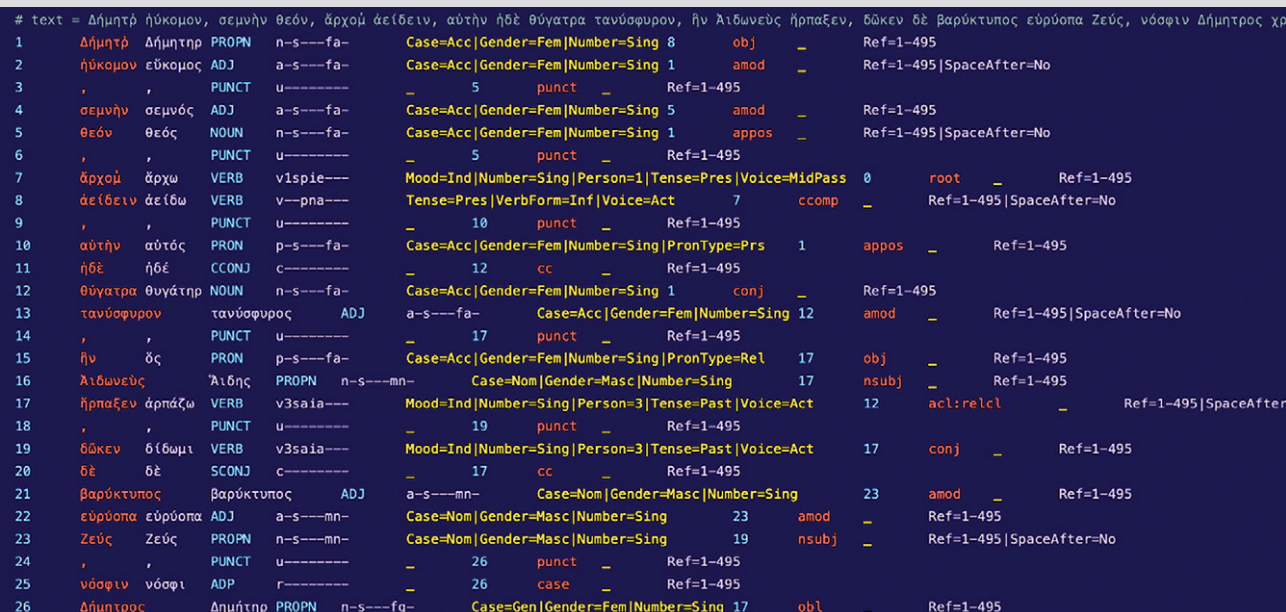
Edited by
ERICA BIAGETTI · CHIARA ZANCHI · SILVIA LURAGHI

Building New Resources for Historical Linguistics

PaviaUniversityPress



Edited by
ERICA BIAGETTI · CHIARA ZANCHI · SILVIA LURAGHI

Building New Resources for Historical Linguistics

PaviaUniversityPress

Il volume è stato pubblicato grazie a finanziamento MIUR nell'ambito del bando PRIN 2015 n. 20159M7X5P_002 «Transitivity and argument structure in flux», PI Silvia Luraghi (Pavia).

Copertina: Cristina Bernasconi, Milano
Impaginazione: Alberto Bellanti, Milano

Via Salasco, 5 - 20136 Milano
Tel. 02/5836.5751 - Fax 02/5836.5753
egea.edizioni@unibocconi.it - www.egeaeditore.it

Pavia University Press
info@paviauniversitypress.it – www.paviauniversitypress.it

Prima edizione: novembre 2021
ISBN volume 978-88-6952-132-4
ISBN ebook 978-88-6952-141-6

Stampa: Logo S.r.l. – Borgoricco (PD)

Table of contents

Introduction: The Value of Digital Resources for Historical Linguistics

Erica Biagetti*, Chiara Zanchi*, Silvia Luraghi*

1. *Ancient languages: from corpora to digital resources*

Ancient languages have a long tradition of literary, philological, and linguistic research, based on data gathered from collections of written texts. Because these data derive from historical records of ancient languages, the latter are known as *Korpussprachen* 'corpus languages' (see, among others, Mayrhofer 1980; Untermann 1983; Haug 2015: 187; Eckhoff et al. 2018b: 300). In this context, the term *corpus* is understood in a broad sense as "a body of naturally occurring language" (McEnery et al. 2006; for a stricter definition of corpus, see for example Sinclair 2005). In this tradition of corpus studies, the digital turn is represented by Father Busa's *Index Thomisticus*, a pioneering electronic collection of all words contained in Thomas Aquinas' *opera omnia* (Busa 1980; Nyhan and Passarotti 2019). From the second half of the 1960s, Father Busa started systematically collecting words of Thomas Aquinas' Latin, initially on punched cards and later on magnetic tapes. The printed version of the *Index* dates back in 1980 and consists of 56 volumes, while a CD-ROM version was released later in 1989.[1] Since Father Busa's time, the efforts to digitalize various linguistic and non-linguistic records of ancient cultures have increased and constitute a subfield of what has come to be known as Digital Humanities (henceforth DH, see Schreibmann et al. 2004, among many others).

DH by definition constitutes an interdisciplinary enterprise, as it is placed at the intersection of digital technologies and the humanities. The creation of digital archives, the use of quantitative methods, and the aim to construct linguistic tools are some among the goals that characterize DH. These goals extend to all traditional disciplines of humanities, such as history, philosophy, linguistics, literature, art, and archaeology, also allowing communication between different humanistic sub-fields (Drucker 2013). Furthermore, DH incorporates both digitized (i.e. remediated) and

* University of Pavia.

1. Currently, a morphosyntactically annotated version of the *Index Thomisticus* is maintained at the Catholic University of the Sacred Heart in Milan, and is being integrated with other linguistic resources available for Latin in the framework of the *LiLa* Project (on which, see Passarotti and Mambrini in this volume).

born-digital materials, as well as both ancient and contemporary phases of a given cultural heritage, consequently encouraging collaboration between diachronic and synchronic traditions of study.

All papers collected in this volume contribute to the large field of DH, as they all deal with the design, the creation, or the development of linguistic resources for historical linguistics. Over the past decades, historical linguistics has witnessed the creation of a large number and variety of linguistic resources, such as digitized and annotated corpora, lexica, databases, and tools for automatic linguistic analysis. To mention but a few large projects engaged in the digitization and encoding of ancient texts, the *Göttingen Register of Electronic Texts in Indian Languages* (GRETIL) is a platform providing standardized machine-readable texts in Indian languages that have been contributed by various institutions. The *Perseus Digital Library* represents the largest collection available to date of texts from Ancient Greek, Latin, and Arabic literature; in addition, it comprises Germanic, 19th-century American, Renaissance materials, issues of the Richmond Times Dispatch, and Humanist and Renaissance Italian Poetry in Latin. An even broader purpose prompted the creation of the *Thesaurus Indogermanischer Text- und Sprachmaterialien* (TITUS), which, as suggested by its name, is a collection of digitized texts covering all branches of Indo-European.

With the increasing number of digitized texts available on the Internet and the possibility to automatically derive them from source texts, the electronic versions of ancient records started being enriched with metadata in the forms of mark-up and/or annotation. Mark-up concerns the textual document as a whole and can be of a stylistic, philological and/or archeological nature: for example, dates and literary genres are annotated in *The Diorisis Ancient Greek Corpus*. Information concerning the place of discovery, editions, textual issues, and hand attribution is associated with the digitalized version of the Mycenean tablets contained in DĀMOS (for additional examples and issues related to mark up, see Section 2).

In contrast, annotation adds linguistic information to text chunks of various length, allows performing fine-grained queries of particular language phenomena, and serves the purpose to help the development of language processing tools. Annotation can concern different levels of linguistic analysis: morphology (with lemmatization and POS tagging), syntax, semantics, and pragmatics. Lemmatization groups together all word forms under a single lemma, thus allowing for queries of different inflected forms without the need of regular expressions. Two other morphological layers, part-of-speech and morphological tagging, allow looking for various combination of morphological features at the word level (to some extent, derivational morphology is also annotated in the corpora of the PROIEL family, on which see Eckhoff and Haug in this volume).

Part-of-speech tags make it possible to query the syntactic distribution of lexical categories. As examples of morphologically annotated corpora of ancient languages, the *Project Wulfila* is a relatively small digital library dedicated to the study of Gothic and Old Germanic languages in general, which provides morphologically annotated editions linked to a digital glossary, POS-tags, and interlinear translations. The corpus MIDIA (*Morphology of Italian in DIAcrony*) is a collection of texts written in Italian which is completely lemmatized and POS-tagged. The selection of the corpus, which extends from the beginning of the 13th to the first half of the 20th century, and the incorporated search tools were designed especially for the study of word formation in Italian from a diachronic point of view, but can also be used for other types of linguistic research.

Syntactic annotation (or syntactic parsing), according to which syntactically parsed sentences are represented and stored as syntactic trees in treebanks, allows looking for groups of words that are syntactically related, even if they are not close to each other in the sentence linear order (Eckhoff et al. 2018b). Semantic and pragmatic annotation adds metadata concerning semantic roles, other semantic information (e.g., animacy of event participants and the semantic class of verbs), and information structure (this information is partly annotated in the corpora of the PROIEL family; Eckhoff and Haug in this volume). Finally, a number of corpora contain genre-specific annotation: for example, *The Chicago Homer*, collecting all literary texts that have been attributed to Homer, is annotated for formulas of the Early Greek epic; *VedaWeb*, on top of morphological metadata, stores metrical annotation of the *RigVeda*, the most ancient Indo-Aryan textual record.

Among corpora enriched with linguistic annotation, treebanks are the most informative ones, providing exhaustive syntactic analysis on top of other annotation layers such as lemmatization, part-of-speech tagging and morphological analysis, and being in turn sometimes enhanced with semantic or pragmatic information (Eckhoff et al. 2018a). Some treebanks of historical languages are synchronic in that they aim to represent one specific stage of the language. If treebanks are enlarged with different texts from different stages of the language, one talks instead of diachronic treebanks. When mature enough and provided with valuable annotation, these treebanks allow researching the scope and effects of diachronic developments and permit large amount of data to be evaluated through statistical methods.

In the past decades, a number of treebanks for historical languages have been created, including Sanskrit, Ancient Greek, Latin, Gothic, Old English, Old Church Slavic, and Hittite among others. While some of these treebanks employ the same annotation schemes as treebanks of modern languages – especially the two *de facto* standards represented by the Penn Treebank phrase-structure format and the Prague Dependency Treebank format –, others developed annotation schemes of

their own or modified existing schemes in order to increase expressivity (Eckhoff et al. 2018a).

For instance, the Ancient Greek and Latin Dependency Treebank 2.0 (AGLDT; Bamman and Crane 2011) is a multi-layered dependency treebank, whose architecture is modelled on the *Prague Dependency Treebank* of Czech. In the AGLDT 2.0, metadata is structured and stored in separated but interlinked morphological, analytical (i.e., syntactic), and tectogrammatical (i.e., semantic/pragmatic) layers. The analytical layer contains dependency syntactic trees and feeds the tectogrammatical layer, which comprises semantic role-labelling, information structure, and anaphora/ellipsis resolution, annotated according to the Praguian linguistic tradition of the *Functional Generative Description* (Sgall et al. 1986). The treebanks of the PROIEL family, instead, were annotated according to a Dependency Grammar scheme enriched with elements of the Lexical Functional Grammar (Bresnan et al. 2015; Eckhoff et al. 2018b). The central aim of the PROIEL project, whose extended name is *Pragmatic Resources in Old Indo-European Languages*, was to establish a parallel treebank of the oldest Indo-European New Testament translations, in order to study how these languages express information structure, that is, their lexical and/or syntactic means to mark such categories as old and new information, contrast, parallelism, topicality and others (Haug 2008). In order to fulfil the aim of this project, the texts were not only provided with morphological and syntactic annotation, but also tagged for a number of different other features known to be relevant for information structure systems, most notably givenness status and anaphoric relations. Since its beginning, the project expanded so as to include languages that were not necessarily/only represented by translations of the Gospels, such as Old Norse, Old Norwegian and Old Swedish (see Eckhoff and Haug in this volume for further details).

Annotated corpora are acknowledged to have several advantages for research purposes (Eckhoff et al. 2018b: 303–304). In the first place, collections of annotated texts have not been assigned metadata with specific research aims, which prevents the risk of circularity. Moreover, the mostly widespread annotation schemes are in principle designed according to non-specific linguistic theories that enjoy large consensus within the linguistic community (Haug 2015). Nor, theoretically, are annotation schemes designed with language-specific features in mind, which means that, in principle, they are portable to large language samples to make linguistic resources comparable among one another.

In the second place, annotated resources allow scholars to automatically retrieve large, and virtually exhaustive, quantitative evidence on linguistic phenomena whose account has been previously based on qualitative evidence and/or partial datasets (Eckhoff et al. 2018b: 303). Furthermore, morphosyntactically annotated corpora require automatic data selection through clear and explicit, though often quite com-

plex, query expressions. This is a crucial factor that makes historical linguistic research replicable (Haug 2015).

In the third place, quantitative evidence has been acknowledged as able to sometimes yield unexpected results, counterintuitive to linguists' naked eye, and to capture and/or prove correlation among linguistic phenomena (Biber 2009; Anthony 2013; for a case in point, concerning the development of configurationality in Ancient Greek and Latin, see Ponti and Luraghi 2018). This happens because morphosyntactically annotated corpora allow linguists to operationalize hypotheses and statistically test correlations.

Beyond facilitating linguistic resarch, annotated corpora allow extracting different types of information from which new linguistic resources can be created. For instance, the *Index Thomisticus Valency Lexicon* (IT-VaLex; McGillivray and Passarotti 2009) is a corpus-driven valency lexicon which was automatically induced from the syntactic layer of the *Index Thomisticus Treebank*. In a similar way, the *Homeric Dependency Lexicon* (HoDeL; cf. Zanchi in this volume), is a verbal lexicon of Homeric Greek whose data are based on the Homeric poems treebanked at AGLDT 2.0.

Many Natural Language Processing (NLP) tools for linguistic annotation are now available for many Indo-European and non-European languages and allow users to extract grammatical information from previously tokenized but unannotated texts. For example, *LemLat* 3.0 is the latest release of a morphological analyzer and lemmatizer for Latin (Passarotti et al. 2018): given an input word form, LemLat produces the corresponding lemma and a number of tags concerning the inflectional paradigm of the lemma and the morphological features of the input word form.

Finally, databases constitute a valuable resource for historical linguistics in that they not only allow storing large amounts of data, but also impose a structure on them, which facilitates access for researchers and application developers. Databases can be of different types and be fed with grammatical, lexical, or semantic information, and even combine different kinds of linguistic and non-linguistic information.

For instance, WordNets are lexical databases aimed to explore the lexicon of a languge, in which information is stored in a relational way. The original WordNet was developed for English at Princeton University and is fully documented in Fellbaum (1998). WordNets comprise nodes for lemmas to which meanings are associated in the form of synsets, i.e., sets of cognitive synonyms accompanied by brief definitions. Lexical relations establish connections among lemmas, whereas synsets are interlinked by means of semantic relations, resulting in a network of meaningfully related words and concepts. WordNets for three ancient Indo-European languages are being developed at the moment, as joint efforts of an international group of scholars: these languages are Latin, Ancient Greek and Sanskrit (see Short in this volume, Minozzi 2017, Franzini et al 2019, Biagetti et al. 2021, and Zanchi et al. 2021.

for further information). As these languages enjoy centuries of attestation and a long tradition of studies, each of the identified synsets is tagged for its periodization(s), literary genre(s), and *loci*, i.e., exemplifying attestations referred to by author(s) and work(s). Furthermore, etymological information is given for each lemma occurring in the database, thus allowing users to investigate whether Latin, Ancient Greek, and Sanskrit cognate words lexicalize comparable arrays of concepts.

2. *Annotating and storing metadata of ancient languages: Issues and perspectives*

While building the linguistic resources presented in this volume, authors were faced with the difficult task of annotating ancient languages and of storing and making these data available for the large public (most languages dealt with belong to the Indo-European family, but Cristofaro and Inglese in this volume show examples of alignment marking from Classical to Mandarin Chinese taken from their diachronic typological database, DemA). In fact, building resources for ancient languages brings about issues of different kinds. On the one hand, issues are linked with the ways in which ancient texts were handed down to the present time, which conditions both size and quality of the available data. In many cases, ancient texts survived up to the present by accidents of history (Joseph and Janda 2003: 15-19). On the other hand, issues are connected with the very nature of ancient languages, which appear to be 'peculiar' from the viewpoint of modern languages, for which current gold standard annotation schemes were originally designed. This brings us to reflect upon the extent to which aimed portability of annotation schemes (see Section 1) is actually mirrored into practice.

Due to the size of the available data, sampling criteria proposed by Sinclair (2005), among many others, cannot be applied to the selection of texts for historical corpora. While the bulk of data for modern languages is potentially unlimited, for ancient languages this is restricted to the few texts from the past that survived the accidents of time. Therefore, the sampling of data for historical corpora cannot rely on the orientation and aim of the corpus under construction, but rather tends to include all available material to maximize representativeness and to compensate for the smaller amount of available texts. This is especially true for those languages such as Gothic, for which we have very poor attestations, whereas for languages with a rich tradition such as Sanskrit, Ancient Greek, and Latin, the choice can be somehow oriented.

In the case of annotated corpora, the size of available data also affects the way in which the annotation can be carried out. For instance, modern treebanks often balance shallow annotation with the availability of huge masses of data and can therefore

exploit machine learning techniques to speed up the annotation process. On the contrary, ancient languages do not provide enough data to train tools such as taggers and parsers, nor to afford the noise produced by automatic shallow annotation. For these reasons, to be of value, the annotation should be carried out manually or be at least manually checked (Eckhoff, et al. 2018a; but cf. Hellwig and Sellmer in this volume on the possibility to employ Deep Learning techniques to accelerate the annotation process).

Frequently, scholars building historical resources are also faced with complications related to editorial issues. Since historical texts usually are testified by different manuscripts, and accordingly come to us in different variants, corpus design necessarily implies choosing the critical edition to record in the corpus and whether to include apparatus information (Eckhoff et al. 2018a). For example, developers of the AGLDT 2.0 (cf. Section 1) devised an 'ownership' model of treebank production and tested it on the work of Aeschylus: this draws on the methods of classical philology to take into consideration different textual variants and to make the personal choices of the annotator explicit (Bamman et al. 2009). Indeed, the long history of philological research on the individual texts differentiate historical treebanks from modern ones in at least two respects. First, while ambiguity is present in all languages, the decisions that annotators make in resolving syntactic ambiguity when dealing with historical texts have been debated for centuries. Furthermore, as we have mentioned, scholarly disagreement can be found not only on the level of the correct syntactic parse, but also on the form of the text itself. Therefore, the ownership model allows encoding multiple annotations for a text, thus allowing scholars who disagree with a specific annotation to encode their disagreement in a quantifiable form (Biagetti 2018: 27).

Hittite texts, which were written in a combination of cuneiform characters and logograms and handed down on clay tablets, represent a good example of the kind of editorial issues that developers must face when building a linguistic resource. In a pilot project for the inclusion of Hittite into Universal Dependencies (UD), Inglese (2015) faced the problems of building a resource which complies both with the current digital annotation standards and with complex philological practices established in the field of Hittitology. In this occasion, general textual information was provided in the comment line of the conll-u format employed by UD;[2] this included reference to the text and the tablet that the sentence belongs to, place of retrieval of the tablet, dating of both the tablet and the text. The MISC field of every form was provided with other philological features, such as integration, language, and transliteration: for instance, the integration feature indicated whether a word was actually attested on the tablet or whether it was restored after other copies or by the editor himself.

2. On the conll-u format, see https://universaldependencies.org/format.htm

The *Hittite Corpus* (Molina 2016) instead tackles the problem of lacunae by assigning a level of brokenness to every sentence in the corpus: brokenness values go from 1 (completely good) to 5 (hard-broken case) and only sentences belonging to the first 3 levels can be syntactically annotated, whereas fully broken fragments are marked as [...] and are excluded from syntactic annotation ("null constituents"). Beside that of lacunae, issues related to the Hittite writing system and its online representation also comprise the need of keeping the transliteration of proper Hittite words written in cuneiform script (in lowercase italics) distinct from that of Sumerian logograms (in capital letters) and of Akkadian words (capital italics).

Finally, ancient languages often feature constructions that are unknown to the modern languages for which the annotation schemes have been designed. In these cases, annotators can decide to customize the scheme itself in order to account for language-specific constructions. One example comes from the syntactic annotation of compounds in Vedic and especially in Classical Sanskrit within the UD scheme. Following a lexicalist approach, UD guidelines attribute compound formation to morphology and recommend not to split compounds into their element when they are univerbated. By contrast, Sanskrit compounds presents cross-linguistically infrequent characteristics (such as high recursiveness or inbound and outbound anaphora) that can be better explained by attributing compound formation to syntax rather than to morphology. In the context of syntactic annotation, this means that compounds should be split into their constituents, and these should be linked with each other by means of dependency relations that usually hold between independent words. This solution was adopted in a pilot study aimed to include Classical Sanskrit in UD (Biagetti 2018), as well as in the annotation of the Vedic Treebank (Hellwig et al. 2020).

In respect to the issues related to 'peculiarities' of ancient languages, another instructive example is the treatment that AGLDT 2.0 accords to preverbs in tmesis position in the treebanked version of the Homeric poems (cf. also Zanchi forthc.). In Homeric Greek, preverbs can occur in the so-called 'tmesis' positions and thus be 'split' from the verbs that they semantically modify by various linguistic items (cf. Zanchi 2019: ch. 3 with references therein for a diachronic interpretation of this preverb positioning). 'Split' preverbs hold an ambiguous status: they retain much of the syntactic freedom of their adverbial origin, but still semantically modify a verb (being closer to preverbs proper) or a noun phrase taken by the verb (being closer to adpositions). This syntactic ambiguity is reflected in an inconsistent annotation in the treebank: in some passages (e.g. Od. 11-64-65), preverbs in tmesis positions are annotated as prepositions (i.e., AUXP) and function as heads of nominal phrases. In contrast, the same items are annotated as other adverbs (i.e., AUXZ), such as logical operators meaning 'not', 'as well', and 'also', in other passages (e.g. Od. 10.559-560). Since Homeric Greek is a language with free word order (thus, 'split' preverbs

do not always occur in fixed positions) and the label AUXZ is ambiguous and frequent, there is no easy and automatic way to fix this issue in the annotation (see further Zanchi and Luraghi 2020).

3. *The aims of this volume*

In the field of computational linguistics, sharing knowledge with other researchers engaged in the creation and development of linguistic resources is essential to avoid multiplying efforts. Furthermore, communication among developers should encourage the use of compatible tools, formats and formalisms in order to increase the interoperability of resources dedicated to the same language as well as to different languages.

This volume collects the papers presented at the workshop *Building New Resources for Historical Linguistics*, which was hosted online by the University of Pavia on 3 November 2020. Its purpose was precisely to provide an opportunity for researchers engaged in the development of linguistic resources for historical linguistics to share their experience and knowledge in the field.

More in detail, the first three papers in the volume are devoted to treebanks.

Eckhoff and Haug familiarize us with the PROIEL treebank family, a set of treebanks of early Indo-European languages annotated according to the PROIEL enriched Dependency Grammar scheme. These treebanks can now be browsed by means of the *Syntacticus* online treebank facility, which also contains generated dictionaries for all the represented languages, including generated paradigms and sentence frames. The paper particularly focuses on the Old Russian dictionary, which has been supplied with Russian and English glosses as an example of how these automatically induced dictionaries can be developed further.

Hellwig and Sellmer describe the design of the Vedic Treebank, a treebank of selected passages from the Vedic literature annotated according to the UD standard. After sketching scope and current coverage of the corpus, the authors motivate the use of Deep Learning techniques for accelerating the annotation process and tackling the lack of trained annotators for Vedic Sanskrit.

Remaining within the Vedic Treebank, Biagetti discusses the annotation process of the *RigVeda*, a collection of religious hymns which constitute the oldest layer of Vedic literature and whose language is strongly conditioned by the poetic and ritual character of the text as well as by its metrical structure. In her paper, she reports some choices made in order to adapt the UD annotation scheme to the characteristics of Rigvedic syntax and takes similative constructions as a case study in order to test to what extent the adopted annotation is informative for the purpose of linguistic research.

With his paper, Scharf concludes the section of the volume dedicated to treebanks with some theoretical considerations. Scharf notes that formal and computational linguistics developed primarily for European languages belonging to the analytic type. Scharf further argues that, to develop universally valid linguistic formalisms, linguistic theories developed to describe languages very different from English should also be taken into account. Scharf pursues this idea showing that the millennial Indian linguistic tradition could offer useful insights to contemporary formal linguistics, and Indian linguistic theories can be formalized and implemented computationally.

With Passarotti and Mambrini's paper, the volume widens the discussion to other linguistic resources and specifically addresses the question of interoperability among them. The authors introduce the architecture of the knowledge base of the *LiLa* (*Linking Latin*) project, which uses the principles, ontologies, and models of the Linguistic Linked Open Data community to connect and make available for users the existing linguistic resources for Latin. In addition, the paper offers a number of practical examples of how specific research questions concerning the Latin lexicon can be better answered using the lexical resources already linked to *LiLa*.

The paper by Carling and colleagues describes the most recently developments of the *Diachronic Atlas of Comparative Linguistics – DiACL*, a lexical database currently at its version 2.1. The improvements concern both the infrastructure and the base data of DiACL. As regards DiACL usage, typological data sets can now have a global coverage and the clicks necessary to navigate through the data has been lowered. As concerns data, inconsistencies among different data sets and etymological trees have been removed, the grammatical and semantic information for Indo-European lexical data is improved, and lacunae in languages and lexemes are filled.

Short's paper is also devoted to the description of a family of lexical databases, the ancient-language WordNets for Sanskrit, Ancient Greek, and Latin. In particular, Short focuses on how a WordNet-based text encoding schema can help linguists address some of the challenges related to polysemy emerging in discourse that are faced while semantically annotating ancient texts. By concrete examples, Short also emphasizes that these considerations are relevant across different disciplines interested in ancient cultures.

The last two papers describe resources focused on valency phenomena, which have been created in the framework of the project *Transitivity and Argument Structure in Flux* (funded by the *Italian Ministry for Education and Research* in the framework of the 2015 PRIN call, grant no. 20159M7X5P). Zanchi's paper presents the *Homeric Dependency Lexicon* (HoDeL), a verbal lexicon of Homeric Greek with a user-friendly interface facilitating the investigation of Homeric verbs, their dependents and other aspects of the Homeric syntax. After discussing the data on which

the lexicon is based, Zanchi illustrates HoDeL incorporated constraints and functionalities and shows how they can be employed by perspective users to answer specific research questions about the Homeric syntax.

Cristofaro and Inglese introduce the Pavia *Diachronic Emergence of Alignment* (DEmA), a new resource for the study of the diachrony of alignment patterns cross-linguistically, aimed to investigate the sources and processes out of which new alignment patterns come into being across languages. In particular, they describe DEmA, its structure and the choices that have been made in its construction to facilitate DEmA's aims: a set of parameters known for playing a role in the development of alignment patterns has been implemented into a searchable format, which also allows for cross-linguistic comparisons.

As illustrated above, some of the papers contained in this volume describe mature resources and discuss their application possibilities. Others instead introduce projects that are still works in progress, presenting their aims, criticalities concerning their construction, and the methodologies employed to tackle them. Thus, though this volume also showed that there is still much work to do in creating digital resources for ancient languages and in making them interoperable, we believe that many advances in this respect are jointly documented in this book. Moreover, a very welcome characteristic of digital resources is that they can be constantly improved, for example by enlarging the data sets, correcting errors in them and refining users' interfaces. Thus, their very nature of digital resources calls for further steps along this longer path.

Websites

AGLDT 2.0: https://perseusdl.github.io/treebank_data/
Ancient Greek WordNet: https://greekwordnet.chs.harvard.edu
DĀMOS: https://damos.hf.uio.no/1
GRETIL: http://gretil.sub.uni-goettingen.de/gretil.html#top
HoDeL: https://su-lab.unipv.it/tasf/index.php/hodel/
IT-VaLex: https://itreebank.marginalia.it/view/IT-valex.php
Latin WordNet: https://latinwordnet.exeter.ac.uk
LemLat 3.0: http://www.lemlat3.eu.
MIDIA: https://www.corpusmidia.unito.it
Perseus Digital Library: http://www.perseus.tufts.edu/hopper/; cfr. also the Scaife Viewer reading environment, which is the first phase of work towards the next version of the Perseus Digital Library, Perseus 5.0: https://scaife.perseus.org.
Prague Dependency Treebank: https://ufal.mff.cuni.cz/pdt3.0
Princeton WordNet: https://wordnet.princeton.edu/download/current-version
PROIEL (syntacticus): http://syntacticus.org

Sanskrit WordNet: https://sanskritwordnet.unipv.it
The Chicago Homer: https://homer.library.northwestern.edu
The Diorisis Ancient Greek Corpus: https://figshare.com/articles/dataset/The_Diorisis_Ancient_Greek_Corpus/6187256
TITUS: http://titus.fkidg1.uni-frankfurt.de/framee.htm?/index.htm
VedaWeb: https://vedaweb.uni-koeln.de
Wulfila Project: http://www.wulfila.be

References

Anthony, Laurence. 2013. A critical look at software tools in corpus linguistics. *Linguistic research* 30(2): 141–161.

Bamman, David, Mambrini, Francesco & Crane, Gregory. 2009. An ownership model of annotation: The Ancient Greek dependency treebank. In Proceedings of the eighth international workshop on treebanks and linguistic theories (TLT 8), Marco C. Passarotti, Adam Przepiórkowski, Savina Raynaud & Frank Van Eynde, 5–15.

Bamman, David & Crane, Gregory. 2011. The Ancient Greek and Latin Dependency Treebank. In *Language Technology for Cultural Heritage*, Caroline Sporleder, Antall van Den Bosch & Kalliopi Zervanou (eds), 79–89. Berlin: Springer.

Biagetti, Erica. 2018. *A dependency treebank of Classical Sanskrit*. MA thesis, University of Pavia.

Biagetti, Erica, Zanchi, Chiara & Short, William M. 2021. Toward the creation of WordNets for ancient Indo-European languages. In *Proceedings of the 11th Global WordNet Conference*, Sonja Bosch, Christiane Fellbaum, Marissa Griesel, Alexandre Rademaker & Piek Vossen (eds), 258–266. EACL/GWC: Global WordNet Association.

Biber, Douglas. 2009. Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In *The Oxford Handbook of Linguistic Analysis*, Bernd Heine & Heiko Narrog (eds), 159–191. Oxford: Oxford University Press.

Bresnan, Joan, Asudeh, Ash, Toivonen, Ida & Wechsler, Stephen. 2015. *Lexical Functional Syntax*. 2nd edition. Hoboken: Wiley Blackwell.

Busa, Roberto. 1980. *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque/consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SJ*. Stuttgart–Bad Cannstatt: Frommann–Holzboog.

Drucker, Johanna (2013). *Intro to Digital Humanities: Introduction*. UCLA Center for Digital Humanities.

Eckhoff, Hanne Martine, Bech, Kristin, Eide, Kristine, Bouma, Gerlof, Haug, Dag T. T., Haugen, Odd E. & Jøhndal, Marius. 2018a. The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52 (1): 29–65.

Eckhoff, Hanne Martine, Luraghi, Silvia & Passarotti, Marco C. 2018b. The added value of diachronic treebanks for historical linguistics. *Diachronica* 35 (3): 297–309.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Franzini Greta, Peverelli, Andrea, Ruffolo, Paolo, Passarotti, Marco C., Sanna, Helena, Signoroni, Edoardo, Ventura, Viviana & Zampedri, Federica. 2019. Nunc Est Aestimandum: Towards an Evaluation of the Latin WordNet. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*, Raffaella Bernardi, Roberto Navigli & Giovanni Semeraro (eds), 1–8. Torino: Accademia University Press.

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

Haug, Dag T. T. 2015. Treebanks in historical linguistics research. In *Perspectives on Historical Syntax*, Carlotta Viti (ed), 187–202. Amsterdam: Benjamins.

Hellwig, Oliver, Scarlata, Salvatore, Ackermann, Elia & Widmer, Paul. 2020. The Treebank of Vedic Sanskrit. In *Proceedings of The 12th Language Resources and Evaluation Conference* (LREC 2020), Nicoletta Calzolari, Frederic Bechet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi et al. (eds), 5137–5146.

Inglese, Guglielmo. 2015. Towards a Hittite Treebank. Basic Challenges and Methodological Remarks. In *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH), 10 December 2015, Warsaw, Poland*, Marco C. Passarotti, Francesco Mambrini & Caroline Sporleder (eds), 59–68.

Joseph, Brian and & Janda, Richard D. (eds). 2003. *The Handbook of Historical Linguistics*. Oxford: Blackwell.

Mayrhofer, Manfred. 1980. *Zur Gestaltung des etymologischen Wörterbuchs einer "Großcorpus-Sprache*. Wien: Akademie der Wissenschaften. Phil-Hist. Klasse.

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

McGillivray, Barbara & Passarotti, Marco C. 2009. The Development of the Index Thomisticus Treebank Valency Lexicon. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education* [LaTeCH – SHELT&R 2009], Lars Borin & Piroska Lendvai (eds), 43–50. Athens: ACL.

Minozzi, Stefano. 2017. Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval. *Antichistica* 14: 123-133.

Molina, Maria. 2016. Syntactic annotation for a Hittite corpus: Problems and principles. In *Proceedings of the Workshop on Computational Linguistics and Language Science* (CLLS 2016), Moscow, Russia, vol. 26.

Nyhan, Julianne & Passarotti, Marco C. 2019. *One Origin of Digital Humanities: Fr Roberto Busa in His Own Words*. Cham: Springer.

Passarotti, Marco C., Ruffolo, Paolo, Cecchini, Flavio M., Litta, Eleonora & Budassi, Marco 2018. LEMLAT 3.0. https://doi.org/10.5281/zenodo.1492133.

Ponti, Edoardo. M. & Luraghi, Silvia. 2018. Non-configurationality in diachrony Correlations in local and global networks of Ancient Greek and Latin. In *Diachronic Treebanks for Historical Linguistics*, Hanne Martine Eckhoff, Silvia Luraghi & Marco C. Passarotti (eds), 70–93. Amsterdam: Benjamins

Schreibman, Susan, Siemens, Ray & Unsworth, John (eds). 2004. *A Companion to Digital Humanities*. Oxford: Blackwell publishing.

Sgall, Petr, Hajičová, Eva & Panevová, Jarmila. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht: D. Reidel.

Sinclair, John. 2005. Corpus and text: Basic principles. In *Developing linguistic corpora: A guide to good practice*, Martin Wynne (ed), 1–16. Oxford: Oxbow Books

Untermann, Jürgen. 1983. Indogermanische Restsprachen als Gegenstand der Indogermanistik. In *Le lingue indoeuropee di frammentaria attestazione. Die indogermanischen Restsprachen. Atti del convegno della Societa italiana di glottologia e della Indogermanische Gesellschaft*, Udine, 22–24 settembre 1981, Edoardo Vineis (ed), 11–28. Pisa: Giardini.

Zanchi, Chiara. 2019. *Multiple Preverbs in Ancient Indo-European Languages.* Tübingen: Narr.

Zanchi, Chiara. Forthc. The *Homeric Dependency Lexicon*: what it is and how to use it. *Journal of Greek Linguistics.*

Zanchi, Chiara & Luraghi, Silvia. 2020. Presenting *HoDeL* - A new resource for research on Homeric Greek verbs. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"* 19, Supplementary volume: 1188-1200 http://www.dialog-21.ru/media/5159/_-dialog2020_supvol.pdf

Zanchi, Chiara, Luraghi, Silvia & Biagetti, Erica. 2021. Linking the Ancient Greek WordNet to the Homeric Dependency Lexicon. In *Computational Linguistics and Intelligent Technologies. Papers from the Annual International Conference "Dialogue 2021"* [Issue 20], 729–737.

Annotation Schemes, Tools and Data in the PROIEL Treebank Family

Hanne Martine Eckhoff*, Dag T. T. Haug**

The article provides a brief description of the PROIEL family of treebanks of ancient languages, originating in the project Pragmatic Resources of Old Indo-European Languages at the University of Oslo. The treebanks contain texts from a wide range of early attestations of Indo-European languages, such as Ancient Greek, Latin, Gothic, Classical Armenian, Old Church Slavonic, Old Russian, Old Norse, Old English, Old French and Old Portuguese. They share an enhanced dependency annotation scheme and a tailored open-source annotation web application. Corpus data are released freely for non-commercial use, and many of the treebank data can be browsed in the Syntacticus web interface, which also includes a module generating dictionaries for each language. The resulting dictionaries exploit the treebank data to generate attested paradigms and valency frames, and may be enhanced with glossing and date metadata. The treebanks are therefore accessible to a wide audience with different needs.

Keywords: treebanks, dependency grammar, classical languages, Indo-European, syntax

1. *Introduction*

The PROIEL family of treebanks (Eckhoff et al. 2018) originated in the research project Pragmatic Resources of Old Indo-European languages in 2008 (PI Dag T. T. Haug, University of Oslo). A main goal of the project was to create a parallel treebank of the Greek New Testament and its earliest translations into Latin, Gothic, Old Church Slavonic and Classical Armenian, and to develop tools and annotation schemes that suited the structure of old Indo-European languages and would allow preserving as much information as possible in the annotation (Haug and Jøhndal 2008). An important aim was also to add annotation that went well beyond morphology and syntax, in particular for information structure and various semantic features, since the main research goal of the project was to account for the linguistic means by which such languages encode pragmatics and information structure, with particular attention to word order, anaphoric expressions, definiteness, participles used as backgrounding devices and discourse particles.

Ever since the beginning of the project, the original PROIEL treebank has been expanded, and new treebanks on the same format have been created by collaborating project groups. The expansions of the PROIEL treebank have mainly been done to achieve diachronic depth for the Classical languages: Caesar and Cicero

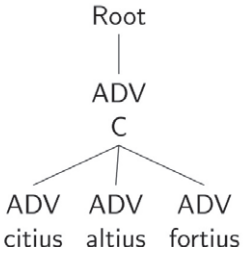
* University of Oxford. ** University of Oslo.

for Classical Latin, Herodotus for Ancient Greek; *Peregrinatio Aetheriae* and *Opus agriculturae* for later Latin and the Chronicles of Sphrantzes for later Greek. Additional treebanks on the same format have been created for early Germanic and Romance (ISWOC, University of Oslo), Old Icelandic (Greinir Skáldskapar, Reykjavik, see Eyþórsson et al. 2014), Old Norwegian (Menotec, Bergen/Oslo), Old Swedish (MAÞIR, Gothenburg) and (additional) Old Church Slavonic, later Church Slavonic recensions, Old Russian and Middle Russian (TOROT, Tromsø/Oxford). Various treebank pilots have also been created, for instance for Hittite and Sanskrit.

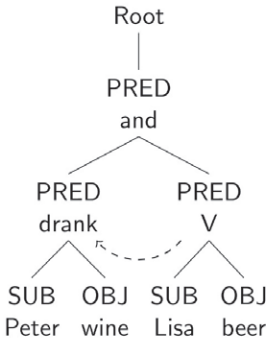
The long-term and collaborative nature of the treebank work has resulted in a well-developed, mature system with detailed annotation schemes and high-consistency annotation. A wide range of freely available and reusable tools has also been developed, tried and tested in large-scale annotation and through a series of studies based on data from the treebanks. At the time of writing, annotation was still ongoing, in some of the treebanks, albeit with limited resources. The following sections will offer a brief description of the PROIEL annotation scheme, annotation tools, browsing tools and the current availability of the existing treebanks.

2. *Dependency grammar annotation scheme*

The PROIEL treebanks are annotated according to an enhanced dependency grammar scheme. As in all dependency treebanks, this means that constituents are not explicitly indicated, and that word order and syntactic relations are independent of each other. Compared to more "classical" dependency schemes, such as the Prague Dependency Treebank scheme and the closely related schemes used for the Latin Dependency Treebank and the Ancient Greek Dependency Treebank, the PROIEL scheme diverges in two important ways to achieve greater expressivity: in allowing empty nodes to represent ellipsis and parataxis, and in allowing secondary edges to indicate structure sharing. In example (1) we see the use of a null conjunction to indicate parataxis, while example (2) uses a null verb to indicate an elided verb. Note that null nominals are not used – elided arguments are dealt with in the dedicated information structure annotation layer (see Haug et al. 2014 for further details).

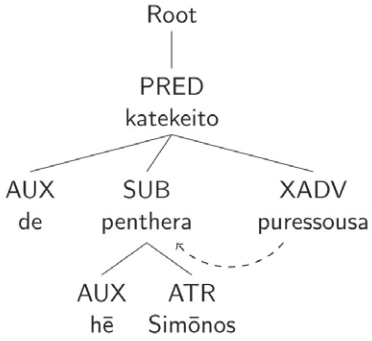


(1) *Citius,* *altius,* *fortius*
fast.CMPR high.CMPR strongly. CMPR
'Faster, higher, stronger.'



(2) *Peter drank wine and Lisa beer.*

Secondary edges are used to indicate structure sharing, e.g. to indicate the external subjects of conjunct participles (control), as seen in (3)



(3) *hē* *de* *penthera* *Simōnos* *katekeito*
the PTC mother_in_law(F).NOM.SG Simon.GEN lie.IMPF.3SG
puressousa
suffer_fever.PTCP.PRS.F.SG

'Now Simon's mother-in-law lay ill with a fever.' (Mk. 1:30)

They are also used to indicate shared dependents and predicate identity in coordinations (see example 2 for the latter).

Secondary edges are not processable by standard dependency parsers, so they must either be ignored or inserted separately. The decision to include them in the annotation scheme was based on the idea that with ancient and scarce sources, which is the norm in the PROIEL-style treebanks, it is more important to express structure than to prioritize computational convenience. The PROIEL dependency scheme shares many features with the enhanced Universal Dependencies format.

Treebanks provide data for larger audiences, but at the same time we see that there is a widening gulf between corpus linguistics and linguistic theory. While

there is a strong trend for dependency treebanks both among corpus linguists and in NLP, dependency grammar is not widely in use as a linguistic theory. While a number of phrase structure treebanks (e.g. the Penn treebanks) do exist, even for historical languages, they typically use flatter tree structures than any syntactician would subscribe to. The difference between dependency treebanks and phrase structure treebanks is thus not as great as it may seem, and the latter are not necessarily much closer to mainstream syntactic theory than the former.

The principle underlying the PROIEL annotation scheme has therefore been to encode no more structure than is common to all frameworks (even if secondary in some frameworks), but to encode *enough* structure to allow reconstruction of theoretically motived structures. Ideally the data in the treebank can then be expanded to structures conforming to a specific theory by adding information from the assumptions of that theory (see Haug 2012 for an attempt to automatically generate LFG c-structures from PROIEL-style dependency graphs).

3. *Annotation tools*

Another cornerstone in the work on the PROIEL-style treebanks is the shared annotation web application, the freely available PROIEL Annotator, written chiefly by Marius L. Jøhndal. The web application was specifically designed to suit the needs of historical treebank projects – since it is often necessary to work with an international team to get the necessary expertise, it was important to allow annotation from an ordinary browser without cumbersome local installation. Since the languages in question are generally low-resourced, the application also aims to exploit existing annotation to provide annotation support, and several of the treebank projects have boosted this support with statistical tagging of part of speech and morphology, as well as automatic lemma guessing (see Eckhoff and Berdicevskis 2015 for an example).

The workflow in the annotation application is as follows:

1. Correction of segmentation and sentence division.
2. Lemmatization and annotation for part of speech and detailed morphology (stored in a 10-place positional tag). The web application provides guesses for previously seen forms – for previously unencountered forms lemmas must be entered manually and part of speech and morphological features chosen from drop-down menus.
3. Dependency annotation assisted by rule-based guesses.
4. Review of the annotation by an experienced annotator.
5. Annotation of information status (old, accessible, new) and anaphoric relations (see Haug, Eckhoff and Welo 2014 for further details).

In addition customized tagging at sentence, lemma and token level may be added, but not via the annotation interface. This option has been used by several of the tree-

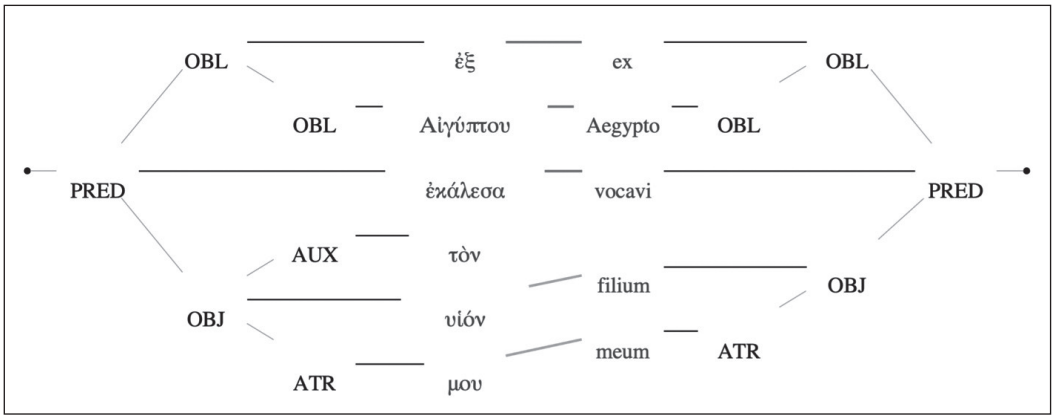
bank projects to add annotation for various semantic features (such as animacy) and derivational morphology (such as aspect morphology in Old Church Slavonic and Old Russian). It is also possible to align parallel treebanks at token level, and this has been done for the original PROIEL New Testament treebank, where all the translations have been aligned with the original Greek text. In the TOROT treebank the Psalterium Sinaiticum has been aligned with the Septuagint Psalms.

4. *Browsing tools*

The data from the PROIEL, ISWOC and TOROT treebanks are released in xml and CoNNL-X format at https://proiel.github.io/, http://iswoc.github.io/ and http://torottreebank.github.io/. There are also partial releases of Universal Dependencies conversions of the PROIEL and TOROT treebanks. However, only a limited audience is able to make use of such data files. An important more recent addition to the PROIEL suite of tools is therefore the treebank browsing interface available at http://syntacticus.org/,[1] where the released treebank data from the PROIEL, ISWOC and TOROT treebanks are published. This makes it possible to disseminate the treebank data to a much wider audience than those who are able to use the raw data files. The browsing interface allows users to read continuous texts and provides different views of the linguistic annotation, including side-by-side views of the dependency trees in aligned texts, as illustrated in Figure 1.

The interface also contains a dictionary module. PROIEL-style treebanks contain a lot of lexicographically interesting information, since morphologically analysed

Figure 1 Parallel Greek and Latin dependency trees: 'Out of Egypt I called my son' (Mt. 2.15)



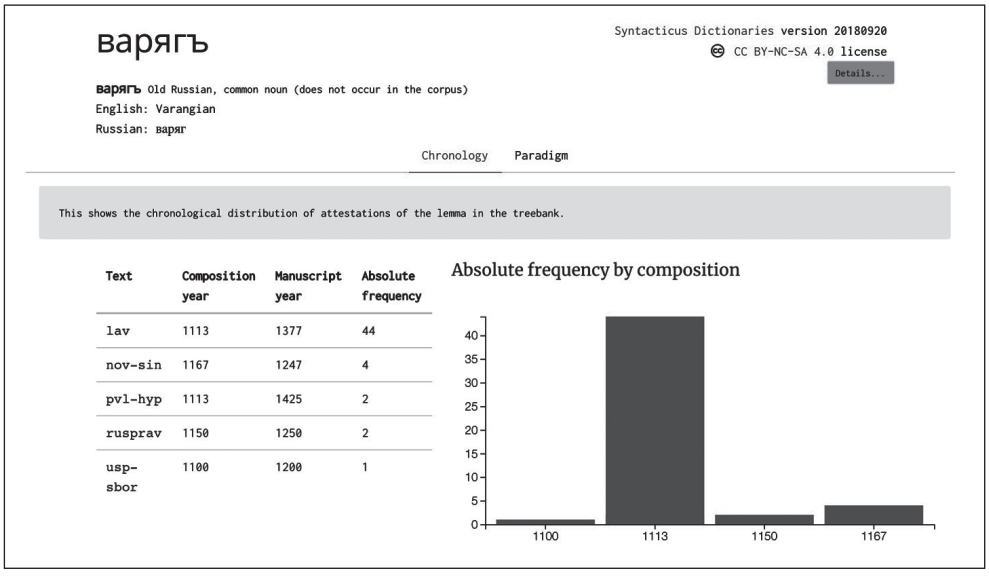
1. Written and maintained by Marius L. Jøhndal.

tokens are assigned to lemmas (which can also potentially be glossed). Furthermore, there is syntactic information about every token and its sentence. It is also possible to add dates of composition and manuscript dates to the metadata of each text in the treebanks. All of these features are exploited in the dictionary module, which provides the following information for each lemma:

a. glossing if added (Figure 2)
b. a timeline with frequencies if date metadata have been added (Figure 2)
c. attested paradigms for inflected items with frequency breakdown and full concordance per attested form-tag combination (Figure 3 and 4)
d. valency frames for verbs (Figure 5)

Such dictionaries are generated for all languages represented in Syntacticus, so that even treebanks without glossing and metadata will have a useful dictionary resource. The Old Russian dictionary, however, has been enhanced with glossing in Russian and English of ca. 8000 lemmas, as well as date of composition and date of manuscript metadata for all texts.[2] Since the dictionary situation for Old Russian is unsatisfactory and fragmented, this is an important addition to the existing resources.

Figure 2 Glossing and timeline for the Old Russian noun *varjagъ* 'Varangian' by year of composition



2. This work was done as part of the project Varangian Rus' Digital Environment project at UiT Arctic University of Norway.

Figure 3 Attested paradigm for the Old Russian noun *varjagъ* 'Varangian' with frequencies per form and live links to concordances

	Singular	Dual	Plural
Nom.	варѧгъ (6)	варѧга (1)	варѧзи (12) варѧгꙑ (1)
Acc.		варѧга (1) варꙗга (1)	варѧгы (8) варѧги (7) варѧгꙑ (2) вариги (1) варѧгы (1)
Gen.			варѧгъ (5)
Loc.			
Dat.			варѧгомъ (4)
Ins.			варѧгы (2) варѧги (1)

Figure 4 Concordance for the Old Russian nominative plural form *варѧзи* (*varjazi)* 'Varangians'

Lav. 4.8	по семуже морю сѣдѧть	варѧзи	сѣмо ко въстоку до предѣла сим
Lav. 4.11	афетово бо и то колѣно	варѧзи	свеи. оурмане русь. агнѧне гал
Lav. 19.7	лѣт .҂ѕ҃.т҃.ѯ҃ .з҃·:· маху дань	варѧзи	изъ заморья. на чюди и на слов
Lav. 19.22	и.	варѧзи	суть.
Lav. 20.15	по тѣмъ городомъ суть находици	варѧзи	
Lav. 23.24	и. бѣша оу него	варѧзи	и словѣни
Lav. 54.8	мнози бо бѣша	варѧзи	хесѧни.
Lav. 78.26	посемь рѣша	варѧзи	володимеру.
Lav. 79.2	и рѣша	варѧзи	
Lav. 79.7	ц҃рю се идуть к тебѣ	варѧзи	
Lav. 140.17	аву же не вѣдущю ѿтьнѣ смр҃ти.	варѧзи	бѧху мнози оу ярослава.
Novgorod Chronicle	а на осѣнь придоша	варѧзи	горою на миръ.

Previous Page 1 of 1 Next page

Figure 5 Valency frames for the Old Russian verb *ubiti* 'kill' with frequencies and live links to concordances

Arguments	Non-reflexive	Reflexive
(none)	56	
OBJ (accusative)	37	
OBJ (genitive)	86	
OBJ	17	
OBJ (personal reflexive pronoun *себе*, genitive)	2	
OBJ (adverb *бесчисльно*)	1	
OBJ (adverb *нѣколико*)	1	
OBJ (preposition *о* + locative)	1	

5. *Summary*

This article has provided a short description of the PROIEL family of treebanks for ancient languages, originating in the University of Oslo project Pragmatic Resources in Old Indo-European Languages. The treebanks share a customized annotation web application and an enhanced dependency annotation scheme tailored for the syntax and rich morphology of these languages. The tools and treebanks are created by and for linguists, and are informed by linguistic theory and tested and improved through extensive research using and enhancing the treebank data. The treebanks are based on well-established standards and guidelines, and extensions to new languages have been done in close cooperation with the original team, which makes them unusually compatible and consistent. The annotation web application is open-source and data are freely shared for non-commercial use. In addition, the Syntacticus browsing interface makes treebank data available to a much wider audience, and includes generated dictionaries for all the represented languages.

Abbreviations

ADV = adverb, ATR = attribute, AUX = auxiliary, C = coordination, CMPR = comparative, F = feminine, GEN = genitive, IMPF = imperfect, NOM = nominative, OBJ = object, PRS = present, PTCP = participle, PRED = predicate, SG = singular, SUB = subject, XADV = conjunct participle

Websites

ISWOC: http://iswoc.github.io/
ISWOC (GitHub): http://iswoc.github.io/
MAÞIR: https://spraakbanken.gu.se/mathir
Menotec: https://www.menota.org/menotec.xml
PROIEL (web app): https://github.com/mlj/proiel-webapp
PROIEL (GitHub): https://proiel.github.io/
Syntacticus: http://syntacticus.org
TOROT: http://torottreebank.github.io/
TOROT (GitHub): http://torottreebank.github.io/
UD format: https://universaldependencies.org/u/overview/enhanced-syntax.html

References

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

Eckhoff, Hanne Martine, Bech, Kristin, Eide, Kristine, Bouma, Gerlof, Haug, Dag T. T., Haugen, Odd E. & Jøhndal, Marius. 2018a. The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52 (1): 29–65.

Eyþórsson, Þórhallur, Karlsson, Bjarki & Sigurðardóttir, Sigríður Sæunn. 2014. Greinir skáldskapar: A diachronic corpus of Icelandic poetic texts. In *Proceedings of Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage –LRT4HDA, workshop at LREC 2014*, Kristín Bjarnadóttir, Mathew Driscoll, Steven Krauwer, Stelios Piperidis, Cristina Vertan & Martin Wynne (eds), 35–41.Reykjavík: University of Iceland.

Haug, Dag T. T. 2012. From dependency structures to LFG representations. In *Proceedings of the LFG12 Conference*, Miriam Butt & Tracy Holloway King (eds), CSLI Publications. https://web.stanford.edu/group/cslipublications/cslipublications/LFG/17/papers/lfg12haug.pdf.

Haug, Dag T. T. & Jøhndal, Marius. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the Sixth International Language Resources and Evaluation* (LREC'08), Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias (eds), 27–34. LREC: European Language Resources Association (ELRA).

Haug, Dag T. T., Eckhoff, Hanne Martine & Welo, Eirik. 2014. The theoretical foundations of givenness annotation. In *Information Structure and Syntactic Change in Germanic and Romance Languages*, Kristin Bech & Kristine G. Eide (eds), 17–52. Amsterdam: Benjamins.

The Vedic Treebank

Oliver Hellwig*, Sven Sellmer**

Vedic Sanskrit is an early Indo-European language in which a large corpus of religious and ritual texts has been transmitted. Despite its linguistic and historical importance, there existed no large-scale digital resources for the syntax of Vedic until recently. This paper gives an overview of the development and the current status of the Vedic Treebank, a syntactic treebank genuinely annotated according to the Universal Dependencies standard. A special focus of this paper is on the use of machine learning techniques that support human experts during the annotation process and are planned to be employed for annotating larger ranges of the Vedic corpus in a fully unsupervised manner.

Keywords: Vedic Sanskrit, treebank, syntax, dependency labeling, diachronic corpora

1. *Introduction*[1]

Vedic Sanskrit, or simply "Vedic", the precursor of Classical Sanskrit, is the oldest attested representative of the Indo-Aryan group and one of the oldest attested Indo-European languages in general. It was in active use from (very roughly) the 15th to the 5th c. BCE in the northern part of the Indian subcontinent. Its name derives from the Vedic textual corpus, in which it is exclusively found. To this day, the texts making up this corpus play a fundamental role for Hinduism and are invaluable sources for our knowledge about the development of Indian culture up to the time of the Buddha (ca. 420–350 BCE). They mostly deal with religious matters in a wide sense, but also touch upon many other aspects of life in Ancient India. It must be stressed that all texts of this huge corpus were composed without the help of writing, the usage of script gaining foot in India only very slowly from the 5th c. BCE onward and making real progress only thanks to the edicts of King Aśoka in the middle of the 3rd c. BCE (see Falk 2018). For many centuries, these texts were memorized and transmitted orally, even after writing had become widespread, and, remarkably enough, thanks to a rigorous teaching system and advanced mnemonics this transmission took place with an extremely small error rate.

* Heinrich Heine University Düsseldorf, Institute for Language and Information; University of Zurich, Department of Comparative Language Science.
** Heinrich Heine University Düsseldorf, Institute for Language and Information; Adam Mickiewicz University in Poznań, Institute of Oriental Studies.

1. The authors were partly funded by the German Federal Ministry of Education and Research, FKZ 01UG2121. The project ChronBMM was funded by the German Federal Ministry of Education and Research, duration: 2021-2023.

The structure of the Vedic corpus can only be sketched very briefly here (see e.g., Gonda 1975 and Gonda 1977 for a detailed overview). First, it is, so to speak, vertically divided into four main divisions (the *Ṛgveda*, *Sāmaveda*, *Yajurveda* and *Atharvaveda*), each of them with numerous schools forming more or less separate textual traditions (see Renou 1947). Second, the corpus is divided according to the text type, with the latter classification having also a chronological value because different types of texts were produced in different historical epochs. To be sure, the absolute dating of these texts is highly disputed so that all dates given here have to be treated as mere approximations. The relative chronology of the main groups, however, is rather uncontroversial. Groups 1–4 of the following overview make up the Veda proper, whereas group 5 includes later and less authoritative texts.

