The DiGreC Treebank

1. Introduction

The DiGreC (DIachrony of GREek Case) treebank[1] has been created as part of the project "Investigating Variation and Change: Case in Diachrony", funded by the Arts & Humanities Research Council (AH/P006612/1). The goal of this project has been to use the Greek language, which furnishes a large quantity of linguistic data over an unusually long span of time, to investigate syntactic phenomena, and to provide a clearer picture of the Greek case system and its changes over time, which has the potential to inform theoretical discussions on the nature of linguistic case. We have chosen to make the data used in this project available to the public in the form of a morphosyntactically and semantically annotated treebank. This article describes the features of this treebank, as well as the data selection principles and methodology involved in its construction.

2. Context

The role of case in grammar has been studied from two different perspectives, each of which has its own need for data. Theoretical discussions of case (e.g. Baker 2015) model the interactions between case and other components of grammar, and make predictions about the types of construction that would be possible in any language; for languages without living native speakers these predictions can only be tested through the exhaustive analysis of large quantities of data. Descriptive grammars of individual languages attempt to categorise constructions and provide guidance on which types of constructions are grammatical; however, for a language such as Ancient Greek, many grammars (e.g. Smyth 1920) predate the development of modern corpora and focus on a relatively small body of literary texts as their data source.

One of our aims has been to provide enough data for the critical evaluation of previous work in both these traditions; however, this requires a data set different from what has been available from existing electronic resources. Large-scale resources such as the *Perseus Digital Library* (Crane, 2020) and TLG (*Thesaurus Linguae Graecae*, Pantelia, 2020) provide extensive quantities of data over a long span of time, but their size makes detailed syntactic annotation impracticable. Conversely, syntactically annotated resources such as the Ancient Greek Dependency Treebank (Bamman and Crane, 2011, http://perseusdl.github.io/treebank_data) and the PROIEL treebank (Haug and Jøhndal, 2008, https://proiel.github.io) comprise much smaller quantities of data; as a result, they may be unsuitable for research involving long-term diachronic analysis and the study of relatively infrequent constructions.

In designing DiGreC, we have struck a balance between these extremes by adopting a 'verb-sensitive' approach of the sort used in corpus-based studies such as Stolk (2017). Rather than attempting to include entire texts, the corpus includes only passages containing selected verbs. This provides a manageable quantity of data, permitting manual review and detailed annotation of the sort described below, while allowing the data set to cover a much broader span of time than would otherwise be possible.

DiGreC also provides semantic annotation of a sort unavailable through most existing resources. Our research makes reference both to morphosyntactic properties and to semantic features such as animacy, as it has been hypothesised that animacy may play a role in verbs' argument selection. DiGreC was designed to allow the searchable tagging of tokens for animacy independently of their other attributes.

3. Methods

As described above, a verb-sensitive approach was employed in selecting data for the treebank. While our focus was on case, simply searching for case-marked nominals would have been impracticable given the total quantity of data, even if this feature were supported in all the existing corpora. As our research questions were primarily on the structure and properties of datives and genitives (as the locus of most diachronic change in the history of Greek), we decided to use the verb-sensitive approach to search for argumental datives and genitives selected by particular verbs. We compiled a list of verbs whose syntactic behaviour was most likely to be of interest for this project, starting from the classifications found in traditional Greek grammars (e.g. Goodwin 1894; Smyth 1920; Tzartzanos 1940), and from the Greek equivalents of verbs listed in semantic classifications such as Levin (1993).[2] Searches were then conducted for forms of these verbs in existing resources, including Perseus and TLG; to include data from as many styles and registers as possible, searches were also conducted using the Papyrological Navigator (Duke University, 2020) and the Packard Humanities Institute's epigraphic database (2020). We are grateful to the University of California, Irvine, for permission to reproduce data from TLG. Data from other sources have been included subject to the terms on which they were originally made available to the public, as described in the treebank documentation. Where automated lemma-based searching was available, this was used to obtain a list of results for all forms of a given lemma; for resources such as the Packard epigraphic database, which does not provide lemmatized data, wildcard searching was used instead to find forms of relevant verb stems. The results of these automated searches, which for many verbs would return thousands or tens of thousands of hits, were then subjected to manual review.

