Pipeline for a data-driven network of linguistic terms

Søren Wichmann
Laboratory of Quantitative Linguistics
Kazan Federal University
wichmannsoeren@gmail.com

Abstract

The present work is aimed at (1) developing a search machine adapted to the large DReaM corpus of linguistic descriptive literature and (2) getting insights into how a data-driven ontology of linguistic terminology might be built. Starting from close to 20,000 text documents from the literature of language descriptions, from documents either born digitally or scanned and OCR'd, we extract keywords and pass them through a pruning pipeline where mainly keywords that can be considered as belonging to linguistic terminology survive. Subsequently we quantify relations among those terms using Normalized Pointwise Mutual Information (NPMI) and use the resulting measures, in conjunction with the Google Page Rank (GPR), to build networks of linguistic terms.

1 Introduction

Linguistics is a discipline rich in terminology. Terminology specific to this domain is needed everywhere from the fine description of individual speech sounds over the categorization of different syntactic constructions to features of language use, and the abundance of terminology stemming from the empirical nature of inquiry itself is compounded by the excess of theoretical approaches, each of which tends to develop its own terminology. Thus, there is no dearth of handbooks of linguistic terms, but they only provide selective glimpses of the vocabulary coming into play when linguists write about languages. Here we take a data-driven (corpus-based) approach to the study of linguistic terminology using a set of 19,761 texts in English that belong to the DReaM corpus of linguistic literature (Virk et al., 2020). These texts consists of full grammars, partial descriptions of certain features, comparative studies, etc. That is, works that describe one or more features of the world's languages. According to the most recent count it spans 4,527 languages (Hammarström et al., 2021). It is important to emphasize that the corpus generally does not include purely theoretical literature. Thus, we are unlikely to come across some term that a theoretician has proposed if its actual usage in descriptions is rare.

This paper has two foci, where the first (1) is the pipeline immediately preceding the harvest of linguistic terms and the second (2) is the analysis of relationships among those terms. As for the first focus (1), we exclude a discussion of all the work that has gone into assembling the corpus, preparing metadata, and running documents through OCR. Instead, we focus on the pipeline for extracting linguistically relevant terms. This pipeline will be presented only summarily, but all steps, both trivial and less trivial ones, will be listed. The second focus (2) is on the relationships among terms. Mapping the relationships between these terms serves two purposes. First, (2a), the online DReaM corpus[1] currently allows for string searches in the available texts. We would like to enhance this functionality with an option to retrieve search results not only for a specific term but also related terms. For instance, in the procedure to be explained below, we find empirically that the term *direct object* is closely related to *indirect object*, and *relative clause* is closely related to *head noun*. A user should be given the option of choosing to include such related terms in a search. Secondly (2b), we want to analyze the network or networks constituted by related terms. A central question here is whether the network(s) can somehow lay the ground for an ontology of linguistic terms.

[1] https://spraakbanken.gu.se/korp/?mode=dream?lang=en

2 Related work

This work pertains to the fields of terminology extraction and automated domain ontology construction. Although the literature in these areas is rich (Medelyan et al., 2013; Qiu et al., 2018; Heylen and Hertog, 2015), it is not the case that an appropriate off-the-shelf tool can be found and applied to the case at hand. Most approaches are directed at cases which are more privileged in terms of the nature of the corpora analyzed. A large proportion of the texts of our sample are replete with OCR errors making the filtering of noise a real issue which is not usually present. Some approaches take recourse to generic resources such as WordNet for establishing concept relations or plugging relations into a wider framework (Navigli and Velardi, 2004; Alrehamy and Walker, 2018). Linguistic terminology, however, is of such a specialized nature that such resources cannot easily be drawn upon. Related to this problem, the common strategy of identifying hypernym-hyponym or is-a relations from texts (Velardi et al., 2004; Alfarone and Davis, 2015) is complicated by the abstract nature of linguistic terminology and the fact that many such relations depends on a particular theoretical framework. For instance, a *subject* can be a kind of topic, argument, position, noun etc. depending on the language, point of view, and theory of grammar. Moreover, such terms are often defined through examples rather than discursively in different grammars. Our approach is minimalist, so we also do not produce a fully POS-tagged corpus as input to term extraction, unlike some other approaches (Bourigault and Jaquemin, 1999).

There seems to be just one published approach similar to ours (Kang et al., 2016). It is similarly a minimalist approach, only relying on the particular corpus of interest. It proceeds from the extraction of terminology to a procedure of relating terms through a vector-based similarity metric. Nevertheless, this approach and ours are only comparable at a general level.