1. *Saṃhitā*s (15th to 11th c. BCE): lit. 'collections', namely of hymns addressed to various deities, and of ritual and magical formulas. In each division there is one *saṃhitā*, which commonly, though imprecisely, is called a 'Veda'. So, e.g., the most important of the *saṃhitā*s, the *Ṛgveda-Saṃhitā* (composed between 1200–1000 BCE), is generally known simply as the *Ṛgveda*.
2. *Brāhmaṇa*s (8th to 7th c. BCE): voluminous prose texts mostly dealing with detailed explanations and discussions of the rituals in which the *saṃhitā* texts are used.
3. *Āraṇyaka*s (7th to 6th c. BCE): prose texts of a partly ritualistic, partly philosophical character that share characteristics of both the *brāhmaṇa*s and *upaniṣad*s.
4. *Upaniṣad*s (7th to 2nd c. BCE): theological and philosophical treatises, the oldest of which are composed in prose, the younger in verses. (Please note that there are very many texts called *'Upaniṣad*s' that do not belong to the Vedic corpus.)
5. *Vedāṅga*s (4th c. BCE to 3rd c. CE). These ancillary texts (lit. 'limbs of the Veda') include numerous treatises on topics like phonetics, metrics, and the like, but most importantly three types of texts, composed in a special, extremely condensed style: the *Gṛhyasūtra*s, *Śrautasūtra*s (manuals of domestic and of solemn rituals), and *Dharmasūtra*s (compendiums of law and customs).

The size of the Vedic corpus has not been precisely established so far. According to our estimation, it may contain up to 3 million lexical units. Though still not small, it is clear that originally it was much larger, but many texts have been lost.

The language used in the Vedic corpus has been an important object of research both in Indo-European linguistics and in linguistics in general right from the time it became known in the West at the beginning of the 19th century. There is no sharp line dividing Vedic from Classical Sanskrit. We rather have to do with a gradual vanishing of typically Vedic features on the syntactic, morphological and lexical level. Classical Sanskrit is a highly standardized language, due to the efforts of the great grammarian Pāṇini (ca. 4th c. BCE) and his predecessors; Vedic Sanskrit, on the oth-

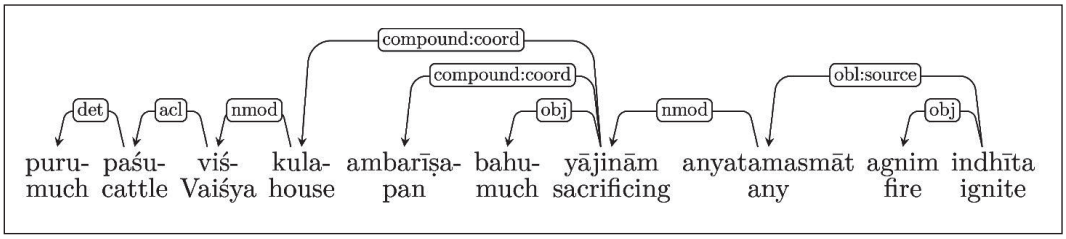
er hand, shows a clear development from the *Ṛgveda* to the youngest texts of the Vedic corpus, and in addition bears traces of a geographical differentiation (Witzel 1989). The language of the younger texts of the last two groups listed above already marks a transitory stage between Vedic and Classical Sanskrit (for an overview of the stages and variants of Sanskrit, see Renou 1956).

The Vedic Treebank (VTB), presented here, forms part of an ongoing project in which the syntactic structures of the Vedic corpus are annotated using the Universal Dependency (UD) standard (Nivre et al. 2016). Its main motivation is a palpable lack of resources: while there exist numerous publications on Vedic syntax (see e.g., the bibliographies in Deshpande and Hock 1991 and Hock 2013), a large-scale database of syntactic annotations that could help to empirically validate linguistic claims was missing until 2020. The annotations made so far can be inspected online at http://sanskrit-linguistics.org/dcs (Hellwig 2010-2021).

2. *Composition and Growth of the VTB*

The VTB is work in progress, and the annotation has seen three major versions so far. The first version (Hellwig et al. 2020) contained extracts from five Vedic texts. Here, the hymns of the Rigveda (RV) and metrical parts from the Śaunaka recension of the Atharvaveda (Śaunakasaṃhitā) represented the oldest layer of the Vedic language, while samples of old Vedic prose were extracted from the Maitrāyaṇīsaṃhitā, Aitareyabrāhmaṇa, Śatapathabrāhmaṇa and Śaunakasaṃhitā 15, the Śatapathabrāhmaṇa sample probably being the youngest among them. On the whole, the first version contained about 4,000 sentences with approximately 27,000 word tokens. An integral part of the initial setup was the composition of an annotation guideline which pays special attention to those cases in which we needed to deviate from the official UD standard, or which often give rise to contending interpretations. Most relevant among these cases is the annotation of compounds which, in contradistinction to languages such as English, can encode complex syntactic and semantic hierarchies (Lowe 2015). In order to capture the rich information encoded in compounds, we annotate their elements as if they occurred in non-composed form. An example for such a compound is the string *puru-paśu-viṭkulāmbarīṣa-bahu-yājinām* from the *Śāṅkhāyanagṛhyasūtra*, a manual of the domestic ritual. This string can be decomposed into seven words, as is shown in Figure 1. While the compound itself is a genitive modifier of the quantifier *anyatamasmāt* 'from any', its basic structure is a coordination of the three words *kula-* 'house', *ambarīṣa-* 'pan' and *yājin-* 'one who sacrifices' which we annotate with the dependency relation *compound:coord*. Two of these three words are further modified: *yājin-* by an object (or adverbial modifier) *bahu-* 'much, many' and *kula-* by *viś-*, the name of a social class, which is again specified by a nested depictive. Annotating such cases using only the officially recommended label *compound* would discard

Figure 1 Compounding at *Śāṅkhāyanagṛhyasūtra* 1.1.8: 'He should light his fire at one of the following places, viz. in the house of a Vaisya who is rich in cattle, at a frying-pan, or (at the fire of) one who offers many sacrifices' (translation by Oldenberg 1886: 13–14)



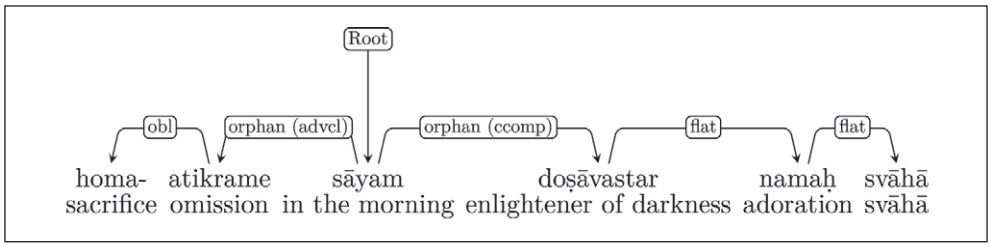
important information encoded in this widely used type of constructions which become even more popular towards the end of the Vedic period.

In the second version, which was created with the active collaboration of Erica Biagetti (Pavia; also see Biagetti in this volume), samples from a new text, the important philosophical treatise called *Bṛhadāraṇyakopaniṣad*, were added to the VTB, and the annotations of the existing five texts were extended, resulting in a total of 6,600 sentences with 47,000 word tokens. In parallel to the annotation, we also consolidated and expanded the annotation guidelines. Most importantly, however, we performed a systematic evaluation of the inter-annotator agreement (IAA) on a set of 96 unsegmented text lines with 1,885 word tokens (Biagetti et al. 2021). The outcomes of this study were surprising for all participants. While annotators of other ancient languages report high values of (un-)labeled attachment agreement scores (UAA, LAA; see e.g., 87,4% UAA and 80,6% LAA for ancient Greek according to Bamman, Mambrini, and Crane 2010), we could only achieve 69,6% UAA (76% with pre-segmented text lines) on the sample annotation. A detailed qualitative study revealed several sources of disagreement. Most importantly, (Vedic) Sanskrit has neither orthographic nor grammatical sentence boundary markers, and differences in the sentence segmentation turned out to be the central source of disagreement. Intricately connected is the question of syntactic co- versus subordination, the status of which is also contested in linguistic research (see e.g., Viti 2008). Other cases of disagreement arose from the fluid scope of Vedic particles and the lack of a diachronic valency lexicon of Vedic verbs. Overall, our observations in this phase of the VTB show that the syntactic data, even if annotated and adjudicated by multiple experts, should be taken *cum grano salis*, and linguistic theories formed on their basis need to be counterchecked in detail.

At the moment, a third version of the VTB is being prepared in the context of the research project ChronBMM. In this project, we aim at gaining a better understanding of the diachronic structure and geographical distribution of the Vedic cor-

pus by inspecting changes in linguistic structures. One important focus of CHRON-BMM are diachronic as well as diatopic developments in the Vedic syntax. As we are planning to cover the whole Vedic corpus from its beginning in the Rigveda until the late Vedic Upaniṣads, which foreshadow the epic and Buddhist literature in terms of their content and language, the scope of the VTB needs to be expanded significantly. At the moment of writing this contribution, 17 more texts have been added to the VTB which now contains approximately 12,000 sentences with 89,000 word tokens. As in the preceding two versions, we are constantly adapting the annotation guidelines, now paying special attention to the language of the sūtra texts which explain, in a concise, often almost unintelligible language, how various rituals have to be performed. As these texts are intentionally optimized for their size (see e.g., Wezler 2001), they abound in elliptic expressions whose content is often open to scholarly discussions and moreover difficult to annotate in a dependency framework. Figure 2 shows another passage from the *Śāṅkhāyanagṛhyasūtra* which prescribes what should be done in case one of the daily sacrifices has been left out. Following the organizational principles of the sūtra literature, the required actions (uttering the mantra) and side conditions (when the sacrifice has been left out) can be supplemented from other parts of the text. Therefore, the present sentence – if it is possible to call this structure a sentence at all – only provides slots for those values that have changed (i.e., the time of the day and the required mantra). While such a structure may still be intelligible for a human reader who has understood how such texts are structured and has memorized the preceding text, a machine learning algorithm for dependency annotation will have problems with joining the syntactically disconnected components, here annotated with the UD label *orphan*. The second word *atikrame*, literally 'at the omission', for example, expects a verb or a verb-like expression as its syntactic head, none of which is present in this sentence. Such challenges complicate the development of a parsing algorithm, as sketched in the next chapter.

Figure 2 Ellipsis at *Śāṅkhāyanagṛhyasūtra* 5.4.3-4: 'When a sacrifice has been left out, (one should utter the mantra) "homage to the one who enlightens the darkness" (when the sacrifice has been left out) in the morning.' Note that the elliptic construction continues in the original text, giving instructions for what should be done when the sacrifice has been left out in the morning



3. *ML approaches*

Manually annotating syntactic structures is time consuming and requires a high level of philological expertise. In order to speed up the annotation process without making compromises regarding the annotation quality, we made use of supportive machine learning techniques right from the beginning of the annotation process. The web interface of the DCS features a machine learning module that proposes a dependency label as soon as an annotator has chosen the dependent and the head of a syntactic relation. This module accesses the lexical and morphosyntactic analysis of the two connected words and feeds this information into a neural network that predicts the most probable label given the current linguistic context. The screenshot in Figure 3 shows one such instance: the dependent word *prāṇena* 'with his breath' (stem: *prāṇa-*) has been dropped on its parent *juhoti* 'he sacrifices' (stem: *hu-*), and the module outputs a list of possible labels sorted by their descending probabilities. Note that the highest scoring relation obl(ique) with 89% probability is the correct answer in this case (further details can be found in Hellwig et al. 2020).

While this built-in labeler simplifies the annotation and helps to avoid errors due to inadvertence, annotators still must construct the syntactic trees in each sentence. For building the third version of the VTB, we therefore concentrate on designing a full-fledged syntactic parser of Vedic Sanskrit which generates syntactic trees along with their labels. Dependency parsing is a well-established subfield of Natural Language Processing (see e.g., Dozat and Manning 2017 for a frequently used baseline

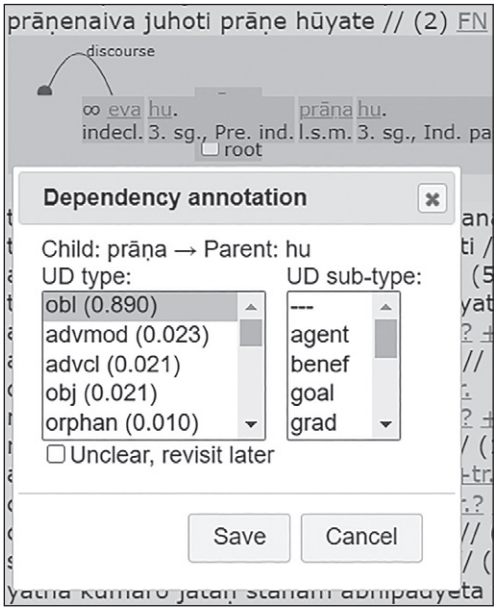
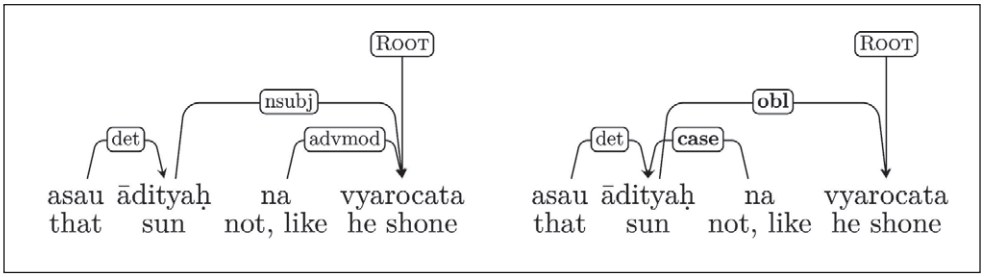


Figure 3 Screenshot of the dependency labeler integrated in the DCS

Figure 4 Alternative interpretations of *Taittirīyasaṃhitā* 2.1.4.1, left: ground truth ('That sun did not shine.'); right: predicted by a dependency parser ('He shone like that sun.'). The parsing algorithm misinterprets na as a particle of comparison, a solution that is diachronically highly improbable



model and Mrini et al. 2019 for the current state of the art in English and Chinese). There also exist recent studies on Sanskrit dependency parsing (Krishna et al. 2020; Sandhan et al. 2021), although these papers report results only for Classical Sanskrit. Since the syntactic structures and the lexicon of Classical Sanskrit differ strongly from those of the Vedic language, models and results described in these papers (e.g., UAS 87.46 / LAS 84.70 in Krishna et al. 2020) cannot simply be transferred to the VTB. Training an off-the-shelf, state of the art dependency parser (Rotman and Reichart 2019) on the VTB in its present form produces an UAS of 75,3% on the development set (LAS: 66,7%). As the human IAA was found to be 76% (see Section 2 above), we suspect that the score of parsing models may not become substantially higher than 80% without adopting the model and the linguistic assumptions it is built on. Such an assumption is, to a certain degree, confirmed when inspecting the output of such a parser. Figure 4 shows the differences between the gold annotations (left subfigure) and the parser output (right subfigure) for a prose sentence from the *Taittirīyasaṃhitā*, an early manual of the Vedic ritual. The parser has generated a valid syntactic tree that even makes sense. It was, however, not aware that the particle *na* can function as a comparison particle only in the old parts of the *Ṛgveda* (and, quite occasionally, still in the *Atharvaveda*), whereas the sentence annotated here belongs to a later layer, namely the early Vedic prose. In this diachronic layer, *na* is almost exclusively used as a negation and thus connected to the governing verb using *advmod*. Appropriately encoding such temporal and text-historical side information in a deep learning framework is therefore a central aspect of our ongoing research, and we are planning to explore existing frameworks for domain adaptation (see e.g., Ganin and Lempitsky 2015) for this task.

4. *Outlook*

As mentioned in Section 2, the version of the VTB described in this paper is a snapshot from the third phase of its development which will be completed in the course of 2021 and integrated in the next official release of the UD treebanks. Since we are planning to manually annotate approximately 500 sentences from each Vedic text contained in the DCS, we are envisaging between 15,000 and 20,000 sentences for the final version. Though still significantly smaller in size than the (combined) treebanks of Ancient Greek and Latin, this release of the VTB will nevertheless provide a sound basis for linguistic and text-historical research in Vedic and its socio-cultural environment.

Abbreviations

DCS = Digital Corpus of Sanskrit, IAA = Inter-annotator agreement, LAA = Labeled attachment agreement, LAS = Labeled attachment score, UAA = Unlabeled attachment agreement, UAS = Unlabeled attachment score, UD = Universal Dependencies, VTB = Vedic Treebank

acl = adjectival clause, advcl = adverbial clause, advmod = adverbial modifier, case = case marker, ccomp = clausal complement, det = determiner, compound:coord = coordinative compound, nmod = nominal modifier, nsubj = nominal subject, obj = object, obl = oblique.

Websites

Digital Corpus of Sanskrit (DCS): http://sanskrit-linguistics.org/dcs
UD: https://github.com/UniversalDependencies

References

Bamman, David, Mambrini, Francesco & Crane, Gregory. 2010. An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8),* Passarotti Marco C., Adam Przepiórkowski, Savina Raynaud & Frank Van Eynde (eds), 5–15.

Biagetti, Erica. This Volume. Erica Biagetti. Annotating the *RigVeda*: Challenges and methodology in parsing the earliest religious poetry of India.

Biagetti, Erica, Oliver Hellwig, Salvatore Scarlata, Elia Hackermann and Paul Widmer. 2021. Evaluating syntactic annotation of ancient languages. Lessons from the Vedic Treebank. *Old World*, Vol. 1. DOI: 10.1163/26670755-01010003

Deshpande, Madhav M. & Hock, Hans Henrich. 1991. A Bibliography of Writings on San-

skrit Syntax. In *Studies in Sanskrit Syntax*, Hans Henrich Hock (ed), 219–67. Delhi: Motilal Banarsidass.

Dozat, Timothy & Manning, Christopher D. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *5th International Conference on Learning Representations*, 1–8.

Falk, Harry. 2018. The Creation and Spread of Scripts in Ancient India. In *Literacy in Ancient Everyday Life*, Anne Kolb (ed), 43–66. Berlin / Boston: de Gruyter.

Ganin, Yaroslav & Lempitsky, Victor. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd ICML*, Francis Vach & David Blei (eds), 1180–9.

Gonda, Jan. 1975. *Vedic Literature (Samhitās and Brāhmanas)*. Vol. 1. A History of Indian Literature 1. Wiesbaden: Otto Harrassowitz.

Gonda, Jan. 1977. *The Ritual Sūtras*. Vol. 1. A History of Indian Literature 2. Wiesbaden: Otto Harrassowitz.

Hellwig, Oliver. 2010-2021. The Digital Corpus of Sanskrit. http://www.sanskrit-linguistics.org/dcs/index.php.

Hellwig, Oliver, Scarlata, Salvatore, Ackermann, Elia & Widmer, Paul. 2020. The Treebank of Vedic Sanskrit. In *Proceedings of the 12th LREC Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, et al. (eds), 5139–48.

Hock, Hans Henrich. 2013. Some Issues in Sanskrit Syntax. In *Proceedings of the Seminar on Sanskrit Syntax and Discourse Structures*, Peter M. Scharf and Gérard Huet (eds), 1–52. Paris.

Krishna, Amrith, Gupta, Ashim, Garasangi, Deepak, Satuluri, Pavankumar & Goyal, Pawan. 2020. Keep It Surprisingly Simple: A Simple First Order Graph Based Parsing Model for Joint Morphosyntactic Parsing in Sanskrit. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Trevor Cohn, Yulan He & Yang Liu (eds), 4791–4797.

Lowe, John J. 2015. The Syntax of Sanskrit Compounds. *Language* 91 (3): 71–115.

Mrini, Khalil, Dernoncourt, Franck, Tran, Quan, Bui, Trung, Chang, Walter & Nakashole, Ndapa. 2019. Rethinking Self-Attention: Towards Interpretability in Neural Parsing. *arXiv Preprint arXiv:1911.03875*.

Nivre, Joakim, De Marneffe, Marie-Catherine, Ginter, Filip, Goldberg, Yoav, Hajic, Jan, Manning, Christopher D, McDonald, Ryan, et al. 2016. Universal Dependencies V1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Johann-Mattis List, Michael Cysouw, Robert Forkel & Nicoletta Calzolari (eds), 1659–1666.

Renou, Louis. 1947. *Les écoles Védiques et La Formation Du Véda*. Paris: Imprimerie Nationale.

Renou, Louis. 1956. *Histoire de La Langue Sanskrite*. Lyon: Edition IAC.

Rotman, Guy & Reichart Roi. 2019. Deep Contextualized Self-Training for Low Resource Dependency Parsing. *Transactions of the Association for Computational Linguistics* 7. MIT Press: 695–713.

Sandhan, Jivnesh, Krishna, Amrith, Gupta, Ashim, Behera, Laxmidhar & Goyal, Pawan.

2021. A Little Pretraining Goes a Long Way: A Case Study on Dependency Parsing Task for Low-Resource Morphologically Rich Languages. *arXiv Preprint arXiv:2102.06551*.

Viti, Carlotta. 2008. The meanings of coordination in the early Indo-European languages. *Revue de Sémantique et Pragmatique* 24: 35–64.

Wezler, Albrecht. 2001. Zu der Frage des 'Strebens nach äußerster Kürze' in den Śrautasūtras. *Zeitschrift Der Deutschen Morgenländischen Gesellschaft* 151: 351–66.

Witzel, Michael. 1989. Tracing the Vedic dialects. In *Dialectes dans les littératures indo-aryennes*, Colette Caillat (ed), 97–265. Paris: Collège de France.

Annotating the RigVeda

Challenges and Methodology in Parsing the Earliest Religious Poetry of India

Erica Biagetti*

This paper discusses the annotation process of the Vedic Treebank (Hellwig et al. 2020), a corpus of selected passages from Vedic Sanskrit literature syntactically annotated according to the Universal Dependencies standard. Special attention is given to problems encountered in the annotation of the *RigVeda*, a collection of religious hymns that constitutes the oldest layer of Vedic literature and whose language is strongly conditioned by the poetic and ritual character of the text, as well as by its metrical structure. After introducing the general principles of annotation in Universal Dependencies, the paper reports the choices made in order to adapt these principles to the characteristics of Rigvedic syntax. To conclude, the paper takes Rigvedic similes as a case study, by showing how they were annotated and discussing to what extent the adopted annotation is informative for the purpose of linguistic research.

Keywords: Vedic Treebank, *RigVeda*, Universal Dependencies, comparative constructions, word order

1. *Introduction*[1]

Vedic Sanskrit (henceforth Vedic) is an ancient Indo-Aryan language, transmitted by a large corpus of poetry and prose texts. Being one of the earliest attested Indo-European (IE) languages and the precursor of Classical Sanskrit, the Vedic corpus is essential for the reconstruction of the early linguistic history of IE and a source for the study of the socio-cultural history of South Asia during the second and first millennium BCE.

The Vedic corpus has been subject to a long tradition of philological and linguistic studies, which stem from the Indian grammarian Pāṇini up to the present day. Although Vedic continues to arouse great interest in both Western and Indian scientific communities, many linguistic and above all syntactic studies have been conducted on limited portions of the corpus and in a non-replicable manner. From this point of view, the gap with other early-attested IE languages is con-

* University of Pavia.

1. I would like to thank Paul Widmer, Oliver Hellwig, and Salvatore Scarlata for welcoming me in their team for the annotation of the Vedic Treebank. In particular, I am grateful to Oliver Hellwig for his constant support and to Salvatore Scarlata for the lengthy discussions on what was the best way to annotate similes.

siderable: indeed, the last two decades have witnessed an increasing digitalization and linguistic annotation of corpora of languages such as Ancient Greek and Latin, who can now take advantage of modern methods of computational linguistics developed for modern languages. With this respect, treebanks such as the AGLDT (Bamman and Crane 2011), the IT-TB (Passarotti 2019) and the PROIEL family of treebanks (Eckhoff et al. 2018) have been welcomed as a valuable resource for linguistic and philological research on ancient languages, because they have the advantage of adding information at various levels of linguistic analysis. When mature enough and provided with valuable annotation, treebanks of historical languages permit large amount of data to be evaluated through statistical methods; furthermore, if enlarged with different texts from different stages of a language, historical treebanks allow researching the scope and effects of diachronic developments (Eckhoff, Luraghi, and Passarotti 2018; see also the introduction to this volume).

In this scenario, the Vedic Treebank (VTB; Hellwig et al. 2020; Hellwig and Sellmer in this volume) is motivated by the need of filling this gap and providing a resource that can be used for data-driven, quantitative research on Vedic syntax.

1.1. The Vedic Treebank

The VTB is a corpus of selected passages from Vedic literature, syntactically annotated according to the Universal Dependencies (UD) standard (Nivre et al. 2016). While it began as a small collection of five Vedic texts (Hellwig et al. 2020), the VTB is now being enlarged in the framework of the research project ChronBMM and will eventually cover the whole Vedic corpus from its beginning in the *Rigveda* until the late Vedic Upaniṣads (see Hellwig and Sellmer in this volume for further details). The first two versions of the VTB were released along with annotation guidelines that keep trace of those cases where the annotation deviates from the UD standard (cf. Hellwig et al. 2020; Hellwig and Sellmer in this volume, with examples on compounds' annotation). Furthermore, the second version came out with a systematic evaluation of the inter-annotator agreement (IAA) on a set of 96 sentences which were independently annotated by three of the authors (Biagetti et al. 2021). While annotators of other ancient languages report high labeled- and unlabeled-attachment agreement scores (UAA, LAA; see e.g. Bamman, Mambrini, and Crane 2009), our IAA-task only achieved 69,6% UAA on the sample annotation. Since the main source of disagreement consisted in sentence boundaries recognition, we carried out a second evaluation on pre-segmented text lines, which achieved a higher score of UAA (76%), whereas LAA remained around 63%. A detailed qualitative study revealed several other sources of disagreement, among which are the distinction between coordinated and subordinated clauses, particles' function and head, and verbal valency.

The performance of the IAA-task was an opportunity to reconsider the potential of treebanks of ancient languages and the peculiarities that distinguish them from treebanks of modern ones. In this paper I focus on the challenges found in annotating the most ancient text in the Vedic corpus, the *RigVeda* (RV), and reflect on the possibility of carrying out reliable quantitative analysis on it.

The paper is organized as follows: in Section 2, I introduce the RV and list some pros and cons of syntactically annotating this text. In Section 3, I summarize the main assumptions on which the Universal Dependencies annotation scheme stands. Taking sentence segmentation as an example, in Section 4.1 I argue that some ambiguities which are inherent in Rigvedic syntax can hardly be reduced to unique dependency relations between words, such as those provided by the UD annotation scheme. In Section 4.2, I describe the annotation process of similative and equative constructions introduced by the particles *ná*, *iva*, and *yáthā/yathā*, to which language-specific sub-relations were added in order to increase the accuracy and informativeness of the annotation. Finally, Section 4.3 tests whether said annotation is informative enough to conduct quantitative analysis of word-order patterns attested in these constructions.

2. *The* RigVeda

The RV is one of the four Vedas, collections of different types of ritual speech that accompanied public and domestic rituals.[2] The RV consists of 1028 hymns, called *sūkta* 'well-spoken (speech)', most of which are devoted to praising the gods associated with the sacrificial rituals that were performed simultaneously to the recitation. The hymns consist of verses composed in strictly regulated meters, amounting to about 10.500 verses.

According to the most accredited hypothesis, the RV, at least in its first nucleus, dates back to the second half of the second millennium BCE (Witzel 1995). A relative dating is less controversial and the division of the collection into ten books (*maṇḍala*, lit. 'circle') reflects the internal chronology of the work.[3] The core of the collection and its oldest part are books II to VII (the so-called Family Books), whereas book X is the most recent. Book VIII and I are generally younger than the Family Books. Finally, book IX differs from the others in that it is organized thematically: it is a liturgical collection of hymns to the god Soma Pavamāna ('purifying itself').

2. See Witzel and Gotō (2007: 427–466) and Jamison and Brereton (2014: 1-83) for detailed descriptions.

3. Invaluable work on the organization and history of the RV was done by Bergaigne (1886, 1887) and Oldenberg (1888: 191–270). For a summary and further explanation see Witzel (1995a, 1997).

The text was composed and transmitted orally for many centuries, even after writing had become widespread. However, the composition of Rigvedic hymns did not follow the principles of oral composition as we know it e.g. from Homeric or Serbo-Croatian epic, which originally lacked a definitive text and were (re-)composed anew at every performance by drawing on the poet's repertoire of formulas, themes, and episodes. Though orally composed and making use of traditional verbal material, each hymn of the RV was composed by a particular poet and fixed at the time of composition, being transmitted in the same form thereafter thanks to a rigorous mnemonic system which kept the error rate to an extremely low value (Jamison and Brereton 2014: 14).

2.1. Pros and cons of annotating Rigvedic syntax

In the long tradition of Vedic studies, it has often been asked whether the RV is a suitable text for conducting syntactic research or whether its ritual and poetic nature could represent an impediment to this purpose. These questions arise again when creating a treebank.

Starting from the advantages of including the RV in a diachronic treebank, the antiquity of the text and the conservative nature of its diction make the RV one of the most suitable sources for comparative studies. It displays features of an inherited Indo-Iranian culture, as shown by the pervasive structural similarities in both language and culture between the RV and Avestan texts. Furthermore, the language of the RV shows many agreements with other ancient IE languages in grammar and lexicon but also in literary sensibility, in that it represents the widespread IE genre of praise poetry (Brereton and Jamison 2020: 4).

Another point in favor of choosing the RV as a corpus for syntactic research is the stability of the text that was handed down to us. Since historical texts often come from the past in several different versions, treebank developers are faced with the problem of which critical edition to choose in making up their corpus (Eckhoff, Luraghi, and Passarotti 2018). Although some may decide to include apparatus information in their treebanks, many syntactic corpora are based on a single version of the text and thus preclude users from accessing philological information when querying the treebank. Thanks to the modality of transmission sketched above, the problem of textual variants is reduced to a minimum in the case of the RV.

Turning now to the possible disadvantages of annotating the RV, one is apparently its poetic style, since poetry is less close than prose to spontaneous language. While it is true that syntactic ambiguity is present in all languages, many aspects of Rigvedic poetry, such as the superimposition of the mythological level to the ritual level in the same passage, as well as the fondness for riddles, show a strong tendency for obscurity and complexity. Often, the poets seem to deliberately exploit grammatical ambiguity in order to reach these effects. For instance, since some grammatical categories are neutralized in particular paradigms, a single form can have two

different functions in two different adjacent constructions at the same time (Brereton and Jamison 2020: 180–181). Take for instance RV 4.1.9a:

(1) *sá* *cetayat* *mánuṣo* *yajñábandhuḥ*
3SG.NOM perceive.INJ.PRS.3SG man(M).ACC.PL/GEN.SG tie-sacrifice(M).NOM.SG
'He [=Agni] makes men perceive [=instructs them], as their tie to the sacrifice of Manu.'[4]

The form *mánuṣaḥ* (*mánuṣo* in the example due to *sandhi*) belongs to the stem *mánus-*, which means both 'man' and 'Manu', the first man and sacrificer. In this nominal paradigm, the ending *-aḥ* marks both the genitive singular and the accusative plural. Since in RV 4.1.9a the form *mánuṣaḥ* occurs between the transitive verb *cetayat* 'he makes perceive' and the compound *yajñá-bandhuḥ* 'tie to the sacrifice', Jamison and Brereton (2014) agree with Geldner (1951) in giving it a double interpretation: in this passage, it is both the object of the preceding verb (accusative plural 'men') and a possessive modifier of the compound (genitive 'of Manu'). While this solution seems to be the most accepted one, treebank annotation forces the annotator to choose only one of the two functions of *mánuṣaḥ*, as shown by the two graphs in Figure 1 and Figure 2.

Figure 1 Labeled graph for *mánuṣaḥ*.ACC.PL 'men' as object of *cetayat*

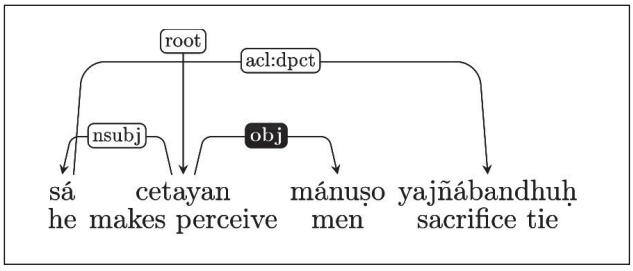
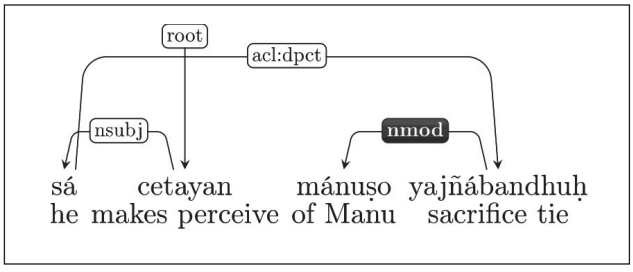


Figure 2 Labeled graph for *mánuṣaḥ*.GEN.SG 'of Manu' as modifier of *yajñábandhuḥ*



4. Unless otherwise stated, translations are taken from Jamison and Brereton (2014).

Another disadvantage might be found in the metrical structure of the text, which allegedly constraints its syntax. However, it has been shown that metric and linguistic units usually coincide in the RV, since clauses tend to be comprised within the border of a verse, hemistich, or text line (*pāda*) (Dunkel 1985; Gunkel and Ryan 2018). For instance, verses consisting of trimeters, such as the *triṣṭubh* and the *jagatī*, contain complex sentences which exhibit coordinative and subordinating strategies. On the contrary, verses consisting of dimeters, such as the *gāyatrī*, tend to incapsulate syntactically simple clauses. Finally, there are very few examples of syntactic enjambment between verses, which occur in highly dramatic contexts for expressive purposes (Brereton and Jamison 2020: 189). Thus, we might say that Vedic meters substantially reflect the articulation of discourse, rather than constrain it (Viti 2007: 30).

Another possible argument against the use of a metrical text such as the RV for syntactic analysis, is the one of formulaicity, since the latter also seems to crystallize syntax or bend it to the rigid template of the formula. Again, this judgment is mitigated by the type of formulaicity that characterizes the RV: as anticipated above, the compositional technique of the RV makes little use of fairly sizable, metrically defined, and invariant formulas (*ready-made surface formulae*; Kiparsky's 1976: 83). Rather, the RV consists of a texture of so-called *deep-structure* or schematic formulas, which make up the poets' repertoire, but which take different instantiations in the text thanks to lexical or grammatical substitution, scrambling, semantic reversal, or metrical variation (Jamison and Brereton 2014: 14, cf. also Jamison 1998).

The aspect of formulaicity that really affects the syntax is the presence of recurrent topics that determine the narrow-shared universe of discourse in the text. Indeed, deep-structure formulas and the shared knowledge of the nature of Rigvedic ritual practices and religious beliefs allow the poets to refer to such knowledge with deliberately elliptical expressions, truncated or twisted formulae, and brief allusions. The consequence of this brachylogical style is perhaps a higher incidence of null arguments or coordination reduction and gapping in the corpus; however, these phenomena are widespread in prose texts too, which suggest that they were allowed by Vedic grammar.

In conclusion, one must be aware of which questions we can ask a treebank of the RV. For instance, if we are interested in investigating the development of subordination in Vedic, we will have to keep in mind that poetry can indeed influence the choice of some linguistic elements, but that it will more probably affects the lexicon rather than functional words such as subordinating conjunctions (Viti 2007: 30). The study of word order is a different matter: while Ryan and Gunkel (2015) have shown that in metrically neutral contexts[5] the genitive precedes the noun in 92% of cases, if

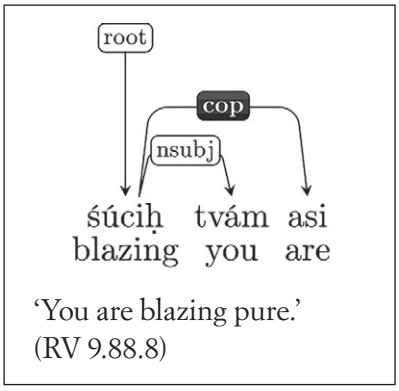
5. Ryan and Gunkel (2015) extracted all swappable bigrams from the RV, that is all bigrams in which both orders are metrically equivalent: e.g. *śubhas patī* ~ **patī śubhas* 'lords of beauty'.

one includes all metrical contexts in a random sample of about 2950 sentences from the Rigvedic treebank, the percentage of GN drops to 66%. This suggests that meter might indeed play a role in determining the frequency of a given word order, but that all attested word orders had to be somehow accepted by the Vedic grammar.

3. *Universal Dependencies*

Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages (Nivre et al. 2016).[6]

Figure 3 Copula annotation



'You are blazing pure.'
(RV 9.88.8)

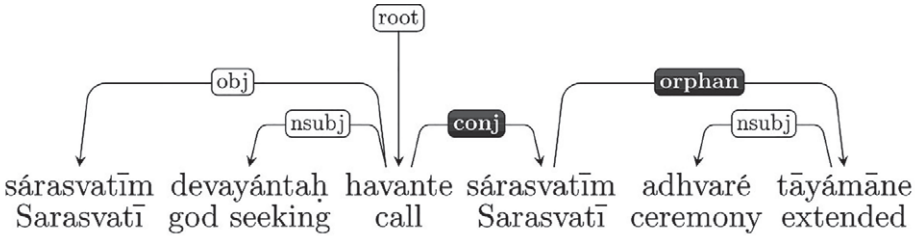
Syntactic annotation in the UD scheme consists of typed dependency relations (*deprel*) between words. The following principles are observed in the annotation in order to maximize parallelism while accounting for cross-linguistic differences. Dependency relations hold primarily between content words, rather than being mediated by function words (*primacy of content words*). Thus, case-marking elements like adpositions and clitic case markers are treated as dependents of the nouns they attach to or introduce. Coordination follows a similar treatment, in that the leftmost conjunct constitutes the head, while other conjuncts as well as the coordinating conjunction depend on it. Finally, auxiliary and copulas depend on the lexical predicate, rather than being the head of the clause (Figure 3).

Even if the major role of syntactic analysis is to represent function, the scheme also provides for some structural analysis, distinguishing between a) nominal phrases, b) clauses headed by a predicate (most commonly verbs, but also adjectives, or nominals), and c) miscellaneous other kinds of modifier words. This distinction is clearly encoded in dependency labels. For example, a verb's adverbial modifier is labeled a) `obl`, b) `advcl`, or c) `advmod` depending on which of the three categories above it belongs to.

The principle of the primacy of content words has consequences on the annotation of ellipsis. Differently from other formalisms, such as the PROIEL scheme, UD does not make use of empty nodes in order to represent ellipsis or gapping. Instead, UD marks all kinds of ellipsis by promoting a member of the elliptical clause to the

6. The latest version (2.8, released on 2021-05-15) consulted during the preparation of this paper includes 202 treebanks of 114 languages.

Figure 4 Annotation scheme for verb ellipsis



'Sarasvatī do those seeking the gods invoke, Sarasvatī when the ceremony is being extended.' (RV 10.17.7)

head position on the base of a "coreness" hierarchy.[7] The promoted member takes the syntactic relation that the elided element would otherwise bear; to signal that the dependency structure is incomplete, all non-promoted dependents of the elided element receive the relation `orphan`. Cf. Figure 4, which represents the treatment of ellipsis in coordination: as a consequence of the elision of the verb *havante* 'they call' in the second conjunct, the object *sárasvatīm* 'Sarasvatī' is promoted to the head position of the coordinate clause (`conj`), whereas the adjunct *tāyámāne* (itself the head of a *locativus absolutus*) depends on it via the relation `orphan`.

4. *Challenges and methodology in the annotation of the RV*

In this section, I will first outline the methodology adopted in addressing syntactic ambiguities which are inherent in the Vedic language, providing examples from the level of sentence segmentation (4.1). Second, taking similative and equative constructions as an example, I will show that it is sometimes possible to customize the annotation in order to increase its granularity and make it suitable for one's own research (4.2). Finally, I test whether the annotation is informative enough to conduct quantitative analysis of word-order patterns attested in these constructions (4.3).

4.1. Tackling ambiguity: sentence boundaries

Sentence segmentation is the task of dividing a string of written language into its component sentences. In English and other languages using punctuation, the full stop/period character is a reasonable approximation, as well as other punctuation

7. Orphaned dependents are considered for promotion in the following order: `nsubj` > `obj` > `iobj` > `obl` > `advmod` > `csubj` > `xcomp` > `ccomp` > `advcl` > `dislocated` > `vocative`.

Figure 5 Hemistich and verse boundaries marked by | and || respectively (RV 1.89.1)

आ नो॑ भद्राः क्रत॑वो यन्तु वि॒श्वतोऽद॑ब्धासो॒ अप॑रीतास उ॒द्भिदः॑ ।
दे॒वा नो॒ यथा॒ सद॒मिद् वृ॒धे अस॒न्नप्रा॑युवो रक्षि॒तारो॑ दि॒वेदि॑वे ॥ १.०८९.०१

á no bhadráḥ krátavo yantu viśvátaḥ-ádabdhāso áparītāsa udbhídaḥ |
devā́ no yáthā sádam íd vṛdhé ásann áprāyuvo rakṣitā́ro divé-dive ||

'Let auspicious ideas come here to us from all sides – undeceivable, uncircumscribable, bursting out – so that the gods will be (ready) to increase us always, will be our unfaltering protectors every day.'

marks such as question mark, colon, and semicolons. Developers of treebank of ancient languages which originally lacked punctuation marks often choose to perform sentence segmentation automatically, on the base of the punctuation adopted by the digital version of the text.

As many other ancient languages, the Sanskrit writing system had originally no punctuation and the editorial conventions adopted for the RV differ from those of Western texts. Here, the so called *daṇḍa* (a single vertical bar |) and double *daṇḍa* (a double bar ||) do not mark syntactic units, but metrical and recitational units such as hemistiches and verses (Figure 5).

Although we have seen above (Section 2.1) that metrical and syntactic units often coincide in the RV, this is not always the case and performing sentence segmentation automatically on the base of hemistich or verse boundaries does not seem to be a safe strategy. For this reason, the VTB reproduces the metrical structure of a hymn, signaling *pāda*, hemistich, and verse boundaries, but leaves it open to the annotator to decide whether these correspond to sentence boundaries. While avoiding applying modern Western conventions to ancient texts has the advantage of preserving their linguistic reality, the absence of "ready-made" sentence boundaries confronts the annotator with the fuzziness that characterizes the distinction between independent sentences, coordinate clauses, and subordinate ones in the RV. This is due to the scarce grammaticalization of clause linkage that characterizes Vedic as well as many other ancient IE languages (Viti 2008).

Asyndetic coordination or juxtaposition is typical of oral languages insofar as intonation can express by itself the coordinating function (Dik 1968: 33). Given the high prestige attributed to orality in ancient India, it is natural to find that the earliest documents abound in asyndetic constructions much more than syndetic devices (Viti 2008: 37). When annotating the RV, the lack of explicit markers for coordina-

Figure 6 Coordinative interpretation of RV 2.2.7

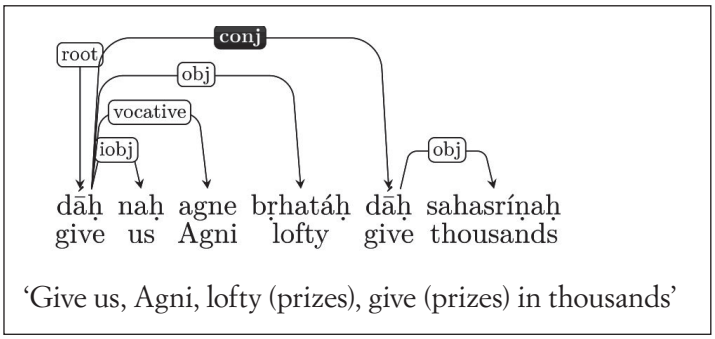
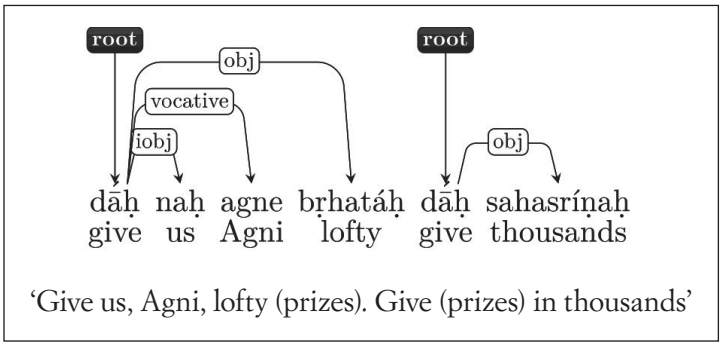


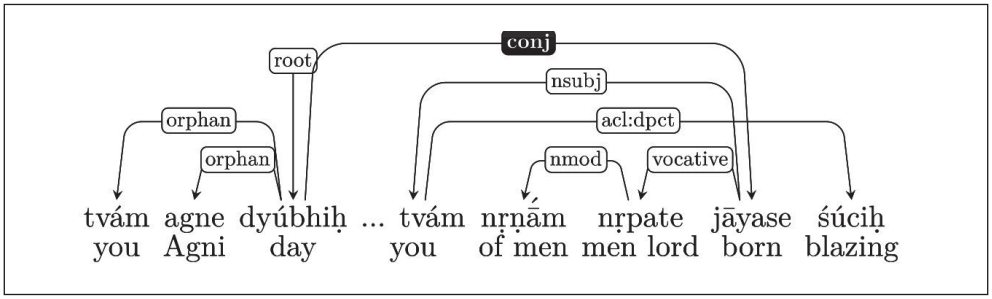
Figure 7 RV 2.2.7 as two self-standing sentences



tion makes it hard to decide whether two clauses should be regarded as coordinated asyndetically (Figure 6) or as self-standing sentences (Figure 7).

Although it is not possible to establish a rationale that is valid in each and every case, some grammatical and stylistic criteria can help the decision. Compare for instance examples (2) and (3), the two initial verses of RV 2.1. Example (2) is a clear case of leftward gapping. Gapping is a kind of ellipsis in coordination, where the omitted constituent is the verb and there are at least two contrasting elements in each clause; these elements are called the contrast-points in the non-gapped clause and the remnants in the gapped one(s) (Hudson 1989: 67; Gaeta and Luraghi 2001: 90). In example (2), the only verb (*jāyase* 'you are born') appears in *pāda* d, where it occurs with the contrast points *tvám* 'you' (the subject) and *śúciḥ* 'pure' (a secondary predicate). *Pādas* a to c, instead, present two gapped clauses each, which in turn contain the remnants, i.e. the emphatically repeated subject *tvám* and an adverbial modifier (cf. Inst. *dyúbhis* 'throughout the days', Abl. *adbhyás* 'from the waters', secondary predicate *āśuśukṣáṇis* 'eager to blaze'). Thus, the whole verse would be annotated as a single sentence, as represented in Figure 8.

Figure 8 Annotation of RV 2.1.1 as a single sentence



(2) a. *tvám* *agne* *dyúbhis* *tvám* *āśuśukṣáṇis*
2SG.NOM Agni.VOC day.INST.PL 2SG.NOM shining.NOM.SG.M

b. *tvám* *adbhyás* *tvám* *áśmanas* *pári*
2SG.NOM water.ABL.PL 2SG.NOM stone.ABL.PL LP

c. *tvám̀* *vánebhyas* *tvám* *óṣadhībhyas*
2SG.NOM forest.ABL.PL 2SG.NOM plant.ABL.SG

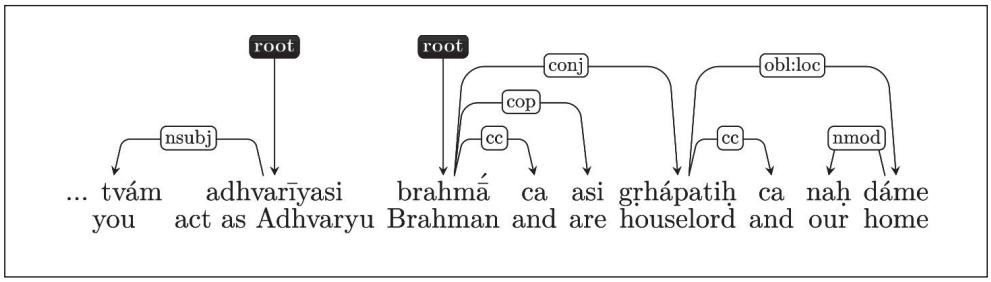
d. *tvám̀* *nr̥̄ṇā́m̀* *nr̥̄ pate* *jāyase*
2SG.NOM man.GEN.PL lord.VOC be-born.IND.PRS.M.2SG

śúciḥ
pure.NOM.M

'You, Agni, (are born) throughout the days, you (are born) eager to blaze here; you (are born) from the waters, you from the stone, you from the trees, you from the plants, you, men-lord of men, are born blazing.' (RV 2.1.1; adapted from Jamison and Brereton 2014)

At first glance, example (3) looks similar to example (2), because here too a verb (*asi* 'you are') in the last *pāda* seems to be shared by a series of gapped clauses. However, a series of grammatical and stylistic considerations point towards a division of the passage into independent sentences (Figure 9), rather than to an analysis as a series of coordinated clauses with gapping. First, unlike lexical verbs, the copula is regularly omitted in Vedic so that there is no need to take *ási* in the last *pāda* as being shared by the other nominal predicates. Second, the symmetry we saw in the previous verse is broken in this second verse by different forms of the pronoun *tvám* (*táva*.GEN ... *tvám*.NOM ...) and by the presence of the lexical verb *adhvarīyasi* 'you act as Adhvaryu' that breaks the series of nominal predicates.

Figure 9 Annotation of RV 2.1.2 as a series of independent sentences



(3)	*táva*	*agne*	*hotráṃ*	*táva*	*potrám*	*ṛtvíyam*
	2SG.GEN	Agni.VOC	duty-of-Hotri.NOM	2SG.GEN	duty-of-Potri.NOM	due.NOM
	táva	*neṣṭráṃ*	*tvám*	*agníd*	*ṛtāyatáḥ* |	
	2SG.GEN	duty-of-Nestri.NOM	2SG.NOM	Agnidh.NOM	pious.PTCP.PRS.GEN.M	
	táva	*praśāstráṁ*		*tvám*	*adhvarīyasi*	
	2SG.GEN	duty-of-Praśastri.NOM		2SG.NOM	act-as-Adhvaryu.IND.PRS.2SG	
	brahmā́	*ca*	*ási*	*gṛhápatiś*	*ca*	
	Brahman.NOM	and	be.IND.PRS.2SG	lord-of-house(M).NOM	and	
	no	*dáme* ||				
	1PL.GEN	house(M).LOC				

'Yours, Agni, is the office of Hotar; yours that of Potar in its turn; yours that of Neṣṭar; you are the Agnidh [=Fire-Kindler] of the one who follows truth. Yours is the office of Praśāstar; you act as Adhvaryu; you are both the Brahman-priest and the houselord in our home.' (RV 2.1.2)

In some cases, however, no grammatical or stylistic consideration seem to be of any help and the only rationale that can be adopted is the one proposed in section 2.1, according to which one hemistich correspond to one sentence, possibly restricted to a *pāda* or extended to the entire verse.

4.2. Metodology: annotating Rigvedic similes

Similative and equative constructions (henceforth: *similes*) encode similarity between a comparee (CPREE) and a standard (STAND) with respect to some action or property, called parameter (PAR), and by means of a standard marker (STM; Haspelmath and

Buchholz 1998; Treis 2017). While equative constructions encode quantitative comparison of equality (4)a, similative constructions encode qualitative comparison, or comparison of manner (4)b.

(4) a. *Peter is as tall as Susan.*
b. *Peter runs like a hare.*

In the RV, similes introduced by the STMs *ná* 'like', *iva* 'id.' and *yáthā/yathā* 'id.' are characterized by lack of the verb in the STAND and by formal and functional parallelism between CPREE and STAND (Bergaigne 1887; Jamison 1982; Pinault 1997). Quantitative and qualitative comparison are encoded by the same constructions and are therefore nearly impossible to distinguish. Studies on Vedic similes (Jamison 1982: 252; Pinault 1997: 310) recognize three main configurations of STAND(s) and CPREE(s). Single similes can take an adjectival predicate as PAR or a verbal one, as in (5). In both cases, the STAND is in the same case as the CPREE.

(5)	*ví*	*ślóka*	*etu*	*pathyà_iva*	*sūréḥ*
	LP	signal_call.NOM.SG	go.IMPV.3SG	pathway.NOM.SG_like	patron.GEN.SG
	PAR-	CPREE-	-PAR	STAND_STM	-CPREE

'Let the signal-call of the patron go forth afar like a pathway.' (RV 10.13.1)

Double similes are characterized by the presence of two parallel elements in the CPREE and in the STAND, as in example (6).

(6)	*matáyaḥ*	*rihánti …*	*índraṁ*	*vatsáṁ*	*ná*	*mātáraḥ*
	thought.NOM.PL	lick.PRES.3PL	Indra.ACC.SG	calf.ACC.SG	like	mother.NOM.PL
	$CPREE_i$-	PAR	-$CPREE_j$	$STAND_j$-	STM	$STAND_i$

'Thoughts lick […] Indra like mothers a calf.' (RV 3.41.5)

Less often, similes may be triple, as shown by example (7), in which both CPREE and STAND consists of three elements: a nominative subject (*yás*.REL and *sū́rya* 'sun'), an accusative argument indicating the path of extension (*páñca kṛṣṭī́ḥ* 'across the five people' and *apás* 'across the waters'), and an instrumental adjunct (*śávasā* 'with strength' and *jyótiṣā* 'with light').

(7)	*sadyáś*	*cid*	*yáḥ*	*śávasā*	*páñca*	*kr̥ṣṭ́íḥ*
	in_one_day	PTC	REL.NOM.SG.M	strength(N).INST.SG	FIVE.ACC.PL.F	people(F).ACC.PL
			CPREE$_i$	CPREE$_z$		CPREE$_j$
	sū́rya	*iva*	*jyótiṣā*	*apás*	*tatā́na*	
	sun(M).NOM.SG	like	light(N).INST.SG	water(F).ACC.PL	stretch.IND.PF.3SG	
	STAND$_i$	STM	STAND$_z$	STAND$_j$	PAR	

'Who just in a single day stretches across the five peoples with his vast power, like the sun across the waters with his light.' (RV 10.178.3ab)

UD guidelines provide annotation standards for phrasal comparatives (8)a and for clausal ones (i.e. with two verbs (8)b). In the former, the standard is linked to the parameter via the relation `obl`, while the standard marker depends on the standard via the relation `case` (Figure 10). The verb of the comparative clause is instead attached to the main verb through the relation `advcl`, the standard marker depending on it via `mark` (Figure 11).

(8) a. *Peter is as tall as Susan.*
b. *I put as much flour as the recipe called for.*

Figure 10 Annotation scheme for phrasal comparatives

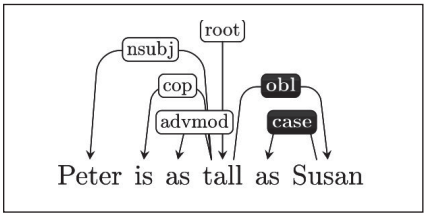


Figure 11 Annotation scheme for clausal comparatives

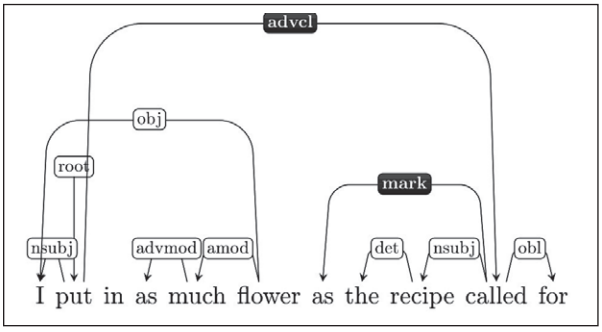
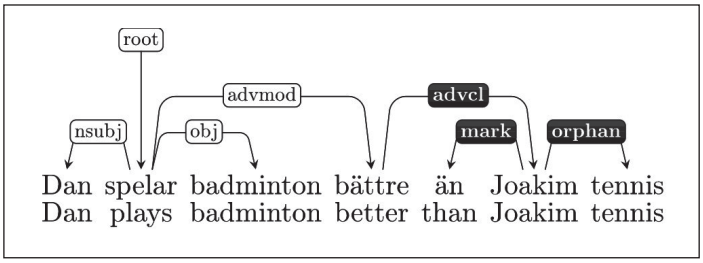


Figure 12 Annotation scheme for gapping in comparison



The annotation of gapping structure in comparative clauses is mentioned in the report of a working group dedicated to comparative constructions. The report provides the Swedish sentence in (9) as an example of gapping in comparative clauses and suggests analyzing such comparative gapping using the `orphan` relation, much like the more widespread coordinate gapping (cf. Figure 12).

(9)	*Dan*	*spelar*	*badminton*	*bättre*	*än*	*Joakim*	*tennis*
	Dan	play.PRS	badminton	better	than	Joakim	tennis

'Dan plays badminton better than Joakim (does) tennis.'

Since the verb is systematically omitted in their STAND clause, Rigvedic similes introduced by *ná*, *iva*, and *yáthā/yathā* are probably not best treated as synchronically involving verb ellipsis or gapping, which are generally optional.[8] However, from a descriptive point of view (i.e. for the purposes of annotation) it is useful to analyze simple similes as cases of verb ellipses in which the promoted element has no dependents, and double and or triple similes as cases of gapping, in which the second remnant is attached to the promoted one with the relation `orphan`.

In UD, there are no relations designed specifically to mark comparative constructions: phrasal comparatives are simply assimilated to other obliques (`obl`), whereas comparative clauses are treated in the same way as other adverbial clauses (`advcl`). Similarly, standard markers take the same *deprel* as other function words such as adpositions (`case`) and subordinating conjunctions (`mark`).

Similes are the most frequent trope found in the RV. Although the literature abounds in contributions on their syntax (Bergaigne 1887, Jamison 1982), origin (Vine 1978, Pinault 1985), or on the distribution of the standard markers (Pinault 1997, Viti 2002), some of these questions have not been answered yet and could be

8. Different theories on the origin of *ná*-similes suggest that ellipsis and gapping in fact played a role in their development (cf. Section 4.3). Similes introduced by *yáthā* also seem to originate from clausal comparison.

addressed anew thanks to a quantitative study on an annotated corpus of similes. However, since the particles *ná*, *iva*, and *yáthā/yathā* have other functions beside that of STM, and since comparison is also expressed by other strategies, it is necessary to increase the informativeness of the annotation, in order to be able to make granular and targeted queries on different types of constructions.

In order to represent the syntax of similes in detail, the VTB distinguishes different subtypes of comparative constructions, listed in Table 1.

