Linguistica 5: Unsupervised Learning of Linguistic Structure

Jackson L. Lee
University of Chicago
1115 East 58th Street
Chicago, IL 60637, USA
jsllee@uchicago.edu

John A. Goldsmith
University of Chicago
1115 East 58th Street
Chicago, IL 60637, USA
goldsmith@uchicago.edu

Abstract

This paper introduces *Linguistica 5*, a software for unsupervised learning of linguistic structure. It is a descendant of Goldsmith's (2001, 2006) Linguistica. Open-source and written in Python, the new *Linguistica 5* is both a graphical user interface software and a Python library. While *Linguistica 5* inherits its predecessors' strength in unsupervised learning of natural language morphology, it incorporates significant improvements in multiple ways. Notable new features include tools for data visualization as well as straightforward extensions for both its components and embedding in other programs.

1 Introduction

The unsupervised learning of linguistic structure has been an important area of investigation in various disciplines. In natural language processing, unsupervised methods have the practical advantage over supervised ones that relatively less training data (which is time-consuming and costly to prepare) is required. In linguistics and cognitive science, a deeper understanding of how linguistic structure can be learned from unstructured data without supervision sheds light on human language acquisition. In this paper, we introduce *Linguistica 5*, a Python-based software for research on the unsupervised learning of linguistic structure. This software is a descendant of Linguistica 4 and its previous versions (Goldsmith, 2001; Goldsmith, 2006) dealing mainly with morphology.[1]

[1] http://linguistica.uchicago.edu/

In the following, we explain the axioms guiding the development of *Linguistica 5* in section 2. In section 3, the dual design of both a graphical user interface (GUI) and a Python library is introduced. Section 4 demonstrates data visualization using the GUI. Section 5 exemplifies how *Linguistica 5* can be used in conjunction with other computational tools in research. Section 6 concludes the paper.

2 Axioms

In the development of *Linguistica 5*, we adhere closely to the axioms of reproducible, accessible, and extensible research.

- **Reproducibility**: Research using *Linguistica 5* is reproducible, in the sense of Claerbout and Karrenbach (1992). *Linguistica 5* is open-source. The source code is publicly hosted at an online repository with detailed documentation (see footnote 1).

- **Accessibility**: Similar to all previous versions, *Linguistica 5* has a graphical user interface to make it accessible to a wide audience. However, *Linguistica 5* significantly departs from them by the introduction of data visualization tools. This is especially important for exploring potentially interesting patterns in large datasets; more on this in section 4.

- **Extensibility**: *Linguistica 5* facilitates extensible research in two ways. First, *Linguistica 5* is highly modular, which makes the addition of new components in further research straightforward. Second, apart from having a GUI, it

is also a Python library, which can be called in other Python programs for computational research; an example is in section 5.

3 Dual interface design: GUI and Python library

Previous versions of Linguistica are written in C++ and built in the Qt framework. These versions are designed to be GUI software out of the box. The major drawback is that the core back end is intimately tied with the GUI code, which makes further development and debugging difficult. To solve this problem, the new *Linguistica 5* takes a radically different approach.

First, we choose Python to be the new programming language for Linguistica, because it has been widely used in computational linguistics and natural language processing for its strengths in fast coding, strong library support for machine learning and other computational tools.

Second, the focus of the *Linguistica 5* development is its back end as a Python library, with a GUI wrapper written in PyQt. This new architecture has several advantages. In terms of the user interface, there are two independent choices. As in previous versions of Linguistica, the GUI allows convenient data analysis – and visualization, a new development in *Linguistica 5* (see section 4). Another novelty is that *Linguistica 5* is a Python library by design. Researchers are able to use *Linguistica 5* in a computationally dynamic and automatic fashion by calling it in their own programs for any research and computational work of their interest; section 5 provides an example.

