Panta rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek

Martina A. Rodda
Scuola Normale Superiore
Piazza dei Cavalieri, 7
56126 Pisa – ITALY
martina.rodda@sns.it

Marco S.G. Senaldi
Scuola Normale Superiore
Piazza dei Cavalieri, 7
56126 Pisa – ITALY
marco.senaldi@sns.it

Alessandro Lenci
CoLing Lab
Università di Pisa
via S. Maria 36
alessandro.lenci@unipi.it

Abstract

English. We present a method to explore semantic change as a function of variation in distributional semantic spaces. In this paper we apply this approach to automatically identify the areas of semantic change in the lexicon of Ancient Greek between the pre-Christian and Christian era. Distributional Semantic Models are used to identify meaningful clusters and patterns of semantic shift within a set of target words, defined through a purely data-driven approach. The results emphasize the role played by the diffusion of Christianity and by technical languages in determining semantic change in Ancient Greek and show the potentialities of distributional models in diachronic semantics.

Italiano. *Si presenta un metodo per indagare il cambiamento semantico come funzione della variazione all'interno di spazi semantici. Questo approccio è applicato per identificare automaticamente aree di cambiamento semantico nel lessico greco antico tra età pre-cristiana e cristiana. Modelli della Semantica Distribuzionale sono usati per identificare cluster e pattern di cambiamento semantico in una lista di parole target, definita con un approccio puramente data-driven. I risultati mostrano il ruolo della diffusione del Cristianesimo e dei linguaggi tecnici nel determinare cambiamenti semantici in greco antico, nonché le potenzialità dei modelli distribuzionali nella semantica diacronica.*

1 Introduction and Related Work

Distributional Semantics is grounded on the assumption that the meaning of a word can be described as a function of its collocates in a corpus. This suggests that diachronic meaning shifts can be traced through changes in the distribution of these collocates over time (Sagi et al., 2011). While some studies focused on testing the explanatory power of this method over frequency- and syntax-based approaches (Wijaya and Yeniterzi, 2011; Kulkarni et al., 2015), more advanced contributions to the field explored how distributional models can be used to test competing hypotheses about semantic change (Xu and Kemp, 2015), or to investigate the productivity of constructions in diachrony (Perek, 2016). The results attest the explanatory power of distributional methods in modeling diachronic shifts in meaning.

In this paper, we propose a method to identify semantic change through the **Representational Similarity Analysis** (RSA; Kriegeskorte and Kievit, 2013) of distributional vector spaces built from diachronic corpora. RSA is a method extensively used in neuroscience to test cognitive and computational models by comparing the geometry of their representation spaces (Edelman, 1998). Stimuli are represented with a representational dissimilarity matrix that contains a measure of the dissimilarity relations of the stimuli with each other. Different matrices are compared to evaluate the correspondence of the representational spaces built from different sources (e.g., behavioral and neuroimaging data). We argue that this method can be applied to compare distributional representations of the lexicon at different temporal stages. The hypothesis is that the elements in the lexical spaces showing larger geometrical variations in time correspond to the lexical areas that have undergone major semantic

changes. To the best of our knowledge, this is the first time RSA is used in diachronic distributional semantics.

Here we present a case study that applies RSA to track patterns of semantic change within the lexicon of Ancient Greek. We focus on the first few centuries AD, when the rise of Christianity caused a deep and widespread cultural shift within the Hellenic world. We predict that this shift will be reflected in the Greek lexicon of the time. In addition to past studies (Boschetti, 2009; O'Donnell, 2005 is a general introduction), we apply a bottom-up approach to the detection of semantic change, with no prior definition of a list of lemmas to be analyzed. The goal is to develop a quantitative "discovery procedure" to detect lexical semantic changes.

From a methodological standpoint, this study aims to show how Distributional Semantics can be applied fruitfully to such a small and literary corpus as the collection of Ancient Greek texts. The results will also highlight the ways in which Distributional Semantics can complement the intuition of the researcher in analyzing semantic change in Ancient Greek, providing a useful tool for future studies in Classics.

