# Brave New World
## Uncovering Topical Dynamics in the ACL Anthology Reference Corpus Using Term Life Cycle Information

**Anne-Kathrin Schumann**

annek‗schumann@gmx.de

## Abstract

One of the main interests in the analysis of large document collections is to discover domains of discourse that are still actively developing, growing in interest and relevance, at a given point in time, and to distinguish them from those topics that are in stagnation or decline. The present paper describes a terminologically inspired approach to this kind of task. The inputs to the method are a corpus spanning several decades of research in computational linguistics and a set of single-word terms that frequently occur in that corpus. The diachronic development of these terms is modelled by means of term life cycle information, namely the parameters *relative frequency* and *productivity*. In a second step, k-means clustering is used to identify groups of terms with similar development patterns. The paper describes a mathematical approach to modelling term productivity and discusses what kind of information can be obtained from this measure. The results of the clustering experiment are promising and well motivate future research.

## 1 Introduction

The discovery of trends and other kinds of topical dynamics is one of the central aims of applied computational linguistics research. It is also of great interest to the digital humanities community for which large text collections are typical sources of information: Which of the many topics mentioned in the corpus are relevant at a given moment in time? How to sort them diachronically, how to model their interplay? These and similar questions, directed towards the ACL Anthology Reference Corpus (ACL ARC) (Bird et al., 2008), form one part of the motivation for the present paper.

A rather more pronounced source of motivation, however, is related to *terminology*, i.e. the study of the specialised lexicon (Wüster, 1979). In terminology, text-linguistic and lexico-semantic approaches (see, for example, Faber and L'Homme (2014)) have been contrasted to knowledge management and its need for abstract, static representations of (specialised) knowledge. Well-known, even if rather different examples of such representations are the Saffron system[1] (Bordea, 2013) and the EcoLexicon[2] (Faber et al., 2016).

The present paper takes a new perspective on terminology by stressing the importance of *temporal dynamics*: Knowledge evolves constantly and this evolution obviously affects concepts and terms as well as the relations that they form. *Term life cycles*, then, are indicative of the evolution of knowledge and a better understanding of them might be helpful in tasks such as information extraction, semantic relatedness analysis, temporal text classification, or trend analysis. Therefore, the present paper aims at finding (preliminary) answers to, at least, one of the following research questions.

1. What are the parameters by which the diachronic development of terms and topics can be described? Is it possible to model diachronic term development patterns or even a term life cycle (e.g. creation, growth, consolidation, and decline)?

2. Is it possible to use knowledge about this life cycle for extracting information (e.g. by distinguishing growing/trending terms from consolidated or dying ones)?

---

[1] http://saffron.insight-centre.org.
[2] http://ecolexicon.ugr.es/en/index.htm.

3. Is it possible to identify terms that exhibit similar development patterns? If yes, are these terms semantically related?

## 2 Related Work

The present investigation is related to various strands of research in terminology and computational linguistics. In a general way, it forms a part of the growing body of scientific work dedicated to the *analysis of scientific text corpora*, an area that has developed a multitude of different approaches (compare, for example, Atanassova et al. (2015)). Text-analytical studies, in their majority, aim at the exploitation of scientific data as a source of knowledge. Typical use cases are term extraction, the analysis of citation networks and co-authorship graphs as well as text classification. Interesting terminological variations on these common themes are the studies by Monaghan et al. (2010), who use terminological methods for the identification of domain experts, and the analysis of the LREC Anthology carried out by Mariani et al. (2014).

*Trend analysis* research is related to our study insofar as we hope to draw conclusions on "trending" or "growing" topics or terms on the basis of term life cycle modelling. Terminology is considered to varying degrees in this kind of research. An example that explicitly accounts for a whole range of term features is the system described by Babko-Malaya et al. (2015). Their complex tool models the emergence of new technologies from a corpus of scientific patents mainly on the basis of non-linguistic sources of information (authors, H-index, affiliation, etc.). However, terms are extracted, too, and characterised, among many other parameters, by the status of authors using them and their maturity as measured by linguistic usage patterns. By far simpler approaches to trend analysis are the studies by Francopoulo et al. (2016) and Asooja et al. (2016). Francopoulo et al. (2016) use machine learning techniques to predict the relative term frequencies of terms extracted from the NLP4NLP corpus (Francopoulo et al., 2015). The work carried by out by Asooja et al. (2016) is similar in that it uses Saffron to extract terms from LREC papers and then combines tf-idf scores with regression modelling to predict the future growth or decline of terms.