As shown in Table 1, the VTB formally distinguishes similes with ellipsis (annotated with `obl` and `case`) from similes with gapping (annotated with `advcl` and `mark`). In addition to the universal dependency taxonomy, UD allows the employment of language-specific extensions that capture peculiar constructions find in a given language or in a group of languages. These extensions are regarded as subtypes of existing UD relations and have the format `universal:extension`: for instance, `obl:manner` stands for the language-specific `manner` extension of the UD relation `obl`. In the VTB, the sublabel `:sim` attached to the relations `case` and `mark` allows the user to easily retrieve all particles that introduce phrasal similes and to distinguish them from those that introduce clausal similes (which take `mark` alone). Finally, the sublabels `:grad` and `:manner` added to the relations `obl` and `advcl` allow, on the one hand, to distinguish between quantitative (`:grad`) and qualitative comparison (`:manner`) and, on the other hand, to distinguish standards of comparison from other kinds of adverbial modifiers. The sublabels `:grad` and `:manner` are given on a lexical basis, e.g. to gradable vs. non gradable adjectives respectively.

Table 1 Comparative constructions with their respective annotation

Construction	Example	Annotation (dependent → relation → head)
predicative sim.	'Agni is like the sun.'	*like* → `case:sim` → *sun*
sim. with ellipsis	'Agni is bright like the sun.'	*like* → `case:sim` → *sun* → `obl:grad` → *bright*
	'The lightning bellows like a cow.'	*like* → `case:sim` → *cow* → `obl:manner` → *bellow*
sim. with gapping	'Thoughts lick Indra like mothers a calf.'	*like* → `mark:sim` → *mothers* → `advcl:manner` → *lick;* *calf* → `orphan` → *mothers*
clausal sim.	'Just as you drank the previous soma drinks, so take a drink today.'	*as* → `mark` → *drank* → `advcl:manner` → *drink;* *previous drinks* → `obj` → *drank;* *so* → `advmod` → *drink*

Table 2 N. of similes for each STM in the VTB and in the whole RV

STM	VTB	RV
ná	506	1300
iva	295	1023
yáthā / yathā	75	75
total	876	2398

4.3. Evaluating the annotation

In this section, I query the treebank in order to test the informativeness of the annotation and whether it can help answering some still open questions about the diachronic development of the syntax of similes. The corpus employed for this case study contains 2958 sentences as a whole and 876 similes that were annotated according to the scheme indicated above (Table 2).

The addressed question concerns the origin of similes introduced by *ná* and their syntax. In the RV, *ná* marks both negation and comparison. The polysemy is not due to homophony but is the result of a semantic shift from negation to comparison.[9] Two hypotheses on the origin of *ná*-similes are found in Vine (1978) and Pinault (1985). According to Vine, the complete overlap of comparative and negative *ná* (which are otherwise in complementary distribution) in slot 9 of the trimeter is of utmost importance for understanding the origin of the former from the latter. Vine suggests that comparative constructions introduced by *ná* may originate from coordinate negative constructions with ellipsis of the verb, since in these constructions the second negation regularly falls on slot 9 of the trimeter. Example (10) is an instance of coordinate negative constructions of this type:

(10)	*ná*	*yáṁ*	*járanti*	*śarádo*	*ná*	*mā́sā*
	NEG	REL.ACC.SG.M	make_old.PRS.3PL	year.NOM.PL	NEG	month.NOM.PL
	*#**ná***	…		…	***ná***	10-11 #

'Whom neither years nor months make old.' (RV 6.24.7; Vine 1978: 181)

On the base of other similarities between Vedic and Baltic as well as Slavic languages in the domain of comparison, Pinault (1985) suggests instead that similes introduced by *ná* may originate from expressions similar to the so-called negative paral-

9. The direction of change is suggested by the fact that cognates of negative *ná* are found in most ancient IE languages (Dunkel 2014 [LIPP]: 546).

lelism. This is a rhetorical device typical of Slavic and Baltic folk literature, which presents the following structure:

(11) (On this birch a cuckoo cries)
It is **not** a cuckoo that cries,
the mother of this one moans.
(fragment from a Russian bylina; adapted from Pinault 1985: 130)

Pinault suggests that some similes in the RV can be interpreted as cases of negative parallelism, where the first verb is regularly omitted: in these expressions, the comparison precedes the predicate (order STAND - PAR), and they have the same meaning both if they are interpreted negatively or comparatively (12):

(12)

vér	***ná***	*druṣác*
bird.NOM.SG	NEG/like	wood_sitting.NOM.SG

camúvor	*ā́*	*asadad*	*dhárih*
cup.LOC.DU	LP	seat.IND.AOR.3SG	tawny.NOM.SG

'It is not a bird sitting in the wood, the tawny one (Soma) has taken his seat in the two cups.' > 'Like a bird sitting in the wood the tawny one (Soma) has taken his seat in the two cups.' (RV 9.72.5d)

Pinault (1985) argues that the negative parallelism is but a relic of an ancient stratum of Rigvedic diction and that not all comparisons must reflect such a construction. Instead, the comparative reading of *ná* must have survived and spread to more recent layers thanks to the existence of two other comparative strategies, which shared the order STAND - PAR with similes introduced by *ná*: these strategies are the so-called comparative compounds (e.g. *vā́ta-jūta-* lit. 'wind.STAND-swift.PAR' = 'swift as the wind') and analytic comparisons with an ablative STAND (e.g. *manáso* 'thought.STAND' *jávīyas* 'swifter.PAR' = 'swifter than thought'; Pinault 1985: 138-143).

Thus while, according to Vine (1978), similes introduced by *ná* originate from constructions in which the PAR (verb) preceded the STAND (slot 9 of trimeters), according to Pinault (1985) they originate from constructions with the opposite order of STAND and PAR.

Pinault's hypothesis has the advantage of combining a syntactic explanation with a semantic one, whereas Vine's hypothesis is convincing only on the syntactic level.[10] Furthermore, typological studies on equative and similative construc-

10. And perhaps not even on the syntactic level, since *iva* too is found as the STM of a negated PAR (see Pinault 1985 for a detailed discussion).

Table 3 N. of STAND – PAR and PAR – STAND orders in similes introduced by *ná*

Order	Abs. n.	%
STAND - PAR	322	63%
PAR - STAND	184	36%
TOTAL	506	

tions have shown that the order STAND - PAR correlates with the O - V order (Andersen 1983; Haspelmath's 2017: 26 *Generalization* 2). Since in Vedic objects tends to precede the verb (cf. Ryan and Gunkel 2015), the STAND - PAR order is the expected one for this language.

As shown by Table 3, the order STAND - PAR predicted by Haspelmath's Generalization 2 is confirmed by a quantitative analysis performed on the treebank (see Appendix 1 for the queries used in this study). However, Table 3 shows that, while the STAND - PAR order is indeed the most frequent in similes introduced by *ná*, the PAR - STAND one is by no means rare.

At this point one might wonder whether the PAR - STAND order is to be attributed to metrical needs (see Section 2.1) or if instead other factors come into play in determining the order of the two elements. Treebank annotation cannot take into account the metrical context in which a similie occurs, but it can help explore the latter hypothesis. A more refined query which distinguishes similes with ellipsis (e.g. 'Agni is bright like the sun', whose STANDs take the deprels `obl:manner` and `obl:grad`), from similes with gapping (e.g. 'Thoughts lick Indra like mothers a calf', whose STANDs take the label `advcl:manner`) yields the following results:

As Table 4 shows, by selecting only similes with ellipsis, the STAND - PAR order increases (69%), thus supporting the predictions made by Generalization 2. If, instead, only similes with gapping are included in the query, the STAND - PAR percentage drops to almost 50%.

A possible explanation of the higher frequency of PAR - STAND order in similes with gapping may come from typological studies on gapping. Mallison and Blake (1981) and Gaeta and Luraghi (2001) have shown that in free word order languag-

Table 4 N. of STAND - PAR and PAR - STAND orders in *ná*-similes with ellipsis and gapping

Order	Similes with ellipsis		Similes with gapping	
STAND - PAR	232	69%	90	52%
PAR - STAND	104	30%	80	47%
TOTAL	336		170	

es such as Ancient Greek and Latin, the specific order of gapping and the relative position of the contrasted constituents seem to depend on pragmatic, rather than strictly syntactic factors. A preference for rightward gapping, i.e. elision of the verb in the second clause (e.g. *Rose studies Greek and John Ø Latin*) has been attributed to the tendency of language processing to favor anaphoric processes over cataphoric ones. This makes leftward gapping (e.g. *Rose Ø Greek and John studies Latin*) cross-linguistically more subject to restrictions with respect to verb position, relative order of the constituents, and type of employed verbs (Hudson 1989; Gaeta and Luraghi 2001: 108).

Treebank annotation allows investigating the order of gapping in Vedic as well. From a preliminary inquiry, it results that the typology of gapping in the RV resembles that of other IE languages like Ancient Greek and Latin.[11] Suffice here to say that, while both orders are attested, in the annotated portion of the RV rightward gapping occurs 41 times, whereas leftward gapping has only 17 occurrences.

Since in similes the verb is always omitted in the STAND clause, similes with PAR - STAND order feature rightward gapping (e.g. *Thoughts lick Indra like mothers Ø a calf*), whereas similes with order STAND - PAR feature leftward gapping (e.g. *Like mothers Ø a calf, thoughts lick Indra*). Thus, the preference for rightward gapping would explain the higher frequency of the PAR - STAND order in similes with gapping than in similes with ellipsis.

5. *Summary and conclusion*

In this paper, I have discussed some challenges encountered in annotating the RV within the VTB. In Section 2, I first introduced the main features of the text, I presented some arguments for and against the possibility and usefulness of syntactically annotating an ancient poetic text such as the RV, and then considered which questions we can ask a treebank of the RV (2.1).

After summarizing the main characteristics of the UD annotation scheme (Section 3), in Section 4 I presented some challenges that the annotator is confronted with when annotating the syntax of the RV: in Section 4.1, taking the problem of sentence segmentation as an example, I have shown the difficulty of reducing the language of the RV to unique dependency relations between words, such as those provided by the annotation scheme. The same discussion could be extended to several other domains of Vedic syntax, such as the scope of preverbs (Casaretto and Schnei-

11. The query employed for this survey only retrieves cases of gapping where the omitted verb is a finite verb. Gapping involving omission of the copula or of non-finite forms needs to be further investigated (the query is reported in the Appendix).

der 2015) or secondary predicates (Casaretto 2020). Taking comparative constructions introduced by the particles *ná*, *iva*, and *yáthā / yathā* as an example, in Section 4.2 I suggested that, thanks to the UD openness to language-specific extensions, it is sometimes desirable to customize the annotation in order to increase its granularity and informativeness.

Finally, in section 4.3, I conducted a small case study on similes introduced by *ná* in order to test the informativeness of the annotation. This case study shows that treebank-backed analysis has the advantage of providing quantitative data and highlighting tendences in a given syntactic phenomenon. For instance, we have seen that the treebank allows to explain the tendency of similes with gapping to choose the PAR - STAND order more often than similes with ellipsis. On the other hand, the kind of information stored in the treebank does not allow to check whether in similes with gapping and PAR - STAND order (that is, rightward gapping), the position of the STM *ná* tends to fall in a given slot of the various meters (e.g. slot 9 of the trimeter as suggested by Vine 1978). Thus, information extracted from the treebank must be used in synergy with other tools that account for the metrical features of the text, such as critical editions, commentaries, and digital resources that allow querying the text according to different types of metadata.[12]

Abbreviations

1 = first person, 2 = second person, 3 = third person, ACC = accusative, AOR = aorist, CPREE = comparee, DAT = dative, DU = dual, F = feminine gender, GEN = genitive, IMPV = imperative, INS = instrumental, LOC = locative, LP = local particle, M = masculine gender, N = neuter gender, NEG = negation, NOM = nominative, PAR = parameter, PL = plural, PF = perfect, PRS = present, PTCP = participle, REL = relative pronoun, SG = singular, STAND = standard, STM = standard marker, VOC = vocative

`acl` = adjectival clause (clause modifier of noun), `advcl` = adverbial clause modifier, `advmod` = adverbial modifier, `amod` = adjectival modifier, `aux` = auxiliary, `case` = case marking, `cc` = coordinating conjunction, `ccomp` = clausal complement, `conj` = conjunct, `cop` = copula, `csubj` = clausal subject, `det` = determiner, `discourse` = discourse element, `:grad` = grade, `iobj` = indirect object, `mark` = marker, `nmod` = nominal modifier, `nsubj` = nominal subject, `obj` = object, `obl` = oblique, `:sim` = simile, `xcomp` = open clausal complement.

12. VedaWeb, for instance, is a web-based, open-access platform which is part of the Cologne South Asian Languages and Texts (C-SALT) and aims to facilitate linguistic and philological research on Old Indic texts. The text corpus is made available in a digitally accessible as well as morphologically and metrically annotated form, searchable for lexicographic and corpus-linguistic criteria.

Websites

Universal Dependencies: https://universaldependencies.org
UD annotation guidelines for comparative constructions: https://universaldependencies.org/u/overview/specific-syntax.html#comparatives
UD report of the working group on comparative constructions: https://universaldependencies.org/workgroups/comparatives.html#working-group-on-comparative-constructions
VedaWeb: https://vedaweb.uni-koeln.de

References

Andersen, Paul Kent. 1983. *Word Order Typology and Comparative Constructions* [Current Issues in Linguistic Theory 25]. Amsterdam: John Benjamins.

Bamman, David, Mambrini, Francesco & Crane, Gregory. 2009. An ownership model of annotation: The Ancient Greek dependency treebank. In *Proceedings of the eighth international workshop on treebanks and linguistic theories* (TLT8), Marco C. Passarotti, Adam Przepiórkowski, Savina Raynaud & Frank Van Eynde (eds), 5–15.

Bamman, David & Crane, Gregory. 2011. The ancient Greek and Latin dependency treebanks. In *Language technology for cultural heritage. Selected papers from the LaTeCH workshop series*, Caroline Sporleder, Antal van den Bosch & Kalliopi Zervanou, 79–98. Berlin, Heidelberg: Springer.

Bergaigne, Abel. 1887. *La syntaxe des comparaisons védiques*. Mélanges Renier, 75–101. Paris: Vieweg.

Biagetti, Erica, Oliver Hellwig, Salvatore Scarlata, Elia Hackermann and Paul Widmer. 2021. Evaluating syntactic annotation of ancient languages. Lessons from the Vedic Treebank. *Old World*, Vol. 1. DOI: 10.1163/26670755-01010003

Brereton, Joel P. & Jamison, Stephanie W. 2020. *The Rigveda: a guide*. Oxford University Press.

Casaretto, Antje. 2020. On Secondary predicates in Vedic Sanskrit. Syntax and Semantics. *International Journal of Diachronic Linguistics and Linguistic Reconstruction* 17: 1–63.

Casaretto, Antje, & Schneider, Carolin. Vedic local particles at the syntax-semantics interface. In Language change at the syntax-semantics interface, Chiara Gianollo, Agnes Jäger & Doris Penka (eds) 223–260. Berlin: De Gruyter Mouton.

Dik, Simon C. 1968. *Coordination. Its implications for the theory of general linguistics*, Amsterdam: North-Holland Publishing Company.

Dunkel, George. 1985. Verse-initial sentence boundary in the Rg̊-Veda: a preliminary overview. In *Grammatische Kategorien. Funktion und Geschichte*, Bernfried Schlerath (ed), 119–133. Wiesbaden: Ludwig Reichert Verlag.

Dunkel, George. 2014. *Lexikon der indogermanischen Partikeln*. Heidelberg: Carl Winter Universitätsverlag.

Eckhoff, Hanne Martine, Bech, Kristin, Bouma, Gerlof, Eide, Kristine, Haug, Dag T. T., Haugen, Odd Einar & Jøhndal, Marius. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1), 29–65.

Eckhoff, Anne Martine, Luraghi, Silvia & Passarotti, Marco C. 2018. Introduction: The added value of diachronic treebanks for historical linguistics research, *Diachronica* 35/2 2018.

Gaeta, Livio & Luraghi, Silvia. 2001. Gapping in classical Greek prose. *Studies in Language*, 25(1), 89–113.

Geldner, Karl Friedrich. 1951. *Der Rig-Veda: Aus dem Sanskrit ins Deutsche übersetzt und mit einem laufenden Kommentar versehen.*

Gunkel, Dieter & Ryan, Kevin. 2015. *Investigating Rigvedic word order in metrically neutral contexts*. Handout. Vienna.

Gunkel, Dieter & Ryan, Kevin. 2018. Phonological evidence for pāda cohesion in Rigvedic versification. In *Language and Meter*, 34–53. Leiden: Brill.

Hale, Mark. 2010. *Some Notes on the Syntax of iva Clauses in Vedic*. Paper presented at the 29th East Coast Indo-European Conference. Cornell.

Haspelmath, Martin & Buchholz, Oda. 1998. Equative and similative constructions in the languages of Europe. In *Adverbial Constructions in the Languages of Europe*, Van der Auwera, Johan (ed), 277–334. Berlin: Mouton de Gruyter.

Haspelmath, Martin & the Leipzig Equative Constructions Team 2017. Equative constructions in world-wide perspective. In *Similative and Equative Constructions: A Cross-linguistic Perspective*, Yvonne Treis & Martine Vanhove (eds), 9–32. Amsterdam, Philadelphia: John Benjamins.

Hellwig, Oliver, Scarlata, Salvatore, Ackermann, Elia & Widmer, Paul. 2020. The Treebank of Vedic Sanskrit. In *Proceedings of The 12th Language Resources and Evaluation Conference* (LREC 2020), Nicoletta Calzolari, Frederic Bechet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi et al. (eds), 5137–5146.

Hellwig, Oliver & Sellmer, Sven. This volume. The Vedic Treebank.

Hudson, Richard A. 1989. Gapping and grammatical relations. *Journal of Linguistics* 25, 57–94.

Jamison, Stephanie W. 1982. *Case disharmony in Rgvedic similes*. *IIJ* 24: 251–271.

Jamison, Stephanie W. 1998. Rigvedic *viśvátaç sīm*, or, Why syntax needs poetics. In *Mír Curad. Studies in honor of Calvert Watkins*, Karl Horst Schmidt (eds), 291–299. Innsbrück: Innsbrücker Beiträge zur Sprachwissenschaft.

Jamison, Stephanie W. & Brereton, Joel P. 2014. *The Rigveda: the Earliest Religious Poetry of India*. New York: Oxford University Press.

Kiparsky, Paul. 1976. Oral Poetry: Some Linguistic and Typological Considerations. In *Oral Literature and the Formula*, Benjamin A. Stolz & Richard S. Shannon III (eds), 73–106. Ann Arbor: Center for Coordination of Ancient and Modern Studies.

Mallinson, Graham & Blake, Barry J. 1981. *Language typology*. Amsterdam: North Holland.

Nivre, Joakim, De Marneffe, Marie-Catherine, Ginter, Filip, Goldberg, Yoav, Hajic, Jan, Manning, Christopher D, McDonald, Ryan, et al. 2016. Universal Dependencies V1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Johann-Mattis List, Michael Cysouw, Robert Forkel & Nicoletta Calzolari (eds), 1659–66.

Passarotti, Marco C. 2019. The Project of the Index Thomisticus Treebank. *Digital Classical Philology*, 10, 299–320.

Pinault, Georges. 1885. Négation et comparaison en védique. *Bulletin de la société de linguistique de Paris* 80, no. 1: 103–144.

Pinault, Georges. 1997. Distribution des particules comparatives dans la Rik-Samhitâ, *Bulletin d'Études Indiennes* 13-14, 307–367.

Treis, Yvonne. 2017. Comparative Constructions: An Introduction. In *Similative and equative constructions: A cross-linguistic perspective* (Vol. 117), Yvonne Treis & Martine Vanhove (eds.). John Benjamins Publishing Company.

Viti, Carlotta. 2002. Comparazione e individuazione: uno studio sugli equativi ṛgvedici *iva* e *ná*. *Archivio Glottologico Italiano*, no. 87.1: 47–87.

Viti, Carlotta. 2007. *Strategies of subordination in Vedic*. Vol. 57. Pavia: FrancoAngeli.

Viti, Carlotta. 2008. The meanings of coordination in the early Indo-European languages. *Revue de sémantique et pragmatique*, 24, 35–64.

Witzel, Michael. 1995. Ṛgvedic History: Poets, Chieftains and Polities. *In The Indo-Aryans of Ancient South Asia*, George Erdosy (ed.), 307–352. Berlin: de Gruyter.

Witzel, Michael, Gotō, Toshifumi, Dōyama, Eijirō & Ježić, Mislav. 2007. *Rig-Veda: das heilige Wissen; erster und zweiter Liederkreis*. Frankfurt am Main and Leipzig: Verlag der Weltreligionen.

Appendix

This Appendix contains all the queries employed for the case study presented in Section 4.3. All queries were written in Udapi query language (https://udapi.github.io).

- Query 1: N. of STAND – PAR and PAR – STAND orders in similes introduced by ná:

```
cat rv.conllu | udapy util.See node='node.deprel in ("advcl:manner",
"obl:manner", "obl:grad") and len([x for x in node.children if
x.lemma == "na"]) == 1'
```

- Query 2: N. of STAND - PAR and PAR - STAND orders in ná-similes with ellipsis

```
cat rv.conllu | udapy util.See node='node.deprel in ("obl:manner",
"obl:grad") and len([x for x in node.children if x.lemma == "na"])
== 1'
```

- Query 3: N. of STAND - PAR and PAR - STAND orders in ná-similes with gapping

```
cat rv.conllu | udapy util.See node='node.deprel == "advcl:manner
and len([x for x in node.children if x.lemma == "na"]) == 1'
```

- Query 4: N. of rightward gapping:

```
cat rv.conllu | udapy util.Eval node='if node.upos != "VERB" and
node.deprel == "conj" and node.parent.feats["VerbForm"] == "" and
len([x for x in node.children if x.deprel == "orphan"]) >= 1: count_
node.lemma +=1' end='pp(self.count)'
```

- Query 5: N. of leftward gapping:

```
cat rv.conllu | udapy util.Eval node='if node.upos != "VERB" and
node.deprel == "root" and len([x for x in node.children if x.deprel
== "orphan"]) >= 1 and len([x for x in node.children if x.deprel ==
"conj" and x.feats["VerbForm"] == ""]) >= 1: count_node.lemma +=1'
end='pp(self.count)'
```

Insights from Pāṇinian Grammar and Theory of Verbal Cognition for Representing Non-Linear Syntax

Developing Language-Neutral Syntactic Representation

Peter M. Scharf*

Formal and computational linguistics developed primarily in the environment of analytic European languages. To develop universally adequate linguistic theory demands investigating sophisticated linguistic theories, structures, and procedures developed to describe languages of a very different character from English. India developed an extraordinarily rich linguistic tradition over more than three millennia that could contribute useful insights to contemporary formal linguistics, and Indian linguistic theories could be formalized and implemented computationally. The Indian cognitive linguists of the seventeenth and eighteenth centuries described the cognition that arises from speech forms from whole sentences down to the level of morphemes. Their analysis reveals a dependency structure of semantic objects which may be projected onto the corresponding speech forms to provide an extremely precise and detailed analysis of verbal relations. By projecting the complex multi-dimensional relations in the realm of thought onto the relatively simple single dimension of speech, theory is able to more efficiently characterize syntactic relations in highly inflected languages with freer word order. An automation of the generation of speech forms in accordance with Pāṇinian rules can generate Sanskrit speech forms with internal dependency relations intact and with external dependency relations in the form of expectancies. This information would be highly useful in developing a Sanskrit parser and refining existing Sanskrit parsers.

Keywords: Sanskrit, Pāṇini, Aṣṭādhyāyī, linguistics, semantics, syntax, dependency relations

1. *Introduction*

Formal and computational linguistics was dominated by English at its inception and developed in subsequent decades primarily in the environment of Western European languages. More recently there has been a concerted effort to undertake formal linguistic analysis of a wide variety of languages, with particular interest in those with dramatically different features, and to enrich linguistic theory to account for linguistic variety. To develop universally adequate linguistic theory demands investigating sophisticated linguistic theories, structures, and procedures developed to describe languages of a very different character from English. India developed an extraordinarily rich linguistic tradition over more than three millennia that remains under-appreciated and under-investigated. The Indian sciences of grammar (Vyā-

* President, The Sanskrit Library.

karaṇa), logic (Nyāya), and ritual exegesis (Karmamīmāṁsā) have much to offer contemporary syntactic theory. The current paper touches upon ways in which Indian linguistics could contribute useful insights to contemporary formal linguistics, and Indian linguistic theories could be formalized and implemented computationally.

In particular, the Indian cognitive linguists of the seventeenth and eighteenth centuries described the cognition that arises from speech forms. In doing so they utilize the associations made between speech forms and the meanings they denote in the generative grammar of Pāṇini. Since these associations are made at the level of morphemes of roots, affixes, nominal bases, and inflectional terminations as well as at the level of words and phrases, the analyses of the structure of verbal cognition reveals a very precise dependency structure of the semantic objects. This dependency structure may be projected onto the corresponding speech forms to provide an extremely precise and detailed analysis of verbal relations. An automation of the generation of speech forms in accordance with Pāṇinian rules can generate Sanskrit speech forms with internal dependency relations intact and with external dependency relations in the form of expectancies. This information would be highly useful in developing a Sanskrit parser and refining existing Sanskrit parsers.

The Indian model of using semantic relations – in the realm of thought – as the source for relations among speech forms has the potential to revolutionize the science of syntax in contemporary linguistics which, under the influence of behaviorism in the preceding century, limited itself to the realm of speech perceptible by the senses and shunned any talk of meanings. The preoccupation with phrase structure analysis that endowed position in an underlying word order with special significance worked well with analytic languages but falls short with languages with freer word order. Analysis of the complex relations of thought as the basis for linguistic relations may restore the priority of thought over speech in linguistic analysis and generate more successful analyses.

2. *Speech and meaning*

Centuries ago, the French linguist Etienne Bonnot de Condillac (1746, 1775) recognized that while thought is complex combining many ideas at once, discourse represents ideas successively and that the order of their representation is not prescribed by nature. Hence he rejected any natural order of words. There is in fact nothing in phonetic forms themselves that indicates any ordering; the ordering of speech forms depends upon their significance, and the structure of complex units of speech depends upon the relations that hold among the concepts signified by their components. Linguistic relations and structures, therefore, do not reside in speech; they reside in cognitive structures in the domain of consciousness.

Following upon the growth in the prestige of the natural sciences with the adoption of empirical methods, the founders of modern linguistics came under pressure to

define linguistics as an empirical science. The result was, as Maat (2012: 434) notes, "a preference for data collection paired with an aversion to a priori speculations," and the embrace of "a relativist perspective". Maat (2012) continues, "In the twentieth century, both the data-oriented approach and the relativist perspective were carried to extremes, resulting in Bloomfieldian behaviorism and the Sapir-Whorf thesis, respectively." In addition, linguistics severed itself from the field of philology in which scholars acquired deep understanding of the thought of authors of various eras and locations by thorough study of their language and culture. Blevins (2012) notes that the efforts of Bloomfield's followers "to redefine linguistics as a science effectively cut the field off from the older philological tradition." Bloomfield (1933: 140) had argued, "The statement of meanings is therefore the weak point in language study," and (1933: 140) that "linguistic study must always start from the phonetic form and not from the meaning." Extending the point, the American structuralist Zellig Harris (1951: 5) began his work *Methods in structural linguistics* writing, "The main research of descriptive linguistics, and the only relation which will be accepted as relevant in the present survey, is the distribution or arrangement within the flow of speech of some parts or features relatively to others." Blevins (2012: 450) describes the tendency towards empiricism and aversion to concern with meaning writing, "The resulting model was purely distributional, concerned with the arrangement of observable units, without regard to any associated meaning or function." In the radical materialism adopted by the psychological school of behaviorism, thoughts and emotions, as unobservable private events, are not accepted as causes of observable behavior, though they may be recognized as epiphenomenal responses. Skinner (1957) advocated the study of behavior, including the use of language, in terms of conditioned responses by reference only to the functional relationships of organisms in their environments without reference to mental structures.

Chomsky revolted against the radical empiricist approach to the study of language as being far too simplistic and considered the seventeenth-century rationalist philosophers his intellectual ancestors (Maat 2012: 434). In his preface to the republication of his review of Skinner's *Verbal behavior*, he (1967) states that the general framework of behaviorism and of empiricism in general that dominated modern linguistics in the first half of the twentieth century "was largely mythology, and that its widespread acceptance was not the result of empirical support, persuasive reasoning, or the absence of a plausible alternative". Instead Chomsky considered that human beings have an innate faculty of language, universal across the species, which accounts for the portion of linguistic knowledge that is inexplicable by experience. While considering the language faculty to be an intrinsic part of human biology, Chomsky (1965: chapter 1) explores the interpretation of a generative grammar as a system of knowledge in the mind of the speaker, and describes language acquisition as a transition between mental states (Freidin 2012: 473). The innate component of

the computational system that constitutes language and is independent of experience, separate from the lexicon, he designated *universal grammar*. Chomsky's research program predominantly focused on determining rules and their sequences in formal grammar in order to account for linguistic data and in locating general rules and principles that would constitute elements of universal grammar. Chomsky (1957, 1955-1956, 1975) accounted for sentences by generating them from basic abstract linear structures by phrase structure and transformation rules. Phrase structure grammar analyzes sentences by progressively dividing phrases into their immediate constituents until the atomic lexical elements are reached. Clauses are initially divided in two into subject and predicate in accordance with traditional predicate logic; usually further analysis similarly descends in binary branching. Rules account for altered word orders as well as the hierarchical structure of components in the sentence.

Despite the fact that Chomsky freed linguistics from narrow constraints of empiricism and behaviorism and saw his approach as turning attention to internal mental considerations, two factors in his program kept linguistic science in the bondage of materialism: (1) his strict separation of syntax from semantics, and (2) his positing a linearly ordered initial state in sentence generation. Concerning the first point, although the motivation behind segregating syntax from semantics was to free considerations of the formal properties of syntax from empirical factors associated with conventions in particular languages at particular times and places, the result was also to separate syntax from more abstract considerations of how humans conceptualize their world, including fundamental conceptions that influence syntactic structure. Concerning the second point, while transformational grammar provides a means of generalizing an abstract structure over alternative constructions such as active, passive and nominalized representations of a verb in relation to agent and object, still one syntactic ordering is adopted as basic and the others accounted for by transformations of it. The basic structure associated with relational significance is thus still described as linear. Moreover, Blevins (2012: 453-454) notes that the overwhelming metalinguistic focus on theoretical concerns and formal devices that the Chomskyan paradigm inspired further divorced linguistics from the study of languages and the philological tradition. This rift deprived linguistics of useful insights that arise from the deep familiarity with other modes of thought that philology and the study of foreign languages engenders. Two developments in linguistics contribute to overcoming these limitations: dependency grammar and cognitive linguistics.

2.1. Dependency grammar

Just two years after the publication of Chomsky's (1957) *Syntactic structures* and the same year of his (1959) review of Skinner's (1957) *Verbal behaviour*, Lucien Tesnière's wife and friends brought out the posthumous publication of his (1959)

Éléments de syntaxe structurale. Tesniere (1959: chapter 6) distinguishes between structural order and linear order. Here Tesnière argues that hierarchical structure precedes linear order in the speaker's consciousness and that the act of speaking transforms this hierarchical order into linear sequence. Conversely, the act of hearing and understanding converts the linear sequence perceived in expressed language into structural order in the hearer's comprehension. The fundamental hierarchical order that Tesnière proposed recognizes the verb as the most important constituent of a clause, recognizes the agent, direct object, and indirect object as its arguments according to the particular valency properties of the verb, and recognizes additional adjuncts unrestricted by verb valency. He places the verb at the top node of a dependency tree and permits multiple descending unordered branches to represent the hierarchical structure. Such unordered hierarchical dependency structure more adequately captures the complex structure of thought expressed in speech than linear phrase structure does.

2.2. Cognitive linguistics

Beginning in the late 1980s, certain linguists reacted against the Chomskyan strict separation of syntax from semantics by bringing the language faculty itself within the realm of more general human cognition. One of the initial proponents of cognitive linguistics, Lakoff (1987: §2.2), states, "For cognitive linguistics meaning is the central issue." Siewierska (2012: 522) writes that cognitive linguists are centrally concerned with "how the mind deals with meaning, i.e. the human conceptual system and its reflection in language." Tsoneva-Mathewson (2009) describes the major assumptions of cognitive linguistics as being "that language is not an autonomous cognitive faculty but an integral part of human cognition and that linguistic knowledge of meaning and form is basically conceptual structure." Siewierska (2012: 518) writes that Langacker, another of the initial proponents of cognitive linguistics, maintains "that all linguistic knowledge (semantic, pragmatic, discourse-functional and crucially structural) is conceptual in nature, a part of semantic space." Langacker (1987: 76) describes this semantic space as "the multifaceted field of conceptual potential within which thought and conceptualization unfold." In contrast to separating syntax from semantics, construction grammar considers syntactic structures to be semantic configurations of conceptual content. Siewierska (2012: 519) writes that in construction grammar, "there are no actual syntactic relations, these being reconceptualized as semantic construals," and that it "has been devoted to providing a conceptual semantics for grammatical categories and relations and developing analyses of a wide range of conceptualization processes." Croft (2001: 236-237) interprets features of language typically interpreted as indicating syntactic relations such as case marking, agreement marking, and word order rather as indicating how

the given syntactic element fits into the semantic interpretation of a given construction and what it contributes to the identification of the construction. Finally, Janda (2015: 134) concisely states how cognitive linguistics brings grammar together with the lexicon into the realm of meaning, "Grammar is an abstract meaning structure that interacts with the more concrete meanings of lexicon."

Tsoneva (2009: §1.1.1) aptly summarizes the history of twentieth century linguistics and its attitude towards meaning as follows:

> The 'cognitive' revolution performed by Chomsky and his followers was a reaction against positivism and behaviorism in human sciences in general and Bloomfieldian linguistics in particular. Behaviorism in America in the period between 1930 and the end of the 1950s studied human behavior including language in terms of habits, stimuli and responses. During this time the study of meaning in language was largely neglected. This is because Bloomfield and his followers, among which was Chomsky's mentor Zeillig Harris, felt that meaning was inherently subjective, directly unobservable and thus beyond the scope of scientific investigation at least for the foreseeable future. In this context Chomsky's professed mentalist approach to linguistic analysis was thought to be the revolution intending to bring 'mind' back into the human sciences after a long cold winter of objectivism.
>
> ...
>
> Chomsky's professed mentalist approach, which was expected to involve meaning i.e. semantics, turned out to be formal systems approach, in which the principal assumption is that the rules of syntax are independent of semantics.

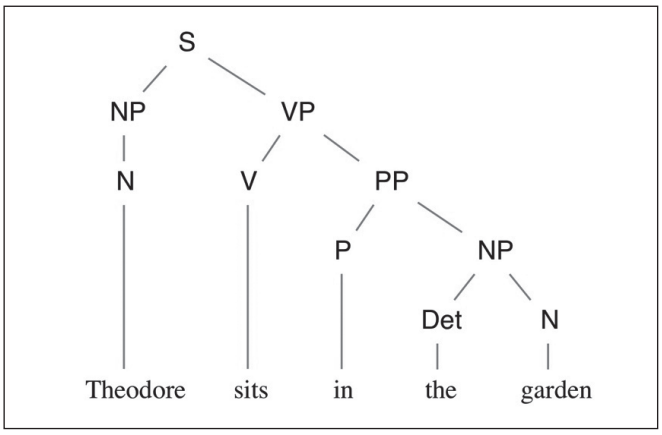
While much of cognitive linguistics is devoted to empirical research concerning the neurophysiology of perception and conceptualization and eschews introspection, its central concern, namely, to understand human linguistic conception, leaves room for contributions from a broad array of avenues. Since it is not possible now to conduct empirical research on ancient Indian native speakers of Sanskrit, the detailed analyses of the conceptual structure inherent in Sanskrit offered by the Indian linguists remains pertinent.

3. *Phrase-structure and dependency trees*

As mentioned above (Section 1, 2, and 2.1), phrase structure analysis usually represents the structure of sentences in binary trees, divides the subject from the predicate as initial immediate constituents and then proceeds to analyze these constituents by further binary analysis. For example, Figure 1 shows the analysis of a simple English sentence.

(1) *Theodore sits in the garden.*

Figure 1 The binary phrase-structure tree for (1)



The phrase-structure tree shows the sentence (S) as the top node, divides the subject noun phrase (NP) from the predicate verb phrase (VP) as initial immediate constituents, divides the verb phrase into a verb (V) and a prepositional phrase (PP), the prepositional phrase into a preposition (P) and the noun phrase (NP) it governs and finally this noun phrase into a determiner (Det) and a noun (N). While recognizing one constituent in each phrase as the head of that phrase, each phrase itself is recognized as a unitary constituent that combines with another constituent in a higher-level phrase. Besides reflecting the hierarchical structure, the ordering of branches reflects the order of words in the sentence.

In contrast, dependency trees fundamentally do not represent the sequence of words in the sentence and represent only the heads of phrases, rather than the phrases as units, as constituent nodes in the tree. Figure 2 shows the dependency tree for the same sentence (1) for which the phrase-structure tree is shown in Figure 1, and Figure 3 shows the dependency tree for the equivalent Sanskrit sentence.

Figure 2 The dependency tree for (1)

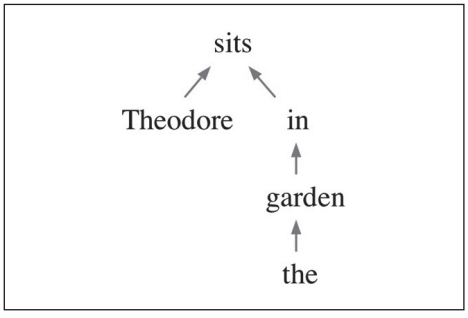
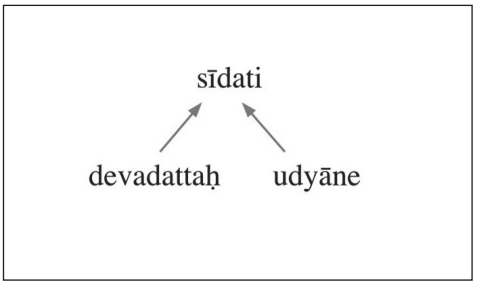


Figure 3 The dependency tree for the Sanskrit sentence equivalent to (1)



The dependency tree takes each lexical item, in this case each word, as a node and represents the dependency of subordinate nodes by a directional arrow relating a subordinate node to the node that governs it. In our trees the subordinate node points towards the node that governs it. Notice that in contrast to the phrase-structure tree, no additional notation is necessary to reveal the head of each phrase because the head is itself taken as a node. Nor is information lacking regarding the other constituents of the phrase a head governs: the subordinate elements in the same phrase are the nodes dependent upon the node representing the head.

The chain of nodes progressively descending from a node includes all the words in the phrase each head governs. In contrast to the phrase-structure tree which makes the subject and predicate equal immediate constituents of the clause, the dependency analysis inaugurated by Tesnière takes the verb as the principal node, and the agent as one of its arguments dependent upon it. While some representations of dependency trees slant arrows at angles to position nodes in the horizontal order in which the words in the sentence occur, such angling is not an essential factor in the dependency tree.

Notice that the Sanskrit sentence (2) equivalent to (1) utilizes a single inflected word *udyāne* instead of the English prepositional phrase *in the garden*.[1]

(2) *Devadatta udyāne sīdati.*

Devadattaḥ	*udyāne*	*sīdati.*
m1s	n7s	pre_a3s
Devadatta	in the garden	sits.

'Devadatta sits in the garden.'

The highly inflected nature of Sanskrit permits representation of six different participants in action in different cases as in (3).

(3) *Devadatto gṛhād yajñadattāya nagare śakaṭena kumbham ānayati.*

Devadattaḥ	*gṛhāt*	*yajñadattāya*	*nagare*	*śakaṭena*
m1s	n5s	m4s	n7s	m2s
Devadatta	from his house	for Yajñadatta	in the city	with a cart

kumbham	*ānayati.*
n3s	pre_a3s
a pot	brings.

Devadatta brings a pot with a cart from his house for Yajñadatta in the city.

1. The first line of examples shows the sentence in Roman transliteration, the second with sandhi analyzed, the third its inflectional identification, the fourth is word-by-word translation, and the fifth its English translation.

Figure 4 The dependency tree for the Sanskrit sentence (3) with six participants in the action

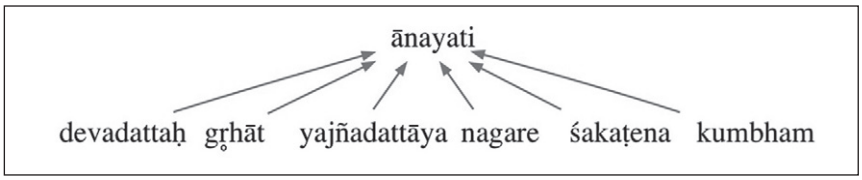


Figure 4 shows the dependency tree for (3). The fact that the different participants in action are represented by inflectional morphology rather than by syntactic position allows a great deal of freedom in word order in Sanskrit. The sentence (4) expresses virtually the same meaning as (3) with an altered word order.

(4) *Devadattaḥ kumbhaṁ śakaṭena gr̥hād yajñadattāya nagare ānayati.*

Devadattaḥ	*kumbham*	*śakaṭena*	*gr̥hāt*	*yajñadattāya*
m1s	m2s	n3s	n5s	m4s
Devadatta	a pot	with a cart	from his house	for Yajñadatta
nagare	*ānayati.*			
n7s	pre_a3s			
in the city	brings.			

'Devadatta brings a pot with a cart from his house for Yajñadatta in the city.'

The dependency tree for (4) shown in Figure 5 differs from the dependency tree for (3) shown in Figure 4 only in the insignificant horizontal placement of the dependent nodes; the hierarchical dependency structure is identical. Figure 6 shows the phrase-structure tree for (3). Notice that the binary phrase-structure tree requires an artificial proliferation of verb-phrase nodes (VP) uniting constituents in single verb phrases due to the ordering of words in (3) where the ordering of these constituents in (4) shows that they are not essentially related to each other, but rather are equally related to the verb as shown in the virtually identical dependency trees in Figure 4 and Figure 5.

Figure 5 The dependency tree for (4) the same Sanskrit sentence as (3) with the words for the six participants in the action reordered

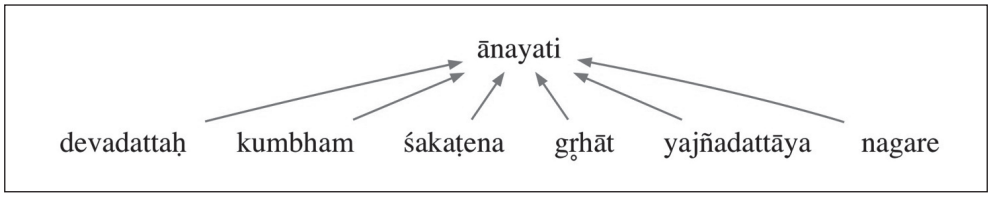
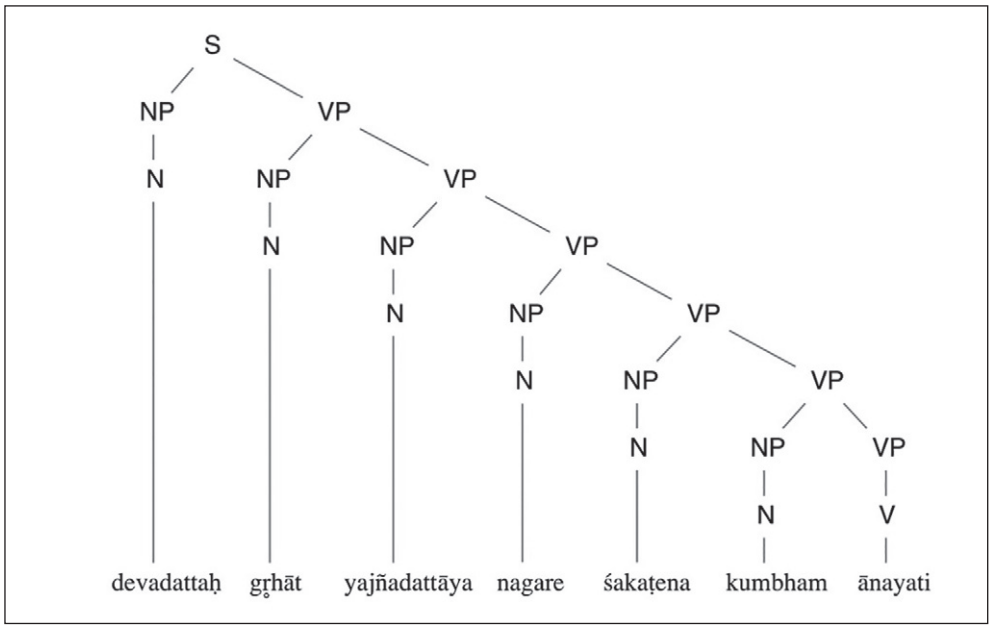


Figure 6 The phrase-structure tree for (3)



4. *Linear versus multidimensional syntactic representation*

Both the phrase structure trees and dependency trees shown in the previous section represent speech forms, either phrases or words, as elements in the hierarchical structure of language. As remarked in Section 2 however, nothing in the phonetics of the speech forms themselves, or their transcription, directly indicates any hierarchical structure. The structure is a feature of the conception of speakers of the language represented in the expressed speech forms. Thus the relations that hold between constituents in that structure are relations among concepts, that is, meanings, not among the speech forms that denote them. Formal linguists have recognized this as have cognitive linguists. As mentioned in Section 2, Chomsky considered generative grammar to be a system of knowledge in the mind of the speaker, and as mentioned in Section 2.1, Tesnière recognized that hierarchical structure precedes linear order in the speaker's consciousness. Despite the fact that these formal linguists recognized the mental nature of hierarchical linguistic structure, the fact appears lost in their formalisms. The trees show relations among speech forms, not among concepts. Particularly in phrase-structure trees, the sequence of these speech forms is paramount. In transformational grammar, position in an underlying sequence was taken to be significant, and investigators typically refer to subject position, object position, etc. as if the location in the sequence itself carried conceptual significance.

The motivation of linguists to attend to expressed speech rather than to mental concepts should be clear enough from the historical sketch of twentieth-century linguistics presented in Section 2: the effort to establish linguistics as a science in the train of empiricism forced them to attend to the aspect of language observable by the senses. Nevertheless, explicit discussion of the nature of thought and speech and of the philosophical presuppositions of attitudes towards the investigation of them will be instructive.

Speech is linear. The expression of sounds by the vocal organs occurs in time in a single dimension. A single dimension is appropriately represented in a line. Relations in a line are restricted to precedence and subsequence; hence sequence is a binary relationship. Therefore, speech is appropriately represented in binary trees.

In contrast, thought is multidimensional. Thought occurs in consciousness. Conceptualization in thought is not inherently limited by spacial and temporal dimensions. Humans conceptualize their experience in at least the four dimensions of space and time, yet poets and philosophers extend these dimensions without limit. Modern physicists conceive of a unified field of eleven dimensions, and mathematicians conceive of an infinite number. Hence thought is multidimensional and appropriately represented in n-space. Relations in n-space are complex. It is artificially constraining to represent complex relations in n-space in binary relationships. Therefore, thought is not appropriately restricted to representation in binary trees.

5. *Projection of thought onto speech*

The hierarchical structure of language is due to the hierarchical structure of conception in the domain of thought, not to any structure in the expressed speech itself. As cognitive scientists strive to infer the structure of human conception in general from empirical data, cognitive linguists strive to infer the structure of linguistic conception from expressed speech with the assumption that the linguistic conception shares structures with human conception in general. The inference of complex structure from simpler expression is not obvious and may not even be possible to achieve in its entirety. It is difficult enough to infer the three-dimensional structure of an object from its two-dimensional shadow on a wall; how much more difficult it is to infer multidimensional structure from its expression in a single dimension! The explicit description of concepts by philosophers and of linguistic concepts by philosophers of language has always guided such inquiry and cannot be taken for granted or assumed to be absent even by the most ardent empiricist. The most preeminent empiricist recognized that it is not possible to reconstruct a conceptual whole just from its discrete parts. In the *Appendix* to his *A Treatise of Human Nature*, David Hume wrote in recognition of just such a failure, "[A]ll my hopes vanish, when I come to explain the principles, that unite our successive perceptions in our thought or consciousness. ... I cannot discover any theory, which gives me satis-

faction on this head." Nor is it possible to reconstruct a higher-level programming language from a sequence of zeroes and ones without knowing the data structures, encodings and rules of the language. The Turing machine's deciphering of the German secret codes during World War II succeeded only with the help of constraints offered by the identification of certain sequences of code with particular people, places, and slogans known independently, and by assumptions about the context of messages being in regard to particular geographical and temporal coordinates. The U.S. Postal Service's automated analysis of addresses on envelopes likewise does not simply read an address as a linear string. It depends upon the constraints offered by prior knowledge about the structure of an address, which cities have which zip codes, which streets are in those cities and zip codes, and the address numbers located on those streets (Govindaraju 1997).

Although it is difficult to infer the structure of multidimensional linguistic conception from its single-dimension expression in speech, the inverse is easy. The projection of multidimensional structures onto fewer dimensions is deducible as for example three-dimensions are projected onto two in projective geometry, or the eleven dimensions of unified field theory onto the four experienced dimensions of space and time. The expression of thought onto speech may analogously be represented by the projection of multidimensional space onto a line. Complex relations may be converted to binary relations just as in digital machines higher level structures are represented ultimately by zeroes and ones. However, projection of a higher number of dimensions onto fewer and conversion of complexity to simplicity may not be lossless.

6. *Philosophical assumptions motivating empirical linguistics*

The description of syntactic roles as associated with position in a linear sequence and the representation of syntax in the binary trees of speech forms are motivated by naïve materialism. In naïve materialism, one assumes that bodies are multiple and discrete, that minds, if such exist at all, are epiphenomenal entities localized within discrete bodies, and that it is not possible for one person, located in one body, to know another person's thoughts directly. The empirical linguist therefore assumes that cognition of the meaning of speech is limited to the content of the speech. Laboratory methodology takes for granted that all relevant content is captured in verbal expression. Likewise, assuming that all relevant content is captured in linguistic transcription, computational linguists assume that they can discover how language works solely by statistical research on corpora of linguistic transcription.

These philosophical assumptions motivating empirical linguistics remain unsubstantiated. In particular, the assumption that language cognition depends only upon what is captured in linguistic expression and transcription is itself without any basis. The very fact of language change demonstrates that subsequent generations

of speakers fail to infer from the linguistic data with which they are presented the same phonological, grammatical, syntactic, and semantic structure that the preceding generation of speakers knew. Even language usage in its real-world context accompanied by explicit educational instruction is unsuccessful in transmitting the linguistic knowledge of one generation to the next.

Naïve materialism assumes materialist reductionism and that science will ultimately explain processes of consciousness and the functioning of living organisms in terms of the science of inanimate physical objects. Materialist reductionism assumes that consciousness is reducible to neurophysiological processes in the brain, that biological processes are explicable in terms of inorganic chemistry, that chemistry is explicable in terms of atomic physics, that atoms are explicable in terms of subatomic particles, and that subatomic particles are explicable in terms of quantum mechanics. However, such an assumption is unwarranted. For materialist reductionism is incompatible with the Copenhagen interpretation of quantum mechanics, the most successful theory of quantum mechanics upon which foundation the entire program of materialist reductionism depends. As D. C. Scharf (1989) has demonstrated, the reduction of macroscopic entities to microscopic objects described in the Schrodinger equation is incompatible with the Copenhagen interpretation of quantum mechanics because "the Copenhagen interpretation of quantum mechanics includes reference to a macroscopic entity (the measuring device). Because it refers to a macroscopic entity, it does not achieve a microscopic explanation of entities. Therefore, Quantum mechanics is incompatible with the reductionist program."

Contrary to the unfounded belief in naïve materialism and materialist reductionism that assumes that structured wholes are explicable in terms of their discrete parts, unified field theories point to the explanation of discrete entities in terms of fundamental non-discrete fields and to a single field underlying all of nature. A comprehensive description of nature suggests that the single unified field include consciousness (Hagelin n.d.). Even without reference to cutting-edge theories of quantum field theory, it is obvious that natural language context includes non-linguistic factors. Actual language understanding arises not from linguistic expression and transcription alone but is always accompanied by a variety of experience.

Considering the issues discussed in this section, a more realistic and honest enlightened linguistics would proceed to describe the structure of linguistic conception in terms of conception itself rather than in terms of speech and would explain language use in terms of conception rather than attempting to explain linguistic conception in terms of expressed speech. Because thought is more complex than speech, cognitive linguists should first examine cognitive structures. Cognitive relations should be expressed directly in terms of hierarchical conceptual structures, not

in terms of the speech forms that denote them. Then one should project the cognitive structure onto the speech forms that denote elements in that conceptual structure. In short, one should explain speech in terms of thought, not thought in terms of speech. This is the approach used by traditional Indian linguists.

7. *Consciousness-based linguistic description*

7.1. Introduction

Indian linguists in the discipline of Vyākaraṇa account for speech in terms of thought, and describe linguistic structure in terms of cognitive structures. Pāṇinian grammatical rules begin with semantic conditions. Under these semantic conditions as well as cooccurrence conditions, they introduce terms that categorize semantic conceptions. Under specific categorial conditions, subsequent rules introduce representative cover symbols and generic affixes which are then subjected to morphophonemic and phonetic changes to produce the actual Sanskrit speech forms. The speech forms are thus accounted for in terms of the conceptions which constitute the initial conditions by a series of rules. All traditional grammatical systems in India, non-Pāṇinian as well as Pāṇinian, use the same approach. Philosophers of language in the Pāṇinian tradition, preeminently Bhartṛhari, investigate in detail the structure of the semantic foundation of speech, the cognitive structures that produce linguistic structure and are expressed in syntactic and morphological structure in speech. For more than two thousand years, the disciplines of logic (Nyāya) and ritual exegesis (Mīmāṁsā) engaged in debate with the discipline of Vyākaraṇa and with each other over the cognitive structures inherent in language. The next section surveys the research on this debate while the following sections touch on a few of the major topics.

7.2. Survey of Indian cognitive linguistic literature

P. M. Scharf (2012: 261-64) briefly surveys the major contributions to and topics covered in the debate among Indian cognitive linguists. A number of secondary works analyze and explain the issues debated in the history of the Indian philosophy of language and the conclusions of the principal parties. Kunjunni Raja (1963) gives a clear presentation of the major points of view in Indian semantics, and Subharao (1969) described the theories of verbal cognition of the major Indian schools of thought. Bhattacharya (1962) is more textually oriented, and Biardeau (1964) is more interpretive. Sastri (1959) provides a general introduction to the topic as treated by Bhartṛhari, while Iyer (1969) provides an extensive summary of the thought presented by Bhartṛhari in his *Vākyapadīya*, and Houben (1995) translates an important chap-

ter, discusses principles for its interpretation, and provides access to recent work on this central figure of Indian philosophy of language. P. M. Scharf (1996) and Aussant (2009) enter into the details of argumentation concerning the semantics of common and proper nouns respectively.

In the seventeenth and eighteenth centuries, cognitive linguists such as Bhaṭṭoji Dīkṣita, Kauṇḍabhaṭṭa and Nāgeśa summarized the conclusions of the discipline of Vyākaraṇa concerning cognitive linguistic structure in terms of the structure of cognitions that arise from various units of speech and their morphological and syntactic constructions. Their works themselves have been the subject of detailed investigation. Several scholars have worked on sections of Kauṇḍabhaṭṭa's longer work, the *Vaiyākaraṇabhūṣaṇa*. Ramakrishnamacharyulu (2015, 2019) edited its first four chapters, the *Dhātvarthanirṇaya*, the *Lakārārthanirṇaya*, the *Nāmārthanirṇaya*, and the *Nāmārthapariccheda*; Gune (1978) translated and analyzed its second chapter, the *Lakārārthanirṇaya*, and Jha (1977, 1998) again translated the first two of these sections. Deshpande (1992) translated and analyzed the chapter concerning the meaning of nominals, the *Nāmārthanirṇaya*. Kauṇḍabhaṭṭa's abridgment of this work, the *Vaiyākaraṇabhūṣaṇasāra*, has likewise been the subject of study. Joshi (1960, 1967) translated the first two chapters, the *Dhātvarthanirṇaya* and the *Lakārārthanirṇaya*, and the last chapter the *Sphoṭanirṇaya*. Joshi (2015) wrote several articles on Kauṇḍabhaṭṭa's thought including on chapters 3 (*Subarthanirṇaya*), 5 (*Samāsaśaktinirṇaya*), and 7 (*Nañarthanirṇaya*), and Rathore (1988) wrote a study of the topics in the work. Das (1990) edited and translated the whole text, though contrary to the claim of the subtitle, the work does not constitute a critical edition: it does not describe any manuscript or provide a critical apparatus, and rarely provides a variant reading. Cardona (n.d.) is currently editing and translating Nāgesa's *Paramalaghumañjuṣā*.

7.3. Indian ontology and semantics

Early Indian ontological discussions mention three fundamental types of entities: 1. *dravya* 'substance', 2. *guṇa* 'quality', and 3. *kriyā* 'action'. However, the Sāṅkhya ontology considers the first to be simply a conglomeration of the second. Nyāya ontology recognizes in addition 4. *jāti* 'generic property', or *sāmānya* 'sameness'. The Vaiśeṣika ontology accepts in addition:

5. *viśeṣa* '(ultimate) particularity',
6. *samavāya* 'inherence', and
7. *abhāva* 'absence'.

Beginning with Pāṇini, the discipline of Vyākaraṇa includes concepts of the first four among the semantic conditions for speech forms. In his commentary *Mahābhāṣya*

on Pāṇini's grammar, Patañjali discusses these, the fourth of which he sometimes refers to as *ākṛti* 'form, class property', and four types of words whose usage is conditioned by concepts of these four types of entities respectively:

1. *yadṛcchāśabda* 'proper names'.
2. *guṇaśabda* 'words for qualities',
3. *kriyāśabda* 'action words', i.e. participles, action nouns, agent nouns, etc., and
4. *jātiśabda* 'common nouns'.

Later Pāṇinian grammarians use the term *pravṛtti-nimitta* 'condition for usage' for the concepts of entities that condition the use of words.

8. *Competing perspectives on linguistic cognitive structure*

The disciplines of grammar (Vyākaraṇa), logic (Nyāya), and ritual exegesis (Karmamīmāṁsā) held different views as to what constituted the principal element in the verbal cognition of a sentence. Nyāya held that the subject, the agent (*kartṛ*), is the principal element and that everything else qualifies him, including his action. In contrast, Karmamīmāṁsā and Vyākaraṇa held the action to be principal, and that everything else qualifies that. The different views in Nyāya and Karmamīmāṁsā are easily explicable from their philosophical presuppositions and principal concerns. Nyāya considered individual selves to be the agents of their activity, and the enjoyers of the results of that activity, and accepted a single divine being, God, as the creator of the world. The creator's activity of creating the world is subordinate to his being. Karmamīmāṁsā is principally concerned with analyzing Vedic injunctions such as (5) and in describing what has to be done in the ritual.