3 Pipeline for term extraction

The following describes the pipeline in numbered steps. Most steps were carried out using R, while a few steps additionally involved Python scripts.

S1. An initial database of text files OCR'ed from linguistic descriptive materials was used. These have been collected and processed by Harald Hammarström over several years (Virk et al., 2020). He also supplied a bibliography file in BibTex style with metadata (henceforth source.bib), which was parsed. The current version of this file is publicly available as part of Glottolog (Hammarström et al., 2020).

S2. When several files were associated with the same bibliographical entry, the besttxt field of source.bib was visited in order to select the best file.

S3. Files tagged in the bibliography as not primarily being grammatical descriptions, but rather lexicographic, ethnographic, etc. works, were removed.

S4. Works having English as the metalanguage (i.e., works written in English, although typically describing some other language) were singled out. Documents using a metalanguage other than English were removed.

S5. All lines having characteristics of something other than running text (tables, lists, short headings, bibliographical entries, etc.) were removed. A machine learning system for recognizing bibliographical entries is under development, but was not actually applied. Remaining lines were concatenated in a single line and subsequently split into sequences delimited by a full stop—in most cases representing sentences, but best described neutrally as 'chunks'. They were then put in a single file, collected.txt.

S6. Another file was created with two columns: one having numbers representing the sentence number in collected.txt and another having the file names. Thus, numbers indexing terms remain cross-referenced with the document where they occur.

S7. Since in linguistics, as in so many other domains, terminology is generally represented by noun phrases rather than just nouns (Nakagawa and Mori, 1998), an NLTK-based shallow parser (Babluki, 2013)[2] was used to identify noun phrases representing the topics (terms) of each sentence.

S8. The list of all terms and their indices was converted to a list of unique lower-cased terms, each with a list of indices. Most recently, this list had 34,437,644 items. Note that at this stage any term is included, not just linguistic terms. (Henceforth we will simply indicate new numbers of items in square brackets and preceded by an arrow as we go through the steps that it took to reduce the list).

S9. Only terms occurring 50 times or more were

[2] Available at https://gist.github.com/shlomibabluki/5539628

retained. [→ 142,729 items].

S10-11. Files were prepared allowing to determine the number of different documents in which a terms occurred. After manual inspection it was decided that a term should occur in at least 6 documents in order to minimize noise and maximize the inclusion of valid linguistic terms. [→ 133,927 items].

In the following three steps a rudimentary form of Named Entity Recognition (NER) is applied. The goal is to remove such entities not belonging to linguistic terminology.

S12. The presence of author names in the list of terms was reduced by matching more than 30k names found in source.bib with the list of terms. [→ 129,791 items].

S13. The presence of language names in the list of terms was reduced by matching more than 30k language names from an earlier version of Ethnologue (Eberhard et al., 2020) with the list of terms. [→ 121,699 items].

S14. The presence of publishers in the list of terms was reduced by matching more than 7k publisher names from source.bib with the list of terms. [→ 121,371 items].

S15. Manual inspection showed noisy terms to often have one of the following symbols in initial position: ‘, /, ¡, =, ¿, @, , —, , , , $. Such terms were found and deleted. [→ 117,648].

S16. Since the number of terms was still very large, at this point we passed from just eliminating negatives (non-linguistic terms) to first identifying positives (linguistic terms). This was done by using a glossary of linguistic terminology (7819 terms, including spelling variants) from the Summer Institute of Linguistics (SIL).[3] 3684 out of the 7819 SIL terms were found to recur among the 117,648 surviving terms in a non-case sensitive matching. We reasoned that a bona-fide linguistic term should bear some distributional similarity to at least one member of the core set of 3684 verified linguistic terms. The amount of similarity could be used as a cut-off for excluding terms not likely to be linguistic in nature. Thus, we measured the Normalized Pointwise Mutual Information (NPMI) (Bouma, 2009) between each of the 117,648 extracted terms and each of the 3684 verified linguistic terms among them, isolating the highest value and using that as a criterion for ‘linguisticality’ of the term. Some manual inspection showed that a maximal NPMI value of 0.5 would allow for a good balance between the inclusion of true positives and computational feasibility. By settling on this cut-off we excluded 98,474 terms, leaving 19,174. The vast majority of the included terms are relevant for the field of linguistics, and a $19,174 * 19,173/2 = 183,811,551$ size object entering into the computation of all pairwise NPMIs (see next section) can be handled efficiently in R.

