IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 3, Number 1
june 2017

Emerging Topics at the Third Italian Conference on Computational Linguistics and EVALITA 2016

aAccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 978-88-99982-64-5

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_3_1

aAccademia university press

Accademia University Press è un marchio registrato di proprietà
di LEXIS Compagnia Editoriale in Torino srl

IJCoL

Volume 3, Number 1
june 2017

Emerging Topics at the Third Italian Conference on Computational Linguistics and EVALITA 2016

CONTENTS

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

Martina A. Rodda *
Scuola Normale Superiore di Pisa

Marco S. G. Senaldi **
Scuola Normale Superiore di Pisa

Alessandro Lenci †
Università di Pisa

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.

1. Introduction

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, Kaufmann, and Clark 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

* Scuola Normale Superiore - Piazza dei Cavalieri 7, 56126 Pisa, Italy. E-mail: martina.rodda@sns.it
** Scuola Normale Superiore - Piazza dei Cavalieri 7, 56126 Pisa, Italy. E-mail: marco.senaldi@sns.it
† CoLing Lab, University of Pisa - Via S. Maria 36, 56126 Pisa, Italy .
E-mail: alessandro.lenci@unipi.it

showing larger geometrical variations in time correspond to the lexical areas that underwent 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), 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, enabling the researcher to discuss and interpret any meaningful patterns that may arise. 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. A distributional approach seems particularly suited to philological research, as it is already common and intuitive for researchers in this field to determine the exact meaning, usage restrictions, and stylistic connotations of a word by analyzing the context in which it occurs, especially when no other sources (such as ancient lexica) are available. Distributional Semantics provides the tools to perform a similar task not just on a much larger scale, but drawing information from the whole corpus; as such, it has the potential to highlight patterns in semantic change that would not otherwise be noticeable.

2. Related Work

The past few years have seen the rise of a series of studies tackling diachronic semantic change via computational methods. As pointed out by Sagi, Kaufmann, and Clark (2011), the increasing availability of computational tools for analyzing and manipulating large data sets and corpora allows for testing hypotheses and detecting statistical trends in a large-scale perspective that does not hinge on the intuitions of the linguist or the philologist. Crucially, most of this research has relied on a diachronic application of the distributional hypothesis (Harris 1954) by modeling semantic shift as a change of the co-occurrence patterns of a given lemma over time.

In Sagi, Kaufmann, and Clark (2011)'s proposal, the semantic narrowing or broadening of English words in the 1150-1710 period is modeled as an increase or decrease in density of the vector space populated by all the token occurrences of a given word in the various decades. The mean cosine similarity between all the token vectors of *dog*, for instance, decreases over time since it shifts from denoting a specific breed of dog to indicating *Canis familiaris* exemplars in general. Contrariwise, the mean cosine similarity between the token vectors of *hound* increases through the decades, since it originally meant 'dog' in general and ended up referring to dogs bred for hunting. Gulordava and Baroni (2011) resort to the American English section of the Google Books Ngram corpus, a collection of more than 5 millions of digitized books that were published between the sixteenth century and today (Michel et al. 2011), to build vector representations for words at two different time spans (the 60s and the 90s). The cosine similarity between the vector of a given word in the 60s space and the vector of the same word in the 90s space is then used as a measure of semantic shift for that term. These two time spans are taken into consideration in light of the major technological innovations that occurred in

the 90s and presumably affected the English lexicon. Such a distributional approach is shown by the authors to complement the results of a simpler frequency-based one, already proposed by Michel et al. (2011), which for instance interprets the increase in relative frequency of a given term over time as a signal of its acquired popularity and therefore of its semantic shift. As Wijaya and Yeniterzi (2011) highlight, such a method falls short of describing the nature of the investigated changes and of spotting more gradual shifts that are not reflected in frequency variations. In their work (Wijaya and Yeniterzi 2011), *k*-means clustering and Topics-Over-Time (Wang and McCallum 2006), a time-dependent topic model, are exploited to observe how and when the topics surrounding a given word change in diachrony. Results clearly bring to light words that change their semantic meaning over time (e.g. *gay* from 'frolicsome' to 'homosexual' around the 70s) and words getting additional meanings (e.g. *mouse* from 'long-tailed animal' to 'computer device' around the 80s-90s).