4 Data visualization

Research along the lines of Linguistica has focused on natural language morphology. Still within the realm of unsupervised learning of linguistic structure, *Linguistica 5* represents an important step forward by attempting to (i) induce structure that goes beyond morphology, and (ii) use it to improve results in morphological learning. Given large datasets, data visualization has become indispensable for both exploring new questions as well as uncovering unknown ones. The increased interest in visualization of linguistic data and resources is reflected by the new, specialized research venues, e.g., the Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources (VisLR) which debuted in 2014. Here we provide an example from our ongoing work.

The area of interest is unsupervised word category induction (see Christodoulopoulos et al. (2010) for a recent review), which potentially offers solutions to challenging problems in fully unsupervised morphological learning (e.g. is the induced morphological paradigm *walk-walks* a verbal or nominal paradigm? And how do we characterize its potential connection with other induced paradigms such as *jump-jumped-jumps*?).

Currently, we are exploring graph-theoretic approaches to the problem of unsupervised word category induction. A central component is to model syntactic neighborhood among words in a given dataset. The current model is implemented as a series of steps for word similarity computation. First, a graph of word similarity for all pairs of word types is computed based on word ngram contexts. The normalized Laplacian graph is derived. Then, the most significant eigenvectors are computed, and the coordinates of words in the vector space based on these eigenvectors are also computed. A new graph of word similarity is obtained based on the Euclidean distance of the word coordinates. Words in this resultant graph are connected to one another in such a way that corresponds to syntactic neighborhood. For instance, the word "the" likely has other articles or determiners such as "a" and "an" as syntactic neighbors that occur in syntactically similar positions. Using the Brown corpus (Kučera and Francis, 1967), several syntactic neighbors for the word types "the", "would", and "after" are in Table 1.

Word	Syntactic neighbors
the	a his their an its this my our that
would	could must will can should may might
after	before like during since without through

Table 1: Syntactic neighbors

Importantly, the syntactic neighbors of a given word are themselves word types in the given dataset. The interconnectedness of words in the syntactic neighborhood results calls for network visualization. This can be done in *Linguistica 5* as one of

the key new features. Figure 1 shows a screenshot of *Linguistica 5* displaying the syntactic word neighborhood network for the most frequent 1,000 word types in the Brown corpus, as rendered by the force-directed graph layout in the JavaScript D3 library (Bostock et al., 2011). Figure 2 zooms in for the cluster of words that would be categorized as modal verbs in Brown such as "could", "would", and "must".

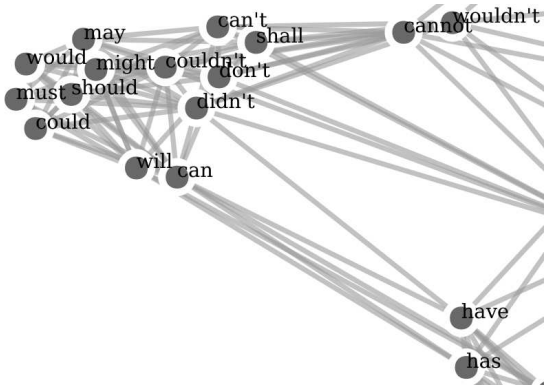

Figure 2: Zooming in Figure 1 for modal verbs

With induced knowledge analogous to word categories in natural language, results of unsupervised morphological learning could be improved. For instance, morphophonology could be learned. Induced morphological signatures (see section 5.1) such as {Ø, ed} (*walk-walked*) and {Ø, d} (*love-loved*) could be aligned for allomorphy across signatures (words with *ed* and *d* belonging to the same word category in this case). While this is work in progress, we have shown that data visualization tools in *Linguistica 5* as exemplified by syntactic neighborhood networks provide insights for new pursuits in research.

5 Embedding Linguistica 5 in other programs

Another new and powerful feature of *Linguistica 5* is that it is a Python library by design and is therefore callable in other Python-based programs. This is significant, because it is now possible to run the Linguistica algorithms dynamically for any data of interest from different sources (either from a local file or from an in-memory Python object).