The list of 19,174 terms along with indices linking them to sentence-like chunks in the collective file containing our database of linguistic literature (further linked to bibliographical references and other metadata) constitutes the basic data for this study. Several steps in the pipeline could be improved. For instance, more work could be done (and is being done) on the identification of bibliographical references in the text, and improvements to and extensions of the NER steps are eminently possible. Moreover, steps taken preceding the pipeline on OCR-error correction and other improvements of the input will increase the performance as well. Finally, it would be helpful if some form of performance evaluation could be developed (Granada et al., 2018). Still, taking into account the likely presence of a few thousand false positives, we have arrived at a list of linguistic terms about twice as large as the handmade SIL list and, most importantly, the list is one that reflects actual usage.

4 Related terms

Given that the list of 19,174 terms is associated with indices representing their occurrence in texts we could compute NPMI values (Bouma, 2009) for all pairs (using our own implementation of the NPMI). Pairs receiving the value -1, meaning that they do not co-occur, were excluded from further consideration. We also computed the Google Page Rank (GPR) for each of the items using the R package igraph (Csardi and Nepusz, 2006). The textual units used for computing NPMI and GPR were the ‘chunks’ (mostly equal to sentences) mentioned earlier.

Analyzing and plotting networks based on these data are useful aids in coming to decisions both about the design of a search functionality involving related terms and the prospects of basing an ontology of linguistic terms on such networks. Figures 1-2 show two clusters of related terms, selected from 3537 clusters. Clustering is based on a two-

[3] Available at https://feglossary.sil.org/english-linguistic-terms (accessed 2019-09-02).

column table where each of the 19,174 terms sits next to the term to which it has the highest NPMI value, here called 'best friend'. The 3537 clusters were extracted using igraph[4]. They range from having 2 to 200 elements, with median size 3 and mean size 5.42. log(size) and log(rank-of-size) is roughly a power-law distributed function (fit: $R^2 = .964$, exponent: -.668). Figures 1 and 2, respectively, are rather typical of a simple and a more complex cluster. The size of a cluster is determined by the availability of neighbors. For instance, the best friend of *voicing* is *degemination*, but there is no term that has *voicing* as its best friend. And all the clusters contain exactly one knot, representing the situation where two terms are each other's best friends. In both figures an arrow indicates relatedness in terms of NPMI and the direction of the error is from the term with the higher GPR to the one with the lower GPR. These directions currently have no real functionality but are included for exploratory purposes.

The clusters tend to be tightly knit around particular areas of linguistic terminology, as in the terms in Figure 1 that refer to processes that consonants may undergo (typically in intervocalic position) and the terms in Figure 2 that refer to elements of the organization of narratives.

We believe that the kind of clustering approach illustrated in Figures 1 and 2 is a useful way of supplying a search machine with suggestions for search terms that are related to the target term. Another possible approach would be to pick the terms that are highest-ranked in terms of their NPMI value, but they would tend to occur in the text returned for the target term by the search machine anyway and would not take the user in new, yet related directions in the same way as the present approach. The choice of how many terms should be returned is a matter of design. Currently even all elements of the largest cluster (200 terms) can be accommodated in a drop-down menu, so no restrictions may be necessary. The order of such a list could be determined by closeness in terms of the number of connecting edges, ties being resolved by GPR values, for instance.

As for the prospects for developing an ontology of linguistic terminology we believe that the present approach could also be productive. The clusters identified already offer themselves as basic components. One challenge is to connect these clusters. It seems that this could be done by finding an 'NPMI friend' of an appropriate member of the cluster in another cluster, and then linking clusters through such single edges.

[4] 'graph from edgelist' and 'decompose' functions

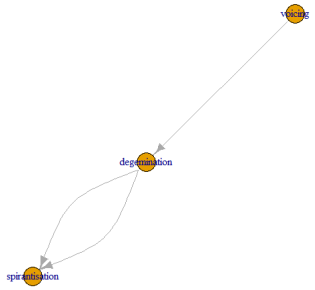

Figure 1: A simple cluster of related terms.

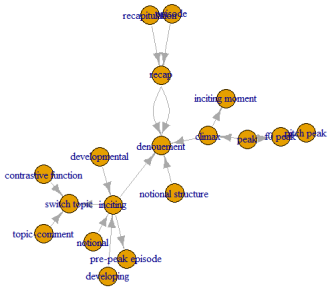

Figure 2: A more complex cluster of related terms.

