of our Lord, the substantial advantages which it has effected for us in our state of religious probation**, seem to be the two following.*
(PPCMBE: FROUDE-1830,2,50.347)

b. Late Modern English (1905)
***Small cutters**, too, **or centre-boards, handled by local amateurs**, will now and again come dashing out…* (PPCMBE: BRADLEY-1905,201.46)

2.4 Factors favouring the splitting of conjuncts

Turning finally to our third research question, what factors trigger the splitting of coordinated subjects, we can use the corpora to investigate at least one possible factor: weight (or length or complexity). Mitchell (1985: §§1464–1472) subsumes split coordinations under a process he calls 'splitting of heavy groups', and it has been shown that weight is a key factor in rightward movement processes in OE in general (Pintzuk & Taylor 2006; Taylor & Pintzuk 2011, 2012a, 2012b, 2014). Thus, despite Perez Lorido's claim that weight is not a factor in the case of split subject coordination,[9] it seems a good candidate. The automatic counting of words and/or nodes in parsed corpora is possible with CorpusSearch, and taking advantage of the coding feature discussed above, measuring weight in terms of number of

8. One example is ***Paleness** sits on every face; **confused tremor and fremescence**; waxing into thunder-peals, of Fury stirred on by Fear*. (PPCMBE: CARLYLE-1837,1,149.338), in which the non-finite clause *waxing…* could be taken as belonging to *confused tremor and fremescence*, in which case the 2nd conjunct is final. The second example, *Was not **both my Topsail Yards** wounded, **and Maintop-Mast**, when I then bore down to the Enemy?* (PPCMBE: HOLMES-TRIAL-1749,41.654) is taken from trial data and represents direct speech. It is similar to examples found in the spoken Switchboard Corpus, as noted in fn. 6.

9. Perez Lorido's numbers show that on average, a split 2nd conjunct is one word longer than a non-split one, but he assumes without testing that this difference is insignificant.

words[10] is generally quite straightforward. Here we test two hypotheses: (i) longer coordinations are more likely to split than shorter ones (based on Mitchell's heavy groups claim); and (ii) the weight of the 2nd conjunct (possibly in comparison with the weight of the 1st conjunct) is a factor in promoting splitting, i.e., heavier 2nd conjuncts are more likely to split, *pace* Perez Lorido.

CorpusSearch counts words within a given node. As non-split coordinations are dominated by a single node (cf. Example (5)), obtaining the number of words in a non-split coordination is completely straightforward. Split coordinations, on the other hand, are not dominated by a single node (cf. Example (3)) with the result that each conjunct must be counted separately and the results summed. For technical reasons related to how CorpusSearch works, it is not easily possible to calculate the length in words of coordinations including embedded clauses (e.g., relative clauses) or for coordinations with shared constituents, as in *the husband and wife* or *the rude savage or uncultured boor*, where the determiner is shared in both cases; these two constructions are thus excluded from the statistics in Tables 4–5 (Appendix 2, A.5).

Table 4 shows the average length of split and non-split coordinations across the three early corpora; the later corpora are omitted as the splitting of coordinations in written texts is essentially over by 1700 (Table 2). In each case the average length of split coordinations is longer than that of non-split coordinations. The extra length is small (about 0.5–1.0 word longer) but significant ($p < 0.05$) for the first two corpora. For the PPCEME, the difference is not significant ($p < 0.1$). This is a potentially interesting difference, but a full exploration is beyond the scope of this paper.

Table 4. Average total length of coordination in words

Corpus	Split	Non-split	Difference
YCOE (800–1150)	6.88	5.99	0.89
PPCME (1150–1500)	7.25	6.46	0.79
PPCEME (1500–1700)	7.96	7.29	0.67
Average*	7.37	6.58	0.78

* The slight discrepancy here is due to rounding errors.

It is certainly not the case that only heavy groups split, as the split three-word examples in (14) show. In our data, three-word coordinations split about 5% of the time, while coordinations of four words and above in length split on average about 15% of the time, showing no particular trend as length increases, as shown in Table 5.

10. Weight/length/complexity can be measured in various ways (cf. Taylor & Pintzuk 2012b and references therein). In practice, weight is such a robust effect that it makes little or no difference what measure is used. Since word count is easy to automate, we use number of words here as the measure of weight.

These data appear to call into question the traditional labelling of this phenomena as 'splitting of heavy groups' (as also noted by Perez Lorido 2009: 35).

(14) a. ***Adam*** *þagyt* ***&*** ***Eua*** *næron onlysde,*
Adam yet and Eve not-were liberated
"Adam and Eve were not yet liberated"
(YCOE: coblick,HomS_26_[BlHom_7]:87.88.1110)

b. *þan shulde* ***pees*** *haue bene,* ***and reste*** *amongus ham, wiþouten eny envy.*
then should peace have been and rest among us without any envy
"then there should have been peace and rest among us without any envy"
(PPCME: CMBRUT3,220.3966)

c. *Did* ***he*** *pull down the Hay* ***or you?*** (PPCEME: LISLE,-E3-H:IV,114C2.104)

Table 5. Percentage of split coordinations by total number of words

Total length in words	N	% split
3	1399	4.9%
4	753	15.3%
5	1571	13.6%
6	751	16.0%
7	620	13.4%
8	416	18.0%
9	341	18.8%
10	157	14.6%

The second question, regarding the length of the 2nd conjunct in particular, is slightly more difficult to test due to the way that coordinations are annotated in the corpora. Most coordinations have the structure illustrated in (5), repeated here as (15), and thus counting the length of the 2nd conjunct, which is entirely dominated by the CONJP node, is straightforward. This number can then be subtracted from the total giving the length of the 1st conjunct as well. Slightly problematic here are coordinations which consist of conjoined single words. These coordinations are annotated as flat structures, i.e., without a CONJP node, as illustrated in (16).[11] In these cases, there is no defined 2nd conjunct that can be counted, and thus the counting of this category has to be done by hand. As in these cases the 1st conjunct

11. This approach to annotating single word coordinations was adopted wholesale from the Penn Treebank in order to save time and effort in the annotation process. In retrospect, it was clearly a mistake not to annotate all coordinations in a consistent manner.

is necessarily one word long, however, this type can be counted in the same way as other unsplit coordinations.

(15)

```
/~*
for the bagpipes and the musicke went to wracke - (ARMIN,-
E2-H:11.98)
*~/
((IP-MAT (CONJ for)
         (NP-SBJ (NP (D the) (N+NS bagpipes))  <-- structured
                                                   coordination
                 (CONJP (CONJ and)
                        (NP (D the) (N musicke))))
         (VBD went)
         (PP (P to)
             (NP (N wracke)))
         (. -))
(ID ARMIN,-E2-H:11.98))
```

(16)

```
/~*
But error and phantasie, do commonlie occupie, the place of
troth and iudgement. (ASCH,-E1-P2:14V.94)
*~/
((IP-MAT (CONJ But)
         (NP-SBJ (N error) (CONJ and) (N phantasie)) < "flat"
                                                       coordination
         (, ,)
         (DOP do)
         (ADVP (ADV commonlie))
         (VB occupie)
         (, ,)
         (NP-OB1 (D the)
                 (N place)
                 (PP (P of)
                     (NP (N troth) (CONJ and) (N iudgement))))
         (. .))
(ID ASCH,-E1-P2:14V.94))
```

A final difficulty is the structure of coordinations with more than two conjuncts. In the corpora, these have the structure given in (17). Two questions arise here, one conceptual and one practical. Conceptually, we need to decide what counts as the 2nd conjunct in non-split coordinations. If we look at the split cases of coordinations with multiple conjuncts, it is clear that the split overwhelmingly occurs after the 1st conjunct. Thus, we should count everything except the 1st conjunct together. This approach, however, leads to a practical problem, because these conjuncts are not dominated by a single node in the annotation and thus can't be counted automatically. In this case, therefore, we do the opposite of what we did with binary coordinations. We count the 1st conjunct (which is dominated by a node) and subtract it from the total, giving the length of the 2nd conjunct.

(17)

```
/~*
And the Lord sayd vnto Aaron, Thou and thy sonnes, and thy
fathers house with thee, shall beare the iniquitie of the
Sanctuary: (AUTHOLD,-E2-P1:XVIII,1N.1125)
*~/
((IP-MAT-SPE (NP-SBJ (NP (PRO Thou))
                     (CONJP (CONJ and)
                            (NP (PRO$ thy) (NS sonnes)))
                     (, ,)
                     (CONJP (CONJ and)
                            (NP (NP-POS (PRO$ thy) (N$ fathers))
                                (N house)
                                (PP (P with)
                                    (NP (PRO thee))))))
             (, ,)
             (MD shall)
             (VB beare)
             (NP-OB1 (D the)
                     (N iniquitie)
                     (PP (P of)
                         (NP (D the) (N Sanctuary)))))
(ID AUTHOLD,-E2-P1:XVIII,1N.1125))
```

With respect to the length of the 2nd conjunct as a factor in favouring splitting, our results confirm Perez Lorido's, as shown in Table 6:[12] on average the 2nd conjunct in a split coordination is one half to one word longer than in a non-split coordination. As with the overall length, this difference is significant ($p < 0.01$) for the two earlier corpora but not for the PPCEME. By contrast the difference in the length of the 1st conjunct between split and non-split coordinations is much smaller and varies in direction, and only the average difference over all three corpora is significant.[13]

Table 6. Average length of the 2nd and 1st conjunct and difference (split – non-split)

Corpus	2nd conjunct			1st conjunct		
	Split	Non-split	Difference	Split	Non-split	Difference
YCOE	5.06	4.08	0.98	1.82	1.92	−0.10
PPCME	5.22	4.35	0.87	2.03	2.12	−0.09
PPCEME	5.51	4.93	0.58	2.46	2.36	0.10
Average	5.18	4.46	0.72	2.00	2.13	−0.13

12. Conjunctions occurring before the 1st conjunct are counted as part of that conjunct; conjunctions occurring before the 2nd conjunct are counted as part of that conjunct.

13. Welch 2 sample t-test: 1st conjunct: YCOE: $p = 0.09$, PPCME: $p = 0.40$, PPCEME: $p = 0.60$, average $p = 0.02$; 2nd conjunct: YCOE: $p = 3.252e\text{-}08$, PPCME: $p = 0.002$, PPCEME: $p = 0.07$, average $p = 1.177e\text{-}07$).

If we look at the percentage of split coordinations by the length of the 2nd conjunct (Table 7), we see that most of the effect is concentrated at the low end, with the percentage rising between 2 and 4 words,[14] but then more or less levelling off, in the same way as in Table 5.

Table 7. Percentage of split coordinations by length of 2nd conjunct

Length of 2nd conjunct in words	N	% split
2	1620	5.1%
3	2047	12.5%
4	1062	17.1%
5	621	16.1%
6	451	19.7%
7	273	19.0%
8	191	15.2%
9	134	15.7%
10	81	11.1%

Thus, length (weight/complexity), particularly of the 2nd conjunct, is clearly a factor in the splitting of coordinations, but the effect appears to be fairly small and makes the most difference for the shortest items. As usual with this factor, it is not clear what it represents; for some discussion of this issue, see Arnold et al. (2000) and Taylor & Pintzuk (2012b).

Another factor that is likely to play a role in split coordinations is information structure, which is well known to influence rightward movement (e.g., Hinterhölzl 2009; Pintzuk & Taylor 2006). Perez Lorido claims that a split 2nd conjunct is defocused, while when non-split it is focused or foregrounded and generally receives more "communicative attention" (2009: 42ff).[15] Kiss (1996) for PDE and Biberauer & Kemenade (2011) for OE/ME propose two subject positions, the higher of which is reserved for specific subjects and the lower for non-specific. Although we would not claim that the conjuncts in split subject coordinations always fill the two subject positions,[16] specificity as a factor in leftward movement of the 1st conjunct is a

14. The coordination is included as part of the count of the 2nd conjunct, and thus a two-word 2nd conjunct is generally made up of a conjunction plus a single word conjunct.

15. We do not find Perez Lorido's analysis convincing due to the lack of any objective measure of the differences he claims in the information structure of split and non-split 2nd conjuncts. However, this does not negate the possibility, indeed the strong probability, that information structure is involved at some level.

16. Both subject positions precede the position normally filled by the finite verb (T or equivalent). In some cases both conjuncts do precede T, as in (3a) and (14a), and thus likely occupy the two subject positions (cf. van Kemenade & Milićev 2012: 249). However, frequently only the 1st conjunct precedes the finite verb.

plausible hypothesis. Finally, it is well known that earlier positions in the clause favour given information and later ones new (see, for example, the 'Given Before New Principle' of Gundel 1988), and thus information structure may also be relevant to this construction. Unfortunately, the annotation of information structure is much more difficult and time-consuming than the annotation of syntactic structure, as well as far less advanced; information structure is not included in the annotation of the English Historical Parsed Corpora Series. As a result most studies which include an information structure component up to now have been done manually (Bech 2001; Taylor & Pintzuk 2011, 2012a, 2012b, 2014). Other corpus projects (e.g., PROIEL, ISWOC) have started to explore methods to annotate information status along with syntax in their corpora, and in the future the spread and ease of carrying out such investigations should increase.

3. Conclusion

In this paper we have demonstrated, via a case study of split coordination, how researchers can track a construction across a long time period and investigate possible hypotheses concerning the frequency of occurrence of the construction over time, its syntactic structure and the factors that influence its use in different contexts. Like many syntactic constructions, split coordinations are extremely difficult to extract without a parsed corpus, since the lexical items and parts of speech that it contains are not limited to phrasal coordinations. In addition, as a low frequency construction, particularly in the later periods, the effort and time needed to find and extract the relevant examples would be difficult to justify. Bringing to bear different types of evidence (structural, discourse/performance effects, rate of change, etc.), we test the hypothesis that split coordination in fact represents two different constructions, one of which, Bare Argument Ellipsis, remains stable over time, while the other – which involves the movement of the 2nd conjunct out of the conjoined phrase – is lost in the post-Middle English period. As a discussion of the latter construction goes beyond the scope of this paper, the interested reader is referred to Taylor & Pintzuk (2017) for an analysis.

Acknowledgements

We would like to thank the audience at the Workshop on Diachronic Treebanks at the 49th meeting of the Societas Linguistica Europaea (Naples Aug. 31–Sept. 3, 2016) and four reviewers for helpful comments and suggestions.

References

Arnold, Jennifer, Anthony Losongco, Thomas Wasow & Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76(1). 28–55. https://doi.org/10.1353/lan.2000.0045

Bech, Kristin. 2001. Word order patterns in Old and Middle English: A syntactic and pragmatic study. PhD thesis, University of Bergen.

Biberauer, Theresa & Ans van Kemenade. 2011. Subject positions and information-structural diversification in the history of English. *Catalan Journal of Linguistics* 10. 17–69. https://doi.org/10.5565/rev/catjl.32

Busquets, Joan. 2006. Stripping vs. VP-ellipsis in Catalan: What is deleted and when? *Probus* 18. 159–187. https://doi.org/10.1515/PROBUS.2006.006

Gundel, Jeanette. 1988. Universals of topic-comment structure. In Michael Hammond, Edith Moravcsik & Jessica Wirth (eds.), *Studies in syntactic typology*, 209–239. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.17.16gun

Hankamer, Jorge & Ivan Sag. 1976. Deep and surface anaphora. *Linguistic Inquiry* 7. 391–426.

Hinterhölzl, Roland. 2009. Information structure and unmarked word order in (Older) Germanic. In Caroline Féry & Malte Zimmerman (eds.), *Information structure from different perspectives*, 282–304. Oxford: Oxford University Press.

Johnson, Kyle. 2006. Gapping. In Martin Everaert & Henk van Riemsdijk (eds.), *The Blackwell companion to syntax*, 407–435. Oxford: Blackwell. https://doi.org/10.1002/9780470996591.ch29

Kiss, Katalin. 1996. Two subject positions in English. *Linguistic Review* 13. 119–142. https://doi.org/10.1515/tlir.1996.13.2.119

Kemenade, Ans van, & Tanja Milićev. 2012. Syntax and discourse in Old English and Middle English word order. In Dianne Jonas & Stephen Anderson (eds.), *Grammatical change: Origins, nature, outcomes (Proceedings of DIGS VIII)*, 237–255. Oxford: Oxford University Press.

Kohonen, Viljo. 1978. *On the development of English word order in religious prose around 1000 and 1200* AD. Åbo: Åbo Akademi Foundation.

Konietzko, Andreas & Susanne Winkler. 2010. Contrastive ellipsis: Mapping between syntax and information structure. *Lingua* 120. 1436–1457. https://doi.org/10.1016/j.lingua.2008.08.009

Mitchell, Bruce. 1985. *Old English syntax*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198119357.001.0001

Munn, Alan. 1993. Topics in the syntax and semantics of coordinate structures. PhD thesis, University of Maryland.

Perez Lorido, Rodrigo. 2009. Reconsidering the role of syntactic "heaviness" in Old English split coordination. *Studia Anglica Posnaniensia* 45. 31–56.

Pintzuk, Susan & Ann Taylor. 2006. The loss of OV order in the history of English. In Ans van Kemenade & Bettelou Los (eds.), *The handbook of the history of English*, 249–278. Oxford: Blackwell.

Reinhart, Tanya. 1991. Elliptic conjunctions: Non-quantificational LF. In Aka Kasher (ed.), *The Chomskyan turn*, 360–384. Oxford: Blackwell.

Reszkiewicz, Alfred. 1966. Split constructions in Old English. In Mieczyslaw Brahmen, Stanislaw Helsztyński & Julian Krzyzanowski (eds.), *Studies in language and literature in honour of Margaret Schlauch*, 313–326. Warsaw: Polish Scientific Publishers.

Ross, John R. 1967. Constraints on variables in syntax. PhD thesis, MIT.

Sielanko, Elzbieta. 1994. Split coordinated structures in late Old English. *Studia Anglica Posnaniensia* 24. 58–72.

Taylor, Ann & Susan Pintzuk. 2011. The interaction of syntactic change and information status effects in the change from OV to VO in English. *Catalan Journal of Linguistics* 10. 71–94.

Taylor, Ann & Susan Pintzuk. 2012a. The effect of information structure on object position in Old English: A pilot study. In Maria-Jose López-Couso, Bettelou Los & Anneli Meurman- Solin (eds.), *Information structure and syntactic change*, 47–65. Oxford: Oxford University Press.

Taylor, Ann & Susan Pintzuk. 2012b. Rethinking the OV/VO alternation in Old English: The effect of complexity, grammatical weight and information structure. In Terttu Nevalainen & Elizabeth Traugott (eds.), *The Oxford handbook of the history of English*, 835–845. Oxford: Oxford University Press.

Taylor, Ann & Susan Pintzuk. 2014. Testing the theory: Information structure in Old English. In Kristin Bech & Kristine G. Eide (eds.) *Information structure and syntactic change in Germanic*, 53–77. Amsterdam: John Benjamins. https://doi.org/10.1075/la.213.03tay

Taylor, Ann & Susan Pintzuk. 2017. Split coordination in Early English. In Bettelou Los & Pieter de Haan (eds.), *Word order change in acquisition and language contact: Essays in honour of Ans van Kemenade*, 155–183. Amsterdam: John Benjamins. https://doi.org/10.1075/la.243.08tay

Traugott, Elizabeth. 1972. *A history of English syntax*. New York: Holt, Rinehart & Winston.

Appendix 1. Corpora

CallHome Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El- Bachouti, Robert Belvin & Ann Houston. OntoNotes Release 5.0 LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

Brown Corpus, Switchboard, Wall Street Journal, Marcus, Mitchell, Beatrice Santorini, Mary Ann Marcinkiewicz, & Ann Taylor. Treebank-3 LDC99T42. Web Download. Philadelphia: Linguistic Data Consortium, 1999.

PROIEL Pragmatic Resources in Old Indo-European Languages: http://www.hf.uio.no/ifikk/english/research/projects/proiel

ISWOC Information Structure and Word Order Change in Germanic and Romance Languages: http://www.hf.uio.no/ilos/english/research/projects/iswoc/

The English Historical Parsed Corpora Series

YCOE: Taylor, Ann, Anthony Warner, Susan Pintzuk & Frank Beths. 2003. York-Toronto-Helsinki Parsed Corpus of Old English Prose. University of York. Distributed through the Oxford Text Archive. http://www-users.york.ac.uk/~lang22/YcoeHome1.htm

YCOEP: Pintzuk, Susan & Leendert Plug. 2002. The York-Helsinki Parsed Corpus of Old English Poetry. Department of Linguistics, University of York. Distributed through the Oxford Text Archive, first edition (http://www-users.york.ac.uk/~lang18/pcorpus.html).

PPCME2: Kroch, Anthony & Ann Taylor. 2000. The Penn-Helsinki Parsed Corpus of Middle English. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4 (http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCME2-RELEASE-4).

PCEEC: Taylor, Ann, Arja Nurmi, Anthony Warner, Susan Pintzuk, & Terttu Nevalainen. 2006. York-Helsinki Parsed Corpus of Early English Correspondence. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

PPCEME: Kroch, Anthony, Beatrice Santorini, & Lauren Delfs. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3 (http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3).

PPCMBE: Kroch, Anthony, Beatrice Santorini, & Ariel Diertani. 2016. The Penn Parsed Corpus of Modern British English. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 1 (http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1).

Appendix 2. Search and coding queries

The data include all subjects of finite clauses which contain a conjunction or a POS-tag CONJP.

The basic set of searches to generate the data in the paper for the Penn Parsed Corpora of Historical English (PPCHE) (YCOE, PPCME, PCEEC, PPCEME, PPCMBE) are the same, except that the label for subjects is different: in the YCOE, the label indicates the nominative case-marking; in the later corpora, the label indicates the subject function. In the set of queries below, a definition is used for 'subject' which will work for both the YCOE and later corpora. The searches for the Penn Treebank corpora, which differ rather more from the PPCHE corpora, are given in § B.

A. Query files for the PPHE corpora

Note that these queries were run using CorpusSearch version 2.21. Using a later version of CorpusSearch may require a different use of the general wildcard * and the digit wildcard #.

The label ***subject*** used in the queries is a definition, and should be replaced with the appropriate label either by hand or by using the following definition file:

```
subject: NP-NOM|NP-NOM-RSP*|NP-NOM-x*|NP-SBJ*
```

A.1 Extract all coordinated subjects in finite clauses

query file: cs.q
input files: all corpus files
output file: cs.out

```
ignore_nodes: null
nodes_only: t
remove_nodes: t
node: IP-MAT*|IP-SUB*
query: (IP-MAT*|IP-SUB* iDoms ***subject***)
    AND (***subject*** iDomsMod NP !\*con*)
    AND (***subject*** iDoms CONJP|CONJ|NEG+CONJ)
```

The second line of the query is to eliminate a few cases of empty 1st conjuncts.

A.2 Separate split and non-split subjects

The print_complement:t command splits the input file into a file of hits that match the query (.out) and a file that doesn't (.cmp). In this case, the non-matching file contains the non-split coordinations.

query file: cs-split.q
input file: cs.out (output file from A.1)
output files: cs-split.out, cs-split.cmp

```
print_complement: t
ignore_nodes: null
node: IP-MAT*|IP-SUB*
query: (IP-MAT*|IP-SUB* iDoms subject)
 AND (subject iDoms CONJP)
 AND (CONJP iDoms \*ICH*)
 AND (IP-MAT*|IP-SUB* iDoms CONJP-#)
 AND (\*ICH* sameIndex CONJP-#)
```

cs-split.out contains all the split coordinated subjects (column 2 of Table 1)
cs-split.cmp contains all the non-split coordinated subjects (column 1 of Table 1)

Rename cs-split.cmp to cs-nonsplit.out for clarity

A.3 Code split coordinated subjects as split, and for the position of the 2nd conjunct (final/non-final), as illustrated in (11) for Table 3

coding query file: c2-position.c
input file: cs-split.out (output file from A.2)
output file: c2-position.cod

```
node: $ROOT
ignore_nodes: COMMENT|CODE|ID|LB|'|\"|,|E_S|.|/
coding_query:

/* code all tokens as split */

1: {
   split: ELSE
   }

/* split: final vs non-final */

2: {
   final: ($ROOT iDoms subject)
         AND (subject iDoms CONJP)
         AND (CONJP iDoms \*ICH*)
         AND (\*ICH* sameIndex CONJP-#)
            AND ($ROOT iDomsLast CONJP-#)
   non.final: ($ROOT iDoms subject)
           AND (subject iDoms CONJP)
           AND (CONJP iDoms \*ICH*)
           AND ($ROOT iDoms CONJP-#)
           AND (\*ICH* sameIndex CONJP-#)
           AND (CONJP-# precedes *)
   }
```

A.4 Code non-split coordinations as non.split and for type (needed for coding length; cf. A.5 below)

Because of the way certain non-split subjects are parsed, it is necessary to know specific information about the type of coordination in order to count the length of the conjuncts. The easiest way to do this is to code for the different types; all others are coded as '/' (i.e., NA = not applicable).

The following types need special treatment:

(18) Coordinations with multiple conjuncts (more than two): coded mult.conjp

(19) Coordinations with shared modifiers (labelled NX in the corpus): coded conj.x

(20) Word-level coordinations, referred to as 'flat', which lack a CONJP: coded flat

coding query file: special-nonsplit.c
input file: cs-nonsplit.out (output file from (A.2))
output file: special-nonsplit.cod

```
node: $ROOT
ignore_nodes: COMMENT|CODE|ID|LB|'|\"|,|E_S|.|/
coding_query:

/* code all tokens as non.split */
1: {
   non.split: ELSE
   }

/* non-split: special types */
2: {
 mult.conjp: ($ROOT idoms subject)
   AND (subject idoms [1]CONJP)
   AND (subject idoms [2]CONJP)
 conj.x: ($ROOT idoms subject)
   AND (subject idoms CONJP)
   AND (CONJP iDoms NX*)
 flat: ($ROOT idoms subject)
   AND (subject idoms CONJ|NEG+CONJ)
 /: ELSE
   }
```

A.5 Code for length of conjuncts

Coding for length is a rather complicated process, as outlined in the paper. The three numbers required are the length of the 1st conjunct (L(C1)), the length of the 2nd conjunct (L(C2)) and the length of the whole subject (L(subject)). If any two of these numbers can be generated automatically by CorpusSearch, the other can be calculated in a spreadsheet. In a few cases, counting has to be done (partly) by hand as it is not possible to generate more than one measurement automatically.

Note the following:

(21) If any length is measured by CorpusSearch as > 30, it is set to 30

(22) Subjects (split or unsplit) containing clauses are coded clause and not included in the length calculations, as noted in the paper

The table below summarizes schematically how lengths can be measured for various types of subjects. In Tables 4–7 in the paper, subjects with shared modification (conj.x) or containing clauses (clause) are omitted.

Type of subject/split	L(C1)	L(C2)	L(subject)
Subject (split or unsplit) containing a clause in any conjunct	coded 'clause'	coded 'clause'	coded 'clause'
Split subject	measured by CS	measured by CS	L(C1)+L(C2)
Unsplit subject with 2 conjuncts	measured by CS	L(subj) – L(C1)	measured by CS
Unsplit subject with more than 1 conjunct mult.conjp	measured by CS	L(subj) – L(C1)	measured by CS
Unsplit subject containing NX conj.x	manually counted	L(subj) – L(C1)	measured by CS
Unsplit flat subjects flat	necessarily 1 word	L(subj) – L(C1)	measured by CS

A.5.1 Code subjects for length of 1st conjunct (C1) and, where possible, length of 2nd conjunct (C2)

Lengths to be calculated in a spreadsheet are coded calculate.

Some errors in the parsing of coordinations in the corpora are detected by the coding below and manually removed from the spreadsheet. A small number of errors, however, are simply counted wrongly. Given the amount of data, this will not appreciably affect the results reported here, and we have not corrected them. Errors of this type are best corrected in the parsed files; alternatively they can be corrected in post-processing.

At this stage, split and non-split coordinations are coded for type and so are processed together.

coding query file: length-1.c
input files: c2-position.cod special-nonsplit.cod (output coded files from A.3, A.4)
output file: length-1.cod

```
node: $ROOT
ignore_nodes: COMMENT|CODE|ID|LB|'|\"|,|E_S|.|/
coding_query:

/* length of 1st conjunct in words */
3: {
 exclude: (CODING col 2 conj.x) /* exclude shared modifier (NX) type */
 \1: (CODING col 2 flat) /* assume length 1 for C1 of flat coordinations */
 clause: ($ROOT idoms subject)
    AND (subject idoms NP|NP-SBJ|NP-NOM)
    AND (NP|NP-SBJ|NP-NOM doms RMV*)
 \1: ($ROOT idoms subject)
    AND (subject idoms NP|NP-SBJ|NP-NOM)
    AND (NP|NP-SBJ|NP-NOM domsWords 1)
 \2: ($ROOT idoms subject)
    AND (subject idoms NP|NP-SBJ|NP-NOM)
    AND (NP|NP-SBJ|NP-NOM domsWords 2)

[coding for lengths 3–28 as above]

 \29: ($ROOT idoms subject)
    AND (subject idoms NP|NP-SBJ|NP-NOM)
    AND (NP|NP-SBJ|NP-NOM domsWords 29)
 \30: ($ROOT idoms subject)
```

```
    AND (subject idoms NP|NP-SBJ|NP-NOM)
    AND (NP|NP-SBJ|NP-NOM domsWords> 29)
 \1: ELSE /* leftovers are badly parsed flat split (and a few errors) */
  }
/* length of 2nd conjunct in words */
4: {
 calculate: (CODING col 1 non.split) /* C2 calculated for nonsplit type */
 clause: ($ROOT idoms subject)
    AND (subject idoms CONJP)
    AND (CONJP idoms \*ICH*)
    AND ($ROOT idoms CONJP-#)
    AND (\*ICH* sameIndex CONJP-#)
    AND (CONJP-# doms RMV*)
 \1: ($ROOT idoms subject)
    AND (subject idoms CONJP)
    AND (CONJP idoms \*ICH*)
    AND ($ROOT idoms CONJP-#)
    AND (\*ICH* sameIndex CONJP-#)
    AND (CONJP-# domsWords 1)
 \2: ($ROOT idoms subject)
    AND (subject idoms CONJP)
    AND (CONJP idoms \*ICH*)
    AND ($ROOT idoms CONJP-#)
    AND (\*ICH* sameIndex CONJP-#)
    AND (CONJP-# domsWords 2)
[coding for lengths 3-28 as above]
 \29: ($ROOT idoms subject)
    AND (subject idoms CONJP)
    AND (CONJP idoms \*ICH*)
    AND ($ROOT idoms CONJP-#)
    AND (\*ICH* sameIndex CONJP-#)
    AND (CONJP-# domsWords 29)
 \30: ($ROOT idoms subject)
    AND (subject idoms CONJP)
    AND (CONJP idoms \*ICH*)
    AND ($ROOT idoms CONJP-#)
    AND (\*ICH* sameIndex CONJP-#)
    AND (CONJP-# domsWords> 29)
 /* no leftovers */
 }
```

A.5.2 Code for length of whole subject, where possible

coding query file: length-2.c
input file: length-1.cod (coded file from (A.5.1))
output file: length-2.cod

```
node: $ROOT
ignore_nodes: COMMENT|CODE|ID|LB|'|\"|,|E_S|.|/
coding_query:
/* total length */
5: {
 clause: (CODING col 3 clause) /* for split subject */
 clause: (CODING col 4 clause) /* for split subject */
```

```
 clause: ($ROOT idoms subject)
    AND (subject doms RMV*)
 calculate: (CODING col 1 split) /* length calculated for split subjects */
 \3: ($ROOT idoms subject)
    AND (subject domsWords 3)
[coding for lengths 4-28 as above]
 \29: ($ROOT idoms subject)
    AND (subject domsWords 29)
 \30: ($ROOT idoms subject)
    AND (subject domsWords> 29)
 error: ELSE /* these really are mostly errors */
 }
```

B. Queries for Penn Treebank corpora (Brown, CallHome, Switchboard, Wall Street Journal)

The following queries will retrieve split and non-split coordinations, respectively, from the Penn Corpora. Note that, in order to use CorpusSearch on the Penn corpora, the format must be altered slightly, as detailed here: http://corpussearch.sourceforge.net/CS-manual/YourCorpus.html

B.1 Split subject coordinations

```
node: $ROOT
query: (NAC-# iDomsFirst CC)
AND (CC iDoms and|or|nor)
AND (CC|CONJP hasSister NP*)
AND (NP-SBJ* iDoms NAC)
AND (NAC iDomsMod -NONE- \*ICH*)
```

B.2 Non-split subject coordinations

```
node: $ROOT
query: (NP-SBJ* iDomsMod NP* CC)
AND (CC iDoms and|or|nor)
```

A corpus approach to the history of Russian *po* delimitatives

Hanne Martine Eckhoff
University of Oxford

This paper illustrates how enriched diachronic treebank data can shed new light on an old and vexed topic, even when that topic is primarily morphological and semantic in nature rather than syntactic. The topic is the rise of the Russian *po* delimitatives, a change seen as crucial in most accounts of the history of Russian aspect, since it represents a major step in generalising the derivational aspect system. Earlier accounts concur that the *po* delimitatives spread fairly recently, too recently for the development to be connected to the loss of the aorist tense, which also had delimitative readings with atelic verbs. Using treebank data from the Tromsø Old Russian and Old Church Slavonic Treebank, enriched with tags for derivational morphology and semantics, I show that the *po* delimitatives were not marginal even in the earliest Slavic sources, either in terms of frequency or semantics, and that they first complemented and then competed with the delimitative aorists. It can thus be claimed that the exotic *po* delimitatives grew organically out of the old Indo-European inflectional aspect system.

Keywords: aspect, delimitative, Slavic, prefixation, treebanks

1. Introduction

The rise of the Russian *po* delimitatives[1] is regarded as crucial to most diachronic accounts of Russian aspect. In the earliest East Slavic texts, such as the Kievan *Primary chronicle*, we find a situation where many of the features of the modern Slavic aspect

1. I use the term 'delimitative' in a wide sense, covering all verb events that are temporally bounded, without a *telos* (similar to the use of the term 'complexive' in the literature on Ancient Greek aspect). In the Slavistic literature the term is (for good reasons) often reserved for temporally bounded events of short duration, while longer durations are referred to as 'perdurative'. In the history of Russian, it is clear that the *po* prefix gradually shifts to specialise with short-duration events (cf. Dickey 2007). This is clearly a topic that should be pursued (see §6 on delimitative contexts in Old East Slavic). However, even in the late Middle Russian dataset, long-duration *po* delimitatives are still found.

https://doi.org/10.1075/bct.113.02eck

system are in place: verbs have a strong tendency to pair up or group into triplets or even larger clusters, where prefixed underived verbs are telic and strongly associated with a perfective-like behaviour, while suffix-derived verbs are strongly associated with an imperfective-like behaviour (e.g., Růžička 1957; Forsyth 1972; Mišina 2017). This is also the case in Old Church Slavonic (OCS), the earliest attestation of Slavic (e.g., Dostál 1954; Amse-De Jong 1974; Eckhoff & Haug 2015). However, many simplex verbs still display behaviour associated with both the perfective and imperfective aspect. The rise of *po* changes this landscape by gradually becoming a general perfectiviser for atelic verbs. This has two important consequences: firstly, almost all verbs now get a partner, and the formerly neutral simplex verbs become clearly imperfective. Secondly, prefixal perfectives need no longer be telic: in many cases *po* perfectivises simplex verbs by adding a purely temporal boundary for the verb event, not a *telos*, since many of these verbs have no inherent, natural endpoint.