*Terminological* studies dedicated to uncovering diachronic aspects of term development are relatively rare. Picton (2011) is an innovative study

dedicated to the description of term life cycles. Working on two very small corpora, Picton uses features such as term frequency, linguistic patterns, term variation, and term productivity to identify term life cycle patterns that can be classified into four categories:

- Novelty and obsolescence (various types of neology and necrology, that is, the disappearance of a concept and its denomination)

- Implantation of terms and concepts, that is, the fact of their being accepted as familiar units in a given domain – the next step after neology

- Centrality: this is a topic-related category containing patterns such as "central topic" and "topic disappearance", that is, terms become obsolete because the dominant paradigm in a given field of expertise changes

- Changes related to the structure of specialised documents, that is, changes caused by terminologically uninteresting reasons

Unfortunately, Picton does not describe a robust analysis or evaluation method for her model. Other related terminological studies are Schumann and QasemiZadeh (2015) as well as Schumann and Fischer (2016). Schumann and QasemiZadeh model the development of the term "machine translation" in the ACL ARC by extracting related terms at two distinct time periods. Schumann and Fischer annotate terms in a diachronic corpus of scientific English and present a pilot study arguing that terms undergo semantic and morpho-syntactic development processes over time.

The present study clearly extends and adds to the cited investigations: The presented approach is not just an attempt at extracting "growing" or "trending" terms, but, in fact, represents a more principled effort towards modelling the evolution of the specialised lexicon. The paper also presents a novel parameter for the description of temporal dynamics in terminology. The scientific goal consists in a better understanding of the evolution of knowledge through the evolution of terms.

## 3 Modelling the Term Life Cycle

This study aims at modelling the life cycles of individual terms in order to learn more about their diachronic development. This is done with the

help of just two parameters, namely term frequency and term productivity. Another important decision is to work on the level of single-word terms. This is not just a pragmatic decision related to the fact that single-word terms have a sufficient amount of occurrences, whereas many multi-word terms may not. We also view single-word terms as representatives of semantic clusters of related, more specific terms or, in the words of Bordea (2013), candidates for "domain models". Consequently, by modelling the life cycles of single-word terms, we hope to model the life cycle of their multi-word child terms as well.

## 3.1 Parameters

As pointed out before, we try to model term life cycles with the help of two parameters, namely term frequency and term productivity, and analyse these parameters in the form of a time series:

- **Term frequency**, that is, the absolute frequency of occurrence of a given term in a given year, normalised by the number of word tokens available from the corpus for that year.

- **Term productivity**, that is, a measure for the ability of a concept (lexicalised as a single-word term) to produce new, subordinated concepts (lexicalised as multi-word terms).

While our take on frequency, though probably unorthodox, may not require any further explanation, a more detailed discussion of "productivity" seems in order here. First of all, productivity is defined only for simple terms, e.g. "word". Productivity, then, is the ability of "word" to participate in the formation of new multi-word terms, e.g. "target word", "input word", etc. We decided to formalise this feature in terms of entropy. In particular, for each year $y$ and single-word term $t$, we calculated the entropy of the conditional probabilities of all $n$ multi-word terms $m$ containing $t$. This is shown in Formula 1:

$$e_{(t,y)} = -\sum_{i=1}^{n} log_2(p_{m_i,y}) \cdot p_{m_i,y} \qquad (1)$$

Entropy is a measure of dispersion and, therefore, adequate for measuring productivity:

- If a term has many derived multi-word terms (MWTs) with similar probabilities, it is very productive and has a high entropy.

- If a term has only a few MWTs, it is not very productive and has a low entropy.

- If a term has only one dominant MWT, it occurs in the form of a fixed expression and has a low entropy.

For calculating the conditional probabilities, we simply took the frequency of a multi-word term $m$ matching the simple-word term $t$ and divided this frequency by the frequency, for a given year, of all $n$ multi-word units pertaining to $t$. This is shown in Formula 2. Here, $f(m)$ denotes the absolute frequency of $m$.

$$p_{m,y} = \frac{f(m)}{\sum\limits_{i=1}^{n} f(m_i)} \qquad (2)$$

## 3.2 Data

All work was carried out on the ACL ARC (Bird et al., 2008), analysed for term occurrences by Zadeh and Handschuh (2014). The corpus was encoded into CWB (Evert and Hardie, 2011) and annotated for terminology from the reference list provided by Zadeh and Handschuh (2014) by means of simple, context-insensitive string matching. This data set was then queried for occurrences of single-word terms. For each year, we extracted frequency information for all single-word terms with an overall absolute frequency of at least 100. This yielded a list of 679 term lemmas. We also extracted frequency-per-year information for multi-word terms, using a regular expression. For calculating productivity, we then had to map multi-word onto single-word units. This was again done with a rather simple string matching procedure and reduced the list of single-word terms under study to 424, since for many terms (e.g. "adaboost", "adjunction", "axiomatization") we did not find any dependent multi-word unit.