2 Materials and Methods

The corpus used for this study is based on the TLG-E (*Thesaurus Linguae Graecae*) collection of Ancient Greek literary texts. The database was divided into two sub-corpora, the first of which contains texts from the 7th to the 1st century BC (pre-Christian era), while the second one spans from the 1st to the 5th century AD (early Christian era). The pre-Christian sub-corpus contains 6,795,253 tokens, while the Christian sub-corpus totalizes 29,051,269 tokens.

The texts were lemmatized using *Morpheus* (Crane, 1991). Any issues with the lemmatization should not have a significant impact on the results unless otherwise stated (cf. Boschetti, 2009, page 60 for a discussion). After filtering for stopwords (mainly particles, pronouns and connectives) and lemmas occurring with a frequency below 100 tokens, the pre-Christian and Christian sub-corpus contain, respectively, 4,109 and 10,052 lemmas, which were used both as targets and dimensions in our vector spaces.

A vector space model was then built for each sub-corpus using the DISSECT toolkit (Dinu et al., 2013). Henceforth, we refer to the pre-Christian era model as the **BC-Space**, and to the Cristian era model as the **AD-Space**. Co-occurrences were computed within a window of 11 words (5 content words to the right and to the left of each target word). Association scores were weighted using positive point-wise mutual information (PPMI) (Evert, 2008); the resulting matrices were reduced to 300 latent dimensions using Singular Value Decomposition (SVD).

2.1 RSA of the distributional vector spaces

We have adapted the RSA method to discover semantic changes between the two vector spaces:
1. we identified the words occurring in both sub-corpora with a frequency higher than 100 tokens, obtaining 3,977 lemmas;
2. we built a representational similarity matrix (RSM) from the BC-Space (RSM_{BC}) and one from the AD-Space (RSM_{AD}). Each RSM is a square matrix indexed horizontally and vertically by the 3,977 lemmas and containing in each cell the cosine similarity of a lemma with the other lemmas in a vector space (this is a minor variation with respect to the original RSA method, which instead uses dissimilarity matrices). A RSM is a global representation of the semantic space geometry in a given period: vectors represent lemmas in terms of their position relative to the other lemmas in the semantic space;
3. for each lemma, we computed the Pearson correlation coefficient between its vector in RSM_{BC} and the corresponding vector in RSM_{AD}.
The Pearson coefficient measures the degree of semantic shift across the two temporal slices. The lower the correlation, the more a word changed its meaning.

3 Discussion of Results

The following section focuses on the words that underwent the biggest changes, i.e. those for which the correlation scores are lower. The primary goal will be to establish whether these words can be clustered into meaningful groups. This would allow us to pinpoint the areas within the lexicon of Ancient Greek that have undergone a significant semantic shift during the early centuries of Christianity.

3.1 Qualitative Analysis

The 50 lemmas with the lowest correlation coefficients were scrutinized in order to establish whether meaningful subgroups emerge. (This list of words is not reproduced here due to space constraints. They are a subset of the 200 words used to build the plot in section 4.3.) The findings in this section, while inevitably limited by

the intuition of the researcher, will provide the starting point for a more sophisticated analysis to be performed in the following sections.

The lemmas under consideration form a somewhat heterogeneous collection, including concrete nouns and relatively common verbs such as ζυγόν (zygón "yoke") and ἕπομαι (hépomai "follow"), as well as some proper nouns. This notwithstanding, a promising subset of words emerges even at this preliminary stage. These are a number of nouns designating eminently Christian concepts, such as παραβολή (parabolé "parable", previously "comparison"), λαός (laós, used for the Christian community as opposed to non-Christians, previously "people"), κτίσις (ktísis "creation", previously "founding, settling").