In the earliest East Slavic texts, as well as in OCS, we also see a still-functional (albeit remodelled) version of the old Indo-European inflectional aspect system. In the past tense system the aorist and imperfect appear to express a viewpoint aspect distinction between perfective and imperfective, and in the participle system we see a similar distinction between so-called past and present participles. The nature of and division of labour between the new affix-derivational and the old inflectional system is a hotly debated issue (e.g., Meillet 1934; van Schooneveld 1951; Dostál 1954; Forsyth 1972; Amse-De Jong 1974; Bermel 1997 and many others). In this article I follow Eckhoff & Haug (2015) in claiming that they both expressed viewpoint aspect.

In fact, in any account of early Slavic aspect, the inflectional aspectual forms serve as important evidence when judging the semantics of individual verbs. Apparently perfective verbs rarely occur as imperfects and present participles, while apparently imperfective verbs rarely occur as aorists and past participles (see Eckhoff & Haug 2015 for OCS data). However, while all early prefixal perfectives are telic, the aorist readily combines with atelic verbs, yielding ingressive or delimitative readings. In the old system, there was thus an established way of encoding temporally bounded atelic events. A tempting hypothesis, then, is that the *po* prefix takes over one of the functions of the aorist and that its frequency is boosted when the aorist is subsequently lost from the system.

This article tests whether this hypothesis is diachronically plausible. An immediate problem is that the previous literature on the subject posits a relatively late date for the expansion of the *po* prefix (Dickey 2007, leaning on Sigalov 1975 and Dmitrieva 2000). However, I argue that the empirical research for these studies was not sufficient and that diachronic treebank data enriched with tags for inflectional morphology and semantics yield a different answer. The article thus also makes a case for the usefulness of enriched treebank data. The article is structured as follows. Section 2 briefly reviews relevant literature. Section 3 describes the data

and methodology. Section 4 is an analysis of the diachronic development of the semantics of *po*. Section 5 is a diachronic analysis of the verb classes found with delimitative aorists and delimitative *po* verbs. Section 6 uses syntactic data to compare the semantics of the delimitative aorist and delimitative *po* verbs in Old East Slavic. Section 7 is the conclusion.

2. Previous approaches

Dickey (2007) sees the rise and expansion of delimitative *po* as something of a mystery. In his view, which is supported by the previous literature, this function of the prefix was only marginally present in early historical times, and then it sharply expanded as late as in the 17th century. He cites Němec (1954), who hypothesises that the original semantics of the *po* prefix encompassed PATH, GOAL and SOURCE, i.e., all main spatial meanings related to motion along a trajectory. By historical times, Dickey (2007: 332) claims, the predominant semantics is PATH, which was also extended to a surface-contact meaning (*posmoliti* "cover with resin"), which tends to be telic since it often implies full coverage. The GOAL meaning of movement toward a landmark (*postignuti* "reach") is also still around. The SOURCE meaning is largely gone but is preserved in verbs that profile the inception of a motion event in time (*poiti* "go, set out"). The delimitative use, he concedes, had also appeared on the stage by historical times. Dickey sees it as a metaphorical extension of the PATH/SURFACE CONTACT sense of the prefix, from the spatial to the temporal domain. However, he concludes that this is a very small and semantically limited group of verbs.

Dickey (2007) did not do systematic empirical work himself but based his conclusions on empirical studies by Sigalov (1975) and Dmitrieva (1991, 2000). Sigalov (1975) claims that the early Slavic delimitatives, which were inherited from Common Slavic, comprised a very small class of what Dickey calls 'basic stative activity predicates', i.e., those referring to statives and low-intensity activities. According to the statistics offered by Dmitrieva (1991) on the basis of dictionaries (which means that she is dealing in type frequencies, not token frequencies), the share of delimitatives among *po* verbs was very low in the Old East Slavic period: only 3.8% were delimitative, while as many as 73.5% were resultative. In Modern Russian she finds that 39.8% are delimitiative,[2] while only 26% are resultative.

Sigalov (1975) suggests the following chronology for the expansion of the *po* delimitatives

2. As Dickey (2007) points out, given the extreme productivity of the delimitative *po* prefix in Modern Russian, 39.8% must be considered a very conservative estimate of the overall share of delimitatives among *po* verbs, since many of them will not make it into the dictionaries.

1. Common Slavic: *po* delimitatives were statives
2. 16th–17th century: spread to indeterminate motion verbs and psychological processes
3. 17th–18th century: spread to speech verbs, sound emission verbs, physical processes

On the basis of these data and proposals, Dickey (2007) asks whether a few stative delimitatives could really have the power to make *po* verbs lose their resultative sense in favour of delimitativity. He concludes that they could not, and he suggests that the late surge in productivity[3] was due to a hyper-frequent model verb: the partial-ingressive *poiti* "go, set out". Dickey claims that this verb could only become a prototype for activities when it paired up with imperfective *iti* "walk, go". His story is thus that *po* went through a prototype shift triggered by a strong model verb.

However, Dickey does not look at competing ways of expressing delimitativity in early Slavic. As stated in the introduction, there was in fact an established way to express this meaning: atelic and aspectually neutral verbs could interact with the aorist to produce this reading (OCS: Eckhoff & Haug 2015; Old East Slavic: Bermel 1995). As far as I am aware, the connection between the delimitative aorist and the delimitative *po* verbs has not been studied. It should be noted, though, that the aorist was lost by the 14th century.[4] If the reported chronologies are correct, a direct causal relationship seems implausible.

This article therefore examines the *po* delimitatives from the earliest attestations up to the 17th century, when the surge in productivity started, according to earlier work.

3. Data and method

The data are taken from the Tromsø Old Russian and OCS Treebank (TOROT) (Eckhoff & Berdičevskis 2015), a diachronic treebank of Russian which belongs to the PROIEL family of treebanks of early attestation of Indo-European languages (Haug & Jøhndal 2008; Eckhoff et al. 2018). At the time of extraction, the treebank contained approximately 160,000 tokens of OCS with morphological and syntactic

3. The term 'productivity' has many possible definitions. In this paper a word-formation pattern will be considered more productive the more lexemes are formed according to the pattern, and the less restricted the input lexeme is to particular semantic classes. For the delimitative *po* verb pattern, this means that the more delimitative *po* verbs we find, and the more semantic classes of base verbs are involved, the more productive the pattern is deemed to be.

4. The precise date of the loss of the aorist in spoken Old East Slavic is much disputed. For a careful discussion of the problem, see Živov (2017: 608–618).

annotation[5] (and a further 50,000 with morphological annotation and lemmatisation only), as well as approximately 230,000 tokens of Old East Slavic and Middle Russian with full morphological and syntactic annotation. The annotation follows the PROIEL dependency scheme, which is an enriched variant of dependency grammar designed to preserve as much linguistic detail as possible in the small but complex historical text sources. In particular, the scheme includes secondary dependencies and empty verb and conjunction nodes to give detailed representations of control, agreement and ellipsis phenomena (see the introduction to this volume).

For the purposes of this article it is especially important to note that OCS, Old Russian and Old East Slavic verb lemmas have been tagged for derivational morphology: prefix, suffix and stem (for more, see Eckhoff & Haug 2015). These tags may be crossed with the detailed morphological tags for tense and mood/finiteness, and they may be used to automatically classify verbs with great precision. In addition, semantic tags of various sorts were added to subsets of the data, as elaborated on further in the following sections.

Table 1. Overview of the four datasets[6]

	Texts	**Number of extracted verbs (excluding *byti* "to be")**
OCS (with Greek parallel)	Codex Marianus, Codex Zographensis	23,538
Old East Slavic (11th–14th century)	Chronicles,[7] Life of Boris and Gleb, Life of Feodosij Pečerskij, Russkaja pravda, charters and treaties	18,071
Early Middle Russian (15th century)	Afanasij Nikitin, Tale of the fall of Constantinople, Tale of Luka Koločskij, Tale of Dracula, Life of Sergij of Radonež', charters	6,240
Late Middle Russian (16th–17th century)	Tale of the taking of Pskov, Domostroj, Life of Avvakum	7,662

5. This includes the Codex Marianus, which is officially released by the PROIEL corpus at https://proiel.github.io/. This article also uses parallel data from the Greek Gospels, likewise taken from the PROIEL corpus. TOROT releases can be downloaded from http://torottreebank.github.io/.

6. The sizes of the four datasets directly reflect the bulk of texts per period in the TOROT at the time of data extraction. Since the two Middle Russian datasets are considerably smaller than the OCS and Old East Slavic ones, it is likely that we miss a number of individual *po* verbs that we might have captured, had the datasets been the same size. However, the relative type and token frequencies would probably not have differed much from what we see now.

7. The Primary Chronicle (Codex Laurentianus version in full, excerpts from the Codex Hypatianus version), excerpts from the First Novgorod Chronicle (Synodal manuscript), the Suzdal' Chronicle (Codex Laurentianus) and the Kiev Chronicle (Codex Hypatianus). For a full overview of the charters and treaties, see https://doi.org/10.18710/PUXWXL.

4. The semantic development of the *po* prefix

The previous literature claims that the delimitative meaning was rare with the *po* prefix in the earliest sources and that there was no sharp rise in the share of delimitatives until the 17th–18th century. To test this claim, I extracted all *po* verbs from my OCS, Old East Slavic, early Middle Russian and late Middle Russian datasets and classified them by their semantics. It should be noted that such classification must always be, to some extent, subjective and that it is certainly possible to argue against some of my classifications. In particular, it was sometimes difficult to distinguish between delimitatives and ingressives.[8]

The *po* verbs were tagged according to the following classification, adapted from Janda et al. (2013) and expanded with spatial semantic tags suggested in Dickey (2012). The classification is illustrated with OCS examples.

1. DELIMITATIVE: *požьdati* "wait for a while" from *žьdati* "wait", *poiskati* "seek for a while" from *iskati* "seek"
2. DISTRIBUTIVE: *pobiti* "throw (a lot, used to describe stoning)" from *biti* "beat", *pozobati* "peck (up a certain amount of seeds)" from *zobati* "peck"
3. GOAL: *poxoditi* "go towards" from *xoditi* "walk"
4. INGRESSIVE: *pomilovati* "take pity on" from *milovati* "pity, be merciful", *pokajati sę* "come to repent" from *kajati sę* "repent"
5. INGRESSIVE MOTION: *povesti* "lead away" from *vesti* "lead", *posъlati* "send (off)" from *sъlati* "send"
6. PATH: *poslědьstvovati* "follow" from *slědьstvovati* "follow"
7. RESULT: *pogreti* "bury" from *greti* "dig", *posešti* "cut down" from *sešti* "cut"
8. SEMELFACTIVE: *počělovati* "kiss once" from *cělovati* "kiss"
9. SURFACE CONTACT: *pomazati* "anoint, rub" from *mazati* "rub", *postьlati* "spread out, lay out" from *stьlati* "spread, lay out"

The expectation from the literature is that delimitatives will be marginal in the semantic network of the OCS *po* verbs. However, we find that this is not the case – in fact, Figure 1 shows that, in terms of type frequency, the DELIMITATIVE meaning is the second most frequent among the *po* verbs (19 lemmas, 18%), second only to RESULT (38%). Figure 2 shows that it is less prominent in terms of token frequency but hardly marginal at 10%. The *po* verbs classified as delimitiative include *požiti* "live", *pomysliti* "think", *poiskati* "seek" and *poslužiti* "serve", among others. For a full breakdown of verb classes, see §5.

8. To give the reader full opportunity to examine the classifications, the classified dataset is published at https://doi.org/10.18710/PUXWXL.

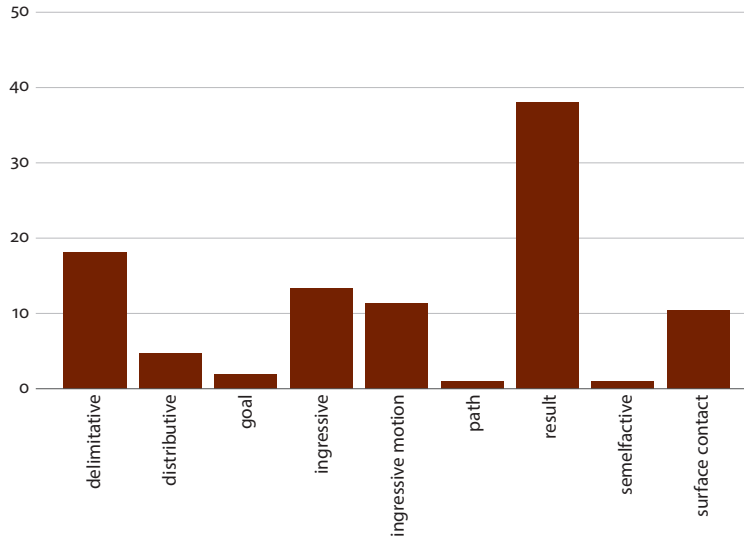


Figure 1. Semantic distribution of *po* verbs in OCS, predominant lemma meaning,[9] per cent (n = 105)

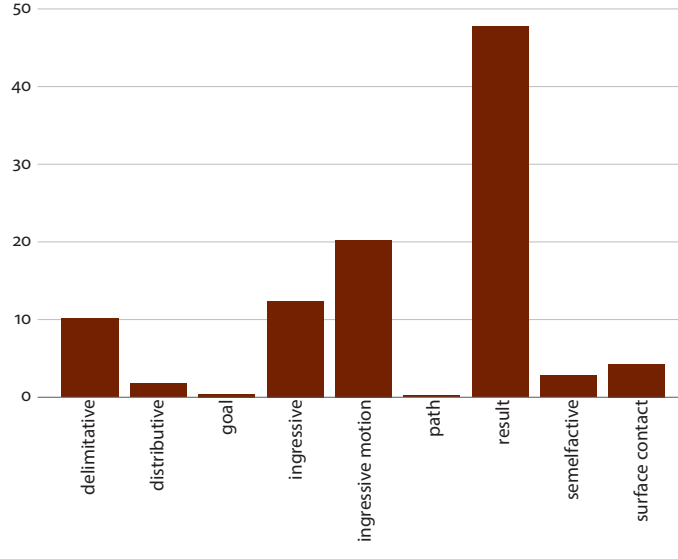


Figure 2. Semantic distribution of *po* verbs in OCS, tokens, per cent (n = 1334)

9. The semantic tags were assigned to the lemmas by relative majority: the most frequent meaning among the lemma's tokens was assigned. The same method was used for the Old East Slavic/Middle Russian data.

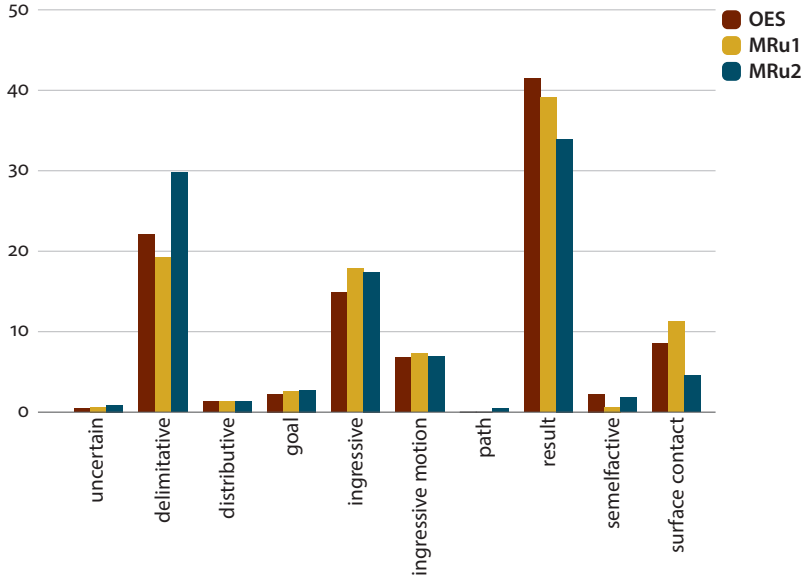


Figure 3. *Po* verb semantics in the history of Russian, predominant meaning of lemmas, per cent; OES = Old East Slavic, MRu1 = early Middle Russian, MRu2 = late Middle Russian

When we look at the Old East Slavic dataset (Figures 3 and 4), we see a similar situation. The DELIMITATIVE meaning is not marginal in terms of frequency even in the earliest attestations. 22.1% of the *po* verbs (49 out of 222) are primarily DELIMITATIVE, including *postojati* "stand", *pojasti* "eat", *poplakati* "cry" and *počitati* "read". For a full breakdown in verb classes, see §5. Looking at token frequencies, their share is somewhat lower, at 11.5% of all *po* verb attestations (268 out of 2323 occurrences) in the Old East Slavic dataset. However, both type and token frequency differ sharply from Dmitrieva's (1991) claim that the share of delimitatives was 3.8% in her Old East Slavic material (cf. §2). The most frequent meaning among the *po* verbs is again RESULT – 41.4% of all *po* lemmas in the Old East Slavic dataset have RESULT as their primary meaning, and 34.8% of the *po* verb tokens are used in a RESULT sense. Looking at the findings from OCS and Old East Slavic together, it is reasonable to claim that the DELIMITATIVE meaning was never marginal to *po* in attested times.

When we look at the diachronic development in type and token frequency (Figures 2 and 3), we do find a significant increase[10] in late Middle Russian (as

10. Increase in type frequency from early to late Middle Russian: $p = 0.0216$. Increase in token frequency from early to late Middle Russian: $p < 0.0001$. Fisher's exact test, two-tailed.

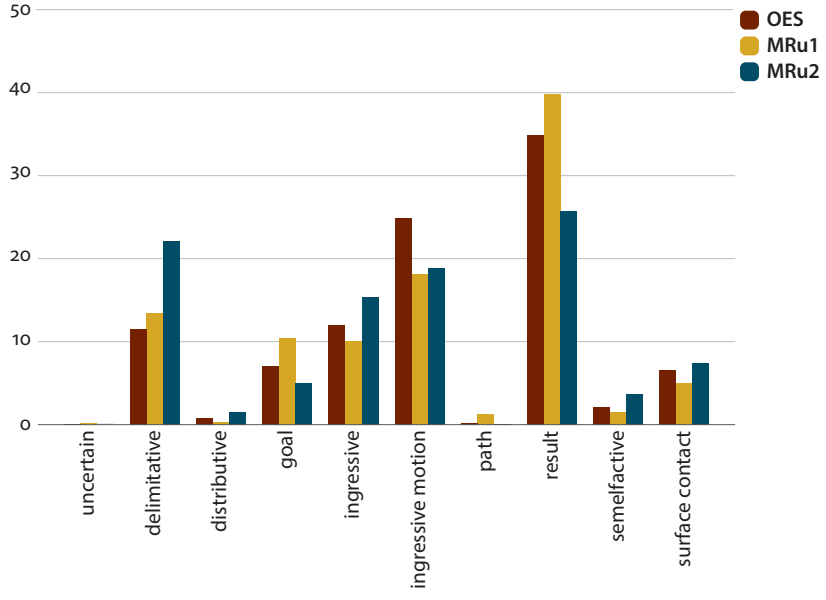


Figure 4. *Po* verb semantics 10th–17th century, tokens, per cent[11]

expected from Sigalov 1975 and Dickey 2007). The increase is not very sharp, and the RESULT meaning is the most frequent in all periods, though the DELIMITATIVE meaning is catching up in late Middle Russian.

All in all, in terms of frequency, the 17th century development is not as radical as one might expect from the literature.

11. All *po* verbs with possible delimitative readings were manually counted. For all other verbs all tokens were deemed to have the predominant semantics of the verb lemma in question (i.e. the semantics of the relative majority of the tokens of that lemma).

5. Aorists vs. *po*: Verb classes across time

We have established that, in terms of frequency, the DELIMITATIVE meaning was not marginal even in the earliest attestations of East Slavic or indeed in OCS, the very first attestation of Slavic overall. However, Sigalov (1975) and Dickey (2007) claim that the earliest *po* delimitative verbs are also SEMANTICALLY restricted, being limited to stative and low-intensity activity verbs. In this section I examine the verb classes found in delimitative *po* formations in all periods under scrutiny.[12] For the OCS and Old East Slavic datasets, I also compare the distributions to those of the delimitative aorists. Thus I will be able to determine (i) whether the early *po* delimitatives were really limited to a small semantically coherent group of verbs, and (ii) whether the same verbs and verb classes were found with both *po* delimitatives and delimitative aorists.

5.1 OCS

According to Sigalov (1975),[13] the expectation is that OCS *po* delimitatives will be statives or low-intensity activities.

However, when we look at the occurrences in the dataset (Table 2), only two of the 19 delimitative *po* verbs are arguably stative: *požiti* "live" and *postradati* "suffer" (tentatively applying the stativity tests in Lakoff 1966 and Dowty 1979). The largest group of verbs consists of plain activities – some of them are low-intensity, such as *požьdati* "wait" and *potrьpěti* "be patient, endure", but most of them are not, such as *poiskati* "seek" and *poslužiti* "serve".

(1) He will gird himself and have them recline at the table.

i	*minǫvъ*	*poslužitъ*	*imъ*
and	pass.PSTP	*po*.serve.*i*.PRS	they.DAT
kai	*parelthōn*	*diakonēsei*	*autois*
and	pass.AORP	serve.FUT	they.DAT

"and will come up and wait on them" (OCS, Mar. Luke 12.37)

12. The verb classes used are more granular than regular actionality class grids, such as Vendler's (1957) simple four-way classification of verbs as ±telic and ±dynamic and variations thereon, for easy comparison with the previous literature.

13. Sigalov takes his data from the file card index of the *Slovar' drevnerusskogo jazyka XI–XVII vv* (Institut russkogo jazyka) and also excerpts from the following texts: *Povest' vremennyx let* (Laurentian Codex), *First Novgorod chronicle* (Synodal manuscript), *Izbornik 1076 goda*, *Gramotki XVII–nač. XVII goda* (Moscow 1969) and *Pamjatniki russkogo narodno-razgovornogo jazyka XVII stoletija* (Moscow 1965). It is not clear whether the analysis of the excerpts from the sources was done in a systematic manner.

Table 2. Verb classes of *po* delimitatives and verbs attested as delimitative aorists/past participles in the OCS dataset

	***Po* delimitatives**	**Verbs occurring as delimitative aorists**
1. States (including positional verbs, but excluding psych verbs)	***požiti*** "live", *postradati* "suffer"	*bolěti* "be ill", *běsъnovati sę* "be possessed by a demon", ***žiti*** "live", *iměti* "have", *ležati* "lie", *mošti* "be able to", *sěděti* "sit", *trěbovati* "need"
2. Plain activities	*pobъděti* "stay awake", *požьdati* "wait", *poiskati* "seek", *pokaditi* "burn incense", *poslužiti* "serve", *potrьpěti* "be patient, endure", *potręsti* "shake"	*biti* "beat, hit", *doiti* "breastfeed", *dělati* "do", *dějati* "do, work", *krьstiti* "baptise (in numbers)", *učiti* "teach", *piti* "drink", *plakati* "cry", *postiti* "fast", *rydati* "sob", *sěti* "sow", *tvoriti* "do, make", *jasti* "eat"
3. Speech verbs	***pomoliti*** "pray", *ponositi* "rebuke, insult", *porǫgati* "mock", *poxvaliti* "praise"	***moliti*** "pray"
4. Psych verbs	*pomysliti* "think"	*věděti* "know", *věrovati* "believe", *mьněti* "think, consider", *xotěti* "want", *čisti* "honour"
5. Motion verbs	*ponesti* "carry for a while"	
6. Secondary derived verbs	*podvižati* "move", *pokyvati* "nod", *pomavati* "wave, make a gesture", *pomyšljati* "think"	

(2) *poištete* *mene* *i* *ne* *obręštete*
po.seek.PRS me and not find.PRS
zētēsete *me* *kai* *oukh* *heurēsete*
seek.FUT me and not find.FUT
"you will look for me but will not find me" (OCS, Mar. John 7.34)

Sigalov (1975: 161–162) argues that both these verbs are resultative in his Old East Slavic sources, but Examples (1) and (2) show their use in contexts with no result, where the limitation is purely temporal – in Example (1) the master will wait on his servants for the duration of the meal, and in Example (2) it is explicitly stated that even though the disciples will look for Jesus for a while, the search will have no result.[14]

14. Example (2) might also be read with a SURFACE CONTACT meaning: "you will search all over the place", but if so, it would be a good bridging context for the purely temporal delimitatives.

In addition to the 'plain' activities, there are also speech verbs (or more precisely activities that involve speech), such as *pomoliti* "pray", a psych verb (*pomysliti* "think") and a motion verb (*ponesti* "carry").[15]

Finally, there are four verbs that are *po*-prefixed but also have a secondary derivational suffix (i.e. row 6 in Table 2). These verbs are not always easy to interpret but appear at least to some extent to be delimitative AND iterative, i.e., the verb describes several temporally bounded bouts of nodding, waving, thinking etc. In Example (3), there are as many delimited head-wagging events as there are passers-by.

(3) *i* *mimo* *xodęštei* *xuděaxǫ* *i.* *i*
and by passing.*i*.PRSP.NOM insult.*i*.IMPERF he.ACC and
pokyvajǫšte *glavami* *svoimi*
po.nod.*va*.PRSP heads.INST their.INST
kai *hoi* *paraporeuomenoi* *eblasphēmoun* *auton* *kinountes*
and the pass_by.PRSP.NOM insult.IMPERF he.ACC move.PRSP
tas *kephalas* *autōn*
the heads.ACC their
"And those who passed by derided him, wagging their heads"
(OCS, Mar. Mark 15.29)

All in all, while OCS *po* delimitatives are not very frequent, they are certainly not limited to statives; in fact there are model verbs in all the most common atelic verb classes.

There are several interesting things to be noted about the TAM distribution of the OCS delimitative *po* verbs and their relationship to their Greek source forms (Table 3). We can start by noting that generally, and even in OCS TAM forms that are not aspectual in themselves (boldfaced in Table 3), the *po* verbs strongly tend to correspond to Greek perfective (aorist) forms. OCS imperfects and present participles are imperfective inflectional forms (cf. Eckhoff & Haug 2015) and also to a very large extent correspond to Greek imperfective forms. However, when we look at what OCS verbs we are dealing with here, we find that they are all formally expected to be imperfective: either they are occurrences of *ponositi* "rebuke, insult", which is formally (but not semantically) a prefixed indeterminate motion verb and expected to be imperfective on that count, or they belong to the group of secondary derived *po* verbs (*pokyvati* "nod", *pomavati* "wave, make a gesture", *pomyšljati* "think"; see Example (3)). There is thus strong support for assuming that the *po* delimitatives are perfective unless they are suffix-derived, i.e., that they follow the most common pattern for prefixed OCS verbs.

15. Technically, *ponositi* "rebuke, insult" is the indeterminate variant of *ponesti* "carry", but since it never occurs in its literal meaning in the dataset, it was classified among the speech verbs.

Table 3. TAM distribution of OCS delimitative *po* verbs and their Greek correspondences; Greek perfect and future are counted as non-aspectual

	Count	Per cent	Greek predominant aspect	Greek tense/mood
Aorist	24	17.5	perfective 92%	aorist 92%, perfect 8%
Imperative	10	7.3	perfective 80%	imperative 100%
Imperfect	17	12.4	imperfective 88%	imperfect 88%, aorist 12%
Infinitive	23	16.8	perfective 83%	infinitive 91%, subjunctive 9%
Past active participle	5	3.6	perfective 100%	participle 100%
Present	37	27.0	perfective 35%	future 35%, present 30%, subjunctive 27%, infinitive 8%
Present active participle	10	7.3	imperfective 100%	participle 100%
Supine	11	8.0	perfective 100%	infinitive 100%

However, the most interesting observation to be made is that the *po* delimitatives are most frequently found in the present tense. Only 30% of these correspond to Greek present-tense forms, and again we observe that these examples are occurrences that belong in the group of secondary derived *po* verbs – in fact they are all occurrences of *pomyšljati* "think".

The majority of the present-tense forms correspond either to the Greek future tense (Examples (1) and (2)) or the Greek subjunctive (4). The occurrences that correspond to the Greek future tense look very much like the perfective future in modern Slavic languages. The occurrences that correspond to Greek aorist subjunctives are mostly found in purpose clauses with *da* "so that" and *donьdeže* "until", where the verb also has a future reference, the wished-for outcome is in the future. Thus, these forms are also similar to modern perfective futures both in form and usage contexts.

(4) Then they brought him children.
da *rǫcě* *vъzložitъ* *na* *nę* *i* *pomolitъ* *sę*
that hand.ACC.DU *vъz*.lay.*i*.PRS on they.ACC and *po*.pray.*i*.PRS REFL
hina *tas* *kheiras* *epithēi* *autois* *kai* *proseuxētai*
that the hands.ACC put-upon.AOR.SUBJ they.DAT and pray.AOR.SUBJ
"(Then children were brought to him) that he might lay his hands on them and pray." (OCS, Mar. Matt 19.13)

There are also a few temporal clauses with *egda* "whenever", but in all of these cases the chosen verb is *ponositi* "rebuke, insult", which formally looks like a derived imperfective verb, and suits the iterative nature of the context.

All in all, while the *po* delimitatives do occur in the perfective inflectional forms aorist and past participle, they seem to be particularly useful in cases when Greek uses a perfective form or a future and OCS has no corresponding inflectional form to cover that particular meaning.

Let us now turn to delimitative atelic aorists in OCS. To find these, I selected all simplex verb lemmas that did not have a clear preference for a particular inflectional aspect, as well as verbs with a preference for the imperfective inflectional aspect but which also had aorist/past participle occurrences. I found that 27 simplex verb lemmas (145 tokens) occur in the aorist or as past participles with a delimitative meaning (for the full dataset, see https://doi.org/10.18710/PUXWXL; cf. Eckhoff & Haug 2015). In Example (5), we have a description of a situation that no longer obtains: Jesus lay in his grave for a certain period of time, but now he is no longer there. The aorist adds a temporal boundary to the simplex positional verb *ležati* "lie".

(5) *vidita město. ideže leža xъ*
see place.ACC where lie.*a*.AOR Christ.NOM
deute idete ton topon hopou ekeito
come see.AOR.IMP the place.ACC where lie.IMPERF
"Come, see the place where He lay." (OCS, Mar. Matt. 28.6)

The 27 simplex verbs can be seen in Table 2. The majority of the verbs are plain activities (such as *dělati* "do", *učiti* "teach", *plakati* "cry"), but there are also a number of stative verbs, including several positional verbs (*žiti* "live", *iměti* "have", *ležati* "lie", *mošti* "be able to", *sěděti* "sit"). There are also psych verbs (*věděti* "know", *věrovati* "believe") and a single speech verb (*moliti* "pray"). All in all, we see that the verbs belong to the same verb classes as the *po* delimitatives, but there are only two overlaps: *žiti* "live" and *moliti* "pray". This is illustrated in Examples (6) and (7). In (6) we again see a present-tense *po* verb translating a Greek future-tense form. A reasonable interpretation might be "Man will not live out his lifespan on bread alone". In (7), on the other hand, we see a past participle of the simplex *žiti* "live" occur with an explicit temporal delimitation, namely "seven years".

(6) *ne o xlěbě edinomь poživetъ čkъ*
not by bread.LOC alone.LOC *po*.live.PRS man.NOM
ouk ep' artōi monōi zēsetai ho anthrōpos
not on bread.DAT alone.DAT live.FUT the man.NOM
"man does not live by bread alone" (OCS, Zogr. Matt. 4.4)

(7) She was advanced in years.

živъši	*sъ*	*mǫžemъ*	*ž*	*lěta*	*otъ*	*děvъstva*	*svoego*
live.PSTP	with	husband.INST	7	years.ACC	from	virginity.GEN	her
zēsasa	*meta*	*andros*	*etē*	*hepta*	*apo*	*tēs*	
live.AORP	with	husband.GEN	years.ACC	7	from	the	
parthenias	*autēs*						
virginity.GEN	she.GEN						

"having lived with her husband seven years from when she was a virgin"
(OCS, Mar. Luke 2.36)

All in all, we see that both *po* verbs and delimitative aorists/past participles appear to be regular ways of expressing delimitativity in OCS. The inflectional way of expressing temporal boundaries on atelic verbs is thus limited to the past-tense and participle system, while the *po* verbs are available in all TAM forms. We see that the *po* verbs are particularly frequent in the present tense, rendering Greek futures and aorist subjunctives. While the delimitative *po* verbs and the verbs occurring as delimitative aorists belong to the same verb classes, the distribution is a bit different, and there are only two directly overlapping verbs. We may perhaps speculate that there was a division of labour between *po* and the aorist, both in the range of verbs and in the choice of TAM forms.

5.2 Old East Slavic

The Old East Slavic dataset is fairly similar to the OCS dataset in several respects. There are both *po* delimitatives and delimitative aorists, so the two are still in competition. The share of delimitatives among the *po* verbs is not significantly larger than what we found in the OCS dataset: 22.4% (50) of the *po* verbs are primarily delimitative (Figure 3), and there are 56 *po* verbs that have at least one delimitative occurrence.

As we see in Table 4, the distribution of delimitative *po* verbs is not what we expect from Sigalov (1975)'s description: most of these verbs are dynamic, and the bulk of verbs are normal-intensity activities, such as *povoevati* "wage war", *potruditi* "work", *poslužiti* "serve" (8).

(8) The sinner observes the righteous man.

i	*poskregčetъ*	*na*	*nь*	*zuby*	*svoimi*
and	*po*.grind.PRS	on	he.ACC	teeth.INST	his.INST

"and grinds his teeth at him" (Old East Slavic, PVL 241.34)

Table 4. Verb classes of delimitative *po* verbs and verbs occuring as delimitative aorists in the Old East Slavic dataset

	***Po* delimitatives**	**Verbs occurring as delimitative aorists**
1. States (including positional verbs, but excluding psych verbs)	***požiti*** "live", ***poležati*** "lie", *posvětiti* "shine", ***postojati*** "stand", *postradati* "suffer", ***posěděti*** "sit"	*alъkati* "hunger", *bolěti* "be ill", *vladěti* "rule", ***žiti*** "live", *iměti* "have", *kъnjažiti* "reign", ***ležati*** "lie", ***stojati*** "stand", ***sěděti*** "sit", *cěsarьstvovati* "reign"
2. Plain activities	*pobljusti* "take care of", ***povoevati*** "wage war", *podvignuti* "move", ***požьdati*** "wait", *pozorovati* "watch, guard", *pomuditi* "wait", *pomьdliti* "wait, slow down", *pooxritati* "sneer at", ***poplakati*** "cry", *poprijati* "be friendly", *poskrъgъtati* "grind (teeth)", *poslužiti* "serve", *posměxati sja* "laugh at, ridicule", *posmějati sja* "laugh", *posъpati* "sleep", ***potruditi*** "work", ***potrjasti sja*** "shake", *potъrpěti* "endure", *potjagnuti* "pull through", *poučiti* "teach, instruct", *počitati* "read", *počrěti* "scoop, draw", *poščupati* "feel for, pinch", ***pojasti*** "eat"	*biti sja* "fight", ***vojevati*** "wage war", *dьržati* "hold", *taiti* "hide", ***žьdati*** "wait", *iskati* "seek", *kopati* "dig", *metati* "throw", *mučiti* "torture", ***plakati*** "cry", *prazdьnovati* "celebrate", *rabotati* "serve", *tvoriti* "do", ***truditi*** "work", ***trjasti*** "shake", ***jasti*** "eat"
3. Speech verbs	*požalovati* "complain, express pity", ***pomoliti*** "pray", *ponositi* "reproach", *porugati* "scold", *poxvaliti* "praise", *poxuliti* "condemn, deplore"	*besědovati* "talk", *zъvati* "call", ***moliti*** "pray", *proročьstvovati* "prophesy"
4. Sound emission verbs	*pogrъměti* "thunder", *potrъtati* "thunder"	*pěti* "sing"
5. Psych verbs	*podivovati sja* "wonder, admire", *pokajati sja* "regret", *pomysliti* "think", *popeči sja* "care about", *poskъrběti* "grieve", *postyditi sja* "be ashamed of"	*věděti* "know", *dumati* "think", *mošči* "be able to", *mьněti* "believe", *xotěti* "want", *čuditi sja* "wonder"
6. Perception verbs	***pozьrěti*** "look at", *poslušati* "listen to, obey"	***zьrěti*** "look at", *sъmotriti* "look at"
7. Motion verbs	***poxoditi*** "walk", *poězditi* "travel"	*broditi* "wander, wade", *vlačiti* "pull", *goniti* "chase", *gъnati* "chase", *letěti* "fly", *ristati* "run", ***xoditi*** "walk", *šьstvovati* "walk, wander", *ěxati* "drive, ride"
8. Secondary derived verbs	*pobarati* "fight", *pokrapljati* "sprinkle", *pomavati* "wave", *pomyšljati* "think", *posypati* "sprinkle", *poučati* "teach, instruct", *pouščati* "incite, encourage", *počrěpati* "scoop, draw"	–

No more than five stative verbs (including positional verbs but excluding psych verbs) were found in the dataset. As in the OCS dataset, there are speech, psych and motion verbs, but there are also sound emission and perception verbs. As in the OCS dataset, we also find suffix-derived *po* verbs which appear to be both delimitative and iterative (9).