We illustrate how *Linguistica 5* can be used as a Python library in conjunction with other tools with an example for computational modeling of human language acquisition, a growing field bringing linguistics, computer science, and cognitive science together (cf. Villavicencio et al. (2013)). We first provide the background on morphological signatures.

5.1 Morphological signatures

Unsupervised learning of morphology in Linguistica revolves around objects known as morphological signatures. A (morphological) signature, in the sense of Goldsmith (2001), is a morphological pattern associated with its stems as induced in some given data. For example, {Ø, s} is a morphological signature very likely to be induced in any sizable English datasets, with possible associated stems such as *walk-, jump-* (which entails that the words *walk, walks, jump, jumps* occur in the data).

Using the Brown corpus (about 50,000 word types from one million word tokens) for written American English, *Linguistica 5* finds over 300 morphological signatures. Those with the most associated stems are shown in the screenshot in Figure 3; the signature {Ø, ed, ing, s} is highlighted, with its associated stems displayed on the right.

	Signature	Stem count
1	NULL/s	2327
2	's/NULL	813
3	NULL/ly	587
4	NULL/d/s	346
5	NULL/d	314
6	ed/ing	197
7	'/NULL	190
8	's/NULL/s	181
9	d/s	175
10	ies/y	173
11	NULL/ed/ing/s	151
12	NULL/ed	134

NULL/ed/ing/s (number of stems: 151)			
abound	administer	affirm	aff
appeal	arrest	assault	att
awaken	award	beckon	be
bloom	bolt	broaden	bu
claw	click	climb	clu
coil	compound	concern	cor
confront	contact	contrast	cra
crown	decay	deck	dia
display	drill	drown	du
eschew	escort	exceed	exc
extend	filter	flounder	fro
haunt	hoot	hover	ho

Figure 3: Signatures with the most stems in the Brown corpus

5.2 On human morphological acquisition

Given that *Linguistica 5* is designed to model unsupervised morphological learning as a major goal, we ask how it can be used to model human morphological learning using child-directed speech data. An important criterion is that for the model to be cognitively plausible, it has to simulate the incremental nature of the input data. This means that the Linguistica algorithm for morphological learning must be