These findings are in line with the idea that the diffusion of Christianity played a substantial role in semantic change in the first centuries AD (cf. Boschetti, 2009). Other Christian terms, such as θεός (theós "God"), ἄγγελος (ángelos "angel", previously "messenger"), πατήρ (patér "father"), υἱός (hyiós "son"), also occur among the 100 words with the lowest correlation coefficients.

Another group of lemmas comprises technical terms whose usage seems to have undergone a specialization or a shift from one domain of knowledge to another. These include words such as ὑπόστασις (hypóstasis "substance", previously "sediment, foundation"), δύναμις (dýnamis "property (of beings)", previously "power"), or ῥητός (rhetós "literal" as opposed to "allegorical", previously "stated").

3.2 Analysis of Nearest Neighbors

To corroborate the intuitions detailed above, the 10 nearest neighbors for each of the last 50 words according to the correlation coefficient were retrieved using DISSECT. The process was repeated for each sub-corpus and the results compared in order to look for visible shifts, especially those involving different semantic domains. A few examples of the results should suffice to confirm the findings in the last section.

For instance, among the nearest neighbors for πνεῦμα (pnêuma "spirit", previously "breath") in the AD-Space we find such words as θεάομαι (theáomai "contemplate"), ἀληθινός (alethinós "true"), κτίσις, υἱός, θεός and so forth, while in the BC-Space the strongest similarity is with terms pertaining to the domain of physics, such as ἀήρ (aér "air"), ὑγρός (hygrós "moist"), θερμός (thermós "hot"). Another clear-cut example is that of δύναμις, whose neighbors change from military terms such as πολιορκία (poliorkía "siege") and στρατόπεδον (stratópedon "encampment, army") to the physical and philosophical domain, with the closest term being ἐνέργεια (enérgeia "activity, actuality", an antonym of δύναμις in its philosophical sense of "potentiality"). The case of δύναμις also shows how nearest neighbor analysis can reveal shifts in the usage of heavily polysemous words.

Not all changes observed through the analysis of nearest neighbors, however, are so easily predictable. Thus, for instance, the neighbors for μοῖρα (môira, another highly polysemous word with meanings spanning from "part" to "destiny") in the AD-Space come exclusively from the domain of astronomy, showing a strong specialization towards a technical usage ("degree" or "division" of the Zodiac). Another remarkable result comes from a geographical adjective, Ποντικός (Pontikós "coming from Pontus"), whose nearest neighbors shift from proper names and philosophical terms in the pre-Christian age (an association due, without doubt, to the usage of "Ponticus" as an epithet for authors, e.g. Heraclides) to names of currency and trade wares, probably as a reflection of the integration of Pontus as a Roman province (with the obvious repercussions on trade) in the 1st century AD.

3.3 t-SNE Plot

As a final analysis, we embedded the RSM_{AD} vectors for the 200 words with the lowest correlation coefficient with the corresponding RSM_{BC} vectors in a two-dimensional space with t-SNE (Figure 1), a technique for dimensionality reduction and data visualization that overcomes some of the limitations of standard multidimensional scaling (van der Maaten and Hinton, 2008). This procedure allows for easy identification of clusters, thus revealing the semantic relation between the most recent meanings of the words that underwent the greatest semantic change.

A number of small clusters can be observed in the plot. Near the left periphery, the most relevant group is composed of terms pertaining to Christian theology (from κύριος kýrios "Lord", λαός and θεός, to παρουσία parousía "Advent" and ποιμήν poimén "shepherd"). The position of ψῦχος (psŷkhos "cold") nearby is due to the mislemmatization of some inflected forms of ψυχή (psyché "soul") under this lemma, as revealed by nearest neighbor analysis. To the left of this group, a small cluster of terms pertaining to Christian exegesis (ῥητός, παραβολή, διασαφέω diasaphéo "illustrate") can be recognized.