(9) *mnogašьdy že i prozvuterъ mltvu tvoritь i vodoju*
frequently PTC even priest.NOM prayer.ACC do.*i*.PRS and water.INST
stoju pokrapljaja
holy.INST *po*.sprinkle.PRSP
"Many times the priest had already prayed and sprinkled holy water (but all in vain)" (Old East Slavic, Life of Feodosij Pečerskij, folio 54b)

The findings in the Old East Slavic dataset contradict the chronology proposed by Sigalov (1975): he claims that *po* delimitatives spread to indeterminate motion verbs and psychological processes in the 16th–17th century and to speech verbs, sound emission verbs and physical processes in the 17th–18th century. However, all of these classes are already well represented in the Old East Slavic dataset.

When we compare with the OCS dataset, we see that the TAM distribution of *po* delimitatives is different. As we see in Table 5, there is no preference for present tense anymore: the most common tense is now the aorist (29.7%), followed by past participles (17.1%), i.e., forms that are presumably also inflectionally perfective. As in OCS, Old East Slavic aorists have the potential for a delimitative (and ingressive) interpretation with atelic verbs (cf. Bermel 1995: 340–341). I argue that the same holds for past active participles (but I have found no examples of past PASSIVE participles with this reading in my dataset, and so I make no claims about them).

To find these delimitative aorists and participles, Old East Slavic data can be sorted according to inflectional aspect in the same way as OCS. The tendency for overtly aspectually marked verbs to stick to the corresponding inflectional aspect is less clear than in the OCS dataset, but the method is still useful to find simplex verbs

Table 5. TAM distribution of Old East Slavic delimitative *po* verbs

TAM form	Count	Per cent
Aorist	80	29.7
Imperative	32	11.9
Imperfect	16	5.9
Infinitive	24	8.9
Past participle	46	17.1
Present	43	16.0
Present participle	16	5.9
***l*-form**	11	4.1

(i.e., verbs without prefixes and/or aspectual derivation suffixes) that are neutral or predominantly imperfective. In the Old East Slavic dataset, 47 neutral simplex verbs occur in the aorist or as a past participle with a delimitative reading (180 occurrences).[16]

The delimitative aorists and past participles are thus fairly common in the Old East Slavic dataset and often occur in contexts that make the temporal boundary on the verb explicit, such as (10) and (11).

(10) *trudixom sja i ne moguče sja dokopati*
work.*i*.AOR REFL and not being_able REFL *do*.digging
"We worked and couldn't finish digging" (Old East Slavic, PVL 210.6–7)

(11) *ždaša za měsjacь*
wait.*a*.AOR for month
"they waited for a month (but he didn't pay them)"
(Old East Slavic, PVL 79.1)

When we look at the simplex verbs, we find all the same verb classes as with the Old East Slavic *po* delimitatives (Table 4).[17] There are statives such as *bolěti* "be ill", *žiti* "live", *iměti* "have"; activities such as *iskati* "seek", *kopati* "dig", *plakati* "cry"; speech, psych and perception verbs; and also a large number of motion verbs.

Recall that, in the OCS dataset the simplex aorist verbs had almost no overlap with the *po* delimitatives. In the Old East Slavic dataset, however, there are 13 overlapping verbs in multiple verb classes (boldfaced in Table 4), such as *žiti* "live", *stojati* "stand", *voevati* "wage war", *plakati* "cry", *moliti* "pray". Recall also that, while *po* delimitatives had a preference for occurring in the present tense in the OCS dataset, they predominantly occur in the aorist in the Old East Slavic dataset. This means that, in the Old East Slavic dataset, we observe a number of 'minimal pairs', that is aorists with delimitative readings and the same base verb, differentiated only by the presence/absence of the *po* prefix. Examples (12) and (13) are very similar and describe temporally delimited waging-war events. However, only Example (12) has the *po* prefix.

(12) *i povoeva okolo kyeva*
and *po*.wage_war.*ova*.AOR near Kiev
"he waged war near Kiev" (Old East Slavic, PVL 57.13)

16. The full classified dataset can be found at https://doi.org/10.18710/PUXWXL.

17. Suffix-derived verbs were excluded by definition, though they do sometimes occur as aorists and past participles with delimitative-iterative readings.

(13) *voevaša* *polovci* *okolo* *zarěčьska*
wage_war.*ova*.AOR Polovecians near Zarečsk
"The Polovecians waged war near Zarečsk" (Old East Slavic, PVL 281.4)

5.3 Middle Russian

The use of the delimitative aorist is nearly non-existent in the Middle Russian sources, and the very few potential examples should probably be treated as archaisms. In this section I therefore only look at the semantic classes of *po* verbs in the two Middle Russian datasets: early (15th century) and late (16th–17th century).

There is no significant increase either in type or token frequency for delimitative *po* verbs in the early Middle Russian dataset. A breakdown of the 36 attested delimitative *po* verbs (28 if we exclude the secondary derived ones) into verb classes (Table 6) shows us that the distribution across verb classes is also very similar to that found in the Old East Slavic dataset. However, we see 20 verbs that were not found in the (considerably larger) Old East Slavic dataset, suggesting that new verbs may nonetheless have joined the *po* verb pattern.

The late Middle Russian dataset is the first dataset where we see a significant increase in type and token frequency of delimitative *po* verbs, as shown in Figures 2 and 3. In the late Middle Russian dataset, there are 71 *po* verbs with at least one delimitative occurrence. 30.4% of the *po* verbs are primarily delimitative.

When we look at the verb classes of the delimitative *po* verbs, we see that they are still the same as in Old East Slavic (Table 7). There are considerable overlaps in the attested verbs in all classes – all classes contain verbs that were attested with delimitative meanings in the Old East Slavic dataset as well (boldfaced) and sometimes also in the OCS dataset (marked with an asterisk). There are also 13 overlaps with the delimitative simplex aorists from the Old East Slavic (marked with † in Table 7), as in Example (14). Nonetheless, there are many clearly delimitative *po*-verbs that were not attested in any of the previous datasets, such as *pomъlčati* "be quiet" (15).

(14) *ašte* *kto* *potrudit* *sę* *v* *semъ* *věce* *crstva*
if someone *po*.work.*i*.PRS REFL in this age kingdom.GEN
radi *nbsnago*
for heavenly.GEN
"… if someone works in this age for the sake of the Kingdom of Heaven"
(Late Middle Russian, Domostroj)

(15) *pomolčalъ* *malenko*
po.be_quiet.PST little
"I was quiet for a while" (Late Middle Russian, Life of Avvakum)

Table 6. Verb classes of delimitative *po* verbs in early Middle Russian (delimitative *po* verbs also found in the Old East Slavic dataset are boldfaced, and the ones that were also found in the OCS dataset are marked with an asterisk; verbs with a base verb that occurred as a delimitative aorist/past participle in the OES dataset are marked with a †)

1. States (including positional verbs)	*pobolěti*† "be ill", *pobyti* "be", ***požiti****†† "live", ***postradati**** "suffer"
2. Plain activities	***požъdati**** "wait", *poiskati** "seek", *pokolěbati sja* "waver", *pokolěbiti* "rock, shake", *porabotati* "work", ***poslužiti**** "serve", *potъlkati* "knock", ***potъrpěti**** "endure", *potjanuti* "pull", ***potrjasti sja****† "shake", ***poučiti*** "teach, instruct", ***počitati*** "read"
3. Speech verbs	*pobešědovati*† "talk", ***pomoliti****† "pray", ***porugati**** "scold", ***poxvaliti**** "praise", *pošьpъtati* "whisper"
4. Sound emission verbs	*postonati* "moan"
5. Psych verbs	***pokajati sja*** "regret", *porasuditi* "consider"
6. Perception verbs	***poslušati*** "listen to, obey", *posmotrěti*† "look"
7. Motion verbs	*poiti* "walk, go ahead", *ponesti* "carry, bear"
8. Secondary derived verbs	*pobivati* "beat", *pokazovati* "display", *pominati* "remember", ***pomyšljati**** "think", *popolaskyvati* "rinse", ***posypati*** "sprinkle", ***poučati*** "teach, instruct", *poxvaljati* "praise"

Table 7. Verb classes of delimitative *po* verbs in late Middle Russian. Delimitative *po* verbs also found in the Old East Slavic dataset are boldfaced, the ones that were also found in the OCS dataset are marked with an asterisk. Verbs with a base verb that occurred as a delimitative aorist/past participle in the OES dataset are marked with a †[18]

1. States (including positional verbs)	***požiti****†† "live", ***poležati***† "lie", *pomъlčati* "be quiet", ***postojati***† "stand", ***postradati**** "suffer", ***posěděti***† "sit"
2. Plain activities	***pobljusti*** "take care of", *pogladiti* "stroke", *podvignuti* "move", *podьržati*† "hold, retain", *požati* "squeeze", *pozvoniti* "ring", *pokaditi** "burn incense", *pokolotiti* "knock", *pokropiti* "sprinkle", *ponakazati* "teach", *popaxati* "plough", *popoloskati* "rinse", ***poslužiti**** "serve", *postegati* "whip", *potъlkati sja* "knock", ***potruditi***† "work", ***potьrpěti**** "endure", *potjanuti* "pull", ***potrjasti sja****† "shake", ***poučiti*** "teach, instruct", ***počitati*** "read", ***poščupati*** "feel for, pinch", ***pojasti***† "eat"

18. Table 7 only contains 70 verbs; the final one is the form *postaja* (Life of Avvakum folio 70v), which has tentatively been lemmatised as *postati* but is unclear. The delimitative semantics seem clear enough, though.

Table 7. (*continued*)

3. Speech verbs	*pobesědovati*† "talk", *poblagodariti* "thank", *pobraniti* "scold", *pogovoriti* "talk", ***ožalovati*** "forgive, endow", *pomolitvovati* "pray", ***pomoliti****† "pray", *poricati* "reproach", ***porugati**** "scold", *poslušьstvovati* "testify", ***poxvaliti**** "praise", ***poxuliti*** "condemn, deplore"
4. Sound emission verbs	*postonati* "moan"
5. Psych verbs	*podumati*† "think", *požaléti* "pity", ***pokajati sja*** "regret", ***pomysliti**** "think", ***popeči sja*** "care about", *poraděti* "care"
6. Perception verbs	*pogljaděti* "look", ***poslušati*** "listen to, obey", *posmotrěti*† "look"
7. Motion verbs	*popoiti* "go away for a bit", ***poxoditi***† "walk"
8. Secondary derived verbs	*pobivati* "beat", *pobirati* "gather", *pobyvati* "be", *pověvati* "wave", *pogljadyvati* "look", *pogovarivati* "talk", *pomanivati* "coax, entice", ***pomyšljati**** "think", *ponašati* "scorn, ridicule", *poskakyvati* "jump, gallop", *poslušivati* "listen", *posmatrivati* "look", ***posypati*** "sprinkle", *potьčivati* "honour, serve", ***poučati*** "teach, instruct", *poxvaljati* "praise", *poxlěbati* "sip", *poxuljati* "complain"

Thus there is no evidence that any new verb classes have joined the *po* verb pattern, although it is likely that a number of individual verbs have, given the significant increase in the type frequency of delimitative *po* verbs. At least some of these (base) verbs have delimitative aorist/past participle attestations in the Old East Slavic dataset, but even more of them occurred in attested variation as early as in Old East Slavic.

6. Delimitative contexts in Old East Slavic

So far this paper has exclusively used lemmatisation, morphological annotation, semantic annotation and sub-word-level annotation of derivational morphology, i.e., annotation levels that are not reserved for treebanks. We can, however, use treebank data directly to explore two closely related questions: Were delimitative aorists and *po* delimitatives really synonymous in Old East Slavic? And were there any competing means of expressing delimitativity? We can operationalise both of these questions by looking at the types of temporal adverbials that come with the two types of delimitatives, since they make explicit the temporal boundary placed on the event.

There were 180 occurrences of aorists or past participles classified as delimitative in the Old East Slavic dataset. 115 of them had an adverbial modifier of some sort. By far the largest group of temporal adverbials (41 examples) were temporal accusatives indicating the (long) duration of the event, as in (16).

(16) *i leža noščь tu*
and lie.*a*.AOR night.ACC there
"and [the corpse] lay there one night"
(Old East Slavic, Suzdal Chronicle, year 6655/1147)

Another sizeable group of temporal adverbials are *vъ*+ACC constructions (11 examples), which generally serve as framesetters and do not delimit the event directly but typically also go with fairly long durations, as in (17).

(17) *V seže vremja voeva kurja s polovci*
in same.ACC time.ACC wage_war.*ova*.AOR Kurya with Polovecians.INST
u perejaslavlja
by Perejaslavl'.GEN
"In the same period Kurya and the Polovecians waged war near Perejaslavl'"
(Old East Slavic, PVL 231.4–5)

There are also five examples expressing long duration with *mъnogo* "much, a lot", but only a single example explicitly expressing *short* duration with *malo* "a little" (18).

(18) *i bivъše sja malo negde. staša novgorodьci*
and fight.PSTP REFL little somewhere stand.AOR Novgorodians.NOM
na ostrově.
on island.LOC
"having fought a little at one point, the men of Novgorod took stand on an island" (Old East Slavic, First Novgorod Chronicle, year 6655/1149)

Thus, it seems that the delimitative aorist/past participle was primarily used for fairly long durations, though short durations were also possible.

There are 253 occurrences of *po* delimitatives in the Old East Slavic dataset, and 105 of them have an adverbial of some type. We do find the same types of temporal adverbials as with delimitative aorists/past participles, but the distribution is quite different. Only four examples have temporal accusatives (19).

(19) *požive že vъ ejuptě. lět .zi.*
po.live.AOR PTC in Egypt.LOC year.GEN.PL 17
"and he lived in Egypt for 17 years" (Old East Slavic, PVL 93.27)

There are also few temporal *vъ*+ACC constructions: again, only four examples were found in the Old East Slavic dataset. The most frequent type of temporal adverbial turns out to be ones expressing SHORT duration: there are 16 examples with the adverbial *malo* "a little, for a short while" (20).

(20) *posědi* *malo*
po.sit-*ě*.IMP2SG little
"Sit for a little while!" (Old East Slavic, PVL 266.2)

Thus, even though both the delimitative aorist/past participle and the delimitative *po* verbs can denote both short and long durations, the distribution of overt duration adverbials suggests that they have different preferences: long durations for the aorists and short durations for the *po* verbs.

Having identified the temporal accusative and short-duration adverbials such as *malo* as typical delimitative contexts, we can also use them to look for other types of verbs occurring in the same contexts. Do we find other verbs than atelic simplices and *po* verbs in the Old East Slavic dataset?

Short-duration adverbials of the *malo* type only very rarely occur with verbs that are neither *po* verbs nor atelic simplices. Only three examples were found, all three prefixed with either *pere* or its Church Slavonic counterpart *prě*, as in (21).

(21) He promised to go, but did not go.
i *perestrjapъ* *malo* *poslušavъ* *žiroslava*
and *pere*.linger.PSTP little *po*.hear.*a*.PSTP Žiroslav.GEN
rekušča *jemu*
saying.GEN he.DAT
"and having lingered a little, he heard Žiroslav saying to him"
(Old East Slavic, Suzdal' Chronicle, year 6656/1148)

With the temporal accusative we find more examples of non-derived verbs with prefixes other than *po* (limited to contexts where the temporal accusative expresses true duration, omitting examples such as *večerъ* "in the evening"). There are 16 examples of *prě* verbs, most of them occurrences of *prěbyti* "stay (for some specified time)" as in (22), and one of *perestrjati*, which we saw in Example (21). The prefix thus appears to be in some use to form perdurative verbs.

(22) *i* *tako* *prebys* *vse* *lěto* *do* *zimy*
and thus *prě*.be.AOR all.ACC summer.ACC until winter.GEN
"and thus he remained all summer until winter"
(Old East Slavic, Suzdal' Chronicle, year 6635/1127)

More intriguingly, the Old East Slavic dataset also contains two examples of temporal accusatives with the aorist of the verb *sъtvoriti* "do, perform", derived from the simplex *tvoriti* with the prefix *sъ*. This prefix otherwise produces telic and apparently perfective verbs. In Examples (23) and (24), however, it seems to take on a meaning close to that of the delimitative aorist.

(23) *vsju noščь moltvu stvoriša*
all.ACC night.ACC prayer.ACC *sъ*.do.*i*.AOR3PL
"they prayed all night" (Old East Slavic, PVL 21.20)

(24) *i togo stvori lět z na svět ne vylazja*
and that.GEN *sъ*.do.*i*.AOR3SG year.GEN.PL 7 on light.ACC not *vy*.climb.*i*.PRSP
"and this he did for seven years without coming out in the daylight"
(Old East Slavic, PVL 192.12–13)

These examples suggest that *po* may have had some competition from *sъ*, which was also a very productive and semantically general prefix.

7. Conclusions

This paper uses new enriched treebank data to revisit a central issue in the development of the modern Russian aspect system: the rise of the *po* delimitative. The paper exploits the fact that the PROIEL and TOROT treebanks are enriched with sub-word-level tagging for derivational morphology, and supplements the rich morphological annotation with semantic tagging but also uses the syntactic tagging to look for delimitative contexts.

There is consensus in the literature that the rise of the *po* delimitative was an important step in generalising the aspect partner system in the history of Russian. When the *po* delimitative gained in productivity, a lot of atelic verbs could suddenly have perfective partners. The results of this paper support the conclusion that this surge in productivity came relatively late. However, the treebank data do not support the common assumption that the *po* delimitatives were wholly marginal and severely semantically restricted until a very late stage: even in OCS and Old East Slavic, around 20% of all *po* verb lemmas had delimitative readings, and these verbs were not restricted to statives but belonged to all verb classes found with late Middle Russian *po* delimitatives.

Previous research also fails to recognise the main competition of the *po* delimitatives in early Slavic: the delimitative usage of aorists and past participles with atelic simplex verbs, arguably the last independent function of the aorist, found both in OCS and Old East Slavic. Given earlier accounts, the chronology did not seem to match up. However, the treebank data support the hypothesis that the *po* delimitatives gradually took over this function from the inflectional aspect system. In the earliest sources we see a situation where verbs from largely the same verb

classes may either occur as delimitative aorists/past participles or form *po* delimitatives: both constructions could convert states and activities into delimitatives. Some verbs are attested in both constructions. There are more of these overlaps in the Old East Slavic dataset than in the OCS one. The TAM distribution of the *po* verbs is also different: in the OCS dataset we see that they predominantly occur in the present tense, while in the Old East Slavic dataset they predominantly occur as aorists, yielding a larger number of 'minimal pairs'. This suggests a development from a division of labour to a situation of free variation. In Middle Russian the aorist is lost, and the *po* delimitatives are left to do the job on their own. Looking at the verb classes of the *po* verbs in the late Middle Russian dataset, we see that they are the same as the ones we found in OCS and Old East Slavic datasets and that many of the verbs are also the same, or correspond to simplex verbs that could formerly occur as delimitative aorists or past participles.

It thus seems fair to say that the development of (productive) Russian *po* delimitatives was boosted by the loss of the delimitative aorist with simplex verbs, since, as outlined in § 6, they coexisted for a considerable while with much the same functions, making the *po* verbs the natural heir of the aorist in this respect. In this sense it may be argued that the exotic Russian *po* delimitatives grew directly out of the old Indo-European aspect system.

References

Amse-De Jong, Tine H. 1974. *The meaning of the finite verb forms in the Old Church Slavonic Codex Suprasliensis: A synchronic study*. The Hague: Mouton.

Bermel, Neil. 1995. Aspect and the shape of action in Old Russian author(s). *Russian Linguistics* 19(3). 333–348. https://doi.org/10.1007/BF01080603

Bermel, Neil. 1997. *Context and the lexicon in the development of Russian aspect* (University of California Publications in Linguistics 129). Berkeley: University of California Press.

Dickey, Stephen M. 2007. A prototype account of the development of delimitative *po-* in Russian. In Dagmar Divjak & Agata Kochanska (eds.), *Cognitive paths into the Slavic domain*, 326–371. Berlin: Mouton De Gruyter.

Dickey, Stephen. 2012. Orphan prefixes and the grammaticalization of aspect in South Slavic. *Jezikoslovlje* 13(1). 71–105.

Dmitrieva, Ol'ga. 1991. Formirovanie semantičeskoj struktury russkogo glagol'nogo prefiksa *po-*. In Lidija I. Barannikova (ed.), *Aktivnye processy v jazyke i reči*, 68–74. Saratov: Izdatel'stvo Saratovskogo universiteta.

Dmitrieva, Ol'ga. 2000. Formirovanie sistemy russkix delimitativnyx glagolov. In T. V. Kočetkova (ed.), *Predloženie i slovo: paradigmatičeskij, tekstovyj i kommunikativnyj aspekty*, 28–33. Saratov: Izdatel'stvo Saratovskogo pedagogičeskogo instituta.

Dostál, Antonin. 1954. *Studie o vidovém systému v staroslověnštině*. Prague: Státní pedagogické nakladatelství.

Dowty, David R. 1979. *Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*. Dordrecht: Reidel. https://doi.org/10.1007/978-94-009-9473-7

Eckhoff, Hanne Martine & Aleksandrs Berdičevskis. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS treebank. *Scripta & e-Scripta* 14–15. 9–25.

Eckhoff, Hanne Martine & Dag Trygve Truslew Haug. 2015. Aspect and prefixation in Old Church Slavonic. *Diachronica* 32:2. 186–230. https://doi.org/10.1075/dia.32.2.02eck

Eckhoff, Hanne, Kristin Bech, Kristine Eide, Gerlof Bouma, Dag Trygve Truslew Haug, Odd Einar Haugen & Marius Jøhndal. 2018. The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52(1). 29–65. https://doi.org/10.1007/s10579-017-9388-5

Forsyth, James. 1972. The nature and development of the aspectual opposition in the Russian verb. *The Slavonic and East European Review* 50(121). 493–506.

Haug, Dag Trygve Truslew & Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In Caroline Sporleder & Kiril Ribarov (eds.), *Proceedings of the language technology for cultural heritage data workshop (LaTeCh 2008)*, 27–34. Marrakech, Morocco. www.lrec-conf.org/proceedings/lrec2008/index.html (last accessed 10 July 2018.)

Janda, Laura A., Anna Endresen, Julia Kuznetsova, Olga Lyashevskaya, Anastasia Makarova, Anastasia, Tore Nesset & Svetlana Sokolova. 2013. *Why Russian aspectual prefixes aren't empty: Prefixes as verb classifiers*. Bloomington: Slavica.

Lakoff, George. 1966. *Stative adjectives and verbs in English*. (Report NSF-17). Cambridge, MA: Harvard Computation Lab.

Meillet, Antoine. 1934. *Le slave commun*. Paris: Champion.

Mišina, Ekaterina A. 2017. K izučeniju perfektivnogo imperfekta v drevnerusskom jazyke (v sopostavlenii so staroslavjanskim). *Russian Linguistics* 41(1). 1–15. https://doi.org/10.1007/s11185-016-9173-x

Němec, Igor. 1954. O slovanské předponě *po-* slovesné. *Slavia* 23. 1–22.

Růžička, Rudolf. 1957. *Der Verbalaspekt in der altrussischen Nestorchronik* (Veröffentlichungen des Instituts für Slawistik 14). Berlin: Akademie-Verlag.

Schooneveld, Cornelis H. van. 1951. The aspect system of the Old Church Slavonic and Old Russian verbum finitum byti. *Word* 7. 96–103. https://doi.org/10.1080/00437956.1951.11659396

Sigalov, Pavel S. 1975. Istorija russkix ograničiteľnyx glagolov. *Trudy po russkoj i slavjanskoj filologii: Serija lingvističeskaja* 24. 141–181.

Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66(2). 143–160. https://doi.org/10.2307/2182371

Živov, Viktor M. 2017. *Istorija jazyka russkoj pis'mennosti*, vol. 1. Moscow: Universitet Dmitrija Požarskogo.

Appendix

The datasets for this study were extracted from the TOROT treebank using Ruby scripts accessing the database and webapp methods directly. Unfortunately, we have not yet been able to provide a similar way to draw detailed datasets of this type that is open to the public.

The basic query for all datasets is very simple. For the OCS study (which recycles the dataset from Eckhoff & Haug 2015), it was "find all tokens of verb lemmas marked with the ISO tag chu in reviewed[19] sentences, as well as their Greek token alignments":

```
Lemma.find_all_by_language_tag('chu').select { |l| l.part_of_speech_tag ==
'V-'
}.map(&:tokens).flatten.each do |v|
     if v.sentence and  !v.sentence.reviewed_at.nil?
           gk = v.token_alignment
           STDOUT.puts […].join(',')
     end
end
```

For the diachronic Russian study, it was "find all tokens of verb lemmas marked with the ISO code orv in annotated sentences":

```
Lemma.find_all_by_language_tag('orv').select { |vl| vl.part_of_speech_tag
=='V-'
}.map(&:tokens).flatten.each do |v|
     if v.sentence.annotated_at != nil
     STDOUT.puts […].join(',')
     end
end
```

After that, the verb *byti* was excluded from both datasets, and the OCS dataset was limited to verbs that had an overt Greek aligned verb. All further limitations were done in the R scripts available at https://doi.org/10.18710/PUXWXL, together with all datasets.

The basic queries can be approximated in the TOROT webapp simply by querying (under "Search") for all occurrences of OCS and/or Old Russian verb lemmas (or alternatively query for the occurrences of verb lemmas in each of the sources used in the study, since the treebank has grown since the data were drawn). It is also possible to query for individual prefixes under "Semantic tags". What cannot be replicated in this way, however, is the output asked for in the "STDOUT.puts […].join(',')" statements, which I used to draw large amounts of information on each verb token: treebank-internal identifiers, lemma, morphology and syntax information (for example argument structure data) as well as derivational morphology tags. For a full overview of the datasets, see the README-file at https://doi.org/10.18710/PUXWXL. While it is possible to download query results in csv or txt format in the TOROT webapp, there is no public way to specify and output the large number of features related to each token found in the datasets – only the token and its left and right context, as well as citation information and the treebank-internal sentence id and token id, will be provided.

---

19. A reviewed sentence has been annotated by one annotator and checked (and if necessary corrected) by another, senior annotator.

All reviewed texts in TOROT are released in xml format at http://torottreebank.github.io/ (the Codex Marianus is released by PROIEL at https://proiel.github.io/). The xml files contain lemmatisation, morphological and syntactic information, as well as information status tags and parallel text token alignment ids where available. They do not contain so-called semantic tags, the customisable tag layer where the derivational tag information is stored. Thus, it is possible to extract a lot of the information in the datasets from these files using an xml query language, but not all of it.

Given the limited public access to easy data extraction, it is especially important to publish detailed and relatively unlimited datasets of the kind used in this study.

Non-configurationality in diachrony

Correlations in local and global networks of Ancient Greek and Latin

Edoardo Maria Ponti and Silvia Luraghi
University of Cambridge / University of Pavia

Non-configurationality is a linguistic property associated with free word order, discontinuous constituents, including NPs, and null anaphora of referential arguments. Quantitative metrics, based both on local networks (syntactic trees and word order within sentences) and on global networks (incorporating the relations within a whole treebank into a shared graph), can reveal correlations among these features. Using treebanks we focus on diachronic varieties of Ancient Greek and Latin, in which non-configurationality tapered off over time, leading to the largely configurational nature of the Romance languages and of Modern Greek. A property of global networks (density of their spectra around zero eigenvalues) measuring the regularity in word order is shown to be strengthened from classical to late varieties. Discontinuous NPs are traced by counting the words creating non-projectivity in dependency trees: these drop dramatically in late varieties. Finally, developments in the use of null referential direct objects are gauged by assessing the percentage of third-person personal pronouns among verb objects. All three features turn out to change over time due to the decay of non-configurationality. Evaluation of the strength of their pairwise correlation shows that null direct objects and discontinuous NPs are deeply intertwined.

Keywords: non-configurationality, treebanks, network analysis, non-projectivity, discontinuous constituents, referential null objects

1. Introduction

In this paper, we show how treebank-based queries and network analysis allow us to measure the development of a number of features of Classical Greek and Latin syntax that are normally considered correlates of non-configurationality, that is, free constituent order, discontinuous NPs and use of null anaphora for definite referential

https://doi.org/10.1075/bct.113.03pon

direct objects. We chose these features as diagnostics for non-configurationality and their decay as a hint to the rise of configurationality based on Baker (2001: 1434), who writes: "[i]n the narrow sense, a nonconfigurational language is one that has … free word order, possible omission of all grammatical functions, and the possibility of having discontinuous NP constituents."

While Classical Greek and Latin displayed these features, both Modern Greek and the Romance languages feature configurational syntax to a large extent. We aim to capture the ongoing rise of configurationality based on two diachronic treebanks of Ancient Greek and of Latin available from the PROIEL project (see §3.1). The analysis is based on quantitative parameters associated with features of non-configurationality, and these allegedly co-vary in time. They are measured both at the local and at the global level (syntactic trees and co-occurrences in single sentences and networks; see §3.2). Our paper is organized as follows. In §2 we discuss the notion of non-configurationality and how it applies to Classical Greek and Latin. In §3 we describe the experiment setup and the data. Section 4 is devoted to the formal definition of individual metrics related to non-configurationality and the assessment of their values. In §5 we present the analysis of these results, and finally we draw some general conclusions in §6.

2. Non-configurationality

The term 'non-configurationality' was introduced in Hale's (1983) study of Warlpiri, in order to account for a number of typical features of this language that make it remarkably different from languages like English. According to Hale (1983), non- configurational languages have a 'flat' structure, or a hierarchical structure at the level of Lexical Structure only, which does not project on Phrase Structure. This observation leads to the conclusion that the VP is not relevant in non-configurational languages, in which, typically, "subjects and objects cannot be identified by word order and simple constituency tests in any straightforward way" (Baker 2001: 1433).

Research on non-configurationality first developed within the Government and Binding framework, but in recent years, as features of non-configurationality have been reported from numerous languages of different genetic and areal affiliations, it has increasingly attracted the interest of typologists (for a survey, see Reinöhl 2016: 23–27, 45–48). As we remarked in §1, typical correlates of non-configurational languages have been shown to be free (i.e., pragmatically determined) word order, discontinuous NPs and extensive null realization of definite referential arguments even when they are not co-referenced on the verb (Austin & Bresnan 1996; Baker 2001).

Current research shows that configurationality should be regarded as a gradient property, as languages may be configurational or non-configurational to different extents. For example, Hungarian has been argued to be non-configurational in clause structure, as it allows free constituent order, but configurational in noun phrase structure, as it does not allow discontinuous NPs (Kiss 1987). Indeed, free constituent order is not necessarily associated with a high degree of non-configurationality: the fact itself that constituency is relevant at the phrasal level implies, for example, that discontinuous NPs are normally not allowed in languages such as Hungarian. We return to this issue in §2.1.

Configurationality can arise as a result of language change and become more extensive over the course of time. Ancient Indo-European languages show typical features of non-configurationality, including among other things free word order, discontinuous NPs and definite referential null objects which are not co-indexed on the verb (Devine & Stephens 2000; Schäufele 1990; Rögnvaldsson 1995; Luraghi 1997, 2003), a weak noun-adjective distinction (see §2.2) and the trend toward increasing configurationality has been described for many of them (Luraghi 2010). As Reinöhl (2016: 45) remarks "Latin and Greek only possess incipient phrasal structures, … (Latin having prepositional phrases and Ancient Greek developing nominal expressions involving articles), … Vedic shows a lack of such structures." According to Hewson & Bubenik (2006), configurationality in Indo-European languages first manifested itself with the increasing grammaticalization of adpositional phrases and the creation of adpositions out of earlier adverbs. Reinöhl (2016), though distancing her views in relevant respects from those of Hewson and Bubenik, also argues that the rise of adpositions brought about configurationality. Ledgeway (2012) shows that the prepositional phrase and, while less developed, the complementizer phrase already existed in Latin. Indeed, adpositional phrases are fully grammaticalized not only in Latin, but also in Classical Greek; for this reason we do not take them into account.

In the following sections, we show how certain correlates of non-configurationality are instantiated in Classical Greek and Latin.

2.1 Word order

Classical Greek and Latin are so-called free word order languages. The position of the verb in the sentence is sensitive to pragmatic factors: it may show author-specific preferences, but it is not restricted from occurring in sentence initial, internal or final position. In particular, concerning the position of the verb in Herodotus' *Histories*, which constitute part of the corpus for this paper, a partial analysis carried out by Dover (1960) yields the following counts for word order patterns: VS (113) vs SV (174); VO (203) vs OV (161), with a preference for post-verbal direct

objects. Other authors analyzed by Dover include Lysias, who shows a preference for pre-verbal direct objects and final verbs, and Plato, who has approximately the same percentage of OV and VO occurrences.

Latin is often referred to as an SOV language, and final verbs do in fact predominate in all authors. However, initial and internal position are also possible options in all literary genres and at all diachronic stages. Caesar, dubbed a 'final position fanatic' (*Fanatiker der Endstellung*) by Linde (1923: 154), has the verb in final position in 84% of main clauses and 93% of subordinate clauses. For Cicero, Linde (1923: 155) found around 50% of final and 50% of non-final verbs in main clauses, with variation among different types of work and a considerably higher proportion of final verbs in subordinate clauses. Similarly, Danckaert (2015: 241) in a survey of various studies of word order in Cicero's works, signals a range of variation from 63.1% to 95.9% in OV sentences (the figures cover both main and subordinate clauses).

Though considered SVO languages, both Modern Greek and the Romance languages allow free constituent order to a varying extent, partly due to extensive use of direct object clitics, which are usually preverbal and can co-index displaced constituents. Changes in clause structure have been observed over the history of these languages (Deligianni 2011; Revithiadou & Spyropoulos 2007, 2008; Salvi 2004; Luraghi 1998, 2010; Ledgeway 2011, 2012). Constraints on the order of constituents necessarily follow the rise of constituency. For this reason, constituent order can be diagnostic for ongoing change from Classical to late varieties of Greek and Latin only if connected with the decay of other correlates of non-configurationality, such as null objects and discontinuous NPs.

2.2 Discontinuous NPs

According to Baker (2001: 1437), discontinuous NPs "are possible only in languages with no more than a weak N[oun]/A[djective] contrast," because syntactically adjectives are predicates of nouns, rather than being dependent. This is similar to the traditional view on Proto Indo-European adjectives, as expressed for example by Meillet & Vendryes (1924: 530): "Adjectives are by no means connected with nouns. They are usually inflected in the same case, same number, and, as distinctive for adjectives, same gender … because they refer to the same entity."[1] Discontinuous constituents in classical varieties of Latin and Greek occur to varying extents depending on literary genres but are well attested in literary prose.