## 3.3 Pilot Study

Picton's typology of diachronic term development patterns does not seem fully convincing since it is, at least, in danger of mixing various levels of analysis (terms, topics, textual aspects). We therefore decided to carry out a pilot study on our data to develop a better understanding of the kinds of dynamics that can be expected to be found. This was done by plotting term frequency and productivity for a number of terms. As a result of this study, we
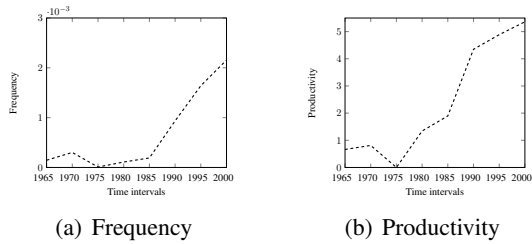
(a) Frequency       (b) Productivity

Figure 1: Frequency and productivity graphs for "corpus".



(a) Frequency       (b) Productivity

Figure 2: Frequency and productivity graphs for "score".



(a) Frequency       (b) Productivity
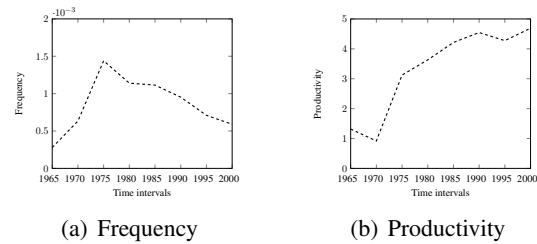
Figure 3: Frequency and productivity graphs for "representation".

expect to find three types of dynamics: growing terms, consolidated terms, and terms in decline.

### 3.3.1 Growing Terms

*Growing* terms exhibit an ongoing increase of both productivity and frequency in 2006, the last year of data in the ACL ARC, that is, none of the two curves has yet started to visibly converge to some maximum. Figure 1 shows frequency and productivity values for "corpus", averaged over intervals of 5 years.[3] Besides "corpus", "cluster", "classification" and "feature" show a similar pattern.

### 3.3.2 Consolidated Terms

*Consolidated* terms still grow in frequency, but not in productivity. One could interpret these terms as belonging to the standard paradigm of computational linguistics (in 2006). Figure 2 exemplifies this for "score": "Scores" are widely cited in many publications, but not many new scores are being developed, while scoring has been the dominant evaluation paradigm already for a while and promises to remain such for the near future. Besides "score", "training" and "translation" exhibit similar patterns.

### 3.3.3 Terms in Decline

*Terms in decline* seem to have reached an upper bound of productivity and are being used less in terms of frequency. Figure 4 shows this for "representation". Such terms might rise again in the future, but in that case, they may already belong to another paradigm, that is, they may have taken on new shades of meaning. Besides "representation", "reasoning" and "grammar" follow a similar pattern.

## 4 Clustering Experiment

### 4.1 Algorithm and Data Representation

To investigate the usefulness of our model for the study of the research questions posed above and to verify the hypotheses derived from the pilot study, we carried out a clustering experiment. The aim was to check whether it is possible to sort the data into three clusters of terms, namely "growing", "consolidated", and "in decline". For this purpose, we used the R implementation (R Core Team, 2013) of the Hartigan and Wong k-means clustering algorithm (Hartigan and Wong, 1979) with 3 centres. Standardized frequency and productivity values for each year and term were passed to the algorithm as a feature vector, each value representing a distinct feature.

### 4.2 Evaluating Clustering Quality

A series of 20 models with 3 centers was calculated. To select the optimal model, we manually labelled all of our 424 observations according to the criteria shown in Table 1. Table 2 shows the distribution of the labels in our data. We do not believe these labels to represent real classes of terms, since the criteria "largest frequency" and "largest productivity" are certainly insufficient for classification. However, we used these labels for approximating the true class distribution when selecting the most reliable from our series of 20 models.

---

[3]6 years for 2000-2006. In the plot, x axis ticks denote the first year of the interval.