(5) *Svargakāmo yajeta.*

Svargakāmaḥ	*yajeta.*
m1s	pop_m3s
one desirous of heaven	should perform a sacrifice.

'One desirous of heaven should perform a sacrifice.'

Karmamīmāṁsā considers all the constituents in the ritual, including the agent, to be subordinate to the ritual action, and even considers the Vedic texts that enjoin ritual actions to be authorless (*apauruṣeya*).

Both Nyāya and Karmamīmāṁsā recognize the bipartite division of a sentence in its discourse structure. The former refers to a subject-predicate structure while the latter refers to a subject-command structure. Both term the subject *uddeśya*. Consistent with their concern with statements versus commands, Nyāya terms the pred-

icate *pratipādya* 'that which is to be made known', while Karmamīmāṁsā terms the command *vidheya* 'that which is enjoined'.

All parties accept the basic semantic analysis of nominals and verbs described by Yāska in his *Nirukta* (circa fifth century BCE), and the distinction he makes between action which has been effected (*siddha*), that is, action as an entity, and action which is to be effected (*sādhya*), that is, action as a process in progress. Finite verbs denote the latter while action nouns denote the former. Yāska wrote:

> *tatraitan nāmākhyātayor lakṣaṇaṁ pradiśanti. bhāvapradhānam ākhyātam. sattvapradhānāni nāmāni. tad yatrobhe bhāvapradhāne bhavataḥ pūrvāparībhūtaṁ bhāvamākhyātenācaṣṭe vrajati pacati iti. upakramaprabhṛtyapavargaparyantaṁ mūrtaṁ sattvabhūtaṁ sattvanāmabhir vrajyā paktir iti.* (*Nirukta* 1.1)
>
> 'They indicate the following definition of a nominal and a verb: a verb has action as principal; a nominal has an entity as principal. Where both have action as principal (i.e. in finite verbs and action nouns), a finite verb, such as *pacati* 'cooks' and *vrajati* 'walks', denotes sequential action. Action nouns denote the action from beginning to end solidified as an entity.'

8.1. The semantics of verbs

All three of the disciplines of Vyākaraṇa, Nyāya and Karmamīmāṁsā recognized that activity could be analyzed into two parts. The first part is the engagement of the agent itself in conduct; the second part is the result (*phala*) of such engagement in the change that takes place. For example, in the action of going (*gamana*) the result is the disjunction of the agent from one place and the conjunction of the agent with another (*pūrvadeśaviyogānyadeśasaṁyoga*). In the action of cooking, the result is the softening (*viklitti*) of the food. In Vyākaraṇa the engagement of the agent is generally given the neutral term *vyāpāra* 'activity', in Nyāya it is termed *kṛti* 'effort', and in Karmamīmāṁsā it is termed *bhāvanā* 'creative engagement'. While in Vyākaraṇa and Karmamīmāṁsā the activity or creative engagement is considered principal in the cognition of an active sentence, in Nyāya, the individual who is the agent of the action and thus the one in whom the effort is located is the principal element.

As mentioned in Section 1, the analysis of the significance of speech forms undertaken by the Indian linguists extends to morphemes of roots, affixes, nominal bases, and inflectional terminations. Thus they all accept that verbal inflectional terminations in a finite verb denote the number (*saṅkhyā*) while there is some disagreement about whether verbal terminations also denote time, and the agent in an active sentence or direct object in a passive sentence. Nyāya and Karmamīmāṁsā hold that the verbal termination also denotes effort (*kṛti*) or creative enterprise (*bhāvanā*).

Table 1 Comparison of verbal cognition in Nyāya, Mīmāṁsā and Vyākaraṇa

darśana view	*dhatu* root	*tiṅ-pratyaya* verbal termination
Nyāya	*phala* result	*kṛti + saṅkhyā + kāla* effort + number + time
Mīmāṁsā	*phala* result	*bhāvanā + saṅkhyā + kāla + kartṛtva* creative enterprise + number + time + agency
Vyākaraṇa	*phala + vyāpāra* result + activity	*saṅkhyā + kāla + kartṛ* number + time + agent

Conversely, Vyākaraṇa holds that the verbal root denotes activity (*vyāpāra*) as well as its result (*phala*) while Nyāya and Karmamīmāṁsā hold that the verbal root denotes only the result (*phala*). Table 1 summarizes the comparison of the cognition of active finite verb forms according to the three disciplines. Departing from the view preferred by the seventeenth century grammarians Bhaṭṭoji Dīkṣita and Kauṇḍabhaṭṭa concerning the finite verb's denotation of time, the grammarian Nāgeśa (circa 1700 CE) considered time, as an essential component of the activity of the agent, to belong properly to the cognition of the root rather that the verbal termination which merely cosignifies it (P. M. Scharf 2021).

8.2. The semantics of nominals

In his *Mahābhāṣya* on *A.* 1.2.64, Patañjali discusses two extreme views concerning the meaning of common nouns: Vyāḍi held that a common noun denotes an individual substance while Vājapyāyana held that it denotes a class property. As P. M. Scharf (1996: 89-91, §1.6) explains, however, he concludes that it denotes both with one or the other predominant in the cognition. He writes:

> It is not the case that an individual substance is not denoted for him for whom a class property is denoted, nor that a class property is not denoted for him for whom an individual substance is denoted. Both are denoted for both. But for each something is principal, something subordinate. For him for whom a class property is the object denoted, the class property is principal and an individual substance is subordinate. For him for whom an individual substance is the object denoted, an individual substance is principal and its class property is subordinate. (Patañjali, *A.* 1.2.64, *vārttika* 53c; P. M. Scharf (1996: 89)).

The generic property clearly predominates in the meaning in sentences that attribute properties or actions to all of a kind such as 1-2 below, whereas the individu-

al clearly predominates in the meaning in sentences that attribute properties or actions to some but not to others such as 3-4 below.

1. The cow is a sacred animal.
2. Cows are sacred animals.
3. The cow followed his owner to the University this morning.
4. The cows are resting in the shade.

Mīmāṁsā held that a common noun was the means of knowledge of the generic property alone and that the listener arrived at knowledge of the individual substrate of the generic property by a separate means of knowledge, presumption (*ākṣepa*, *arthāpatti*). The presumption in comprehending 3, for instance, would be that because the action of following is impossible for a generic property, an individual which is the substrate of it must be intended. In contrast, Nyāya accepted that the common noun made known the individual qualified by the generic property by the means of knowledge of verbal testimony directly.

As in the case of verbs, the Indian linguists extend the analysis of the significance of nominals to morphemes. The grammarian Kauṇḍabhaṭṭa considers various views concerning the significance of nominal bases and inflectional terminations. The stem of a common noun denotes up to five things inclusively in the following order:

1. class property (*jāti*),
2. individual (*vyakti*),
3. gender (*liṅga*),
4. number (*saṅkhyā*),
5. participant in the action denoted by the verb (*kāraka*).

Most Indian cognitive linguists accept that the nominal base denotes no more than the first three and that the nominal termination denotes number and either the participant in action or the relation of its participation.

8.3 Conditions for kārakas and nominal terminations

8.3.1. Conditions for kārakas and nominal terminations according to Pāṇini

Pāṇini accounted for the semantics and syntax of the relation of various participants in the action denoted by the verb and their denotation by nominal terminations in two stages. In the first, he categorizes general semantic conditions, and exceptional semantic conditions under certain cooccurrence conditions, by introducing a kāraka term for the participant under those conditions. The kāraka classification conditions the introduction of derivational affixes termed *kṛt* and *taddhita*, compound-

Table 2 The general meaning of nominal terminations (*vibhaktis*) according to Pāṇini

vibhakti	*kāraka*	meaning
1. *prathamā*		
2. *dvitīyā*	*karman* 'object'	*kartur īpsitatamam* 'most desired by the agent'
3. *ṭṛtīyā*	*kartṛ* 'agent'	*svatantra* 'independent'
	karaṇa 'instrument'	*sādhakatama* 'most efficacious'
4. *caturthī*	*sampradāna* 'indirect object'	*karmaṇā yam abhipraiti* 'whom he intends with the direct object'
5. *pañcamī*	*apādāna* 'source'	*dhruvam apāye* 'fixed in relation to departure'
6. *ṣaṣṭhī*	*śeṣa* 'remainder'	*sambandha* 'relation'
7. *saptamī*	*adhikaraṇa* 'locus'	*ādhāra* 'substrate'

ing, and verbal inflectional terminations. After introducing verbal terminations for items given the principal kāraka terms of agent (*kartṛ*) or object (*karman*), he then introduces nominal terminations for the participants not yet denoted. Additional rules account for the use of nominal terminations that do not involve participation in the action. Nominal terminations are provided in triplets of singular, dual, and plural terminations termed *vibhakti*. Table 2 shows the general purely semantic conditions for the various kāraka terms and the triplets of nominal terminations (*vibhakti*) conditioned by them.

According to Pāṇini, the first triplet (*prathamā vibhakti*) denotes only the particular meaning of the base, gender and number and does not denote any participant in the action. The sixth triplet generally denotes a relation other than participation in the principal action of the sentence, though it frequently denotes the agent or object of an action denoted by the verbal root in a nominal derivate ending in a kṛt affix and exceptionally denotes other kārakas in relation to specific verbs.

Pāṇini includes a number of additional specific semantic and cooccurrence conditions for the introduction of kāraka terms. These are shown along with the general semantic conditions in Table 3. The kāraka conditions under which Pāṇini then introduces triplets of nominal terminations are shown in Table 4. He also introduces nominal terminations directly under additional semantic and cooccurrence conditions such as after nominal bases denoting time or distance, after a nominal base occurring in connection with a preposition (*karmapravacanīya*) or direction word, or to indicate a cause. These conditions are not discussed in detail here.

Table 3 Specific kāraka conditions according to Pāṇini

kāraka	condition	meaning
kartṛ 'agent'		*svatantra* 'independent'
	causative	*prayojaka* 'instigator'
karman 'object'		*kartur īpsitatamam* 'most desired by the agent'
		tathāyuktaṁ anīpsitam 'undesired by the agent'
	double acc.	*akathitam* 'unstated'
	causative of going, knowing, eating, reciting, intransitive	*aṇi-kartṛ* 'the agent of the base root'
	causative of *hṛ*/*kṛ*	*aṇi-kartṛ* 'the agent of the base root'
	preverb + *krudh*/*druh*	*yam prati kopaḥ* 'the object of anger'
	div	*sādhakatama* 'most efficacious'
	adhi + *śī*/*ās*/*sthā*	*ādhāra* 'substrate'
	abhi-ni + *viś*	*ādhāra* 'substrate'
	upa/*anu*/*adhi*/*ā* + *vas*	*ādhāra* 'substrate'
karaṇa 'instrument'		*sādhakatama* 'most efficacious'
sampradāna 'indirect object'		*karmaṇā yam abhipraiti* 'whom he intends with the direct object'
	pleasing	*prīyamāna* 'the one being pleased'
	praise, ...	*jñīpsyamāna* 'whom one wants to make known'
	debt	*uttamarṇa* 'the creditor'
	desire	*īpsita* 'what is desired'
	anger	*yam prati kopaḥ* 'the object of anger'
	request	*yasya vipraśnaḥ* 'the one whose fate is examined'
	assenting	*pūrvasya kartā* 'the author of the proposition'
	hiring	*sādhakatama* 'most efficacious'
apādāna 'source'		*dhruvam apāye* 'fixed in relation to departure'
	fear	*bhaya-hetu* 'the cause of fear'
	defeat	*asoḍha* 'what one is unable to overcome'
	blocking	*īpsita* 'the desired object'
	hiding	*yenādarśanam icchati* 'by whom one does not want to be seen'
	instruction	*ākhyātṛ* 'the teacher'
	production	*prakṛti* 'the original material'
	arising	*prabhava* 'the source'
adhikaraṇa 'locus'		*ādhāra* 'substrate'

Table 4 Special as well as general *kāraka* conditions for *vibhaktis* according to Pāṇini

vibhakti	kāraka	condition
1. prathamā		
2. dvitīyā	karman	
3. tṛtīyā	kartṛ	
	karaṇa	
	karman	*hu* Vedic
	karman	*sam* + *jñā*
4. caturthī	sampradāna	
	karman	locomotion
	karman	elided infinitive or agent noun of purpose
	karman	inanimate object of *man-ya* in metaphorical insult
5. pañcamī	apādāna	
	apādāna	*stoka*/*alpa*/*kṛcchra*/*katipaya* denoting a property rather than a substance
6. ṣaṣṭhī	kartṛ	with a *kṛt*-derivate (except a participle, indeclinable, desiderative adjective, adj. ending in the affix *uka*, affixes in the meaning of the affix *khal*, a habitual agent noun ending in the affix *tṛn*, debtor or future agent nouns ending in *aka* or *in*)
	kartṛ	with a participle ending in the affix *kta* used in present time, or to denote a locus
	karman	with a *kṛt*-derivate
	karman	'miss'/*day*/*īś*
	karman	*kṛ* 'prepare'
	karman	'afflict' (except *jvar*)
	karman	*nāth* 'wish for'
	karman	*jas* caus./*ni-pra* + *han*/*naṭ* caus./*krāth* caus./*piṣ* 'injure'
	karman	*vi-ava* + *hṛ*/*paṇ* 'play, buy, sell'
	karman	*div* 'play'
	karman	*pra* + *iṣ*/*bru* ipv a2s, offering an oblation to a deity
	karaṇa	*jñā* 'make use of'
	karaṇa	*yaj* Vedic
	adhikaraṇa	time within which n-times
7. saptamī	adhikaraṇa	

8.3.2. Conditions for kārakas and nominal terminations according to Kauṇḍabhaṭṭa

While Pāṇini classifies semantic objects under kāraka terms by a series of rules that introduce the kāraka terms in a variety of special senses and cooccurrence environments in addition to a general semantic condition, Kauṇḍabhaṭṭa defines each kāraka. He frames an abstract statement that universally characterizes each. The following are his definitions for the first three kārakas:

1. kartṛ: *dhātv-artha-vyāpārāśraya*.
   agent: the substrate of the activity (*vyāpāra*) denoted by the root.
2. karman: *kriyā-janya-phalāśraya*
   object: the substrate of the result (*phala*) generated by the action.
3. karaṇa: *avyavadhānena phala-janaka-vyāpārāśrayaḥ*
   instrument: the substrate of the activity that generates the result immediately.

While framing general definitions of each kāraka, Kauṇḍabhaṭṭa does not neglect describing the variety captured by Pāṇini's rules. On the contrary drawing upon the long history of the discussion of the various types of each kāraka and examples of each type, he summarizes these discussions. Hence he describes three types of agent (*kartṛ*) and seven types of object (*karman*) as sumarized with examples in Table 5 and Table 6 respectively.

Table 5 Three types of *kartṛ*

kartṛ	meaning	example
śuddha	simple	*mayā hariḥ sevyate* 'Hari is served by me.'
prayojaka	motivator	*kāryate hariṇā* 'It is made to be done by hari.'
		gamayati kṛṣṇaṁ gokulam 'He causes Kṛṣṇa to go to school.'
karma-kartṛ	object-agent	*taṇḍulaḥ pacyate svayam eva* 'Rice cooks by itself.'

Table 6 Seven types of *karman*

karman	meaning	example
nirvartya	created	*ghaṭaṁ karoti.* 'He makes a pot.'
vikārya	transformed	*suvarṇaṁ kuṇḍalaṁ karoti.* 'He makes gold a bracelet.
prāpya	reached	*ghaṭaṁ paśyati.* 'He sees a pot.'
udāsīna	neutral	*tṛṇaṁ spṛśati.* 'He touches grass' (accidentally while traveling).
dveṣya	averse	*viṣaṁ bhuṅkte.* 'He eats poison'
anākhyāta	double acc.	*gāṁ dogdhi.* 'He milks the cow.'
anyapūrvaka	co-occurrence	*krūram abhikrudhyati.* 'He is angry at the cruel man.'
		krūrāya krudhyati. 'He is angry at the cruel man.'

9. Pāṇinian dependency trees

The previous sections have described the detailed analysis that Indian cognitive linguists provide regarding the cognitive structure of linguistic meaning, its organization into kāraka categories and the introduction of morphemes under semantic and cooccurrence conditions to represent the semantic objects so categorized. The hierarchical structure described belongs to the linguistic cognition, i.e. to the sphere of meaning, rather than to the speech forms. Yet because meaning is expressed in speech due to the intimate connection between words and their meanings, speech forms are taken to represent a structure which in fact belongs to the sphere of meaning. Because words are readily recognized units of speech, words have generally been taken as the units to represent the hierarchical structure of meaning as expressed in language in dependency trees.

9.1. Word dependency trees

Sections 2.1 and 3 introduced dependency grammar and dependency trees. As described in 2.1, Tesnière (1959) considered the verb as the principal constituent of a clause, and other elements as its arguments and adjuncts. He represented the hierarchical structure of the clause by placing the verb at the top node of a dependency tree and other elements, including the agent, in multiple descending unordered branches beneath it. Section 3 gave examples of dependency trees that take each word as a node. The dependency tree for (3) shown in Figure 4 and the dependency tree for (3) shown in Figure 5 show the verb as the top node and words for all six participants in the action of the verb recognized by the Indian grammarians (*kārakas*) as its dependent nodes. Because the Indian linguists describe various relations by which a semantic object is subordinate to another, for example, of how participants in action (*kārakas*) are subordinate to action (*kriyā*), information is available to label the edges that show these subordinating relations. By projecting that cognitive information supplied by the Indian linguists onto the speech forms, one can label the edges between words to indicate the specific subordinating relation by which each node is subordinate to another. Because of the prevalent convention of using words as nodes in dependency analysis of various languages, the same procedure was adopted for Sanskrit. Kulkarni, Pokar, and Shukl (2010) initiated the use of dependency trees with labeled edges to show the dependency structure of words in Sanskrit sentences, and Kulkarni and Ramakrishnamacharyulu (2013) worked out a useful set of relations discernible by automated parsing that could be used to label edges.

Figure 7 shows a typical dependency tree for a simple active Sanskrit sentence with words as the nodes and with labeled edges. The paraphrase of the cognition clearly indicates that Devadatta 'Theodore' is the agent and the rice (*odana*) the object of the activity of cooking (*pāka*) which is principal. Projecting this cognitive analysis onto whole words yields the word-level syntactic dependency structure shown in Figure 8.

(6) *devadatta odanaṁ pacati.*

Devadattaḥ	*odanam*	*pacati.*
m1s	m2s	pre_a3s
Theodore	rice	cooks

'Theodore cooks rice.'

(7) *devadatta-kartṛ-kaḥ odana-karma-kaḥ pāka-anukūla-vyāpāraḥ*
devadatta-kartṛ-kaḥ odana-karma-kaḥ pāka-anukūla-vyāpāraḥ

'Activity favorable to cooking that has Theodore as its agent and rice as its object.'

Sentence (8) is the passive corresponding to (6).

(8) *Devadatteṇa odanaḥ pacyate.*

Devadatteṇa	*odanaḥ*	*pacyate.*
m3s	m1s	pre_p3s
By Theodore	rice	is cooked

'Rice is cooked by Theodore.'

The grammarian Kauṇḍabhaṭṭa would paraphrase the analysis of the verbal cognition that arises from this passive sentence exactly as shown in (7) for the active. The dependency tree of the corresponding words is shown in Figure 8.

Dependency trees with words as nodes such as shown in Figure 7 and Figure 8 do not accurately capture the detailed analysis that Indian cognitive linguists provide regarding the cognitive structure of linguistic meaning. The very fact that both the active and passive share the identical cognitive paraphrase in (7) shows that the different representations by the tree for the active sentence shown in Figure 7 and the tree for the passive sentence shown in Figure 8 are not accurate. Instead this simple

Figure 7 The word-unit dependency tree for the active sentence (6)

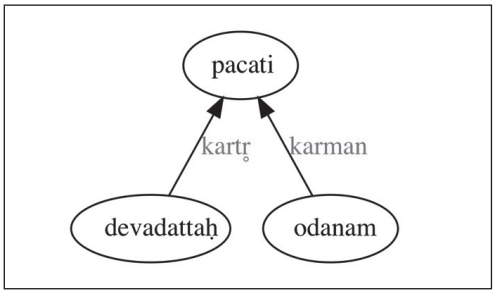


Figure 8 The word-unit dependency tree for the passive sentence (8)

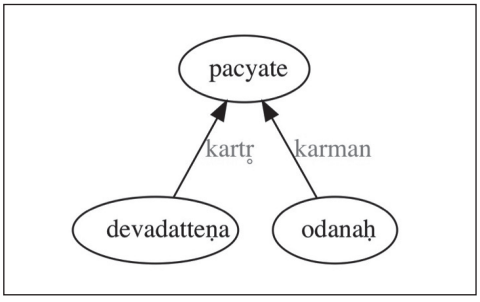
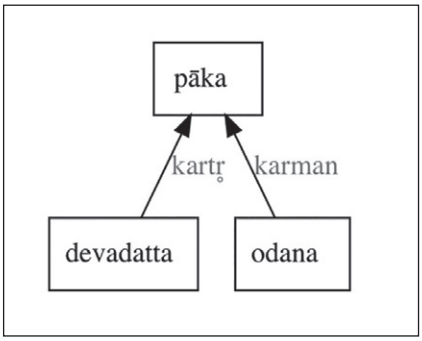


Figure 9 The cognitive dependency tree for the cognitive paraphrase in (7) for the active and passive sentences devadatta odanaṁ *pacati.* and (8)



paraphrase would be properly represented in the same dependency tree for both using terms for the cognitive concepts themselves. Figure 9 shows such a tree using boxes instead of ovals to represent concepts directly.

Yet the word trees do not even fully capture the information provided in the paraphrase of the cognition in (7), which itself is just a brief summary. In (7) a separation is made between the activity of the agent and the resulting change that takes place in the object. The dependency trees in Figure 7 and Figure 8 take the verb as a single unit to represent both. In addition, the nominative *devadattaḥ* is taken to represent the agent in Figure 7 when instead the Pāṇinian analysis recognizes that in the active the verbal termination *ti* denotes the agent, and in Figure 8, the nominative *odanaḥ* is taken to represent the object, when instead the Pāṇinian analysis recognizes that in the passive the verbal termination *te* denotes the object. In both the active and passive, the nominative termination indicates only that the meaning it denotes has the relation of non-difference (*abheda*) with the meaning denoted by the verbal termination. Moreover, the brief paraphrase provided in (7) does not mention the particular character of the result of cooking, does not indicate the particular relation the agent has with the activity and the object has with the result, does not mention the number or gender of the agent or object, and does not mention the generic property of the latter, all of which detail is clearly described by the Indian cognitive linguists. Pāṇinian grammar also specifically describes the sense of the verbal stem-forming affixes, *śap* in the active, and *yak* in the passive.

9.2. Cognitive dependency trees

As has been emphasized, the hierarchical structure of language belongs to the sphere of meaning rather than to the speech forms. Sections 8.1 and 8.2 described the detailed analysis that Indian cognitive linguists provide regarding the cognitive structure in the verbal cognition of verbs and nominals. Section 8.3 described Pāṇini's organization of semantic concepts into kāraka categories and the introduction of morphemes under semantic and cooccurrence conditions to represent the semantic objects so categorized, and also described Kauṇḍabhaṭṭa's definitions of kārakas.

Because the hierarchical structure described belongs to the cognition rather than to the speech forms, to properly represent this structure a dependency tree

should include the meanings as nodes rather than the words. It is only because the Indian linguists have provided precise rules for the introduction of particular morphemes under specific semantic conditions that the cognitive hierarchy is mappable onto speech forms at all; however, as has been shown in section Section 9.1, it is not possible to map the cognitive concepts accurately onto whole words. Because the Indian linguists introduce morphemes under specific conditions, information is available to show how these morphemes denote the cognitive concepts; however, the mapping of these morphemes onto the cognitive concepts is not trivial. Hence, in order to represent the information provided by Indian cognitive linguists accurately in dependency trees it is necessary to represent the cognitive concepts with their specific technical terms directly rather than to use the words that denote these concepts in ordinary usage or the morphemes provided by Pāṇinian grammar.

Indian cognitive linguists generally do not compose a paraphrase of verbal cognition in a way that provides all the detail of the cognitive structure; rather they focus on what is relevant in the context of the discussion. Yet it is possible to construct a complete paraphrase in accordance with what they explain about conceptual structure. Such a detailed paraphrase of the cognition of both the active and passive sentences (6) and (8) in accordance with the analysis of the grammarian Kauṇḍabhaṭṭa is shown in (9).

(9) *odanatva-jāti-puṁliṅga-ekatva-saṅkhyā-viśiṣṭa-odana-abhinna-karma-āśrita-viklitti-anukūlaḥ puṁliṅga-ekatva-saṅkhyā-viśiṣṭa-devadatta-abhinna-kartṛ-āśritaḥ vartamāna-kālakaḥ vyāpāraḥ*

'The activity of cooking in present time that resides in the agent which is no different from Theodore who is qualified by masculine gender and singular number, which activity is favorable to softening that resides in the direct object which is no different from rice which is qualified by the generic property riceness, masculine gender and singular number.'

Figure 10 shows an accurate dependency tree of the cognitive concepts so paraphrased. In this cognitive dependency tree, the activity (*vyāpāra*) of cooking is principal in the cognition. This activity would involve such actions as putting firewood on the fire, blowing on the fire, pouring water and rice grains into a pot, putting the pot on the fire, etc. Its result (*phala*) in the softening (*viklitti*) of the rice is shown in a node subordinate to the principal activity by the relation of being its result (*phalatā*). The occurrence of the activity in present time (*vartamāna kāla*) is likewise shown in a node subordinate to the principal activity by the relation of being the temporal substrate of that activity (*kālatā*). The agent (*kartṛ*) which is the substrate (*āśraya*) of the activity (*vyāpāra*), and the object (*karman*) which is the substrate (*āśraya*) of the result (*phala*) are shown in nodes subordinate to the principal activity and the

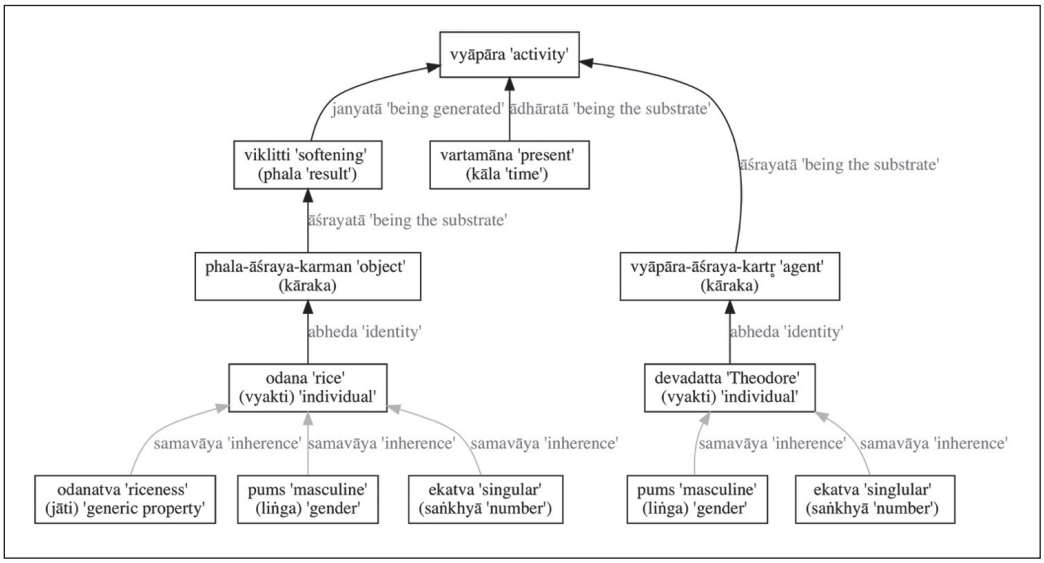
result respectively by the relation of being a substrate (*āśrayatā*). These characterizations of the entities of rice and Theodore are indicated by the relation of identity (*abheda*) to the individual (*vyakti*) rice (*odana*) and person of Theodore. Finally, in the bottom rank of the dependency tree the generic property (*jāti*), gender (*liṅga*) and number (*saṅkhyā*) inherent in these individual objects are shown: the generic property riceness (*odanatva*), masculine gender (*pums*), and singular number (*ekatva*) in the rice, and masculine gender (*pums*), and singular number (*ekatva*) in Theodore (*devadatta*).

Although grammarians generally paraphrase relations as non-directional, the logicians offer paraphrases of relations that reside in one relatum and are directed towards the other. Hence, while a grammarian would describe the relation between the activity of the agent and the result generated in the object by the phrase *janya-janaka-sambandha* 'the relation between the generator and the generated', a logician would describe the relation of the activity towards its result by the phrase *viklitti-niṣṭha-janyatā-nirūpita-vyāpāra-niṣṭha-janakatā* 'the property of being a generator located in the activity realized in relation to the property of being generated located in the result'. Alternatively, the logician would describe the relation of the result towards the activity by the phrase *vyāpāra-niṣṭha-janakatā-nirūpita-viklitti-niṣṭha-janyatā* 'the property of being generated located in the result realized in relation to the property of being a generator located in the activity'. Because we wish to show the direction of dependency in our diagrams, we use the latter logician's phrase to describe the relation of the dependent node to the node on which it depends.

Figure 11 shows the mapping of morphemes onto the cognitive concepts in the cognitive dependency tree shown in Figure 10. The morphemes are shown in ovals, the cognitive concepts in boxes, and the signification of concepts by speech forms by dotted arrows. Denotation is shown by a black dotted arrow while cosignification is shown by a grey dotted arrow. The root *pac* denotes both the principal activity (*vyāpāra*) of cooking as well as its result (*phala*) in the softening (*viklitti*) of the rice. The verbal termination (*ti*) denotes the agent (*kartṛ*), its singular (*ekatva*) number (*saṅkhyā*) and, in Kauṇḍabhaṭṭa's view, the time (*kāla*) of the principal activity (*vyāpāra*). The agent is cosignified, as indicated by the grey arrow, by the stem-forming affix (*a*). The nominal base (*devadatta*) denotes the individual so named and his masculine gender. The nominative singular nominative termination (*s*) denotes singular number but does not indicate any kāraka since the agent is already denoted by the verbal termination (*ti*). The accusative nominal termination (*am*), however, denotes the object (*karman*) as well as the singular number of the rice denoted by its base (*odana*). This common noun denotes the generic property inherent in the rice as well as its gender and their substrate.

According to Kauṇḍabhaṭṭa, the cognition that results from the passive sentence (8) is identical to the cognition that results from the active sentence (6) as shown in Figure 10. Figure 12 shows the mapping of morphemes onto the cognitive concepts.

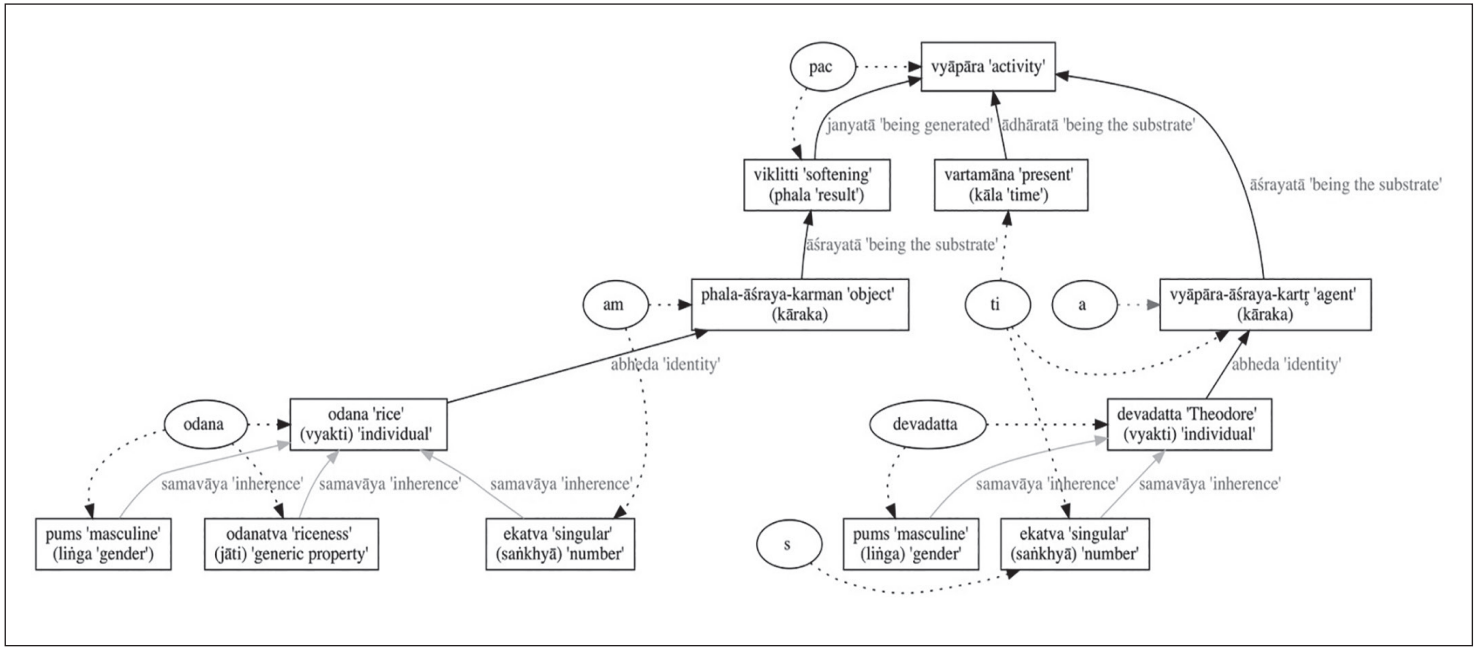
Figure 10 The cognitive dependency tree for the detailed cognitive paraphrase in (9) for the active and passive sentences (6) and (8) according to Kauṇḍabhaṭṭa



As in the active, the root *pac* denotes both the principal activity (*vyāpāra*) of cooking as well as its result (*phala*) in the softening (*viklitti*) of the rice. In the passive, the verbal termination (*te*) denotes the object (*karman*) and its singular (*ekatva*) number (*saṅkhyā*) rather than the agent and its, but, according to Kauṇḍabhaṭṭa, as in the active, also denotes the time of the principal activity (*vyāpāra*). The object is cosignified, as indicated by the grey arrow, by the passive stem-forming affix (*ya*). The nominal base (*devadatta*) denotes the individual so named and his masculine gender. The instrumental singular nominative termination (*ā*) denotes singular number and the agent (*kartṛ*) because it has not been denoted by the verbal termination. The nominative nominal termination (*s*) on the nominal base (*odana*), however, does not indicate any kāraka since the object has already been denoted by the verbal termination (*te*). The nominative nominal termination also denotes the singular number of the rice denoted by its base. As in the active sentence, the base of this common noun denotes the generic property inherent in the rice as well as its gender and their substrate.

There are minor differences among Pāṇinian grammarians concerning verbal cognition. As discussed by P. M. Scharf (2021), Nāgeśa considers time, as an essential feature of the principal activity, to be denoted by the root rather than by the verbal termination. Nāgeśa also considers that the cognition that arises from a passive sentence differs from the cognition that arises from an active sentence in that in the cognition of the passive the result (*phala*) is the principal element and the activity (*vyāpāra*) that brings it about is subordinate to it.

Figure 11 The mapping of morphemes in the active sentence (6) onto the cognitive dependency tree for the detailed cognitive paraphrase in (9) according to Kauṇḍabhaṭṭa



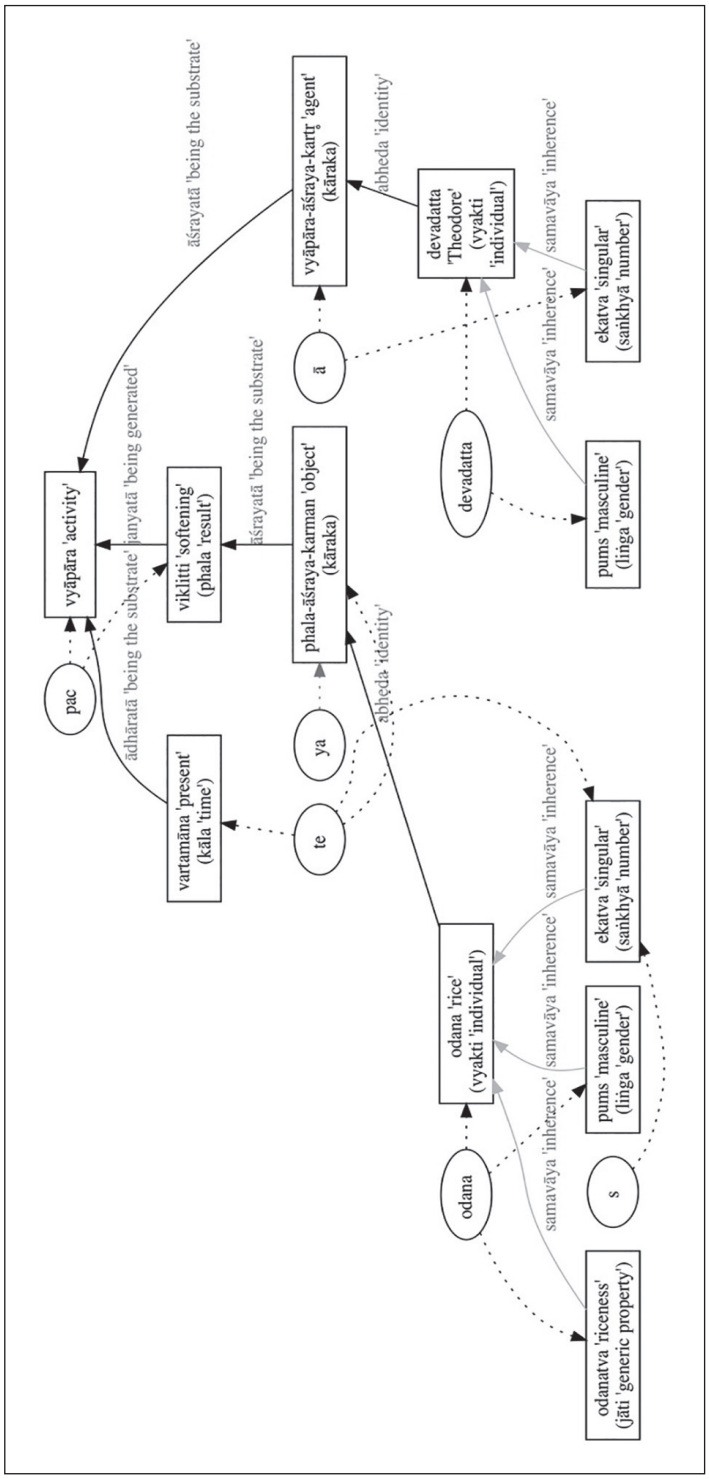


Figure 12 The mapping of morphemes in the passive sentence (8) onto the cognitive dependency tree for the detailed cognitive paraphrase in (9) according to Kauṇḍabhaṭ

10. *Conclusion*

Language permits the expression of complex multidimensional cognitive structures in the realm of thought in the single dimension of auditory speech and its representation in writing. Language is thus a phenomenon that involves an intimate connection between thought structures in consciousness and perceptible indications of it in physical dimensions. The effort to establish linguistics as an empirical science in the twentieth century led linguists to eschew contemplation of structures of thought and to restrict attention to the aspects of language immediately perceptible by the five senses. With attention focused on speech and its representation in writing, meaning took a subordinate place. The attention to experimental methods that culminated in behaviorism also led to a rift between linguistics on one side, and philology and cultural studies on the other. Although the concern of formal linguistics with morphological and syntactic structures brought attention to certain internal and universal aspects of language, the preoccupation with formalism as represented in the structure of expressed speech resulted in further neglect of the realm of meaning, and further distancing from philology and cultural expertise. Phrase-structure analysis, and transformational grammar in particular, imbues the linear sequence inherent in expressed speech with structural significance. Binary phrase structure trees in particular artificially group constituents under nodes inherent in the formalism that are devoid of any real linguistic import. Dependency analysis dispenses with the artificiality of binary division and the preoccupation with sequence and hence allows more freedom to represent complex cognitive linguistic structures in the domain of meaning. Cognitive linguistics expands the concern of linguistic analysis to broader neurological and cognitive concerns shared by psychological and cultural disciplines. The avenues of cognitive linguistics that are not restricted to empirical experimental methodology permit investigation of cognitive structures described by experts intimately familiar with the thought, language, and culture of different geographical and historical contexts.

The cognitive linguists in the long history of the sophisticated disciplines concerned with the analysis of language in India offer detailed analyses of the cognitive structures expressed in various languages of India, particularly in Sanskrit. The disciplines of grammar (Vyākaraṇa), logic (Nyāya), and ritual exegesis (Karmamīmāṁsā) engaged in debates concerning verbal cognition for more than two millennia. Pāṇinian grammar accounts for Sanskrit usage in a generative grammar that begins with concepts in the consciousness of the speaker, organizes those concepts in cognitive structures, and maps speech forms onto elements in those cognitive structures before applying morphophonemic and phonetic operations to produce linguistically valid expressions in linear speech. Pāṇinian commentators beginning with Kātyāyana and Patañjali in the third and second centuries BC, philosophers of language such

as Bhartṛhari, and the Indian cognitive linguists of the seventeenth and eighteenth centuries, Bhaṭṭoji Dīkṣita, Kauṇḍabhaṭṭa, and Nāgeśa, summarize the conclusions of the grammarians concerning the cognitive structures resulting from verbal cognition from the perspective of the listener. Their linguistic analysis restores the realm of consciousness to a position of priority in linguistic analysis.

The conclusions of the grammarians regarding cognitive linguistic structures and the mapping of speech forms onto cognitive structures by Pāṇinian generative grammar offer precise apparatus to represent the cognitive linguistic structures expressed in Sanskrit. Their intellectual contributions supply the materials necessary to represent the hierarchical structure of thought directly and to indicate its expression in speech by means of that cognitive structure. The cognitive dependency trees shown in Section 9.2 exemplify this procedure. Through a computational implementation of Paitāmbarī, the formalization of Pāṇini's grammar in XML I produced over the past few years and P. M. Scharf (2016) described, I plan to produce a comprehensive Pāṇinian lexicon enriched with the representation of internal dependency relations and with external dependency relations in the form of expectancies. The result will facilitate the precise dependency analysis of Sanskrit sentences, and the development of a Sanskrit parser that constructs cognitive dependency trees.

Abbreviations

1 = nominative, 2 = instrumental, 3 = accusative/3rd person, 4 = dative, 5 = ablative, 7 = locative, a = active, Det = determiner, n = neuter, N = noun, m = masculine/middle, NP = noun phrase, P = preposition, PP = prepositional phrase, pre = present, pop = present optative, s = singular, S = sentence, V = verb, VP = verb phrase

References

Allan, Keith, ed. 2012. *Oxford handbook of the history of linguistics.* Oxford: Oxford University Press.

Aussant, Émilie. 2009. *Le nom propre en Inde: Considérations sur le mécansime référentiel.* ENS Éditions.

Bhattacharya, Bishnupada. 1962. *A Study in Language and Meaning: A Critical Examination of Some aspects of Indian Semantics.* Calcutta: Progressive Publishers.

Biardeau, Madeleine. 1964. *Théorie de la connaissance et philosophie de la parole dans le brahmanisme classique.* Le monde d'outre-mer passé et présent, Première série, études 23. Paris; Le Haye: Mouton.

Blevins, James P. 2012. Chapter 18. American descriptivism ('structuralism'). In *Oxford handbook of the history of linguistics*, Keith Allan (ed), 436–455.

Bloomfield, Leonard. 1933. *Language.* New York: Holt, Rinehart and Winston.

Cardona, George, ed. *Nāgeśabhaṭṭa's Vaiyākaraṇasiddhāntaparamalaghumañjuṣā: critically edited with an annotated English translation.* Unpublished.

Chomsky, Noam. 1955–1956. *The logical structure of linguistic theory*. Manuscript.

Chomsky, Noam. 1957. *Syntactic structures.* The Hague; Paris: Mouton.

Chomsky, Noam. 1959. A review of B. F. Skinner's Verbal Behavior. *Language* 35: 26–58. URL: http://cogprints.org/1148/.

Chomsky, Noam. 1965. *Aspects of the theory of syntax.* Cambridge, Mass.: MIT Press.

Chomsky, Noam. 1967. Preface to the reprint of A Review of B. F. Skinner's Verbal Behavior. In Leon A. Jakobovits and Murray S. Miron (eds), *Readings in the psychology of language*, 142–143. Prentice-Hall.

Chomsky, Noam. 1975. *The logical structure of linguistic theory.* New York: Plenum.

Condillac, Etienne Bonnot de. 1746. *Essai sur l'origine des connoissances humaines.* Amsterdam: Pierre Mortier.

Condillac, Etienne Bonnot de. 1775. *Cours d'étude pour l'instruction du Prince de Parme;* Tome premier, Grammaire. Parme: Imprimérie Royale. [Reprint: Stuttgart-Bad Canstatt: Frommann-Holzboog, 1986.]

Croft, William. 2001. *Radical construction grammar.* Oxford: Oxford University Press.

Dās, Karuṇāsindhu (ed. and trans.) 1990. *A Pāṇinian approach to philosophy of language: Kauṇḍabhaṭṭa's Vaiyākaraṇabhūṣaṇasāra critically edited and translated into English.* 1st ed. Calcutta: Sanskrit Pustak Bhandar.

Deshpande, Madhav. 1992. *The meaning of nouns: semantic theory in classical and medieval India = Nāmārtha-nirṇaya of Kauṇḍabhaṭṭa.* Studies of classical India 13. Ph.D. diss. University of Pennsylvania, 1972. Dordrecht: Kluwer Academic Publishers. [Reprint: New Delhi: D.K. Printworld, 2007.]

Freidin, Robert. 2012. Chapter 19. Noam Chomsky's contribution to linguistics: a sketch. In *Oxford handbook of the history of linguistics*, Keith Allan, 456–487. Oxford: Oxford University Press.

Gune, Jayashree. 1978. *The meaning of tenses and moods: the text of Kauṇḍabhaṭṭa's Lakārārthanirṇaya, with introduction, English translation, and explanatory notes.* Ph.D. diss. University of Pennsylvania, 1974. Pune: Deccan College Postgraduate and Research Institute.

Hagelin, John. Is consciousness the unified field? *Science & non-duality.* URL: https://www.science and nonduality.com/videos/john-hagelin-is-consciousness-the-unified-field/. [Video.]

Harris, Zellig S. 1951. *Methods in structural linguistics.* Chicago: Chicago University Press.

Houben, Jan E. M. 1995. *The Saṁbandha-Samuddeśa (chapter on relation) and Bhartṛhari's Philosophy of Language.* Groningen: Egbert Forsten.

Iyer, K. A. Subramania. 1969. *Bhartṛhari: A study of the Vākyapadīya in the light of the Ancient Commentaries.* Deccan College Building Centenary and Silver Jubilee Series 68. Pune: Deccan College.

Janda, Laura A. 2015. Cognitive linguistics in the year 2015. *Cognitive Semantics* 1.1: 131–154.

Jha, Vashishtha Narayan, trans. 1977. *The vaiyākaraṇabhūṣaṇa of Koṇḍabhaṭṭa;* vol.1, *Dhātvarthanirṇaya: English translation.* 1st ed. Sri Garib Das Oriental Series 221. Delhi: Sri Satguru Publications.

Jha, Vashishtha Narayan (trans.) 1998. *The Vaiyākaraṇabhūṣaṇa of Koṇḍabhaṭṭa;* vol.2, *Lakārārthanirṇaya: English translation.* Sri Garib Das Oriental Series 244. Delhi: Sri Satguru Publications.

Joshi, Shivaram Dattatray (ed. and trans.) 1960. *Kauṇḍa Bhaṭṭa on the meaning of Sanskrit verbs: an English translation and annotation of the Vaiyākaraṇabhūṣaṇasāra, Chapter 1 with the introduction.* 3 parts. Nagoya studies in Indian culture and Buddhism: Saṁbhāṣā 14, 16, 17. Ph.D. diss. Harvard University, 1960. Chikusa-ku, Nagoya: Department of Indian Philosophy, Nagoya University.

Joshi, Shivaram Dattatray, ed. and trans. 1967. *The Sphoṭanirṇaya: chapter XIV of the Vaiyākaraṇabhūṣaṇasāra of Kauṇḍa Bhaṭṭa: edited with introduction, translation, and critical and exegetical notes.* Publications of the Centre of Advanced Study in Sanskrit; class C 2. Pune: University of Poona.

Joshi, Shivaram Dattatray. 2015. Kauṇḍa Bhaṭṭa. In *Encyclopedia of Indian Philosophies; The philosophy of the grammarians*, H. G. Coward and K. Kunjunni Raja (eds), 255–308. Princeton, N.J.: Princeton University Press; Delhi: Motilal Banarsidass.

Kulkarni, Amba P., Pokar, Sheetal & Shukl, Devanand. 2010. Designing a constraint based parser for Sanskrit. In Girish Nath Jha (ed), *Sanskrit Computational Linguistics: 4th International Symposium, New Delhi, India, December 2010, Proceedings*, 70–90. Lecture notes in computer science 6465. Berlin; Heidelberg: Springer-Verlag.

Kulkarni, Amba P. & Ramakrishnamacharyulu, K. V.. 2013. Parsing Sanskrit texts: Some relation specific issues. In Malhar Kulkarni and Chaitali Dangarikar (eds), *Recent Researches in Sanskrit Computational Linguistics: fifth international symposium proceedings 4–6 January 2013, IIT Bombay, India*, 191–212. New Delhi: D. K. Printworld.

Kunjunni Raja, K. 1963. *Indian Theories of Meaning.* Adyar Library Series 91. Second edition: 1969. Madras: The Adyar Library and Research Center.

Lakoff, George. 1987. *Women, fire and dangerous things: what categories reveal about the mind.* Chicago: University of Chicago Press.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar;* vol. 1, *Theoretical prerequisites.* Stanford: Stanford University Press.

Maat, Jaap. 2012. "Chapter 17. General or universal grammar from Plato to Chomsky. In Keith Allan (ed), *Oxford handbook of the history of linguistics*, 413–435.

Madhvanath, Sriganesh, Kleinberg, Evelyn & Govindaraju, Venu. 1997. Empirical design of a multi-classifier thresholding control strategy for recognition of handwritten street names. *International Journal of Pattern Recognition and Artificial Intelligence* 11.6: 933–946. World Scientific Publishing Company.

Ramakrishnamacharyulu, K. V., ed. 2015. *The Vaiyākaraṇasiddhaṭṭa: with the Nīrañjanī commentary by Ramyatna Shukla and Prakāśa explanatory notes by K. V. Ramakrishnamacharyulu.* Critically edited. Vol. I. Regards sur l'Asie du Sud / South Asian Perspectives 6; Shree Somnath Sanskrit University Shastragrantha Series 2. Pondichery: Institut Français de Pondichéry; Shree Somnath Sanskrit University.

Ramakrishnamacharyulu, K.V. (ed) 2019. *The Vaiyākaraṇasiddhāntabhūṣaṇa of Kauṇḍabhaṭṭa: with the Nīrañjanī commentary by Ramyatna Shukla and Prakāśa explanatory notes by K. V. Ramakrishnamacharyulu.* Critically edited. Vol. II. Collection Indologie 139; Shree Somnath Sanskrit University Shastragrantha Series 5. Pondichery: Institut Français de Pondichéry; Shree Somnath Sanskrit University.

Rathore, Sandhya. 1988. *Kauṇḍa Bhaṭṭa's Vaiyākaraṇabhūṣaṇasāra: an analytic study.* New Delhi: Indian Council of Philosophical Research. ISBN: 81-85636-41-9.

Sastri, Gaurinath. 1959. *The Philosophy of Word and Meaning: Some Indian approaches with special reference to the philosophy of Bhartṛhari.* Calcutta Sanskrit College Research Series 5. [Reprint: Calcutta: Century Press, 1983.]

Scharf, David C. 1989. Quantum measurement and the program for the unity of science. *Philosophy of Science* 56.4: 601–623.

Scharf, Peter M. 1996. *The Denotation of Generic Terms in Ancient Indian Philosophy: Grammar, Nyāya, and Mīmāṁsā.* Transactions of the American Philosophical Society 86, part 3. Philadelphia: American Philosophical Society.

Scharf, Peter M. 2012. Chapter 11: Linguistics in India. In Keith Allan (ed), *Oxford handbook of the history of linguistics*, 230–264. Oxford: Oxford University Press.

Scharf, Peter M. 2016. An XML formalization of the Aṣṭādhyāyī. In Amba Kulkarni (ed), *Sanskrit and computational linguistics: select papers presented at the 16th World Sanskrit Conference in the 'Sanskrit and the IT world' section 28 June – 2 July 2015, Sanskrit Studies Center, Silpakorn University, Bangkok*, 77–102.

Scharf, Peter M. 2021. On the source of the cognition of time in verbs. In Shyamanand Mishra (ed), *Studies in honor of S.N. Mishra*. Varanasi: Venkatesh Prakashan. Forthcoming.

Siewierska, Anna. 2012. Chapter 21. Functional and cognitive grammars. In Keith Allan (ed), *Oxford handbook of the history of linguistics*, 506–524. Oxford: Oxford University Press.

Skinner, Burrhus Frederic. 1957. *Verbal behavior.* New York: Appleton-Century-Crofts.

Subha Rao, Veluri. 1969. *The philosophy of a sentence and its parts.* New Delhi: Munshi Ram Manohar Lal.

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale.* Paris: Librairie C. Klincksieck. [Revised and corrected second edition, 1966.]

Tsoneva-Mathewson, Snezha T. 2009. Cognitive linguistics. In Vesna Muhvić-Dimanovski and Lelija Sočanac (eds), *Encyclopedia of life support systems*, vol. 4. Oxford: EOLSS Publishers/UNESCO. URL: http://www.eolss.net/sample-chapters/c04/e6- 91-12.pdf.

Linking Latin

Interoperable Lexical Resources in the LiLa Project

Marco C. Passarotti*, Francesco Mambrini*

This paper introduces the overall architecture of the LiLa Knowledge Base, which makes distributed language resources for Latin interoperable on the Web through the application of principles, ontologies and models developed by the Linguistic Linked Open Data community. In particular, the paper focuses on some linguistic aspects of the Latin lexicon that the lexical resources already linked to LiLa allow to investigate, showing how the network of connections that the LiLa Knowledge Base builds between lexical and textual resources for Latin is bigger than the parts considered singularly.

Keywords: Latin, lexicon, lemmatization, Language Resources, linguistic Linked Open Data, interoperability

1. *Introduction: The quest for interoperability of (research) data*

A recent trend that has gained traction in the area of scientific infrastructures is the emphasis on reusability and accessibility of scholarly data. A growing consensus has emerged on a set of principles that are now popularized in the often-quoted acronym FAIR – Findability, Accessibility, Interoperability and Reusability (Wilkinson et al. 2016). One of the purposes behind these guidelines is to overcome obstacles in the discovery and reuse of data, a problem that is particularly urgent, as the current COVID-19 pandemic has proved, in fields like the bio-medical sciences, where an effective and quick access to information is of the essence. Nevertheless, the emphasis to adopt models that lead to more integrated and discoverable digital datasets is gaining momentum in the community of language resources too. In particular, the growing interest in standards for representing linguistic collections as Linked Open Data (LOD) is also a response to the need for more carefully documented and more interconnected data in the field.

Latin and the ecosystem of digital projects of linguistic tools, lexica and corpora dedicated to that language represents a small but compelling example of the importance of such initiatives, as well as of the limitations that they intend to overcome. Over the last decade, the amount and diversity of the (often freely) available resources for Latin has grown exponentially.[1] However, most tools and collections

* Università Cattolica del Sacro Cuore.

1. See Passarotti et al. (2020) for an overview of the currently available language resources for Latin.

of textual or lexical material still live in insulated online environments, such as institutional websites, and are often unknown beyond the circle of the already knowledgeable experts.

Even though discoverability is a serious issue, more damaging still is the lack of interoperability. In the last years the community of Latin language learners and researchers has witnessed the publication, to name just a few interesting resources, of a Latin WordNet in at least two different projects, a series of Latin treebanks (i.e. corpora with word-by-word morphosyntactic annotation), and many other text collections with some forms of linguistic annotation, like lemmatization. However, how would a user leverage the combined power of these datasets to, for instance, discover all the subjects of verbs belonging to a certain WordNet synset? The problem can be readily summarized in the following terms: although digital corpora and lexical resources intuitively deal with the same *entities*, all connections between them exist (if at all) only in the mind of the human user.

The LiLa project was built to answer this very issue, by creating an infrastructure to link potentially all the resources that provide information about the same entities; by taking such steps, the project aims to respond to the challenge of interoperability highlighted by the FAIR best practices. In order to connect all the resources that attach some information to Latin words, LiLa builds a Knowledge Base, meant as a network of structured information about lemmas, the canonical forms that are used (or may potentially be used) by digital language resources to lemmatize word forms or to index dictionary entries.

In this paper we first introduce the model of the LiLa Knowledge Base and its architecture; in the following sections then we focus on some linguistic aspects of the Latin lexicon that the lexical resources already linked to LiLa allow investigating. Finally, we briefly address the question of why and how the whole, i.e. the network of connections that the LiLa architecture builds between those lexical resources and the corpora, is potentially more powerful than a simple sum of its parts.

2. *The LiLa Knowledge Base*

2.1. The role of lemmatization

As was said, an impressive array of digital resources for the study of Latin is currently available over the internet. The most obvious types of datasets in this respect are the digital libraries of Latin texts from all genres, media and periods, including such diverse typologies of documents as Late-Latin legal charters, inscriptions, ecclesiastical, historical and technical treatises, as well as the works of literature from the Classical era. A second group of resources that can be identified includes lexicons, both in the form of retro-digitized editions of printed dictionaries, and of digi-

tal-born databases. A third class includes tools for either automatic linguistic analysis and Natural Language Processing (NLP), or language learning, such as applications for generating exercises on vocabulary or syntactic constructions.

This situation is in fact an ideal use case for applying the paradigm of Linked Open Data. The expression "Linked Open Data" (LOD) points to a set of guidelines for the publication of "smarter" data on the web, which are interlinked through connections that can be semantically queried. Among others, two tenets that are particularly relevant for our discussion are: (1) the prescription to use Uniform Resource Identifiers (URIs), i.e. unambiguous and stable identifiers compliant to a formalized syntax, as the name of the data points; possibly, those URIs should be in the form of HTTP Uniform Resources Locators (URLs) that can be looked up in a web browser; (2) to link data across different data collections, so that information about the same entity from multiple sources may be attainable.