1. L'adjectif n'est nullement lié au substantif. Il est généralement au même cas, au même nombre, et, ce qui est le trait caractéristique de l'adjectif, au même genre …, mais parce qu'il s'applique au même objet.

Examples of discontinuous NPs from Latin are (1) and (2); see further *duas legiones … novas* in (7).

(1) ***aliquo*** *te* *cum* *hoc* *rei publicae* ***vinculo*** *esse coniunctum*
INEF.ABL 2SG.ACC with DEM.ABL state:GEN link:ABL bind:INF.PF.P
"(that) you were bound to him by some responsibility for the state."
(Cic. *Mur*. 64)

(2) *neque quisquam agri modum certum aut* ***fines***
nor INDEF.NOM land:GEN measure:ACC certain:ACC or border:ACC.PL
habet ***proprios***
have:PRS.3SG own:ACC.PL
"Nor has anyone a fixed quantity of land or his own individual limits."
(Caes. *Gal.* 6.22.2)

Classical Greek NPs differ from Latin mainly due to the existence of fully grammaticalized definite articles. This should point toward a higher degree of configurationality. Note however that even NPs with definite articles allow for various types of discontinuity (see Devine & Stephens 2000 for an exhaustive description). In particular, occurrences in which the definite article is separated from the noun it determines, when a constituent is sentence initial, as in (3), should receive separate treatment.

(3) *ho* *dè* *khrusós* *hoûtos* *kaì* *ho* *árguros*
ART.NOM PTC gold:NOM DEM.NOM and ART.NOM silver:NOM
kaléetai *Gugádas*
call:PRS.M/P.3SG Gygian:NOM
"This gold and the silver are called Gygian." (Hdt. 1.14.3)

Notably, items that can stand between the definite article and the noun are so-called postpositives, that is, second position, or P2, particles that may bear a graphic accent but prosodically behave as clitics. We return to this issue in §4.1.

Often, discontinuity is caused by the occurrence of a clitic, as in (4). In this example, the direct object clitic *min* is not only separated from the verb *apopémpseie* "had sent", it also splits up the NP *hoîon ándra* "such man". Notably, *min* is often described as a P2 clitic. However, in Classical Greek especially, pronominal P2 clitics could be placed elsewhere in the sentence, as described in Goldstein (2016), even though they did not show any special preference for a specific type of constituent; see further Luraghi (2013).

(4) *thōmázein* *te* *autoû* *par'* ***hoîón*** *min* ***ándra***
wonder:INF.PRS PTC there by such:ACC 3SG.ACC man:ACC
apopémpseie
send:OPT.AOR.3SG
"(The herald) wondered what sort of man he had been sent to." (Hdt 5.92f3)

The occurrence of discontinuous NPs is a major difference between Classical Greek and Latin on the one hand and Modern Greek and Romance on the other. As the data in §4.1 show, the number of discontinuous NPs drops by 86.2% from Classical to Late Greek and by 89.48% from Classical to Late Latin. One of the few discontinuous NPs in the Late Greek corpus is *héteron doûlon* "another servant" in (5a). Interestingly, the Latin translation in (5b) does not mirror the same discontinuity. Modern Greek in (5c) and Italian in (5d) also contain continuous NPs.

(5) a. *kaì prosétheto* ***héteron*** *pémpsai* ***doulon***
and add:AOR.MID.3SG other:ACC send:INF.AOR servant:ACC
"He sent yet another servant." (Luke 20.11)
b. *Et addidit* ***alterum servum*** *mittere*
c. *Apéstile ke páli* ***énan állon dulo***
d. *Mandò* ***un altro servo***

2.3 Definite referential null objects

Null arguments are common in Latin and in Classical Greek. Both languages make extensive use of null subjects; however, as subjects are extensively co-indexed on finite verbs through a complex morphological system of agreement, the occurrence of null subjects is not indicative of non-configurationality.

Much more significant is the occurrence of null referential direct objects, as direct objects are not co-indexed on the verb. Null referential direct objects occur in different syntactic and discourse conditions, as in Examples (6) and (7) (see Luraghi 1997, 2003; Keidana & Luraghi 2012).

(6) *Epexêlthon* ***hoí*** ***te*** ***epíkouroi*** ***kaì autôn***
march:AOR.3PL ART.NOM.PL PTC mercenary:NOM.PL and DEM.GEN.PL
Samíōn ***sukhnoí***$_i$ *dexàmenoi* *dè* *toùs*
Samian:GEN.PL many:NOM.PL engage:PTCP.AOR.NOM.PL PTC ART.ACC.PL
Lakedaimoníous ep' olígon khrónon épheugon opísō
Spartan:ACC.PL on little:ACC time:acc flee:IMPF.3.PL back
hoì dè epispómenoi ***Ø***$_i$ *ékteinon* ***Ø***$_i$.
ART.NOM.PL PTC pursue:PTCP.AOR.MID.NOM.PL kill:IMPF.3PL
"The mercenaries and many of the Samians themselves sallied out near the upper tower on the ridge of the hill and withstood the Lacedaemonian advance for a little while; then they fled back. The Lacedaemonians pursuing **them** destroyed **them**." (Hdt. 3.54.2)

(7) *Caesar* ***duas*** ***legiones***$_i$ *in citeriore Gallia* ***novas***
Caesar:NOM two:ACC legionACC.PL in hither:ABL Gaul:ABL new:ACC.PL
conscripsit et inita aestate, in interiorem
enroll:PF.3SG and begin:PTCP.PF.ABL summer:ABL in inner:ACC
Galliam qui ***Ø***$_i$ *deduceret, Quintum Pedium*
Gaul:ACC REL.NOM lead:SBJ.IMPF.3SG Quintus:ACC Pedius:ACC
legatum misit
lieutenant:ACC send:PF.3SG
"Caesar enrolled two new legions in Hither Gaul and at the beginning of the summer he sent Quintus Pedius, lieutenant-general, to lead **them** into Inner Gaul." (Caes. G. 2.2)

In particular, in cases in which the same direct object is shared by two coordinated clauses (object sharing), deletion in the second clause seems to be mandatory. In Latin, no exceptions have been found in Classical authors. The only occurrences in which a direct object can be repeated in coordinated clauses appear in contexts in which more than one possible antecedent is available, as in (8), or in which the direct object in the second conjunct is accented for emphasis, as in (9). In the latter case, the emphatic pronoun typically hosts the clitic conjunction *=que*. Occurrences similar to these from other Latin authors are thoroughly discussed in Luraghi (1997).

(8) ***litteras*** *scripsi hora decima Cerialibus*
letter:ACC.PL write:PF.1SG hour:ABL tenth:ABL Cerealia:ABL.PL
statim ut tuas legeram sed ***eas***
immediately as POSS.2PL.ACC.PL read.PPF.1SG but 3SG.ACC.PL
eram daturus ut putaram postridie
be:IMPF.1SG give:PTCP.FUT.NOM as think:PPF.1SG next.day
"I wrote (you) **a letter** at four o'clock in the afternoon of the Cerealia as soon as I received yours [possible conflicting antecedent], and I was thinking of giving it the next day (to the first available person)." (Cic. *Att.* 2.12.4)

(9) *postero autem die Caesar …* ***Vettium*** *in rostra*
next:ABL however day:ABL Caesar:NOM Vettius:ACC in roster:ACC.PL
produxit ***eum=que*** *in eo loco constituit quo*
bring:PF.3SG 3SG.ACC and in DEM.ABL place:ABL place:PF.3SG where
Bibulo consuli adspirare non liceret
Bibulus:DAT consul:DAT hope:INF NEG be.allowed:SBJ.IMPF.3SG
"However, the next day Caesar took Vettius on the rostra and placed him in a position in which Bibulus, though being consul, was not allowed to stand." (Cic. *Att.* 2.24.3)

In the case of direct object sharing, Classical Greek too mostly features pronouns in the second coordinated clause when the context contains more than one possible antecedent (see Luraghi 2003; Keydana & Luraghi 2012). However, in Herodotus one also finds at least one exception, discussed in Keydana & Luraghi (2012: 119), featuring overt realization of a pronominal direct object in the second conjunct and the coordinating conjunction *kaí*. In (10), the direct object is overtly realized in the second conjunct, featuring the adversative particle *allá*. In fact, in this passage the adversative character of the second coordinate clause may have favored repetition of the direct object (note further that the statue is the topic of a long stretch of discourse in the preceding context).

(10) ***tòn*** *dè* ***andriánta*** ***toûton*** *Dḗlioi* *ouk* *apḗgagon*
ART.ACC PTC statue:ACC DEM.ACC Delian:NOM.PL NEG remove:AOR.3PL
allá ***min*** *di' etéōn* *eíkosi* *Thēbaîoi …* *ekomísanto*
but 3SG.ACC for year:GEN.PL twenty Theban:NOM.PL bring:AOR.MID.3PL
epì *Dḗlion*
toward Delion:ACC
"But the Delians never carried **that statue** away; twenty years later the Thebans brought **it** to Delium." (Hdt. 6.118.3)

Referential null objects disappeared with the rise of configurationality in Romance (Luraghi 1998; Ledgeway 2012) and Modern Greek (Revithiadou & Spyropoulos 2007, 2008), which developed a system of pronominal clitics.

3. Methodology

We use a treebank corpus to explore non-configurationality in Latin and Ancient Greek. As mentioned in §1, we observe the relevant features of non-configurationality on both the local and the global levels of linguistic networks, following the distinction proposed by Čech et al. (2011). The local level consists of the syntactic dependency trees and word order of the individual sentences (see §3.1). The global level consists of a single network constructed from a treebank with a technique pioneered by Ferrer-i-Cancho & Solé (2001): each distinct lemma corresponds to a node, whereas each distinct relation between a pair of lemmas corresponds to an edge directed from one lemma to the other. This relation can be either co-occurrence, meaning that a word follows another in the linear order of a sentence, or dependency, meaning that a word is the parent of another in a syntactic tree. Global networks mirror holistic properties of a language, possibly different from the sum of the properties of the local networks (Solé et al. 2010; Baronchelli et al. 2013). Local networks in turn are better suited for identifying fine-grained phenomena.

As a consequence, both levels are necessary to capture non-configurationality, as it affects both a language variety as a whole and some of its specific constructions. In the rest of this section, we describe the data from Latin and Ancient Greek that we used for this paper, and we outline the method used to generate co-occurrence and dependency networks from the data.

3.1 The corpus

The data come from the collection of dependency treebanks developed within the PROIEL project (*Pragmatic resources in old Indo-European languages*; see Haug & Jøhndal 2008). A dependency treebank is a corpus of texts annotated with dependencies at the syntactic layer. Sentences are represented as trees where each word corresponds to a node: top-down relations indicated by edges convey grammatical relations between a head and a dependent. In addition, nodes are arranged on the left-to-right dimension to convey the linear order, i.e., the precedence relations.

PROIEL contains treebanks for several ancient Indo-European languages and for different varieties of the same language. We selected four of these treebanks, Ancient Greek and Latin, both Classical and Late. The amount of tokens was equalized to the count of the smallest treebank (67,247 tokens) approximated to the closest sentence boundary. The four treebanks consist of the texts listed in Table 1 (in parentheses we indicate the actual span of the text used for this work).

Table 1. Texts composing the treebanks

	Classical Greek	**Late Greek**	**Classical Latin**	**Late Latin**
Author	Herodotus	Septuagint	Caesar and Cicero	Jerome
Title	*Histories* (I.1–VII.83)	*New Testament* (Matthew I.1 – Acts of the Apostles V.10)	*The Gallic War* (I.1–VII.77) and *Letters to Atticus* (I.1–VI.9)	*Vulgate* (Genesis I.1 – Acts of the Apostles XIV.11)
Date	440–429 BCE	49–150 ca. CE	58–50 BCE and 68–43 BCE	382–413 CE

We preferred the PROIEL treebanks over the Perseus collection (Bamman et al. 2009) for a series of reasons. In the first place, the two sets of treebanks cannot be merged, as they rely on incompatible annotation schemes. Also, Ancient Greek texts in the Perseus treebank are mostly poetry, and further, the Latin texts are limited in size. Hence we consider the texts in the PROIEL treebank the best approximation available for the relevant language varieties, even though we are aware of the fact that they are not entirely representative, as they are mostly limited to single authors, and more variables than just diachrony separate them, notably social status of the authors and literary genre of the texts.

3.2 Network induction

A network is a graph consisting of a set of nodes *V* and a set of edges *E*. A network can be induced from a treebank by setting up an equivalence between (i) nodes and properties of words (e.g., their form, lemma, part-of-speech tag, etc.) and (ii) edges and word relations (e.g., precedence in linear order, dependency in syntax, etc.). Ferrer-i-Cancho et al. (2004) developed a method to create networks based on lemmas as nodes and dependencies as edges. Networks are useful because they mirror global properties of a language that can be hidden in local structures (the dependency trees). For example, linguistic networks of children's speech show a sudden change from tree-like structures to scale-free, small-world structures (i.e., hierarchical and highly connected) around the age of two years (Solé et al. 2010). This happens simultaneously with the appearance of functional words and inflectional morphology. Furthermore, linguistic networks cast light on the nature of linguistic universals and variation. Čech et al. (2011), for example, demonstrated that verbs behave as hubs (i.e., nodes with many connections) in linguistic networks cross-linguistically, providing evidence in support of Tesnière's predicate-centric theory (Tesnière 1959).

Although recent work on linguistic networks has focused on dependency networks, these completely obscure another range of properties of languages: those affecting the linear order of words. Free word order (including the fluctuation in the verb position) is such a property. Hence, in this work we explore linguistic networks based on co-occurrence and dependencies in order to investigate to what extent they diverge. Thus far, networks based on linear order have exploited collocations, i.e., co-occurrences more frequent than chance (Ferrer-i-Cancho & Solé 2001; Kapustin & Jamsen 2007), as an approximation of dependency-based networks. Our method does not filter out any co-occurrence: each lemma corresponds to a node, whereas an edge is created between every two adjacent lemma instances for co-occurrence networks and between every head-dependent pair for dependency networks. The edges of these graphs are oriented, in order to distinguish right and left (for linear order) or top and bottom (for syntactic dependencies) contexts of the nodes, and loops (edges departing from and arriving to the same node) are forbidden.

Figure 1 displays the dependency tree (left) and the equivalent lemma-based co-occurrence network (center) of Example (11). The latter is constructed by creating a node for each distinct lemma of the words in the tree (e.g., *et* for *et* and *voco* for *vocabis*). Then a directed edge is created between two nodes if one immediately follows the other in the tree (e.g., *et* and *voco*). Repeating this procedure over the whole treebank of Late Latin results in a global network (right). For the sake of visualization, this is shown in such a way that the more an edge tends to the center, the

more frequent it is (note that this choice of display has no effect on the properties of the network). These graphs can be equivalently specified by a binary adjacency matrix *A*. Each row and each column of the matrix corresponds to a separate node. For a node pair *i* and *j*, the cell $A_{i,j}$ is filled with 1 if there exists an edge between them, otherwise it has 0.

(11) *Et* *vocabis* *nomen* *eius* *Iesum*
and call:FUT.2SG name:ACC 3SG.GEN Jesus:ACC
"And you will call him Jesus." (Matthew 1.21)

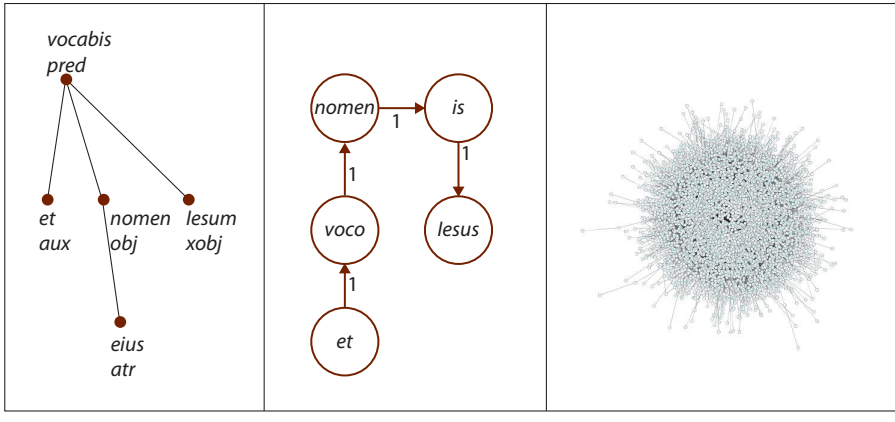


Figure 1. An example of syntactic tree (left), its equivalent network representation (center) and the global network resulting from a similar transformation of the whole treebank

We created both co-occurrence and dependency networks of lemmas for each treebank. Basic information about them is summarized in Table 2, which reports the total number of nodes and edges. Note that the late varieties have lower figures for both, because of inherent properties of the texts. In particular, for the same number of tokens, they have a smaller set of lemmas, which on average have a higher frequency compared to the set of classical varieties.

Table 2. Number of nodes and edges of the induced networks

Name	Nodes (lemmas)	Edges
Classical Greek	5398	34076
Late Greek	3025	20788
Classical Latin	4824	39311
Late Latin	3071	25021

4. Metrics and results

As we argued in §2, non-configurationality has been related to free word order, discontinuous NPs and null anaphora of definite referential direct object. In this section, we propose some metrics to assess quantitatively both the pervasiveness of these phenomena in a language variety and the difference between two diachronic stages. These metrics hinge on properties of global networks or queries constrained by word order and syntactic dependencies in local networks.

4.1 Free word order

The feature of unconstrained word order is problematic for many reasons. In the first place, it surfaces in languages that are not, strictly speaking, non-configurational (Luraghi 2010). In addition, neither a clear formal definition nor any sound method to measure this feature is available. Futtrel et al. (2015) propose a measure of freedom of word order (i.e., argument order with respect to the verb) based on conditional entropy, that is, the uncertainty in determining a word sequence given an unordered dependency tree. They demonstrate that classical varieties of Ancient Greek and Latin are the languages with the highest entropy in terms of word order and head-dependent directionality among the treebanks in the Universal Dependencies collection (Nivre et al. 2016), including Modern Greek and Romance languages. However, this measure turns out to be unreliable if ranging over all the syntactic relations because of the difficulties in estimating entropy statistically and avoiding data sparsity due to the long-tail distribution of linguistic phenomena.

With respect to the position of verbs, we counted whether objects depending on a verb follow it (VO) or precede it (OV) in Table 3, obtaining results in line with the expectations of §2.1. For Ancient Greek, we observe a shift from indifference regarding OV or VO order to a clear preference for VO. In Latin, the preference changes by swapping the order from OV to VO.

Moreover, following Ponti (2016), we propose as an alternative identifying across-the-board word order freedom with the 'irregularity' of a global network.

Table 3. Counts of objects following (VO) or preceding (OV) verbs in the four treebanks

	VO	OV
Classical Greek	3762	3409
Late Greek	4501	2158
Classical Latin	1473	5471
Late Latin	4884	2791

The gist is that, if a lemma is allowed to appear in more contexts, then its neighborhood in the network is more idiosyncratic and does not match the neighborhood of similar words. For instance, if a verb like *fero* usually appears before a set of object nouns, they will be linked together in the network. If another verb like *accipio* shows similar behavior, then its neighbors will overlap (at least in part) with *fero*. However, if both verbs can occur in any position in the sentence, no syntactic regularity forbids the neighbors to be different.

This measure of irregularity is more reliable than other topological properties of networks such as Clustering Coefficient or Average Minimum Path Length, because those properties may be skewed by the size of the network. Indeed, on account of idiosyncratic properties of the texts, both networks of the late varieties appear to have a smaller number of nodes and edges compared to classical varieties (see Table 2), although they are generated from texts of comparable length. This implies that texts in late varieties consist of fewer lemmas and on average each lemma appears more frequently (possibly creating more edges). This boosts the connectedness of the corresponding network artificially.

Irregularity can be assessed quantitatively through spectrum analysis, which consists in estimating the eigenvalues of the binary adjacency matrix of a global network. λ is an eigenvalue for this matrix on condition that there is a non-zero vector x (named eigenvector) that can satisfy the equation: $A\mathbf{x} = \lambda\mathbf{x}$. The spectrum of A is the density of the set Λ including all the eigenvalues $\lambda_1 \ldots \lambda_n$ and their multiplicities (the number at which an identical eigenvalue repeats): the set cardinality amounts to the number of rows/columns in the matrix. The density is a function over a continuous random variable (in this case, eigenvalues) and represents the likelihood that the variable values fall within a certain range. This likelihood is evaluated as the integral over the function values within that range.

Spectrum analysis has been proven to be useful in unraveling grammatical regularities that are independent from pure frequency by Choudhury et al. (2010). In fact, there are methods such as the Dorogovtsev-Mendes growth model (Dorogovtsev et al. 2000) to generate artificial networks that are indistinguishable from real networks created from corpora with respect to their topological properties. Crucially, however, real and artificial networks differ in their spectra. In particular, the density of the former is higher around zero. This happens because grammatical constraints make the neighborhoods of nodes in real networks more regular (see above). In other words, the rows (or equivalently columns) of elements with similar grammatical behavior are more similar. As the eigenvectors and eigenvalues can be interpreted geometrically as the direction coordinates and the factor of a transformation, respectively, then the more the factor tends to 0, the more reduced is the extent of the transformation. The lesser this extent, the more regular a matrix.

To evaluate freedom in word order, we calculated the spectra of co-occurrence and dependency networks (treated as unoriented in order that eigenvalues are real numbers). We plot the spectra for Ancient Greek in Figure 2 and the spectra for Latin in Figure 3.

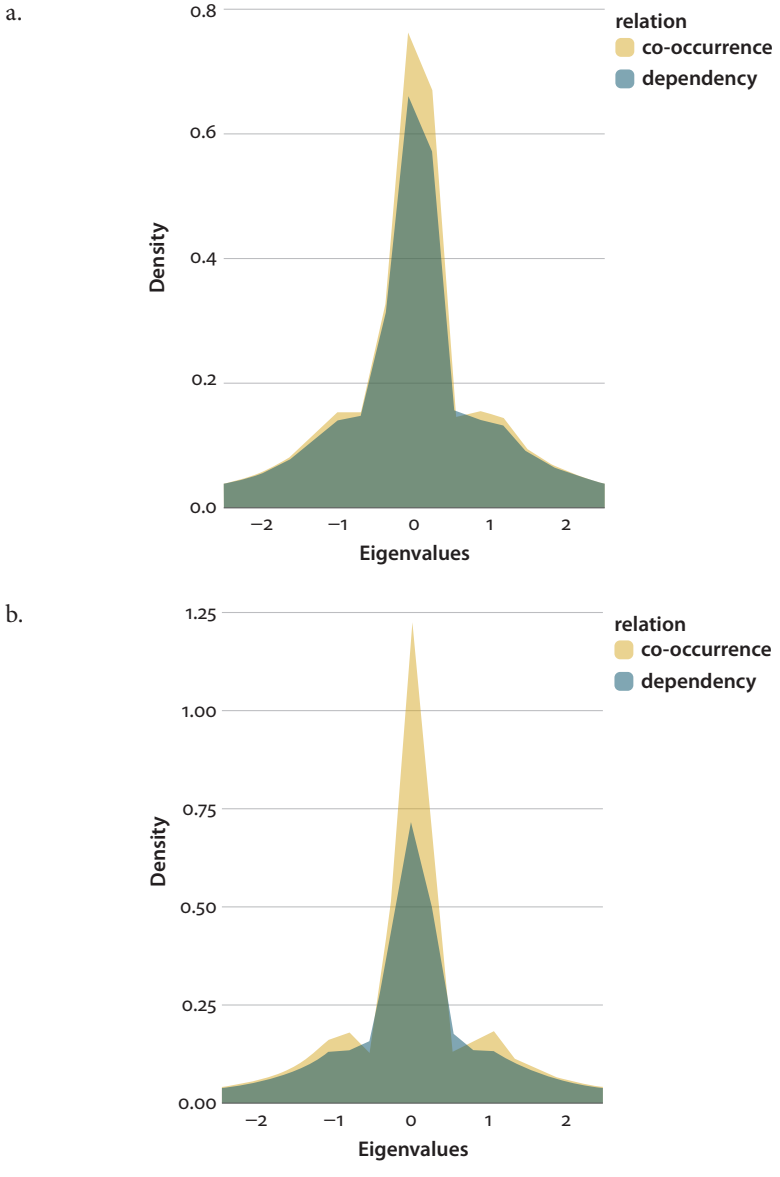
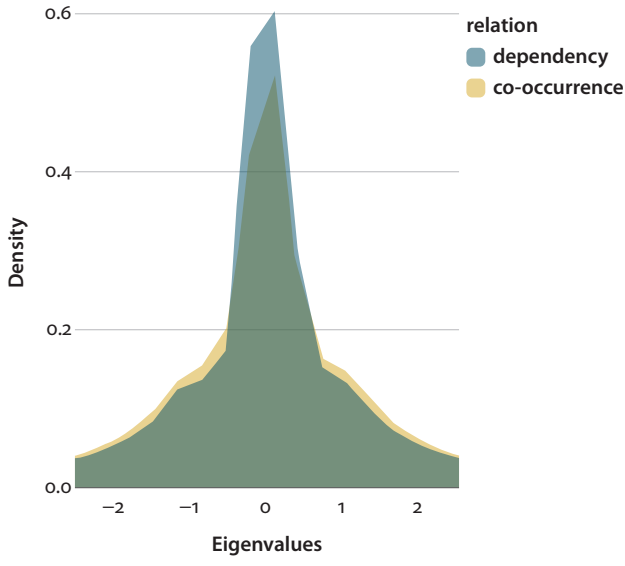


Figure 2. Spectra for co-occurrence (yellow) and dependency (green) networks for Classical (a) and Late (b) Greek

a.



b.

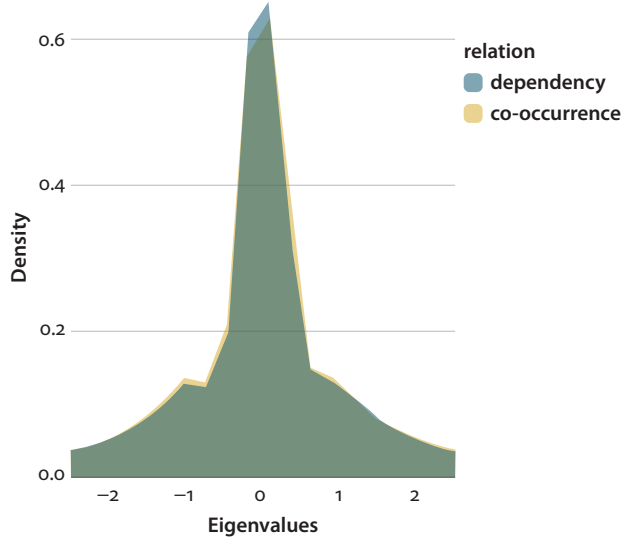


Figure 3. Spectra for co-occurrence (yellow) and dependency (green) networks for Classical (a) and Late (b) Latin

In Table 4 we compare the numerical values for the densities of $\lambda = 0$ (null factor of transformation).

Table 4. Density at $\lambda = 0$ of the co-occurrence and dependency adjacency matrixes

Variety	Co-occurrence	Dependency
Classical Greek	0.74	0.63
Late Greek	1.21	0.69
Classical Latin	0.50	0.59
Late Latin	0.60	0.63

As shown in Table 4, the density of the eigenvalues around 0 grows in late varieties for co-occurrence networks. A higher value means a higher regularity in the network; in turn, we maintain this to be a proxy for a more rigid word order. In particular, the metric value increases by 63.5% in Ancient Greek and by 20% in Latin. The soundness of our method is demonstrated by the stability of the density at 0 across languages and time for dependency networks. In fact, we expect that these syntactic relations, being universal, enforce a constant set of constraints on the word combinations.

4.2 Discontinuous NPs

We measured NP discontinuity by analyzing local networks, i.e., dependency trees. We considered two types of constituents: NPs consisting of (i) article + noun for Ancient Greek or (ii) attributive adjective + noun for both languages. These constituents were counted as discontinuous on condition that their components were separated in the linear order by at least an element that does not belong to their subtree. More formally, if a node lies between other nodes but does not share a common sub-tree with them, it creates non-projectivity and is said to be 'in a gap' (Marcus 1965). Note that in both the cases a head-dependent relation holds between the components of the constituent.

Mambrini & Passarotti (2013) have shown that the amount of non-projective trees in Ancient Greek and Latin is higher than in any modern Indo-European language they were able to test – based on treebank availability – on account of the occurrence of discontinuous constituents. In Ancient Greek, this phenomenon is even more relevant: the main causes listed by the authors are clitics and other P2 particles, which occupy the second position in the sentence and thus split any constituent spanning across that position. Mambrini & Passarotti (2013) found another source of non-projectivity in the displacement to the left of arguments and adjuncts of the verb either for pragmatic purposes or from a subordinate clause to the main clause. For nouns, interrogative pronouns and predicative adjectives contribute to non-projectivity the most.

In late varieties of Ancient Greek and Latin, however, continuous NPs grammaticalized. This trend has been charted quantitatively: Gulordava & Merlo (2015) estimated the percentage of adjacent heads and dependents in NPs. Diachronically they observed a sharp decrease in non-adjacent modifiers in both Ancient Greek and Latin. However, this criterion does not necessarily imply a discontinuity, since heads and dependents can be separated by words that depend on either of them and belong to the same constituent. In this work, we devise a metric limited to actual discontinuities.

We measured the absolute frequency of constituents separated by at least a 'node in a gap' in the four language varieties using the Tree Query extension of the TrEd Tree Editor[2] software.[3] In order to make the treebanks readable by this software, we pre-processed them by converting them into the Prague Markup Language (PML), a data format based on XML intended for storing linguistically annotated data. In the queries, we searched nodes preceding/following a noun and following/preceding an adjective or an article with the dependency relation of attribute or determiner, respectively. We also ensured that the adjective or article depends on the noun and that the intervening nodes are not part of this subtree. As an example, consider the sentence in (12) and the ensuing tree retrieved by the query in Figure 4.

(12) *Qui* *diutissime* *impuberes* *permanserunt,* *maximam*
REL.NOM.PL long chaste:NOM.PL remain:PRET.3PL highest:ACC
inter *suos* *ferunt* *laudem*
among POSS.3PL.ACC.PL carry:PRS.3PL praise:ACC
"Those who have remained chaste for the longest time, receive the greatest commendation among their people." (Caes. G. 6.21)

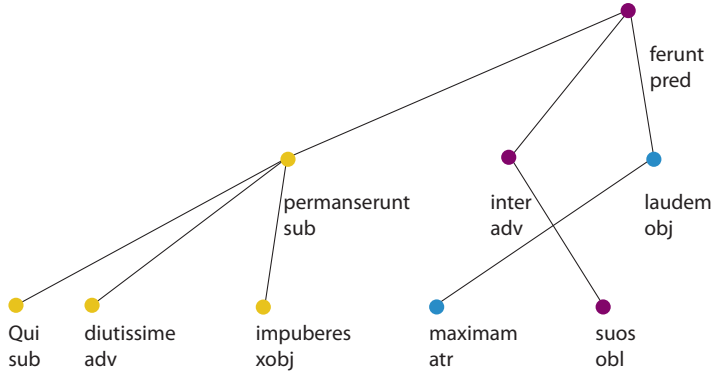


Figure 4. Discontinuous constituent in Classical Latin retrieved by the query

2. https://ufal.mff.cuni.cz/tred/

3. We excluded discontinuous constituents in coordinated constructions to simplify the query.

In (12), a NP *maximam … laudem* "highest praise" (in blue in Figure 4) is separated by three words in a gap *inter suos ferunt* "receive among them" (in purple), which belong to a different subtree.

The results of the queries appear in Table 5 (the total includes right-headed pairs (AdjN) and left-headed pairs (NAdj)).

Table 5. Counts of pairs divided by at least a word in a gap (i.e., creating discontinuities) by kind of constituent

Language variety	Noun phrases (determiner)	Noun phrases (adjective)
Classical Greek	400	85 + 206 = 291
Late Greek	112	12 + 23 = 35
Classical Latin	–	221 + 58 = 279
Late Latin	–	25 + 18 = 43

Table 5 shows that the number of discontinuous constituents drops dramatically in late varieties for both kinds of nominal constituents compared to classical varieties. Discontinuous NPs with determiners drop by −72% in Ancient Greek; discontinuous NPs with adjectives drop by −87.27% in Ancient Greek and −84.58% in Latin. These results confirm the grammaticalization of constituency, which is often held to be true but seldom assessed quantitatively in the literature (e.g., Ledgeway 2011, 2012: 31–58).

As we saw in §2.2, Classical Greek NPs had definite articles while Latin did not. Table 5 indicates a major difference between the drop in frequency of split constituents depending on whether it is the article that is separated from the noun or whether it is an adjective. In the former, the frequency is less reduced than in the latter. This depends on the fact that, as remarked in §2, when an NP with a definite article occurs initially in the sentence, a P2 sentence particle is placed immediately after the article, as in (3). Indeed, a qualitative analysis shows that the nodes in a gap between the article and the noun in Ancient Greek derive from the occurrence of such particles, as shown in Table 6, in which we considered the top five nodes in a gap by the frequency (in parentheses) of their lemma.

Table 6. Ranking by frequency of words in a gap between articles and nouns in the varieties of Ancient Greek

Classical Greek	Late Greek
dé (254)	*dé* (177)
te (76)	*oûn* (16)
mén (48)	*gár* (15)
dḗ (43)	*mén* (14)
gár (41)	*te* (9)

4.3 Referential null objects

To estimate the number of referential null objects, no direct source for information is available either in local or in global networks, since co-reference and implicit nodes are not present in all the treebanks. Instead, this measure was approximated by the percentage of third person personal pronouns among the objects of verbs. We assumed that, if the object is mandatory, it must surface as a pronoun in some constructions.[4] On the other hand, when null anaphora of the object is possible, no object at all is expressed. As a consequence, the loss of null objects in late varieties is expected to be related to the skyrocketing of the rate of personal pronouns among objects. As the nominative of first and second person pronouns is especially frequent in discourse, we limit our observations to third person pronouns. Table 7 contains their count, the total object count and their ratio. Also, we counted the number of verb pairs in coordination with the following requirements: both must govern objects, the first being a noun and the second a third person personal pronoun. Although this query retrieves false positives where the two objects are not co-referent, it nevertheless indicates a clear diffusion of this pattern in late varieties.

Table 7. Counts of the objects (total and subsets)

	CG	LG	CL	LL
objects	7171	6659	6944	7675
of which **3rd person personal pronouns**	295 (4.11%)	789 (11.85)	167 (2.40)	763 (9.94)
of which **in coordinated constructions**	5 (0.07%)	58 (0.87%)	10 (0.14%)	57 (0.74%)

The results in Table 7 strongly support our hypothesis. This change also entails a side effect in global networks. Each node in these networks is associated with a degree, which equals the number of its edges. The nodes with the highest degree are defined hubs, and their deletion alters the topology of the linguistic networks dramatically because this would make them highly disconnected. Since in late varieties personal pronouns occur more often and they collocate with an open, paradigmatically rich class like verbs, they increase the number of edges of their corresponding nodes in global networks. This is equivalent to saying that they

4. One could object that a direct object can be realized by an NP rather than a (personal) pronoun. Note however that null direct objects occur in contexts in which they are easily recoverable from the context (see Luraghi 1997, 2003 on specific conditions), and in such conditions overtly realized objects are normally weak pronouns.

increase their 'hubness'. Table 8 shows the top six lemmas by degree. Note that since these figures concern lemmas, and not specific forms, all forms of pronouns are included, most notably the nominative.