4

|  | Largest Frequency | Largest Productivity | Label |
|---|---|---|---|
| Year | 2005-2006 | 2005-2006 | g(rowing)-g |
|  |  | 1990-2004 | g-c(onsolidated) |
|  |  | earlier than 1990 | g-d(ying) |
|  | 1990-2004 | 2005-2006 | cg |
|  |  | 1990-2004 | cc |
|  |  | earlier than 1990 | cd |
|  | earlier than 1990 | 2005-2006 | dg |
|  |  | 1990-2004 | dc |
|  |  | earlier than 1990 | dd |

Table 1: Manual labels for data.

| Label | Number | % |
|---|---|---|
| cc | 118 | 28 % |
| dc | 105 | 25 % |
| cd | 58 | 14 % |
| dd | 51 | 12 % |
| cg | 36 | 8 % |
| dg | 31 | 8 % |
| gg | 15 | 4 % |
| gc | 7 | 1 % |
| gd | 3 | 1 % |

Table 2: Label distribution in data.

Evaluation of clustering results was then performed by means of a simple variation of accuracy calculation: For each label, we assumed that the cluster with the majority of observations represented the "real" class for this label. Accuracy was calculated for each label as the proportion of correct class assignments and overall accuracy was calculated as the average over all 9 labels. Since this leads to overestimation for labels with only a few observations (e.g. *gd*), we also devised a weighted accuracy score.

### 4.3 Best Model

Our best model reached 84 % of accuracy (weighted accuracy: 75 %) and distributes labels over clusters as shown in Table 3. From the table it appears that there is a rather neat distinction between cluster 1 – terms with "dying" frequencies, that is, terms whose largest relative frequency was observed before 1990 – and cluster 3: terms with active or, at least, consolidated productivity values. Cluster 2 is more difficult to interpret. The last row of the table also shows that the terms are distributed relatively evenly over the three clusters.

| Label | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| cc | 15 | 20 | 83 |
| cd | 4 | 54 | 0 |
| cg | 5 | 0 | 31 |
| dc | 69 | 18 | 18 |
| dd | 14 | 37 | 0 |
| dg | 20 | 4 | 7 |
| gc | 0 | 0 | 7 |
| gd | 0 | 3 | 0 |
| gg | 0 | 0 | 15 |
| **Terms** | **127** | **136** | **161** |

Table 3: Best model clustering result.

### 4.4 Typical Terms

So far, our results seem to confirm the existence of a term life cycle with distinct stages such as growth and decline. However, from a digital humanities point of view, it is more interesting to identify "typical" terms for each cluster. We did this by calculating, for each term, its Euclidean distance from the center of its respective cluster. This is shown in Formula 3, where $e$ is the Euclidean distance for each term, $f$ is its feature vector and $c$ is the vector representing the cluster center. $n$ is the number of features passed to the function.

$$e = \sqrt{\sum_{i=1}^{n}(f_i - c_i)^2} \qquad (3)$$

Table 4 gives an overview of the resulting typicality ranking for each of the three clusters. The table displays the terms with the 10 shortest distances from the center (for each cluster) and the terms with the 5 largest distances. The distance values are also given. Columns F and P display the year in which a given term reached its highest frequency or productivity value, respectively. Figures 4 and 5 plot standardised frequency and productivity values for the top-3 terms for clusters 1 and 3 against the cluster centers (labelled as "Cluster 1" and "Cluster 3", respectively).

## 5 Interpretation of Results

### 5.1 Results of First Experiment

The results presented in the previous sections confirm that "typical" terms for cluster 1 are indeed terms with a long-standing history. Many of them were used more actively in the 1970s and 1980s than in later years. Some of them indeed exhibit