In our particular case of Latin corpora, dictionaries, lexica and NLP tools, all the resources are not only conceptually linked to the same "entities," but they also use comparable steps to identify them. Such "entities" are the words of the Latin lexicon, and the way words are identified in corpora, recognized by NLP tools in their input texts, and indexed in dictionaries is via lemmatization. Lemmas, then, are the ideal candidates to provide links across all the types of language resources, according to the principles of the LOD paradigm.

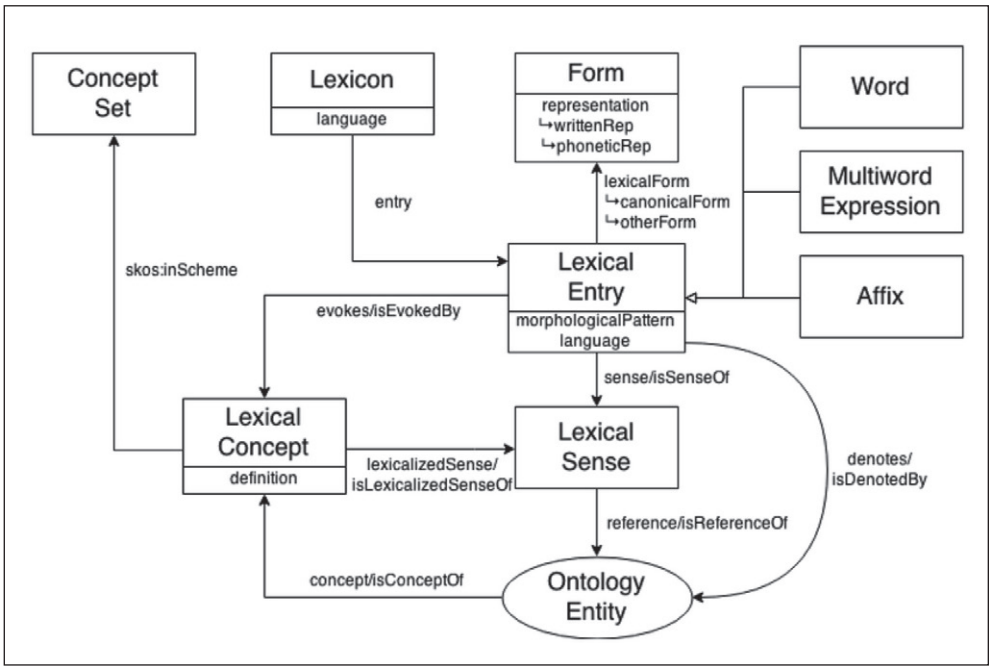
In standard Latin lexicography and corpus annotation, lemmatization is defined as the task of reducing the multiple inflected forms of a word to a form conventionally recognized as canonical. Accordingly, to lemmatize a noun form (e.g., the genitive singular *lupi*) means to reduce it to the nominative singular (*lupus* 'wolf'). Thus, the approach that LiLa adopted in order to connect the different resources is precisely to rely on this process: a corpus with a series of lemmatized tokens, as well as the output of NLP software that includes lemmatization, together with entries in lexicons that are indexed under a lemma, are all making statements about the same objects.

2.2. Form and meaning: LiLa and the OntoLex-Lemon model

While the emphasis on the practical task of lemmatization is peculiar to it, the lexically-based approach of LiLa and its emphasis on the special relation between canonical forms and words is entirely compatible with one of the best established model adopted by the Linguistic LOD community.

The OntoLex-Lemon module (Cimiano et al. 2020: 45-60), developed by the W3C Ontolex Group, has now become the *de-facto* standard for the representation of lexical resources. Figure 1 illustrates how the ontology provides a simple, but sophisticated vocabulary to describe lexical items, such as words, multi-word

Figure 1 The OntoLex-Lemon core model



expressions and affixes. The "Lexical Entry," the central concept in the core model, can be defined in both its formal and semantic properties. In the upper part of the diagram, the entry is in relation with a series of its (inflected) forms, which, in turn, have at least one (or more) written representations and, possibly, a phonetic representation. The semantic aspect of a word can be captured either in terms of the relation of denotation towards an entity defined in a formal ontology or knowledge base (for example, an entry in DBPedia representing a Wikipedia page), or by a reference to an evoked mental concept ("Lexical Concept"). In both cases, as shown in the diagram, the relation between the lexical item and the concept or the entity can be either expressed directly and/or be mediated via a "Lexical Sense."[2]

The OntoLex core model provides a suitable framework for LiLa. In particular, the working hypothesis about lemmatization can be converted into a formal definition that aligns itself with the rest of the classes and properties of the on-

2. See the definition of Lexical Sense in the official documentation at https://www.w3.org/2016/05/ontolex/#lexical-sense-reference.

tology. According to the schema of Figure 1, a lemma is defined as an instance of an OntoLex Form that can be linked to a Lexical Entry via the property "canonical form".

This design choice carries important consequences. To begin with, in OntoLex, a lexical entry cannot be assigned more than one part of speech (POS). Accordingly, if a word is licensed to being used in more than one syntactic function (as, for instance, an adverb or an adjective) and being annotated with different POS, then it must be differentiated into two different lexical entries. Moreover, a lexical entry cannot have more than one canonical form, but canonical forms can have more than one written representation. For Latin, this feature is particularly useful, as it can readily accommodate multiple variant and non-standard spellings of a wordform, which, in the case of a language with more than two millennia of written attestations, are particularly abundant. Thus, for instance, we can attribute to the lemma of the adjective *exspes* 'without hope' both the quoted spelling and the variant *expes*.[3] In the OntoLex ontology, however, written representations are modeled as data properties, i.e. properties that link resources to data values like strings or numbers; data properties do not point to other resources, and therefore cannot become in turn subjects of other statements. As a consequence, written representations cannot be assigned any other property, and it is impossible, within the current version of OntoLex, to make statements about them, such as in which testimonia a given variant spelling is attested, from what date or place, or how many occurrences of each of the variants are documented.

2.3. The Lemma Bank

The backbone of the network of resources in LiLa is made of a set of lemmas (called Lemma Bank) that is sufficiently large as to allow for all resources that deal with any kind of Latin texts or lexical collections to identify the forms used for lemmatization. According to the principles of LOD, the lemmas in the LiLa Lemma Bank are all identified by a unique identifier, which complies to the format of URIs. Moreover, each of them is described by a series of features and a series of relations that are formalized in the dedicated LiLa ontology.[4]

Among the linguistic features attached to lemmas, a special importance is given to the POS. As said, whenever a form is susceptible of multiple interpretations in terms of POS assignment, the solution within the OntoLex-Lemon model is to distinguish as many lexical entries as the POS concerned and, therefore, as many ca-

3. See http://lila-erc.eu/data/id/lemma/102584.
4. See https://lila-erc.eu/ontologies/lila/.

nonical forms. Accordingly, for instance, LiLa has three lemmas with written representation *cum* 'with, along, as', corresponding to the preposition, the adverb and the conjunction.[5] Other features includes the relevant morphological tags (e.g. gender and number for nouns) and the verbal or nominal inflection type, according to the definitions of traditional grammars.[6]

In some cases, deciding whether the orthographic and morphological variation related to a single lemma or multiple instances, each with its own URI, proved more challenging. Purely orthographic variations of the canonical form, that do not modify even a single trait of the morphological analysis, as in the case of *expes/exspes* quoted above, clearly entail a single lemma with multiple written representations. Whenever the variation brings about also a different morphological interpretation or a change in the inflectional category, on the other hand, we decided to create distinct instances. This is often the case with verbs attested with either a deponent or an active inflection, such as *somnio* and *somnior* 'to dream'.[7]

By applying these criteria, we generated the Lemma Bank of LiLa out of the lexical base provided by the database of the morphological analyzer LEMLAT 3.0 (Passarotti et al. 2017). As the software includes independent word lists targeted to the analysis of Classical Latin, Medieval Latin and proper names respectively, a considerable amount of repeated lemmas had to be identified and collapsed under a singular item.

Currently, the LiLa lemma bank includes 196,365 canonical forms, with a total of 232,340 written representations, ready to be linked to lexical resources or lemmatized texts.

3. *Lexical resources in LiLa*

At the moment of writing, six lexical resources are connected to the Lemma Bank of the LiLa Knowledge Base. Table 1 provides an overview of them. Although their coverage in terms of Latin lexical entries is variable, and in some cases quite low, these resources account for a rather wide spectrum of lexical and semantic phenomena.

The following subsections discuss how the linguistic aspects that each of the

5. See respectively http://lila-erc.eu/data/id/lemma/97201, http://lila-erc.eu/data/id/lemma/97207, and http://lila-erc.eu/data/id/lemma/97202.

6. See for instance the definition for the first verbal conjugation in the LiLa ontology at: http://lila-erc.eu/ontologies/lila/v1r.

7. See http://lila-erc.eu/data/id/lemma/125124 (*somnio*), and http://lila-erc.eu/data/id/lemma/125123 (*somnior*). For a more detailed discussion of the different classes of lemmas and of the properties linking them in the LiLa ontology see Passarotti et al. (2020).

Table 1 Lexical resources currently in LiLa

Title	Content	Status	Tot Entries
WFL	Word formation and derivation	Completed	36,138
Brill EDLIL	Etymology (I.-E. and Proto-Italic)	Completed	1,452
IGVLL	Etymology (Greek loan words)	Completed	1,759
Latin Affectus	Polarity	Ongoing	1,998
Latin WordNet	Word senses and synsets	Ongoing	1,424
Vallex 2.0	Valency lexicon	Ongoing	1,064

lexical resources currently in LiLa attempts to describe are represented by applying the LOD principles and the Semantic Web ontologies that were chosen to model the data.

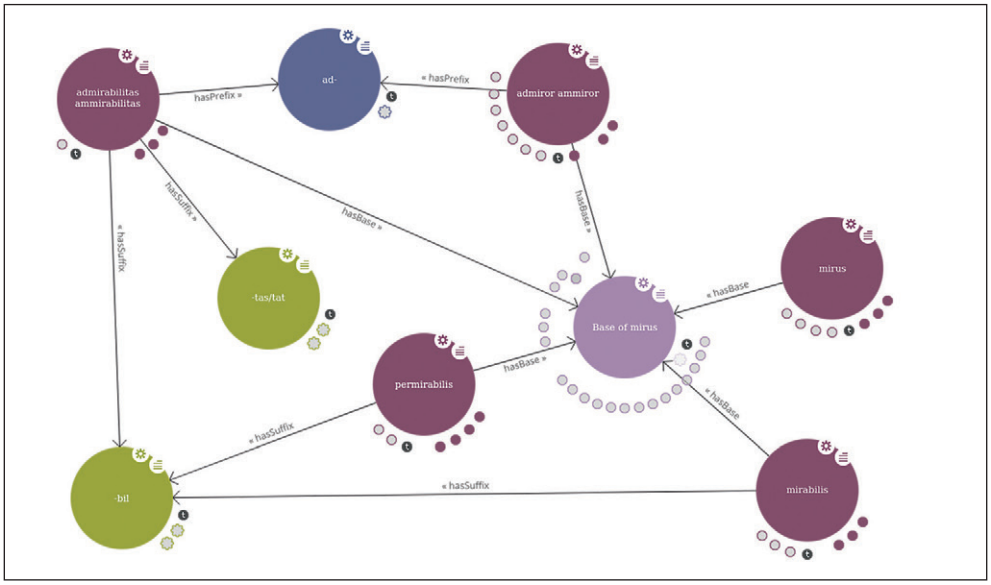
3.1. Word formation

Information on how Latin words are formed and are analyzable in terms of derivational processes is linked to the LiLa Knowledge Base in two different forms (Litta et al. 2019, Litta et al. 2020). The data used in both representations come from the Word Formation Latin (WFL) lexicon, a database where Latin words are described (and related to each other) in connection with word-formation rules. Following a step-by-step morphotactic approach, each process of word formation is regarded as the application of one rule (Litta 2018).

On the one hand, information on derivation is already attached to the canonical forms stored in the LiLa Lemma Bank. A total of 36,250 lemmas from the collection are linked to two special classes of morphemes that are recognizable in their derivational process. Affixes, further sub-specified as either prefixes or suffixes, are connected to forms where each of them is identifiable at any step in the derivational history of the word, so that, for instance, the prefix *per-* links forms such as *pernobilis* 'very famous', *perueho* 'to convey (through)', but also *imperfectus* 'imperfect'.[8] Lexical bases, on the other hand, are those morphemes that are left once all the affixes have been removed, and correspond to the lexical element that is shared by all the derivational family: so, for instance, the base of *ueho* 'to

8. For the prefix *per-* see http://lila-erc.eu/data/id/prefix/14, where all the 843 connected lemmas in the Lemma Bank are also listed.

Figure 2 Affixes, bases, and lemmas in the LiLa Lemma Bank



transport' links lemmas like *perueho*, *conuector* 'one who carries', or *inuecticius* 'imported, exotic'.[9]

The result of this representation is a network of derivational information like the one shown in Figure 2, which represents a lexical base surrounded by a series of connected canonical forms, together with two suffixes (*-bil* and *-tas/tat*) and one prefix (*ad-*) that are involved in the formation of the connected forms.

The output-oriented and descriptive model adopted in the LiLa Lemma Bank does not include any information on derivation processes (in terms of both word formation rules and order of their application), in accordance with the paradigm of Construction Morphology (Booij 2010, Litta, et al. 2020). At the same time, the LiLa Knowledge Base leverages the OntoLex ontology, with the help of some classes taken from its Morph extension that is currently under development (Klimek et al. 2019), in order to link also the entries and the word formation rules as represent-

9. See http://lila-erc.eu/data/id/base/134, with the 104 lemmas connected. Note that, although the OntoLex-Lemon ontology allows representing the morphemes as regular lexical entries with their own canonical form, we did not adopt this representation. Indeed, canonical forms of lexical entries *must* have at least one written representation, but, at the current stage of the work, we are not sure whether lexical bases comply to this constraint, as it is disputed which canonical form is to assign to lexical bases (a root? a stem?). Affixes and bases are therefore independent concepts of the LiLa ontology, not linked to OntoLex. In particular, lexical bases are just used as connectors between the lemmas that belong to the same derivational family in the Lemma Bank.

ed and applied WFL. In such representation, the LiLa lemmas are linked (via the OntoLex property "canonical form") to the lexical entries of the WFL resource. In their turn, each of these entries can be the source (input) and/or the target (output) of a word-formation relation, which is linked to a word-formation rule. In WFL 239 rule types are defined, distinguishing compounding from derivational rules, which are in turn sub-specified as suffixation, prefixation and conversion. In the LOD representation of WFL, classes of rules are described also in terms of POS of their input and output, such as for instance a suffixation rule that outputs an adjective from a verb.[10] To go back to the example mentioned above, the canonical form *imperfectus* from the LiLa Lemma Bank is linked to its lexical entry in WFL, which is, in turn, put in relation with both the verb *perficio* 'accomplish' and with *imperfectio* 'imperfection'. With the former, *imperfectus* is the output of a verb(participle)-to-adjective rule involving the negative prefix *in-*. With the latter, the relation is produced by a rule of the type adjective to noun that involves the suffix *-(t)io(n)*.[11]

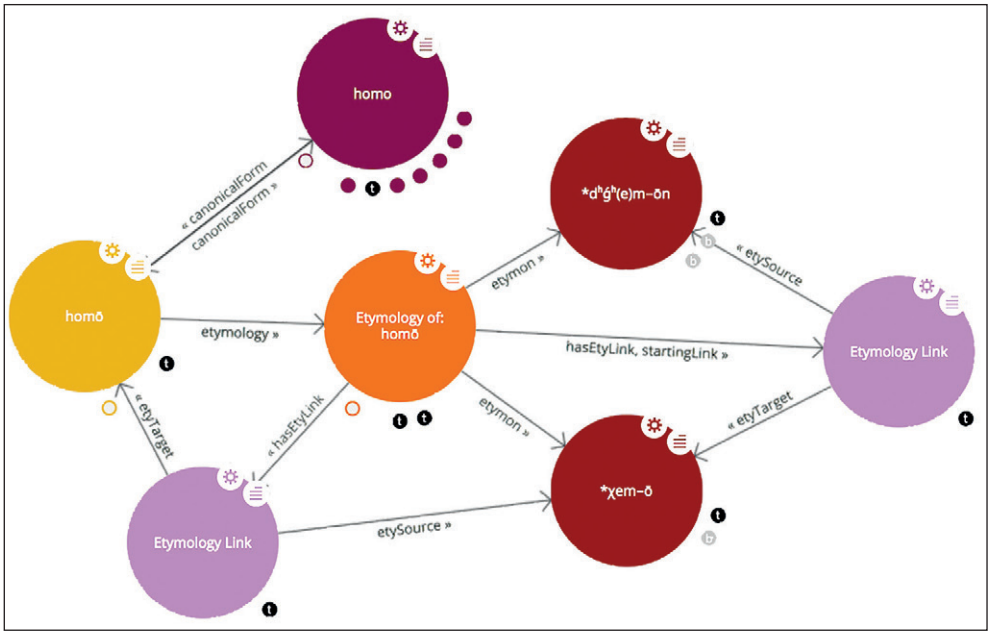
3.2. Etymology

The lemonEty ontology (Khan 2018) extends the OntoLex-Lemon model with classes and properties to express the etymological relations between words and forms. The module introduces a special sub-class of the OntoLex Lexical Entry called "Etymon", which includes all those lexical items that are used to discuss an etymological hypothesis, and that generally belong to a different language or a different diachronic phase as the entry whose etymology is being discussed. Reconstructed Indo-European words or borrowed terms from neighbor languages in an etymological dictionary of Latin are all possible examples of etymons. Etymologies are also defined as resources (in the technical sense that they are entities provided with a URI and which can become subjects or objects of statements). Instances of the class Etymology reify a scientific hypothesis about the origin of an entry and consist of a set of "etymology links" that connect a source to a target. One special advantage of this modeling strategy is the fact that both reified etymological hypotheses and links can be assigned any type of descriptive properties, from a bibliographical reference, to possible truth values. The full sequence of the argumentative steps on which the etymology relies can also be expressed, using a formalism such as the CRM_{inf} (Stead et al. 2019; Mambrini and Passarotti 2020).

Etymology links can be further specified in terms of the relation type that they postulate between a source word and a target. The prototypical instances are inheritance relation from an ancestor language or borrowing. As a matter of fact, LiLa

10. See http://lila-erc.eu/ontologies/lila/wfl/Suffixation/VerbToAdjective.
11. The WFL lexicon in LiLa can be accessed at https://lila-erc.eu/data/lexicalResources/WFL/Lexicon.

Figure 3 Etymology of *homo* in LiLa (according to de Vaan 2008)



makes use of both types of links to express the etymological hypotheses advanced in two lexical resources that are connected to the Knowledge Base.

The entries of the *Etymological Dictionary of Latin and the other Italic Languages* (de Vaan 2008) are all connected to etymologies that encompass a series of links to Proto-Indo-European and Proto-Italic source etymons. Figure 3 represents the etymology of *homo* 'man, human being' in LiLa, as reconstructed by de Vaan (2008).[12] The reified etymological hypothesis is represented by the node at the center of the picture ("Etymology of: homo"); the etymology connects the lexical entry ("homō") to a chain of etymological links (the red nodes) that go from the Proto-Indo-European reconstructed ancestor **dʰǵʰ(e)m-ōn* back to the Latin word via the properties etySource and etyTarget.

The retro-digitized *Index Graecorum Vocabulorum in Linguam Latinam Translatorum* (IGVLL, Saalfeld 1874) integrates these data with a list of loan words from Ancient Greek. In this case too, we chose to model the information with the lemon-

12. See http://lila-erc.eu/data/lexicalResources/BrillEDL/id/etymology/116. The *Etymological Dictionary* by de Vaaan can be accessed in LiLa at https://lila-erc.eu/data/lexicalResources/BrillEDL/Lexicon. Note that LiLa does *not* include a full version of the printed dictionary, but only the etymological links between the Latin words and the I.-E. and Proto-Italic etymons. The lexical entries are linked to their pages on the website of the publisher, so that subscribing readers can access the full text of the dictionary.

Ety extension of the OntoLex core ontology. The lexical entries of IGVLL are also linked to reified etymologies, which consist of one single etymology link from the Greek to the Latin word.[13]

3.3. Polarity

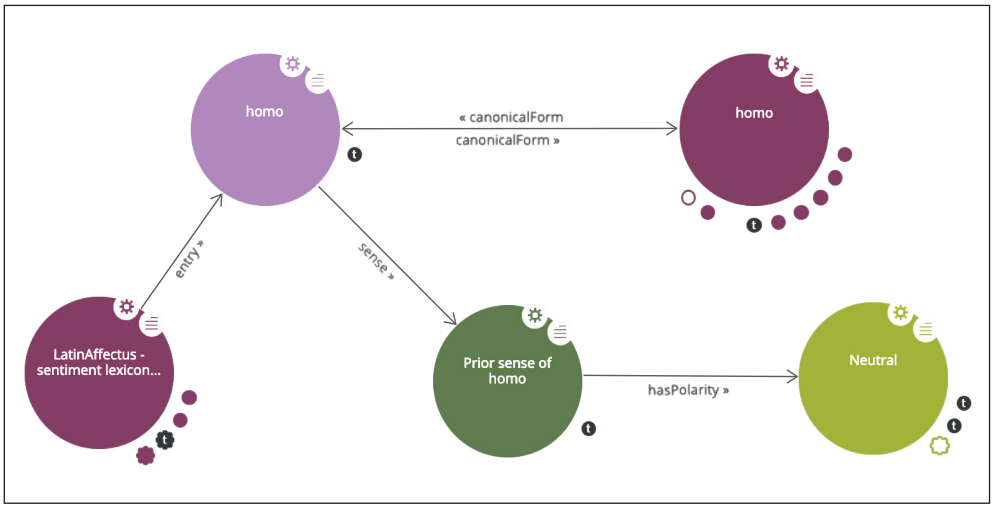
As shown in Figure 1, the OntoLex-Lemon model provides a flexible set of properties and classes to describe the plurality of senses and meanings of a word. Whether the lexical entry is set in relation to an evoked mental concept or a denoted entity, these relations can be either direct and/or mediated through a lexical sense.

The *LatinAffectus - sentiment lexicon for Latin* is a lexical resource that records the prior polarity of a selection of Latin adjectives and nouns (Sprugnoli et al. 2020a). By "prior polarity" we intend the positive or negative value associated to an item in the lexicon of a language, independently from the actual usages in context. Therefore, the polarity value is attached to a single, general sense of a word, and it is measured on a scale of five scores: -1, -0.5 (negative pole), 0 (neuter), +0.5, +1 (positive).

The scores were originally assigned manually by experts working independently, whose annotation underwent an extensive reconciliation phase, then extended with information from derivational morphology (Sprugnoli et al. 2020b). Further iterations of manual annotation and reconciliation are in progress.

Figure 4 shows how the polarity values provided by LatinAffectus are repre-

Figure 4 Polarity of *homo* from LatinAffectus in LiLa



13. The IGVLL lexicon in LiLa can be accessed at https://lila-erc.eu/data/lexicalResources/IGVLL/Lexicon.

sented in LiLa. In particular, Figure 4 shows the polarity of the noun *homo*. The word does not carry any *a-priori* positive or negative connotations, and is therefore recognized as neutral (score of 0). The node for the lexical entry in LatinAffectus is linked to the lemma *homo* in the Lemma Bank through the property "canonical form" and to its prior sense via the property sense. In turn, the prior sense of *homo* is linked to its polarity value (Neutral) via the property "has polarity."[14]

3.4. Senses, synonyms and valency

WordNet is a lexical database of English that groups certain categories of words (nouns, adjectives, verbs and adverbs) into sets of cognitive synonyms known as "synsets" (Fellbaum 1998). Although originally developed for English, several projects have extended the application of the synsets to the lexicons of many more languages (Pianta et al. 2002; Bond and Foster 2013). In 2004, Minozzi (2017) created a Latin WordNet with a total 9,378 lemmas, spread across 8,973 synsets, that where automatically classified using the Italian and English WordNet and bilingual dictionaries to match the Latin words. This dataset represents a foundational resource, but its usefulness is limited by a series of shortcomings, such as the arbitrary selection of the included lemmas and the existence of wrong connections to synsets inherited from English (Franzini et al. 2019). More recently, a larger Latin WordNet including more than 70,000 entries, has been developed, by following the same automatic procedure as the one built by Minozzi.[15] The precision and recall of the synset assignment of this Latin WordNet still has to be assessed.

The English WordNet is also one of the largest datasets that were converted and distributed as LOD.[16] Particularly, an official RDF version of the Princeton WordNet is available, which uses OntoLex-Lemon to model the relations between words, senses and synsets (McCrae et al. 2014; Cimiano et al. 2020: 215–28). The synset is there interpreted as an OntoLex Lexical Concept, i.e. as an "abstraction, concept, or unit of thought that can be lexicalized by a given collection of senses."[17]

Starting from Minozzi's Latin WordNet and the RDF Princeton distribution, the LiLa team has worked on two different tasks. Firstly, we decided to revise manually as many lemma-synset associations from the available Latin WordNet as possible, in order to correct the instances of misalignment (precision) and to integrate the senses established in Latin lexicography that were not represented in the origi-

14. The LatinAffectus lexicon in LiLa can be accessed at https://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon.
15. See Short in this volume.
16. See Cimiano et al. (2020: 217) for a history and an overview of the different projects dealing with the publication of the WordNet(s) as LOD.
17. https://www.w3.org/2016/05/ontolex/#lexical-concept.

nal version (recall) (Franzini et al. 2019). Secondly, our goal was to publish this refined resource as LOD, following the model of the RDF WordNet closely. Since this double effort goes on in parallel, the published LOD version of the Latin WordNet[18] now includes 1,424 lexical entries, distributed among 5,220 synsets.[19]

Following the OntoLex-Lemon model (see Figure 1), the relation between a word and a synset is mediated through a lexical sense. A second resource for Latin that is being actively developed for inclusion into the LiLa network also draws on the list of word senses associated with the entries of the Latin WordNet. Passarotti et al. (2016) built the first version of a valency lexicon, named Latin Vallex, on the evidence of the syntactic annotation from two Latin treebanks, namely the *Index Thomisticus* Treebank (Passarotti 2019), and the Latin Dependency Treebank (Bamman and Crane 2006). All valency-capable lemmas occurring in the semantically annotated portion of the two treebanks are assigned one lexical entry and one valency frame in Latin Vallex.

The structure of Latin Vallex is closely modeled on that of the Czech PDT-VALLEX (Hajič et al. 2003). Each entry of the lexicon consists of a sequence of frame entries that contain each a sequence of frame slots corresponding to the arguments of the given lemma. Each frame slot is assigned a semantic role labeled with the same tags used for the semantic annotation of the Prague Dependency Treebank (Mikulová et al. 2006). In the current stage of the work, in order to enhance the coverage of the Latin Vallex, the process of creation of the valency frames is running independently from the treebank annotation and is fully intuition-based. The task is currently being performed manually: the valency frames included in the first version of Latin Vallex have been updated, cleaned or rectified. Currently, 1,064 lexical entries have been annotated, for a total of 8,327 valency frames.

Valency frames are strictly linked to senses: for each recognized sense of a valency-capable word, a frame is established intuitively, and assigned the set of its obligatory complements. The senses to be annotated are taken directly from the repertoire of word senses in the Latin WordNet; thus, each entry-synset pair for the valency-capable words in the Latin WordNet is annotated (or will be annotated, once the work is completed) with a valency frame.

As the core module of OntoLex is not sufficiently expressive to capture the predicate structure of a lexical entry, we have adopted the PreMOn extension to model the information in the Latin Vallex and to map the entities to other schemas such as the Latin WordNet (Corcoglioniti et al. 2016). The property and classes that are

18. See http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon.

19. Note that the LiLa dataset also includes all relations between synsets that are stipulated in the Princeton WordNet (like antonymy, hypernymy and hyponymy). In total, the LiLa LatinWordNet provides information on 22,742 synsets.

needed to describe the valency frames are formalized in a dedicated extension of the PreMOn core ontology.

Following the model of PreMOn, each different frame of any given entry in Latin Vallex is an instance of the Valency Frame class. The arguments involved in the valency frames of Latin Vallex are called "frame slots," and are defined as a subclass of PreMOn's Semantic Roles. The slots, which are defined locally for each semantic class, correspond to the so-called "functors" (i.e. semantic values of syntactic dependency relations) of the Functional Generative Description (Mikulová et al. 2006). One of the main use cases of PreMOn was the mapping of different predicate models, namely those for PropBank (Palmer et al. 2005), NomBank (Meyers et al. 2004), VerbNet (Schuler 2005) and FrameNet (Baker et al. 1998). Therefore, the ontology is ideally suited to express the link between the word-synset pairs and the predicate analyses in the Latin Vallex. The PreMOn core module defines a special reification of the relation between a given semantic class and a lexical entry, called "Conceptualization." The linking itself is performed with instances of the class Mapping, which is defined as a set of conceptualizations, semantic classes, or semantic roles. Following this schema, which is also applied in the PreMOn data itself,[20] we match the words-synsets pairs of Latin WordNet and the predicate analyses in Latin Vallex by means of mapping instances linking the corresponding conceptualizations.

Figure 5 shows the complex of the WordNet and Valency annotation for one of the 12 senses recorded for the Latin verb *dono* 'to give, donate', namely the one connected to the synset 00887463-v of the Princeton WordNet (version 3.0).[21] The lexical entry (yellow node at the center of the image) is connected to both a valency frame (left-hand side) and a synset (on the right). A mapping node (in purple, directly below the entry)[22] connects the two conceptualizations.[23]

4. C*onclusion. Parts of a whole: interoperability in LiLa*

The diagram in Figure 6 provides a plastic representation of the interconnection between different layers of information linked to a canonical form in the LiLa Lemma Bank. The lemma of the adjective *malus* 'bad, evil' is described with a series of fea-

20. See for instance the mapping between a synset and a predicate analysis for the English verb "to leave out" at http://premon.fbk.eu/resource/sense-Ep7UGYgbEXbB3B2uGhZamc.

21. The synset encompasses the English lemmas: "devote, commit, give, dedicate, consecrate", with the following definition: "give entirely to a specific person, activity, or cause". See http://wordnet-rdf.princeton.edu/pwn30/00887463-v. Note that the figure also shows a second synset that is recorded as hyperonym of 00887463-v.

22. http://lila-erc.eu/data/lexicalResources/LatinVallex/id/Mapping/wn-val-l_100087_00887463-v.

23. The Latin Vallex connected to Latin WordNet in LiLa can be accessed at https://lila-erc.eu/data/lexicalResources/LatinVallex/Lexicon.

Figure 5 WordNet and valency annotation for a sense of *dono*

tures, some of which (namely, the POS and the inflection paradigm) are represented in the image. Moreover, the lemma is linked to a lexical base that is common to all the derived words belonging to the same derivational family of *malus*, like *malitia* 'malice', *maleficus* 'evil-doing, nefarious', or the rare verb *maleficio* 'to practice black magic' (shown in Figure 5).

In addition to the properties of the lemma, the canonical form is directly linked to three lexical entries from as many different resources (yellow nodes). The entry for *malus* from the etymological dictionary by de Vaan (2008) lists the inheritance relations from the Indo-European and Proto-Italic reconstructed forms. The entry from the WFL lexicon is connected to several formations in which the adjective is involved, the one with the verb *maleficio* being the only one represented in the diagram. Finally, on the left-hand side of the lemma, the entry for *malus* in the LatinAffectus lexicon registers the a-priory negative sense of the adjective.

The series of connections illustrated in Figure 6 (which, by the way, omits reference to the Latin WordNet or Vallex, as no information of the sort is available for the lemma in question) is already sufficient to provide a plastic visualization of the strong "network effect" that the model adopted by LiLa achieves. One of the most immediate applications to leverage the power of interoperability is to cross the information from one resource to the another in order to study the Latin lexicon. Traditionally, for instance, etymological dictionaries like de Vaan's (2008) do not discuss *all and every* word whose roots can be traced back to an Indo-European ancestor. Rather, the authors proceed by identifying a key lemma for a whole entry, where all the lexical items that are derived from it by regular word-formation processes are also listed. Even such list of "derivatives" is far for complete, both for the chronological limits that the dictionary authors would set to their work, and for the obvious limitations of space (in printed books) and time (available to the compiler). In a LOD scenario, these two tasks can be decoupled and assigned to two different resources, one dedicated to etymology, the other to derivational morphology. Students and scholars interested in a full list of items in the lexicon that trace back their etymology to a certain Indo-European root can interrogate the two datasets simultaneously.[24] Other possibilities offered by the interconnections between lexical resources include, for instance, a study on the semantic aspects of derivational processes. Indeed, the coverage of the LatinAffectus lexicon was extended by targeting words associated with morphemes capable of altering or conveying a polarity value, such as the prefix *in-* with negative meaning (Sprugnoli et al. 2020b).

24. See Litta et al. (2020: 177–82) for a comparison between the data on derivative words in the dictionary of de Vaan (2008) and in LiLa.

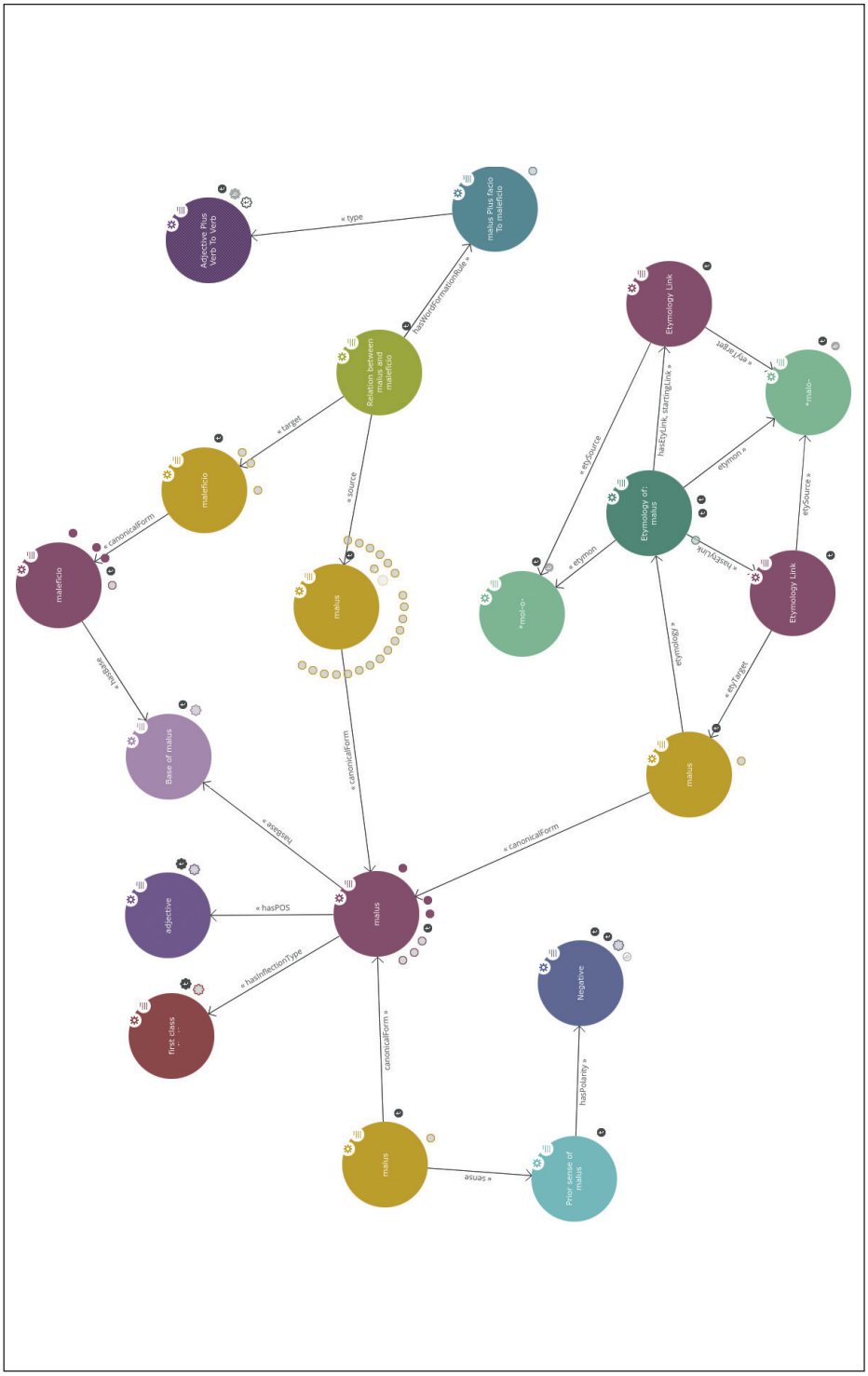


Figure 6 *Malus* (adjective) in LiLa: properties and lexical resources

Further possibilities are opened when tokens from textual corpora are integrated into the network. At present, all lemmatized corpora register the lemma of each token as a string associated to the form in the text; the same type of output is produced by automatic lemmatizers. A connection to the LiLa network is obtained when this lemma string is associated unambiguously to one of the lemmas in the Knowledge Base, for instance by matching it to one of the written representations of the canonical forms. Mambrini and Passarotti (2019) report on the results of a preliminary experiment of matching: up to 81.52% of the tokens in the Latin PROIEL UD corpus (v. 2.3) could be unambiguously associated with a LiLa lemma with a simple string match. Considering the central role played by textual resources in LiLa, the project developed a tool to automatically link a Latin raw text (i.e. without any linguistic annotation) to the LiLa Knwoledge Base. The tool, called Text Linker, makes use of an automatic lemmatizer, built upon a large training corpus that collects more than 6 million words taken from Latin texts of different eras.[25]

One of the added values of the LiLa Knowledge Base is interoperability between the different kinds of information about words provided by lexical resources (ranging from mono-/bilingual definitions to etymologies, polarity, morphology etc.) and their actual usage in texts stored in corpora, which makes of LiLa the natural venue where publishing any available or newly created language resource for Latin. By applying the principles of the LOD paradigm, it is today possible to interlink the (meta)data from any Latin resource, thus exploiting to the best its specific contribution in relation to the overall picture. This feature is essential when dealing with ancient languages that can be studied only through the attestations that survived throughout the centuries. Furthermore, interoperability between resources in LiLa is achieved by using (and sometimes extending) data models, categories and ontologies widely adopted in the larger community of Linguistic LOD. This design strategy is what makes Latin resources speak "the same language" as the resources of many other languages, both ancient and modern.

25. The training corpus was compiled by joining texts from various resources, including the LASLA corpus, the Latin treebanks available in Universal Dependencies, a subset of the Computational Historical Semantics corpus and the full text of *Confessiones* by Augustinus. Lemmatization criteria were harmonized among the corpora and the Universal POS tags were assigned (https://universaldependencies.org/u/pos/index.html).

Websites

Computational Historical Semantics corpus: https://comphistsem.org/home.html
English WordNet: https://wordnet.princeton.edu/
LASLA corpus: http://web.philo.ulg.ac.be/lasla/
Latin PROIEL UD corpus (v.2.3): http://hdl.handle.net/11234/1-2895
Latin WordNet: https://latinwordnet.exeter.ac.uk
Text Linker (beta version): http://lila-erc.eu:8080/LiLaTextLinker
UD: https://universaldependencies.org

References

Baker, Collin F., Fillmore, Charles J. & Lowe, John B. 1998. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, 86–90. Montreal, Quebec: Association for Computational Linguistics. https://doi.org/10.3115/980845.980860

Bamman, David & Crane, Gregory. 2006. The Design and Use of a Latin Dependency Treebank. In *TLT 2006: Proceedings of the Fifth International Treebanks and Linguistic Theories Conference*. Prague: Institute of Formal and Applied Linguistics.

Bond, Francis & Foster, Ryan. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Volume 1*, Hinrich Schuetze, Pascale Fung & Massimo Poesio (eds), 1352–1362. Sofia, Bulgaria: Association for Computational Linguistics.

Booij, Geert. 2010. Construction Morphology. *Language and Linguistics Compass* 4: 543–55.

Cimiano, Philipp, Chiarcos, Christian, McCrae, John P. & Gracia, Jorge. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Cham: Springer. https://doi.org/10.1007/978-3-030-30225-2

Corcoglioniti, Francesco, Rospocher, Marco, Aprosio, Alessio P. & Tonelli, Sara. 2016. PreMOn: A Lemon Extension for Exposing Predicate Models as Linked Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 877–884. Portorož, Slovenia: European Language Resources Association.

Fellbaum, Christiane (ed). 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Franzini, Greta, Peverelli, Andrea, Ruffolo, Paolo, Passarotti, Marco C., Sanna, Helena, Signoroni, Edoardo, Ventura, Viviana & Zampedri, Federica. 2019. Nunc Est Aestimandum. Towards an Evaluation of the Latin WordNet. In *Sixth Italian Conference*

on Computational Linguistics (CLiC-It 2019), Raffaella Bernardi, Roberto Navigli & Giovanni Semeraro (eds), 1–8. Bari, Italy: CEUR-WS.org.

Hajič, Jan, Panevová, Jarmila, Urešová, Zdeňka, Bémová, Alevtina, Kolárová, Veronika & Pajas, Petr. 2003. PDT-VALLEX: Creating a Large-Coverage Valency Lexicon for Treebank Annotation. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2)*, Joakim Nivre & Erhard W. Hinrichs (eds), 57–68. Växjö: Växjö University Press.

Khan, Anas F. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information* 9: 304. https://doi.org/10.3390/info9120304

Klimek, Bettina, McCrae, John P., Ionov, Maxim, Tauber, James K., Chiarcos, Christian & Bosque-Gil, Julia. 2019. Challenges for the Representations for Morphology in Ontology Lexicons. In *Proceedings of Sixth Biennial Conference on Electronic Lexicography, ELex 2019*, Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubíček, Simon Krek & Carole Tiberius (eds), 570–591. Brno: Lexical Computing. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_33.pdf

Litta, Eleonora. 2018. Morphology Beyond Inflection. Building a Word Formation-Based Lexicon for Latin. In *Formal Representation and the Digital Humanities*, Paola Cotticelli-Kurras & Federico Giusfredi (eds), 97–114. Newcastle upon Tyne: Cambridge Scholars Publishing.

Litta, Eleonora, Passarotti, Marco C. & Mambrini, Francesco. 2019. The Treatment of Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*, Eleonora Litta, Marco C. Passarotti, Žabokrtský Zdeněk & Ševčíková Magda, 35–43. Prague: Institute of Formal and Applied Linguistics, Charles University.

Litta, Eleonora, Passarotti, Marco C. & Mambrini, Francesco. 2020. Derivations and Connections: Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. *The Prague Bulletin Of Mathematical Linguistics* 115: 163–86.

Mambrini, Francesco & Passarotti, Marco C. 2019. Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 13th Linguistic Annotation Workshop*, Annemarie Friedrich, Daniz Zeyrek & Jet Hoek (eds), 71–80. Florence, Italy: Association for Computational Linguistics.

Mambrini, Francesco & Passarotti, Marco C. 2020. Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Globalex Workshop on Linked Lexicography. LREC 2020 Workshop*, Ilan Kernerman, Simon Krek, John P. McCrae, Jorge Gracia, Sina Ahmadi & Besim Kabashi (eds), 20–28. Paris: European Language Resources Association.

McCrae, John P., Fellbaum, Christiane & Cimiano, Philipp. 2014. Publishing and Linking WordNet Using Lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, Christian Chiarcos, Petya Osenova, John P. McCrae & Cristina Vertan. Reykjavik, Iceland: European Language Resources Association.

Meyers, Adam, Reeves, Ruth, Macleod, Catherine, Szekely, Rachel, Zielinska, Veronika, Young, Brian & Grishman, Ralph. 2004. The NomBank Project: An Interim Report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, 24–31. Boston: Association for Computational Linguistics.

Mikulová, Marie, Bémová, Allevtina, Hajič, Jan, Hajičová, Eva, Kolářová, Veronika, Kučová, Lucie, Lopatková, Markéta et al. 2006. Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank. Annotation Manual. 30. Prague: UFAL. https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf

Minozzi, Stefano. 2017. Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval. In *Strumenti digitali e collaborativi per le Scienze dell'Antichità*, Paolo Mastrandrea (ed), 123–133. Venezia: Ca' Foscari. https://doi.org/10.14277/6969-182-9/ANT-14-10

Palmer, Martha, Gildea, Daniel & Kingsbury, Paul. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31: 71–106. https://doi.org/10.1162/0891201053630264

Passarotti, Marco C. 2019. The Project of the Index Thomisticus Treebank. In *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, Monica Berti (ed), 299–319. Berlin: De Gruyter.

Passarotti, Marco C., Budassi, Marco, Litta, Eleonora & Ruffolo, Paolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, Gerlof Bouma & Yvonne Adesam (eds), 24–31. Gothenburg: Linköping University Electronic Press.

Passarotti, Marco C., Mambrini, Francesco, Franzini, Greta, Cecchini, Francesco M., Litta, Eleonora, Moretti, Giovanni, Ruffolo, Paolo & Sprugnoli, Rachele. 2020. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici* 58: 177–212. https://doi.org/10.4454/ssl.v58i1.277

Passarotti, Marco C., Saavedra, Berta G. & Onambele, Christophe. 2016. Latin Vallex. A Treebank-Based Semantic Valency Lexicon for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 2599–2606. Portorož, Slovenia: European Language Resources Association.

Pianta, Emanuele, Bentivogli, Luisa & Girardi, Christian. 2002. MultiWordNet: Developing an Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*. Vol. 152. Mysore, India: Global WordNet Association.

Saalfeld, Günther A. 1874. *Index graecorum vocabulorvm in linguam latinam translatorum quaestiunculis auctus*. Berolini: Berggold.

Short, William M. This Volume. WordNets, Sembanks, and the Challenge of Semantic Polyvalency.

Schuler, Karin K. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD

dissertation, University of Pennsylvania. https://repository.upenn.edu/dissertations/AAI3179808

Sprugnoli, Rachele, Mambrini, Francesco, Moretti, Giovanni & Passarotti, Marco C. 2020a. Towards the Modeling of Polarity in a Latin Knowledge Base. In *WHiSe 2020 Workshop on Humanities in the Semantic Web 2020*, Alessandro Adamou, Enrico Daga, & Albert Meroño-Peñuela (eds), 59–70. Heraklion, Greece: CEUR-WS.org.

Sprugnoli, Rachele, Passarotti, Marco C., Corbetta, Daniela & Peverelli, Andrea. 2020b. Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 3078–3086. Marseille, France: European Language Resources Association.

Stead, Stephen, Doerr, Martin, Ore, Christian-Emila, Kritsotaki, Athina et al. 2019. *CRMinf: The Argumentation Model, Version 0.10.1 (Draft)*. http://new.cidoc-crm.org/crminf/sites/default/files/CRMinf%20ver%2010.1.pdf

de Vaan, Michiel. 2008. *Etymological Dictionary of Latin: And the Other Italic Languages*. Leiden and Boston: Brill.

Wilkinson, Mark D., Dumontier, Michel, Aalbersberg, Ijsbrand J., Appleton, Gabrielle, Axton, Myles, Baak, Arie, Blomberg, Niklas et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.1

New Functions and Updates of the Resource DiACL – Diachronic Atlas of Comparative Linguistics (Version 2.1)

Gerd Carling*, Rob Verhoeven*, Filip Larsson*,
Olof Lundgren*, Linus Nilsson

The paper describes the new functions and updates of the database DiACL, version 2.1. Typological data sets can now have a global coverage, instead of previously being confined to a Focus area. In addition, the online interface to both typological data sets and (lexical) word lists has been updated to reduce the number of mouse clicks necessary to navigate through the data. Inconsistencies between different data sets have been removed, aiming at a similar standard for all sets. In addition, the grammatical and semantic information for Indo-European lexical data is improved, and lacunae in languages and lexemes are filled. The structure of etymological trees is improved and standardized, which has also been done with the connections between lexemes and concept meanings. The improvements and updates will create a more uniform standard and quality of both the infrastructure and the content of the database.

Keywords: database, phylogenetics, historical linguistics, Indo-European, Amazonian

1. *The database DiACL – Diachronic Atlas of Comparative Linguistics*

The DiACL – Diachronic Atlas of Comparative Linguistics database is an open access database harboring types of data that are frequently used for the purpose of computational phylogenetic research, i.e., lexical data (basic vocabulary and culture vocabulary), and grammar typology data (morphosyntactic features and variants) (Carling 2017; Carling et al. 2018). A rationale of the database is providing data sets that reflect, with a high degree of granularity and accuracy, accrued wisdom from traditional historical-comparative linguistics. This includes the adaptation of comparative linguistics to computational methods of analyzing data by means of data character coding that encapsulates historical-comparative information. These concern, e.g., geographic position and temporal extent in prehistory, or known prehistoric language by lexical borrowing.

The database uses *language* (not dialect) as a unique identifier, and this concept includes contemporary, historical, as well as reconstructed languages. Languages are definable units, which can be constrained by time and space, but with a varying de-

* Lund University.

greeof certainty. Therefore, the unique identifier *language* is connected to metadata information that includes time period, spatial extent (focal point/ polygons), position in a cladistic reference tree, and reliability (modern language/ dead language (well documented)/ dead language (fragmentary)/ reconstructed language). All this information is retrieved by conventional methods and is sourced in scientific literature.

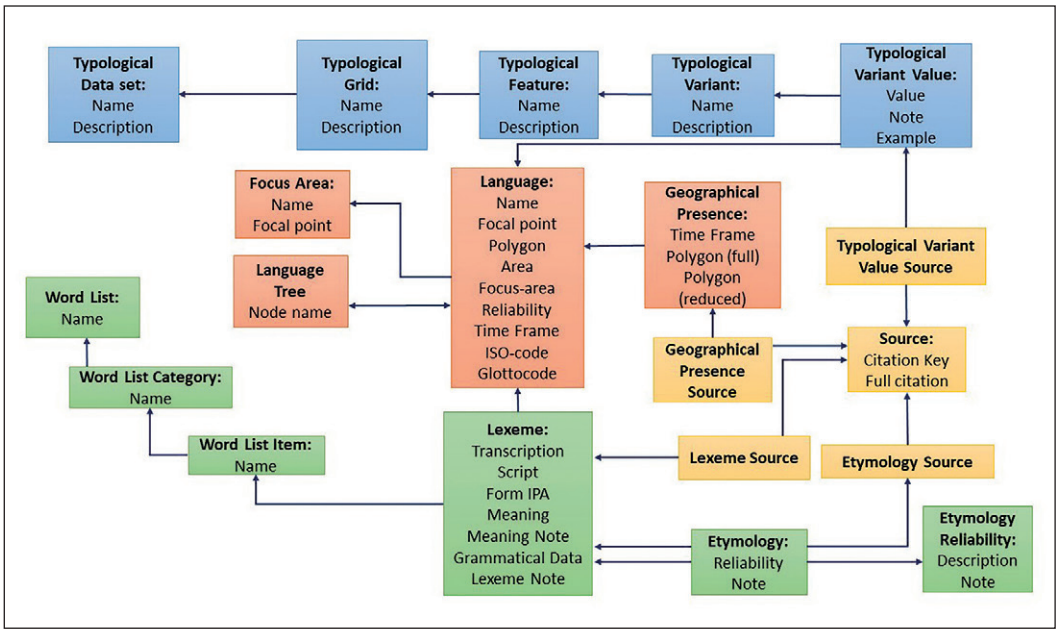
A uniquely identified *language* is linked to various tables with linguistic data. Feature organization and data character coding are implemented according to a hierarchical model of increasing detail and possible local adaptation, which is implemented both in lexical and typological data (Carling et al. 2018). This hierarchical model aims at capturing variability at the functional side of the data, enabling definitions and recoding of data sets for testing various models of analyzing data. Additional tables contain: 1) Language metadata (including geographic position, temporal extension, reliability, and language familytree topology), 2) Lexical and etymological data, including various types of concept lists, such as Swadesh lists and culture vocabularies, organized into semantic classes, 3) Typological and morphosyntactic data sets, organized into a four-levelled hierarchy, and 4) Source data (Literary sources or Informants) (Figure 1).

In its current state, the data of the DiACL database mainly contains data from three geographical areas, Eurasia, Amazonia, and Austronesia. Data sets are available from approximately twenty families, but the database is a living repository and research instrument, meaning that the set of languages and families can be extended.

2. *Demands for changes for the new version DiACL (2.1)*

In previous descriptions of the database and its functionality (Carling 2019), the areal focus of the resource is highlighted. The notion of "Focus area" as a central concept to all (typological) data is a central node in earlier versions of how tables were organized (Carling 2019: 25) However, this focus has been changed in the new, updated version of the database. This has been done for several reasons. Most importantly for the reason that the research agenda has been modified. A substantial part of previous and partly ongoing projects, in which the DiACL database serves as a repository, focuses on ancient languages and comparative-historical reconstruction (Carling and Cathcart 2021a, 2021b). The Indo-European language family and adjacent families serve as a central part of the research in this area. In addition, lexical data is oriented more towards locally adapted culture vocabularies (Carling et al. 2019) rather than universal concept lists, such as Swadesh or Leipzig-Jakarta lists (Tadmor and Haspelmath 2010). Other involved families include Austronesian and Tupí, but the focus of data compilation is similar to Eurasian families: locally adapted culture vocabularies and adapted typological features rather than Swadesh lists and typologi-

Figure 1 Overview of the organization of tables in the database DiACL, version 2.1. Orange: Language and language metadata, Blue: Typology/morphosyntax, Yellow: Sourcing, Green: Lexicon



cally universal feature sets. Even though this functionality is still present in the new version of the database, ongoing and planned research required a modification of the database structure to account for a more general, global perspective. This has motivated us to introduce some changes and improvements of the database structure.

3. *Description of the new database structure of DiACL (2.1)*

The core of the database is the entity *Language*, which contains languages along with their attributes. All other parts of the database are linked to it (directly or indirectly, when it concerns Lexemes). The Language types include contemporary languages, extinct languages, and reconstructed language states (e.g., Proto-Indo-European, Proto-Tupí). Several tables link directly to the central Language table, with which they constitute the section of the database that pertains to languages and their metadata, such as Focus area, Language tree, and Geographical presence. Outside of this core, pertaining to languages per se, the database has (much like in previous versions), three subsections: 1) Lexicology, 2) Typology, and 3) Source. Language metadata on the table Language (Figure 1) include a standardized name, ISO 693-3 code, Glottolog code (new field), alternative names, location, time frame, language area,

and reliability. As in previous versions of the database, Location gives a focal point, which renders the most prototypical geographic center for a language. Alternative names give various names of the language in different descriptions or in different languages. Time frame gives an estimation of the period within which a language is spoken. Language area is a more detailed classification of language areas, compared to Focus area. Reliability has four distinctions: Modern language, Dead (well attested), Dead (fragmentary), and Reconstructed. Dead (well attested) is the label for languages with corpora large enough to provide data equivalent to living languages, whereas Dead (fragmentary) is the label for languages with scarce documentation.

The most important difference in contrast to the previous version is that Focus area is no longer a central concept to the typological data. Previously, besides being linked to by the Language entity, Focus area was the central entity of which the Typology section was dependent. The dependency of typological data on a Focus area has now been removed. Focus area does remain linked to the Language entity, acting as its attribute, similar to the other language metadata that is recorded.

In addition to the changes in the core structure, the Typology subsection is organized under a more general concept of "data set". A "data set" in Typology can be any predefined set of grammar features, from general data sets with a global coverage (e.g., word order in all the world's languages), similar to WALS features (Dryer and Haspelmath 2013), or they can be locally and typologically highly specialized (e.g., evidentiality in Tupí). This is similar to the arrangement that was already present for the Lexicology section. A "data set" in the Lexicology section is a concept list, labeled Word List, which can be a list of general and global coverage (e.g., a Swadesh list), or a highly locally and semantically specialized list (e.g., Metallurgy in Central American languages). The internal structure is further discussed below.

Apart from these changes, the organization of the Typology subsection remains the same as before. The Typology subsection is organized according to a hierarchical structure of Grid (general grammatical domain), Feature (more specialized feature), Variant (even more specialized variant of a feature), and Value (binary value of feature variant) (Carling et al. 2018). The typological data sets can be downloaded with their binary values in JSON and Excel formats, which preserve the hierarchical organization of the data sets from the database. The binarized data can then be recoded or adapted to ask various types of research questions (Carling and Cathcart 2021a, 2021b).

The Lexicology subsection, like in the previous versions of the database, has the function of a comparative lexical cognacy database. A comparative lexical cognacy database is a resource of lexemes of languages connected to concepts (Dellert et al. 2020; Rzymski et al. 2020), which *also* connects lexemes to cognacy trees. In the new version of DiACL, like in the previous one, lexemes are organized by concepts lists, labeled Word Lists (Poornima and Good 2010). Word Lists can be organized

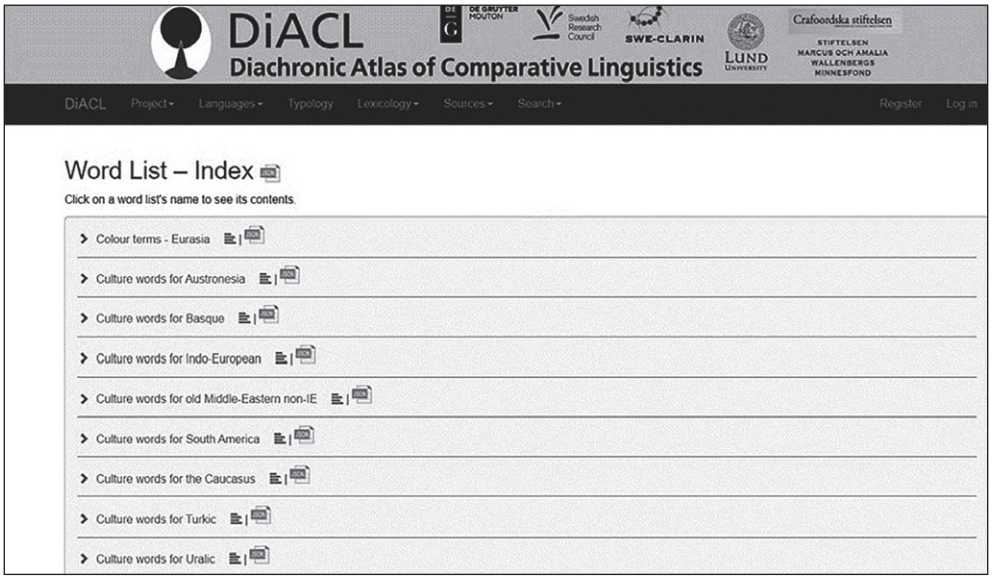
hierarchically, with a general level Word List (a concept list defined by any criteria, local or functional), Word List Category (any semantic subgroup of a concept list), and Word List Item (a defined concept). Linked to Word List Items are the Lexemes, each of which belong to a particular language. The most important attribute of the Lexeme table is the Transcription form field, which is a required field and gives the transcribed form of the lexemein Latin script. Next follows a field Transliteration form, which allows the Lexeme to be captured in its (original) writing script, such as Georgian or Cyrillic script. Furthermore, there is a possibility to include the IPA transcription of a lexeme (this field is at current state not filled for any language). The Meaning field records the full meaning of a lexeme, not the connected concept meaning, accounting for polysemy in languages. The following field Meaning note records information connected to the meaning of the lexeme. Next, there is a field for Grammatical data, which may record information about inflection/conjugation of the lexeme, such as the gender of nouns. Finally, a field Note gives a possibility to add relevant data that does not fit into any other field.