Kulkarni et al. (2015) compare the frequency-based approach with a syntactic one, which tracks variations in the probability distribution of part of speech tags given a target word in the different time snapshots of a corpus, and a best-performing word embeddings-based one (Mikolov et al. 2013), which learns word vectors for different time periods, warps the vector spaces into a unique coordinate system and creates a distributional time series for every word to assess its semantic displacement across time. With respect to Wijaya and Yeniterzi (2011), they also propose an algorithm for detecting the exact semantic change point in the time series built for each word with each of the three methods presented above. Their approach is also shown to be scalable and applicable to spotting shifts in different time spans, namely a century of written books with Google Book Ngrams, in years of Twitter blogging and in a decade of Amazon movie reviews. Diachronic distributional semantics is instead employed by Xu and Kemp (2015) to corroborate the *parallel change* law with respect to the *differentiation* one as for the semantic behavior of synonyms in time. Synonymic pairs like *impending* and *imminent* therefore tend to semantically evolve in parallel rathen than going different routes, maybe by virtue of analogical forces that aim at maintaining relationship patterns between words.

Another application of the distributional approach to a diachronic corpus is carried out by Perek (2016), who investigates the productivity of the "V *the hell out of* NP" construction from 1930 to 2009. The vectors of the verbs occurring in this construction are analyzed with multidimensional scaling and clustering to pinpoint the preferred semantic domains of the construction in its diachronic evolution, while a mixed effects logistic regression analysis shows the density of the semantic space of the construction around a given word in a certain period to be predictive of that word joining the construction in the subsequent period.

Hamilton, Leskovec, and Jurafsky (2016b) evaluate the performance of different kinds of word embeddings (PPMI, SVD, word2vec) in detecting attested historical semantic shifts (e.g. *broadcast* from 'scatter' to 'transmit') on cross-linguistic data by measuring changes in pair-wise similarities and the semantic displacement of a given lemma across time and run a series of regression analyses that reveal two general statistical laws of semantic change, namely that frequent words evolve at a slower rate and polysemous ones mutate faster. In a second study (Hamilton, Leskovec, and Jurafsky 2016a), they make use of both a global and a local neighborhood measure of semantic change to disentangle shifts due to cultural changes from purely linguistic ones. While the first index, which measures the cosine similarity between the vectors of the same word in consecutive decades, fares better in spotting purely linguistic changes for verbs, the second one, which keeps track of the changes in the nearest neighbors of

a word over time, performs better in detecting culturally motivated changes on nouns (e.g. *virus* from 'infectious disease' to 'unauthorized and harmful computer program').

3. Materials and Methods

3.1 The Corpus

The corpus used for this study is based on the TLG-E (*Thesaurus Linguae Graecae*) collection of Ancient Greek literary texts. This corpus does not include inscriptions or private letters and/or non-literary papyri, but it does include several fragmentary texts in both poetry and prose genres. Texts were divided into two sub-corpora, the former spanning from the 7th to the 1st century BC (pre-Christian era), while the latter 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.

Table 1
Percentage distribution of the main textual genres in the BC era and the AD era subcorpora (please keep in mind that a given text may belong to more than one genre at once).

Genre	BC era	AD era
Epic poetry	2.3%	0.3%
Historiography	13.79%	15.43%
Iambus and lyric	13%	6.24%
Tragedy	6.7%	2.6%
Comedy	12.88%	0%
Philosophy	14.86%	47.87%
Astronomy	2.54%	7.10%
Medicine	3.84%	19.17%
Mathematics	5.71%	2.09%

As Table 1 clearly shows, the two subsections are rather heterogeneous as regards the distribution of the main textual genres that compose them. When inspecting percentage values, please keep in mind that a given text may partake of more than one genre at once. As we can see, while the percentage of poetical texts (epic, iambic and lyric poetry) and theatrical texts (tragedy and comedy) diminishes from the BC to the AD era, the AD centuries are characterized by a greater diffusion of philosophical and technical (e.g. astronomical and medical) writings, with the exception of mathematical writings, that decrease from 5.71% to 2.09%. The percentage of historiographical works, finally, does not appear to vary considerably across the centuries.