The manual review performed after searching was used to determine which of the examples identified should be included in the corpus. As our methodology is not quantitative, no attempt was made to provide a data set of the size necessary for quantitative analysis; moreover, to provide data that could be used directly for such analysis, without the need for techniques such as weighting, it would be necessary to control for variables such as date, genre and register, which in turn would involve limiting the size of overrepresented categories and thus excluding potentially valuable data. Instead, samples were chosen to provide a representative overview of the constructions in which a verb could occur. Although the treebank may include all, or almost all, attestations of rare verbs, for high frequency verbs, where there are often a large number of syntactically parallel examples, only a subset of examples have been chosen. During the manual selection of these examples, we have tried to illustrate as fully as possible the different cases and case combinations occurring with a given verb, and to exemplify how the verb's behaviour differs in different voices, showing which arguments can become the subject of a passive construction and what other differences, if any, exist between active and passive constructions.[3] This selection involved the manual classification of examples on the basis of morphological features such as voice, a process which in the case of TLG was facilitated by the option to group results automatically by verb form; for each construction type, a number of examples were then chosen, with the aim of providing enough data to be representative while minimizing redundancy and keeping the size manageable. Where possible, we have given priority to early attestations and to those with the least potential ambiguity in their syntactic structure, although we have also included late examples where diachronic change or continuity is relevant to the phenomena under study. Our approach to the selection of examples has been geared towards the study of phenomena that can be described in binary terms, such as grammaticality. If a construction can be found in the treebank, this shows that it was attested in natural language; if it does not occur, this indicates that we could not find any such examples anywhere in the data sources described above. Table 1 includes a list of verbs for which such exhaustive verb-sensitive searches are currently represented. Although the distribution of other verbs has not been reflected in the data-selection process, many other verbs of course occur in the treebank data.

Table 1: Verbs used for verb-sensitive searching

Once text samples were selected for inclusion, they were subjected to automatic morphological tagging and lemmatisation. The tagger used for this work was TnT (Brants 1998), which was trained on tagged data from PROIEL; as described below, the data formats for this project build upon those introduced by PROIEL. The full Greek dataset was used, comprising an excerpt from Herodotus, the New Testament, and Sphrantzes' *Chronicles*; this was converted to a list of words tagged with the combined part-of-speech and morphological information from PROIEL. After training and testing, the accuracy rate was found to be approximately 75%; accordingly, manual correction was performed on the data at a later stage in order to improve accuracy further.

For lemmatisation, the Morpheus program was used (Crane 1991, https://github.com/alpheios-project/morpheus). This tool was designed to make use of a lexical database of Ancient Greek and to base its lemmatisation on an analysis of the morphological structure of a word; although Morpheus can also be used for morphological tagging, TnT was preferred for this purpose because of its greater flexibility. In contrast to the PROIEL convention, homographs belonging to the same part of speech were not distinguished in any way. It was decided that the number of potentially problematic lexemes is quite small, and that distinguishing, e.g., *δέω* ‘bind’ and *δέω* ‘lack’ as *δέω#1* and *δέω#2* would be of little help for searching unless the user knew in advance which number was associated with the meaning desired.

Next, syntactic annotation was performed, using the dependency-grammar notation employed in the PROIEL corpus. This format is compatible with existing tools, such as the converters and visualisers described below. However, it also allows the relations among constituents in a sentence to be described in a relatively abstract, theory-neutral manner, without commitment to the underlying cognitive basis of the structures depicted. One of the aims of this research project has been to use the data collected to formulate a more accurate representation of these structures, from a generative perspective, than currently exists; however, this corpus was intended to function as a starting point from which such representations could be pursued as a goal.