| Cluster 1 | | | | Cluster 2 | | | | Cluster 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Terms | Distance | F | P | Terms | Distance | F | P | Terms | Distance | F | P |
| interpretation | 5.16 | 1981 | 2006 | report | 4.50 | 2005 | 1965 | annotation | 3.33 | 2004 | 2004 |
| parsing | 5.38 | 1983 | 1992 | anchor | 4.51 | 1995 | 1965 | corpus | 3.40 | 2005 | 2006 |
| representation | 5.43 | 1975 | 1998 | lexicalization | 4.59 | 1994 | 1965 | cluster | 3.41 | 2002 | 2006 |
| process | 5.44 | 1975 | 2004 | internet | 4.60 | 2004 | 1965 | smooth | 3.49 | 2006 | 2006 |
| syntax | 5.45 | 1980 | 2005 | unigram | 4.60 | 2003 | 1965 | classifier | 3.50 | 2003 | 2006 |
| formalism | 5.49 | 1987 | 1992 | synset | 4.65 | 1998 | 1965 | ranking | 3.51 | 2004 | 2004 |
| case | 5.56 | 1983 | 1994 | perplexity | 4.69 | 1989 | 1965 | method | 3.61 | 2003 | 2006 |
| backtrack | 5.59 | 1987 | 1965 | collocate | 4.71 | 1998 | 1965 | n-gram | 3.64 | 2005 | 2004 |
| semantic | 5.61 | 1982 | 2006 | pcfg | 4.72 | 1999 | 1965 | measure | 3.70 | 2006 | 2000 |
| understanding | 5.62 | 1975 | 1994 | cd-rom | 4.73 | 1999 | 1965 | corpora | 3.71 | 2002 | 2006 |
| … | | | | … | | | | … | | | |
| device | 8.23 | 1965 | 2003 | grammaticality | 8.19 | 1989 | 2004 | character | 7.56 | 1980 | 2003 |
| transformation | 8.33 | 1967 | 1998 | phrasing | 8.21 | 1967 | 2002 | hownet | 7.60 | 2002 | 2002 |
| natural-language | 8.33 | 1982 | 1983 | array | 8.26 | 1967 | 1992 | paragraph | 7.62 | 1991 | 2004 |
| linguist | 8.51 | 1969 | 1983 | grouping | 8.32 | 1965 | 1996 | morph | 7.64 | 1965 | 2001 |
| comprehension | 8.61 | 1978 | 1983 | concordance | 8.37 | 1969 | 1997 | summarizer | 7.64 | 2000 | 2002 |

Table 4: Typical terms for all three clusters.

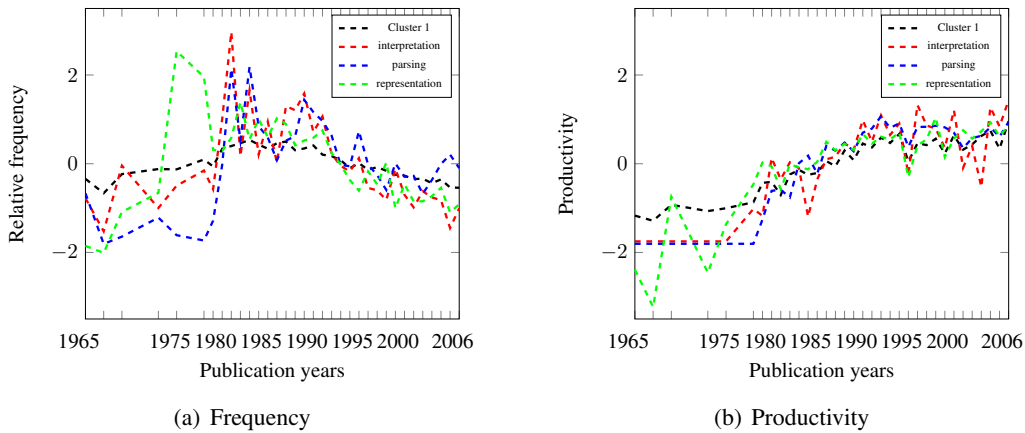

(a) Frequency



(b) Productivity

Figure 4: Frequency and productivity development for typical terms in cluster 1.

decreasing productivity, so they can really be considered terms "in decline". Others, such as "syntax" or "interpretation", seem to have lost importance in terms of frequency, however, they continue to give rise to multi-word terms and they may also have taken on new or other shades of meaning over the intervening years [4]. For these reasons, it might be reasonable to consider them "consolidated" terms rather than terms "in decline", that is, these terms form a part of the standard vocabulary of computational linguistics. Table 5 in the appendix seems to support this interpretation. While some of the top-50 terms for cluster 1 seem indeed outdated (e.g. "prolog"), others denote research topics that were more active in the past (e.g. "formalism", "grammar"), but still cannot be considered irrelevant today. Still others seem to be part of the background vocabulary without which computational linguistics cannot exist (e.g. "sentence", "meaning").

The terms typical for cluster 3 exhibit a very different pattern of development. Their history starts in the 1990s (at any rate, not earlier than in the second half of the 1980s). They then rise quickly and steadily and continue to grow in 2006 when our period of observation ends. It seems straightforward to predict further growth for them and, indeed, today, 10 years later, we know that terms like "corpus", "classifier", and "n-gram" still play an important role in computational linguistics research. In fact, Table 5 confirms that the top-50 terms of cluster 3 almost exclusively represent the statistical paradigm of computational linguistics and we are actually surprised that they are so easily identifiable. These terms almost seem to constitute a kind of newspeak that is associated not only to new topics, but also to new methods and, possibly, a new generation of researchers.