Texts were lemmatized using *Morpheus* (Crane 1991). This parser is estimated to reach approximately 80% accuracy in lemmatizing Ancient Greek (Boschetti 2009, p. 60). Minor issues with the lemmatization are therefore to be expected, and will be mentioned and discussed in the Results section. Generally speaking, they seem to fall into two categories. The most basic issue arises when some inflected forms of a lemma are erroneously lemmatized separately (examples are visible in Table 5, where the comparative and superlative of the adjective ταχύς "takhýs; swift", e.g., appear as distinct lemmas); this kind of mis-lemmatization, however, should not have a significant impact on the semantic analysis, since said redundant lemmas will effectively have the

same meaning, and can be expected to behave in similar ways. Cases where forms of a word are lemmatized under an entirely unrelated lemma could, on the other hand, affect the results in a more significant way, but they appear to be very rare (the main example that can be detected in our data concerns forms of ψυχή "psykhé; soul" being erroneously lemmatized under ψῦχος "psŷkhos; cold": see section 4.3 below).

3.2 Building the Distributional Vector Spaces

Distributional Semantic Models (Lenci 2008; Turney and Pantel 2010) implement the distributional hypothesis advanced by Harris (1954), whereby linguistic expressions that are similar in meaning tend to occur in similar contexts. In these models, target linguistic expressions are represented as vectors in a high-dimensionality space, while each dimension of the vectors records the co-occurrence statistics of the target elements with some contextual features, e.g. the content words occurring in a fixed contextual window on the left and on the right of the target. By virtue of their representation with distributional vectors, words are encoded as points in a semantic space (Sahlgren 2006), and geometric measures of vector similarity or distance, like cosine (Turney and Pantel 2010), are exploited to model their semantic similarity. Like previous applications of distributional semantics to Ancient Greek (Boschetti 2009), we built two vector spaces from the TLG corpus, one from the pre-Christian subsection (**BC-Space** henceforth) and one from the Christian subsection (**AD-Space** henceforth).

After filtering out stop-words (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, Pham, and Baroni 2013). 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) (Turney and Pantel 2010), a statistical association measure that computes if two words x and y co-occur more often than expected by chance and sets to zero the negative results:

$$PPMI(x, y) = max(0, log \frac{P(x, y)}{P(x)P(y)}) \quad (1)$$

The resulting matrices were reduced to 300 latent dimensions with Singular Value Decomposition (SVD) (Deerwester et al. 1990).

3.3 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.

4. Discussion of Results

The following section focuses on the words that underwent the biggest changes, i.e. those with the lowest correlation scores. The primary goal is 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 underwent a significant semantic shift during the earliest centuries of Christianity.

4.1 Qualitative Analysis

The 50 lemmas with the lowest correlation coefficients were scrutinized by hand, 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 some adverbs and relatively common verbs such as ἕπομαι "hépomai; follow", as well as some proper nouns. This notwithstanding, two promising subsets of words emerge even at this preliminary stage (see the examples in Table 2).

Table 2
Some examples of lemmas undergoing the most substantial semantic change

Lemma	BC era meaning	AD era meaning
CHRISTIAN TERMS		
παραβολή parabolé	'comparison'	'parable'
λαός laós	'people'	'the Christians'
κτίσις ktísis	'founding'	'creation'
TECHNICAL TERMS		
ὑπόστασις hypóstasis	'foundation'	'substance'
δύναμις dýnamis	'power'	'property (of beings)'
ῥητός rhetós	'stated'	'literal (vs. allegorical)'

The first group comprises several 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 to drive 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. The shift undergone by words such as τόκος ("tókos; childbirth") is also likely to be connected to their occurrence in Christian contexts, even though it is hard to define this as a "meaning shift" *stricto sensu*. Such cases, and the theoretical issues they bring about, will be discussed separately in section 4.2, in light of the results of the nearest neighbor analysis.