Like in the previous version of the database, Etymology is a function of the database that connects Lexemes to each other in a cognacy relationship. Etymology accounts for etymologies by linking two Lexemes, defined as Ancestor Lexeme or Descendant Lexeme. Relationship between those two Lexemes can be labeled in several ways: Unspecified (the etymological relation has not yet been processed within the database), Inherited (there is a secured cognacy relation which pertains to lineage), Probably borrowed (likely borrowing), Certainly borrowed (certain borrowing), Uncertain origin, Wanderwort (word most likely borrowed, but the exact source and direction cannot be defined), and Derivation (the lexeme is derived by morphological derivation).

At the frontend of the database, Etymologies are visible as trees, in which the user can move back and forth within etymologies. An important detail of the way in which we chose to implement etymological constructions are the Stub languages, which are used to define the roots of lexical etymologies, beyond the reconstructed families. Stub languages consist of Stub words, generally defined by a concept meaning and a cognate number. Stub lexemes normally have proper reconstructions in proto-languages as descendants, but they may also lead directly to attested languages. The organization of Stub languages, Stub words, and etymologies is same as in the previous version of the database (Carling 2019), but in the current process of cleaning the database, Stub languages have become increasingly important (see further below).

The new version DiACL (2.1) also includes a change in the interface. This change serves the immediate purpose of improving the ergonomics: the hierarchical structure of organizing typological and lexical data implied a relatively high number of

Figure 2 Screenshot of the new function of collapsing and expanding data sets, displayed for Word Lists in the Lexicology section



mouse clicks to overview data. Data sets – typological data sets and lexical word lists – can now be expanded and collapsed by one single click (Figure 2). Just as the typological data discussed previously, also the lexical data (including all information of lexemes and cognacy), can be downloaded as JSON or CSV files (see further below).

4. *Current updates: gender coding, etymological cleaning, and cleaning of connections between lexemes and concept lists*

Current updates of the data in the database are motivated both by ongoing research as well as by a general aim to make the information of the database more standardized and of higher quality. An ongoing research project on clitic pronouns in Austronesian and Indo-European (Swedish Research Council, 2020-2022) expands the typological section with new data sets. However, these additions will not be described in this paper, which will focus on the lexical section and the improvements of gender information, cognacy coding, and links between lexemes and concept meanings.

Much of earlier redundancy and error in the lexical section, including lexeme doublets, faulty etymologies, mistakes and inconsistencies in the lexeme information, was caused partly by irregular policies. However, most importantly, mistakes and inconsistencies were caused by internal problems in the Excel sheets with data that constituted the data repository before the creation of the database, which were

migrated in batches to the database. The database infrastructure was built during the period 2013-2017, and the construction of the resource early on met a number of technical obstacles (Lenardič 2020). During the construction of the infrastructure, data was fed into Excel sheets and during the migration of these Excel sheets to the database, inconsistencies in the way data had been recorded resulted in several errors. The cleaning of these errors has been a long process, which is still ongoing. We will describe these processes in greater detail. Due to funding from Swedish Research Council (2019-02967) and Åke Wiberg Foundation (2019-02967), we are able to complete and clean the Swadesh and culture lists of Indo-European, as well as to complete the coding of gender in this data. The previous version of the database contained a Swadesh 100-list for 153 Indo-European languages. The aim of compiling this list, even though there are similar lists available, such as IE-LEX (Chang et al. 2015) or its continuation IE-COR, was to have a basic vocabulary list for phylogenetic purposes, in case other similar resources were made unavailable. The original list that we compiled was very simple; it consisted of an Excel-sheet with a list of languages and cognacies, containing only lexemes of the target languages, with no additional meaning, and excluding loans. This list was used for phylogenetic inference in several publications (Carling 2019; Cathcart et al. 2018) and was migrated to the database. In the process of migrating this data set to the database, the missing meaning of individual lexemes was automatically given as the concept meaning of the Swadesh term (and flagged as such). All other information (e.g., grammatical information) was lacking in the data. In addition, the orthography of the data was not controlled and standardized, and since several sources were often used for one language, there was a lot of orthographic inconsistency. In the current cleaning, we go back to the dictionary sources for each language and check and complete the information for each lexeme. An important aspect of the coding, with immediate relevance to the research project, is the coding of gender of nouns. In the Indo-European family as a whole, there is a relatively large variation in the typology of gender (Matasović 2004). We record gender as it is given in dictionaries, standardized by the abbreviations m=masculine, f=feminine, n=neuter, a=alternans (masculine in singular, feminine in plural), c=common gender (masculine/feminine gender against neuter). In case of several genders of a noun, the symbols given, separated by a comma, reflect the order of gender preference of that lexeme. In genderless languages or cases where gender is not known, nouns are marked by s=substantive (=noun, to avoid confusion with n=neuter). In addition, lexemes that are loans are added (and marked as loans accordingly). The semantic information of lexemes is completed by filling the exact and complete (including polysemy) information of meaning in the field "Meaning". Synonyms are separated by comma and different meanings by semicolon, e.g., 'wolf, grey wolf; thief'. To avoid unnecessary redundancy due to inconsist-

ency between different dictionary sources, polysemous meanings are standardized to exclude too detailed and redundant variants, which may be the case of dictionaries such as *The Oxford English Dictionary* or *A Greek-English lexicon* (Liddell and Scott 1901). In addition, metaphorical uses of lexemes, such as TABLE LEG for LEG are not included (and removed if they have been included).

In the previous version of the database, the recorded lexemes for culture differ substantially from the lexemes recorded for the Swadesh lists. The former Lexemes had more complete grammar information (word class, gender, etc.), meanings were given according to the dictionary meaning, and cognacy trees had greater detail, including reconstructed forms. An important issue was that etymological trees also included lexemes where the meaning has changed (Carling 2019: 179ff.). Compared to the Swadesh data, the culture vocabulary data set had fewer languages (around 100 Indo-European languages), leading to a substantial discrepancy between the two sets. Another important difference, caused by the process of migration, was that in the culture list of Indo-European, all lexemes of cognates, including reconstructed lexemes as well as lexemes that had changed their meaning, were connected to the Word List Item (i.e., concept meaning), leading to strange connections between concept meanings and lexemes in the database. This caused many inconsistencies in the data base. In the current update, these differences have been removed. Lexemes that belong to etymological trees but which have changed their meaning are disconnected from the Word List Items (but not from the etymologies).

Another issue in the database, partly caused by the process of migration, was the organization at the root of etymological trees or cognacies. Most cognacies were originally created during the migration of Excel sheets. Already from the start, the idea was to let all cognacies root in a joint precursor (which we referred to as "Top node" internally), defined as a Stub lexeme in a Stub language, in which the Stub language was named after the Word List as well as the family, such as "Stub PIE Swadesh" or "Stub Culture Indo-European" (Carling 2019: 178ff.). Apart from organizing etymologies as well as concept lists, the Stub languages play an important role when concept lists are downloaded into JSON or CSV files: in the downloading process, the user is requested to select first a Word List and then a Top Node language. The downloading harvests the information from the root (Top node) to the leaves (Lexeme) of etymological trees. The resulting file gives the full information of lexemes, as well as the previous ancestor and the root of the etymological tree (see Table 1). However, the migration caused many inconsistencies here. Very often, the earliest precursor of etymological trees was not a Stub lexeme in a Stub language, but rather a reconstructed form in a proto-language. Forms in proto-languages are supposed to be rendered as a reconstruction, and if not possible, as the concept meaning and a number (e.g., woman-18). Stub lexemes are given as the concept meaning and a

Table 1 Overview of fields included in downloaded JSON and CSV files of Word Lists of the Lexicology section

Field	Comment
LexemeId	Database ID of Lexeme
Language	Name of Language
LanguageId	Database ID of Language
X	Latitude
Y	Longitude
Language Reliability	Reliability of Language
WordListItem	Word List Item = Concept meaning
WordListItemId	Database ID of Word List item
Transcription	Transcription of Lexeme
Transcription (no markup)	Transcrption of Lexeme
Ipa	IPA of Lexeme
Meaning	Full meaning of Lexeme
LexemeNote (no markup)	Note field
Grammatical Data	Grammatical information (e.g., gender)
Top node (no markup)	Transcription of earliest precursor in etymological tree
Top Node's LexemeId	Database ID of earliest precursor in etymological tree
Parent (no markup)	Ancestral Lexeme (immediate precursor) in etymological tree
Parent's LexemeId	Database ID of ancestral Lexeme (immediate precursor)
Reliability	Status of link between ancestral Lexeme and Lexeme (Inherited, Borrowed etc.)
Source Language	Language of Ancestral Lexeme (important, e.g., in case of borrowings)

number. In the previous version, this was both inconsistent and incomplete, leading to a scenario where it was not possible to download the full data amounts of Word lists without knowledge of the inconsistencies in the database. We account for this in the new version, and in addition, all Indo-European etymologies are screened for mistakes. New lexemes from languages, e.g., languages that lack culture vocabulary data, are added to etymological trees.

5. *Conclusion*

The current updates of the database and changes in the content of lexemes, described in this paper, aim to serve as a basis for future research. There are several improvements and updates in the current version of the database (DiACL 2.1). Some of them

target improvements and updates of the database structure and online interface. A substantial change is the removal of Focus area as an organizational unit, according to which all typological datasets had to be defined. This allows for adding data sets of global coverage. This means that lexical and typological data can be organized into data sets, which can have any content and extension in terms of area and family. An update of the interface facilitates reading and overviewing the data, reducing the number of mouse clicks. The new function, which is similar for lexical and typological data, lists the data sets with a possibility to collapse and expand all sublevels in the data. If desired, maps can be displayed to show the expansion of features, but these maps do not occupy large parts of the screen in the display mode. Changes to the proper data aim to improve quality and detail, as well as to reduce differences between and within data sets. Lexemes have full meaning and complete grammatical information, including gender for nouns. Etymological trees are corrected and improved, and redundant links between concepts and lexemes are removed. In addition, the structure of etymological trees is improved: all etymological trees have their roots in a Stub lexeme in a Stub language. Upon downloading data, lexemes are compiled from the concepts meanings, and etymologies are identified from the Stub language concepts in the database. These improvements and changes of the database structure will facilitate future research, in which there is a larger freedom to formulate research questions and define data sets, as well as a possibility to extract complete and representative sets for analyzing in computationally more powerful programs, such as ArcMap or R.

Websites

ArcMap: https://desktop.arcgis.com/
IE-COR: https://www.shh.mpg.de/dlce-research-projects/ie-cor-database
R: https://www.rstudio.com/

References

Carling, Gerd. 2017. DiACL – *Diachronic Atlas of Comparative Linguistics Online.*

Carling, Gerd. 2019. *Mouton Atlas of Languages and Cultures*. Vol. 1: *Europe and West, Central and South Asia*. Berlin - Boston: Mouton de Gruyter.

Carling, Gerd & Cathcart, Chundra. 2021a. Evolutionary dynamics of Indo-Eurpean alignment patterns. *Diachronica* 38(3): 358–412.

Carling, Gerd & Cathcart, Chundra. 2021b. Reconstructing the evolution of Indo-European grammar. *Language* 97(3): 1–38.

Carling, Gerd, Cronhamn, Sandra, Farren, Robert, Aliyev, Elnur & Frid, Johan. 2019. The causality of borrowing: Lexical loans in Eurasian languages, *PLoS ONE* 14(10): e0223588. https://doi.org/10.1371/journal.pone.0223588v.

Carling, Gerd, Larsson, Filip, Cathcart, Chundra A., Johansson, Niklas, Holmer, Arthur, Round, Erich & Verhoeven, Rob. 2018. Diachronic Atlas of Comparative Linguistics (DiACL) – A Database for Ancient Language Typology, *PLoS ONE* 13(10): e0205313. https://doi.org/10.1371/journal.pone.0205313.

Cathcart, Chundra, Carling, Gerd, Larsson, Filip, Johansson, Niklas & Round, Erich. 2018. Areal pressure in grammatical evolution. *Diachronica*, 35(1): 1–34.

Chang, Will, Hall, David, Cathcart, Chundra & Garrett, Andrew. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1): 194–244.

Dellert, Johannes, Daneyko, Thora, Münch, Alla, Ladygina, Alina, Buch, Armin, Natalie, Clarius, Grigorjew, Ilja et al. 2020. NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources & Evaluation* 54 (1): 273–301.

Dryer, Matthew S. & Haspelmath, Martin. 2013. WALS Online.

Lenardič, Jakob 2020. Tour de CLARIN: Interview with Gerd Carling, Tour de CLARIN (clarin.eu: CLARIN).

Liddell, Henry George & Scott, Robert. 1901. *A Greek-English lexicon*. Oxford: Clarendon.

Matasović, Ranko. 2004. *Gender in Indo-European.* Heidelberg: Winter.

Poornima, Shakthi & Good, Jeff. 2010. Modeling and Encoding Traditional Wordlists for Machine Applications. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, ACL 2010. Uppsala: Association for Computational Linguistics, 1–9.

Rzymski, Christoph, Tresoldi, Tiago, Greenhill, Simon J., Wu, Mei-Shin, Schweikhard, Nathanael E., Koptjevskaja-Tamm, Maria, Gast, Volker et al. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data* 7(1): 13.

Tadmor, Uri & Haspelmath, Martin. 2010. Borrowability and the notion of basic vocabulary. *Diachronica* 27(2): 226–46.

WordNets, Sembanks, and the Challenge of Semantic Polyvalency

William Michael Short*

In this paper, I describe the design and architecture of the ancient-language WordNets and then consider some of the challenges that are faced in the creation of a corpus of semantically annotated Sanskrit, Greek and Latin texts utilizing WordNet constructs. Very specifically, I consider the problematics of textual polyvalency for semantic annotation in electronic corpora, and how a WordNet-based text encoding schema can help take account of the natural multiplicity of meanings in discourse, and avoid predetermining the interpretation of texts in the course of annotation (an aspect of corpus-building that remains undertheorized). A series of examples from texts of different periods of Latin literature not only illustrates the different kinds polyvalency that may require representation via any adequate annotation scheme, but also emphasizes that these considerations are relevant to corpus-building across classical studies, medieval and renaissance studies, Neo-Latin studies, and other disciplines.

Keywords: WordNet, semantic annotation, encoding, polyvalency, polysemy, Sanskrit, Greek, Latin

1. *Introduction*

The creation of an interlinked, interoperable, and inter-reliant system of classical-language WordNets, building on the pioneering work of the (now defunct) MultiWordNet project of the Fondazione Bruno Kessler, is the on-going concern of an international team of scholars at the University of Exeter, the University of Pavia, and Harvard's Center for Hellenic Studies.[1] These projects are bringing rich semantic data to the landscape of digital resources for Sanskrit, Greek, and Latin – including information about the sense of words and their relations, lexical field organization, and high-order structures of meaning such as metaphors and metonymies – under the perspective of up-to-date theories of meaning (especially Roschian prototype theory, and Lakovian conceptual metaphor theory in cognitive semantics). What is more, they do so in a way that enables cross-linguistic comparison of such structures, by utilizing a single, shared set of primitive sense definitions ("synsets") to describe word meaning in all three languages where commonalities can be observed, or to pinpoint divergences in semantic configurations. By making a standard

* University of Exeter.

1. For discussion of the latest unified efforts to create WordNets for the ancient languages, see Biagetti, Zanchi and Short (2021).

API (application programming interface) available across all three systems, they permit any user or computer application to programmatically access lexical and semantic content in a consistent manner, regardless of language (or simultaneously for all languages). This means that NLP tools already available for the ancient languages can automatically and immediately take advantage of the WordNet data to improve their functionality, accuracy, and scope.[2]

In this paper, I provide basic context around the creation of the ancient-language WordNets and then move on to consider one of the most promising – but also most challenging – aspects of this overall endeavor: namely, the creation of a corpus of semantically annotated Sanskrit, Greek and Latin texts utilizing WordNet constructs. Very specifically, I consider the problematics of textual polyvalency for semantic annotation in electronic corpora, and how a WordNet-based text encoding schema can help take account of the natural multiplicity of meanings in discourse, and avoid predetermining the interpretation of texts in the course of annotation (an aspect of corpus-building that remains undertheorized). I give a series of examples from texts of different periods of Latin literature not only to illustrate the different kinds polyvalency that may require subtle variations in the annotation scheme, but also to emphasize that these considerations are relevant to corpus-building efforts across disciplines (classical studies, medieval and renaissance studies, Neo-Latin studies, and so on).

2. *The WordNet Framework*

The WordNet framework provides a robust scaffold for building computational systems that describe the semantic structures of a language at different levels (See Fellbaum 1998 and 2017). In a WordNet, the lexemes (and, to a lesser extent, phrasal lexical items) of a given language are assigned to one or more "synsets", which correspond to its various senses. Each synset – represented, minimally, by a unique identifying alphanumeric string and a gloss in English – can therefore be seen as constituting a set of semantically related words (a *synonym set*). For example, in the English WordNet, the synset glossed as "an individual building where a person resides" comprises the words *house* and *home*. *Home* is also included in the synset glossed as "place of emotional attachment, where one resides," thus capturing this word's special se-

2. The WordNet API will be integrated into the Classical Language Toolkit (CLTK) to bring semantic data to text processing pipelines of all kinds. One project that depends on the ancient language WordNets is *Cylleneus,* a "semantic" search engine for Sanskrit, Greek, and Latin. In effect, *Cylleneus* enables texts in Greek and Latin to be queried via the meanings of words (expressed in English or potentially any other language for which a WordNet is available), as well as their syntactic configurations.

mantic features and differentiating it from *house*. A WordNet can also include information about semantic relations between synsets, such as antonymy, hypernymy, or holonymy, as well as lexical relations.[3] Synsets may also be grouped into broader semantic fields or conceptual domains called semfields. In the WordNets of the classical languages, words from Sanskrit, Greek, and Latin are keyed to a single pool of synset identification numbers, where appropriate.[4] Thus, in the Latin WordNet *domus* and *domicilium* are tagged with the same synset identifier as *house* and *home* in the English WordNet, just as *oikos* in the Greek WordNet and *niketa* in Sanskrit are – to the extent that the meanings of these terms overlap. Of course, as the senses of words in these languages are often idiosyncratic, the WordNets also include many language-specific synsets in order to model the semantic system of the individual languages as accurately as possible.

Besides a focus on modelling the distinct meaning structures of Sanskrit, Greek, and Latin, sensitivity to different kinds of figurative senses and figurative relations between and amongst words of the lexicon is a distinguishing feature of the ancient language WordNets, which goes well beyond the Princeton (English) WordNet framework.[5] Whereas the original WordNet specification treats semantic structure as basically flat, we recognize the importance of figurative relations both at the level of word sense organization and at the level of overall organization of the semantic system, in terms of supra-lexical metaphorical and metonymic mappings. Etymological information is added in order to specify etymological senses, as distinct from conventionalized usages of words. For example, the Latin word *considero* is conventionally used to mean 'to consider; reflect'; however, its etymological sense (< *con-*

3. Unfortunately, in many cases these distinctions have not always been strictly and rigidly maintained. Very often WordNets treat antonymy, for example, as a relation between words (like English *hot* and *cold*): however, strictly speaking, antonymy is a relation between senses. *Hot* is not an antonym of *cold* when it means 'very attractive'; and *cold* does not have *hot* as its antonym in the sense of 'emotionally distant'. Thus, we can only say that *hot* and *cold* are antonymous in their temperature-related senses. In the ancient language WordNets, antonymy is considered primarily a semantic relation and only secondarily a relation between lemmas.

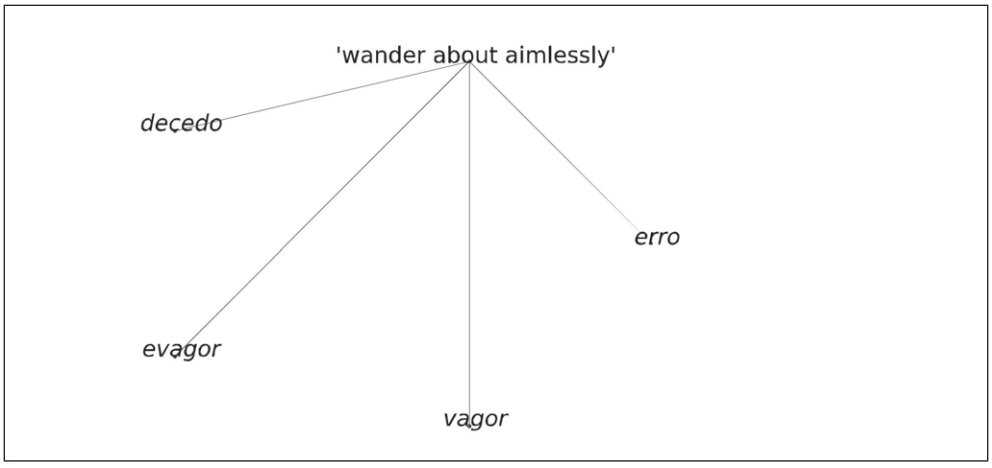
4. A 9,000-lemma WordNet for Latin was created in 2008 by Stefano Minozzi (Minozzi 2009). Minozzi's data has been incorporated into the University of Exeter's *Latin WordNet* and expanded to include more than 70,000 lemmas. Marco C. Passarotti's *Linking Latin* (see also Passarotti and Mambrini in this volume) project has also contributed greatly to the refinement of Minozzi's data by revising by hand-checking over 6,500 synset assignations. Efforts to develop an Ancient Greek WordNet were launched in 2012 by Eleonora Sausa (cf. Sausa 2012). In the same period, Bizzoni et al. (2014)'s work centered on annotating the lexicon of Homer with synset data; this has now been integrated into the vastly more comprehensive *Greek WordNet*, which attempts to model the semantics of ancient Greek diachronically. The Sanskrit WordNet has been partially derived from the semantic annotations found in the Digital Corpus of Sanskrit.

5. See esp. Buccheri et al. (forthcoming) on the improvements that have been made to the WordNet frame for capturing figurative meaning structures both at the level of lexical semantics and across the conceptual system.

'completely' + *sed-* 'sit'?) is something closer to 'sit completely (in a place)' (in other words, its conventional meaning is in fact metaphorically derived from its physical etymological meaning). Another innovation has been to vastly enrich the repertoire, and usage, of semantic field (or "semfield") designations. Semfields are large sets of related synsets and can be conceived as conceptual domains covering very broad fields of semantically related terms. In the original MultiWordNet, synsets could be assigned to any of a set of 126 categories of this kind (e.g. "economy", "military", "architecture"). To deliver the necessarily granularity of conceptual domains, the ancient language WordNets instead use the Dewey Decimal Classification System as a topic index. Although this system of topic designations includes many domains that are largely irrelevant to ancient texts (like "History of South America" or "Extraterrestrial Worlds"), it provides suitably discrete category distinctions in areas like "Greek and Roman religion", "Latin literature", "Philosophy", and "Geography of the Ancient World". We have also taken into account the sorts of lexical constructions that tend to typify morphologically complex languages like Greek and Latin, by adding new relations such as parasynthesis, composition, and inclusion (see again Biagetti et al. 2021: 4-5).

Consider the semantic network of Latin *erro* as encoded in the Latin WordNet.[6] In Latin, the word *erro* possesses the literal meaning 'to wander about aimlessly'. In this sense, the word is closely synonymous with *vagor* and *evagor*, as well as (somewhat more remotely) *decedo*. Thus, an approximation of the semantic network in which *erro* is embedded can be represented as something like Figure 1:

Figure 1 Semantic network of *erro* via its literal 'wandering' sense



6. On the metaphor of "wandering" for mistakenness in Latin, see Short (2013).

where the links of literal signification between *erro*, *vagor, evagor* and *decedo* indicate that these words are synonymous *in this sense*. However, *erro* can also be used metaphorically in the sense of 'make a mistake', as in example (1):

(1) Ter. *Heaut.* 105.
erras, si id credis
'You're mistaken, if you believe that.'

In this sense, *erro* is not in fact synonymous with *vagor*, or indeed any of its other literal synonyms, which never in extant texts demonstrate anything like this 'mistaking' meaning. (Of course, *decedo, vagor,* and *evagor* possess their own unique figurative profiles; for instance, *decedo* can mean 'to die', a sense it does not share with any of the 'wandering' terms). Still, numerous semantic linkages can be traced between these words and others. For instance, *erro* possesses a second metaphorical sense, 'to digress' (that is, to turn aside from the main subject or attention or course of argument), which both *vagor* and *evagor* share, thus presenting partial figurative overlap with *erro*. On the other hand, *vagor* has at least one metaphoric sense that *erro* does not: it can mean 'to vary' (i.e. to be subject to change) and through this sense could be linked to other nodes within the semantic network (represented by, for example, the literal sense of *dubito* or the metaphorical sense of *fluctuo*).

Erro's 'mistaking' sense, meanwhile, leads to another area of the semantic network entirely. As a general category, *erro* can refer to 'mistakes' of all different kinds, including intellectual uncertainty or misapprehension, moral faults, deception, and even fear or madness. Frequently it also refers to mistakes of language – in other words, it has the meaning of English *misspeak* – as in (2):

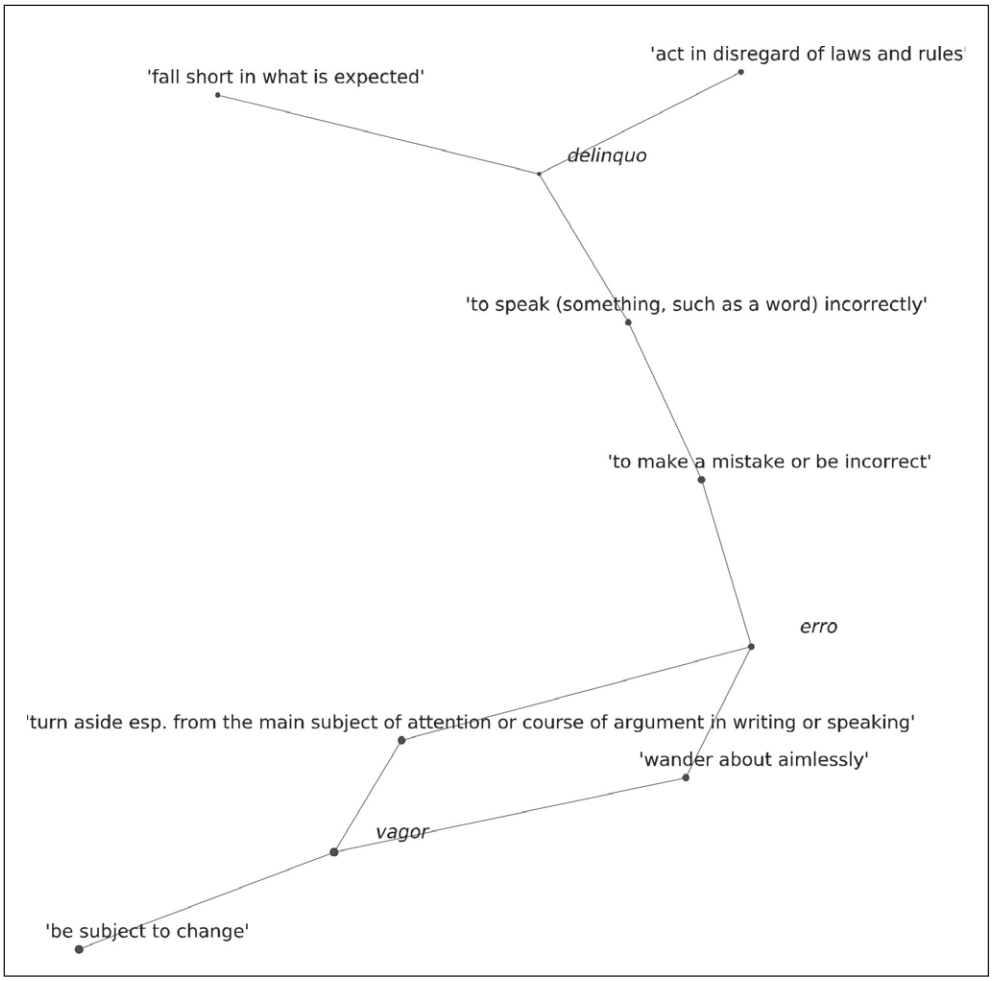
(2) Quint. *IO.* 7.3.17
nam est etiam periculosum, cum, si uno verbo sit erratum, tota causa cecidisse videamur
'For it is also dangerous when we seem to have lost the whole case, if a single word has been misspoken.'

In this more specific figurative sense, *erro* is synonymous with the verb *delinquo*: cf., e.g. (3):

(3) Quint. *IO.* 1.5.49
sunt quaedam cognata . . . qui alia specie quam oportet utetur, non minus quam ipso genere permutato deliquerit
'There are some nouns which are cognate . . . and he who uses the wrong *species* in connection with one of these will be guilty of the same offence as if he were to change the *genus*.'

Though hardly anyone would consider *delinquo* a proper (lexemic) synonym of *erro* because of their very different literal meanings, *erro* and *delinquo* can nevertheless be linked together in the semantic network through a relationship of superordination between their respective metaphorical senses. This graph fragment can be illustrated as in Figure 2, where the (shared or individual) senses of *erro, vagor* and *delinquo* have been given as glossed by WordNet synsets: in other words, as represented practically in the Latin WordNet (synset identifiers have not been given for ease of reading; however, in the WordNet each of these glosses is paired with a unique alphanumeric identifier which properly constitutes the synset. Furthermore, although portrayed here as a graph network, the WordNet itself is representation-independent and is presently implemented as a relational database).

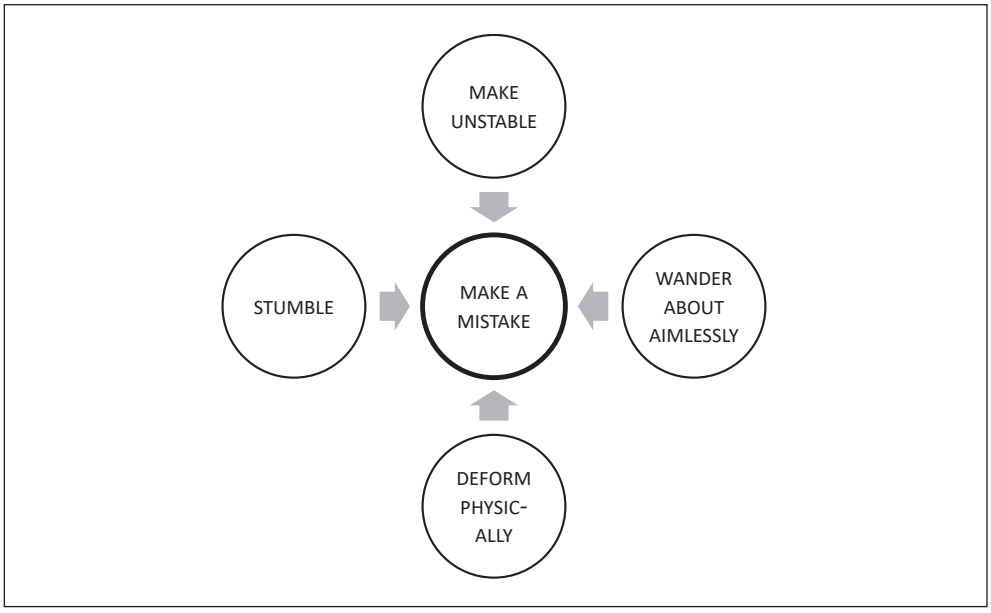
Figure 2 Partial network of literal and figurative senses of *erro, vagor* and *delinquo*



Importantly, the ancient language WordNets also include information on metonymic and metaphorical relations that operate at a supralexical level – that is, outside of, and at a higher conceptual order than, the semantic structure of any particular word – so to be able to consider the network via relations between senses themselves. In other words, it is possible to isolate portions of the conceptual system without taking into account the lexical instantiation of figures (or, by the same token, to consider the semantic system of Sanskrit, Greek, and Latin in terms of figurative relations between word senses). The same graph fragment illustrated above for *erro*, *vagor*, and *delinquo* could be described only in respect to the senses that form a network of literal and figurative meaning relations. Thus, 'wander about aimlessly' can be connected through metaphorical links to 'be subject to change' (in the semantic structure of *vagor*) and to 'make a mistake' (in *erro*'s), which in turn would be connected by a metonymical link to 'speak incorrectly' (in *delinquo*'s). Further dimensions of the figurative network can then be traced. For example, the network of domains other than 'wandering' that metaphorically structure 'making a mistake' in Latin is represented as in Figure 3.

For each of the concepts used metaphorically of 'making a mistake' – stumbling, deforming, making unstable and so on – the system is designed to contain information about corresponding linguistic expressions that reflect the mappings, including cases where an expression might be considered to instantiate two or more metaphors simultaneously. For instance, the metaphor 'stumbling is making a mistake'

Figure 3 Partial network of domains that metaphorically structure the concept of 'making a mistake' in Latin



is instantiated in the meaning structure of *fallor* (from **fal-*, lit. 'trip, cause to fall'), *offendo* (lit. 'dash against something'), *labor* (lit. 'slip; fall'), and *pecco* (denominative from **ped-k-*, 'falling, (a) fall'), which are synonymous with *erro* in its figurative 'mistaking' sense, while *perperam* (literally, 'oblique'), *perversus* (lit., 'overturned'), and *depravatus* (< *pravus*, lit., 'crooked'), *and deformis* ('misshapen') and *mendosus* (< *menda*, 'a blemish (of the face)') reflect the 'making unstable' and 'deforming' metaphors. Equally, the system is capable of accommodating information about connections to domains other than 'mistaking' that are metaphorically structured by each of these concepts. Thus, the ancient language WordNets have the capacity to capture what in cognitive linguistics is called the "range" of the target and the "scope" of the source, or the sets of domains for which any particular concept serves as either a metaphorical source or a metaphorical target. This representation of the network – considering the figurative relations that underpin Latin or Greek or Sanskrit expression in isolation from any specific linguistic instantiation – amounts, to my lights, to a view of the conceptual system of this language that goes far beyond what can be reconstructed from a conventional dictionary.

3. *Towards a "sembank" of the classical languages*

Building a large-scale corpus of semantically annotated Latin texts – in other words, a "sembank" for Latin, whose texts encode representations of the *meanings* of words or larger textual units (compound forms, fixed phrases, idioms, and so on) utilizing WordNet constructs – constitutes a critical and potentially very powerful branch of the ancient-language WordNet endeavor. This presents a significant challenge but would yield real dividends in terms of analytic possibilities. In my own line of research, at least, to be able to process texts computationally and at scale on the basis of the meanings of words (as well of the kinds of grammatical constructions in which they appear) would in fact make comparative analysis of figurative expressions (and thus the reconstruction of cultural patterns of thought) far easier, by reducing the number of queries required to cover the possible lexical make-up of metaphorical expressions. Consider, for instance, the common metaphor in Latin that construes WAR in terms of FIRE, as in (4):

(4) Tac. *Hist.* 2.86
flagrabat ingens bellum
'A huge war was (literally) burning.'

With available corpora, and assuming a source-domain vocabulary of *adolere, (ad) uro, aestuare, ardere, fervere, flagrare, incendere,* and *torrere* and a target-domain vocabulary of *bellum*, *certatus, certamen, colluctatio, concertatio, conflictus, congres-*

sio, congressus, dimicatio, proelium, pugna and *Mars*, more than 255 discrete queries would need to be performed to even begin to cover this figurative structure. With a semantically annotated corpus, on the other hand, to query any portion or even all of Latin literature for occurrences of collocations of these terms would be trivial.[7]

Sembanking is like treebanking, but for semantic rather than syntactic data. In a treebank, texts are annotated via a consistent set of tags, which are meant to describe their syntactic properties and relations under a certain theory of grammar (see, e.g. Abeillé 2012). In a sembank, texts are annotated instead according to their semantic properties, at one or more levels of description, using a tagging structure that is guided by some theory of meaning (cf. Baker, Fellbaum and Passonneau 2017). Typically, semantic annotation is viewed as an adjunct to syntactic annotation and involves the encoding of semantic roles in a corpus rather than of contextual sense definitions.[8] In a WordNet-based sembank, semantic annotations are provided for all appropriate tokens in a corpus (excluding stop words and members of closed class parts of speech apart from prepositions), and these annotations will correspond to synsets, semfields, and so on, drawing on any of the constructs and structures delivered by the WordNet architecture. But creators of language corpora have not yet widely integrated semantic annotations into text mark-up. More challenging, in my view, is the fact that most current encoding practices tend to adhere – explicitly or implicitly – to the principle of "one token, one tag," in the sense that generally tokens are not (and cannot be) tagged with more than one annotation of a given type.[9] In this sense, encoding schemas and practices are determinative (and intentionally so): they require tokens to be annotated and require annotators to make selections, in order to supply *the* correct reading for a given token. However, the philological and literary-critical tradition recognizes that texts are normally open to multiple readings, arising from divergent transmission histories, genuine interpretive differences due to lexical polysemy, intentional ambiguities, imaginative expression (including punning), and so on. Semiosis, we know, is open, multi-level and multi-dimension – if not actually "unlimited!"[10]

In many cases, a univocal tagging schema that adheres to the principle of "one tag per token" would not actually be problematic. Consider the opening phrase of the preface to Cato's *De agricultura* (5):

---

7. *Cylleneus* is designed to do exactly this: given a semantically annotated corpus, the engine can find occurrences of collocations of concepts as well as of lexemes. The engine is also capable of working with unannotated corpora, however in this case results are often very fuzzy.

8. Cf. Palmer, Kingsbury and Gildea (2005); Baker, Fillmore and Lowe (1998); Prasad et al. (2005). However, increasingly corpora have begun to include token-level word-sense annotations of different kinds, via WordNet constructs: cf. Miller et al. (1994 and 1993); Boschetti (2019).

9. This deterministic principle is actually seen as a foundational element of computational linguistics: cf., e.g. Manning and Schutze (1999: 139-145).

10. The term is from Peirce (1931-1966: vol. 1, p. 339).

(5) Cat. *Agr.* pr. 1
est interdum praestare mercaturis rem quarere
'It is true that to obtain money by trade is sometimes more profitable.'

Using any conventional text mark-up schema, annotation of this passage does not present particular difficulties. The text is relatively certain, and the meaning of the words are hardly in dispute. It is possible to conceive an XML schema, for instance, that could capture the sense(s) of the words as they occur here, which can be easily and straightforwardly represented in terms of WordNet synsets. The mark-up would consist of a sequence of tokens, with corresponding tags that represent the meaning of each word in terms of a unique synset (the gloss is optional, as this and other information could be retrieved programmatically via the WordNet API using the synset ID alone). This could be mocked as:

```
<sentence n="1" cite="pr.1" urn="cts:latinLit:phi0022.phi001">
   <token n="1" form="est" synset="v#01775163" gloss="have an
existence, be extant" />
   <token n="2" form="interdum" synset="r#00020741" gloss="on
certain occasions'" />
   <token n="3" form="praestare" synset="v#01246259"
gloss="value more highly'" />
   <token n="4" form="mercaturis" synset="n#00707408"
gloss="the commercial exchange of goods and services" />
   <token n="5" form="rem" synset="n#09639711" gloss="the
most common medium of exchange" />
   <token n="6" form="quaerere" synset="v#01513874"
gloss="come into possession of" />
</sentence>
```

Using a tagging schema of this kind, it would also be possible to annotate meanings at a higher order of structure: for instance, *rem quaerere* is a conventionalized idiom and as a sense-bearing unit could be independently tagged with synset v#01564908, "earn on some commercial or business transaction", distinct from the senses of its constituent elements. This would only require an additional mark-up element (like `<phrase>`) capable of incorporating one or more tokens and likewise taggable with a synset idenficiation number, as in, e.g.:

```
  <phrase synset="v#01564908" gloss="earn on some commercial or
business transaction">
    <token n="5" form="rem" synset="n#09639711" gloss="the most
common medium of exchange" />
    <token n="6" form="quaerere" synset="v#01513874"
gloss="come into possession of" />
  </phrase>
```

Something similar can be said for the *variae lectiones* that characterize the manuscript traditions of some ancient texts. A schema for semantic annotation must be able to take account of such variations, which may have consequences for meaning. Take the phrase in Vegetius's *De re military* (6):

(6) Veg. *De re mil.* 1.4
non enim tantum celerius, sed etiam perfectius imbuuntur, quae discuntur a pueris
'What (things) are learned by young children, not only are more quickly but also more completely absorbed.'

The *editio princeps* preserves the reading of several MS (the Palatine *ε* and *β*), namely, *imbuuntur*, 'are absorbed'.[11] At the same time, however, an alternate line of transmission preserves *imbibuntur*. In this case, the difference in contextual meaning between *imbibuntur* 'are drunken in' and *imbuuntur* 'are soaked in, imbrued' is not large and could in fact be annotated by the same metaphorical sense, v#00403772, "acquire or gain knowledge or skills". In the following hypothetical mark-up, this compatibility between the metaphorical senses of these words in context is captured by the token-level synset attribution, whereas discrete `<reading>` tags preserve the literal senses of the different manuscript readings (the "figure" tag indicates that the reading is metaphorical; different values could be used for other figurative significations, like metonymy).

```
  <token n="8" synset="v#00403772" gloss="acquire or gain
knowledge or skills" figure="#">
    <reading n="1" form="imbuuntur" codices="εβ"
synset="v#00318860" gloss="fill or imbue totally" />
    <reading n="2" form="imbibuntur" synset="v#00799582"
gloss="drink to the last drop" />
  </token>
```

More interesting, perhaps, is a case like (7):

(7) Col. *De re rust.* 1.pr.7
(prodigio simile est ut . . .) idque sperneretur genus amplificandi relinquendique patrimonii.
'(It is amazing that . . .) this method of increasing and passing on an inheritance should be despised.'

11. On the history of the text of Vegetius, see Allmand (2011).

While one branch of transmission preserves *relinquendi* ('passing on'), another branch preserves the equally plausible – but semantically entirely different – *retinendi* ('holding on to') – entailing a fairly radical change in meaning and requiring discrete semantic annotations. In the mock annotation below, this is achieved through distinct `<reading>` annotations that capture the fact that the two possibilities do not share, at the level of contextual interpretation, any overlapping sense: they are independent alternatives.

```
<token n="6">
   <reading n="1" form="relinquendi" synset="v#01525812"
gloss="give to or transfer possession of" />
   <reading n="2" form="retinendi" synset="v#01828422"
gloss="keep or maintain in unaltered condition; cause to
remain or last" />
</token>
```

Alongside the *variae lectiones* characterizing the transmission histories of many ancient texts, there are also interpretive differences that arise at the point of reception. I mean those many cases where, completely apart from any divergences in the actual state of the text, modern scholarship recognizes the possibility of different readings. Thus, for example, Catullus's expression *pudicitiam matris indicet ore* in (8):

(8) Cat. *Carm.* 61.217-218
et facile insciis noscitetur ab omnibus / et pudicitiam suae matris indicet ore
'He will be easily recognized by all, and will declare the fair fame of his mother by his *os*.'

has famously been the site of scholarly controversy, as *os* can be interpreted either as 'face' (i.e. appearance) or, more specifically, as 'mouth' and then by metonymical extension as 'speech'. The latter hypothesis acknowledges that in Roman culture it is specifically a person's manner of speech that functions as a mechanism of identity (in certain contexts more than physical appearance); and that a person's "way of speaking" was in fact transmitted as a kind of genetic inheritance, through the paternal bloodline (Bettini 1999: 189-98). We will never know what Catullus "really" intended the meaning of this phrase to be (and even if we could know, reader-response theory tells us that we should not in any case privilege "authorial meaning!"), and an adequate annotation scheme would therefore need to be able to represent the two readings of the meaning of *os* simultaneously, as – perhaps – synsets n#03683012, "outward or visible aspect of a person or thing" and n#05319899, "communication by word of mouth" e.g.:

```
<token n="6" form="ore" lemma="os" morpho="n-s---nb3-">
   <reading n="1" ref="Harrison 1996" synset="n#03683012"
gloss="outward or visible aspect of a person or thing"
figure="-" />
   <reading n="2" ref="Bettini 1999" synset="n#05319899"
gloss="communication by word of mouth" figure="~" />
</token>
```

Of course, such alternative readings are mutually exclusive. Either we read the text of Columella as *retinendi* or we read it as *relinquendi*: although the text is uncertain, there may be principled reasons for selecting one or the other *lectio* and treating this as the "correct" text. Likewise, we may choose to read the sense of *os* in Catullus either as 'face, appearance' or as 'mouth' (metonymically: 'speech') for principled hermeneutic reasons. These differences could thus be easily encoded as discrete documents, using, for instance, some form of "stand-off annotation" (Celano 2019). In a stand-off annotation system, different and different kinds of annotations are "added" to a main text in separate documents, which are ultimately linked to the so-called "base text," which is meant to be unchangeable. But this may not be satisfactory in many cases, because, as Philip Lieberman has remarked, "Language is inherently ambiguous and uncertain. That is the problem and the power of the system" (Lieberman 1984: 82). Indeed, ambiguity (or perhaps better, polyvalency) has in fact emerged as a central concern of the critical study of language and literature (cf. e.g. Empson 1930).

Take puns, which abound in Latin literature. A pun functions by simultaneously activating two independent meanings: indeed, the efficacy of a pun rests precisely in its audience's ability to simultaneously consider the multiple senses of a word or phrase. Consider the Latin word, *ius*, for instance – which can mean both 'law' and 'soup'.[12] Varro pokes fun at the Roman aristocracy's mania for fish keeping by joking that (9):

(9) Var. *RR*. 3.17.4
hos piscis nemo cocus in ius vocare audet
'No cook dares summon these fish to *ius*.'

where the pun suggests that fish are almost treated as having legal rights, so that making a soup of them (*ius*) would be like hauling them into court (also *ius*)! Plautus, too, takes a shot at the lawyering profession by punning on the double meaning of this word, which he pairs with the equally polysemous *coctus*: as *coctus* has the

12. For this and other examples, see Fontaine, McNamara and Short, "Introduction" (2018).

sense of 'learned, knowledgeable' in addition to its alimentary meaning (*iuris coctiores, Poen.* 586), the joke is that lawyers are 'rather knowledgeable of the law' just as lowly cooks are of soup. These are not alternative readings: they are "readings" that are meant to be simultaneously activated by the reader, and whose simultaneity in fact determines the joking effect of the text (even if we are not entirely able to find these jokes funny). An adequate annotation schema would need to be able to represent both readings equally: e.g.:

```
<token n="6" form="ius" lemma="ius" morpho="n-s---na3-">
   <reading n="1" synset="n#05638174" gloss="liquid food
especially of meat or fish or vegetable stock often containing
pieces of solid food" />
   <reading n="2" synset="n#02511574" gloss="a room in which
a law court sits" />
</token>
```

A more complex case might combine semantic with syntactic differences. A good example is given by philosopher Alanus de Insulis's reflection on the vanity of life, *Omnis mundi creatura:*

> *Omnis mundi creatura*
> *quasi liber et pictura*
> *nobis est in speculum.*
> *nostrae vitae, nostrae mortis,*
> *nostri status, nostrae sortis*
> *fidele signaculum.*
>
> 'Every creature of the world' *or* 'The creation of the whole world
> like a book, or painting,
> is as a mirror for us:
> of our life, our death,
> of our condition, of our fate.
> a constant reminder.'

Interpretation of this early twelfth-century poem revolves around exactly a morphological ambiguity, which interacts with a lexical ambiguity, in the opening stanza. The crucial ambiguity lies in the first line, where the phrase *omnis mundi creatura* can be read according to two distinct grammatical configurations, and thus with two distinct meanings, that depend on the choice between taking the verbal noun *creatura* in the concrete sense of 'creature' or the abstract sense of 'creation', along with the morphological ambiguity of the form *omnis*, which can agree either with it or with *mundi*. At the two extremes, the line can thus mean either 'every creature in

the world' or 'the creation of the whole world'. Which is the "correct" reading that should be encoded via annotation? Arguably, both are correct. In fact, the ambiguity is precisely the point, since the wavering between the two possible readings – 'every creature in the world' or 'the creation of the entire world' – captures an important theme of the poem as a whole. If the poem is meant to emphasize the brevity and (ultimately) insignificance of human life, as the extended simile of the blooming and then withering rose in the following stanzas shows, then the dual perspective offered by the opening line's twofold interpretation could hardly be more fitting. The character of our existence on this planet is reflected both in the particularizing perspective ("every creature") and in the globalizing perspective ("all of creation") of God's work. It is not the case, then, that we must choose between one or the other interpretation as what the poet "really" meant. The meaning of the poem – that in respect of all other life on earth, as well as in view of the immensity of Creation, our births and deaths mean but little – rests in the simultaneous availability of the opening line's two readings. Through the ambiguity, Alain intends for the reader to understand that human life is vain and fleeting, however one looks at it. The ambiguous readings operate together, rather than against one another, to create this meaning. To choose one or the other reading – as a simplifying "one tag per token" annotation scheme inevitably forces us to do – would thus effectively impoverish, if not actually destroy, the meaning of the poem.

An adequate annotation scheme would thus again need to take into account all this syntactic and semantic information, and deliver a structure capable of accommodating different readings at the token level: for instance, by using a construct like `<semtagm>` to represent a syntactic unit with its own discrete semantic properties: thus, e.g.:

```
<semtagm author="Alanus de Insulis" title="Omnis mundi
creatura" cite="1.1">
<reading n="1">
<token n="1" form="omnis" lemma="omnis" morpho="aps---fn3i"
synset="a#02160157" gloss="each and all of the members of a
group considered singly and without exception" />
<token n="2" form="mundi" lemma="mundus" morpho="n-s---mg2-"
synset="n#06691078" gloss="everything that exists anywhere" />
<token n="3" form="creatura" lemma="creatura" morpho="n-
s---fn1-" synset="n#00008019" gloss="a living organism
characterized by voluntary movement" />
</reading>
<reading n="2">
<token n="1" form="omnis" lemma="omnis" morpho="aps---mg3i"
synset="a#00482100" gloss="constituting the undiminished
entirety" />
<token n="2" form="mundi" lemma="mundus" morpho="n-s---mg2-"
```

```
synset="n#06691078" gloss="everything that exists anywhere" />
<token n="3" form="creatura" lemma="creatura" morpho="n-s--
-fn1-" synset="n#00154105" gloss="God's act of bringing the
universe into existence" />
</reading>
</semtagm>
```

Consider one final example, from a light-hearted epigram of John Owen found in Nicolas Mercier's anthology (under the section *de argutia mixta* 'on mixed wordplay'). The epigram in question is entitled *in Amorem nudum* – which immediately frames interpretation of the poem in terms of lexical ambiguities, as *amor* can of course refer to the mythological Cupid or to love and indeed to sexual acts:

quae villis natura feras et gramine campos
ornat, aues pluma, vellere vestit oues,
denique frigidulo quodcunque sub aëre nasci
contigit, innata veste vel arte tegit;
vestiuit nudum cur omnia praeter amorem?
quo nudus magis est, hoc minus alget amor.

'What nature adorns beasts with pelts and fields with grass,
clothes birds in feather and sheep in fleece,
whatsoever happens to be born under the cold sky
covers by in-born garb or art.
Why did nature clothe everything but Love?
The more naked Love is, the less it languishes.'

The poem contains numerous poetic flourishes, but especially, in its outward form, the insistent phonic repetition of *-v-* in the initial lines (*villis* . . . *aves* . . . *vellere vestit* . . . *veste vel* . . . *vestiuit*), which interplays with *-s-* and *-t-* in pentameter lines before giving way to *-n-* and *-m-* in the closing couplet (*nudum* . . . *omnia* . . . *amorem* . . . *nudus magis* . . . *minus* . . . *amor*). This creates an almost jingly quality around forms of the root *vest-* and forms of *nudus* that emphasizes the key thematic contrast of the poem: clothing and nudity. In fact, the sound play helps drive the poem to its humorous punchline: that nature has given every creature in the world some form of clothing ('by nature' or 'by art') while leaving Cupid—Love—Sex naked. Why? Because, with scarcely concealed sexual overtones, Cupid is (paradoxically) "less cold" when "more naked". Thus, interpretation of the poem turns generally on the duality of meaning in *amor*, but also on the specific polyvalency of the verb form *alget* in the last line. On the one hand, it can be read literally and concretely: Cupid is "less cold" when "more naked". On the other, it can be read abstractly and figuratively, in terms of the very fre-

quent LOVE IS FIRE metaphor, in which the intensity of love (or sex) is mapped to an intensity of heat, and conversely, an absence of passion is mapped to the coldness: thus, something like Love "is less unimpassioned" or "less diminished" when there is more nakedness.

This is a case, very similar to the previous example in Alain de Lille's poem, where a word's polyvalency – its having multiple meanings – actually determines alternative (and, I believe, intentionally simultaneous) readings of the text, in conjunction with other lexical elements. In other words, the "ambiguity" of *alget* does not operate in isolation; its meaning interacts – and co-varies – with the meanings of other polyvalent words (and with other features) to afford different opportunities of interpretation, in this case one more literal and one more figurative. The difference, of course, is that in the case of *Omnis mundi creatura* the semantic reinterpretation was in some sense prompted by the syntactic ambiguity: whether we read *creatura* as 'creature' or as 'creation' depends on how *omnis* was analyzed. In this case, by contrast, there is no syntactic ambiguity to signal the difference in interpretation. The multiplicity of meaning is, as it were, baked in. But at the same time, again, it seems evident that the two readings are meant to be activated together. There is not a choice to take *either* the literal (physical, bodily) sense of *Amor minus alget* – 'Cupid gets less cold' – *or* the more figurative (erotic) sense of this expression – 'sex is less unimpassioned', but the two work hand in hand. The whole effect, and joke, of the poem depends on it. Of course, as there is no single synset in the Latin WordNet (or in any WordNet or in any possible system of semantic description), any encoding schema must be able to accommodate multiple parallel annotations, as in:

```
<semtagm author="Ioannes Ovvenus" title="Epigrammata"
subtitle="In Amorem nudum" cite="2.88.6">
<reading n="1a">
   <token n="7" form="alget" lemma="algeo"
morpho="v1spia--2-" synset="v#00054128" gloss="be cold" />
   <token n="8" form="amor" lemma="amor" morpho="n-s---mn3-"
synset="n#06906245" gloss="god of love; son of Aphrodite;
identified with Roman Cupid" />
</reading>
<reading n="1b">
         <token n="7" form="alget" lemma="algeo"
morpho="v1spia--2-" synset="v#00167689" gloss="make less
active or intense" />
         <token n="8" form="amor" lemma="amor" morpho="n-s--
-mn3-" synset="n#05567842" gloss="the arousal of feelings of
sexual desire" />
</reading>
</semtagm>
```

4. *Concluding remarks*

I have dwelt on the range of possible types of polyvalency in Latin literature – instances where words have multiple readings, arising either through genuine ambiguity, differences in interpretation, divergences in reception, and so on – in order to indicate some of the challenges that are presented in the creation of semantically annotated corpus (a "sembank"), if the desire is to represent as accurately as possible the sense of words in context. Some cases (like Cato's *De agricultura*) will be easily captured through simple one-to-one attributions of meaning – represented here in terms of discrete WordNet synsets –, others will require annotators to introduce multiple different tags to represent alternative readings (like Catullus's poem), and still others will necessitate an encoding schema capable of capturing meanings that co-exist simultaneously in the text – puns, creative usages of syntactic ambiguity, innuendos (like in Alain de Lille or John Owen). While WordNet constructs (especially synsets) deliver a robust as well as flexible system for representing meanings of the lexicon in machine-readable and machine-actionable form, an equally robust and flexible system is needed for characterizing meanings as they emerge through imaginative literary expression (or any type of discourse) – and not only at the level of lexical sense attribution.

Websites

Ancient Greek WordNet: https://greekwordnet.chs.harvard.edu
Cylleneus: https://github.com/cylleneus/cylleneus
Digital Corpus of Sanskrit (DCS): http://www.sanskrit-linguistics.org/dcs/
Latin WordNet: https://latinwordnet.exeter.ac.uk
Sanskrit WordNet: https://sanskritwordnet.unipv.it

References

Abeillé, Anne. ed. 2012. *Treebanks. Building and using parsed corpora*. Dordrecht: Kluwer.

Allmand, Christopher. 2011. *The De Re Militari of Vegetius*. Cambridge: Cambridge University Press.

Baker, Collin, Fillmore, Charles & Lowe, John. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Collin Baker, Charles Fillmore & John Lowe, 86–90. Montreal: Université de Montréal.

Baker, Collin, Fellbaum, Christiane & Passonneau, Rebecca. 2017. Semantic Annotation of MASC. In *Handbook of Linguistic Annotation,* Nancy Ide & James Pustejovsky (eds), 699–717. Dordrecht: Springer.

Bettini, Maurizio. 1999. *The Portrait of the Lover*. Trans. by Laura Gibbs. Berkeley: University of California Press.

Biagetti, Erica, Zanchi, Chiara & Short, William M. 2021. Toward the creation of WordNets for ancient Indo-European languages. In *Proceedings of the 11th Global Wordnet Conference*, Piek Vossen & Christiane Fellbaum, 258–266. Global Wordnet Association.

Bizzoni, Yuri, Boschetti, Federico, Diakoff, Harry, Del Gratta, Riccardo, Monachini, Monica & Crane, Gregory. 2014. The Making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis, 1140–1147. Reykjavik: European Languages Resources Association.

Boschetti, Federico. 2019. Semantic Analysis and Thematic Annotation. In *Digital Classical Philology*, Monica Berti (ed), 321–340. Berlin: Mouton de Gruyter.

Buccheri, Alessandro, De Felice, Irene, Fedriani, Chiara & Short William M. Forthcoming. Semantic analysis and frequency effects of conceptual metaphors of emotions in Latin. From a corpus-based approach to a dictionary of Latin metaphors. *Journal of Latin Linguistics*.

Celano, Giuseppe G.A. 2019. Standoff Annotation for the Ancient Greek and Latin Dependency Treebank. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, 149–153. Association for Computing Machinery.

Empson, William. 1930. *Seven Types of Ambiguity*. London: Chatto and Windus.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Fellbaum, Christiane. 2017. WordNet: An electronic lexical resource. In *The Oxford Handbook of Cognitive Science*, 301–314. Oxford: Oxford University Press..