Table 8. Ranking of the most connected nodes of global networks by number of edges

Position	CG	LG	CL	LT
1	*ho*	*ho*	*et*	*et*
2	*kaí*	*kaí*	*sum*	*sum*
3	*dé*	***autós***	*qui*	***is***
4	*eimí*	*dé*	*que*	*in*
5	*hoûtos*	*eimí*	***is***	*autem*
6	***autós***	*egṓ*	*ego*	*qui*

Third person pronouns (*autós* for Ancient Greek, *is* for Latin) climb the ranking in the late varieties compared to the classical varieties.[5] The evidence from both the local and global networks points towards a change in the role of third person pronouns: in particular, their rising tendency to appear as verb objects points toward the overt realization of the object even in contexts in which its referent can be recovered from discourse.

5. Discussion

In the previous sections, we drew three straightforward conclusions, which can be summarized as follows:

- Free word order. Words in late varieties show a higher regularity in the co-occurrence patterns.
- Constituents. The number of discontinuous NPs drops dramatically in late varieties.
- Null referential direct objects. Third-person personal pronoun objects increase in number inside the treebanks, and third person pronouns increase in degree inside the global networks in late varieties.

Is the variation across these parameters correlated? In order to assess this, we calculated the Pearson correlation coefficients of each possible pair of metrics. This

5. Note that we are giving the lemmas of pronouns. We are, however, well aware of the fact that in no Ancient Greek variety does the nominative *autós* function as a third-person anaphoric pronoun: this function is limited to non-nominative forms.

parameter measures the linear correlation between a pair of random variables, i.e., values observed for a given parameter: in our case, the values within a single language. Pearson correlation coefficients measure the covariance of the two random variables normalized by the product of their standard deviations. A matrix of Pearson correlation coefficients for the measures of non-configurationality features is shown in Table 9. In the bottom-left half we report the strength of the correlation (rho). The possible values range between −1 (perfect negative correlation) and 1 (perfect positive correlation), passing by 0 (no correlation). The top-right of the table, instead, shows the statistical significance. Such confidence in the correlation is expressed by p-values, i.e., the probabilities (from impossibility 0 to certainty 1) that the correlation has emerged by chance (in other words, that the null hypothesis is true). Note that we excluded NPs with articles because part of the values are not applicable, as Latin has no articles, hence the measure cannot be computed.

Table 9. Pearson coefficients (bottom-left) and their significance expressed as p-value (top-right) of the correlations among the variables measured in both global and local networks

	Density	**Discontinuous NPs w/ adjective**	**3rd person personal pronouns in coordination**
Density	–	$p = 0.4692$	$p = 0.4884$
Discontinuous NPs w/ adjective	$\rho = -0.53$	–	$p = \mathbf{0.0007}$
3rd person personal pronouns in coordination	$\rho = 0.51$	$\rho = \mathbf{-0.99}$	–

Based on the figures in Table 9, we found a correlation to be statistically significant and strong. Indeed, the count of third-person personal pronouns in coordination correlates negatively with the number of non-projective adjective-noun pairs. Since we linked the former with the absence of the null anaphora of objects, it turns out that the occurrence of null referential direct objects and the occurrence of discontinuous NPs are interdependent.

On the other hand, metrics related to word order freedom do not offer any evidence to support a correlation with the other metrics: this might be due to the independence of word order freedom from non-configurationality. In fact, word order is relatively unconstrained also in some languages that are usually taken to be configurational, such as most Romance languages, hence it is comparatively less revealing than other features. In general, the correlations show that various developments commonly considered typical of increasing configurationality besides being parallel in time are indeed part of the same ongoing change.

6. Conclusions

We have argued that correlates of non-configurationality found in Classical Greek and Latin were declining in late varieties of the two languages and that their decline can be measured by using appropriate metrics based on network analysis, both at the local level (syntactic dependency trees and word order of single sentences) and at the global level. In this way, we have identified clues to increasing configurationality, consisting in a decrease in the freedom of word order, in the almost complete disappearance of discontinuous NPs and in the increase of pronominal direct objects (mirror image of the simultaneous decrease of null direct objects). Moreover, we have found a significant and strong correlation between discontinuous NPs and null direct objects, demonstrating that these variables co-vary over time.

Nevertheless, there remain some caveats with respect to the data, metrics and variables considered. As mentioned in §3.1, diachrony is not the only variable explaining differences in the texts: style and genre can also influence the freedom of word order. In Ancient Greek, discontinuity within NPs shows a lesser decrease in cases in which definite articles occur, due to the frequency of P2 particles that were routinely placed between the article and the noun. Moreover, while we have proposed the presence of third person personal pronominal direct objects as a proxy for the absence of null direct objects in specific contexts in which their referent is recoverable from discourse, the ideal metric for null objects would also require a pragmatic level of annotation, which is only partly available for these languages currently. Finally, a larger sample in future experiments would be required to corroborate the evidence for this correlation. In particular, more languages should be taken into account, and the treebanks for Ancient Greek and Latin should be extended by manual annotation or syntactic parsing (Gulordava & Merlo 2015; Ponti & Passarotti 2016).

Acknowledgements

We would like to thank the editors of *Diachronica* and the reviewers for their invaluable suggestions, which made our work more thorough in both the theoretical assumptions and the technical implementation. Edorado Ponti also wishes to thank Marco Passarotti for having introduced him into the fascinating field of network analysis.

References

Austin, Peter & Joan Bresnan. 1996. Non-configurationality in Australian languages. *Natural Language and Linguistic Theory* 14. 215–268. https://doi.org/10.1007/BF00133684

Baker, Mark. 2001. Configurationality and polysynthesis. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook*, vol. 2, 1433–1441. Berlin: Mouton de Gruyter.

Bamman, David, Francesco Mambrini & Gregory Crane. 2009. An ownership model of annotation: The Ancient Greek dependency treebank. In Marco Passarotti, Adam Przepiórkowski, Savina Raynaud & Frank van Eynde (eds.), *Proceedings of the eighth international workshop on treebanks and linguistic theories (TLT 8)*, 5–16. Milan: EDUCatt.

Baronchelli, Andrea, Ramon Ferrer-i-Cancho, Romualdo Pastor-Satorras, Nick Chater & Morten H. Christiansen. 2013. Networks in cognitive science. *Trends in Cognitive Sciences* 17(7). 348–360. https://doi.org/10.1016/j.tics.2013.04.010

Čech, Radek, Ján Mačutek & Zdeněk Žabokrtský. 2011. The role of syntax in complex networks: Local and global importance of verbs in a syntactic dependency network. *Physica A: Statistical Mechanics and its Applications* 390(20). 3614–3623. https://doi.org/10.1016/j.physa.2011.05.027

Choudhury, Munmun, Dipak Chatterjee & Animesh Mukherjee. 2010. Global topology of word co-occurrence networks: Beyond the two-regime power-law. In Aravind K. Joshi, Chu-Ren Huang & Dan Jurafsky (eds.), *Proceedings of the 23rd international conference on computational linguistics: Posters*, 162–170. Stroudsburg, PA: Association for Computational Linguistics.

Danckaert, Lieven. 2015. Studying word order changes in Latin: Some methodological considerations. In Carlotta Viti (ed.), *Perspectives on historical syntax*, 233–250. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.169.09dan

Deligianni, Efrosini. 2011. Modern Greek word order in the process of syntacticization: Preliminary evidence from Late Byzantine and Early Modern Greek. In Katerina Chatzopoulou, Alexandra Ioannidou & Suwon Yoon (eds.), *Proceedings of the 9th international conference on Greek linguistics (ICGL 9)*, 440–455. Columbus: The Ohio State University.

Devine, Andrew & Laurence Stephens. 2000. *Discontinuous syntax: Hyperbaton in Greek*. Oxford: Oxford University Press.

Dorogovtsev, Sergey N., José Fernando F. Mendes & Alexander N. Samukhin. 2000. Structure of growing networks with preferential linking. *Physical review letters* 85, no. 21:4633. https://doi.org/10.1103/PhysRevLett.85.4633

Dover, Kenneth. 1960. *Greek word order*. Cambridge: Cambridge University Press.

Ferrer-i-Cancho, Ramon & Richard V. Solé. 2001. The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences* 268(1482). 2261–2265. http://dx.doi.org/10.1098/rspb.2001.1800

Ferrer-i-Cancho, Ramon, Ricard V. Solé & Reinhard Köhler. 2004. Patterns in syntactic dependency networks. *Physical Review E* 69, no. 5:051915.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In Joakim Nivre & Eva Hajičová (eds.), *Proceedings of the third international conference on dependency linguistics (Depling 2015)*, 91–100. Uppsala: Uppsala University.

Goldstein, David. 2016. *Classical Greek syntax: Wackernagel's Law in Herodotus*. Leiden: Brill.

Gulordava, Kristina & Paola Merlo. 2015. Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and Ancient Greek. In Joakim Nivre & Eva Hajičová (eds.), *Proceedings of the third international conference on dependency linguistics (Depling 2015)*, 121–130. Uppsala: Uppsala University.

Hale, Ken. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory* 1. 5–47.

Haug, Dag T. T. & Marius L. Jøhndal. 2008. Creating a parallel treebank of the Old Indo-European Bible translations. In Caroline Sporleder and Kiril Ribarov (eds.), *Proceedings of the workshop on language technology for cultural heritage data (LaTeCH 2008)*, 27–34. Marrakech, Morocco.

Hewson, John & Vit Bubenik. 2006. *From case to adposition: The development of configurational syntax in Indo-European languages*. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.280

Kapustin, Victor & Anna Jamsen. 2007. Vertex degree distribution for the graph of word co-occurrences in Russian. In Chris Biemann, Irina Matveeva, Rada Mihalcea & Dragomir Radev (eds.), *Proceedings of the second workshop on TextGraphs: Graph-based algorithms for natural language processing*, 89–97. Rochester, NY: Association for Computational Linguistics.

Keydana, Götz & Silvia Luraghi. 2012. Definite referential null objects in Vedic Sanskrit and Ancient Greek. *Acta Linguistica Hafniensia* 44(2). 116–128.

Kiss, E. Katalin. 1987. *Configurationality in Hungarian*. Dordrecht: Reidel.

Ledgeway, Adam. 2011. Morphosyntactic typology and change. In Martin Maiden, John Charles Smith & Adam Ledgeway (eds.), *The Cambridge history of the Romance languages, vol. 1: Structures*, 382–471, 724–734. Cambridge: Cambridge University Press.

Ledgeway, Adam. 2012. *From Latin to Romance*. Oxford: Oxford University Press.

Linde, Paul. 1923. Die Stellung des Verbums in der lateinischen Prosa. *Glotta* 12. 153–178.

Luraghi, Silvia. 1997. Omission of the direct object in Classical Latin. *Indogermanische Forschungen* 102. 239–257.

Luraghi, Silvia. 1998. Omissione dell'oggetto diretto in frasi coordinate: Dal latino all'italiano. In Paolo Ramat & Elisa Roma (eds.), 183–196.

Luraghi, Silvia. 2003. Definite referential null objects in Ancient Greek. *Indogermanische Forschungen* 108. 169–196.

Luraghi, Silvia. 2010. The rise (and possible downfall) of configurationality. In Silva Luraghi & Vit Bubenik (eds.), *Continuum companion to historical linguistics*, 212–229. London: Continuum.

Luraghi, Silvia. 2013. Clitics. In Silvia Luraghi & Claudia Parodi (eds.), *The Bloomsbury Companion to Syntax*, 165–193. London: Bloomsbury.

Mambrini, Francesco & Marco Passarotti. 2013. Non-projectivity in the Ancient Greek dependency treebank. In Eva Hajičová, Kim Gerdes & Leo Wanner (eds.), *Proceedings of the second international conference on dependency linguistics (Depling 2013)*, 177–186. Prague: Matfyz Press.

Marcus, Solomon. 1965. Sur la notion de projectivité. *Mathematical Logic Quarterly* 11(2). 181–192. https://doi.org/10.1002/malq.19650110212

Meillet, Antoine & Joseph Vendryes. 1924. *Traité de grammaire comparée des langues classiques*. Paris: Champion.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning et al. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (ed.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 16)*, 1659–1666. Portorož: European Language Resources Association.

Ponti, Edoardo M. 2016. Divergence from syntax to linear order in Ancient Greek lexical networks. In Zdravko Markov and Ingrid Russel (eds.), *Proceedings of the twenty-ninth international FLAIRS conference*, 187–193. Palo Alto: AAAI Press.

Ponti, Edoardo M., & M. Passarotti. 2016. Differentia compositionem facit: A slower-paced and reliable parser for Latin. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 16)*, 683–688. Portorož: European Language Resources Association.

Reinöhl, Uta. 2016. *Grammaticalization and the rise of configurationality in Indo-Aryan*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198736660.001.0001

Revithiadou, Anthi & Vassilios Spyropoulos. 2007. *A typology of Greek clitics with special reference to their diachronic development*. Ms., University of the Aegean. http://ling.auf.net/lingBuzz/000496 (last accessed on 11/07/2018.)

Revithiadou, Anthi & Vassilios Spyropoulos. 2008. Greek object clitic pronouns: A typological survey of their grammatical properties. *Language Typology and Universals* 61(1). 39–53. https://doi.org/10.1524/stuf.2008.0005

Rögnvaldsson, Eiríkur. 1995. Old Icelandic: A non-configurational language? *North-Western European Language Evolution* 26. 3–29. https://doi.org/10.1075/nowele.26.01rog

Salvi, Giampaolo. 2004. *La formazione della struttura di frase romanza*. Tübingen: Niemeyer. https://doi.org/10.1515/9783110945508

Schäufele, Steven. 1990. Free word-order syntax: The challenge from Vedic Sanskrit to contemporary formal syntactic theory. Urbana-Champaign: University of Illinois at Urbana-Champaign dissertation.

Solé, Richard V., Bernat Corominas Murtra, Sergi Valverde & Luc Steels. 2010. Language networks: Their structure, function, and evolution. *Complexity* 15(6). 20–26. https://doi.org/10.1002/cplx.20326

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.

Text form and grammatical changes in Medieval French

A treebank-based diachronic study

Alexandra Simonenko, Benoît Crabbé and Sophie Prévost
FWO/UGhent / LLF/CNRS/Paris Diderot, USPC / LATTICE, CNRS/ENS/ Paris Sorbonne Nouvelle, USPC & PSL

This paper presents a treebank-based study of the effect the text form (prose vs. verse) has on the course of two grammatical changes in Medieval French: the loss of null subjects and the loss of OV word order. By means of statistical analysis, we demonstrate that naive estimates of the spread of overt subjects and VO orders give the impression that there is a significant difference between the rates of development in prose vs. verse. By contrast, estimates based on an abstract grammar competition model which distinguishes between grammar-ambiguous surface forms (overt personal subjects, null subjects in coordination contexts) and grammar-unambiguous surface forms (overt expletive subjects, null subjects in non-coordination contexts) show prose-verse parallelism, prose having an earlier change onset, in line with traditional intuitions. At a more general level, these results suggest that the product of the interaction of a particular grammar with universal pragmatic laws is constant, which can be observed if the factors responsible for variation in grammatical choices are controlled for.

Keywords: prose vs. verse in language change, Constant Rate Effect, null subjects, word order change, Medieval French, treebanks

1. Introduction

This paper investigates the effects of the text form (prose vs. verse) on diachronic changes in Medieval French using the treebanks MCVF and the Penn Supplement to MCVF (≈1,5 million words, Penn scheme annotation).[1] Despite the common intuition that prose is more 'advanced' than contemporary verse with respect to

1. Word counts are based on the version of the Penn Supplement available as of September 2017.

https://doi.org/10.1075/bct.113.04sim

grammatical changes, by virtue of not being subject to the versification constraints, in the absence of statistical models based on large-scale corpora, the magnitude of the difference has remained unknown. Estimates for the decline of pro-drop based on smaller data samples strongly suggest that the distinction is indeed real (Prévost 2018). To estimate the prose-verse lag is especially important for studies modelling language evolution based on written sources. Grammatical factors influencing the speed of language change have to be disentangled from metagrammatical ones associated with conscious stylistic manipulations.

We estimate the prose-verse lag for different types of grammatical changes by means of statistical analysis. Specifically, we examine the trajectory of two changes: the decline of null subjects (morphosyntactic) and the shift from OV_{fin} to $V_{fin}O$ orders (syntactic) across text forms by modelling each change as the evolution of a binary variable whose values correspond to competing grammars (Kroch 1989; for an overview of much subsequent work, see Pintzuk 2004). That is, we estimate the effect of time on the probability that a finite clause has an overt pronominal subject (as opposed to a covert one); as well as the probability that a finite transitive clause with a nominal object exhibits VO (rather than OV) order.

The relevance of this work is threefold. First, it makes a methodological contribution to the study of language change by considering metagrammatical factors potentially affecting the rates of various grammatical changes. Estimating the rate of change has been central to a series of historical analyses pioneered by Kroch (1989), who first suggested that grammatical changes should be analysed not by directly comparing various data points but by comparing the behaviour of well-understood mathematical functions fitted to relevant data sets. The Constant Rate Hypothesis of Kroch (1989) states that a grammatical change progresses at the same rate (or, more accurately, at not significantly different rates) in different GRAMMATICAL contexts. The hypothesis relies on fitting logistic regression models to binary variables. It has been shown to hold for a number of grammatical changes across GRAMMATICAL contexts and is known as the Constant Rate Effect (see Pintzuk 2004 for an overview).[2] The hypothesis says nothing, however, about how changes spread in contexts which contrast in metagrammatical characteristics, such as prose vs. verse, and rightly so, since by definition such contrasts may be associated with conscious manipulations of linguistic features. This means that, to an extent, all bets are off as to what may happen to a given language change in text sources affected by such manipulations, such as versified texts. This study thus charts new territory by means of a large-scale quantitative investigation of the effects of a metalinguistic distinction between prose and verse on the course of grammatical changes spanning the

2. We are not aware of any counterexamples to the hypothesis, that is, developments of clearly the same nature proceeding at different rates in different grammatical contexts.

whole medieval period. A major research question we address here is whether a grammatical change has the same trajectory across metalinguistically different environments. A statistical analysis relying on data from large annotated corpora allows us to demonstrate that grammatical changes proceed in parallel ways in prose and verse, provided that strictly grammatical features are isolated from features susceptible to pragmatic/stylistic variation. In our case one such 'volatile' feature is the use of subordinate clauses, which varies greatly (and in a temporally unstable way) between verse and prose. Our results are meant to be fully replicable: the full set of queries we use is given in online Appendix 1 and the lists of relevant lexical items are provided in online Appendix 2.

Second, this study paves the way for overcoming the issue of a text form/time correlation. For some periods verse may be the only or the dominant form in the available texts, which makes it crucial to understand its potential effects on the course of grammatical changes. For instance, the available body of Medieval French texts is characterised by the prevalence of texts in verse until approximately the end of the 12th century. It needs to be stressed that given the form/time correlation, the only way to estimate the effect of text form on linguistic changes is by means of statistical extrapolation, which, in turn, is only possible if we can estimate parameters of interest, such as the rate of null subjects, at time points for which we have data. Estimating those necessarily requires exhaustive annotation of text samples, which essentially amounts to using an existing treebank or creating a new one. We do not see any other way which would allow us to make conclusions about the text form/time correlation.

Thirdly and finally, this project contributes to a better understanding of specific linguistic phenomena, that is, subject omission and word order, by examining their interaction with text forms. We get a better handle on factors governing these phenomena by relating them to the features which characterise a given text form.

In what follows we first consider the loss of null subjects, then we turn to the loss of OV order (in finite clauses with a non-clitic direct object).

2. The loss of null subjects

We begin by considering the decline of subjectless finite clauses during the medieval period across text forms. Early Medieval French (henceforth MF) is commonly recognised as being (at least partially) a pro-drop language, whereas late MF lost this property completely except in cases of subject ellipsis under coordination. This change is well documented (Foulet 1928; Fontaine 1985; Hirschbühler 1992; Schøsler 2002; Kaiser 2009; Zimmermann 2014; Marchello-Nizia 2018; Prévost 2018; Simonenko et al. 2018). We model it by estimating the distribution of the

variable SUBJECT, which takes the value *yes* if a clause has an overt personal pronominal subject and *no* otherwise, in a sample including all finite clauses with either an overt personal pronominal or null subject (total of 76,150).[3] All clauses are tagged for the date of the manuscript they belong to. We fit these data to a logistic regression model $P(\text{Subject} = \text{yes} \mid \text{Date} = d) = \frac{e^{\alpha+\beta d}}{1+e^{\alpha+\beta d}}$ plotted in Figure 1.[4] Here crosses correspond to the estimated probability of an overt pronominal subject in each text in prose (brown) and verse (yellow). The lines correspond to the predicted probability of an overt pronominal subject based on logistic regression estimates, again, in prose (brown), verse (yellow), and the two text forms combined (blue). Parameter estimates of the model are given in Tables 1 (prose), 2 (verse) and 3 (overall).[5]

Table 1. Logistic regression estimates for overt pronominal subjects in prose

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−0.3562	0.1439	−2.474	0.01
COEFFICIENT	0.0016	0.0001	14.72	$<2 \times 10^{-16}$

Table 2. Logistic regression estimates for overt pronominal subjects in verse

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−4.6863	0.2226	−21.04	$<2 \times 10^{-16}$
COEFFICIENT	0.0038	0.0002	20.30	$<2 \times 10^{-16}$

Table 3. Logistic regression estimates for overt pronominal subjects overall

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	-6.976×10	0.2226	−21.04	$<2 \times 10^{-16}$
COEFFICIENT	6.223×10^{-3}	7.701×10^{-5}	80.81	$<2 \times 10^{-16}$

3. We exclude imperatives and wh-clauses targeting subjects because of their idiosyncratic subject syntax, as well as clauses introduced by connectives *et* "and" and *si* (often difficult to translate, best rendered as "then"). Connectives license subject ellipsis almost at a constant rate throughout the medieval period as well as in Modern French, and therefore they should not be considered as possible pro-drop environments. There are a handful of other conjunctive adverbs capable of licensing subject ellipsis, such as *puis* "then", but since those are much less frequent than *et* and *si*, we do not exclude them.

4. We use logistic regression as has been traditional for modelling historical data since Kroch (1989) (see also Kauhanen & Walkden 2018).

5. For the details of the interpretation of logistic regression parameters we refer the interested reader to Agresti (2002), as well as to Kroch (1989) who offers a very concise introduction of the use of logistic regressions in linguistic analysis.

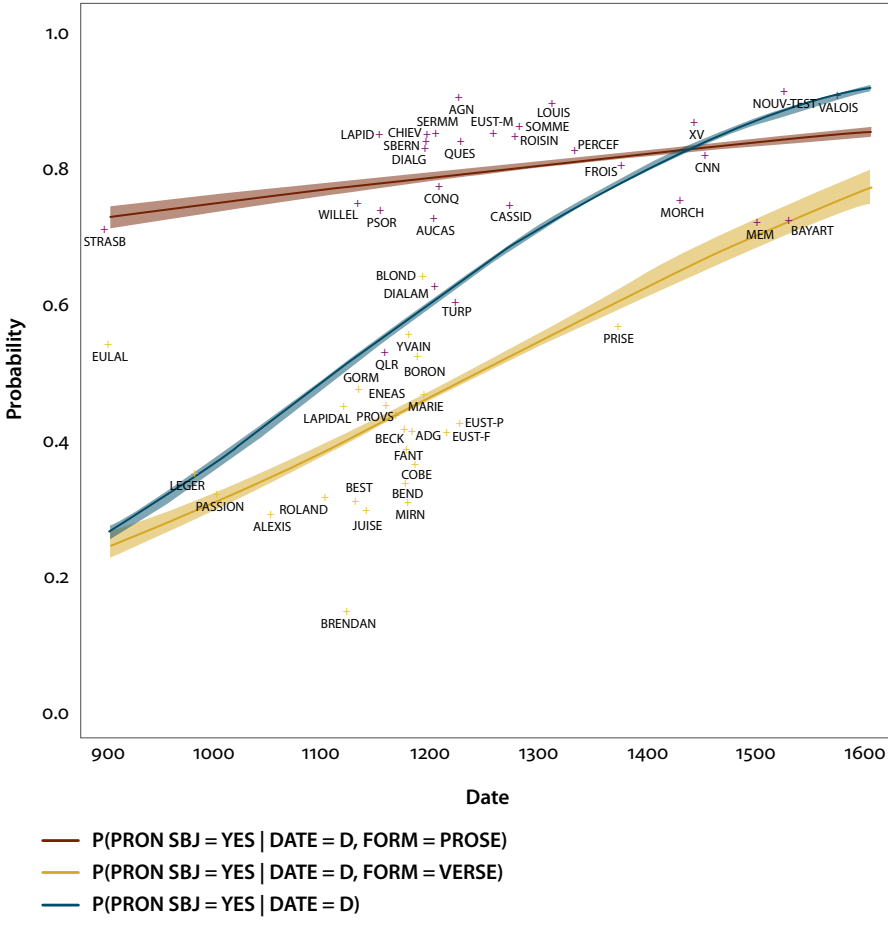


Figure 1. Overt subject emergence in prose and verse

The greater intercept for prose indicates that the change manifested itself first in this text form, in line with the traditional intuition. Looking at the coefficients, we see, first, that the trends are rather different in verse and prose, and, second, that prose is more advanced than verse in terms of the probability of a pronominal subject being overt throughout the medieval period. This contrast is not surprising in itself, given that the prose/verse distinction is of a metalinguistic nature, so we have no *a priori* reasons to expect to find the Constant Rate Effect here. However, investigating what it is about prose that makes it favour overt subjects can be a fruitful line of inquiry since it can shed light on the grammar of null subjects. Interestingly, according to Walkden & Rusten (2017), during the Old English period which features the tailing off of the null subject decline, it is also verse that favours null subjects. Walkden & Rusten (2017: 465) conclude that "null subjects in O[ld] E[nglish] can be seen

mainly as a feature of the poetry".[6] They suggest that metrical requirements imposed on versified texts could have favoured deletion of unstressed monosyllabic pronominals. They also quote Mitchell (1985: 992–993), who suggests that null subjects help poetry "to achieve compression and to give the poetry its characteristic texture". As a matter of speculation, we can say that subject (non)omission is a parameter which can be engaged for metrical purposes (adding or subtracting a syllable whenever needed).[7] However, this topic will have to await a focused quantitative study which would test whether (non)omission of pronominal subjects in verse was aligned with metrical requirements in a non-random way.

2.1 Abstract grammar-based analysis

Before concluding, however, that the emergence of overt pronominal subjects was happening at significantly different rates in verse and prose, let us consider what these surface patterns mean in terms of grammatical shifts. Assuming a model of diachronic variation in terms of grammar competition (between two or more grammars), let us say that the replacement of null personal pronominal subjects by overt ones corresponded to the replacement of a grammar which had a structural component licensing null subject, such as an Agr(eement) head (Jelinek 1984; Barbosa 1995; Alexiadou & Anagnostopoulou 1998) by a grammar without such a head. Specifically, the output of the first grammar, let us call it the AgrP-Grammar, contained both null and overt personal pronominal subjects but only null expletive subjects (as is the case in modern incontestable pro-drop languages such as Italian). The output of the second – let us call it the Tense Phrase-Grammar (TP-Grammar) – had only overt subjects (in contexts not licensing subject ellipsis), whether personal pronominal or expletive. Thus the only subject type which can unambiguously be classified as belonging to the output of one grammar or another are expletive subjects. A null expletive corresponds to the AgrP-Grammar, an overt one to the TP-Grammar. Moreover, because both grammars are, by hypothesis, categorical as to whether expletives are overt or null, we can expect that the (non) expression of expletives is entirely a function of the probability of a given grammar to be used at a given point in time and is not something a given speaker has control of once (s)he has chosen a generating grammar for a given illocutionary act. This

6. Walkden & Rusten (2017: 465) show that in the earliest Old English texts the share of null subjects in verse is about 12%, as opposed to ca. 2% in prose.

7. Old French and Old English poetry were both based on qualitative metre, the most widespread metres being iamb and trochee.

means that while the expression of some personal pronominal subjects in verse could have been the result of metrical adjustments or other stylistic factors, with expletives this possibility is eliminated. We therefore model the spread of overt expletives only, across prose and verse.

We fit finite clauses with either null or overt expletive subjects (total of 11,495) to the model $P(\text{Subject} = \text{yes} \mid \text{Date} = d) = \frac{e^{\alpha + \beta d}}{1 + e^{\alpha + \beta d}}$ plotted in Figure 2.

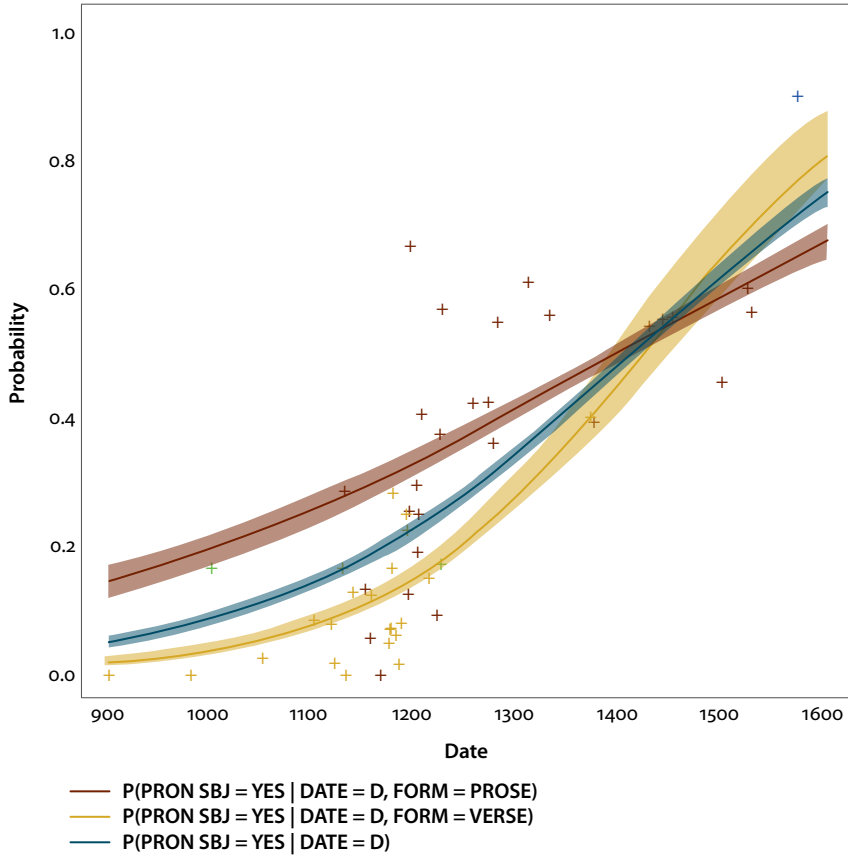


Figure 2. Overt expletive subject emergence in prose and verse

Table 4. Logistic regression estimates for overt expletive pronominal subjects in prose

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−5.0264	0.2842	−17.69	$<2 \times 10^{-16}$
COEFFICIENT	0.0036	0.0002	17.36	$<2 \times 10^{-16}$

Table 5. Logistic regression estimates for overt expletive pronominal subjects in verse

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	-1.115×10^{01}	6.377×10^{-01}	−17.48	$<2 \times 10^{-16}$
COEFFICIENT	7.851×10^{-03}	5.204×10^{-04}	15.09	$<2 \times 10^{-16}$

We observe in Figure 2 a striking parallelism between verse and prose for the time period for which we have good confidence of estimation (until around 1300). This confirms our grammar competition-based prediction that expletive expressions are 'out of reach' for metalinguistic manipulations, since those presumably cannot override the boundaries of grammaticality. To quote Kroch (1989: 36), this shows "the controlling effect of abstract grammatical analyses on patterns in usage data". Specifically, an analysis in terms of grammatical options rather than in terms of direct surface forms allows us to separate what appears to be a properly grammatical change from the effects of metalinguistic prose/verse distinction, even though the nature of the latter remains to be explained. We will see below that an abstract syntactic analysis has a similar clarifying effect on the disappearance of the OV order.

2.2 Direct vs. narrative discourse

Let us explore another perspective and consider overt pronominal subject emergence in MF across discourse types, that is, direct vs. narrative. It is well established that these two registers differ quantitatively with respect to a number of grammatical characteristics (e.g., Dufter 2010 and references therein; Marchello-Nizia 2012; Lagorgette & Larrivée 2013; Guillot-Barbance et al. 2017; Glikman & Mazziotta 2013; Prévost 2018). The two types are illustrated in (1). Figure 3 visualises logistic regression models estimating the emergence of overt subjects (both personal and expletive) in direct discourse vs. narrative for verse and prose.

(1) *Respondet l' altre: "Mal i diz."*
responds the other bad there say
"The other one responds, 'You are wrong'." (1000-PASSION-BFM-P,113.216)

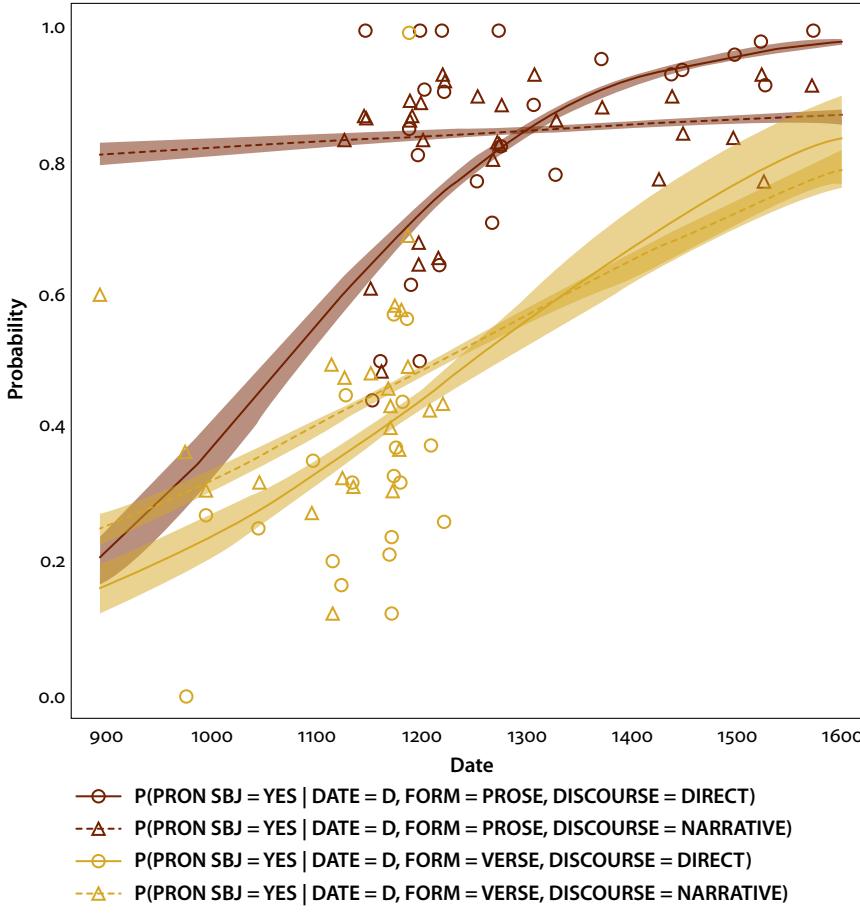


Figure 3. Overt subject emergence in prose and verse & direct and narrative discourse

The estimates of the logistic regression models are given in Tables 6–9.