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"). When the lemmas in this group refer to metaphysical concepts or exegetical terms, the influence of Christian thought may also be present. Within this category as well, one finds cases such as ἐνιαύσιος ("eniáusios; annual"), where the meaning of the word can hardly be assumed to have changed in the strictest sense, but its context of usage (as will be made clear by the nearest neighbor analysis in the next section) has shifted towards technical literature.

Together, the most clear-cut examples of these two groups (including those for which a semantic shift will be recognizable thanks to the nearest neighbor analysis performed in the following section) account for about half of the 50 words that underwent the most substantial semantic change. There is, of course, a measure of subjectivity in judging which words shifted towards a Christian or technical meaning; the findings in this section, however, can be supported through a more refined analysis.

4.2 Analysis of Nearest Neighbors

Nearest neighbor analysis proves especially useful when it comes to detecting shifts in meaning that would not be predictable through simple observation. Thus, for instance, the neighbors for μοῖρα ("môira", another highly polysemous lemma, with meanings spanning from "part" to "destiny") in the AD-Space come exclusively from the domain of astronomy and geometry (see Table 4; note that διάμετρον "diámetron; daily ration" is likely to be a lemmatization error for διάμετρος "diámetros; diameter"), showing a strong specialization towards a technical usage ("degree" or "division" of the Zodiac). Similarly, among the neighbors for the apparently anodyne noun ζυγόν ("zygón; yoke") one finds the constellations Λέων ("Léon; Leo"), Σκορπίον ("Skorpíon; Scorpius"), Παρθένος ("Parthénos; Virgo"), and Τοξότης ("Toxótes; Sagittarius"), revealing a shift in usage towards the astronomical sense, where Ζυγόν is the name of the constellation and Zodiac sign "Libra". This word, however, is the only name of a constellation that appears among the last 50 lemmas according to the correlation coefficient; in any case, the presence of words such as ὑποτάσσω ("hypotásso; to set, to submit"), δούλειος ("dóuleios; slavish"), and φορτίον ("fortíon; load") among the nearest neighbors in the AD-space shows that the astronomical meaning did not become as predominant as in the case of μοῖρα.

A similar surprising result comes from the 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

Table 3
Examples of nearest neighbors in the BC- and AD-space

πνεῦμα **'breath'** → **'spirit'**	
BC-space NNs	**AD-space NNs**
ἀήρ aér 'air'	θεάομαι theáomai 'to contemplate'
ὑγρός hygrós 'moist'	ἀληθινός alethinós 'true'
θερμός thermós 'hot'	αἰών aión 'aevum'
ψυχρός psykhrós 'cold'	κτίσις ktísis 'creation'
ὑγράζω hygrázo 'to be wet'	υἱός hyiós 'son'
θερμαίνω thermáino 'to heat'	θεός theós 'God'
πυκνός pyknós 'compact'	πατήρ patér 'God the Father'
ἀναπνοή anapnoé 'breathing'	δοξάζω doxázo 'magnify'
ψυχρόομαι psykhróomai 'to be chilly'	οἰκονομία oikonomía 'administration'
θερμότης thermótes 'heat'	πληρόω pleróo 'to fill'
δύναμις **'power'** → **'property (of beings)'**	
BC-space NNs	**AD-space NNs**
προάγω proágo 'to lead forward'	ἐνέργεια enérgeia 'activity'
πολιορκία poliorkía 'siege'	μετέχω metékho 'to partake of'
ἀθροίζω athróizo 'to gather'	ἐνεργέω energéo 'to be in action'
στρατόπεδον stratópedon 'encampment'	κινητικός kinetikós 'related to motion'
στρατιώτης stratiótes 'soldier'	φύς phýs 'son'
παράταξις parátaxis 'line of battle'	οὐσία ousía 'substance'
ἀναζεύγνυμι anazéugnymi 'to yoke'	ἰδιότης idiótes 'specific property'
καταπλήσσω kataplésso 'to strike down'	φύσις phýsis 'nature'
Καρχηδόνιος Karkhedónios 'Carthaginian'	ποιότης poiótes 'quality'
ἀναλαμβάνω analambáno 'to take up'	δισσός dissós 'twofold'

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. This is not, strictly speaking, a shift in meaning, but in real-word reference and usage; as such, it is parallel to cases such as θεός, where the most relevant change is in the cultural context.