Fontaine, Michael, McNamara, Charles & Short, William M. 2018. *Quasi Labor Intus: Ambiguity in Latin Literature.* The Paideia Institute.

Lieberman, Philip. 1984. *The Biology and Evolution of Language.* Cambridge, MA: Harvard University Press.

Manning, Chris & Schütze, Heinrich. 1999. *Foundations of Statistic Natural Language Processing.* Cambridge, MA: MIT Press.

Miller, George, Leacock, Claudia, Tengi, Randee & Bunker, Ross T. 1993. A Semantic Concordance. In *Proceedings of the 3 DARPA Workshop on Human Language Technology.*

Miller, George, Chodorow, Martin, Landes, Shari, Leacock, Claudia & Thomas, Robert. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*, 240–243.

Minozzi, Stefano. 2009. The Latin WordNet Project. In *Latin Linguistics Today*, Peter Anreiter and Manfred Kienpointner (eds), *Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, Innsbrucker Beitrage zur Sprachwissenschaft, Volume 137, 707–716.

Passarotti, Marco C. & Mambrini, Francesco. This volume. Linking Latin. Interoperable lexical resources in the *LiLa* project.

Palmer, Martha, Kingsbury, Paul & Gildea, Daniel. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1): 71–106.

Peirce, Charles. 1931-1966. *Collected Papers of Charles Sanders Peirce*, Charles Hartshorne, Paul Weiss & Arthur W. Burks. Cambridge, MA: Belknap.

Prasad, Rashmi, Joshi, Aravind, Dinesh, Nikhil, Lee, Alan, Miltsakaki, Eleni & Webber, Bonnie. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, Belz & Varges 25–32. Stroudsburg, PA: Association for Computational Linguistics.

Sausa, Eleonora. 2012. *Toward an Ancient Greek wordnet*. https://www.yumpu.com/en/document/view/6198686/toward-an-ancient-greek-wordnet-eleonora-sausa-

Short, William M. 2013. Getting to the Truth: Metaphors of Mistakenness in Greek and Latin. *Arion* 21(2): 111–40.

HoDeL: a Dependency Lexicon of Homeric Greek

Chiara Zanchi*

This paper presents the *Homeric Dependency Lexicon* (HoDeL), a new verbal lexicon of Homeric Greek with a very user-friendly interface facilitating the investigation of Homeric verbs, their dependents and other aspects of the Homeric syntax. The paper grounds HoDeL in the framework of linguistic resources available online for the study of Ancient Greek, introduces the notion of valency in Dependency Grammar as well as previous approaches to build valency lexica of both modern and ancient languages. Moreover, the paper discusses the architecture of the *Ancient Greek Dependency Treebank* (AGDT 2.0) and the theory of valency underlying it to lay the ground for explaining how HoDeL was built. HoDeL was induced from the analytical layer of AGDT 2.0, extracting all dependents tagged as SBJ, OBJ, PNOM, and OCOMP with a set of SQL queries. Extracted data were then stored in a relational database that interacts with users' interface. The paper then illustrates HoDeL incorporated constraints and functionalities and shows how they can be employed by perspective users to answer specific research questions about the Homeric syntax.

Keywords: HoDeL, valency lexica, Homeric Greek, AGDT 2.0, dependency treebank, syntax

1. *Introduction*[1]

The Ancient Greek (henceforth, AG) handed down by the *Iliad* and the *Odyssey*, the so-called Homeric poems, is the most ancient literary variety of AG that survived up to the present. The Homeric poems were probably recorded in writing in the 8th century BC but are acknowledged to preserve at least two centuries older layers of AG: they constitute an inherently diachronic corpus. Their language mostly represents an archaic Eastern Ionic, admixed with a dialectal pastiche of Mycenean and Aeolic traits, as well as with other linguistic features that can hardly be associated with any AG dialect (e.g., Horrocks 2010: 44). This admixture is due to the very nature of the poems: though their authorship is traditionally ascribed to Homer, the *Iliad*

* University of Pavia.

1. HoDeL was created at the Department of Humanities of the University of Pavia and funded by the Italian Ministry of Education and Research in the framework of the project *Transitivity and Argument Structure in Flux* (2015 PRIN call, grant no. 20159M7X5P), coordinated by Michela Cennamo and Silvia Luraghi. Chiara Zanchi and Paolo Ruffolo are the main responsible for its creation and worked under Silvia Luraghi's supervision. In various ways and at different times many people contributed to HoDeL, to whom I am thankful: Federico Boschetti, Giuseppe G. A. Celano, Giulia D'Agostino, Marco Forlano, Francesco Mambrini, Nahumi Nugrahaningsih, Marco Passarotti, Edoardo M. Ponti, Eleonora Sausa.

and the *Odyssey* have been demonstrated to be examples of oral poetry (Lord 1960; Parry 1971). Due to their antiquity, their chronologically multi-layered nature and their linguistic richness, the Homeric poems are crucial both for historical comparative research (cf., e.g., Watkins 1976) and for perspective diachrony of later Greek.

Accordingly, a number of online linguistic resources, most notably *The Chicago Homer* and the relevant sections of the *Perseus Digital Library* and of the *Ancient Greek Dependency Treebank* (AGDT 2.0), contain digitalized versions of the Homeric poems enriched with their English translations and various types of annotations. However, neither of these resources can easily be used to investigate Homeric syntax and specifically Homeric verbs, their dependents and constituent order. This paper presents a linguistic resource that has been built precisely to fill this gap: HoDeL (*The Homeric Dependency Lexicon*), an online linguistic resource greatly facilitating the investigation of Homeric syntax through an extremely user-friendly interface. HoDeL is currently hosted at *The Pavia linguistic resources repository* and can be freely queried online.

The paper is structured as follows. Section 2 provides the background to frame HoDeL: it introduces some among the existing linguistic resources of AG, the notion of valency and methodologies to build valency lexica. Section 3 details the architecture of AGDT 2.0 and the theory of valency underlying it and accounts for the construction of HoDeL. Section 4 illustrates the main queries and filters incorporated in HoDeL interface, while Section 5 exemplifies their usage with concrete examples of research questions that HoDeL helps addressing. Section 6 contains the conclusions.

2. *Background: Annotated corpora of Ancient Greek and valency lexica*

Several electronic corpora of AG texts are nowadays available online. In this paper, I name just a few (alphabetically ordered): AGDT 2.0 (*Ancient Greek Dependency Treebank*), *The Chicago Homer* (Early Greek epic), DĀMOS (Mycenaean; Aurora 2015), DFHG (*Digital Fragmenta Historicorum Graecorum*), *The Diorisis Ancient Greek Corpus* (Vatri and McGillivray 2018), EAGLE (*Electronic Archive of Greek and Latin Epigraphy*), the *Perseus Digital Library* (Bamman and Crane 2008), PROIEL (*Pragmatic Resources of Old Indo-European Languages*; Haug and Jøndal 2008),[2] SEMANTIA (313 papyrological texts from the Duke Databank of Documentary Papyri; Vierros 2018), TITUS (*Thesaurus Indogermanischer Text- und Sprachmaterialien*), and TLG (*Thesaurus Linguae Graecae*). In some of these corpora, the

2. Additional morphosyntactically annotated corpora have stemmed from the PROIEL project, specifically, ISWOC for old Romance and Germanic languages and TOROT for Old Slavic languages (Eckhoff and Berdičevskis 2015; Eckhoff et al. 2018; see also Eckhoff and Haug in this volume).

digitized version of the texts is enriched with mark-up and/or annotation. To lay the ground for HoDeL, the most relevant type of metadata is the multi-layered annotation that provides linguistic information at the levels of morphology and syntax: the morphosyntactically annotated corpora, i.e., treebanks (on which, see the papers by Eckhoff and Haug, Hellwig and Sellmer, Biagetti in this volume), can be used to induce other linguistic resources, such as valency lexica.

The term "valency" was borrowed from chemistry into linguistics by Lucien Tesnière (see Àgel and Fisher 2015 for precursors) in the framework of Dependency Grammar (Tesnière 1959). In this context, valency refers to the ability of verbs to combine with and to determine the form of a fixed number of participants called *actants*. In Tesnière (1959), *actants* are contrasted with *circumstants*, free modifiers of verbs. The English terms *complement* and *adjunct* are meant to replicate Tesnière's *actant-circumstant* distinction (Matthews 1981). Frequently, *actants/complements* are also called *arguments*, especially in the US linguistic tradition of the last quarter of the 20th century. However, the term *argument* holds an ambivalent status between semantics and syntax: it can indicate all inherent roles that occupy a place in the semantic relationality of a concept, such as *mouth* in the event of *eating*, or constituents headed by verbs, such as *an apple* with the verb *to eat* in the sentence *Chiara eats an apple*. This ambiguity is fundamental to account for a number of mismatches concerning the theory of valency underlying HoDeL (see Section 2.2). From Tesnière onward, linguistic theories have put valency at the core of their research agenda and, notably, have disagreed about its very nature: valency has been regarded as a concept fundamentally syntactic, semantic, or both. However, until the relatively recent developments of Construction Grammar (cf. Fried and Boas 2005 for an overview), there has been at least one point of substantial agreement: valency is a property of verbs. As such, valency information can be stored in dictionoaries and (verbal) lexica.

Valency-related information is contained in traditional dictionaries of AG. For example, the Liddell-Scott-Jones dictionary (LSJ), under the entry *manthánō* 'learn', provides its different meanings, along with morphological and semantic features of the participants associated with these meanings: *manthánō*+ACC 'learn something', *manthánō*+DAT 'understand somebody', *manthánō*+INF 'learn something', *manthánō+apó/ek/pros/pará*+GEN 'learn from something/ somebody'. Valency-dedicated lexicography began with Helbig and Schenkel's (1991[1969]) *Wörterbuch zur Valenz und Distribution deutscher Verben* (on German, see also Schumacher et al. 2004; on English, see Herbst et al. 2004; a contrastive valency lexicon of Dutch-French-English is Colleman et al. 2004). Manually collected valency lexica are now available for ancient Indo-European languages as well: Happ (1976) contains a non-exhaustive list of Latin verbs and their valency patterns as evidenced in a corpus of 800 sentenc-

es sampled from Cicero's *Orationes*. In her PhD dissertation, Frigione (2015) collected and published the materials for a verbal lexicon of Old Church Slavic.

Intuition-based valency lexica as those introduced above are necessarily partial in coverage and time-consuming to build. Therefore, scholars have increasingly exploited annotated corpora to automatically extract verbal valency patterns and their frequencies from texts (but see Passarotti and Mambrini in this volume). The first attempts in this direction were hybrid: for example, PropBank (Kingsbury and Palmer 2002) and FrameNet (Ruppenhofer et al. 2006) were initially built with an intuition-based method and later refined with corpus-driven data.

Primarily automatic approaches to valency lexica have been first carried out for modern languages (see VALEX for English, Korhonen et al. 2006; LexShem on French, Messiant et al. 2008). Nowadays, automatically derived valency lexica exist for ancient Indo-European languages as well. For example, Bamman and Crane (2008) and McGillivray (2013: 31–60) used the morphosyntactically annotated texts in the *Perseus Digital Library* and in the *Latin Dependency Treebank* (LTD; Bamman and Crane 2006) to induce valency lexica for Latin. Similarly, the syntactic subcategorization lexicon for Thomas Aquinas' Latin, IT-VaLex (McGillivray and Passarotti 2009), was derived from the *Index Thomisticus Treebank* (IT-TB; McGillivray and Passarotti 2015). McGillivray and Vatri (2015) used the same method documented in McGillivray and Passarotti (2009) and McGillivray (2013: 31–60) to induce a valency lexicon from AGDT 1.0 (to my knowledge, these lexica have never been made available online). Finally, a semantic valency lexicon of Latin, *Latin Vallex*, was derived from a semantically-annotated subset of LTD and IT-TB (Passarotti et al. 2016). Currently, the PROIEL-style treebanks, which can be consulted through the *Syntacticus* interface, also come with generated dictionaries that include a comprehensive list of valency frames extracted from treebanks (see Eckhoff and Haug in this volume).

HoDeL is also a lexicon of Homeric Greek verbs that has been built upon the IT-VaLex model. The treebank on which HoDeL is based and its construction are illustrated in the next section.

3. *Building HoDeL*

1.1. The Ancient Greek Dependency treebank and its theory of valency

The data of HoDeL are based on the Homeric poems treebanked at AGDT 2.0. The architecture of AGDT 2.0 is modelled on the *Prague Dependency Treebank* of Czech, the groundbreaking project in the field of Dependency Grammar (PDT 3.0; Hajič et al. 1999). Like other dependency treebanks (see Eckhoff and Haug, Hellwig and Sellmer, and Biagetti in this volume for details), PDT 3.0 and AGDT 2.0 share the following features: they are predicate-centred, contain the same number

of nodes as the number of words and allow for trees with more than two branches (on the differences between constituent vs. dependency treebanks and among different dependency treebanks, see Biagetti 2018: 9–37). PDT 3.0 and AGDT 2.0 are multi-layered dependency treebanks: metadata is structured and stored in separated but interlinked morphological, analytical and tectogrammatical layers. The analytical (i.e., syntactic) layer contains dependency syntactic trees and works as a basis for the tectogrammatical (i.e., semantic) layer, which stores semantic role-labelling, information structure, and anaphora/ellipsis resolution, annotated in the framework of the Praguian linguistic tradition of the *Functional Generative Description* (Sgall et al. 1986).

The first release of *Ancient Greek and Latin Dependency Treebank* is documented in Bamman and Crane (2011). Currently, the AG section of the treebank contains 557,922 tokens and the following works: *Iliad*, *Odyssey*, *Hymn to Demeter*, Aeschylus' tragedies, Sophocles' *Ajax*, *Antigone*, *Electra*, *Oedipus Tyrannus*, *Trachinae*, Hesiod's *Theogony*, *Works and Days*, *Shield of Heracles*, Plato's *Euthyphro*, Lysias' *On the Murder of Erathostenes*, *Against Alcibiades* 1 and 2, *Against Pancleon*, Plutarch's *Alcibiades* and *Lycurgus*, and sections of Aesop's *Fables*, Athenaeus' *Deipnosophists*, Diodorus Siculus' *Library,* Herodotus' *Histories*, Polybius's *Histories*, Pseudo Apollodorus's *Library*, and Thucydides' *Histories*. All works are annotated for the morphological and analytical layers, whereas the tectogrammatical annotation is available only for subsections of Aesop's *Fables* and Diodorus Siculus' *Library* (Celano 2019: 283–284, 288). Thus, no tectogrammatical annotation is available for the Homeric poems.

The theory of valency underlying AGDT 2.0 – and PDT before it –is the theory of valency of the *Functional Generative Description* (Panevová 1994). The *Functional Generative Description* regards valency as the ability of linguistic elements to open positions for other linguistic elements, called *complementations*. Complementations are grouped in two ways: (i) inner participants vs. free modifications; (ii) obligatory vs. optional complementations. Inner participants (e.g., actor, patient, addressee, origin) are verb-specific and can occur only once per verb, whereas free modifications (e.g., time, place, goal, instrument, etc.) are not verb-specific and can be added freely to the sentence. The distinction between inner participants and free modifications seems to be semantically based, while that between obligatory and optional complementations is grounded in the syntactic notion of obligatoriness. These two classifications do not overlap: there can exist *optional inner participants*, such as instrument with verbs of cutting, and *obligatory free modifications*, such as origin with motion verbs. In PDT, this valency theory is accounted for at the tectogrammatical layer, which also contains anaphora and ellipsis resolution, as mentioned above. Overall, then, the theory of valency of the *Functional Generative Description* seems to be based on semantic criteria rather than on syntactic ones. The only notion that

seems to distinguish syntactic arguments from adjuncts, which is mentioned in the PDT guidelines, is obligatoriness.

As there is no tectogrammatical layer in AGDT 2.0, its valency theory is discussed in the guidelines of the analytical layer. On the one hand, the guidelines state that AGDT 2.0 inherits its valency theory from PDT. On the other hand, however, the critical notion of obligatoriness is not dealt with at a sufficient level of granularity. Thus, this lack, as well as the lack of the tectogrammatical layer in AGDT 2.0, results in a number of mismatches in the argument vs. adjunct annotation with respect to the most widely accepted theories of syntactic valency (for details, Zanchi and Luraghi 2020 and Zanchi forthc.).

1.2. How has HoDeL been built?

HoDeL has been induced from the analytical layer of the *Iliad* and *Odyssey* treebanked at AGDT 2.0 (for details, see Zanchi and Luraghi 2020 and Zanchi forthc.). The guidelines of the analytical layer of AGDT rely on those of PDT with some *addenda* to enhance representativeness (Celano 2019: 285–286). From this layer, a series of SQL queries extracts all verbal forms and dependents labeled as SBJ (Subject), OCOMP (Object Complement), PNOM (Predicate Nominal) and OBJ (Object). The latter tag includes all verbal arguments except SBJ and arguments labeled as OCOMP and PNOM and hence comprises accusative, dative, genitive nouns or pronouns, prepositional phrases, infinitive verbs, accusative+infinitive constructions, and other types of subordinate clauses that can function as verbal objects (for details on these tags, see Celano 2019: 286–287 and the guidelines of the analytical layer of AGDT 2.0). All extracted dependents are considered part of verbal valency according to the guidelines of AGDT 2.0. Argumental dependents have been then recorded in a spreadsheet, from which a relational database has been built. The relational database in turn interacts with users' interface.[3]

In contrast, we did not extract dependents that the AGDT 2.0 guidelines do not consider belonging to the verbal valency: specifically, the tags ADV (adverbials providing the event with background information), ATR (NP modifiers) and ATV/ATVV (non-governed complements, i.e., predicative noun phrases/adjectives which may morphologically agree with their head noun, but qualify the whole event denoted by the verb).

3. The original query algorithm and its implementation were conceived to build IT-VaLex (McGillivray and Passarotti 2009). To induce HoDeL, the queries have been adapted to the AGDT 2.0 tagset. An earlier version of HoDeL, released in 2016 (Zanchi et al. 2018), was based on a previous version of the treebank (AGDT 1.0) and lacked transliteration and English translation. Moreover, for this release of HoDeL, we improved the quality of the base data of AGDT 2.0, as documented in Zanchi (forthc.).

The scrutiny of extracted data revealed a number of mismatches in the argument vs. adjunct distinction with respect to the most widely accepted theories of syntactic valency. For HoDeL users, this means keeping in mind that the valency theory of the *Functional Generative Description* lies behind both the PDT 3.0 and the AGDT 2.0 and that such theory allows for the categories of *optional inner participants* and *obligatory free modifications*, which resulted to be particularly problematic for annotation. The resulting issues are thoroughly discussed in Zanchi and Luraghi (2020) and Zanchi (forthc.), and are briefly summarized and exemplified in what follows:

a. agent participants of passive verbs

(1) *ēdè* *phílēthen* ***ek*** ***Diós***
and love.AOR.3PL.PASS out_of Z.GEN

'…and they were loved **by Zeus**.' (Il. 2.668-669)

b. dative instruments and beneficiaries

(2) *Autàr* ***Odussêos*** *talasíphronos* *oú* *pot'*
PTC O.GEN steadfast.GEN NEG ever
éphasken *zōoû* *oudè* *thanóntos* *epikhthoníōn*
say.IMPF.3SG alive.GEN NEG die.PTCP.AOR.GEN earthly_one.GEN.PL
teu *akoûsai*
INDF.GEN hear.AOR.INF

'Yet **concerning Odysseus** steadfast heart, whether living or dead, he said he had heard **from no man** on earth.' (Od.17.114-115)

(3) *hṓs* *té* *pou* *è* *autòs* *pareòn*
as PTC anywhere or self.NOM be_present.PTCP.PRS.NOM
è *állou* *akoúsas*
or other.GEN hear.PTCP.AOR.NOM

'As though you had been present yourself or had heard (it) **from someone else**.' (Od. 8.491)

c. origin participants

(4) *ek* *dé* *moi* *aukhèn* ***astragálōn***
out_of PTC 1SG.DAT neck.ACC vertebra.GEN.PL
eágē
break.AOR.3SG.PASS

'My neck (lit. 'the neck to me) was broken away from the vertebrae' (Od. 11.64-65)

Agents of passive verbs, such as *ek Diós* 'by Zeus' in (1), are annotated as OBJ, i.e., as part of the verbal valency. This happens because, at a semantic level, agent participants are *inner* participants of transitive verbs. However, passive voice is usually acknowledged to be a syntactic valency decreasing strategy, which removes agents from argument structure and makes them *optional* (e.g., Siewierska 2005). Thus, *ek Diós* 'by Zeus' is an *optional inner participant*. In (2) and (3), the same dative plural, *ophthalmoîsi* 'with (my) eyes', is inconsistently tagged as OBJ (as a syntactic argument in (2)) and as ADV (as an adjunct in (3)), in dependence of forms of the same verb, *eîdon*, the aorist suppletive form of *horáō* 'see'. Again, this inconsistency results from the fact that 'eyes' are *optional inner* participants in the event of seeing. As mentioned in Section 3.1., according to the *Functional Generative Description*, origin is classified as an *obligatory free modification*. Thus, it is not surprising that in (4), the origin participant, encoded by *ek*+GEN, is tagged as OBJ in dependence of a verb of breaking, *ágnumi* 'break'. This annotation, however, is problematic for multiple reasons: first, it regards origin as a syntactic argument of *ágnumi* 'break', which it is not; second, it treats the initial local particle *ek* (see, e.g., Zanchi 2019: 82–86 on this terminology) as a preposition governing the genitive *astragálōn*, which is not necessarily the case either.[4]

As outlined in Section 3.1., the analytical layer of AGDT 2.0 does not account for elliptical structures and does not include empty nodes for null arguments: in the dependency treebanks modelled on PDT, null arguments are integrated at the tectogrammatical layer. AG is a pro-drop language: by default it omits topical subjects, which are indexed on verbs through personal endings. Moreover, AG, as well as other ancient Indo-European languages, preferably or obligatorily selects null referential objects in certain syntactic and pragmatic contexts, including conjunct participles, coordinated verbs and clauses, and yes/no questions (Luraghi 2003; Haug 2012; Keydana and Luraghi 2012; Sausa and Zanchi 2015). Both null subjects and null referential objects occur frequently in the Homeric poems and, crucially, fill slots of verbal valency. The fact that they are not included in the syntactic trees of the analytical layer of AGDT 2.0 results in an incomplete account of the valency of a number of verbs.[5]

4. On the annotation of local particles in so-called 'tmesis' positions in AGDT 2.0, see Zanchi (forthc.). Other issues in the annotation that are due to peculiarities of the Homeric language and lemmatization, see Zanchi and Luraghi (2020) and Zanchi (forthc.).

5. The problematic issues regarding passive agents, null objects, and others were noted by the creators of the PROIEL family treebanks (see Eckhoff and Haug in this volume; Haug and Jøhndal 2008; Haug 2012; Eckhoff and Berdičevskis 2015; Eckhoff et al. 2018). Although based on a version of Dependency Grammar, the tagset of the PROIEL treebanks includes additional labels and relations to improve descriptiveness, such as the label AG for passive agents and specific relations for elliptic structures. For further details on the issue of integrating null participants, see Zanchi (forthc.).

Table 1 Greek characters-Beta code correspondences

Greek	α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ	μ	ν	ξ	ο	π	ρ	σ,ς	τ	υ	φ	χ	ψ	ω
Beta Code	a	b	g	d	e	z	h	q	i	k	l	m	n	c	o	p	r	s	t	u	f	x	y	w
Greek	ά	ὰ	ᾶ	ὰ	ὰ	ᾴ		ᾳ		ᾱ		ᾱ		ᾰ		ᾰ		ᾳ	ᾰ		ᾳ			
Beta Code	a/	a\	a=	a)	a(	a)/		a(/		a)\		a(\		a)=		a(=		a|	a)/|		a)=|			
An * should be used to type capital letters (e.g. Βρισηις → *brishi/s)																								

4. *How to do things with HoDeL: A practical guide*

After detailing the data on which HoDeL relies and the methods employed to build it, I offer in this section a practical guide for HoDeL users. The section is divided in subsections that respond to specific practical questions.

4.1. How to type Greek characters and to visualize transliterations?

In order to type Greek characters, HoDeL users should employ Beta Code, as in the *Perseus Project* and in TLG. The correspondences between Greek fonts and Beta Code are reported in Table 1. The least intuitive Greek-Beta Code correspondences are highlighted in grey. Example (5) shows how the first line of the *Odyssey* looks in Greek characters (5)a, Beta-Code (5)b and in transliteration (5)c.

(5) a. ἄνδρα μοι ἔννεπε μοῦσα πολύτροπον ὃς μάλα πολλὰ
b. a)/ndra moi e)/nnepe mou=sa polu/tropon o(\s ma/la polla\
c. ándra moi énnepe moûsa polútropon hòs mála pollà
'Tell me, Muse, about the wily man who (wandered) long and far' (Od. 1.1)

Users can choose to visualize either the Greek script or its transliteration by flagging 'greek' or 'trans' in the 'Display' box at the top of HoDeL homepage (Figure 1).

Figure 1 How to visualize transliterations

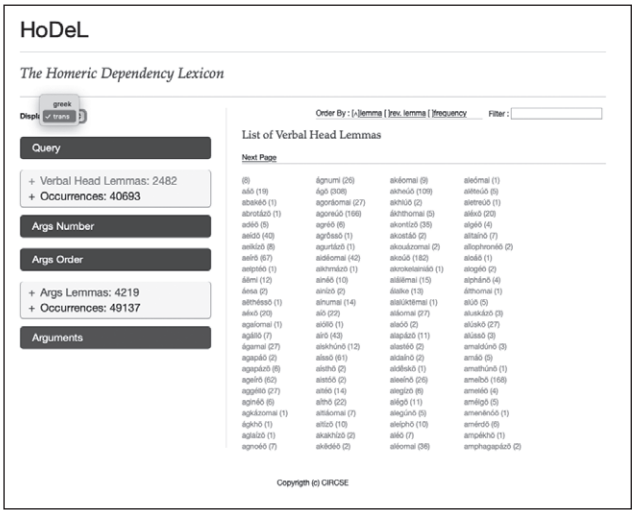
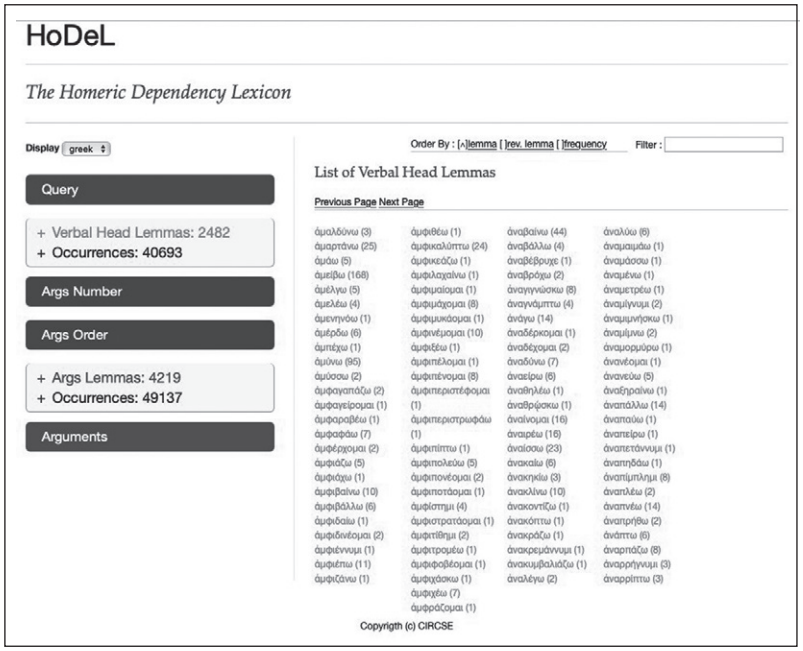


Figure 2 The homepage of HoDeL



4.2. What is shown in HoDeL homepage?

HoDeL homepage shows a list of Homeric verbs alphabetically ordered. After each lemma, its frequency is provided (Figure 2). If preferred, users can choose to visualize Homeric verbs by reverse alphabetical order or by increasing frequency, by flagging either the [^]**rev. lemma** or the [^]**frequency** box on top right of the home page.

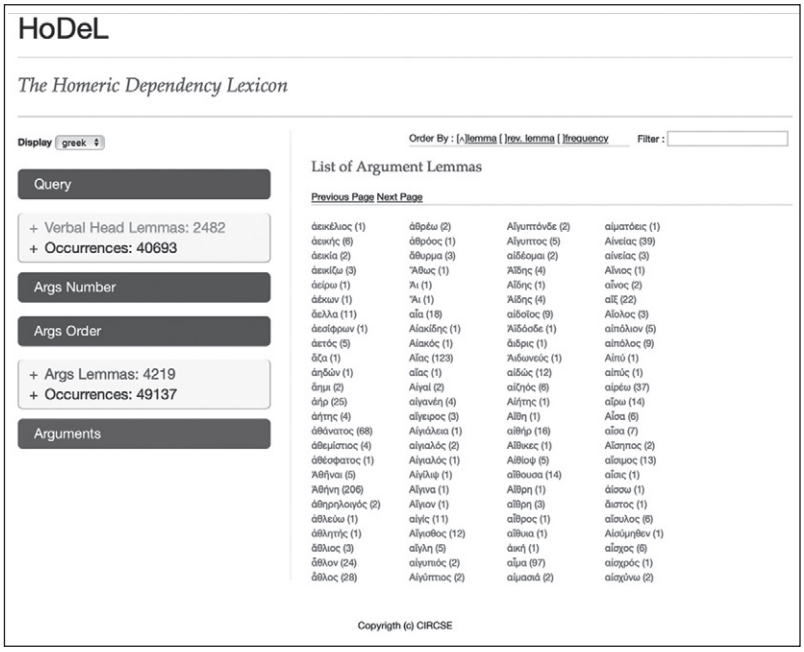
By default, HoDeL gives frequency information concerning verbal lemmas and their dependents tagged as SBJ, OBJ, PNOM, and OCOMP, and specifically:

– 2,482 = type frequency of verbal heads;
– 40,693 = token frequency of verbal heads;
– 4,219 = type frequency of dependent lemmas;
– 49,137 = token frequency of dependent lemmas.[6]

As users add filters to their queries, HoDeL always provides these and other frequency counts.

6. Note that the token frequency of dependent lemmas is higher than the token frequency of verbal heads (i.e., 49,137 > 40,693). This is because if two dependent lemmas are taken by a certain verb *in the same occurrence*, that occurrence is listed twice in the dependent occurrence count, i.e., one for each dependent.

Figure 3 List of dependent lemmas



Users can also visualize all lemmas that depend on Homeric verbal heads, as shown in Figure 3. A number of verbs occur in this list: these verbs function as main verbs in dependent SBJ or OBJ clauses.

4.3. How to search for specific verbs and dependents?

Both lists in Figures 2 and 3 contain clickable lemmas. For example, by clicking on a verbal lemma of the list in Figure 2, e.g., *akoúō* 'hear', users obtain all its forms occurring in the Homeric poems, the ordered contexts of these occurrences (automatically chunked by an algorithm that exploits punctuation marks), and syntactic subtrees representing the queried verb and its argumental dependents (Figure 4). In the output passages and subtrees (see Figures 4–5), the verbal forms and the dependents are circled.

HoDeL summarizes the selected query filters in the box 'Query' and at the top of the output page (see Figure 5) and provides users with frequency information: the verb *akoúō* 'hear' occurs 182 times in the Homeric poems and takes 86 different argument lemmas. In turn, the argument lemmas have a token frequency of 210. Again, the token frequency of argument lemmas is higher than the token frequency of *akoúō* (see fn. 6).

Figure 4 Frequency information and syntactic subtrees of *akoúō* 'hear'

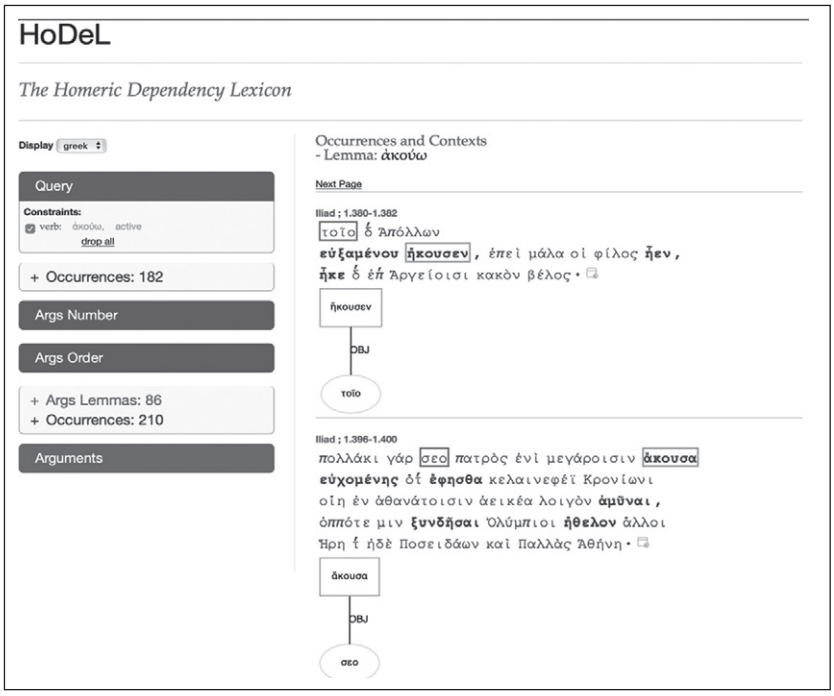


Figure 5 Summary of active constraints

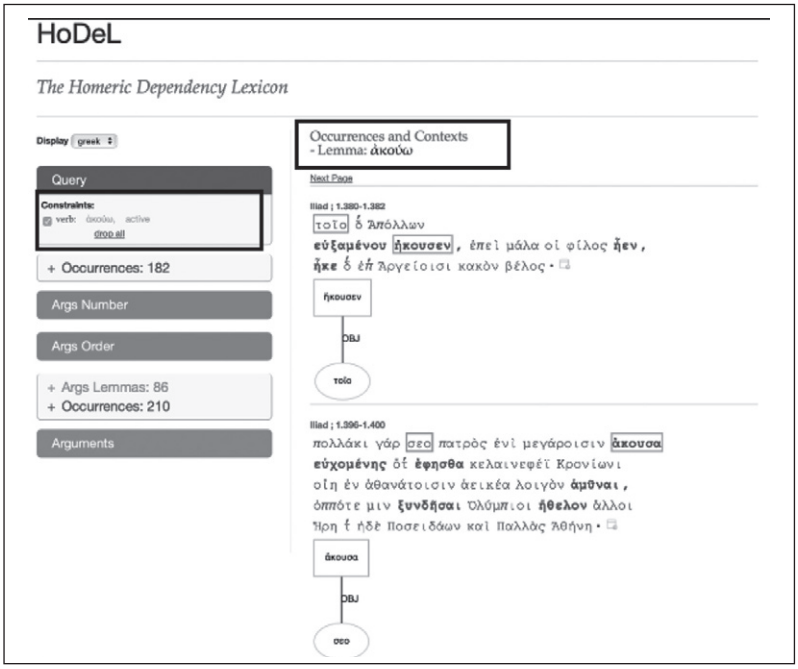
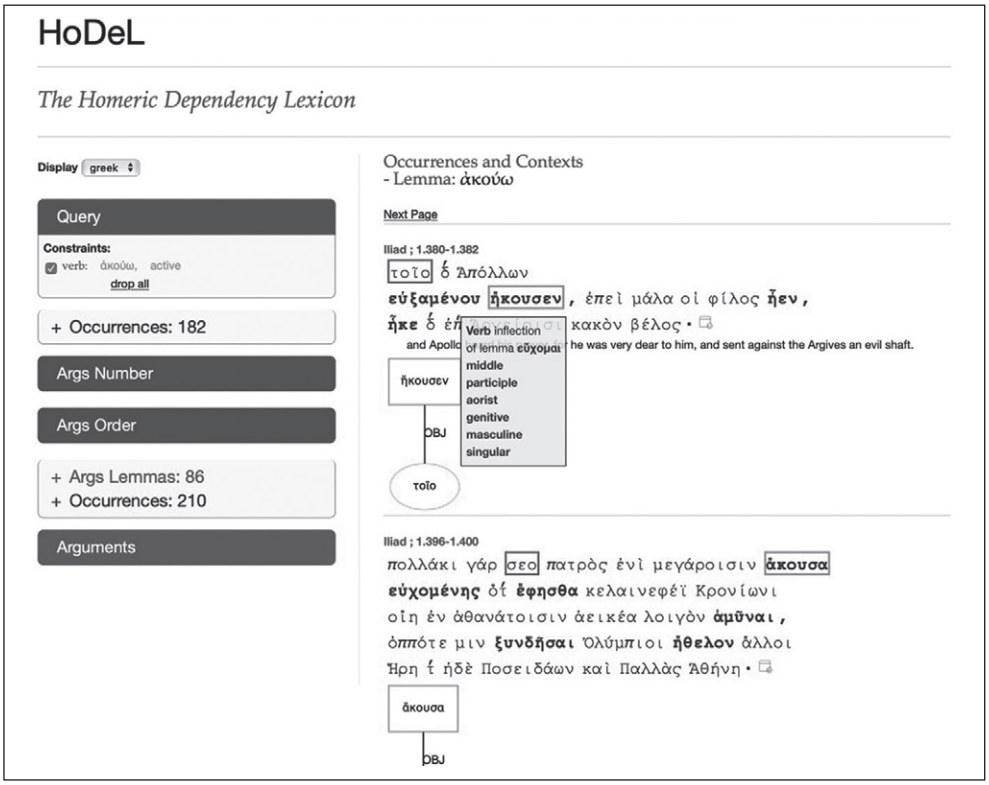


Figure 6 Visualizing morphological metadata and English translations



4.4. How to visualize morphological information and English translations?

By pointing at a word in the output contexts, users obtain morphological annotation as stored in the morphological layer of AGDT 2.0 (Figure 6). For example, Figure 6 shows that the form *euxaménou* is the genitive masculine singular of the aorist middle participle of the verb *eúkhomai* 'pray'. Furthermore, if users click on the folder after the Greek text, HoDeL provides the corresponding English translations. The latter have been automatically aligned with the Greek text using an algorithm that exploits punctuation marks and text chunks contained in the texts provided at the *Perseus Digital Library*. The automatic alignment has been manually checked and, when necessary, modified according to the translation available at *The Chicago Homer* (Figure 6).[7]

7. The translation available at the *Perseus* project are those by Murray (1919, 1924). *Chicago Homer*'s translations are those by Lattimore (1951, 1967).

4.5. How to use the 'Args Number' *and* 'Args Order' *constraints?*

The box 'Args Number' shows frequency information concerning the number of arguments taken by verbal heads and concerning the syntactic relations (SBJ, OBJ, PNOM, OCOMP) of arguments taken by verbal heads. For example, as shown in Figure 7, *akoúō* 'hear' can take from zero (37 occurrences) to three arguments (2 occurrences). When this verb is the head of two dependents, the latter can have different syntactic functions, called 'Subcat.(egories)' in the resource. By flagging one of these categories ('No. Args') and subcategories (i.e., argument number *and* functions), users obtain filtered passages and subtrees. Note that the categories and subcategories suggested by the system are corpus-induced for each selected verb.

The constraint 'Args Order' allows users to investigate constituent order in Homeric Greek. As shown in Figure 8, at a lower level of granularity, attested relative orders of verbs (*akoúō* 'hear' in this case) and OBJ dependents, together with their frequencies (attested verb-OBJ orders are labeled as 'Cat.(egories)' in the interface), are provided for users. At a higher level of granularity, for each attested verb-OBJ order, the relative positioning of other argumental dependents, such as SBJ, can be taken from the lexicon (in this case, attested orders are labeled as 'Subcat.(egories)'). As seen in the 'Args Number' example, these orders are given by the system based on patterns attested in the corpus. Both categories and subcategories of orders can be flagged to obtain filtered contexts and subtrees.

4.6. How to use the 'Arguments' *constraint?*

Users can search the Homeric verbs by argument relation and case/mood using the box 'Arguments'. In Figure 9, the attested functions and forms of arguments taken by the verb *akoúō* 'hear' are shown, and each attested 'Cat.(egory)' and 'SubCat.(egory)' can be flagged to obtain filtered examples and relative subtrees.

4.7. How to type in and search for specific verbs and dependents?

By clicking on the box 'Query', a window opens in which the requested lemma can be typed using Beta Code (Figure 10; cf. Table 1). In Figure 10, I typed the verb *akoúō* a)koúw in the 'Verbal Head Lemma' box: its relative subtrees can be seen by clicking 'Submit', the button that launches queries.

Additional filters are incorporated in the 'Query' box as illustrated in Figure 10. First, users can work only on a single Homeric poem, by using the drop-down menu 'Poem'. Similarly, they can search for verbs in a specific morphological 'Voice' (available options = active : passive : middle : medio-passive). In addition to verbal lemmas, users can also search for specific argument lemmas, by typing them in the 'Ar-

Figure 7 Frequency information on argument number

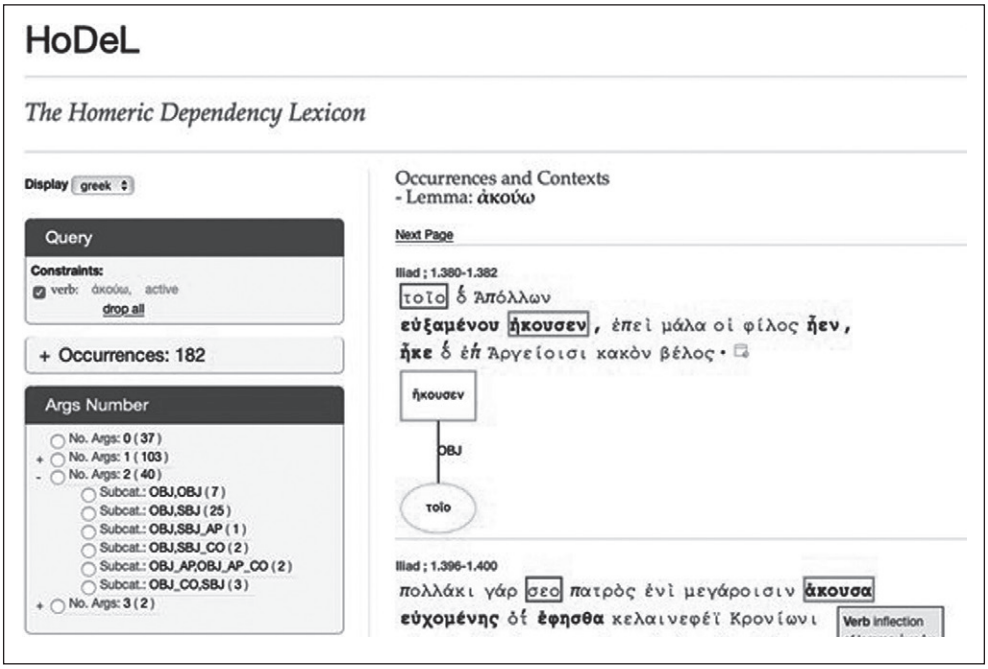


Figure 8 Word order information on verbal arguments

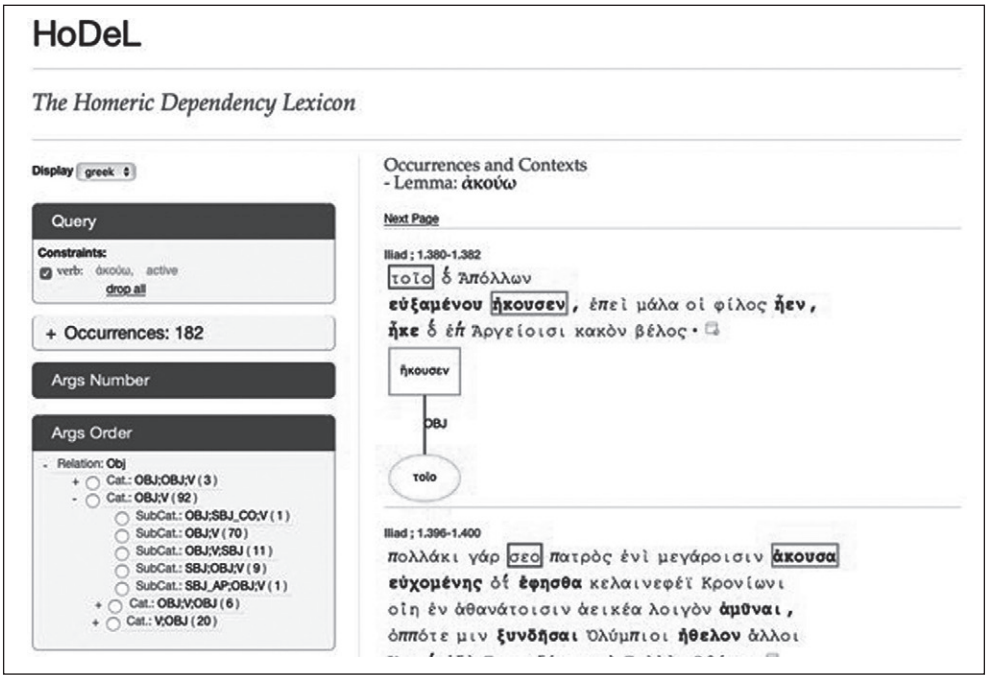


Figure 9 Filtering arguments by relation and morphological information

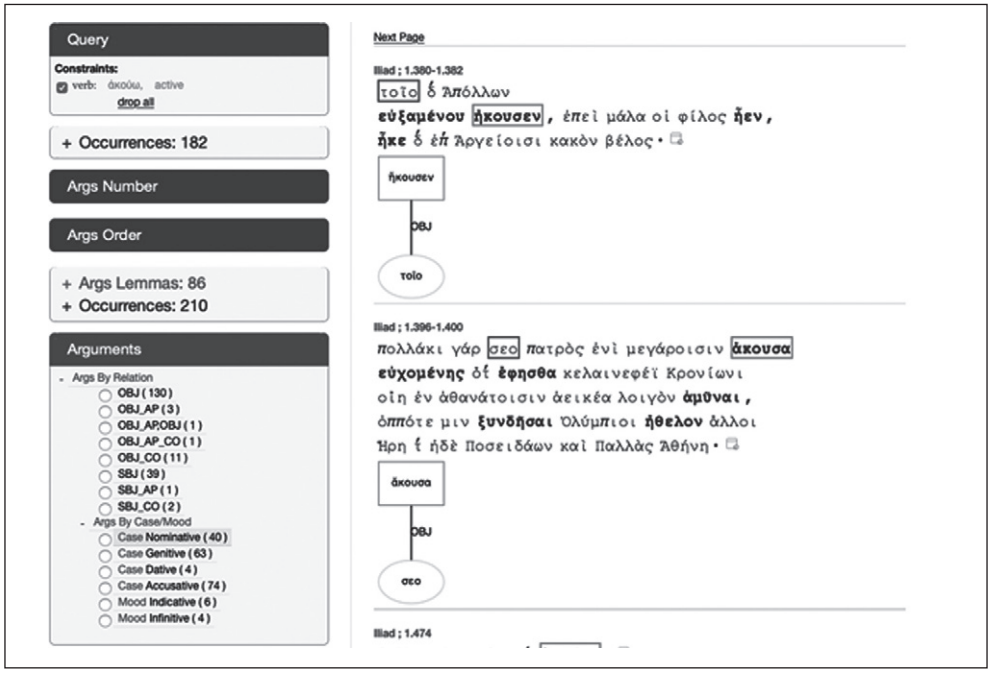
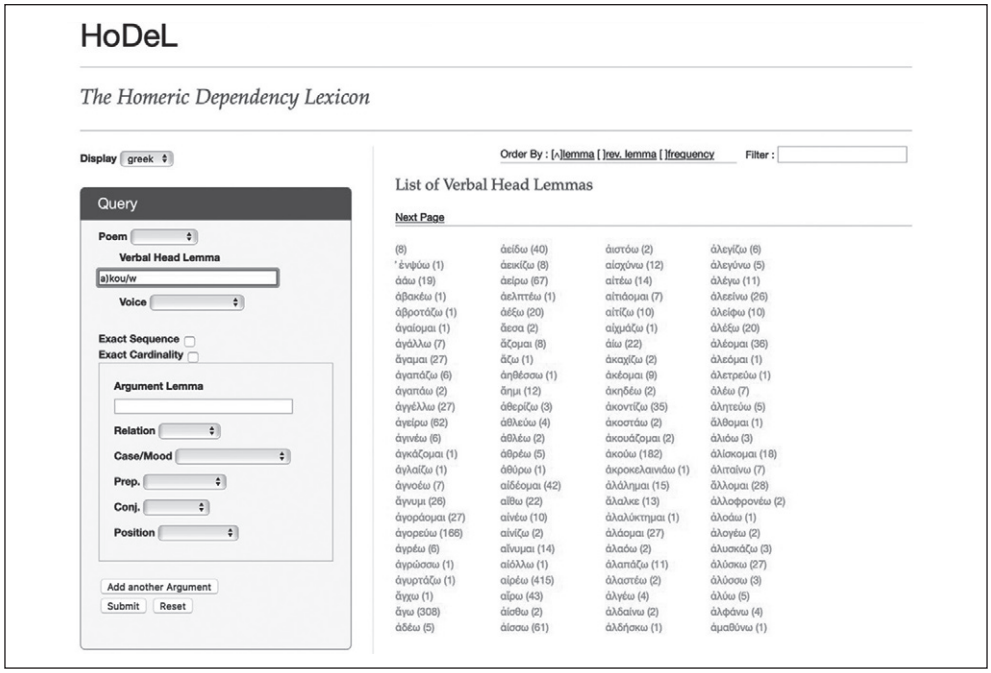


Figure 10 Typing in the lemma *akoúō* 'hear'



gument Lemma' box. Users can also filter their output contexts and subtrees based on some features of the argument lemmas, by using the drop-down menus provided under the 'Argument Lemma' box. Specifically, they can search by 'Relation' (available options = Sbj : Obj : Pnom : Ocomp), by 'Case/Mood' (available options for Case = nominative : genitive : dative : accusative : vocative; available options for Mood = Indicative : Subjunctive : Infinitive : Imperative : Participle : Optative), by 'Preposition' (a data driven list of the Ancient Greek lemmas that are annotated as prepositions in AGDT 2.0 is given by the resource, both in Greek and in Latin scripts), by 'Conjunction' (a data driven list of the Ancient Greek lemmas that are annotated as conjunctions in AGDT 2.0 is automatically given by the resource, both in Greek and in Latin script), and by 'Position' (an argument can occur before verb : after verb: b./a. verb). All these parameters can be combined with one another and can be associated with a typed verbal and/or argument lemma. As seen before, users should remember to click the 'Submit' button to run their queries.

HoDeL also allows users to search for more than one argument at one time: to do this, one should employ the 'Add another Argument' button. By clicking on it, an additional 'Argument Lemma' box appears, together with the related drop-down menus for choosing argument features. Each additional argument can be deleted using the 'Delete this argument' button. When searching for more than one argument, the options 'Exact Sequence' and 'Exact Cardinality' become useful: the former searches for the exact sequence of arguments as listed in the form below; the latter searches for the exact number of arguments as listed in the form below, regardless of their order. The button 'Reset' clears the form.

5. *Research questions that can be addressed using HoDeL: Examples*

After illustrating how HoDeL was built, its basic functionalities, and the kinds of data it contains (Sections 3–4), I now show some examples of how the lexicon can help researchers to operationalize specific research questions on Homeric syntax.

The main advantage of using HoDeL lies in the fact that it allows users to carry out corpus-based quantitative studies on Homeric Greek without learning the complex formalisms necessary to directly query the treebanks of AG. Currently, AGDT 2.0 can be queried online from the web-repository of *Universal Dependencies* using a language called PML-Tree Query. However, this method has the disadvantage of not allowing specific texts to be singled out from the rest of the treebank: so, for example, Homeric texts cannot be investigated separately from later diachronic varieties of AG.[8] To focus on the Homeric texts independently from the rest of the tree-

8. On the GitHub page of AGLDT 2.0, it is stated that the treebanks can also be queried online

bank, one has to download the whole treebank in.xml format, separate the Homeric texts, convert them into another format (e.g, to the .pml or the ConLL-U formats), download a query tool (to my knowledge, the *Tree Editor – TrEd* is one of the few supporting the .pml format, while *Udapi,* Popel et al. 2017, allows working with the ConLL-U format) and finally query the texts using the supported formalism. The other main treebank of AG, PROIEL, can be likewise queried via *Universal Dependencies* or via INESS Search (a reimplementation of Tiger Search), but it does not contain Homeric Greek. Thus, HoDeL fills a gap among the available linguistic resources: it offers an extremely user-friendly interface to perform corpus-based research on the Homeric poems.

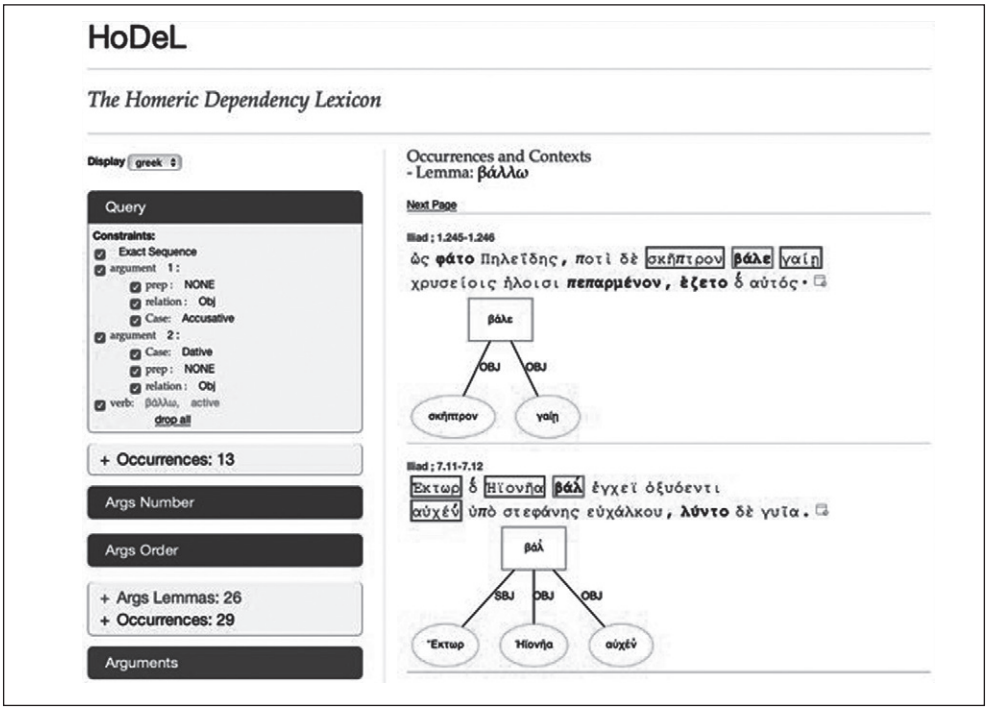
To begin with, HoDeL can be used to automatically retrieve all relevant examples of the construction under investigation. For example, the 'Args Order' option (Figure 8) can be employed to obtain the frequency distribution of sentences attesting to the VSO, SVO, and SOV orders in Homeric Greek. This data could contribute to shedding light on a number of still-open issues regarding Homeric word order and information structure (cf., e.g., Beschi 2018 with references).

The functionality 'Arguments' (Figure 9) can be used to extract all coordinated subjects and objects by selecting the relevant argument relations, specifically, SBJ_CO, OBJ_CO, SBJ_AP_CO, and OBJ_AP_CO (_CO means 'coordinated', while _AP_CO means 'appositive and coordinated'). If the outputs of this filter are crosschecked with those of the 'Args Order' filter, researchers can effortlessly obtain frequency information on positioning patterns of coordinated subjects and objects with respect to verbs: do coordinated elements tend to surface in the same position, be it preverbal or postverbal, or do coordinants tend to be 'split' by verbs? How do these ordering patterns correlate with verbal agreement in the case of coordinated subjects? What do these ordering patterns reveal about verbal government of coordinated objects?

Beyond facilitating queries related to word order, HoDeL can also be useful to detect passages containing infrequent patterns of the Homeric language, which would require a time-consuming manual reading of the poems to be detected. For example, by searching for a specific preverbed verbal lemma in the 'Query' box and combining it with the prepositional phrase headed by the same local particle, one can easily find attested instances of preverb repetition outside the preverbal context. This information can be used to account for the different paces of grammaticalization or lexicalization paths undergone by different AG preverbs (see, e.g., Zanchi 2017): the local particles that allow for repetition are more lexicalized or grammaticalized into preverbs and prepositions.

via *Structural Search* and *Tündra*. Currently, however, neither of the two links given seems to work (http://perseusdl.github.io/treebank_data/; last access: 2021-05-30).

Figure 11 Accusative and dative dependents taken by *bállō* 'throw, hit'



The option 'Add another argument' can be employed to investigate ditransitive verbs that feature argument structure alternations, such as the transfer verb *bállō* 'throw, hit' (Figure 11). This verb can mean 'throw something (ACC) toward something else / someone (DAT)', as in Il. 1.245-246, or 'hit someone (ACC) with something (DAT), as in Il. 7.11-12. Both passages are shown in Figure 11.

Note that in Il. 7.11-12, the instrumental dative is labeled as OBJ, in spite of its uncertain argumental status in the domain of syntactic valency (cf. Section 3.2). Thus, the OBJ tag may well be imprecise from a theoretical standpoint, but this annotation has the welcome advantage that it demonstrates the suitability of HoDeL for this study and similar ones. Indeed, HoDeL is richer than a strictly syntactic valency lexicon and allows investigations of the behavior of event participants whose argumental status is controversial, such as those regarded as *optional inner complementations* in the view of the *Functional Generative Description*.

6. *Conclusions*

This paper presented HoDeL, a new lexicon intended to ease and refine the researching of Homeric verbs and their dependents. The data on which the lexicon

is based and the methodology that has been employed to build it were documented and framed within the larger picture of morphosyntactically-annotated corpora of AG and valency lexica of ancient and modern Indo-European languages. The basic functionalities and incorporated constraints of the HoDeL online interface were illustrated and accompanied by suggestions about how to interpret frequency counts. Finally, the paper showed how the lexicon can be employed to easily operationalize diverse research questions concerning the Homeric syntax, and how its user-friendly interface and incorporated filters and queries allow scholars with basic computational skills to perform advanced corpus-based studies on the Homeric language. In addition, the paper demonstrated how HoDeL may be used to search for morphological information, transliteration and aligned translations of the AG passages, which also greatly facilitate the interpretation of the output results.