Last but not least, cluster 2 is not as easily in-

---

[4] Note that terms with 1965 as the most "productive" year actually have 0 productivity over the whole period of observation.
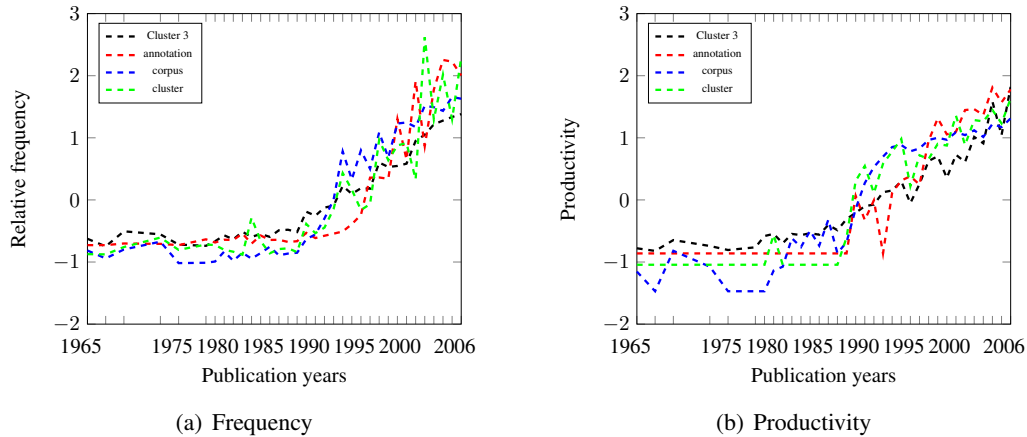
(a) Frequency

(b) Productivity

Figure 5: Frequency and productivity development for typical terms in cluster 3.

terpretable. Many of the terms in this cluster actually have zero productivity over the whole period of observation (for example, "unigram" has 1965 as the year of its "largest" productivity, meaning that the 0 value (=0 or 1 collocation(s)) set for this year was not overwritten by any larger value in any of the following years). We believe that this is, at least, in part a result of our processing decision to attribute multi-word terms to only one simple term (see Section 3.2 for more detailed information) in order to avoid double-counting. However, it seems that this leads to a loss of relevant information.

## 5.2 Double-Counting

To check the effect of this detail, we ran the experiment a second time, with the double-counting option set: Now, multi-word units could be assigned to more than one single-word term. First of all, this lead to a very considerable increase of the data set that now holds 592 terms[5]. It also contains more "growing" terms (labels containing the letter *g*) and less clearly "dying" ones (label *dd*). Moreover, this slight shift in the data set seems to be echoed in the clustering result in the sense that the cluster of "growing" terms now holds a larger share of the data. Accuracy slightly decreased to 0.80 (weighted: 0.73).

In fact, however, changing how multi-word units are attributed to single-word terms does not affect the general result of the experiment. Table 6 shows that clusters 1 and 3 exhibit only slight changes in comparison to the first experiment. Still, the result looks more convincing than

in the first experiment. For example, terms like "internet", "unigram", "synset", and "perplexity" are now are in cluster 3, as we would expect. Cluster 2 also turns out to be more interesting in this experiment, at least in the sense of being more readily interpretable. Already among the top-10 terms for this cluster we now find:

- Terms with non-standard orthography (e.g. "word-net"). The example term's counterpart "wordnet" is in cluster 3.

- Regional variants of terms that are less popular. An example is "tokenisation". The term has 196 corpus hits. Its counterpart "tokenization" is in cluster 3 and has 1256 corpus hits.

- Infrequent terms such as "sbar" with only 242 hits in the corpus.

- "Terms" that are the result of defective lemmatisation (e.g. "classifiers"). The example term's counterpart "classifier" is in cluster 3.

- Terms that are actually proper names and, therefore, less likely to form multi-word units (e.g. "umls").

Cluster 2, then, really is a residual class of unproductive rather consolidated terms, as was expected after the pilot study. However, it provides interesting insights into the features that distinguish preferred terms from their non-preferred variants. We also believe that the finding that proper names are less likely to form multi-word units – if it can be shown to hold in general – can be useful in entity recognition.