4. Data

The treebank data exist in three distinct formats: as a single XML file, an alternative CSV version, and as a web-based interface to a relational database generated on the server from XML input. All these formats include the same basic data, comprising excerpts from 655 texts, for a total of 3385 sentences and 56,440 word tokens. The total time span represented ranges from the Homeric epics (c. 8th century BC) to early Modern Greek authors such as Theodosius Zygomalas (17th century AD). Each text has metadata including its identifier in the TLG cataloguing system, or an equivalent identifier for papyri and inscriptions; the author; the title; the approximate date of composition; and a URL for the original source of the text. Where a searchable version of a text is provided only by the TLG but a PDF copy of a public-domain edition is readily available, we have in some cases provided a reference to the latter for the convenience of users.

The XML file has been deposited in Ulster University's institutional repository, hosted on the Elsevier PURE system (https://pure.ulster.ac.uk/files/87540132/digrec.xml). The dataset has also been assigned a persistent DOI linking to its location in this repository (https://doi.org/10.21251/59fd3210-83fe-4d1c-8d18-f2cd1168ccd6). The XML data format uses the PROIEL 2.0 schema, making it interoperable with existing tools. Not only are there interfaces designed specifically for this format, such as Syntacticus, but the PROIEL project provides tools for converting this schema to a number of other formats, including CoNLL, TigerXML, and Tiger2. Through the use of such tools, our data can be used with a range of other corpus systems.

In the XML file, annotation is associated with tokens as attributes. The annotations for each token indicate its lemma, part of speech, morphological features, and syntactic dependencies. In keeping with the PROIEL specification, Greek text is stored in UTF-8 format using Unicode Normalization Form C; for our purposes, this form differs from other Unicode forms primarily in that characters with an acute accent as the sole diacritic are stored using the Modern Greek ‘tonos’ codes (e.g. ά = 03AC) rather than the polytonic Greek ‘oxia’ codes (e.g. ά = 1F71).

Manual semantic annotation has also been added, to categorise forms as animate, inanimate, or ‘propositional’. This last category is used for forms such as infinitives and clauses that refer to propositions rather than entities, and allows hypotheses to be tested in which the two classes behave differently. In the XML file, semantic tags are represented as attributes on token elements, in accordance with the informal schema extension used by the original PROIEL tools. The primary principle on which the tagging is based has been to add one tag for each referential expression. Accordingly, tags have been added to all nouns and to independent adjectives, but not to adjectives modifying a noun either attributively or as a predicate; infinitives and subordinate clauses have been given the ‘propositional’ tag only when they refer to propositions that form arguments of other verbs, but not in other constructions such as infinitives of purpose.

We have also set up a web site, located at http://cid.ulster.ac.uk, to provide a user interface for working with the data directly. This site is based on the PROIEL web application, but has been extensively modified and customised to optimise it for our data set; for example, the DiGreC site includes new functionality for working with animacy, instead of the PROIEL interfaces relating to information structure. The site allows searching for tokens based on morphological, syntactic, and semantic annotation, singly or in any combination; it also displays the syntactic annotation in a graphical, tree-based format. Although the underlying data format is UTF-8, it is possible to search for text using either Greek text or BetaCode; in both these formats, accents will be ignored, so that variation in accentuation will not prevent the identification of relevant forms. As the figures show, the web site displays data in a format which is more readily human-readable than the original XML. The code for the site is available at https://github.com/mdm33/digrec, and is distributed under the GNU General Public License version 2.

Figure 1: Sample XML representation

Figure 2: Web visualisation[4]

The GitHub site also contains the CSV data files. These provide the same data in a format similar to that used by the relational database on the server; however, the CSV version has been slightly simplified, to reduce the number of separate tables. Most of these tables have been combined into a single file, tokens.csv; however, additional files are used for the index of texts (sources.csv) and details of the ‘slash notation’ used for certain syntactic relationships such as subject/predicate (slashes.csv). With these files is included an up-to-date list of the verbs exhaustively represented in the corpus. We are grateful to a reviewer for the suggestion to make this material available.