Specialization towards a narrower usage is not, however, the only possible route of semantic change for technical terms: some of these appear to have moved from one domain to another. The case of πνεῦμα, whose semantic domain shifts from physics to metaphysics and philosophy (see Table 2 above), has already been discussed. Another example is σύμπτωμα ("sýmptoma" with the generic meaning of "chance occurrence"), whose top three neighbors in the BC-space are λογισμός ("logismós; calculation, reasoning"), θεωρέω ("theoréo; to contemplate"), and προερέω ("proeréo; to predict"); in the AD-space, in their place we find πυρετέω ("pyretéo; to be feverish"), νόσημα ("nósema; disease"), and πυρετός ("pyretós; fever"), revealing a shift from the philosophical to the medical domain (i.e. from "property" to "symptom"). Another example, this time spanning the technical and Christian domains, is παραβολή ("parabolé; parabola, parable", among other possible meanings), whose neighbors in the BC-space mostly have to do with geometry, while in the AD-space they pertain to the domain of biblical and literary exegesis. The nearest neighbors of ῥητός, one of the lemmas that had already

Table 4
Examples of nearest neighbors for astronomical terms

μοῖρα **'part, portion'** → **'degree, division (of the Zodiac)'**

BC-space NNs	AD-space NNs
ἕπομαι hépomai 'to follow'	ἔγγιστος éngistos 'nearest, next'
δύω dýo 'to plunge in, to enter'	ζῳδιακός zodiakós 'Zodiac'
μένος ménos 'might, spirit'	ἰσημερινός isemerinós 'equinoctial'
κέω kéo 'to lie down, to rest'	πάροδος párodos 'passage, entrance'
γαῖα gâia 'earth'	διάμετρον diámetron ['diameter'*]
ἀστήρ astér 'star'	τμῆμα tmêma 'section, sector'
ἦμαρ êmar 'day'	Κριός Kriós 'Aries'
τόσος tósos 'so much (as)'	μεσουρανέω mesouranéo 'to culminate'
λείπω léipo 'to leave'	κέντρον kéntron 'center'
αὐτίκα autíka 'at once'	μεσημβρινός mesembrinós 'of noon, southern'

* see in-text discussion.

ζυγόν **'yoke'** → **'Libra'**

BC-space NNs	AD-space NNs
κέω kéo 'to lie down, to rest'	ὑποτάσσω hypotásso 'to set; to submit'
ὦμος ômos 'shoulder'	δούλειος dóuleios 'slavish'
ἕπομαι hépomai 'to follow'	Λέων Léon 'Leo'
πούς póus 'foot'	κυριεύω kyriéuo 'to be lord'
μέση mése 'middle string'	φορτίον fortíon 'load'
μέσος mésos 'middle'	δουλεύω douléuo 'to be slave'
δόρυ dóry 'shaft, spear'	Σκορπίον Skorpíon 'Scorpius'
μοῖρα môira 'part, portion'	Παρθένος Parthénos 'Virgo'
λαιός laiós 'left'	Τοξότης Toxótes 'Sagittarius'
γόνυ góny 'knee'	ἐλεύθερος eléutheros 'free'

been singled out as promising examples of a shift towards a technical meaning through qualitative analysis, show a similar evolution from the mathematical to the exegetical domain.

There are also sporadic cases where the shift in meaning seems to be from a more technical usage in the BC-space to a more generalized meaning in the AD-space. A representative example is the verb δίειμι ("díeimi; to go through"). Its nearest neighbors in the BC-space all come from the domain of physics, and are indeed strongly specialized towards adjectives indicating properties of matter (see Table 5; some more minor issues with lemmatization make an appearance here, with the same adjective being categorized as two different lemmas, but since these lemmas seem to behave in a similar fashion, the impact on the results can be supposed to be minimal). In the AD-space, the physical domain seems to have disappeared entirely, with the synonym διέρχομαι ("diérkhomai; to go through") now taking pride of place among the nearest neighbors. Of course, it is also possible that the appearance of this kind of pattern for a limited number of lemmas might be due to the different size of the two sub-corpora.