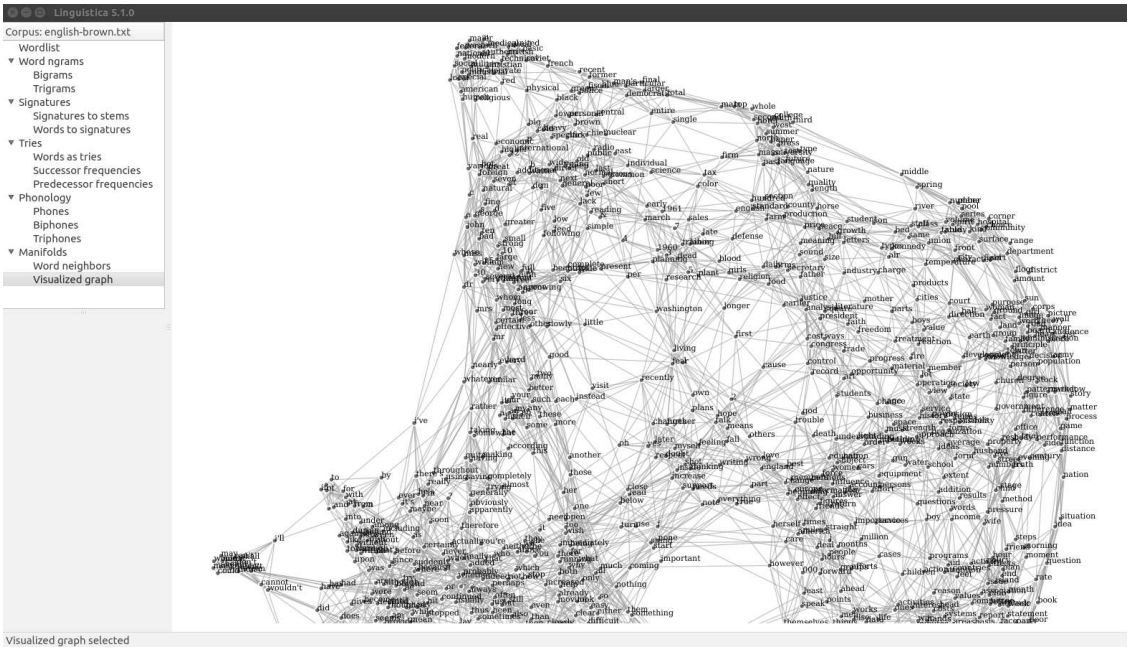

Figure 1: Syntactic word neighborhood network in *Linguistica 5*

called and applied flexibly over some growing data.

Concretely, we tested *Linguistica 5* for its ability to model morphological acquisition using Eve's data in the Brown portion (Brown, 1973) of the CHILDES database (MacWhinney, 2000), an idea sketched in Lee (2015). The child-directed speech (CDS) at different ages of the target child in the data was extracted by the PyLangAcq library (Lee et al., 2016) and fed into *Linguistica 5*. Table 2 shows the results of morphological signature induction from growing word types up to the ages of 18, 21, and 24 months, respectively.

Age	# word types	Induced signatures
18 mths	610	{'s Ø}{Ø s}
21 mths	1,246	{'s Ø}{Ø s}{Ø ing}{ll s}
24 mths	1,601	{'s Ø}{Ø s}{Ø ing}{ll s}{'s Ø s}

Table 2: Morphological signatures from CDS to Eve

The classic study of first language acquisition by Brown (1973) reports that the first three morphological patterns acquired by English-speaking children are the third-person singular inflection {Ø, s}, the possessive {'s, Ø}, and the progressive {Ø, ing}. Table 2 shows these are patterns that *Linguistica 5* successfully discovers in Eve's child-directed speech. Other induced signatures are {ll, s} (as in *she'll-she's*) and {'s, Ø, s}, a more complex pattern found when more data becomes available to the learner. The results for modeling language acquisition here contrast sharply with those from the Brown corpus in section 5.1, for the much larger amount of input data and results in the latter. But of particular interest is the *incremental* nature of learning in the former case. The fact that *Linguistica 5* is a Python library makes it possible to devise tools embedding it for multiple learning iterations run automatically.

In this section, we have shown how *Linguistica 5* can be used jointly with other programs for highly dynamic computational research, which is complementary to its GUI counterpart for exploratory ground work.

6 Conclusions

Linguistica 5 opens new doors to reproducible, accessible, and extensible research in unsupervised learning of linguistic structure. Building on the strengths of its predecessors, *Linguistica 5* incorporates novel elements of data visualization as well as employs a flexible and modular architecture to allow its integration into other projects and to maximize continual research and development.

Acknowledgments

This work was completed in part with resources provided by the University of Chicago Research Computing Center and Big Ideas Generator. We would also like to thank the three anonymous reviewers for very helpful comments and suggestions.

References

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309.

Roger Brown. 1973. *A first language: The early stages.* Harvard University Press.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised pos induction: How far have we come? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Jon Claerbout and Martin Karrenbach. 1992. Electronic documents give reproducible research a new meaning. In *Proc. 62nd Ann. Int. Meeting of the Soc. of Exploration Geophysics*, pages 601–604.

John A. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

John A. Goldsmith. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12(4):353–371.

Henry Kučera and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence.

Jackson L. Lee, Ross Burkholder, Gallagher B. Flinn, and Emily R. Coppess. 2016. Working with CHAT transcripts in Python. Technical Report TR-2016-02, Department of Computer Science, University of Chicago.

Jackson L. Lee. 2015. Morphological paradigms: Computational structure and unsupervised learning. In *Proceedings of NAACL HLT 2015 Student Research Workshop*.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ.

Aline Villavicencio, Thierry Poibeau, Anna Korhonen, and Afra Alishahi, editors. 2013. *Cognitive Aspects of Computational Language Acquisition*. Springer.