---

[5]Note, however, that this does not mean that fewer multi-word units were considered in the first experiment. They were just attributed to a smaller set of single-word terms.

7

# 6 Conclusion

It is tempting to discuss the 2006-state of computational linguistics on the basis of our results, however, we leave this discussion to digital humanities researchers. As a side note, we only remark that our results clearly illustrate the rise of the statistical paradigm and the extent to which it has lead to the creation of not only new methods for doing computational linguistics, but also of a new language to talk about it. In fact, the right-hand sides of Tables 5 and 6 seem to be slightly more uniform in their concentration on mathematical methods than the left-hand sides of the tables which present a mixture of linguistic topics, discussions of processing problems ("prolog", "disk", "processor", etc.), and methods that used to be more important in the more distant past. It would be an interesting research task to investigate whether this apparent increase in uniformity can be confirmed in a large-scale study and, if this is the case, how it relates to the Kuhnian notion of "normal science" (Kuhn, 1962). With regard to the research questions posed in the beginning of this paper we find the following:

1. Our study confirms that terms, their semantics and relevance for a domain, change over time, and that frequency and productivity are useful parameters for the description of such changes. Consolidation and growth seem to be common term development patterns. However, there certainly must be more features than the two used here (e.g. those used in trend research), or more types of development patterns, since our clustering experiment did not result in a clean separation of the three expected classes. We also find that terms remain productive in many cases even if they are used less. Extinction, then, may actually be an exceptional case: Knowledge develops continuously and complete ruptures are uncommon.

2. It seems relatively straightforward to predict future growth for terms with a stable growth pattern. In our experiments, growth patterns were identified with simple methods, however, our approach is not able to predict disruptive, sudden changes in a domain. On the other hand, there is no reason why state-of-the-art terminological methods should not be combined with our method for an in-depth analysis of terms, their development, and their relations. In our current experiments, we did not even look at features such as term co-occurrence, linguistic patterns, etc., but we plan to do so in the future. Finally, studying the interactions between various features might be beneficial for the development of more powerful applications. For example, one might hypothesize that a sudden increase of term productivity is a predictor of a future frequency increase. Clearly, more work is wanted in that direction.

3. Clustering seems to be quite useful for finding terms with similar trajectories and we believe that our method can be used in conjunction with co-occurrence-based approaches, in particular, for the purpose of search space reduction. We expect that more sophisticated modelling will lead to even more interesting results – especially with respect to the modelling of semantically related terms.

## References

Kartik Asooja, Georgeta Bordea, Gabriela Vulcu, and Paul Buitelaar. 2016. Forecasting Emerging Trends from Scientific Literature. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association.

Iana Atanassova, Marc Bertin, and Philipp Mayr, editors. 2015. *Mining Scientific Papers: Computational Linguistics and Bibliometrics*. Co-located with ISSI 2015. CEUR Workshop Proceedings.

Olga Babko-Malaya, Andy Seidel, Daniel Hunter, Jason C. HandUber, Michelle Torrelli, and Fotios Barlos. 2015. Forecasting Technology Emergence from Metadata and Language of Scientific Publications and Patents. In *Proceedings of the 15th International Conference on Scientometrics and Informetrics (ISSI 2015)*, Istanbul, Turkey. Boğaziçi Universitesi.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association.

Georgeta Bordea. 2013. *Domain adaptive extraction of topical hierarchies for Expertise Mining*. Ph.D. thesis, National University of Ireland, Galway.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, UK. University of Birmingham.

Pamela Faber and Marie-Claude L'Homme. 2014. Lexical semantic approaches to terminology. Special issue of Terminology, 20 (2).

Pamela Faber, Pilar León-Araúz, and Arianne Reimerink. 2016. Ecolexicon: New Features and Challenges. In *GLOBALEX 2016: Lexicographic Resources for Human Language Technology*, co-located with LREC 2016, Portorož, Slovenia. European Language Resources Association.

Gil Francopoulo, Joseph Mariani, and Patrik Paroubek. 2015. Nlp4nlp: The Cobbler's Children Won't Go Unshod. *D-Lib Magazine: The magazine of Digital Library Research*, 21(11/12).

Gil Francopoulo, Joseph Mariani, and Patrick Paroubek. 2016. Predictive Modeling: Guessing the NLP Terms of Tomorrow. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association.

J. A. Hartigan and M. A. Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 1(28):100–108.

Thomas S. Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.

Joseph Mariani, Patrick Paroubek, Gil Francopoulo, and Olivier Hamon. 2014. Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland. European Language Resources Association.