Finally, like in the qualitative analysis, we find examples of lemmas where the shift seems to have to do with a different context of usage rather than thorough meaning

Table 5
Nearest neighbors for δίειμι

δίειμι **'to go through'**	
BC-space NNs	**AD-space NNs**
πυκνός pyknós 'compact'	διέρχομαι diérkhomai 'to go through'
λεπτόν leptón 'thin'*	θάσσων thásson 'swifter'*
λεπτός leptós 'thin'*	διεξέρχομαι diexérkhomai 'to pass through'
ξηρά xerá 'dry'*	ἀξιόλογος axiólogos 'remarkable'
παχύς pakhýs 'thick'	ὁπόσος hopósos 'as much (as)'
ψυχρός psykhrós 'cold'	τάχιστος tákhistos 'very swift'*
ξηρός xerós 'dry'*	χωρίον khoríon 'place'
ὑγρός hygrós 'moist'	διέξειμι diéxeimi 'to pass through'
μανός manós 'sparse'	ἀποχωρέω apokhoréo 'to go away'
ὑγρότης hygrótes 'moisture'	πλεῖστος pléistos '(the) most, greatest, largest'

* see in-text discussion.

change. Perhaps the most clear-cut case is the locative adverb αὐτόθεν ("autóthen; from this very spot, immediately"), whose nearest neighbors in the BC-space are entirely generic (including words such as ἄγνυμι "ágnymi; to break" and ναῦς "náus; ship"), while in the AD-space they seem to pertain mostly to the domain of logical and mathematical reasoning (with words such as ὑπόθεσις "hypóthesis; hypothesis", ἀκόλουθος "akólouthos; following, consequent", and ἀποδείκνυμι "apodéiknymi; to prove, to demonstrate"). In this case, just as for τόκος in section 4.1, it is hard to posit a "meaning shift" of any sort, but we can envisage a technical context of usage becoming predominant.

Cases in which the change in context does not seem to straightforwardly translate to a shift in meaning, draw attention to one of the subtlest implications of the results presented here. Given the small dimensions of the corpus, it is sometimes difficult to rule out an influence of the genre of the texts analyzed on the distribution of results — for instance, the impact of technical usage on the meaning of many of the terms that underwent the most significant semantic change might be connected to the presence of a higher number of philosophical and technical treatises in the AD-space. As we showed in Section 3.1, the percentage of works classified as "philosophical" in the TLG categorization system does indeed rise steeply in the AD-corpus (47.87%, as opposed to 14.86% in the BC-corpus), but the increase is less noticeable for other technical genres (e.g. astronomical writings, 2.54% to 7.10%, and medical writings, 3.84% to 19.17%), while the percentage of mathematical writings is actually lower in the AD-corpus (2.09%) than in the BC-corpus (5.71%). Note that, since the same work can be categorized as belonging to more than one genre in the TLG, percentages for different genres need to be kept apart. Further research should undoubtedly highlight the effect of corpus composition; a focus on shorter periods of time might be of interest for future studies, since, for instance, the rise of technical prose writing is widely recognized as being a characteristic of the Hellenistic Age (cf. e.g. Gutzwiller (2007, p. 154-167). Note that, for the aims of this study, texts from this period are included in the BC-space, not the AD-space). A documented change in the proportion of different possible usages of a word, however, is in itself a very informative result, especially in a field such as Classics, where

the analysis of (literary) texts is paramount. Indeed, the shift towards Christian usage for several terms can in itself be described as the introduction of an entirely new genre of Christian writings, but this would sidestep the issue that there has been a noticeable change in the usage of these words (and, by definition, their meaning, according to the Distributional Hypothesis).

4.3 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 using 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. While the analysis in the previous sections was aimed at detecting patterns of semantic shift between the BC-space and AD-space, the purpose of the t-SNE plot is to investigate whether there is any significant relationship between the meanings of the words that underwent such a shift; because of this difference in purpose, the information contained in the plot is limited to one semantic space. For the same reasons, the potential issues about the composition of the corpus and the impact of genre, as sketched at the end of section 4.2 above, are not relevant for the discussion here.