Table 6. Logistic regression estimates for overt pronominal subjects in prose (direct discourse)

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−8.2267	0.4258	−19.32	$<2 \times 10^{-16}$
COEFFICIENT	0.0076	0.0003	22.98	$<2 \times 10^{-16}$

Table 7. Logistic regression estimates for overt pronominal subjects in prose (narrative)

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	0.8758	0.1556	5.626	1.85×10^{-08}
COEFFICIENT	0.0006	0.0001	5.716	1.09×10^{-08}

Table 8. Logistic regression estimates for overt pronominal subjects in verse (direct discourse)

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−8.1465	0.2187	−37.24	$<2 \times 10^{-16}$
COEFFICIENT	0.0057	0.0002	35.29	$<2 \times 10^{-16}$

Table 9. Logistic regression estimates for overt pronominal subjects in verse (narrative)

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−5.9545	0.6172	−9.64	$<2 \times 10^{-16}$
COEFFICIENT	0.0047	0.0005	9.01	$<2 \times 10^{-16}$

Focusing on direct discourse, the change appears to proceed in a parallel way in verse and prose, with prose as expected being more advanced than verse. We also see that in verse there is virtually no difference between direct speech and narrative. It is more difficult to interpret the virtual stability of the rate of pronominal subject expression in prose narrative, as opposed to prose direct discourse, where the change progresses along an expected curve. As a consequence, it looks as though until approximately the end of the 13th century, prose narrative is more advanced than direct discourse, and then the situation reverses. This contrasts with the results of Glikman & Mazziotta (2013: 77), who report more overt subjects in direct discourse (in a sample of clauses from one text). This difference in results, however, may be due to a methodological difference: we exclude subjects omitted under coordination, while Glikman & Mazziotta (2013) included them. This suggestion is supported by the fact that in the sample examined by Glikman & Mazziotta (2013: 79) we find more connectives such as *et* "and" in narrative (and therefore more contexts for subject ellipsis) than in direct discourse. This methodological point aside, our result runs counter to the commonly accepted idea that direct speech is more advanced than narrative with respect to the progress of grammatical changes. It has been largely acknowledged that direct speech (whatever the state of a language is) displays linguistic features closer to spoken language than narrative does, although it cannot be strictly equated with the latter. Because linguistic changes are expected to be more advanced in spoken language than in written language, it is expected that innovating features appear first in direct speech.

Recall that we run into a similar issue with the rate of pronominal subject expression in prose in general (expletive and personal subjects and direct discourse and narrative combined) in §2. One feature which potentially sets apart prose narrative from both prose direct discourse and verse (narrative and direct discourse) is the frequency of subordinate clauses, which are known to favour subject expression significantly more than matrix ones (Adams 1987; Franzén 1939; Foulet 1928;

Hirschbühler 1992; Prévost 2018; Roberts 2014; Vance 1997; Zimmermann 2014, among others). If this feature does indeed set them apart, the apparently stable high rate of pronominal subject expression in prose narrative may be due to a larger share of subordinate clauses in prose narrative than in any other text form we have examined and to the fact that the change comes to completion earlier in subordinate clauses. This hypothesis can be tested if we check for the relative frequency of subordinate clauses in different text forms. The relevant numbers are given in Table 10.

Table 10. Frequency of clause types across text forms

	MATRIX	MATRIX QUESTIONS	SUBORDINATE
PROSE NARRATIVE	0.53 (56964)	0.00	0.47 (50831)
PROSE DIRECT DISCOURSE	0.82 (13466)	0.02 (319)	0.16 (2647)
VERSE NARRATIVE	0.61 (33615)	0.00	0.39 (21159)
VERSE DIRECT DISCOURSE	0.89 (11638)	0.01 (116)	0.11 (1385)

In order to further test for the influence of discourse type and text form on the rate of subordinate clauses, we ran a logistic regression model on a dependent variable CLAUSE TYPE with the values *matrix* and *subordinate* (ignoring the very infrequent matrix questions) with the predictor variables FORM *(prose, verse)* and DISCOURSE TYPE *(narrative, direct)*. As the summary of the model's parameters in Table 11 shows, both predictors are highly significant, with narrative affecting the probability of a subordinate clause positively and verse negatively. That is, prose narrative comes out as the environment favouring subordinate clauses the most, which can explain the high rate of pronominal subject expression in this environment.

Table 11. Logistic regression estimates for clause type

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−0.9503	0.0218	−43.50	$<2 \times 10^{-16}$
DISCOURSE TYPE (NARRATIVE)	2.2958	0.0229	100.32	$<2 \times 10^{-16}$
FORM (VERSE)	−0.5466	0.0175	−31.15	$<2 \times 10^{-16}$

In Figure 4 we plot models fitting the distribution of the variable CLAUSE *(matrix, subordinate)* in prose and verse. The rate of subordinate clauses appears to be increasing in verse.[8] Tables 12 and 13 show parameter estimates for the the models.

8. We cannot test for the significance of the difference between the model's coefficients in verse and prose due to insufficient data for verse in the later periods.

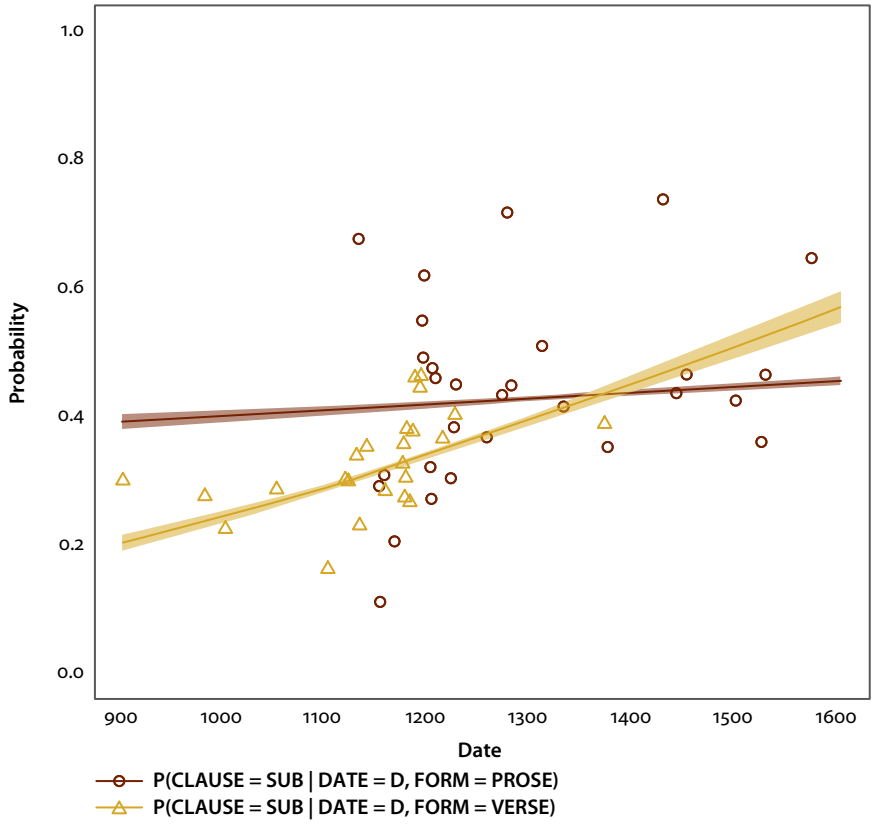


Figure 4. Subordinate clauses in prose and verse

Table 12. Logistic regression estimates for subordinate clauses in prose

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	-7.739×10^{-1}	6.542×10^{-2}	−11.83	$<2 \times 10^{-16}$
COEFFICIENT	3.722×10^{-4}	4.864×10^{-5}	7.65	1.97×10^{-14}

Table 13. Logistic regression estimates for subordinate clauses in verse

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−3.481	0.136	−25.50	$<2 \times 10^{-16}$
COEFFICIENT	0.0023	0.0001	20.46	$<2 \times 10^{-16}$

In view of these results, let us focus our attention on matrix clauses alone. As Figure 5 shows, if limited to this environment, the picture conforms to the traditional expectation of a faster change in environments approximating speech, that is, in direct discourse. Model estimates are given in Tables 14–17.

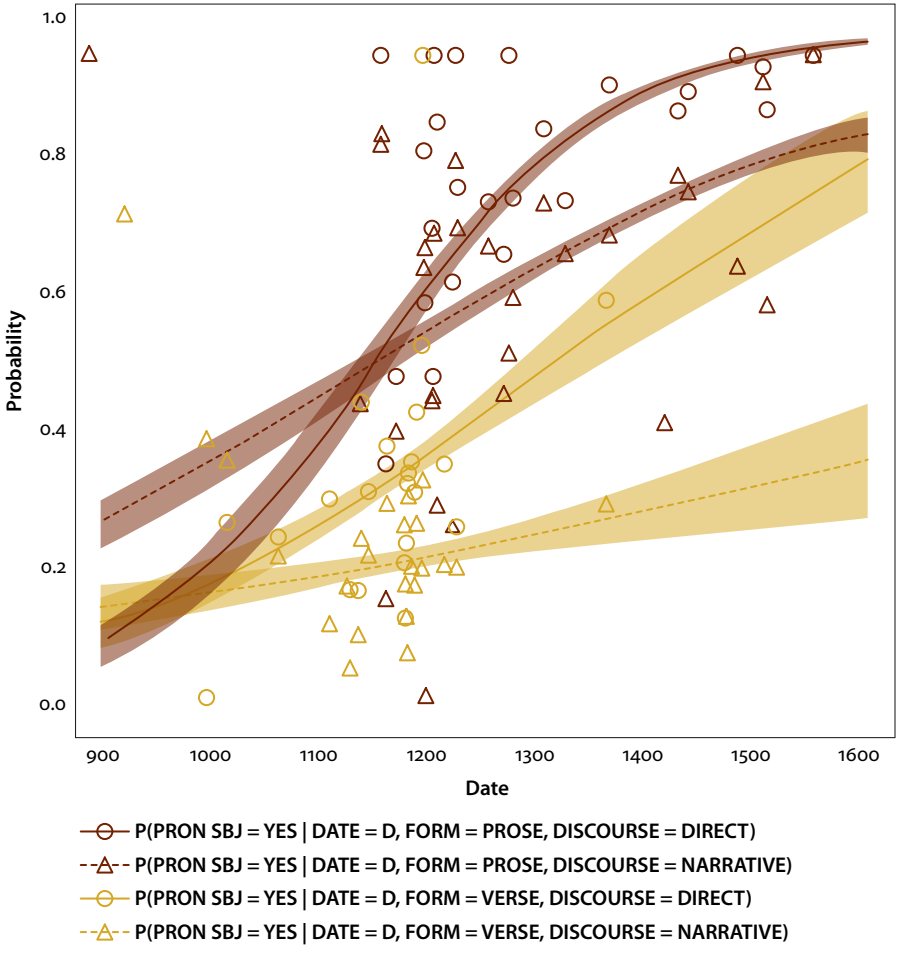


Figure 5. Overt subject emergence in prose and verse and direct and narrative discourse (matrix clauses)

Table 14. Logistic regression estimates for overt pronominal subjects in prose (direct discourse, matrix)

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	-1.035×10	5.017×10^{-1}	−20.63	$<2 \times 10^{-16}$
COEFFICIENT	9.070×10^{-3}	3.879×10^{-4}	23.38	$<2 \times 10^{-16}$

Table 15. Logistic regression estimates for overt pronominal subjects in prose (narrative, matrix)

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−4.5319	0.2704	−16.76	$<2 \times 10^{-16}$
COEFFICIENT	0.0039	0.0002	19.58	$<2 \times 10^{-16}$

Table 16. Logistic regression estimates for overt pronominal subjects in verse (direct discourse, matrix)

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−6.1587	0.6605	−9.32	$<2 \times 10^{-16}$
COEFFICIENT	0.0047	0.0006	8.42	$<2 \times 10^{-16}$

Table 17. Logistic regression estimates for overt pronominal subjects in verse (narrative, matrix)

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−3.2436	0.5023	−6.45	1.07×10^{-10}
COEFFICIENT	0.0017	0.0004	3.98	6.69×10^{-05}

Summarising up to this point, in this study of pro-drop across text forms we have first established that, if surface forms are counted indiscriminately, that is, all kinds of null subjects together and without distinguishing discourse types, prose appears to have a very different change profile, with pronominal subject expression rates being very high from the earliest texts on. If not predicted, this is at least not surprising in the two-grammar competition model where the old grammar allows for overt subjects under some pragmatically defined conditions. This pragmatic flexibility can arguably be exploited differently in different text forms. Once we look at the data in which the output of the two grammars is assumed to be categorically distributed, namely, clauses with expletive subjects (i.e., always null for the old grammar and always overt for the new one), the prose/verse distinction virtually disappears, as predicted by our grammar competition model. That is, once pragmatic factors are excluded, we find a grammatical parallelism between the two text forms. Another

way to uncover this parallelism is to look at the environment which is assumed to approximate oral speech the most, direct discourse. The rates of overt pronominal subjects are similar for verse and prose in this environment. We conclude that the major source of non-parallelism in other contexts is the uneven distribution of subordinate clauses, known to favour subject expression. If limited to matrix clauses, the change develops in parallel ways across prose and verse in pragmatically similar environments (either narrative or direct discourse). The influence of pragmatic factors on the change is thus stable across text forms (as manifested by the absence of dramatic differences between rates of change in matrix clauses) if we properly control for the grammatical environments with which these factors interact, such as the distinction between matrix and subordinate clauses.

3. OV_{fin} decline in prose vs. verse

Early MF is known to have greater word order flexibility than Modern French, in particular, in allowing for both $V_{fin}O$ and OV_{fin}, with the latter option disappearing with time (Marchello-Nizia 1995; Vance 1997; Labelle & Hirschbühler 2005; Labelle 2007; Zaring 2011; Marchello-Nizia & Rouquier 2012; Kroch & Santorini 2014). Examples below illustrate the OV_{fin} option unavailable in Modern French.

(2) [*lei*]$_{obj}$ *consentit*$_{v}$ *et* *observat*$_{v}$
law agreed and observed
"he respected and observed the law" (0980-LEGER-V,XII.82)

(3) [*Ja mais*]$_{adv}$ [*ledece*]$_{obj}$ *n'avrai*$_{v}$
never joy won't.have
"I will never have joy" (10XX-ALEXIS-V,99.892)

(4) [*Li quens Rollant*]$_{sbj}$ [*Gualter de l' Hum*]$_{obj}$ *apelet*$_{v}$
the king Roland Walter of the Hum called
"The king Roland called Walter of Hum" (1100-ROLAND-V,65.779)

In what follows we examine the effects of the verse/prose distinction on how this change proceeded.

3.1 From OV to VO: Simple estimates

We first model this change by estimating the distribution of the variable $V_{fin}O$ (with the values *yes* and *no*) in a set of finite clauses with non-clitic direct objects excluding imperatives and *wh*-clauses targeting subject or object (total of

40,120). Some studies focus on tracking specifically base-generated OV orders. For instance, Kroch & Santorini (2014) in their study of the OV decline take into account only some non-finite clauses and exclude cases where the VO order could have been generated from OV by V-to-T or V-to-C movement. In contrast, we are examining the loss of object movement to the left-periphery, that is, to the left of the finite verb, assuming that a finite verb is at least as high as T.[9] That is, disregarding the question about the headedness of the VP, we suggest that the 'old' grammar, inherited from Late Latin, allowed movement of direct objects to the clausal left-periphery, while the new grammar that eventually took over did not allow for this sort of movement and generated only VO sequences.[10] We also assume that the old grammar could generate VO, or V1 ('verb-first'), orders in those cases where the verb moved higher than any of the arguments. This order is illustrated in (5). We assume, for now, that the old grammar generated such orders at some constant rate associated with a particular set of pragmatic conditions.[11] This assumption will be important in the discussion since a VO string is ambiguous as to which grammar generated it.

(5) *Baisset sun chef,*
lowered his head
"He lowered his head." (1100-ROLAND-V,9.112)

For now let us abstract away from the exact structural positions of the arguments and simply look at the distribution of OV/VO sequences over time.

Figure 6 visualises logistic regression models of the $V_{fin}O$ variable for prose, verse and the two forms combined. The slope of the model corresponds to the rate of replacement of the old grammar by the new one, assuming that the new grammar generated only VO while the old one generated OV plus (a constant rate of) VO.[12] Since the rate of 'old' VO is assumed to be constant, it should not matter for the slope comparison.

9. Interestingly, though, the progressions of $V_{nonFin}O$ reported in Kroch & Santorini (2014) and $V_{fin}O$ presented here turn out to be quite similar if we consider prose and verse combined.

10. However, there seems to be no reason to assume that OV was a predominant order even in Late Latin (e.g., Passarotti et al. 2015 and references therein).

11. An underlying assumption here is that the product of the interaction between a given grammar and universal pragmatic laws is constant in the absence of external perturbing factors.

12. Modern French makes use of OV order under very restricted conditions discussed in Abeillé et al. (2008).

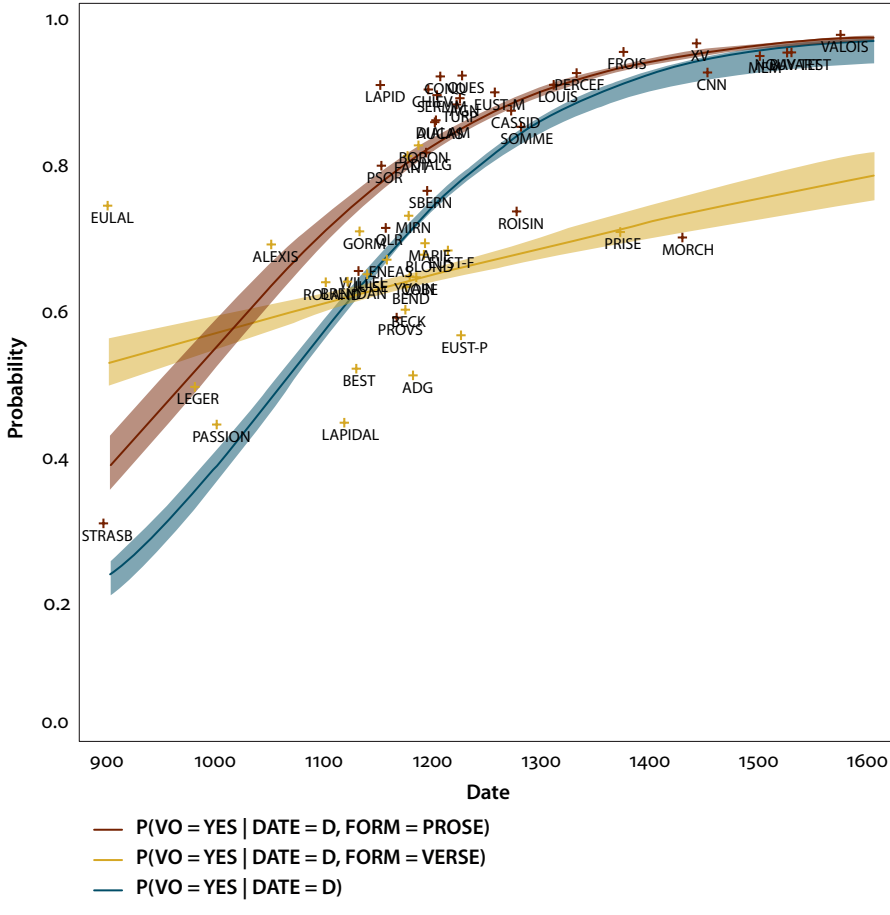


Figure 6. $V_{fin}O$ in prose and verse

The parameter estimates for our model $P(V_{fin}O = \text{yes} \mid \text{Date} = d) = \frac{e^{\alpha + \beta d}}{1 + e^{\alpha + \beta d}}$ are given in Tables 18–20.

Table 18. Logistic regression estimates for $V_{fin}O$ in prose

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−7.0739	0.2963	−23.87	$<2 \times 10^{-16}$
COEFFICIENT	0.0072	0.0002	30.92	$<2 \times 10^{-16}$

Table 19. Logistic regression estimates for $V_{fin}O$ in verse

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−1.5419	0.2952	−5.22	1.76×10^{-7}
COEFFICIENT	0.0017	0.0003	7.10	1.22×10^{-12}

Table 20. Logistic regression estimates for $V_{fin}O$ overall

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−8.6574	0.1796	−48.19	$<2 \times 10^{-16}$
COEFFICIENT	0.0081	0.0001	54.83	$<2 \times 10^{-16}$

While the rate of OV_{fin} for verse is almost constant over time (coefficient close to zero), prose appears to be more advanced than verse in the transition to $V_{fin}O$, at least during the 12th century, for which we have data points for both prose and verse (although verse is still better represented). One way to interpret the logistic regression parameters we obtained is to say that the temporal 'window' available for verse is such that we cannot really observe the decline of OV_{fin} in verse. This would be due to a problem of the corpus text sample, since we know that OV_{fin} ends up disappearing almost completely even from verse.

3.2 Abstract grammar-based analysis

Before we concede that there is an unsurmountable data sampling problem responsible for the difference or that there is actually a significant difference between the rates of change in prose and in verse during the available time window, let us consider another analytical possibility. Recall that our calculations of the rate of change from the 'old' OV to the 'new' VO order involved an assumption that, even though both grammars can generate VO, we count all VO as 'new' assuming that those that are generated by the old grammar (as V1 configurations) constitute a fixed proportion in the overall output of the old grammar at any given point. Thus miscounting them as produced by the new grammar does not affect the rate of the spread of the innovative grammar. That is, even though, because of the added VO counts, the new grammar's probabilities would be "bumped up" at any given time point, this bumping up would be a constant over the whole medieval period and independent of the prose/verse distinction, and thus it could be neglected for the purposes of comparing the overall rates of change in prose and verse. However, if this assumption is wrong, that is, if for some reason the bumping-up effect varies depending on the text form and/or time, this could be a source of non-parallelism between prose and verse. In what follows we show that the original assumption is indeed problematic and that we do need to sort out VOs. The main culprit turns out to be the VO orders with 'true' pro-drop (that is, not cases of ellipsis under coordination), because (non-expletive) pro-drop rates vary depending on the prose/verse distinction (as we show in §2).

To discuss the possible effect of grammar-ambiguous VOs, we need to be more specific about what kind of competing grammars we assume and what orders they can generate, including the position of the subject.

3.2.1 *Grammar A ('old')*

Table 21 gives an overview of the evolution of word order in transitive finite clauses with non-clitic objects.[13] The obvious general trend is the steady increase in SVO at the expense of all other permutations. Another immediate observation is the rarity of OSV and VOS orders, which we therefore exclude from detailed examination.

Table 21. Word order evolution in transitive clauses with non-clitic objects

	OSV	OV	OVS	SOV	SVO	VO	VOS	VSO
1100	0.00 (6)	0.26 (411)	0.08 (127)	0.04 (65)	0.19 (306)	0.40 (649)	0.00 (6)	0.02 (39)
1200	0.00 (60)	0.22 (3756)	0.05 (923)	0.05 (860)	0.32 (5487)	0.29 (4852)	0.01 (175)	0.05 (879)
1300	0.00 (25)	0.06 (343)	0.04 (225)	0.02 (128)	0.50 (2837)	0.28 (1598)	0.01 (36)	0.09 (515)
1400	0.01 (60)	0.05 (390)	0.02 (153)	0.01 (80)	0.56 (4749)	0.28 (2382)	0.01 (62)	0.07 (553)
1500	0.01 (28)	0.02 (92)	0.02 (100)	0.01 (32)	0.66 (3225)	0.25 (1208)	0.00 (19)	0.04 (193)
1600	0.00 (6)	0.01 (18)	0.01 (26)	0.00 (1)	0.74 (1829)	0.21 (516)	0.00 (9)	0.03 (81)

A note is in order on the scope of this investigation. The (evolution of) clausal structure in MF has been the subject of much attention in the literature (Vennemann 1974; Harris 1978; Fleischman 1992; Roberts 1993; Marchello-Nizia 1995; Vance 1997; Lafond 2003; Labelle & Hirschbühler 2005; Rouveret 2004; Mathieu 2006; Labelle 2007; Zaring 2011; Simonenko & Hirschbühler 2012; Kroch & Santorini 2014, to name just a few). Our focus here is limited to the disappearance of preverbal non-clitic objects in transitive finite clauses, and we are concerned only with the position of the main arguments. Most importantly, we are interested in how this change manifested itself depending on the text form, a topic which has not yet been explored at all in a systematic fashion, as far as we know. That is, such issues as the (un)availability of V3 in Old French, the syntax of different subordinate clauses and matrix and embedded questions and many other puzzles of MF syntax are left out of the present picture.

We assume that the old grammar is characterised by an articulated left-periphery which involves an agreement projection, Agreement Phrase (AgrP), as well as (at least) two information structure-related projections, Focus Phrase (FocP) and Topic Phrase (TopP). In the following we briefly discuss our assumptions concerning

13. The reason we excluded pronominal clitic objects is that their syntax even in the earliest texts is already that of verbal clitics, meaning that they are much more syntactically constrained compared to nominal arguments, whereas pronominal subjects do not entirely cliticise until later. Specifically, the position of non-emphatic object pronominals is strictly dependent on the position of the verb: they immediately precede the verb if the verb is not clause-initial, and they immediately follow it when the verb is clause-initial, a generalisation known as the Tobler-Mussafia law. For a detailed corpus-based study of the syntax of object clitics, see Simonenko & Hirschbühler (2012). We also excluded clauses with subject or object wh-dependency because of their idiosyncratic argument syntax.

the structures underlying each surface order.[14] Our eventual goal is to be able to classify as many surface strings as possible as generated by the old or by the new grammar, in order to track the disappearance of the OV-generating old grammar across text forms.

OVS

We begin our inventory of the configurations made available by the old grammar with OVS, (6). We assume that OVS corresponds to the object and subject placement in the Specifier of a discourse-oriented functional projection Topic Phrase (SpecTopP) and the Specifier of the Tense Phrase (SpecTP), respectively. This is accompanied by the movement of the verbal (complex) head to the functional head Agr, which hosts subject person and number features, as in Figure 7.

(6) [*Messe e matines*]$_{obj}$ *ad* [*li reis*]$_{sbj}$ *escultet.*
mass and matines has the king attended
"The king has attended mass and matines." (1100-ROLAND-V,54.647)

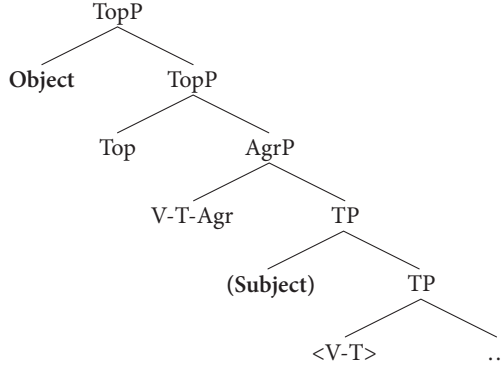


Figure 7. Grammar A generating OVS

Obviously, in order to test the adequacy of this representation, we cannot directly probe the information structure of MF by gathering speakers' judgments. However, as a proxy, we can look at the formal properties of the noun phrases involved, such as the presence/absence and semantics of determiners and modifiers, on the assumption that determiner types correlate with the information-structure statuses of arguments. Specifically, a number of determiners, such as definite and possessive ones, are commonly assumed to trigger presuppositions, that is, contraints on what

14. We abstract away from fine details of the structure below the TP level, such as the presence of modal, aspectual and agent-introducing projections. Angled brackets indicate movement traces and regular brackets indicate the possibility of argument omission.

kind of information a context should entail in order for the utterance in question to be felicitous in that context.

Table 26 in the appendix below gives the distribution of head types in direct object noun phrases in OVS configurations, and Table 27 presents the distribution of determiner types with nominal objects. We put the adjective *tel* "such" in a separate category because of its frequency and special semantics. Noun phrases with such modifiers normally have an antecedent and therefore can be assumed to be demonstrative-like.

Below we compare these results with the determiner distribution in other syntactic configurations and show that there is a remarkably high incidence of demonstratives, both as heads and as pre-nominal determiners. Simonenko (2017) provides a semantic argument as well as arguments from synchronic studies that demonstratives are very likely to be shifted topics and that the position in question was likely associated with prosodic prominence (see also Rainsford 2011: 216 for MF). This corresponds to the Top label of the relevant head in Figure 7.

SOV and OV

Another eventually disappearing configuration is SOV, (7), for which we assume the structure illustrated in Figure 8 where the subject and the object occupy the SpecTopP and the SpecFocP, respectively.

(7) [*Li reis Marsilie*]$_{sbj}$ [*le poign destre*]$_{obj}$ *i perdiet*
the king Marsile the fist right there lost
"The king Marsile lost there his right fist." (1100-ROLAND-V,200.2782)

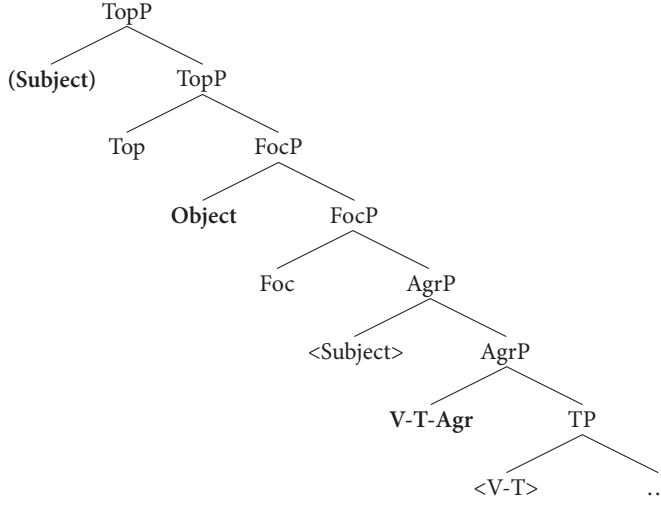


Figure 8. Grammar A generating SOV

Head types and determiner types with nominal objects in this configuration are distributed as in Tables 28 and 29 (see appendix below) respectively, where we find a much lower rate of demonstratives than in OVS configurations.[15]

Under this configuration we subsume OV orders with null subjects. Specifically, we assume that, if they are not contrastive, subjects are null in the old grammar. The distribution of head types and determiners in object phrases is remarkably similar in SOV and OV configurations, as a comparison between Tables 30–31 (see appendix below) on the one hand and Tables 28–29 on the other shows. In fact, if we exclude full object pronouns, the difference in the distribution of the other heads types between OV and SOV is not statistically significant at the 0.05 threshold ($\chi^2 = 3.57$, $df = 3$, $p = 0.31$).[16] Another observation which suggests that OV and SOV should be grouped together in terms of clause structure is a similar rate of possessive determiners, which in both cases is much higher than in OVS orders. This can be viewed as a consequence of the requirement that a possessive pronoun co-indexed with the subject be c-commanded by the latter. Finally, the rate of object pronominalisation is significantly higher in SOV than in OV, or, in other words, when the subject is overt, the object is more likely to be pronominal.[17] Recall, however, that these are non-clitic objects, which means that they were most likely contrastively focused (otherwise a clitic variant would have been chosen), which is reflected in their position in Figure 8.

VSO

We assume that VSO orders, as in (8), have an in-situ object inside of VP and a subject in the canonical subject position in the Specifier of TP, as in Figure 9. The Specifier of the TopP in such configuration is occupied by an indirect object or a non-argument constituent.

(8) De Guenelun atent [li reis]$_{sbj}$ [nuveles]$_{obj}$…
from Guenelun awaits the king news
"The king awaits news from Guenelun…" (1100-ROLAND-V,53.642)

15. The difference is highly statistically significant ($\chi^2 = 84.6$, $df = 1$, $p = 3.53 \times 10^{-20}$).

16. We had to remove free relatives from consideration because the number of observations was too small.

17. This can be related (at least in cases where the subject is nominal) to the first Preferred Argument Structure constraint identified by Du Bois (2003: 34): "Avoid more than one lexical core argument".

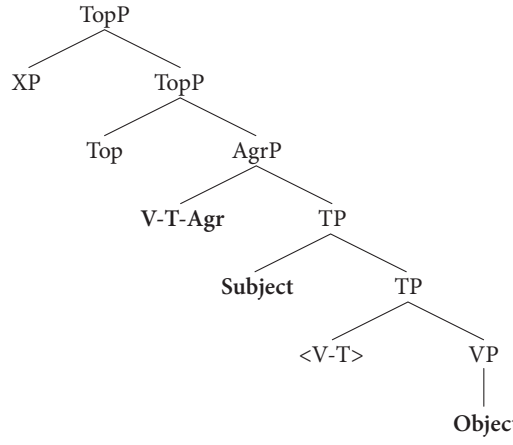


Figure 9. Grammar A generating VSO

The distribution of determiners with the objects is given in Tables 32 and 33 (see appendix below).

SVO and VO

Finally, for the old Grammar A, let us consider the pair SVO and VO. As far as Grammar A goes, we assume that these orders resulted from a structure as in Figure 10. An overt subject occupies the Specifier of the Topic projection. The SVO string, however, is ambiguous, as it could also be the output of the new grammar, as will be illustrated in §3.2.2. In our estimates of the disappearance of OV we will not try to disambiguate SVO and will count them all as the output of the new grammar.

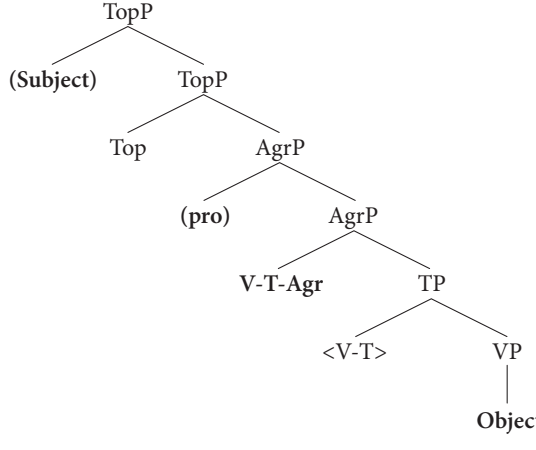


Figure 10. Grammar A generating SVO & VO

3.2.2 *Grammar B ('new')*

We assume that, in contrast to Grammar A, Grammar B lacks an articulated left-periphery and an agreement head. It is also characterised by obligatory subject expression, with the subject by default occupying SpecTP. It is well established that MF underwent verbal agreement syncretisation (Bettens 2015; Buridant 2000; Dees et al. 1980; Foulet 1935; Jong 2006; Marchello-Nizia 1992; Morin 2001; Simonenko et al. 2018). As a result, Modern French finite verbs do not distinguish between 1st, 2nd and 3rd person singular in present indicative. The only subject-less (non-imperative) finite clauses Grammar B generates are those where the subject is elided under coordination with the preceding clause, just like in Modern French. A simple declarative clause with a transitive predicate could thus be schematised as in Figure 11.

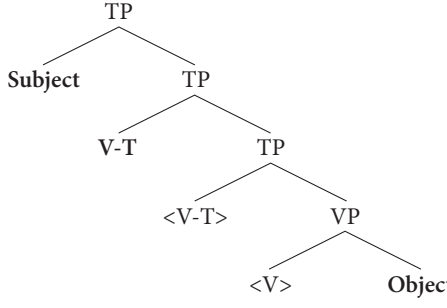


Figure 11. Grammar B generating SVO

3.3 Transition from Grammar A to B

We now classify all strings as generated by Grammar A or B. Seeing the loss of OV as resulting from the loss of a grammar with an extended left-periphery is in line with the tradition of analysing word order changes in MF as reflecting a transition from Topic-initial to Subject-initial utterance organisation (Vennemann 1974; Harris 1978; Marchello-Nizia 1995: 100). Similarly, Labelle & Hirschbühler (2005) suggested that during the medieval period French lost an information structure-related projection in the clausal left-periphery.