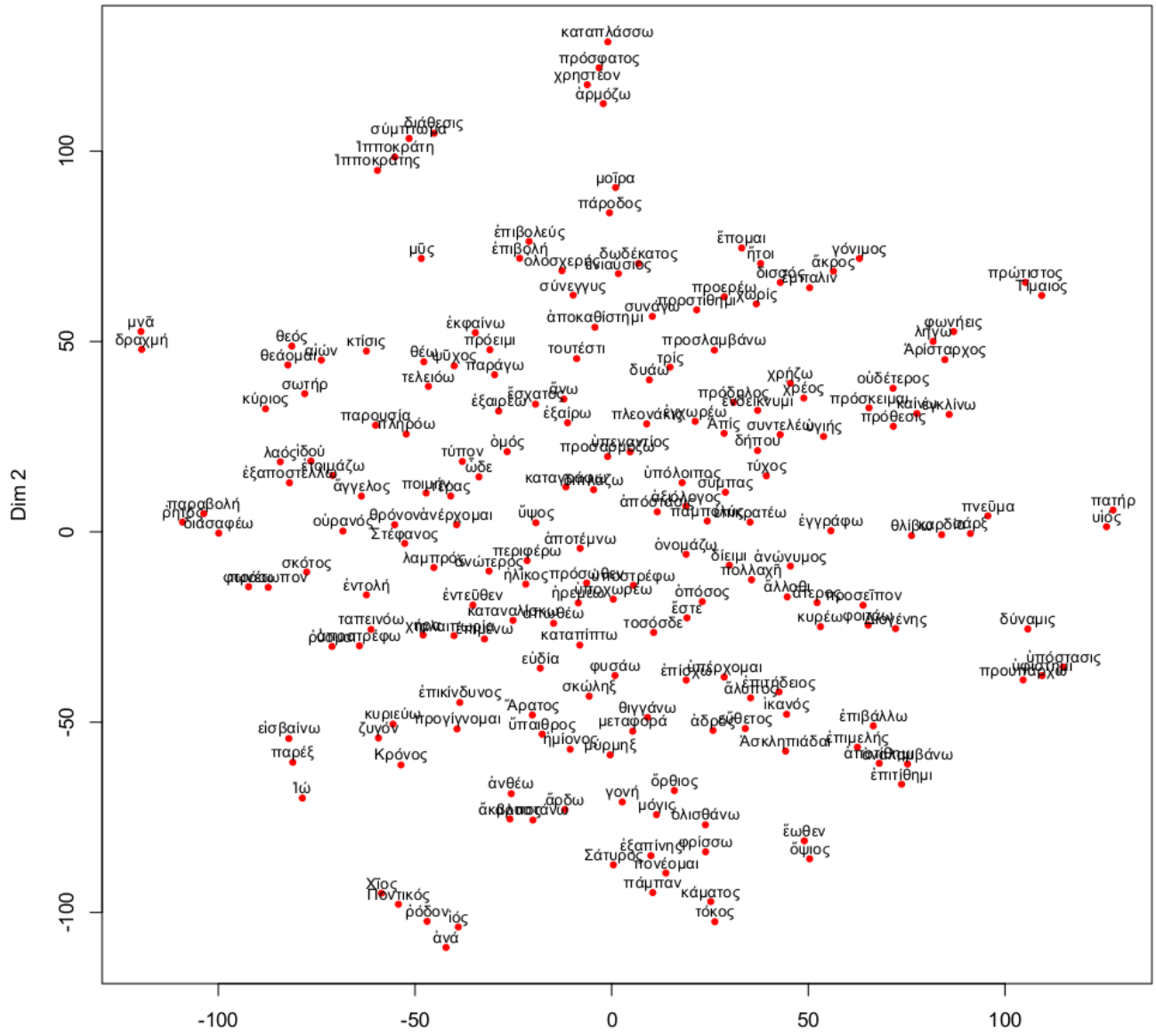

Figure 1. Relative positions within the AD-Space of the 200 words with the lowest correlation scores. Dimensionality reduction was performed using t-SNE (van der Maaten and Hinton, 2008).