5 Conclusion

In this paper we have demonstrated a pipeline for extracting terms from a thematically coherent text corpus, in this case a corpus of descriptive linguistic literature (to refer back to the outline in the Introduction this was Focus 1). We then went on to show that a simple clustering method, relying on single 'best friends' in terms of Normalized Pointwise Mutual Information (NPMI), is a useful basic step for designing a search machine suggesting search terms related to the target term (Focus 2a) and also has potential for helping in the construction of an ontology (Focus 2b).

We place importance on the fact that the pipeline for the extraction of domain-specific terms was fully automated, apart from some shortcuts where we used list of terms from external sources to prune the list.

Future work not already mentioned above, will go into developing a more systematic evaluation procedure, applying a similar pipeline to texts in languages other than English, and connecting the output in a ways such as to create both a multilingual search machine and a multilingual ontology.

Acknowledgments

Our research was carried out under the auspices of the project "The Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage" (NWO proj. no. 335-54-102) within the European JPI Cultural Heritage and Global Change programme (http://jpi-ch.eu/). It would not be possible without the DReaM corpus and associated metadata painstakingly compiled by Harald Hammarström.

References

Daniele Alfarone and Jesse Davis. 2015. Unsupervised learning of an IS-A taxonomy from a limited domain-specific corpus. In *IJCAI'15: Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1434—1441.

Hassan H. Alrehamy and Coral Walker. 2018. Semcluster: Unsupervised automatic keyphrase extraction using affinity propagation. In *Advances in Computational Intelligence Systems: Contributions Presented at the 17th UK Workshop on Computational Intelligence, September 6–8, 2017, Cardiff, UK*, pages 222–235, Cham. Springer.

Shlomi Babluki. 2013. An efficient way to extract the main topics from a sentence. https://thetokenizer.com/2013/05/09/efficient-way-to-extract-the-main-topics-of-a-sentence/. Technical report.

Gerlof Bouma. 2009. Normalized (point-wise) mutual information in collocation extraction. In *Proceedings of GSCL*, pages 31–40. Gesellschaft für Sprachtechnologie und Computerlinguistik.

Didier Bourigault and Christian Jaquemin. 1999. Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 15–22, Bergen. Association for Computational Linguistics.

Gabor Csardi and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2020. *Ethnologue: Languages of the World.* SIL International, Dallas, TX.

Roger Granada, Renata Vieira, Cassia Trojahn, and Nathalie Aussenac-Gilles. 2018. Evaluating the complementarity of taxonomic relation extraction methods across different languages. https://arxiv.org/abs/1811.03245.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. *Glottolog 4.3*. Max Planck Institute for the Science of Human History, Jena. (Available online at http://glottolog.org).

Harald Hammarström, One-Soon Her, and Marc Allassonnière-Tang. 2021. Keyword spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions. In *Swedish Language Technology Conference 2020 (SLTC 2020)*. NEJLT.

Kris Heylen and Dirk De Hertog. 2015. Automatic term extraction. In *Handbook of Terminology, Vol. 1*, pages 203–221, Amsterdam. John Benjamins Publishing Company.

Yong-Bin Kang, Pari Delir Haghigh, and Frada Burstein. 2016. TaxoFinder: A graph-based approach for taxonomy learning. *IEEE Transactions on Knowledge and Data Engineering*, 28:524–536.

Olena Medelyan, Ian H. Witten, Anna Divoli, and Jeen Broekstra. 2013. Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures. *WIREs Data Mining Knowl. Discov.*, 3:257–279.

Hiroshi Nakagawa and Tatsunori Mori. 1998. Nested collocation and compound noun for term recognition. In *Proceedings of the First Workshop on Computational Terminology*, pages 64—70, Montreal. Université de Montréal.

Roberto Navigli and Paola Velardi. 2004. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics*, 30(2):151–179.

Jing Qiu, Lin Qi, Jianliang Wang, and Guanghua Zhang. 2018. A hybrid-based method for Chinese domain lightweight ontology construction. *International Journal of Machine Learning and Cybernetics*, 9:1519–1531.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2004. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Shafqat Mumtaz Virk, Harald Hammarström, Markus Forsberg, and Søren Wichmann. 2020. The DReaM corpus: A multilingual annotated corpus of grammars for the world's languages. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 878–884.