5. Conclusion

The DiGreC treebank represents an attempt to make the data from our project accessible to and reusable by other researchers. As described above, this treebank provides syntactically and semantically annotated data from a more diverse range of texts, over a broader time span, than many existing resources. Although it does not exhaustively represent the full surviving body of Ancient Greek texts, it can be used by researchers seeking examples of specific constructions, for research not only on those aspects of grammar on which we have focused but on the many other phenomena which our data embody (e.g. tense, aspect, modality, number). In addition, this resource will continue to evolve; we will expand the data in the treebank to increase the number of verbs exhaustively represented, as we investigate outstanding questions such as the role of prefixes and prepositions in assigning case. We hope that such a resource will be of lasting value to many others in the field of linguistics.

Notes

1. As described below, the syntactic trees represented in this corpus are based on the dependency grammar format used in corpora such as the PROIEL Treebank, rather than the generative-style format used in corpora such as the Penn treebanks.
2. Special attention was given to ditransitive verbs (verbs with both a direct and an indirect object), owing to the existence of previous studies such as Conti Jiménez (1998) focusing on monotrantitives.

3. The existence of alternations such as dative–nominative and genitive–nominative is important for theories that identify structural case based on its participation in such alternations (Chomsky 1986, Vergnaud 1977)
4. For reasons of space, the illustration shows a minimally simple tree. More complex sentences are represented by multi-level trees with multiple links among the different nodes.

References

Baker, M. C. (2015). *Case: Its principles and its parameters*. Cambridge: Cambridge University Press.

Bamman, D., & Crane, G. (2011). The Ancient Greek and Latin Dependency Treebanks. In C. Sporleder, A. van den Bosch, & K. Zervanou (Eds.), *Language Technology for Cultural Heritage: Theory and Applications of Natural Language Processing* (pp. 79–98). Berlin: Springer. https://doi.org/10.1007/978-3-642-20227-8_5

Brants, T. (1998). *TnT — Statistical part-of-speech tagging*. http://www.coli.uni-saarland.de/~thorsten/tnt/

Chomsky, N. (1986). *Knowledge of language*. New York: Praeger.

Conti Jiménez, L. (1998). Zum Passiv von griechischen Verben mit Genitiv bzw. Dativ als zweitem Komplement. *Münchener Studien zur Sprachwissenschaft* 58, 13–50.

Crane, G. (1991). Generating and parsing Classical Greek. *Literary and Linguistic Computing*, 6(4), 243–245. https://doi.org/10.1093/llc/6.4.243

Crane, G., ed. (2020). *Perseus Digital Library*. Medford, MA: Tufts University. http://www.perseus.tufts.edu/

Duke University. (2020). *Papyrological Navigator*. http://papyri.info

Goodwin, W. W. (1894). *A Greek grammar*, rev. edn. London: Macmillan.

Haug, D. T. T., & Jøhndal, M. L. (2008). Creating a parallel treebank of the Old Indo-European Bible Translations. In C. Sporleder & K. Ribarov (Eds.), *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)* (pp. 27–34).

Levin, B. (1993) . *English verb classes and alternations*. Chicago: University of Chicago.

Packard Humanities Institute. (2020). *Searchable Greek Inscriptions*. http://epigraphy.packhum.org

Pantelia, M., ed. (2020). *Thesaurus Linguae Graecae® Digital Library*. Irvine, CA: University of California, Irvine. http://www.tlg.uci.edu/

Smyth, H. W. (1920). *A Greek grammar for colleges*. New York: American Book Company.

Stolk, J. V. (2017). Dative alternation and dative case syncretism in Greek. *Transactions of the Philological Society*, 115(2), 212–238. https://doi.org/10.1111/1467-968X.12098

Tzartzanos, A. A. (1940). *Grammatikē tēs Archaias Hellēnikēs glōssēs*. Athens: Organismos Ekdoseōs Scholikōn Vivliōn.

Vergnaud, J.-R. (1977). Letter to Noam Chomsky and Howard Lasnik on “Filters and control”. In R. Freidin, C. Otero., & M. Zubizaretta (Eds.), *Foundational Issues in Linguistic Theory* (pp. 3–15), Cambridge, MA: MIT Press, 2008.