The upper portion of the plot houses technical terms from the domains of medicine (the uppermost groups), astronomy and geometry, while philosophical terminology is found in the outer right area. Some smaller groups are also noticeable, such as μνᾶ (mnâ "mina") and δραχμή (drakhmé "drachma"), both units of currency, on the left, and πρώτιστος (prótistos "the very first") and Τίμαιος (the proper name Tímaios, Latin Timaeus), both connected to (Neo-)Platonic philosophy, on the right.

All in all, despite a certain amount of noise, the plot in Figure 1 supports the findings detailed so far. We can see how the main semantic changes in the Greek lexicon between the pre-Christian and Christian era affected the domains of religion (in a broader sense) and/or technical language. Within these domains, some more fine-grained relations between words that underwent significant semantic shifts can be observed.

4 Conclusion

This paper shows how Distributional Semantics can be used as an exploratory tool to detect semantic change. In this case study on Ancient Greek, the proposed method based on distributional RSA not only confirms the hypothesis that the diffusion of Christianity was a crucial cause of semantic change in the Greek lexicon, but also allows for the identification of unexpected patterns of evolution, such as the apparent specialization in the usage of technical terms. This last phenomenon could also be influenced by the fact

that the AD-corpus is richer in philosophical and technical treatises; however, a documented change in the proportion of different possible usages of a word is in itself a very informative result, especially in a field such as Classics, where the analysis of (literary) texts is paramount. Further research should undoubtedly highlight the effect of corpus composition. A focus on shorter periods of time might be of interest, since, for instance, the rise of technical prose writing is a characteristic of the Hellenistic Age (cf. e.g. Gutzwiller 2007, pages 154-167).

From a methodological standpoint, the fact that the results obtained from such a small corpus of purely literary texts are both meaningful and informative is of great relevance. Furthermore, the choice to adopt a data-driven approach proved fruitful, in that it brought to light directions of change that were not expected *a priori*. For traditional research in Classics, a computational approach to the lexicon of Ancient Greek is compelling because it provides new information about a language for which the judgments of native speakers are unavailable (cf. Perek, 2016). The results of this study show how Distributional Semantics can complement the assertions of the philologist, as well as help discover patterns of lexical change that would otherwise be impossible to grasp beyond an intuitive level.

References

Boschetti, Federico. 2009. A Corpus-based Approach to Philological Issues. PhD Thesis, University of Trento, Trento.

Crane, Gregory. 1991. Generating and parsing Classical Greek. *Literary and Linguistic Computing*, 6(4):243–245.

Dinu, Georgiana, Nghia The Pham and Marco Baroni. 2013. DISSECT – DIStributional SEmantics Composition Toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 31–36, Sofia.

Edelman, Shimon. 1998. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21:449–467.

Evert, Stefan. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 1212–1248, Berlin.

Gutzwiller, Kathryn J. 2007. *A guide to Hellenistic literature* (Blackwell guides to Classical literature). Blackwell Publishing, Oxford.

Kriegeskorte, Nikolaus and Roger A. Kievit. 2013. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web* (WWW ‘15), pages 625–635, Firenze.

Van der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

O’Donnell, Matthew Brook. 2005. *Corpus Linguistics and the Greek of the New Testament* (New Testament Monographs, 6). Sheffield Phoenix Press, Sheffield.

Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, 54(1):149–188.

Sagi, Eyal, Stefan Kaufmann and Brady Clark. 2011. Tracing semantic change with Latent Semantic Analysis. In Kathryin Allan and Justyna A. Robinson, editors, *Current Methods in Historical Semantics*, pages 161–183, Boston, MA.

Wijaya, Derry Tanti and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web* (DETECT ‘11), pages 35–40, Glasgow.

Xu, Yang and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (CogSci 2015), Pasadena, CA.