A number of small clusters can be observed in the plot. Near the left periphery, the most relevant group (in blue) is composed of terms pertaining to (Christian) theology, from κύριος (“kýrios; Lord”), λαός and θεός, to παρουσία (“parousía; Advent”), ποιμήν (“poimén; shepherd”), τέρας (“téras; sign, portent”), and οὐρανός (“ouranós; heaven”). The position of ψῦχος (“psŷkhos; cold”) near this cluster is due to the mis-lemmatization of some inflected forms of ψυχή (“psykhé; soul”) under this lemma, as revealed by nearest neighbor analysis (see section 3.1 above). To the left of this group, a small cluster of terms (in light blue) pertaining to Christian exegesis (ῥητός, παραβολή, διασαφέω “diasaphéo; to illustrate”) can be recognized. At the far right of the plot, diametrically opposed to the previous clusters, another small group of Christian terms can be recognized; this includes πατήρ, ὑιός, πνεῦμα, and potentially καρδία (“kardía; heart”) and σάρξ (“sárx; flesh”).

The upper portion of the plot (in green) houses technical terms from the domains of medicine (the upper-most group, spanning the personal name Ἱπποκράτης “Hippokrátes; Hippocrates”, the nouns διάθεσις “diáthesis; condition” and σύμπτωμα, the verb καταπλάσσω “kataplásso; to apply a plaster/poultice”, and the adjective πρόσφατος “prósphatos; fresh”), astronomy and geometry (difficult to distinguish, from μοῖρα and πάροδος “párodos; passage” to ἄκρος “ákros; top-most” and δισσός “dissós; two-fold”).

Philosophical terminology (in red) can be found in the lower right area (δύναμις, ὑπόστασις, etc.), while a separate cluster of terms pertaining to moral philosophy (ἐπιτήδειος “epitédeios; suitable”, ἱκανός “hikanós; sufficient”, ἐπιμελής “epimelés; careful”, all clustering around the crucial term ἄλυπος “álypos; without pain, painless”) is visible nearer to the center of the plot (in brown). Some smaller groups are also noticeable, such as μνᾶ (“mnâ; mina”) and δραχμή (“drakhmé; drachma”), both units of currency, on the left (in orange), and πρώτιστος (“prótistos; the very first”) and Τίμαιος (the proper name Tímaios, Latin Timaeus), both connected to (Neo-)Platonic philosophy, on the right (in red). All in all, despite the inevitable amount of noise, the plot in Figure 1 supports the findings detailed so far. We can see how the main