Fergal Monaghan, Georgeta Bordea, Krystian Samp, and Paul Buitelaar. 2010. Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food. In *Proceedings of the 9th International Semantic Web Conference, Semantic Web Challenge*, Shanghai, China.

Aurelie Picton. 2011. Picturing short-period diachronic phenomena in specialised corpora: A textual terminology description of the dynamics of knowledge in space technologies. *Terminology*, 17(1):134–156.

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Anne-Kathrin Schumann and Stefan Fischer. 2016. Compasses, Magnets, Water Microscopes: Annotation and Analysis of Terminology in a Diachronic Corpus of Scientific Texts. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association.

Anne-Kathrin Schumann and Behrang QasemiZadeh. 2015. Tracing Research Paradigm Change Using Terminological Methods: A Pilot Study on "Machine Translation" in the ACL Anthology Reference Corpus. In *Proceedings of 11th International Conference on Terminology and Artificial Intelligence*, Granada, Spain. CEUR Workshop Proceedings.

Eugen Wüster. 1979. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*, volume 1: Textteil of *Schriftenreihe der Technischen Universität Wien; 8*. Springer, Wien.

Behrang Q. Zadeh and Siegfried Handschuh. 2014. The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, co-located with COLING 2014, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

| Cluster 1 | Cluster 3 |
| --- | --- |
| interpretation | annotation |
| parsing | corpus |
| representation | cluster |
| process | smooth |
| syntax | classifier |
| formalism | ranking |
| case | method |
| backtrack | n-gram |
| semantic | measure |
| understanding | corpora |
| mechanism | optimization |
| logic | estimation |
| theory | regression |
| knowledge | precision |
| prolog | learn |
| concept | annotator |
| interface | validation |
| instantiation | document |
| ambiguity | evaluation |
| meaning | entropy |
| user | model |
| parse | score |
| analyser | prune |
| grammar | token |
| predicate | train |
| implementation | label |
| structure | algorithm |
| verb | summarization |
| reasoning | training |
| denotation | approach |
| hardware | sampling |
| discourse | ontology |
| signal | statistic |
| generation | distribution |
| debug | probability |
| quantifier | weighting |
| inferencing | co-occurrence |
| procedure | approximation |
| disk | tag |
| event | markup |
| unification | disambiguation |
| utterance | chunk |
| sentence | word |
| synthesis | nlp |
| vocabulary | mining |
| inheritance | bigram |
| fact | technique |
| linguistic | likelihood |
| conjunct | bootstrapping |
| inference | voting |

Table 5: Top-50 terms for clusters 1 and 3 in first experiment.

| Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|
| grammars | word-net | clustering |
| parsing | sbar | annotation |
| interpretation | svms | ranking |
| logic | bagging | smooth |
| formalism | grounding | precision |
| mechanism | classifiers | learning |
| process | tf-idf | classifier |
| theory | umls | corpus |
| case | runtime | n-gram |
| interface | tokenisation | rank |
| parser | negra | bootstrap |
| semantic | k-nn | cluster |
| representation | retrieve | method |
| structure | minipar | regression |
| understanding | collocate | measure |
| knowledge | interoperability | treebank |
| unification | hmms | cross-validation |
| parse | f-score | entropy |
| ambiguity | adaboost | wordnet |
| processing | caching | corpora |
| prolog | technologies | segmentation |
| meaning | knn | optimization |
| concept | recogniser | learn |
| user | ptb | annotator |
| grammar | basque | bootstrapping |
| implementation | comlex | label |
| generation | tokenizer | unigram |
| verb | cd-rom | estimation |
| reasoning | genia | weighting |
| mean | television | document |
| predicate | collapse | tagging |
| message | word-segmentation | validation |
| syntax | usability | model |
| discourse | synchronization | evaluation |
| database | standardization | prune |
| synthesis | nucleus | token |
| signal | superarv | chunk |
| lambda | key-word | backoff |
| spell | measuring | ontology |
| processor | pagerank | nlp |
| understand | parse-tree | sampling |
| composition | lemmatization | summarisation |
| utterance | hypothesize | summarization |
| vocabulary | timeml | score |
| inheritance | nonterminal | chunking |
| natural-language | silence | bigram |
| text-to-speech | questionnaire | training |
| morphologic | translations | algorithm |
| linguistic | katakana | tag |
| event | retirieving | approach |

Table 6: Top-50 terms for clusters 1, 2, and 3 in second experiment.