String type	Generating grammar
OVS	Grammar A
SOV & OV	Grammar A
VOS	Grammar A
VSO	Grammar A
'true' VO (i.e., subject omitted not under coordination)	Grammar A
'false' VO (i.e., subject omitted under coordination)	Grammar A or B
SVO	Grammar A or B

We are now in a position to model the transition from a grammar with a rich left periphery to a grammar without one as the distribution of a binary variable Grammar with values A and B, where all OVS, SOV, OV, VOS, VSO and true VO are classified as Type A and all SVO as Type B, with false VO being excluded from consideration.[18] We fit the following logistic regression model to our data: P(Grammar = B | Date = d, Form = f) = $\frac{e^{\alpha + \beta d}}{1 + e^{\alpha + \beta d}}$, and the result is visualised in Figure 12.

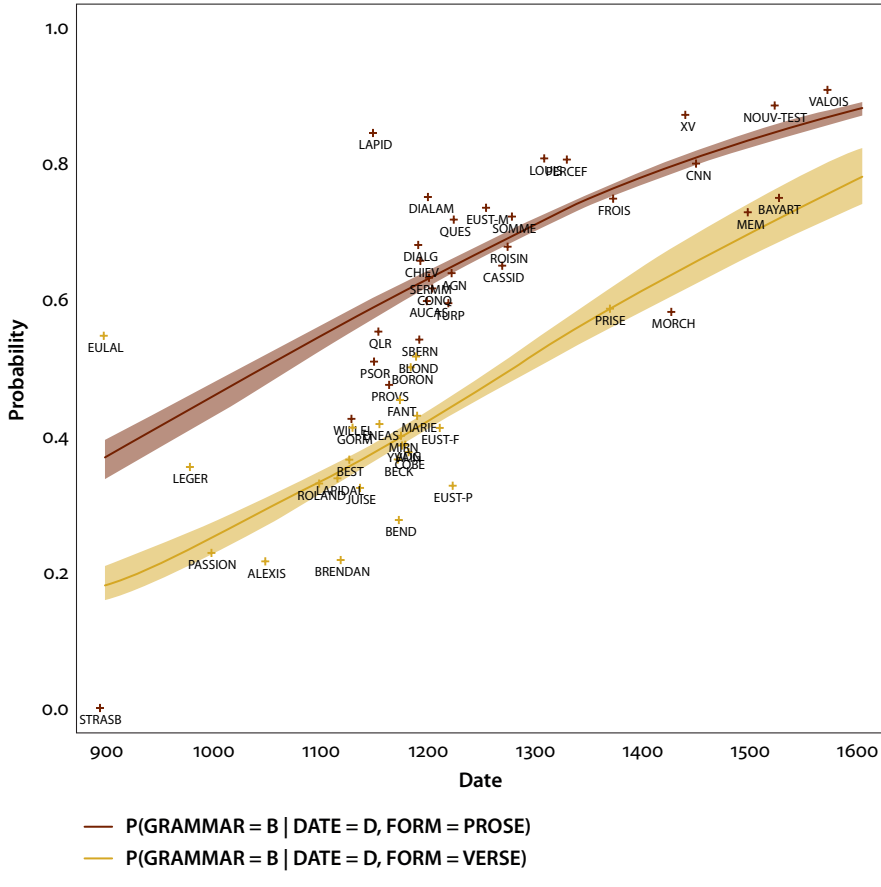


Figure 12. Type B Grammar emergence

18. We count all SVO as generated by Grammar B on the assumption that Grammar A generated such strings at a rate which was stable both across time periods and across text forms.

Table 22. Logistic regression estimates for Grammar B in prose

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−3.860	0.1870	−20.64	$<2 \times 10^{-16}$
COEFFICIENT	0.0036	0.0001	25.87	$<2 \times 10^{-16}$

Table 23. Logistic regression estimates for Grammar B in verse

	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)
INTERCEPT	−5.0760	0.3393	−14.96	$<2 \times 10^{-16}$
COEFFICIENT	0.0039	0.0003	13.89	$<2 \times 10^{-16}$

Estimated this way, the passage to the SVO grammar proceeds in parallel in prose and verse. An explicit testing for the difference in the slope parameter by means of a comparison based on an analysis of deviance of a model where the coefficient parameter can vary depending on the prose/verse distinction with one where the coefficient is not sensitive to these contexts reveals that the prose/verse parameter does not significantly contribute to better predict the data ($\chi^2 = 0.93$, $df = 2$, $p = 0.62$). In other words, the distinction between the rate of change in verse and prose is not statistically significant.

The main change in our estimates compared to just tracking the distribution of OV/VO orders, as in §3.1, is counting true VO as belonging to the same grammar as OV. Before doing a more fine-grained investigation of word-order changes, we speculated that, if the rate of true VO were different in verse and prose because of the difference in the null subject rate, this would have affected our simplistic estimates of the passage from OV to VO, since the rate of the latter would have been disproportionately bumped up in verse. In contrast, classifying true VO as generated by the old Type A grammar results in the rates of change being now very similar across text forms. In other words, given that null subjects are more frequent in verse (see Figure 1), counting all VO including the true subjectless ones as generated by Grammar B leads to an overestimation of the probability of the latter in verse in the first periods, where null subjects were still very frequent. Comparison between Tables 24 and 25 (the only centuries for which there is enough data in both text forms are considered) makes it obvious that the main difference between prose and verse is the relative frequency of true OV and VO orders: the frequency of these orders is higher in verse, but it drops in the 14th century. This is consistent with what we know about the decline of null subjects (and thus true OV and VO), and this, we suggest, is the source of the non-parallelism between prose and verse we initially observed in Figure 6.

Table 24. Word order in transitive clauses with non-clitic objects in prose

	OSV	OV	OVS	SOV	SVO	VO	VOS	VSO
1200	0.00	0.04	0.09	0.08	0.58	0.08	0.02	0.11
1300	0.01	0.02	0.05	0.02	0.67	0.09	0.01	0.13
1400	0.01	0.01	0.02	0.00	0.79	0.06	0.01	0.10

Table 25. Word orders in transitive clauses with non-clitic objects in verse

	OSV	OV	OVS	SOV	SVO	VO	VOS	VSO
1200	0.01	0.18	0.07	0.07	0.40	0.20	0.01	0.06
1300	0.01	0.23	0.05	0.08	0.37	0.22	0.01	0.04
1400	0.00	0.15	0.02	0.07	0.60	0.13	0.00	0.03

4. Conclusions

We have examined two changes affecting different components of the Medieval French grammar across two text forms, prose and verse. First, we quantified the changes as variation in two surface forms, an 'old' and a 'new' variant. For the change in subject expression, that meant quantifying occurrences of null vs. overt subjects and, for the word order change, quantifying instances of OV_{fin} vs. $V_{fin}O$ orders. In both cases this approach revealed a puzzling non-parallelism between verse and prose, namely, either prose or verse would appear almost to stagnate across the medieval period. This does not accord with the obvious fact that in both text forms all changes came to completion much earlier than today's French.

We then shifted from estimating surface form competition to a more abstract modelling of variation as a competition between two grammars for which we assumed a certain mapping between abstract representations and surface forms. For the null subjects case we assumed an old grammar which could generate both null and overt personal pronominal subjects and a new one generating only overt ones. On these assumptions only expletive subjects unambiguously signalled which grammar was used. Estimated as the variation in null/overt expletives, the change progresses in a parallel fashion in prose and verse. These results suggest that, in the grammar allowing for null subjects (the old grammar), the expression of PERSONAL pronominal subjects depends on the text form and, therefore, is not subject to strict grammatical constraints. This is a welcome result given that in modern null subject languages the conditions on the use of overt personal subjects are commonly defined in information-structural or pragmatic terms (e.g.,

aboutness-shift in Italian and Spanish, Frascarelli 2007 and Jiménez-Fernández 2016, respectively) and given that the structuring of discourse depends largely on how the speaker chooses to relate a semantic representation to the utterance context. We further discovered that a major grammatical factor influencing such pragmatic choices is the clause type, matrix vs. subordinate: once we control for it, we see a prose/verse parallelism in the emergence of the overt personal pronominal subject. This suggests that pragmatic factors interact with grammatical choices in a stable way across time, which may be interpreted as an indication of the universality of pragmatic reasoning.

We also find that the difference in personal pronominal subject expression between prose and verse had repercussions for the estimation of the loss of the OV order as a simple competition between OV and VO. A higher rate of null subjects in verse resulted in what seemed like a very early dominance of VO. Once recast in terms of abstract grammars whereby the old grammar could generate subjectless VO sentences (and other argument permutations) and the new one only SVO, we once again see parallel changes in prose and verse.

This project demonstrates that suitably large treebanks make it possible to engage tools of statistical analysis to test some of the traditionally accepted impressionistic and/or intuitive claims in the literature, strengthening the empirical basis of the field.

References

Abeillé, Anne, Danièle Godard & Frédéric Sabio. 2008. Deux constructions à SN antéposé en français. In Jacques Durand, Benoît Habert & Bernard Laks (eds.), *Congrès mondial de linguistique française (CMLF08)*, 2361–2376. Paris: Institut de Linguistique Française.

Adams, Marianne. 1987. From Old French to the theory of pro-drop. *Natural Language & Linguistic Theory* 5(1). 1–32. https://doi.org/10.1007/BF00161866

Agresti, Alan. 2002. *Categorical data analysis*. John Wiley & Sons. https://doi.org/10.1002/0471249688

Alexiadou, Artemis & Elena Anagnostopoulou. 1998. Parametrizing AGR: Word order, V-movement and EPP-checking. *Natural Language & Linguistic Theory* 16(3). 491–539. https://doi.org/10.1023/A:1006090432389

Barbosa, Pilar. 1995. Null subjects. University of Massachusetts, Amherst dissertation.

Bettens, Olivier. 2015. Chantez-vous français ? Remarques curieuses sur le français chanté de Moyen Âge à la période baroque (accessed 12 july, 2018). http://virga.org/cvf/.

Buridant, Claude. 2000. *Nouvelle grammaire de l'ancien français*. Paris: Sedes.

Dees, Anthonij, Steintje Meilink, Karin van Reenen-Stein & Pieter van Reenen. 1980. Un cas d'analogie: l'introduction de -e à la première personne du singulier de l'indicatif présent des verbes en -er en ancien français. *Rapport/Het Franse Boek* 50. 105–110.

Du Bois, John W. 2003. Argument structure: Grammar in use. In J. W. Du Bois (ed.), *Preferred argument structure. Grammar as architecture for function*, 11–60. Amsterdam: John Benjamins. https://doi.org/10.1075/sidag.14.04dub

Dufter, Andreas. 2010. Subordination et expression du sujet en ancien français. In *Actes du XXVe Congrès International de Linguistique et de Philologie Romanes (Innsbruck, 3–8 septembre 2007)*, vol. 2, 443–458.

Fleischman, Suzanne. 1992. Discourse and diachrony: The rise and fall of Old French SI. In Marinel Gerritsen & Dieter Stein (eds.), *Internal and external factors in syntactic change*, 433–73. Berlin: Walter de Gruyter.

Fontaine, Carmen. 1985. Application de méthodes quantitatives en diachronie: L'inversion du sujet en français. Université du Québec à Montréal MA thesis.

Foulet, Lucien. 1928. *Petite syntaxe de l'ancien français*. Paris: Champion, troisième édition revue. Réédition 1982.

Foulet, Lucien. 1935. L'extension de la forme oblique du pronom personnel en ancien français. *Romania* 61–62. 257–315. https://doi.org/10.3406/roma.1935.3758

Franzén, Torsten. 1939. Étude sur la syntaxe des pronoms personnels sujets en ancien français. Uppsala University dissertation.

Frascarelli, Mara. 2007. Subjects, topics and the interpretation of referential pro. *Natural Language & Linguistic Theory* 25(4). 691–734. https://doi.org/10.1007/s11049-007-9025-x

Glikman, Julie & Nicolas Mazziotta. 2013. Répresentation de l'oral et syntaxe dans la prose du *Queste del saint Graal (1225–1230)*. In Dominique Lagorgette & Pierre Larrivée (eds.), *Représentations du sens linguistique 5*, 69–90. Université de Savoie.

Guillot-Barbance, Céline, Bénédicte Pincemin & Alexei Lavrentiev. 2017. Représentation de l'oral en français médiéval et genres textuels. In *Autour du clivage langue parlée/langue écrite du latin au français*, Clermont-Ferrand, France.

Harris, Martin. 1978. *The evolution of French syntax: A comparative approach*. Longman.

Hirschbühler, Paul. 1992. L'omission du sujet dans les subordonnées V1: les CNN de Vigneulles et les CNN anonymes. *Travaux de Linguistique* 24. 25–46.

Jelinek, Eloise. 1984. Empty categories, case, and configurationality. *Natural Language & Linguistic Theory* 2(1). 39–76. https://doi.org/10.1007/BF00233713

Jiménez-Fernández, Ángel. 2016. When discourse met null subjects. *Borealis-An International Journal of Hispanic Linguistics* 5(2). 173–189. https://doi.org/10.7557/1.5.2.3727

Jong, Thera de. 2006. *La prononciation des consonnes dans le français de Paris aux 13ème et 14ème siècles*. Utrecht: LOT/Netherlands Graduate School of Linguistics.

Kaiser, Georg A. 2009. Losing the null subject. A contrastive study of (Brazilian) Portuguese and (Medieval) French. In G. A. Kaiser & E.-M. Remberger (eds.), *Null-subjects, expletives, and locatives in Romance*, 131–156. Konstanz: Fachbereich Sprachwissenschaft, Universität Konstanz.

Kauhanen, Henri & George Walkden. 2018. Deriving the Constant Rate Effect. *Natural Language & Linguistic Theory* 36(2). 483–521. https://doi.org/10.1007/s11049-017-9380-1

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1(3). 199–244. https://doi.org/10.1017/S0954394500000168

Kroch, Anthony & Beatrice Santorini. 2014. On the word order of Early Old French. Handout for a talk at the *7th Conference on Syntax, Phonology and Language Analysis (SinFonIJA 7)*, Graz, Austria.

Labelle, Marie. 2007. Clausal architecture in Early Old French. *Lingua* 117(1). 289–316. https://doi.org/10.1016/j.lingua.2006.01.004

Labelle, Marie & Paul Hirschbühler. 2005. Changes in clausal organization and the position of clitics in Old French. In Montserrat Batllori, Maria-Lluïsa Hernanz, Carme Picallo & Francesc Roca (eds.), *Grammaticalization and parametric variation*, 60–71. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199272129.003.0004

Lafond, Larry. 2003. Historical changes in verb-second and null subjects from Old to Modern French. In D. Eric Holt (ed.), *Optimality Theory and language change*, 387–412. Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-010-0195-3_14

Lagorgette, Dominique & Pierre Larrivée. 2013. *Représentations du sens linguistique 5*. Université de Savoie.

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

Marchello-Nizia, Christiane. 1995. *L'évolution du français. Ordre des mots, démonstratifs, accent tonique*. Paris: Armand Colin.

Marchello-Nizia, Christiane. 2012. L'oral représenté: un accès construit à une face cadrée des langues 'mortes'. In Bernard Combettes, Céline Guillot, Evelyne Oppermann-Marsaux, Sophie Prévost & Amalia Rodrígez Samolinos (eds.), *Le changement en français. Etudes de linguistique diachronique*, 247–264. Bern: Peter Lang.

Marchello-Nizia, Christiane. 2018. De S0 à SV: Vers le sujet obligatoire et antéposé en français, les dernières phases d'un changement. *Journal of French Language Studies* 28(1). 1–19. https://doi.org/10.1017/S0959269517000023

Marchello-Nizia, Christiane & Magali Rouquier. 2012. De (S)OV à SVO en français: où et quand? L'ordre des constituants propositionnels dans la Passion de Clermont et la Vie de saint Alexis. In Monique Dufresne (ed.), *Constructions en changement. Hommage à Paul Hirschbühler*, 111–155. Presses de l'Université de Laval.

Mathieu, Éric. 2006. Stylistic Fronting in Old French. *Probus* 18(2). 219–266. https://doi.org/10.1515/PROBUS.2006.008

Mathieu, Éric. 2009. From local blocking to cyclic agree. In Jila Ghomeshi, Ileana Paul & Martina Wiltschko (eds.), *Determiners: Universals and variations*, vol. 147, 123–158. John Benjamins. https://doi.org/10.1075/la.147.04mat

MCVF. 2010. Corpus MCVF annoté syntaxiquement, (2005–2010), dirigé par France Martineau, avec Paul Hirschbühler, Anthony Kroch et Yves Charles Morin.

Mitchell, Bruce. 1985. *Old English syntax*. Oxford: Clarendon Press. https://doi.org/10.1093/acprof:oso/9780198119357.001.0001

Morin, Yves Charles. 2001. La troncation des radicaux verbaux en français depuis le Moyen Age. Études diachroniques. *Recherches Linguistiques de Vincennes* (30). 63–85.

Passarotti, Marco, Barbara McGillivray & David Bamman. 2015. A treebank-based study on Latin Word Order. In Gerd Haverling (ed.), *Latin Linguistics in the Early 21st Century. Acts of the 16th International Colloquium on Latin Linguistics*, 338–350. Uppsala: Uppsala Universitet.

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

Pintzuk, Susan. 2004. Variationist approaches to syntactic change. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 509–528. Wiley Online Library.

Prévost, Sophie. 2018. Increase of pronominal subjects in Old French: Evidence for a starting-point in Late Latin. In Anne Carlier & Céline Guillot (eds.), *Latin tardif -français ancien: Continuités et ruptures Coll*. Beihefte zur Zeitschrift für romanische Philologie, 169–198. Berlin: De Gruyter.

Rainsford, Thomas Michael. 2011. The emergence of group stress in medieval French. University of Cambridge dissertation.

Roberts, Ian. 1993. *Verbs and diachronic syntax: A comparative history of English and French*. Dordrecht: Kluwer.

Roberts, Ian. 2014. Taraldsen's Generalization and language change: Two ways to lose null subjects. In Peter Svenonius (ed.), *Functional Structure from Top to Toe*, vol. 9 The Cartography of Syntactic Structures, 115–148. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199740390.003.0005

Rouveret, Alain. 2004. Les clitiques pronominaux et la périphérie gauche en ancien français. *Bulletin de la Société de Linguistique de Paris* 99(1). 181–237. https://doi.org/10.2143/BSL.99.1.541914

Schøsler, Lene. 2002. La variation linguistique: Le cas de l'expression du sujet. In Rodney Sampson & Wendy Ayres-Bennett (eds.), *Interpreting the history of French, A festschrift for Peter Rickard on the occasion of his eightieth birthday*, 187–208. Amsterdam: Rodopi.

Simonenko, Alexandra. 2017. Determiners as a probe into diachronic information structure & the Definiteness cycle. Handout for a talk at *Formal diachronic semantics* 2, Saarland University, Saabrücken, November 20–21 (accessed 12 July, 2018). https://www.academia.edu/35249405/Determiners_as_a_probe_into_diachronic_information_structure_and_the_Definiteness_cycle.

Simonenko, Alexandra & Anne Carlier. 2016. The evolution of the French definite article: from strong to weak. Handout for a talk at *Going Romance*. Johann Wolfgang Goethe-Universität, Frankfurt. 08–10/12/2016 (accessed July 12, 2018). https://www.academia.edu/28473104/The_evolution_of_the_French_definite_article_from_strong_to_weak.

Simonenko, Alexandra, Benoît Crabbé & Sophie Prévost. 2018. Agreement syncretisation and the loss of null subjects: Quantificational models for Medieval French. accepted for *Language Variation and Change* (accessed July 12, 2018). http://ling.auf.net/lingbuzz/003491.

Simonenko, Alexandra & Paul Hirschbühler. 2012. Placement de clitiques dans les propositions V1 et évolution de la structure de la proposition en ancien français. In Monique Dufresne (ed.), *Constructions en changement* Les voies du français, 11–53. Sainte-Foye: Presses de l'Université Laval.

Vance, Barbara. 1997. *Syntactic change in Medieval French* Studies in Natural Language and Linguistic Theory. Dordrecht: Kluwer. https://doi.org/10.1007/978-94-015-8843-0

Vennemann, Theo. 1974. Topics, subjects and word order. In J. M. Anderson & Charles Jones (eds.), *Historical linguistics: Proceedings of the first international conference on historical linguistics*, vol. 1, 339–376. Amsterdam: North Holland.

Walkden, George & Kristian Rusten. 2017. Null subjects in Middle English. *English Language and Linguistics* 21(3). 439–473. https://doi.org/10.1017/S1360674316000204

Zaring, Laurie. 2011. On the nature of OV and VO order in early Old French. *Lingua* 121. 1831–1852. https://doi.org/10.1016/j.lingua.2011.07.008

Zimmermann, Michael. 2014. *Expletive and referential subject pronouns in Medieval French*. Walter de Gruyter. https://doi.org/10.1515/9783110367478

Appendix. Head types and determiners with direct objects

Table 26. Head types in object phrases in OVS

HEAD TYPE	
FREE RELATIVE	0.03 (5)
NOUN	0.47 (735)
PERSONAL PRONOUN	0.03 (53)
PRONOUN WITH A CP-COMPLEMENT	0.09 (147)
DEMONSTRATIVE	**0.37** (584)
PROPER NOUN	0.02 (30)

Table 27. Determiners with nominal objects in OVS

DETERMINER	
DEFINITE	0.19 (139)
DEMONSTRATIVE	**0.21** (155)
tel	0.09 (65)
POSSESSIVE	0.07 (50)
QUANTIFIER	0.15 (111)
INDEFINITE	0.02 (16)
PARTITIVE	0.01 (6)
ZERO	0.26 (193)

A zero determiner in MF is not to be equated with indefiniteness. The spread of overt determiners was another change that progressed gradually over the medieval period (e.g., Simonenko & Carlier 2016), and in the earlier texts bare nouns occurred frequently in the contexts which in Modern French require a definite determiner, a demonstrative or a possessive pronoun (Mathieu 2009).

Table 28. Head types in object phrases in SOV

HEAD TYPE	
FREE RELATIVE	0.002 (2)
NOUN	0.78 (913)
PERSONAL PRONOUN	0.06 (70)
PRONOUN WITH A CP-COMPLEMENT	0.05 (60)
DEMONSTRATIVE	0.05 (59)
PROPER NOUN	0.05 (62)

Table 29. Determiners with nominal objects in SOV

DETERMINER	
DEFINITE	0.25 (234)
DEMONSTRATIVE	0.06 (54)
tel	0.02 (20)
POSSESSIVE	0.19 (170)
QUANTIFIER	0.1 (92)
INDEFINITE	0.03 (26)
PARTITIVE	0.001 (1)
ZERO	0.34 (316)

Table 30. Head types in object noun phrases in OV

HEAD TYPE	
FREE RELATIVE	0.003 (17)
NOUN	0.83 (4158)
PERSONAL PRONOUN	0.01 (54)
PRONOUN WITH A CP-COMPLEMENT	0.06 (281)
DEMONSTRATIVE	0.06 (284)
PROPER NOUN	0.04 (216)

Table 31. Determiners with nominal objects in OV

DETERMINER	
DEFINITE	0.26 (1086)
DEMONSTRATIVE	0.02 (91)
tel	0.02 (82)
POSSESSIVE	0.2 (845)
QUANTIFIER	0.14 (572)
INDEFINITE	0.03 (142)
PARTITIVE	0.005 (21)
ZERO	0.3 (1319)

Table 32. Head types in object phrases in VSO

HEAD TYPE	
FREE RELATIVE	0.01 (27)
NOUN	0.84 (1909)
PERSONAL PRONOUN	0.00 (1)
PRONOUN WITH A CP-COMPLEMENT	0.1 (215)
DEMONSTRATIVE	0.01 (22)
PROPER NOUN	0.04 (86)

Table 33. Determiners with nominal objects in VSO

DETERMINER	
DEFINITE	0.25 (478)
DEMONSTRATIVE	0.03 (64)
tel	0.02 (39)
POSSESSIVE	0.17 (330)
QUANTIFIER	0.14 (278)
INDEFINITE	0.04 (81)
PARTITIVE	0.01 (32)
ZERO	0.31 (607)

Table 34. Head types in object phrases in SVO

HEAD TYPE	
FREE RELATIVE	0.01 (167)
NOUN	0.84 (15652)
PERSONAL PRONOUN	0.002 (30)
PRONOUN WITH A CP-COMPLEMENT	0.08 (1534)
DEMONSTRATIVE	0.02 (417)
PROPER NOUN	0.034 (633)

Table 35. Determiners with nominal objects in SVO

DETERMINER	
DEFINITE	0.27 (4293)
DEMONSTRATIVE	0.03 (578)
tel	0.02 (248)
POSSESSIVE	0.16 (2611)
QUANTIFIER	0.11 (1738)
INDEFINITE	0.04 (687)
PARTITIVE	0.02 (287)
ZERO	0.33 (5210)

Appendices

Appendices 1 and 2 can be found online at: https://doi.org/10.1075/dia.00008.sim.additional

Spoken Latin behind written texts

Formulaicity and salience in medieval documentary texts

Timo Korkiakangas
University of Oslo

This study uses treebanking to investigate how spoken language infiltrated legal Latin in early medieval Italy. The documents used are always formulaic, but they also always contain a 'free' part where the case in question is described in free prose. This paper uses this difference to measure how ten linguistic features, representative of the evolution that took place between Classical and Late Latin, are distributed between the formulaic and free parts. Some variants are attested equally often in both parts of the documents, while perceptually or conceptually salient variants appear to be preserved in their conservative form mainly in the formulaic parts.

Keywords: treebank, salience, documentary Latin, legal Latin, Early Middle Ages, scribe, diplomatics, formulaicity, language acquisition

1. Introduction and objectives

In this paper, I use syntactically annotated data made available in a dependency treebank to explore how spoken-language features find their way into written texts in historical text corpora with conservative text genres. This is an important question because, with historical language data, the extent to which the written surface reflects the reality of the spoken language is usually unknown. At the same time, the very same texts are often the only material available for tracking spoken-language developments. This study discusses the subject by quantitatively examining the mechanisms that determined why certain spoken-language features crept into the documentary Latin of early medieval Italy – and why others did not. In this way, the study seeks to establish methods for using written historical texts for the study of spoken language. The aim is thus methodological, but the analyses of specific constructions chosen to illustrate the approach also shed new light on legal Latin data.

https://doi.org/10.1075/bct.113.05kor

The methodology involves a corpus study to isolate linguistic features that are sensitive (or insensitive) to the formulaicity of the documents. All spoken and written communication relies heavily on prefabricated fixed or semi-fixed expressions, i.e., formulae (MacKenzie & Kayman 2018). In early medieval Italian documents, 'formulae' refers to standard phrases and clauses that recur in multiple documents of the same type. These phrases and clauses, which guaranteed the juridical validity of the contents, can be identified by comparing documents with each other. In traditional diplomatic terminology, formulaic phrases are called, for example, invocation, inscription and corroboration (Guyotjeannin et al. 1993: 72–85). Yet, documents also contain one or more non-formulaic slots where the distinctive characteristics of the matter at hand, such as the extent of a property being sold, are described in detail. Sabatini (1965) observed that the language utilised in these slots differs considerably from that of the centuries-old formulae.

Documentary Latin, a term used, for example, by Sabatini, represents an ideal data set for a treebank study because it allows us to contrast linguistically conservative formulae with linguistically innovative non-formulaic passages, which are assumed to draw on the early medieval spoken idiom. This kind of examination of two evolutionary stages of Latin, i.e., conservative features derived from Classical Latin and innovative features developed by the 8th/9th century AD, is an indirect way of addressing diachronic variation within a relatively synchronic treebank (see §2).

The theoretical framework adopted here highlights the role of salience, a factor that is used below to determine the distributions of conservative and innovative forms and constructions. In this endeavour, findings of second language acquisition studies prove to be helpful because it can be argued that the scribes, native speakers of an early Romance vernacular, learnt written documentary Latin as a second language (L2). Conceptual salience is defined as the cognitive prominence of a (syntactic) construction.

The following section (§2) describes the data utilised, followed by a section (§3) which defines formulaicity. After that, §4 discusses the theoretical background and research setting while §5 presents the linguistic features to be examined. The quantitative results are presented and interpreted in §6: §6.1 explains how formulaicity and salience are related to each other and §6.2 analyses each linguistic feature in terms of the two notions. §7 is the conclusion.

2. Data

This study utilises the Late Latin Charter Treebank (LLCT), a syntactically annotated corpus of original private documents, i.e., charters, from Central Italy (519 texts, c. 226,000 words; Korkiakangas & Passarotti 2011).[1] The treebank method allows for the study of all the grammatical domains from lexicon to syntax, surpassing the possibilities provided by simple morphologically tagged text corpora. The LLCT documents, consisting mainly of contracts about transferring property, were written in historical Tuscia, a region which comprises most of modern Tuscany, between 714 and 869. The time span is considered too short to enable a normal diachronic approach. However, the method is ideal for contrasting conservative classical and innovative Late Latin features.

LLCT is based on three freely available copyright-free diplomatic editions and is available for download (see LLCT and online appendix for queries). In terms of morphological and syntactic annotation, LLCT is based on the Ancient Greek and Latin Dependency Treebank (AGLDT) style codified in the *Guidelines for the syntactic annotation of Latin treebanks* (Bamman et al. 2007). Since documentary Latin is a non-standard variety that often contains ambiguous morphological and syntactic features, Korkiakangas & Passarotti (2011) introduce a set of additions and modifications to the *Guidelines*, designed for Classical Latin.

In early medieval Italy, scribes did not copy documents from model document collections, as was done later in the Middle Ages, but rather reproduced the wording of documents from memory (Schiaparelli 1933: 3). This, together with the fact that classical standards were obviously not strictly required for documentary texts, led to considerable linguistic variation, fruitful for variationist and diachronic studies of the linguistic situation at a time when the transition from Latin to an Italo-Romance vernacular must have been well advanced to all appearances in spoken language. Indeed, documentary Latin is a variety of non-standard Latin that has several features proven to originate from the spoken language of the time, either through direct reflection or indirectly by way of misinterpretations of classical legal Latin (Korkiakangas 2016: 240). Obviously, the term 'non-standard' calls for a definition of 'standard'. No standardisation of language in the modern sense of the word was practised, i.e., the grammar and spelling were not canonised by the authorities. Nevertheless, the essentially classical orthography and morphology seem to have still served as a valued model for the best-written LLCT texts. Thus, there was rather a clear idea of a standard, in terms of a substantial consensus about 'correct' or 'accepted' morphology and spelling (Korkiakangas 2016: 36).

1. LLCT is currently being enlarged with ca. 200,000 extra newly annotated documentary Latin words (LLCT2).

3. Formulaicity

Documentary texts consist of two parts: formulaic expressions and the so-called disposition, the declarative part, in free prose. As already noted, formulaic expressions guaranteed the juridical validity of early medieval Italian documents. Most formulae date from the imperial Roman chancery tradition. Sabatini (1965) emphasised that each document also contains non-formulaic case-specific parts which record non-universal features, such as descriptions of the transferred property or ownership central to the current legal act. This information primarily lies in the disposition, but may be scattered within and between the very formulae. Given that the scribes could not rely on prefabricated phrases when composing the case-specific descriptions, there appears a distinct difference in language between these so-called free parts and the formulaic parts which were anchored to the age-old legal tradition, alien to the everyday language. The free parts understandably have recourse to the spoken idiom, while the formulaic parts reflect the spoken language only by way of hypercorrections.

Sabatini relied upon the free/formulaic dichotomy to support his theory on the formation of the Italian plural forms (Sabatini 1965: 978–987). My hypothesis is that this distinction is indicative of linguistic change and variation on other levels of linguistic representation as well. So far, the lack of any annotated corpus of documentary Latin has made large-scale quantitative research impossible. To overcome this, I have provided LLCT with annotation that distinguishes the free parts, most notably the disposition, from the formulaic parts, i.e., the rest of the document (Korkiakangas & Lassila 2013). The following quotes from a sales contract (*CDL* 26, Lucca, March 720) illustrate a typical free sentence from the disposition (1) and a typical formulaic sentence called *sponsio* (2). The quotes show that the distinction between free and formulaic is not clear-cut: both sentences contain both free and formulaic elements. The free/non-formulaic words are underlined.

(1) *Consta me Aufrid v(ir) d(evotus) hanc die vendedisset et vendedi, tradedisset et tradedi vobis Aunuald, Teutpald, Leonaci, Petronaci, Teutp(er)t, Dommuli, Vuilifrid, Nandulo, Geminiano clerico, Teuderaci ortu meum quem avire videor ante s(an)c(t)o Selvestre, qui latere tene prope curte vel orte s(an)c(t)i Selvestri, rectu casa Domnici vel de filio Iovanni.*
"It is manifest that I, Aufrid, *vir devotus*, in the present day sell and hand over to you Aunuald, Teutpald, etc., my orchard which I am known to have in front of the church of Saint Sylvester and which has its border close to the court and garden of Saint Sylvester, by the house of Domnicus and his son Iohannes."

(2) *Et, sicot non crido, ut si ego aut eridis meus vos molestaverimus aut da qualivet homine vobis defensare non potuero, spondeo vobis cunponere dupla condicionem*. "And, which I do not believe, if I or my heirs molest you or if I cannot defend you from whoever man, I promise to compensate double the price."

Although free parts usually contain some formulaic elements and vice versa, the formulaicity status is assigned in LLCT to each sentence because otherwise syntactic features, which operate on a sentence level, become difficult to examine (Korkiakangas 2016: 25–29).

4. Theoretical background and research setting

I have selected ten linguistic features traditionally assumed to reflect language changes in Late Latin. The features are described in §5, and §6 shows which of these features are sensitive and which are insensitive to formulaicity. My working hypothesis is that, if a certain innovative feature is in a statistically significant way more frequent in the free parts of documents, it is more likely to reflect the current state of the spoken language. Conversely, a conservative feature is expected to occur more frequently in the formulaic parts. The other logical possibility, that an innovative feature is more frequent in the formulaic parts or a conservative feature in the free parts, did not happen with the features examined.

To find out which types of features are sensitive to formulaicity, the ten features were chosen to cover the grammatical landscape broadly. To examine all kinds of features from lexicon to syntax within a single framework, I adopt here a cognitive view of grammar in terms of a syntax-lexicon continuum (Table 1). The syntax-lexicon continuum is a uniform model of grammatical representation which locates constructions on a continuum according to their generality. Atomic means that an item cannot be further divided into meaningful parts unlike in complex

Table 1. The syntax-lexicon continuum (Croft & Cruse 2004: 255; Broccias 2012: 738)

Rank	Grammar domain/Construction type	Traditional name	Example
5	Complex and (mostly) schematic	Syntax	noun verb noun (= transitive construction)
4	Complex and (mostly) specific	Idiom	pull one's leg
3	Complex but bound	Morphology	noun-s
2	Atomic and schematic	Word class	pronoun, adjective
1	Atomic and specific	Word/lexicon	*this, green*

constructions, whereas a schematic construction subsumes specific constructions, like *adjective* subsumes *green*.[2]

The selected linguistic features concern the following stages or construction types, as they are called in the construction grammar tradition, of the syntax-lexicon continuum: atomic and specific (lexicon), complex with bound morphemes (morphology) and complex and schematic (syntax). This ordering of grammatical domains partly overlaps with the classification of morphemes into free and bound lexical morphemes and free and bound functional morphemes (Croft & Cruse 2004: 254–256). Lexical morphemes carry meaning by themselves (e.g., *dog*), whereas functional morphemes (e.g., *of*) specify the relationship between other morphemes. Free morphemes are free-standing words (e.g., *dog, of*) while bound morphemes occur only as part of other words (e.g., *form* in *transform*, *-s* in *dogs*). It needs to be emphasised that the syntax-lexicon continuum is a simplification of a complicated linguistic reality, like all organisational schemes intended to capture the whole of grammar.