For the future, we plan to continue improving the quality of the base data contained in AGDT 2.0. In addition, the HoDeL team is working to link the lexicon with other lexical resources of AG, such as the growing *Ancient Greek WordNet* (Biagetti et al. 2021; Short in this volume). As shown in Zanchi et al. (2021), the enhanced access to data and the extreme user-friendliness of HoDeL can be exploited to integrate sentence frames in the metadata associated to each verbal entry of the *Ancient Greek WordNet*.

Abbreviations

ACC = accusative, AG = Ancient Greek, AOR = aorist, GEN = genitive, Il. = Iliad, IMPF = imperfect, INDF = indefinite, INF= infinite, NEG = negation, NOM = nominative, Od. = Odyssey, PASS = passive, PL = plural, PRS = present, PTC = particle, PTCP = participle, SG = singular, 1= first person, 3 = third person

Websites

AGDT 2.0 (Ancient Greek Dependency Treebank): https://perseusdl.github.io/treebank_data/

AGDT 2.0 (Ancient Greek Dependency Treebank) guidelines of the analytical layer = https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/guidelines/Greek_guidelines.md#prg_ann

Ancient Greek WordNet: https://greekwordnet.chs.harvard.edu

The Chicago Homer: https://homer.library.northwestern.edu

DĀMOS (Database of Mycenaean at Oslo University): https://damos.hf.uio.no/1

DFHG (Digital Fragmenta Historicorum Graecorum): http://www.dfhg-project.org

The Diorisis Ancient Greek Corpus: https://www.turing.ac.uk/research/publications/diorisis-ancient-greek-corpus
EAGLE (Electronic Archive of Greek and Latin Epigraphy): http://www.edr-edr.it/it/Link_it.php
HoDeL (Homeric Dependency Lexicon), resource: https://hodel.unipv.it/hodel-res/
HoDeL (Homeric Dependency Lexicon), guidelines: https://su-lab.unipv.it/tasf/wp-content/uploads/2021/01/HoDeL_guidelines.pdf
ISWOC (Information Structure and Word Order Change in Germanic and Romance Languages): http://www.hf.uio.no/ilos/english/research/projects/iswoc/; http://iswoc.github.io
IT-TB (Index Thomisticus Treebank): https://itreebank.marginalia.it/itvalex
IT-VaLex (Index Thomisticus Valency Lexicon): https://itreebank.marginalia.it/itvalex; https://github.com/CIRCSE/ITVALEX
LSJ (Liddell-Scott-Jones Dictionary): http://www.perseus.tufts.edu/hopper/resolveform?redirect=true&entry=fe/rw
Perseus Digital Library: http://www.perseus.tufts.edu/hopper/collections
PDT 3.0 (Prague Dependency Treebank): https://ufal.mff.cuni.cz/pdt3.0
PML- Tree Query: https://ufal.mff.cuni.cz/pmltq
PROIEL (Pragmatic Resources of Old Indo-European Languages): https://www.hf.uio.no/ifikk/english/research/projects/proiel/
SEMANTIA: https://sematia.hum.helsinki.fi/user/
Syntacticus: http://syntacticus.org
The Pavia linguistic resources repository: https://su-lab.unipv.it/tasf/
TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien): http://titus.fkidg1.uni-frankfurt.de/framee.htm?/texte/texte2.htm
TLG (Thesaurus Linguae Graecae): http://stephanus.tlg.uci.edu
TOROT (The Tromsø Old Russian and OCS Treebank): http://torottreebank.github.io/; https://nestor.uit.no
TrEd (Tree Editor): https://ufal.mff.cuni.cz/tred/
Universal Dependencies: https://universaldependencies.org

References

Ágel, Vilmos & Fischer, Klaus. 2015. Dependency Grammar and Valency Theory. In *The Oxford Handbook of Linguistic Analysis* (2 edn.), Bernd Heine & Heiko Narrog (eds), 223–256. Oxford: OUP.

Aurora, Federico. 2015. DĀMOS (Database of Mycenaean at Oslo). Annotating a fragmentarily attested language. In *Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics* (CILC2015), Pedro A. Fuertes-Olivera et al. (eds), 21–31. *Procedia - Social and Behavioral Sciences* 198.

Bamman, David & Crane, Gregory. 2006. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, Joakim Nivre & Jan Hajič (eds), 67–78. Prague: ÚFAL.

Bamman, David & Crane, Gregory. 2008. Building a dynamic lexicon from a digital library. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL 2008), Christoph Becker, Hannes Kulovits, Andreas Rauber & Hans Hofman (eds), 11–20.

Bamman, David & Crane, Gregory. 2011. The Ancient Greek and Latin Dependency Treebank. In *Language Technology for Cultural Heritage*, Caroline Sporleder, Antall van Den Bosch & Kalliopi Zervanou (eds), 79–89. Berlin: Springer.

Beschi, Fulvio. 2018. The Ancient Greek sentence left periphery. A study on Homer. *Journal of Greek Linguistics* 18: 172–210.

Biagetti, Erica. 2018. *A dependency treebank of Classical Sanskrit*. MA thesis, University of Pavia.

Biagetti, Erica. This volume. Annotating the *RigVeda*: Challenges and methodology in parsing the earliest religious poetry of India.

Biagetti, Erica, Zanchi, Chiara & Short, William M. 2021. Toward the creation of WordNets for ancient Indo-European languages. In *Proceedings of the 11th Global WordNet Conference*, Sonja Bosch, Christiane Fellbaum, Marissa Griesel, Alexandre Rademaker & Piek Vossen (eds), 258–266. EACL/GWC: Global WordNet Association.

Celano, Giuseppe G.A. 2019. The Dependency Treebanks for Ancient Greek and Latin. In *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, Monica Berti (ed), 279–298. Berlin: De Gruyter.

Colleman, Timothy, Defrancq, Bart, Devos, Filip & Noël, Dirk. 2004. *The Contragram Dutch–French–English Contrastive Verb Valency Dictionary*. Ghent: University of Ghent.

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

Eckhoff, H. Martine, Bech, Kristin, Eide, Kristine, Bouma, Gerlof, Haug, Dag T. T., Haugen, Odd E. & Jøhndal, Marius. 2018. The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52 (1): 29–65.

Eckhoff, H. Martine & Haug, Dag T. T. This volume. Annotation schemes, tools and data in the PROIEL treebank family.

Fried, Mirjam & Boas, Hans C. (eds) 2005. *Grammatical constructions*: *back to the roots*. Amsterdam: Benjamins.

Frigione, Chiara. 2015. *Lexematik, Konstruktionsmuster und Argumentstruktur im altkirchenslavischen Verbum: Materialien für eine Lexikosyntax des Slavischen / Lessematica, modello costruzionale e struttura argumentale nel verbo paleoslavo: Materiali per una lessico-sintassi verbale paleoslava*. PhD dissertation, Universität zu Köln / Università per Stranieri di Siena.

Hajič, Jan, Panevová, Jarmila, Buráňová, Eva, Urešová, Zdeňka & Bémová, Alevtina (in cooperation with) Kárník, Jiří, Štěpánek, Jan, Pajas, Petr. 1999. Annotations at Analytical Level: Instructions for Annotators.

< https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html > (2021-05-30).
Happ, Heinz. 1976. *Grundfragen einer Dependenz-Grammatik des Lateinischen*. Göttingen: Vandenhoeck & Ruprecht.
Haug, Dag T. T. 2012. Syntactic conditions on null arguments in the Indo-European Bible translations. *Acta Linguistica Hafniensia* 44(2): 129–141.
Haug, Dag T. T. & Jøhndal, Marius L. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data* (LaTeCH 2008), Caroline Sporleder & Kiril Ribarov (eds), 27–34.
Helbig, Gerhard & Wolfgang, Schenkel. 1991 [1969]. *Wörterbuch zur Valenz und Distribution deutscher Verben*. Tübingen: Niemeyer.
Hellwig, Oliver & Sellmer, Sven. This volume. The Vedic Treebank.
Herbst, Thomas, Heath, David, Roe, Ian F. & Götz, Dieter. 2004. *A Valency Dictionary of English.* Berlin: de Gruyter.
Horrocks, Geoffrey C. 2010. *Greek. A history of the language and its speakers*. Oxford: Blackwell.
Keydana, Götz & Luraghi, Silvia. 2012. Definite referential null objects in Vedic Sanskrit and Ancient Greek. *Acta Linguistica Hafniensia* 44(2): 116–128.
Kingsbury, Paul & Palmer, Martha. 2002. From Treebank to Propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Manuel González Rodríguez & Carmen Paz Suarez Araujo (eds), 1989–1993. Las Palmas-Gran Canaria: ELRA.
Korhonen, Anna, Krymolowski, Yuval & Briscoe, Ted. 2006. A Large Subcategorization Lexicon for Natural Language Processing Applications. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk & Daniel Tapias (eds), 1015–1020. Genoa: ELRA.
Lattimore, Richmond A. 1951. *The Iliad of Homer*. Chicago: University of Chicago Press.
Lattimore, Richmond A. 1967. *The Odyssey of Homer.* New York: Harper and Row.
Lord, Albert B. 1960. *The singer of tales*. Cambridge, MA: Harvard University Press.
Luraghi, Silvia. 2003. Definite referential null objects in Ancient Greek. *Indogermanische Forschungen* 108: 169–196.
Matthews, Peter H. 1981. *Syntax*. Cambridge: Cambridge University Press.
McGillivray, Barbara. 2013. *Methods in Latin Computational Linguistics*. Leiden: Brill.
McGillivray, Barbara & Passarotti, Marco C. 2009. The Development of the Index Thomisticus Treebank Valency Lexicon. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education* [LaTeCH – SHELT&R 2009], Lars Borin & Piroska Lendvai (eds), 43–50. Athens: ACL.
McGillivray, Barbara & Passarotti, Marco C. 2015. Accessing and using a corpus-driven Latin Valency Lexicon. In *Latin Linguistics in the Early 21st Century. Acts of the 16th International Colloquium on Latin Linguistics, Uppsala*, June 6th–11th , 2011, Gerd V. M. Haverling (eds), 289–300. Uppsala: Uppsala Universitet.

McGillivray, Barbara & Vatri, Alessandro. 2015. Computational valency lexica for Latin and Greek in use: a case study of syntactic ambiguity. *Journal of Latin Linguistics* 14(1): 101–126.

Messiant, Cédric, Korhonen, Anna & Poibeau, Thierry. 2008. LexSchem: A Large Subcategorization Lexicon for French Verbs. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias (eds), 533–538. Marrakech: ELRA.

Murray, A. T. 1919. *Homer Odyssey, 2 volumes.* Translated by A. T. Murray. Revised by George E. Dimock. Loeb Classical Library 104. Cambridge, MA: Harvard University Press.

Murray, A. T. 1924. *Homer. Iliad, 2 volumes.* Translated by A. T. Murray. Revised by William F. Wyatt. Loeb Classical Library 170. Cambridge, MA: Harvard University Press.

Panevová, Jarmila. 1994. Valency Frames and the Meaning of the Sentence. In *The Prague School of Structural and Functional Linguistics. A short introduction*, Philip A. Luelsdorff (eds), 223–243. Amsterdam: Benjamins.

Parry, Adam (ed). 1971. *The making of Homeric verse: the collected papers of Milman Parry*. Oxford: Oxford University Press.

Passarotti, Marco C., González Saavedra, Berta & Onambele, Christophe. 2016. Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 2599–2606. Portorož: ELRA.

Popel, Martin, Žabokrtský, Zdeněk & Vojtek, Martin. 2017. Udapi: Universal API for Universal Dependencies. In *Proocedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, Marie-Catherine de Marneffe, Joakim Nivre, Sebastian Schuster (eds), 96–101. Gothenburg: Association for Computational Linguistics.

Ruppenhofer, Josef, Ellsworth, Michael, Petruck, Miriam R. L., Johnson, Christopher R. & Scheffczyk, Jan. 2006. *FrameNet II. Extendend Theory and Practice*. <http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126> (2021-05-30).

Sausa, Eleonora & Zanchi, Chiara. 2015. Non-accusative null objecs in the Homeric Dependency Treebank. In *Proceedings of the Workshop on Corpus-Based Research in the Humanities,* Marco C. Passarotti, Francesco Mambrini & Caroline Sporleder (eds), 107–116. Warsaw: Institute of Computer Science of the Polish Academy of Sciences.

Schumacher, Helmut, Kubczak, Jacqueline, Schmidt, Renate & de Ruiter, Vera. 2004. *VALBU—Valenzwörterbuch deutscher Verben*. Tübingen: Narr.

Sgall, Petr, Hajičová, Eva & Panevová, Jarmila. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht: D. Reidel.

Short, William M. This volume. WordNets, Sembanks, and the Challenge of Semantic Polyvalency.

Siewierska, Anna. 2005. Passive constructions. *WALS Online* <http://wals.info/chapter/107> (2021-20-04).

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

Vatri, Alessandro & McGillivray, Barbara. 2018. The Diorisis Ancient Greek Corpus. *Research Data Journal for the Humanities and Social Sciences* 3(1): 55–65.

Vierros, Marja. 2018. Linguistic Annotation of the Digital Papyrological Corpus: Sematia. In *Digital Papyrology II*, Nicola Reggiani (ed), 105–118. Berlin: De Gruyter.

Watkins, Calvert. 1976. Toward Proto-Indo-European syntax: Problems and pseudo-problems. In *Proceedings of the Chicago Linguistics Society: Papers from the parasession on diachronic syntax*, Sanford B. Steever, Carol A. Walker & Salikoko S. Mufwene (eds), 305–326. Chicago: Chicago Linguistics Society.

Zanchi, Chiara. 2017. New evidence for the Source-Goal asymmetry. Ancient Greek preverbs. In *Space in Diachrony*, Silvia Luraghi, Tatiana Nikitina & Chiara Zanchi (eds), 147–178. Amsterdam: Benjamins.

Zanchi, Chiara. 2019. *Multiple Preverbs in Ancient Indo-European Languages.* Tübingen: Narr.

Zanchi, Chiara. Forthc. The *Homeric Dependency Lexicon*: what it is and how to use it. *Journal of Greek Linguistics.*

Zanchi, Chiara & Luraghi, Silvia. 2020. Presenting *HoDeL* – A new resource for research on Homeric Greek verbs. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020"* [Issue 19. Supplementary volume], 1188–1200.

Zanchi, Chiara, Sausa, Eleonora & Luraghi, Silvia. 2018. HoDeL, a Dependency Lexicon for Homeric Greek: Issues and Perspectives. In *Formal Representation and the Digital Humanities*, Paola Cotticelli-Kurras & Federico Giusfredi (eds), 221–246, Cambridge: Cambridge Scholars Publishing.

Zanchi, Chiara, Luraghi, Silvia, & Biagetti, Erica. 2021. Linking the Ancient Greek WordNet to the Homeric Dependency Lexicon. In *Computational Linguistics and Intelligent Technologies. Papers from the Annual International Conference "Dialogue 2021"* [Issue 20], 729–737.

Introducing DEmA: the Pavia Diachronic Emergence of Alignment Database

Sonia Cristofaro*, Guglielmo Inglese**

The Pavia *Diachronic Emergence of Alignment* (DEmA) database is a new resource for the study of the diachrony of alignment patterns cross-linguistically. In this paper, we offer a description of DEmA, its structure and the choices that have been made in its construction. The main goal of DEmA is to offer a platform that makes it possible to investigate the sources and processes out of which new alignment patterns come into being across languages. In order to do so, each instance of the emergence of a construction with a new alignment pattern is decomposed into a number of well-defined parameters pertaining to the initial situation in the language, the developmental mechanisms leading to the new alignment pattern, and the effects of the change. These various parameters are effectively implemented into a searchable format. This systematization enables users to easily retrieve and compare various type of information concerning the emergence of alignment patterns in the world's languages.

Keywords: alignment pattern, diachronic typology, grammaticalization, historical linguistic, parameters of language change, database

1. *Introduction*[1]

Over the past decades, typologists have repeatedly stressed the importance of taking diachronic information into consideration when explaining cross-linguistic regularities (see recently Grossman and Polis 2018; Cristofaro 2019; Haspelmath 2019). Unfortunately, resources providing information on how specific phenomena develop over time cross-linguistically are not numerous. Progress in grammaticalization studies and historical linguistics has brought to light an increasing body of evidence regarding the possible origins of different alignment patterns. Information about these processes is, however, scattered across specialized publications, and often not easily comparable from one language to another, nor accessible to non-specialists.

In this paper, we introduce the Pavia *Diachronic Emergence of Alignment* (DEmA) project. The project aims to build a comprehensive open access database

* Sorbonne Université. ** KU Leuven – FWO.

1. The *Pavia Diachronic Emergence of Alignment* (DEmA) project has been carried out at the Department of Humanities of the University of Pavia, Section of Theoretical and Applied Linguistics. The project was also funded by the Italian Ministry for Education and Research (MIUR) in the framework of the 2015 PRIN call, project 'Transitivity and argument structure in Flux' (grant no. 20159M7X5P).

on the emergence of alignment patterns cross-linguistically, so as to complement existing typological databases on alignment, for example the three WALS chapters devoted to this topic (Comrie 2013a, 2013b; Siewierska 2013), which only provide information about synchronic patterns.

The data in DEmA is systematized in such a way that one can readily search and compare various type of information pertaining to the role of the different components at play in the emergence of new alignment patterns. In particular, we propose to decompose the emergence of alignment patterns into three notionally distinct domains: the initial stage of the language, the developmental mechanism, and the results of the change.

The paper is organized as follows. In Section 2, we briefly outline current issues in the diachronic study of alignment patterns and discuss the possible research questions that DEmA will make it possible to explore. Section 3 focuses on the structure of DEmA: we first describe the parameters relevant to the initial situation of the language (Section 3.1) and developmental mechanisms (Section 3.2). We then move to the parameters describing the effects of the change on the global alignment of the language (Section 3.3). Section 4 deals with the practical aspects of how queries can be carried out in DEmA.

2. *Alignment patterns in diachrony*

By alignment pattern is meant here, in a maximally general sense, any possible grouping of the three argument roles A, S, and P (Comrie 1989; Dixon 1994), in terms of case marking (nominal inflection, adpositions, clitics), indexation, or other morphosyntactic phenomena.

Progress in grammaticalization studies and the study of language change cross-linguistically means that a comparatively large body of data is now available on the emergence of alignment patterns in a variety of languages across different families and geographical areas (see, for example, Gildea 1998 on Carib; König 2008 on African languages; Bubenik 1998, Haig 2008 and 2017, Verbeke 2013 on Indo-Aryan). This evidence, however, has not yet been integrated into a comprehensive overview of the possible sources and developmental mechanisms that can give rise to particular alignment patterns (for example, accusative, ergative, or active) from one language to another. An early study in this direction is Harris and Campbell (1995: chap. 9), which, however, concentrates on possible mechanisms of alignment change, rather than the specific alignment patterns emerging through each mechanism, or the source constructions that can give rise to individual patterns. Another strand of cross-linguistic research (e.g. Heine and Kuteva 2002 [now Kuteva et al. 2019]; Kulikov 2006) has focused on the etymology of particular case markers, irrespective of the contexts and developmental mechanisms that lead to these markers

evolving from particular source elements, or the consequences of this process for the alignment patterns of the language.

In general, research on the emergence of alignment patterns in individual languages has shown that individual patterns typically emerge from pre-existing constructions, through various mechanisms of constructional reinterpretation or, sometimes, phonological change. The main goal of DEmA is to provide an expanding platform where the available evidence on these processes is integrated in a typologically informed framework that makes it possible to compare different processes from one language to another, so as to obtain data both on the emergence of alignment patterns in particular languages, and on the possible sources and developmental processes leading to the emergence of particular alignment patterns cross-linguistically. This type of data can be used to address different research questions about the diachronic origins of alignment (Harris and Campbell 1995; Gildea 1998; Mithun 2005; Creissels 2008; Cristofaro 2012, 2013, 2014; Zúñiga 2018 among others):

1. What source constructions give rise to particular alignment patterns cross-linguistically?
2. What developmental mechanisms lead from particular source constructions to particular alignment patterns?
3. What is the relationship between the properties of particular source constructions and developmental mechanisms and the properties of the resulting alignment pattern, in terms for example of what argument roles are or are not encoded in the same way, or the distribution of the pattern across different contexts (NP-based and TAM-based alignment splits, or other types of splits)?
4. The same alignment patterns (for example, ergative or accusative alignment) originate from different source constructions and through different developmental mechanisms in different cases. Can individual patterns be explained in terms of some overarching principle that applies to all instances of the pattern, or should different instances of the pattern be explained in terms of different principles depending on the developmental processes involved?

3. *The organization of DEmA*

In DEmA, each entry is a process that has led to the development of a construction with a new alignment pattern in some language, as described in published sources. At present, we focus on monotransitive alignment (i.e. alignment of one- and two-place verbs) only.

In line with a number of cross-linguistically oriented accounts (see, for example, Harris and Campbell 1995: Chap. 9), the development of a new alignment pattern

is conceived as a process that takes place within particular constructions, for example through the reinterpretation of the argument structure of these constructions, or through the development of a new marker for A, P, or S arguments as a result of grammaticalization. This process will lead to the development of a particular alignment pattern for the construction in question, and may have different effects depending on the original alignment pattern of the language in the relevant grammatical domain. For example, the development of a new perfective construction with ergative alignment may lead to a TAM based split if non-perfective constructions use a non-ergative pattern. If these constructions have ergative alignment, however, the language will remain consistently ergative.

The most innovative feature of DEmA is that it allows for a fine-grained research of the various components involved in the emergence of new alignment patterns. In particular, DEmA is structured so as to provide information about three different domains:

1. The initial situation in the language, including both the original alignment pattern of the language and a detailed description of the source construction involved in the emergence of the new alignment pattern.
2. Developmental mechanisms, that is, the nature and dynamics of the change that gives rise to the new alignment pattern.
3. The effects of the process of change, including the alignment pattern that develops in the construction undergoing the change and the effects of this development on the global alignment pattern of the language.

For each of these domains, DEmA offers multiple searchable fields, which are described in detail in the reminder of this section.

1.1. The initial situation in the language

This domain pertains to the situation in the language before the emergence of the new alignment pattern. Two distinct fields are provided:

1. **Original alignment pattern**: This refers to the alignment patterns originally attested in the language, along with any constraints in the distribution of these patterns, e.g. accusative, ergative, TAM or NP based splits, and the like.

Only the alignment pattern pertaining to the grammatical domain involved in the process of change is taken into account. For example, if a process of change involves alignment in indexation, only the alignment pattern originally found for indexation in the language (and not, for example, case marking alignment) is taken into account.

2. **Source construction**: This refers to the construction that serves as the basis for the development of the new alignment pattern.

In this field, we focus on the specific elements that undergo change in the development of the new alignment pattern (for example, particular lexical items that grammaticalize into case markers, particular adpositions or case affixes that undergo a change in their grammatical function). While we try to standardize the terminology used in the description of different source constructions cross-linguistically, this field contains highly heterogenous and language-specific descriptions. This is due to the fact that, for each language, different semantic, pragmatic or morphosyntactic properties of the source construction must be taken into account that play a role in the development of the new alignment pattern.

As an example, consider the development of accusative case marking alignment through the reinterpretation of a construction involving the verb *bǎ* 'take' in Mandarin Chinese. The entry for this change in DEmA is shown in Figure 1.
The language originally had neuter case marking alignment, that is, A, S, and P arguments were not distinguished in terms of case marking. In constructions of the type 'take X (and) VERB (X)', where the 'take' verb and some other verb share a P argument, the 'take' meaning was lost, so that *bǎ* evolved into a marker for its former

Figure 1 The emergence of accusative alignment in Mandarin Chinese in DEmA

The Pavia DEmA (Diachronic Emergence of Alignment) Database

Home Languages Search How to use DEmA Credits and contacts

Home › Languages › Mandarin Chinese

Mandarin Chinese

Glottocode: mand1415

Genalogical classification: Sino-Tibetan

Original alignment pattern: Neuter

Source construction: Constructions of the type 'take X (and) VERB (X)', where the verb bǎ 'take' and some other verb share a P argument, as in (1).

Developmental mechanism: The verb bǎ 'take' is reinterpreted as a marker for the shared P argument, and the original biclausal construction is reanalyzed as a monoclausal construction 'ACC X VERB'.

Resulting construction: Transitive construction with a P argument overtly marked by bǎ.

Type of change: Grammaticalization

Alignment in the resulting construction: Nominative-Accusative

Global alignment pattern following the change: Nominative-Accusative

Grammatical domains: Case marking

Symmetry: Asymmetric

Type of data: Historical data

References: Li & Thompson 1974

direct object, 'ACC X VERB'. This is shown by the contrast between the two sentences in (1) and (2), which illustrate, respectively, the use of *bǎ* as a lexical verb and its use as a direct object marker.

(1) Classical Chinese (Sino-Tibetan; Li and Thompson 1974: 202)[2]

Yù	*qīng*	*bǎ*	*tīan*	*zhǐ*	*ruì-lìng*	*yǐ*	*zhēn*
Yu	himself	take	heaven	POSS	mandate	to	conquer
yǒu	*Miáo*						
PTCL	Miao						

'Yu himself took the mandate of heaven to conquer Miao.' (Mè-zǐ, 5th century BCE)

(2) Mandarin Chinese (Sino-Tibetan; Li and Thompson 1974: 203)

Tāmen	*bǎ*	*Zhāg-sān*	[...]	*jǐantao*	*le*	*lǐan*	*xîaoshi*
They	ACC	Zhang-san		scrutinize	ASP	two	hours

'They scrutinized Zhang-san for two hours.'

In the DEmA entry for this process, the field 'original alignment pattern' has 'Neuter', whereas the source construction field provides a description of the construction that gave rise to the accusative pattern: "constructions of the type 'take X (and) VERB (X)', where the verb *bǎ* 'take' and some other verb share a P argument."

The need to distinguish between the source construction and the original alignment pattern attested in the language for the relevant grammatical domain is motivated by the fact that (i) the processes that give rise to a new alignment pattern take place within particular constructions, and may be independent of the alignment patterns previously attested in the language, but (ii) the global effects of individual processes in the language will depend on these patterns. For example, ergative patterns have been shown to develop as intransitive resultative constructions with an oblique NP are reinterpreted as transitive ones, so that the S argument in the intransitive construction becomes a P argument, whereas the oblique NP becomes an A argument ('X is VERBed by Y' > 'Y ERG VERBed X': Gildea 1998, among others). This process will give rise to ergative alignment for resultative constructions, and is independent of the original alignment of S arguments, for example whether they are aligned with A (accusative alignment) or P (ergative alignment). The original alignment of S arguments, however, will determine the global effects of the process in the language. If S arguments were originally aligned with P arguments, the process will only lead to the development of an additional ergative pattern in the language,

2. Glosses and translations of examples are generally taken from the sources. A list of all abbreviations can be found at the end of this paper.

specialized for resultative constructions. By contrast, if S arguments were originally aligned with A arguments, this alignment will be retained for non-resultative constructions, leading to a split between accusative alignment in non-resultative constructions and ergative alignment in resultative ones.

A well-known example of this development comes from Indo-Aryan languages (see Dahl and Stroński 2016 with extensive references), where a tense-based split-ergative system arose through the reinterpretation of Old Indo-Aryan resultative participial constructions with nominatively marked S and instrumental A, as in (3), as transitive constructions with ergative marking on A, as in (4). Notably, while there is a general consensus that the participial construction with *-ta* in (3) served as the basis for the emergence of a new ergative pattern, whether the ergative postposition *=ne* of Modern Indo-Aryan languages, such as Hindi in (4), is a direct continuant of the Old Indo-Aryan instrumental case marking *-eṇa* remains a matter of dispute (Verbecke and De Cuypere 2009).

(3) Vedic (Indo-European; Dahl and Stroński 2016: 18)
ha-tā́ *índr-eṇa* *paṇay-aḥ*
kill-PPP.NOM.PL.M Indra-INS Pani-PPP.NOM.PL.M
śay-adhve
lie_down-2PL.PRS.MID

'You Panis lie down smashed by Indra.'

(4) Hindi (Indo-European; Dahl and Stroński 2016: 12)
laṛke=ne *kitāb* *paṛhī*
boy=ERG book(F).ABS read.PST.PRF.F.SG

'The boy has read the book'

1.2. Developmental mechanisms

For this domain, we provide a number of fields pertaining to various aspects of the processes whereby the source construction gives rise to a new alignment pattern:

1. **Developmental mechanism**: This field features a description of the mechanisms whereby the source construction gives rise to the new alignment pattern.

For example, the developmental mechanism whereby the Classical Chinese verb *bǎ* 'take' develops into an accusative marker in Mandarin Chinese is described in DEmA as follows "The verb *bǎ* 'take' is reinterpreted as a marker for the shared P argument, and the original biclausal construction is reanalyzed as a monoclausal construction 'ACC X VERB'."

2. **Intermediate stages**: This is an optional field that is used in case the historical scenario can be described as unfolding in a number of distinct steps.

In some cases, for example, a new alignment pattern initially develops in particular constructions, and is subsequently extended to other constructions. A case in point is the development of a new split intransitive system in Series II verbs in Georgian. As discussed by Harris (2010: 213-216), these verbs originally had ergative alignment, but later developed a split intransitive pattern. This process started from transitive constructions with light (semantically generic) verbs such as 'do, make' and an incorporated object. These constructions were reinterpreted as intransitive ones, e.g. 'gave a shout > shouted', as in (5). In the resulting intransitive construction, the S arguments maintains the same marking of the A argument from which it is derived, leading to an accusative pattern initially restricted to the verbs that were derived in this way. A second step in the process was the extension of this pattern to all active intransitive verbs in Series II. As other intransitive verbs in the series maintained ergative alignment, this gave rise to a split intransitive pattern.

(5) Georgian (Kartvelian; Harris 2010: 215)

gagad-q'o	*q'ovel-man*	*er-man*
shout-make	all-ERG	people-ERG

'All the people shouted, gave a shout.'

3. **Type of change**: This field provides a typological classification of different types of developmental mechanisms.

While this classification involves abstracting away from the details of individual processes of change (for which the user is referred to the relevant sources), it aims to relate these processes to the general mechanisms of change traditionally discussed in grammaticalization studies and historical linguistics. We identify five main types of change (note that multiple such mechanisms may be at play for individual types of change): grammaticalization, reinterpretation of argument structure, extension, phonological change, loss.

A. *Grammaticalization*: An element not originally used to encode grammatical relations (e.g. a verb form, a demonstrative, a topic marker) grammaticalizes into a marker for A, S, or P arguments (Lehmann 2015).

An example of this change is the development of an accusative marker from a 'take' verb in Mandarin Chinese, as described above in (1) and (2). In this case, the grammaticalization of the 'take' verb into a direct object marker leads to the development of dedicated marking for P arguments, whereas A and S arguments remain undifferentiated, yielding an accusative pattern.

B. *Reinterpretation of argument structure*: A new alignment pattern emerges through the reinterpretation of the argument structure of the source construction.

This type of change, which has also been described as *reanalysis* (Harris and Campbell 1995: Chap. 4; De Smet 2009), is illustrated by Hanis Coos. In this language, an ergative marker *x̣=* is derived from an instrumental marker. Mithun (2005) submits that this is a result of a reinterpretation processes that took place in two types of constructions: passive sentences with 1st/2nd person P and a 3rd person oblique A marked with *x̣=*, as in (6)a, and transitive sentences with an instrumental NP likewise marked with *x̣=* and no overt 3rd person A, as in in (6)b. Passive constructions such as (6)a are the only possible strategy to encode combination of 1st/2nd person P and 3rd person A in the language. As a consequence, the distinction between active and passive is blurred in these contexts, so that the passive construction can be reinterpreted as a transitive construction with the oblique agent becoming an A argument. Similarly, given the lack of an overt A argument in (6)b, in this construction the originally instrumental NP can be reinterpreted as an A. In both cases, the reinterpretation of the source constructions leads to a new alignment pattern, in which the original instrumental/oblique marker *x̣=* is reinterpreted as an ergative marker for A arguments, as in (6)c.

(6) Hanis Coos (Coosan; Mithun 2005: 87, 84)

a. *x̣ = lau* *kwanɬ* *tə=n=tsx̣ewé-i:ɬ* *tə=x̣* *hú:mis*
OBL=that_one seems-will that=1SG=kill-PASS that=OBL woman
'I may be killed by that woman.'

b. *k'win-t* *x̣ = miɬ:aqətš*
shoot-TRANS OBL=arrow
'(He) shot at him with an arrow.'

c. *x̣ = yiqántštextbarime:x̣* *mæ* *hanƛ* *eʔkwinai:ɬ*
ERG=last people shall they_see_thee
'The last generation shall see you.'

C. *Extension*: The markers used for particular argument roles are extended to other roles (e.g. from A to S) or the same roles in other contexts (e.g. from the S arguments of particular intransitive verbs to the S arguments of other intransitive verbs).

Consider the case of Bats (Harris 2010: 210-213). In origin, Bats had distinct indexes for 1st/2nd person A and S roles, as in (7)a and (7)b, respectively. Later on, the index for A was also analogically extended to the S of intransitive verbs with A-like properties (possibly as a result of contact with Georgian), leading to the rise of a new accusative pattern for these verbs. However, this extension did not take place with

other intransitive verbs, which retained P-like coding in an ergative pattern. As a result, Bats developed a system of split intransitivity, with S arguments of some verbs coded like A and others like P argument of transitive verbs, as comparison between (7)b and (7)c shows.

(7) Bats (Nakh-Daghestanian; Harris 2010: 212)

a. *p'ay* *b-eyɬ-n-as* *ħo*
kiss.NOM CM-give-AOR-1SG.ERG 2SG.DAT
'I gave you a kiss.' 'I kissed you.'

b. *(so)* *vož-en-sŏ*
1SG.ABS fell-AOR-1SG.ABS
'I fell down, by accident.'

c. *(as)* *daħ* *y-apx-yail-n-as*
1SG.ERG PV CM-undress-AUX-AOR-1SG.ERG
'I took my clothes off.'

D. *Phonological change*: These are cases in which a new alignment pattern emerges as phonological changes lead either to the development of specialized forms for particular argument roles or to the loss of existing specialized forms.

The first scenario is illustrated by Louisiana Creole (Haspelmath and the APiCS Consortium 2013). In origin, pronouns for A, S and P roles were undifferentiated in this language. However, A/S pronouns underwent phonological reduction, possibly on account of their higher frequency. As a consequence, the form of A/S pronouns became different from that of P pronouns, yielding an accusative pattern, as shown in Table 1.

The development of a new alignment pattern through the loss of existing forms for particular argument roles is illustrated by English (Blake 2001: 176-178). In Old English, some inflectional classes of nouns retained a distinction between nominative and accusative case in the singular, the former used for A and S and the latter for P. As shown in Table 2, the distinction was realized differently for distinct noun classes. The distinction between nominative and accusative cases was disrupted by two phonological changes. On the one hand, unstressed vowels were reduced to schwa, so that NOM *talu* and ACC *tale* both became /'talə/. On the other hand, word final *-n*

Table 1 Pronominal declension in Louisiana Creole French

Person	A, S	P
1SG	*mo*	*mwa*
2SG	*to*	*twa*

Table 2 Core case marking in Old English

Case	'name'	'tale'
NOM	*nama*	*talu*
ACC	*naman*	*tale*

was lost, so that ACC *naman* became identical to NOM *nama*. The result of the loss of case distinction was the emergence of a new neuter alignment pattern for nouns.

E. *Loss*: This refers to cases where an existing marker for some argument role was lost in the language, but there is no clear evidence that this was due to phonological change.

The emergence of a new alignment pattern as a consequence of loss has been discussed for Tākestāni, a Tāti dialect. Like many modern Indo-Iranian languages, Tāti dialects feature a TAM-based alignment split. In the past tense, argument roles are arranged ergatively: A arguments receive dedicated ergative marking, while S and P arguments are unmarked and are indexed on the verb, as in (8)a-b. In addition, A arguments may, under certain conditions, also trigger the occurrence of A-indexing clitics.

(8) Eshtehārdi (Tāti dialect) (Indo-European; Rasekh-Mahand and Izadifar 2016: 141; Yarshater 1969: 230)

i. *Maryam-ā* *Hasan* *beza(d)*
Maryam(F)-ERG Hasan(M) hit.PST.3SG.M
'Maryam hit Hasan.'

ii. *bābā-š* *bemárda*
father(M)-3SG.POSS.M die.PST.3SG.M
'His father has died.'

In Tākestāni, past transitive constructions have undergone several changes that have led to the emergence of a new alignment pattern. These changes are partly due to loss. In particular, ergative case marking for A and verbal indexes for P were lost, as shown by the comparison between (8)a and (9)b. As a result, past tense transitive constructions show a new tripartite alignment pattern (Rasekhahand and Izdifar 2016 for discussion): S is the only argument that triggers agreement with the verb, P is the only available host for A-clitics, and A triggers the use of A-clitics. The pattern is shown in (9)a-b.

(9) Tākestāni (Indo-European; Rasekh-Mahand and Izadifar 2016: 148)

a. *ā* *ketāb* *xeyli* *sext* *ve*
that.M book(M) very hard be.PST.3SG.M
'That book was very hard.'

b. *a* *jā* *ketāb=em* *bo*
1SG that.OBL book=1SG bring.PST
'I brought that book.'

1.3. The effects of the process of change

For this domain, a number of fields are provided that describe the effects of the process of change leading to the development of the new alignment pattern:

1. **Resulting construction**: This field is similar to the 'Source construction' field in that it features a description of the construction resulting from the process of change.

For example, the reinterpretation of the 'take' verb construction in Mandarin Chinese illustrated in (2) above yields a transitive construction with a P argument overtly marked by *bǎ*.

2. **Alignment in the resulting construction**: This field reports the alignment pattern in the construction resulting from the process of change.

For example, if an intransitive resultative construction of the type 'X is VERBed by Y' is reinterpreted as a transitive one 'Y VERBed X', as is the case of Hanis Coos in (6), this will give rise to ergative alignment, because X becomes a P argument and is encoded in the same way as the S argument from which is derived, whereas Y becomes an A argument with dedicated marking, because it retains the marking used for the oblique NP from which it is derived.

3. **Global alignment pattern following the change**: This field describes the global alignment pattern resulting from the combination of (i) the new alignment pattern of the construction resulting from the change and (ii) the alignment pattern of other constructions within the same grammatical domain.

For example, some processes of change may give rise to new perfective constructions with ergative alignment. If non-perfective constructions have other alignment patterns, however, the language will end up with a TAM-based alignment split, rather than a global ergative alignment pattern, as discussed for Hindi in (4).

Another example showing why it is useful to distinguish between alignment in the resulting construction and global alignment pattern following the change comes from Galela (Holton 2008). This language originally had nominative alignment in indexation. A new alignment pattern as a result of the reinterpretation of intransitive constructions with third person non-human indefinite A arguments and experiencer P arguments indexes, as in (10)a. In these constructions, the indexes for A arguments were progressively lost, and the construction was reinterpreted as an intransitive one, e.g. 'something angers her' > 'she is angry'. As a result, the original P index was reinterpreted as an S index, as shown in (10)b.

(10) Galela (North Halmahera; Holton 2008: 272)

a. *i-mi-tosa*
3SG.A.NONHUM-3F.SG.P-angry
'Something makes her angry'

b. *mi-pereki*
3F.SG.P-old
'She is old'

This change led to the emergence of a new ergative alignment pattern for the relevant intransitive verbs. This is shown by examples (11)a-b, where the same index *ni-* is used for S and P argument as opposed to a distinct A index *wo-*. As the S arguments of other intransitive verbs retains A-like marking, however, at a global level the process results into split-intransitivity.

(11) Galela (North Halmahera; Holton 2008: 261)

a. *ni-kiolo*
2SG.P-asleep
'You are asleep'

b. *wo-ni-doto*
3M.SG.A-2SG.P-teach
'He teaches you'

4. **Constraints**: This field is optional and provides further specification about possible distributional restrictions for the alignment splits resulting from the process of change.

If there is a TAM or NP based split, for example, the field will specify the exact properties of the split (e.g. perfective constructions vs. non-perfective ones, pronouns vs. nouns, inanimate nouns vs. other NP types).

5. **Grammatical domain**: This refers to the grammatical domain involved in the process of change, for example case marking, indexation, or word order.

Particular processes of change may involve multiple grammatical domains, e.g. both case marking and indexation. An example is Tākestāni in (9) where the emergence of a new alignment pattern is the result of the loss of both ergative case marking and verbal agreement.

6. **Symmetry**: This refers to the morphosyntactic encoding of argument roles in the construction resulting from the change.

Symmetric encoding means that all roles are encoded though the same strategy (e.g. overt case marking, overt indexation, whereas asymmetric encoding means that different roles are encoded through different strategies (zero vs. overt case marking, zero vs. overt marking in indexation).

An example of asymmetric marking is accusative alignment in Mandarin Chinese in (2): A and S roles are unmarked whereas P receives overt marking by means of *bǎ*. Symmetric marking can be found in case marking in Modern English, in which A, S and P are all equally unmarked (see discussion of the data in Table 2), and in the indexing pattern of Tobelo in (12), where all roles are variously marked by indexation on the verb.

(12) Tobelo (North Halmahera; Holton 2003: 22)

a. *to-ni-gohara*
1SG.NOM-2SG.ACC-hit
'I hit you'

b. *to-tagi*
1SG.NOM-go
'I go.'

4. *How to use DEmA*

DEmA allows for fine-grained searches of the various components involved in the emergence of alignment patterns. Users can browse data in DEmA in two ways.

1. **By language:** the full list of languages included in DEmA is provided in the **Languages** section, as shown in Figure 2. By clicking on each entry, users can visualize all the fields with the relevant information on the emergence of a new alignment pattern in that specific language.

2. **By field:** our Search engine allows for queries on various fields, as shown in Figure 3. Users can simultaneously combine queries for multiple fields. Fields are divided into two categories based on the type of query parameter that they allow:

a. *Free text query*: users can freely enter their textual query in these fields (these are e.g. 'Language', 'Source Construction', 'Constraints').

An important free text query field is the **Keywords** field. Each Language is characterized by a number of keywords. These are intended as generic shortcuts for the various aspects of the historical process described in each entry and are meant to reflect the terminology most commonly used in the literature to refer to that specif-

Figure 2 The DEmA Languages interface

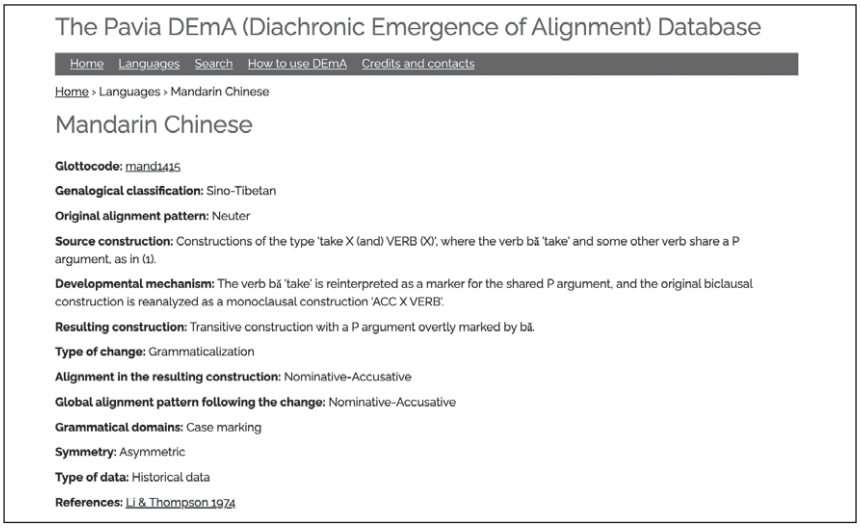
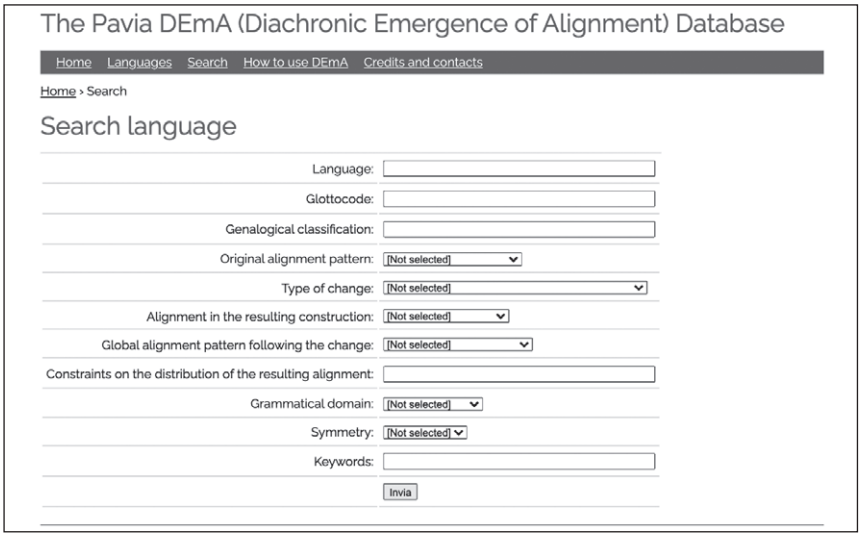


Figure 3 The DEmA Search interface



ic process. Possible keywords include, for example, 'ergative', 'split ergativity', 'nominalization', 'passive', 'resultative construction'.

b. *Selectable option query*: users can select one of the pre-existing options (e.g. 'Alignment in the resulting construction' features only a few options, such as Nominative-Accusative and Ergative-Absolutive).

4. *Conclusions*

In this paper, we have offered an overview of the structure of the Pavia *Diachronic Emergence of Alignment* (DEmA) database. The database will be hosted by the University of Pavia, and will be available together with other linguistic resources developed at the Section of Theoretical and Applied Linguistics through the *The Pavia linguistic resources repository*.[3] Once released, the database will be fully searchable, allowing users to query the database for all parameters and combinations thereof. The database is also expandable, and we encourage scholars working on the diachrony of alignment to make their data available through DEmA.

At a more general level, the architecture of DEmA is unique in that it offers a theoretically well-grounded and explicit systematization of several parameters pertaining to language change (e.g. source constructions, type of change, type of data), so that these can be effectively implemented into a searchable format. In this respect, we hope that DEmA will also provide a suitable model for future typological resources dealing with the diachrony of other grammatical domains.

Abbreviations

1 = first person; 2 = second person, 3 = third person, A = agent, ABS = absolutive, ACC = accusative, AOR = aorist, ASP = aspect, AUX = auxiliary, CM = (gender-)class marker, DAT = dative, ERG = ergative, F = feminine gender, INS = instrumental, M = masculine gender, MID = middle voice, NOM = nominative, NONHUM = non-human, OBL = oblique, P = patient, PASS = passive, PL = plural, POSS = possessive, PPP = perfect passive participle, PRF = perfect, PRS = present, PST = past, PTCL = particle, PV = preverb, SG = singular, TRANS = transitive

Websites

The Pavia Linguistic Repository: https://su-lab.unipv.it/tasf/

References

Blake, Barry J. 2001. *Case*. 2nd edition. Cambridge: Cambridge University Press.
Bubenik, Vit. 1998. *A historical syntax of late middle Indo-Aryan (Apabrahmśa)*. Amsterdam: John Benjamins.

3. The DEmA interface has been built by Dr. Alessio Palmero Aprosio, whom we thank for his technical assistance.

Comrie, Bernard. 1989. *Language universals and linguistic typology.* 2nd edition. Oxford: Basil Blackwell.

Comrie, Bernard. 2013a. Alignment of Case Marking of Full Noun Phrases. In *The World Atlas of Language Structures Online*, Matthew S. Dryer & Martin Haspelmath (eds). Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/98> (2021-05-19)

Comrie, Bernard. 2013b. Alignment of Case Marking of Pronouns. In *The World Atlas of Language Structures Online*, Matthew S. Dryer & Martin Haspelmath (eds). Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/99> (2021-05-19)

Creissels, Denis. 2008. Direct and indirect explanations of typological regularities: The case of alignment variations. *Folia Linguistica* 42: 1–38.

Cristofaro, Sonia. 2012. Cognitive explanations, distributional evidence, and diachrony. *Studies in Language* 36: 645–670.

Cristofaro, Sonia. 2013. The referential hierarchy: Reviewing the evidence in diachronic perspective. In *Languages across Boundaries: Studies in the Memory of Anna Siewierska*, Dik Bakker & Martin Haspelmath (eds), 69–93. Berlin & New York: Mouton de Gruyter.

Cristofaro, Sonia. 2014. Competing motivations and diachrony: What evidence for what motivations? In *Competing motivations in grammar and usage*, Brian MacWhinney, Andrej Malchukov & Edith Moravcsik (eds), 282–98. Oxford: Oxford University Press.

Cristofaro, Sonia. 2019. Taking diachronic evidence seriously: Result-oriented vs. source oriented explanations of typological universals. In *Explanation in typology*, Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis & Ilja A. Seržant (eds), 25–46. Berlin: Language Science Press.

Dahl, Eystein & Stroński, Krzysztof. 2016. Ergativity in Indo-Aryan and beyond. In *Indo-Aryan Ergativity in Typological and Diachronic Perspective*, Eystein Dahl & Krzysztof Stroński (eds), 1-37. Amsterdam: John Benjamins.

De Smet, Hendrik. 2009. Analysing reanalysis. *Lingua* 119(11): 1728-1755.

Dixon, R. M. W. 1994. *Ergativity*. Cambridge: Cambridge University Press.

Gildea, Spike. 1998. *On reconstructing grammar: Comparative Cariban morphosyntax*. Oxford: Oxford University Press.

Grossman, Eitan & Polis, Stéphane. 2018. Swimming against the typological tide or paddling along with language change? Dispreferred structures and diachronic biases in affix ordering. *Journal of Historical Linguistics* 8(3): 388–443.

Haig, Geoffrey. 2008. *Alignment Change in Iranian Languages: A Construction Grammar Approach*. Berlin & New York: Mouton de Gruyter.

Haig, Geoffrey. 2017. Deconstructing Iranian Ergativity. In *The Oxford Handbook of Ergativity*, Jessica Coon, Diane Massam & Lisa D. Travis (eds), 465–500. Oxford: Oxford University Press.

Harris, Alice C. 2010. Origins of Differential Unaccusative/Unergative Case Marking: Implications for Innateness. <http://works.bepress.com/alice_harris/11/> (2021-05-19).

Harris, Alice C. & Campbell, Lyle. 1995. *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.

Haspelmath, Martin. 2019. Can cross-linguistic regularities be explained by constraints on change? In *Explanation in typology*, Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis & Ilja A. Seržant (eds), 1–23. Berlin: Language Science Press.

Haspelmath, Martin & the APiCS Consortium. 2013. Alignment of case marking of personal pronouns. In *Atlas of Pidgin and Creole Language Structures Online*, Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds). Oxford: Oxford University Press.

Heine, Bernd & Kuteva, Tania. 2002. *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Holton, Gary. 2003. *Tobelo*. Munich: Lincom Europa.

Holton, Gary. 2008. The rise and fall of semantic alignment in Northern Halmahera, Indonesia. In *The typology of semantic alignment*, Mark Donohue & Søren Wichmann (eds), 252–276. Oxford: Oxford University Press.

König, Christa. 2008. *Case in Africa*. Oxford: Oxford University Press.

Kulikov, Leonid. 2006. Case systems in a diachronic perspective. In *Case, valency and transitivity*, Leonid Kulikov, Andrej Malchukov & Peter de Swart (eds), 23–47. Amsterdam: John Benjamins.

Kuteva, Tania, Heine, Bernd, Hong, Bo, Long, Haiping, Narrog, Heiko & Rhee, Seongha. 2019. *World Lexicon of Grammaticalization*. 2nd edition. Cambridge: Cambridge University Press.

Lehmann, Christian. 2015. *Thoughts on grammaticalization.* 3rd edition. Berlin: Language Science Press.

Li, Charles N. & Thompson, Sandra A. 1974. An explanation of word order change SVO→ SOV. *Foundations of Language* 12(2): 201–214.

Mithun, Marianne. 2005. Ergativity and language contact on the Oregon Coast: Alsea, Siuslaw, and Coosan. In *Proceedings of the Berkeley Linguistic Society* 26. *Special session on syntax and semantics of the indigenous languages of the Americas*, Andrew K. Simpson (ed), 77–95. Berkeley (CA): Berkeley Linguistic Society.

Rasekh-Mahand, Mohammad & Izadifar, Rahele. 2016. Compensating ergative alignment loss in Takestani. In *Further topics in Iranian linguistics, Proceedings of the 5th international conference on Iranian linguistics, held in Babmberg on 24-26 August 2013*, Jila Ghomesh, Carina Jahani & Agnès Lenepveu-Hotz (eds), 135–155. Leuven: Peeters.

Siewierska, Anna. 2013. Alignment of Verbal Person Marking. In *The World Atlas of Language Structures Online*, Matthew S. Dryer & Martin Haspelmath (eds). Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/100> (2021-05-19)

Verbeke, Saartje. 2013. *Alignment and Ergativity in New Indo-Aryan Languages*. Berlin & New York: Mouton de Gruyter.

Verbeke, Saartje & De Cuypere, Ludovic. 2009. The rise of ergativity in Hindi: Assessing the role of grammaticalization. *Folia Linguistica Historica* 43: 367-389.

Yarshater, Ehsan. 1969. *A Grammar of Southern Tati Dialects*. The Hague & Paris: Mouton.

Zúñiga, Fernando. 2018. The diachrony of morphosyntactic alignment. *Language and Linguistic Compass* 12: e12300.

Contributors

Erica Biagetti

Università di Pavia
Dipartimento di Studi Umanistici
Sezione di Linguistica Teorica e Applicata

Strada Nuova 65
I-27100 Pavia, Italy

erica.biagettio1@universitadipavia.it

Gerd Carling

Lund University
The joint Faculties of Humanities
and Theology

Helgonabacken 12,
221 00 Lund, Sweden

gerd.carling@ling.lu.se

Sonia Cristofaro

Sorbonne Université
Faculté des Lettres
Laboratoire STIH

1 rue Victor Cousin
75005 Paris, France

sonia.cristofaro@sorbonne-universite.fr

Hanne Martine Eckhoff

Lady Margaret Hall
University of Oxford

Norham Gardens
Oxford, OX2 6QA, United Kingdom

hanne.eckhoff@mod-langs.ox.ac.uk

Dag T. T. Haug

University of Oslo
Department of Philosophy, Classics,
History of Art and Ideas

Blindernveien 31
Georg Morgenstiernes hus
0313 Oslo, Norway

d.t.t.haug@ifikk.uio.no

Oliver Hellwig

Heinrich Heine University Düsseldorf,
Institute for Language and Information
University of Zurich, Department of
Comparative Language Science

Thurgauerstrasse 30
CH-8050 Zurich, Switzerland

hellwig7@gmx.de

Guglielmo Inglese

KU Leuven
Faculty of Arts

Blijde-Inkomststraat 21
3000 Leuven, Belgium
guglielmo.inglese@kuleuven.be

Filip Larsson

Lund University
The joint Faculties of Humanities
and Theology

Helgonabacken 12,
221 00 Lund, Sweden

filip.larsson@ling.lu.se

Olof Lundgren

Lund University
The joint Faculties of Humanities and Theology

Helgonabacken 12,
221 00 Lund, Sweden

lundgren8@gmail.com

Silvia Luraghi

Università di Pavia
Dipartimento di Studi Umanistici
Sezione di Linguistica Teorica e Applicata

Strada Nuova 65
I-27100 Pavia, Italy

silvia.luraghi@unipv.it

Francesco Mambrini

Università Cattolica del Sacro Cuore
Dipartimento di Scienze linguistiche e Letterature straniere
CIRCSE Research Centre

Largo Gemelli, 1
20123 Milan, Italy

francesco.mambrini@unicatt.it

Linus Nilsson

Lund University
The joint Faculties of Humanities and Theology

Helgonabacken 12,
221 00 Lund, Sweden

linus.nilsson@ling.lu.se

Marco C. Passarotti

Università Cattolica del Sacro Cuore
Dipartimento di Scienze linguistiche e Letterature straniere
CIRCSE Research Centre

Largo Gemelli, 1
20123 Milan, Italy

marco.passarotti@unicatt.it

Peter M. Scharf

President, Sanskrit Library
Providence, Rhode Island, USA

scharf@sanskritlibrary.org

Sven Sellmer

Heinrich Heine University Düsseldorf,
Institute for Language and Information
Adam Mickiewicz University in Poznań,
Faculty of Modern Languages and Literatures, Institute of Oriental Studies

ul. Grunwaldzka 6
61-874 Poznań, Poland

sven@amu.edu.pl

William M. Short

University of Exeter
Department of Classics & Ancient History
The Old Library
Prince of Wales Road

Exeter, Devon
EX4 4SB, United Kingdom

W.Short@exeter.ac.uk

Rob Verhoeven

Lund University
The joint Faculties of Humanities and Theology

Helgonabacken 12,
221 00 Lund, Sweden

rob.verhoeven@ling.lu.se

Chiara Zanchi

Università di Pavia
Dipartimento di Studi Umanistici
Sezione di Linguistica Teorica e Applicata

Strada Nuova 65
I-27100 Pavia, Italy

chiara.zanchio1@unipv.it

This volume collects the papers originally presented at the workshop *Building New Resources for Historical Linguistics*, held at the University of Pavia in November 2020. The purpose of this workshop was to provide an opportunity for researchers engaged in the development of linguistic resources for historical linguistics to share their experience and knowledge. Sharing is crucial in computational linguistics in order to avoid multiplying efforts and encourage the use of compatible tools, formats, and formalisms to increase the interoperability. Reflecting the purpose of the original workshop, this book introduces the reader to different projects aimed at creating, developing and linking linguistic resources for historical linguistics. While some of the papers in the volume describe mature resources and discuss their possible application, others introduce resources that are still in progress, presenting their aims, the challenges faced in their construction, and the methodologies employed to tackle them. The different types of resources described in the volume include syntactically annotated corpora (*treebanks*), dependency lexica, as well as lexical and typological databases. Furthermore, some of the papers are concerned with the thriving field of Linguistic Linked Open Data, the current up-to-date standard to link linguistic resources.

Erica Biagetti is postdoctoral researcher at the University of Pavia, where she works on the creation of two lexical databases (*WordNets*) for Ancient Greek and Sanskrit. Her research interests include Indo-European linguistics with a focus on Sanskrit, computational linguistics, and digital humanities.

Chiara Zanchi is assistant professor at the University of Pavia, where she teaches Linguistic data analysis lab. Her research interests include Indo-European linguistics with a focus on Ancient Greek, pragmatics of both ancient and modern Indo-European languages, and cognitive linguistics.

Silvia Luraghi is professor of linguistics at the University of Pavia. Her research interests include language change in a usage-based perspective, language reconstruction, Indo-European linguistics, language typology, and cognitive linguistics.

www.paviauniversitypress.it