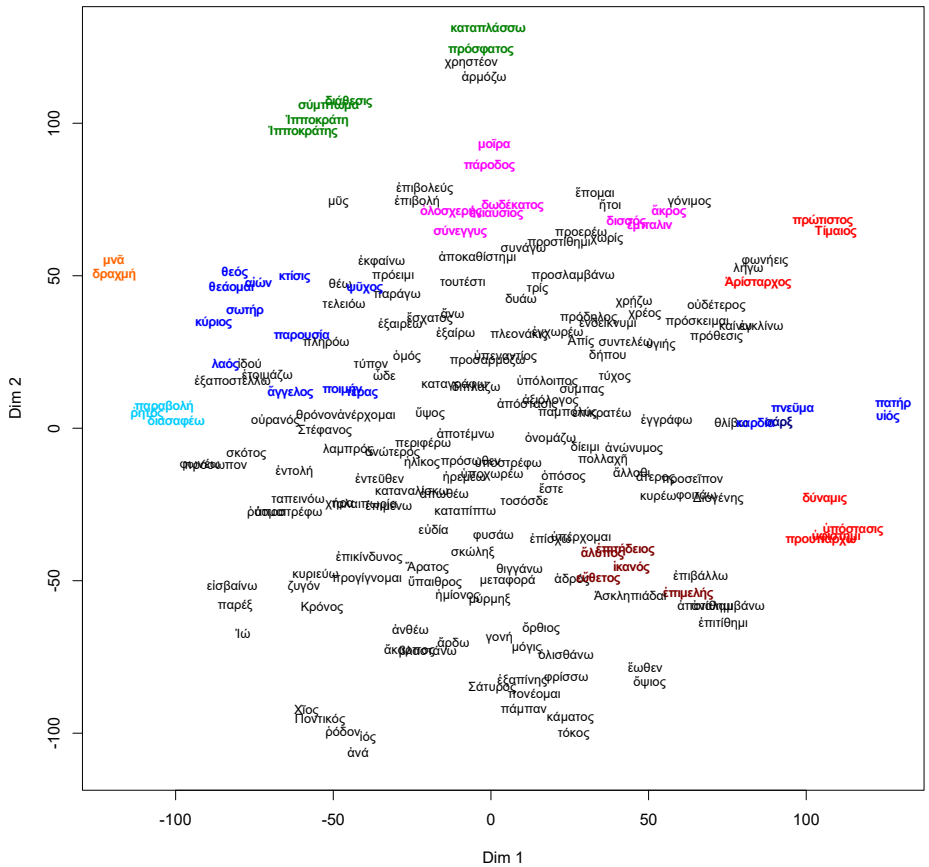

Figure 1
Relative positions within the AD-Space of the 200 words with the lowest correlation scores. Dimensionality reduction was performed using t-SNE.

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 went through a significant semantic shift can be observed.

5. Conclusions and future work

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 specialization in the usage of technical terms. 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. The nearest neighbor analysis performed in section 4.2 brought to light several patterns of change, which proved informative both as concerns the evolution of some semantic domains between the BC- and AD-space, and

the potential effects of the composition of the corpus (in itself a potentially interesting source of information for Ancient Greek). The t-SNE plot, by showing how the words that underwent the most relevant meaning shifts tend to form semantically-motivated clusters, provided a further opportunity to detect areas of the lexicon that underwent significant semantic change.

As far as broader methodological issues are concerned, 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 findings of the philologist, as well as help discover patterns of lexical change that would otherwise be impossible to grasp beyond an intuitive level. Nonetheless, a few issues remain open and could benefit from a more fine-grained investigation in future studies. First and foremost, it could be interesting to observe which parts of speech tend to change first, e.g. whether nouns or verbs (Dubossarsky, Weinshall, and Grossman 2016), and whether specific genres are more prone to change than others. Secondly, a targeted study of a more restricted period right after or right before the advent of Christianity (rather than the twelve-century time span considered here) could help confirm that the shifts we detected were primarily due to the spread of Christianity itself, which would have then represented a major breaking point, and rule out the possibility that a more natural and broad-spectrum change was already taking place.

References

Boschetti, Federico. 2009. *A Corpus-based Approach to Philological Issues*. Ph.D. thesis, University of Trento.

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

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391.

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,Bulgaria, August, 4-9.

Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman. 2016. Verbs change more than nouns: A bottom-up computational approach to semantic change. *Lingue e linguaggio*, 15(1):7–28.

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

Gulordava, Kristina and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, Scotland, July 31.

Gutzwiller, Kathryn J. 2007. *A Guide to Hellenistic Literature*. Blackwell Publishing.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2122, Austin, Texas, USA, November 1-5.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, Berlin, Germany, August 7-12.

Harris, Zellig S. 1954. Distributional structure. *Word*, 10(2-3):146–162.

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 World Wide Web Conference*, pages 625–635, Florence, Italy, May 18-22.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26tth International Conference on Neural Information Processing System*, pages 3111–3119, Lake Tahoe, Nevada, USA, December 5-10.

O'Donnell, Matthew Brook. 2005. *Corpus Linguistics and the Greek of the New Testament*. Number 6. Sheffield Phoenix Press.

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*. Mouton de Gruyter, pages 161–183.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis.

Turney, Peter D. and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

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

Wang, Xuerui and Andrew McCallum. 2006. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433, Philadelphia, Pennsylvania, USA, August 20-23.

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*, pages 35–40, Glasgow, United Kingdom, October 24-28.

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*, pages 2703–2708, Pasadena, California, July 22-25.