Importantly for the present study, the ranking inherent in the syntax-lexicon continuum (see Column 1 in Table 1) can be considered to reflect the cognitive effort involved in recognising the linguistic domains in question, at least under certain conditions. The assumption is that, roughly speaking, a higher ranking means greater mental effort required by a language learner to adopt the features that pertain to that domain, due to the higher complexity of these features. This higher complexity is assumed to result from the higher-ranking construction types being generalisations based on a large number of exemplars and lower-level generalisations. This is expected to apply principally to language-learning situations, not to the language processing of L1 language users in general. These assumptions are in harmony with the overall picture sketched by the studies on the L2 acquisition order of grammatical categories: lexical morphemes appear to be acquired before functional morphemes and, within each of these groups, free morphemes are acquired before bound ones (Zobl & Liceras 1994: 172–175; Goldschneider & DeKeyser 2001: 28). An effectively similar picture arises from the processability-informed theories of language acquisition, which assume a complexity-based processing hierarchy: learners are supposed to first acquire the relations between lemmas, then those within words (lexical morphology), within phrases, between phrases and, finally, between clauses (Pienemann 1999: 7–9). I exploit the complexity ranking combined with the morpheme classification (free/bound) to account for the conceptual salience in this study (see §6).

2. For the theoretical motivation and a detailed definition of the terminology, see Croft & Cruse (2004: 247–256).

These features all have one variant associated with Classical Latin and another associated with Late Latin or Italo-Romance, e.g., GENITIVE CASE FORM VS. PREPOSITIONAL PHRASE WITH *DE*, respectively. I call these diachronic variants 'conservative' and 'innovative', respectively. Although the starting point and the endpoint of the development are known, the chronology often remains uncertain: it is not always clear to what extent the innovative variant has established itself and ousted the conservative variant in the spoken language. For example, it is known that the replacement of the genitive case form by the prepositional phrase with *de* was a gradual process which took centuries and arguably was not yet fully completed by the time of LLCT (Valentini 2017: 47ff.).

Defining conservative and innovative variants is not only problematic with respect to diachrony. The need to define variant pairs often leads to forced dichotomies which cannot take into account various nuances connected with register or other preconditions, amply examined in various studies (e.g., Valentini 2017 for genitive/*de* + PP). With some features, such as the dative case form, the conservative variant was not replaced by any single innovative construction, but rather by a plethora of (partly) new means of expressing recipiency. In these cases, no complementary distribution between conservative and innovative variant can be established, and only the conservative variant lends itself to quantification. Although this variant cannot be meaningfully compared to any other, its relative frequency in the total word count can be calculated.

5. Linguistic features

The examined features are presented in the order they appear in Table 2 (§6).

1. NON-CLASSICAL LEMMAS: 81 innovative lemmas, e.g., *barba* "uncle", *cambium* "exchange", *fossatum* "ditch", *petia* "piece" (see the full list provided in the online appendix). This feature is not part of the analysis proper but is introduced as a "calibration variable". Lexicon cannot contribute genuinely to the present analysis because vocabulary is not optional, like grammatical variants, but is determined by the propositional content of the phrase. For example, several non-classical agricultural words, such as *tessero* "to mark boundaries with signs" or *cavallarius* "horsekeeper", are relevant in free parts, where the highly varying case-specific traits of the transferred property are described. Instead, formulaic parts establish the universal legal circumstances of the act and cultivate classical technical terminology, such as *indictio* "dating cycle" or *confirmatio* "confirmation". Thus, formulaic and free parts seem to consist of different vocabulary, which is assumed to be visible in the formulaicity distribution of

conservative and innovative lemmas. This is not the case with morphological and syntactic features, where the distribution of conservative and innovative variants is not predetermined by the propositional content but is expected to depend on the different prestige attributions between formulaic and free parts. The innovativeness of the lemmas has been verified by cross-checking them in Lewis & Short's *Latin dictionary*[3] and in Du Cange's *Glossarium mediae et infimae Latinitatis*.[4]

2. FUTURE PERFECT FORM: e.g., *apparuerit* "it/he/she will have appeared". The classical future perfect survives in Romance only sporadically and mainly in idiomatic expressions (Lausberg 1962: 205–206; cf. Weber 1924: 60–62, who calls the form 'conditional'). In documentary Latin, the form is utilised to anticipate a future execution of what was agreed on by the contracting parties.
3. DATIVE PLURAL IN *-BUS*: e.g., *potestatibus* "to dominions". Along with the general decline of the Latin case system, the dative was increasingly replaced by other means of expressing recipiency, such as the prepositional phrase with preposition *ad* "to" (Adams 2013: 278–294). Since the *ad* + PP is utilised in LLCT for a plethora of adnominal relationships of ambiguous interpretation, I could not calculate the relative frequency of the dative form and *ad* + PP. Thus, instead I counted the percentage of the dative form in the total number of words. The Romance personal pronominal system retains the indirect object forms as clitics (Salvi 2011: 322–324). Therefore, pronouns are excluded from this investigation. Only the 3rd, 4th and 5th declensions with the phonologically/graphically substantial (see § 6.2) ending *-bus* are examined.
4. ADNOMINAL POSSESSION: genitive case form vs. prepositional phrase with *de*, e.g., *regis* vs. *de rege* "of a/the king". The *de* + PP competed with the original possessive strategy, the genitive case, and, in the Romance languages, became the main means of expressing adnominal possession (Valentini 2017; Adams 2013: 267–274). The majority of the possessive constructions are still expressed by the case form in both formulaic and free parts of LLCT.
5. PHRASAL COMPLEMENTATION: accusative plus infinitive (ACI) vs. complementiser clause, e.g., *Sichiprandum*.ACC *scribere*.INF *rogavi* vs. *rogavi ut*.COMP *Sichiprandus scriberet* "I asked S. to write". In Latin, the innovative complement clause, introduced by a complementiser, had long rivalled the accusative and infinitive construction, the principal means of complementation in Classical Latin. The Romance languages have generalised the complementiser pattern (Zamboni 2000: 119–120; Ledgeway 2012: 244ff.). Considered a classical

3. Available through the Perseus Word Study Tool (http://www.perseus.tufts.edu/hopper/morph?la=la).

4. http://ducange.enc.sorbonne.fr/

prestige feature, ACI is the prevailing complementation strategy in LLCT, likely because it was part of some documentary formulae.

6. ABSOLUTE CONSTRUCTIONS: e.g., *iuvante Deo* "God willing". It is agreed that the Latin absolute constructions were typical of (classical) literary texts and hardly occurred in spoken language (Väänänen 1981: 166–168). They continued to be utilised as stylistic prestige features in LLCT, apart from being part of certain formulae, such as *regnante* + the name of the king "under the reign of N.". Here only constructions with participles are examined.
7. SECOND-PERSON SINGULAR: form with classical *-s* vs. without *-s*, e.g., *teneas* vs. *tenea* "you should hold". The classical 2nd-person singular ending of all the active indicative and subjunctive tense forms, except for the perfect, was *-s*. All the word-final consonants were either lost or weakened by the Early Middle Ages in the spoken language of Italy (Väänänen 1981: 67–69; Adams 2013: 132ff.), and the second-person singular of the subjunctive came to end in /a/ while, in the indicative, the outcome was /i/ (Maiden 1996).
8. DATIVE SINGULAR: e.g., *potestati* "to dominion". Like the dative plural, the dative singular form was being replaced by other constructions. See (3) above.
9. SUBJECT CASE ENCODING: nominative vs. accusative subject, e.g., *ista portio*. NOM *sit in potestate tua* vs. *ista portionem*.ACC *sit in potestate tua* "let this parcel be in your possession". In Late Latin, the accusative is known to have extended partially to the subject function as a symptom of a major reorganisation of grammatical relations, i.e., so-called semantic alignment, where the Agent-like arguments tended to be encoded with the nominative and the Patient-like arguments with the accusative (Ledgeway 2012: 328–335; Korkiakangas 2016: 57–74). Here, only those non-pronominal 3rd-declension subjects are counted where the morphological contrast between the nominative (*portio*) and the accusative (*portione*(*m*)) remained intact in Late Latin for phonological reasons (Korkiakangas 2016: 111). The few person and place names were also excluded because their case endings are often ambiguous.
10. VERB/OBJECT ORDER: e.g., *casam donavit* (OV) vs. *donavit casam* (VO) "he/she/it donated a house". The most typical verb/direct object order in Classical Latin was OV, although much variation existed. As time passed, the originally mostly pragmatically conditioned Latin order became increasingly syntactically motivated and gradually turned into the VO order predominant in the Romance languages. OV still remained frequent for a long time and obviously had stylistic overtones (Ledgeway 2012: 225–235; Zamboni 2000: 101–102). Here, only clauses with one non-coordinated finite verb and non-pronominal direct object are examined because they are prototypical and unambiguous. In LLCT, clauses with coordinated verbs tend to be long and complex, with verbs occurring before and after direct object(s). This leads to an ambiguity

about which verbs the object(s) belong(s) to (Korkiakangas 2016: 196). Main and subordinate clauses are treated equally. Pronominal objects are discarded because they have peculiar syntactic characteristics of their own, such as the relative pronoun's typical clause-initial position.

6. Results and their interpretation

Table 2 presents the examined features in two groups according to whether they appear to be sensitive to the formulaic/free distinction or not. For each feature that only has a conservative variant, its share in the total of words of LLCT is presented (parts per thousand values, ‰). For the features which allow the identification of both conservative and innovative variants, the distribution of the conservative and

Table 2. The examined features with their relative frequencies in formulaic and free parts of documents

Statistically significant sensitivity to formulaicity		**N**	**‰ in total of words**		**% distribution**		**Sig. level**
Domain	**Measured variant**		**Formulaic**	**Free**	**Formulaic**	**Free**	
lexicon	81 innovative lemmas	767	1.5	9.1	–	–	$p < 0.001$
morphology	future perfect form	2,315	12.3	3.5	–	–	$p < 0.001$
	dative plural form	154	0.8	0.4	–	–	$p = 0.004$
morphology/ syntax	adnominal genitive form	8,027	37.2	30.6	90.3	69.2	$p < 0.001$
	adnominal *de* + PP	1,441	4.0	13.6	9.7	30.8	
syntax	ACI	982	5.6	1.8	84.7	57.8	$p < 0.001$
	conjunction clauses	235	0.9	1.3	15.3	42.2	
	absolute constructions	916	4.9	1.5	–	–	$p < 0.001$
Statistically non-significant sensitivity to formulaicity		**N**	**‰ in total of words**		**% distribution**		**Sig. level**
Domain	**Measured variant**		**Formulaic**	**Free**	**Formulaic**	**Free**	
morphology	2nd person singular *-s*	248	1.0	1.4	89.0	87.6	n.s.
	2nd person singular not *-s*	32	0.1	0.2	11.0	12.4	
	dative singular form	3,878	15.8	21.4	–	–	n.s.
syntax	nominative subjects	278	1.2	1.3	74.8	65.8	n.s.
	accusative subjects	107	0.4	0.7	25.2	34.2	
	OV order	1,118	3.8	8.5	66.4	62.5	n.s.
	VO order	611	1.9	5.1	33.6	37.5	

innovative variant is presented (percentages). The formulaic parts contain 169,520 words and the free parts 56,314 words. The statistical significance is the *p* value of the chi-squared test (95% confidence interval). When calculating the chi-squared test for those variables that present only the conservative variant, the frequencies are compared to thousands of words. N indicates the population size, i.e., the number of occurrences.

The variables NON-CLASSICAL LEMMAS, FUTURE PERFECT FORM, DATIVE PLURAL IN *-BUS*, ADNOMINAL POSSESSION, PHRASAL COMPLEMENTATION and ABSOLUTE CONSTRUCTIONS show a statistically significant dependence with the formulaicity variable. No statistically significant dependence is attested between formulaicity and the variables SECOND-PERSON SINGULAR, DATIVE SINGULAR, SUBJECT CASE ENCODING and VERB/OBJECT ORDER.

6.1 Formulaicity and salience

This and the following section interpret the results of the quantitative analysis. The results seem to support the intuitive postulate adopted by earlier scholarship, namely, that the scribes did draw from different linguistic repositories when writing formulaic and free parts of documents (Sabatini 1965). This becomes evident on the basis of lexicon. NON-CLASSICAL VOCABULARY was utilised as a 'calibration variable'. It was thought that, if the lexicon variable showed a clear difference between free and formulaic parts, it would be reasonable to carry out the formulaicity analysis with other, more complex domains of grammar. This assumption proves to be sound on the basis of the numbers of Table 2: the innovative lemmas occur six times more often in free parts than in formulaic parts.

Apart from lexicon, formulaicity is likely to affect the distributions of other features as well. The question is whether there is something common to all the features that show a statistically significant difference between formulaic and free parts in Table 2. The numbers reveal that the complexity ranking based on the syntax-lexicon continuum alone is not enough to explain the behaviour of the examined features because the same linguistic domains may be sensitive or insensitive to formulaicity. The concept of salience becomes useful at this point. A widely used notion in semiotics, social psychology and sociolinguistics, salience is a gradient property which operates on the physical world/cognition interface. In linguistics, salience can refer to the characteristics of the linguistic input/output itself or to those external-world factors that cause some parts of the input/output to become salient, such as the referent of the linguistic expression being bright-coloured or interesting to the language user (Cintrón-Valentín & Ellis 2016). In this paper, salience is understood in terms of how prominent or noticeable certain lexical items,

morphemes or syntactic constructions appear to a language learner in the linguistic input. Several quantitative and experimental L2 acquisition studies focus on the role of salience in language acquisition. For example, the eye-tracking measurements of Cintrón-Valentín & Ellis (2016) prove that the low perceptual salience of certain short inflexional morphemes essentially contributes to L2 learners' difficulty in learning them.

These findings based on modern language learning situations can be fruitfully extrapolated to early medieval Latin. I suggest that the statistically significant features of Table 2 involve a variant which is either conceptually (in terms of its grammatical prominence) or perceptually (in terms of phonetic or graphic substance) more salient than those which do not show a statistically significant formulaicity distribution. In other words, the features with statistically significant sensitivity to formulaicity are salient forms or constructions in terms of one or the other of the mentioned salience types, or, if they involve two or more variants, at least one of the variants is salient.

The criterion of perceptual salience, the amount of formal prominence or noticeability in terms of phonetic/graphic substance, seems to apply to the lowest ranking domains of the syntax-lexicon continuum, here words and morphemes. Most words are, as such, phonetically/graphically perceptible units that convey lexical meaning, whereas in morphology, the grammatical information is carried by morphemes of differing phonetic/graphic perceptibility. This is not the case with the highest ranking domain, i.e., syntax, where it often makes no sense to speak about perceptual salience. In syntactic constructions, free-standing words or phrases are linked to each other by an underlying rule, and in Latin each word involved is usually encoded by a certain bound morpheme. Thus, the salience of a variant of a syntactic construction must rather be thought of as conceptual salience, which I define here as the grammatical prominence or noticeability of the syntactic rule to the language learner. I have adopted the term 'conceptual' to distinguish the just-described salience from 'semantic' salience, a vague term utilised to cover a vast variety of uses from the prominence of discourse referents to that of extra-linguistic entities (e.g., Chiarcos et al. 2011: 1–3).

To give an example, a construction is more salient conceptually if it involves free morphemes instead of bound morphemes, as in the case of COMPLEMENTISER CLAUSE VS. ACCUSATIVE AND INFINITIVE. It is true that in this case both variants can also be considered perceptually salient given that they consist of (multiple) free-standing words. Instead, the rule conditioning the SUBJECT ENCODING is considered non-salient because it involves only bound morphemes and is thus less noticeable. As regards perceptual salience, it may emerge, among other things, from the amount of phonetic substance, stress level or usual serial position in a sentence (Dulay & Burt 1973: 409). In the case of written texts, graphic substance,

i.e., characters, may determine salience, especially when the phoneme/grapheme relationship is weak, as it certainly was between written and oral codes in early medieval Italy.

Given that the early medieval Tuscan scribes very likely spoke a variety which might already be described as a Romance vernacular, they learnt documentary Latin in practice as a second language. Indeed, the innovative and conservative features examined here can be seen as characteristics of the scribes' native L1 and of the literary Latin L2 to be learnt, respectively.[5] The L2 studies about the acquisition order of grammatical morphemes lend this study a useful framework which also seems to be extendible to predominantly syntactic features. Goldschneider & DeKeyser (2001) have shown that the acquisition order of certain English morphemes is largely explained by perceptual salience, semantic complexity, morphophonological regularity, syntactic category and frequency. The authors claim that these five factors all constitute aspects of salience in a broad sense of the word (Goldschneider & DeKeyser 2001: 35).

The early medieval documentary scribes had imbibed the basics of Latin spelling and morphology when learning to write, but that was not enough for writing legal documents. They also had to adopt the formulae, and this was likely done by reading existing legal documents. The imperfect command of certain formulaic passages indicates that many scribes had memorised the formulae rather superficially, without profound comprehension of them. However, Classical legal Latin enjoyed a high prestige as the language of law.[6] The scribes knew they were supposed to use this venerable variety when writing documents, especially the formulae. I suggest that, during and after the original memorisation process, the learners who were to become scribes recognised the perceptually or conceptually salient conservative features more readily as Classical Latin forms than the less salient ones and remembered to reproduce those more often in formulaic parts, which were considered vital for legal validity. This is also likely to work the other way around: the salient innovative features, which were felt to belong to the spoken language, were recognised as stigmatised, i.e., having a kind of negative prestige, and consequently avoided in the formulaic parts. The perceptually or conceptually non-salient features, instead, went unnoticed by the scribes and failed to be attributed a (positive or negative) prestige.

---

5. On the other hand, the picture is complicated by the fact that Latin was still apparently considered the expected written form of the language people spoke at the time. On the metalinguistic change between Latin and vernacular, see Wright (1991).

6. For the challenges involved in the reconstruction of prestige patterns in historical language varieties, see Sairio & Palander-Collin (2012) and Adams (2013: 841ff.).

As a consequence, non-salient features appear to be distributed evenly (i.e., due to chance) between formulaic and free parts. The drive to recognise and imitate words and expressions that were considered echoes of the dignified legal tradition and correct grammar seems to result from the scribes experiencing external or internal normative pressures, especially when writing the formulaic parts. Apparently, this aspiration to classical grammar was a common phenomenon even though Italian documentary Latin seems to have been a recognised genre sanctified by the long traditions and not subject to similar corrective interventions of the authorities, as it was in Carolingian Gaul (Bartoli Langeli 2006: 28ff.).

6.2 Analysis of the morphological and syntactic features

I now examine the evidence to justify the above theoretical considerations. The distributions of the variants of the morphological features examined here seem to be plausibly explained by their perceptual salience, i.e., by the amount of phonetic/graphic substance of the feature or of one involved variant with respect to the other variant. The DATIVE SINGULAR form *potestati* "to possession" differs only by one character from, for example, the genitive singular form *potestatis* or from the accusative singular form *potestate*(*m*).[7] Since the case system had already largely collapsed and the use of the dative and genitive was, in all likelihood, no longer supported by the spoken language, it is probable that a form like *potestati* with the dative morpheme *-i* did not stand out enough from the paradigm where the most common form must have been something like *potestate*. This accusative-based form had possibly become the nearly all-purpose or default form in late spoken Latin and competed, perhaps, only with the nominative form *potestas*. It then became the only form of the noun in the Romance languages.[8] Assuming thus that *potestat-* was the most typical stem of the paradigm, morphemes resulting in forms such as *potestati* or *potestate* cannot be considered perceptually salient.

Instead, the DATIVE PLURAL form *potestatibus* "to possessions" differs more clearly from the other case forms of the word: the ending *-*(*i*)*bus* is both graphically[9] and phonologically more substantial than, for example, the respective singular ending *-i*. Indeed, the enduring prestige of *-bus* is witnessed by its hypercorrect use in subject and object function (see also Sornicola 2012: 57–58), as in (3).

---

7. Word-final vowels were likely to be vague in Late Latin and the same applies to the *-s* as well. The final *-m* had ceased to be pronounced very early on (Adams 2013: 62, 128–147).

8. For the two-case system, see Zamboni (2000: 110–115) and Korkiakangas (2016: 74–79).

9. However, the two last letters were sometimes abbreviated in handwriting by a loop resulting roughly in *-*(*i*)*b*^{9}.

(3) […] *ut p*(*er*) *singulos annos ego, dum vixero, et s*(*upra*)*s*(*crip*)*ti nepotes mei vel heredib*(*us*) *eor*(*um*) *dare et reddere debeam*(*us*) *ad ep*(*iscopu*)*m* […] *unum sol*(*idum*) (*CDL* 285)
"[…] so that every year I, as far as I live, and my mentioned descendants as well as their heirs have to give to the bishop […] one *solidus*"

The same explanation also applies to the 2nd-person singular and the future perfect forms. The SECOND-PERSON SINGULAR morpheme *-s* in the form *teneas* "you should hold" is only one phonologically and graphically non-substantial sound/character and is not particularly discernible from the most frequent forms of the same paradigm *teneat* (3rd person) and *teneam* (1st person). This minor distinction does not seem to have been enough to guarantee the scribes' attention and the subsequent recognition of the form's classical prestige in a situation where all the singular persons of the subjunctive were likely to end in /a/ in the spoken language. Instead, the FUTURE PERFECT affix *-er*(*i*)*-*, with the phonologically persistent *r*, forms an entire syllable and results, thus, in clearly more substantial forms, such as *apparuerit* "he/she/it will have appeared". These forms were easily associated with prestige. The future perfect leaps out from the verbal paradigm as one of the graphically longest inflexional forms, along with the subjunctive pluperfect. In other words, a single morpheme is perceptually salient if it stands out from the horizon of expectation consisting of the most common or typical forms of the paradigm.

How about features which involve two or more variants? I argue that for the formulaicity distribution to be statistically significant, at least one of the variants has to be salient, perceptually or conceptually. If the conservative variant is perceptually salient, the same mechanism applies as with the morphological features above. Instead, if only the innovative variant is salient, it is avoided because it is recognised as stigmatised. Note that motivations of this kind can be perceived only as statistical tendencies because several simultaneous conflicting motivations are involved as well. Some scribes were more aware of classical grammar than others, and specific linguistic features were given highly idiosyncratic prestige attributions. Therefore, the distributions of the innovative and conservative features are nowhere near fully determined by formulaicity, as can be seen in Table 2. What is important is the statistically significant difference in relative frequency between the formulaic and free parts.

In the case of ADNOMINAL POSSESSION, i.e., the genitive form replaced by a prepositional phrase with *de*, an average genitive case form (except for the infrequent genitive plurals) is rather non-salient perceptually due to its minor phonetic/graphic substance, whereas the *de* + PP, which consists of the preposition and its complement, is easier to notice both perceptually, because it is two words, and conceptually, because the words are two free morphemes (instead of a bound one in the genitive) (e.g., Zobl & Liceras 1994: 172). So, although a learner may not

have recognised the non-substantial genitive form as a classical prestige form and, indeed, may not even have been aware of the two variants being in complementary distribution with each other in terms of prestige, he may have been able to induce a rule to avoid the innovative *de* + PP because it did not occur in prestigious texts or it had been defamed by the school master. In this way, scribes could learn to shun the salient innovative variants when writing. However, especially in the free parts of documents, where the scribes had to resort to their own linguistic instinct rather than to ready-made formulae, they sometimes let the PP creep in. Of course, several LLCT scribes did not succeed well in attributing stigma to the PP given that the PP is found in 9.7% of cases in the formulaic parts.

The case of PHRASAL COMPLEMENTATION is essentially the same, although here both variants, the infinitival construction and the complementiser clause, are syntactic constructions that involve several words. I maintain that the innovative variant, the complementiser clause, is here also the more salient one. This is because the structure of the complementiser clause only consists of free morphemes, consequently rendering the structure less abstract. The accusative plus infinitive construction, instead, is based on the syntactic interplay of the bound morphemes in the subject noun and infinitive. In any event, the attribution of prestige status is enabled. ABSOLUTE CONSTRUCTIONS are salient only conceptually, although they may involve words which are perceptually salient *per se*, but this is not relevant for its recognition as a prestigious syntactic construction. On the other hand, it has to be remembered that the range of the absolute constructions attested in LLCT is lexically limited.

Let us now look at the two syntactic features that show no statistically significant sensitivity to formulaicity, i.e., SUBJECT CASE ENCODING and verb/object order. The statistical non-significance indicates that the use of the syntactic cases nominative and accusative as the case forms of the subject is not dependent on formulaicity. The morphological difference between the 3rd-declension nominative and accusative forms that were examined, such as *portio* and *portione*(*m*), might be perceptual *per se*, although attributing the salience status to either of the two forms would be difficult.[10] What is relevant, however, is that the underlying principle of the subject encoding (alignment of the arguments of the verb according to semantic or syntactic criteria) is particularly abstract and difficult to grasp and, consequently, not conceptually salient. Intuitively, the rule that conservatively assigns the nominative case to all the subjects of finite verbs, rather than only to semantically active ones (the Late Latin way), is not to be learnt as easily as, for example, the rule 'remember to put the genitive and not the *de* + PP'. In addition,

10. The salient one is perhaps *portio* because it differs by its stem from the rest of the paradigm and, moreover, *portione* was probably the all-purpose form of the day.

the rule of subject case encoding involves only bound morphemes, which keeps the variants perceptually non-salient.

As with subject encoding, learning a VERB/OBJECT ORDER which differs from that of one's native tongue also calls for a profound understanding of syntactic functions, a matter hardly promoted by school teaching. Indeed, syntactic issues seem to have passed largely unheeded in the Latin grammatical tradition. The linearisation of the verb and the object complement is abstract and involves bound morphemes to encode the constituent, so I consider it conceptually non-salient.[11] On the other hand, OV was still clearly the prevalent order in LLCT. Therefore, it can be asked to what extent the Romance-type VO had spread in the spoken language. Perhaps the crucial stabilisation of the VO order took place only after the period examined here. Indeed, in many Late Latin texts, the word order still essentially follows the same pragmatic constraints as in earlier Latin (Spevak 2010). The possibility cannot be excluded, however, that the persistently favoured classical clause-final position had kept the OV order perceptually salient, at least to some scribes and with certain substantial verb forms (Ledgeway 2012: 229ff.; Dulay & Burt 1973). All this said, the verb/object order is perhaps not as felicitous an indicator as one would have wished.

7. Conclusion

This study has examined the role of formulaicity in the distribution of conservative and innovative linguistic features in documentary Latin. The scribes had memorised the formulaic parts, which were considered the juridical heart of the document. Thus, the scribes reproduced many conservative, classical forms and constructions predominantly in these parts and, correspondingly, avoided using innovative spoken-language features in them. However, this sensitivity to formulaicity seems to be limited to features that the scribes recognised as prestigious or non-prestigious. The results of this study support a view that the recognition of prestige, i.e., a feature being Classical Latin, required a certain type of prominence from the linguistic variants. Based on the analysis of ten linguistic features, I have argued that this prominence can be assimilated to perceptual and conceptual salience, the former being salience in terms of phonetic/graphic substance and the latter in terms of noticeability of the underlying grammatical (syntactic) rule.

11. Note that, before the stabilisation of the syntactically motivated (S)VO order, the Late Latin word order was also likely to be affected by the semantic realignment of grammatical relations, mentioned with the subject case encoding (Ledgeway 2012: 335–336; Korkiakangas 2016: 212–216).

I have argued that the formulaicity distribution of the features seems to be related to the cognitive prominence of those features (syntax-lexicon continuum and free/bound morphemes), such that the morphological features are explained by perceptual salience and the syntactic features by conceptual salience. According to this view, the domains of grammar that rank the highest on the syntax-lexicon continuum involve abstract and, thus, unnoticeable syntactic rules. The scribes, native speakers of a Romance-type variety, often did not recognise these due to lack of syntactically-informed education (Black 2001: 64–70), hence the non-salient syntactic constructions' statistically non-significant distributions between formulaic and free parts.

By clarifying the role of salience, this study has identified one mechanism that makes it possible to examine spoken language-related features in conservative written genres. The salience approach can also be applied to assessing the status of language use in other historical treebanks, provided that they have been written by non-native speakers. All the same, it would be desirable to verify the validity of the salience approach delineated here on a larger and more varied repertoire of linguistic features in a further study. In this way, it could be evaluated whether the concept of perceptual and conceptual salience can still be reduced to other, even simpler linguistic motivations: whether there is a systematic association between perceptually and conceptually salient features and, for example, phonetically/phonologically motivated and semantically motivated language change, respectively.

The results highlight the use of treebanks for historical linguistics. In particular, philological annotation, such as that concerning the free/formulaic parts in LLCT, makes it possible to subject relatively well-known data sets to detailed quantitative analysis that would be unimaginable without treebanking. At the same time, this study exemplifies how essentially synchronic data can be used for diachronic research.

Acknowledgements

I am deeply grateful to the reviewers for their valuable comments. I also thank Dr. Hilla Halla-aho for commenting on a previous version of this article.

Appendix

The online appendix is available at https://doi.org/10.1075/dia.00009.kor.additional

References

Adams, James Noel. 2013. *Social variation and the Latin language*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511843433

Bamman, David, Marco Passarotti, Gregory Crane & Savine Raynaud. 2007. *Guidelines for the syntactic annotation of Latin treebanks* (v. 1.3). http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf (3 June, 2017.)

Bartoli Langeli, Attilio. 2006. *Notai: scrivere documenti nell'Italia medievale*. Roma: Viella.

Black, Robert. 2001. *Humanism and Education in Medieval and Renaissance Italy*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511496684

Broccias, Cristiano. 2012. The syntax-lexicon continuum. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford handbook of the history of English*, 735–747. Oxford: Oxford University Press.

Chiarcos, Christian, Berry Claus & Michael Grabski. 2011. Introduction: Salience in linguistics and beyond. In Christian Chiarcos, Berry Claus & Michael Grabski (eds.), *Salience: Multidisciplinary perspectives on its function in discourse*, 1–28. Berlin: Gruyter. https://doi.org/10.1515/9783110241020

Cintrón-Valentín, Myrna C. & Nick C. Ellis. 2016. Salience in second language acquisition: Physical form, learner attention, and instructional focus. *Frontiers in Psychology* 7. 1284. https://doi.org/10.3389/fpsyg.2016.01284

Croft, William & Alan D. Cruse. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511803864

Dulay, Heidi C. & Marina K. Burt. 1973. Should we teach children syntax? *Language Learning* 23. 245–258. https://doi.org/10.1111/j.1467-1770.1973.tb00659.x

Goldschneider, Jennifer M. & Robert M. DeKeyser. 2001. Explaining the 'natural order of l2 morpheme acquisition' in English: A meta-analysis of multiple determinants. *Language Learning* 51. 1–50. https://doi.org/10.1111/1467-9922.00147

Guyotjeannin, Olivier, Jacques Pycke & Benoît-Michel Tock. 1993. *Diplomatique médiévale*. Paris: Brepols.

Korkiakangas, Timo. 2016. *Subject case in the Latin of Tuscan charters of the 8th and 9th centuries*. Helsinki: Societas Scientiarum Fennica.

Korkiakangas, Timo & Matti Lassila. 2013. Abbreviations, fragmentary words, formulaic language: Treebanking medieval charter material. In Francesco Mambrini, Marco Passarotti & Caroline Sporleder (eds.), *Proceedings of the third workshop on annotation of corpora for research in the humanities*, 61–72. Sofia: Bulgarian Academy of Sciences.

Korkiakangas, Timo & Marco Passarotti. 2011. Challenges in annotating Medieval Latin charters. *Journal of Language Technology and Computational Linguistics* 26. 103–114.

Lausberg, Heinrich. 1962. *Romanische Sprachwissenschaft, II: Formenlehre*. Berlin: Gruyter.

Ledgeway, Adam. 2012. *From Latin to Romance: Morphosyntactic typology and change*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199584376.001.0001

LLCT = Late Latin Charter Treebank. Available in pml.xml format at: https://doi.org/10.5281/zenodo.1197357

MacKenzie, Ian & Martin A. Kayman (eds.). 2018. *Formulaicity and creativity in language and literature*. London: Routledge.

Maiden, Martin. 1996. On the Romance inflectional endings *i* and *e*. *Romance Philology* 50. 147–182.

Pienemann, Manfred. 1999. *Language processing and second language development: Processability theory*. Amsterdam: John Benjamins. https://doi.org/10.1075/sibil.15

Sabatini, Francesco. 1965. Esigenze di realismo e dislocazione morfologica in testi preromanzi. *Rivista di Cultura Classica e Medievale* 7. 972–998.

Sairio, Anni & Minna Palander-Collin. 2012. The reconstruction of prestige patterns in language history. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The handbook of historical sociolinguistics*, 626–638. Chichester: Blackwell. https://doi.org/10.1002/9781118257227.ch34

Salvi, Giampaolo. 2011. Morphosyntactic persistence. In Adam Ledgeway, Martin Maiden & John C. Smith (eds.), *The Cambridge history of the Romance languages, vol. 1: Structures*, 318–381. Cambridge: Cambridge University Press.

Schiaparelli, Luigi. 1933. Note diplomatiche sulle carte longobarde II: Tracce di antichi formulari nelle carte longobarde. *Archivio Storico Italiano* 19. 3–34.

Sornicola, Rosanna. 2012. Bilinguismo e diglossia dei territori bizantini e longobardi del Mezzogiorno: le testimonianze dei documenti del IX e X secolo. *Quaderni dell'Accademia Pontaniana* 59. 1–102.

Spevak, Olga. 2010. *Constituent order in Classical Latin prose*. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.117

Väänänen, Veikko. 1981. *Introduction au latin vulgaire*. Paris: Éditions Klincksieck.

Valentini, Cecilia. 2017. L'evoluzione della codifica del genitivo dal tipo sintetico al tipo analitico nelle carte del Codice diplomatico longobardo. Firenze: Università degli Studi di Firenze dissertation.

Weber, Shirley Howard. 1924. *Anthimus, De observatio[ne] ciborum: Text, commentary, and glossary, with a study of the Latinity*. Leiden: Late E.J. Brill.

Wright, Roger. 1991. The conceptual distinction between Latin and Romance: Invention or evolution. In Roger Wright (ed.), *Latin and the Romance languages in the Early Middle Ages*, 103–113. University Park: The Pennsylvania State University Press.

Zamboni, Alberto. 2000. *Alle origini dell'italiano: dinamiche e tipologie della transizione dal latino*. Roma: Carocci.

Zobl, Helmut & Juana Liceras. 1994. Functional categories and acquisition orders. *Language Learning* 44. 169–180. https://doi.org/10.1111/j.1467-1770.1994.tb01452.x

Subject index

Index of languages

Index of authors

Over the last few decades, the widespread diffusion of digital technology has increased availability of primary textual sources, radically changing the everyday life of scholars in the humanities, who are now able to access, query and process a wealth of empirical evidence in ways not possible before.

Also for ancient languages, corpora enhanced with increasingly complex layers of metalinguistic information, such as part-of-speech tagging and syntactic annotation (called 'treebanks') are now available. In particular, diachronic treebanks, which provide data for a language across several historical stages of a given language, allow for a new approach to diachronic studies of syntactic phenomena where scholars previously had to content themselves with empirical work on a much smaller scale.

This volume brings together a set of papers that report research on various diachronic matters supported by evidence from diachronic treebanks. The contents of the papers cover a wide range of languages, including English, French, Russian, Old Church Slavonic, Latin and Ancient Greek. Originally published as special issue of *Diachronica* 35:3 (2018).

ISBN 978 90 272 0798 2

John Benjamins Publishing Company