# Mapping the Natural Language Processing Domain: Experiments using the ACL Anthology

**Elisa Omodei*, Jean-Philippe Cointet**, Thierry Poibeau*****

* LATTICE-CNRS and ISC-PIF, 1 rue Maurice Arnoux – F-92120 Montrouge France (*elisa.omodei@ens.fr*)
** INRA SenS and ISC-PIF, 5 boulevard Descartes 77454 Marne-la-Vallée (*jean-philippe.cointet@polytechnique.edu*)
*** LATTICE-CNRS, 1 rue Maurice Arnoux – F-92120 Montrouge France (*thierry.poibeau@ens.fr*)

**Abstract**

This paper investigates the evolution of the computational linguistics domain through a quantitative analysis of the ACL Anthology (containing around 12,000 papers published between 1985 and 2008). Our approach combines complex system methods with natural language processing techniques. We reconstruct the socio-semantic landscape of the domain by inferring a co-authorship and a semantic network from the analysis of the corpus. First, keywords are extracted using a hybrid approach mixing linguistic patterns with statistical information. Then, the semantic network is built using a co-occurrence analysis of these keywords within the corpus. Combining temporal and network analysis techniques, we are able to examine the main evolutions of the field and the more active subfields over time. Lastly we propose a model to explore the mutual influence of the social and the semantic network over time, leading to a socio-semantic co-evolutionary system.

**Keywords:** ACL Anthology, semantic network, social network

## 1. Introduction

The statistical analysis of large scientific repositories of texts is a popular research theme since the 1960s with the first scientometric studies ("the science of analyzing and measuring science" as defined in Wikipedia). Initial research was mostly based on lists of authors and keywords provided with scientific papers. The growing availability of digital data has put new life into this research field, allowing analysts to build conceptual maps of different media, including scientific text archives. New methods are also being explored:

- Natural language engineering techniques are now available to analyse the text itself (and not only metadata),
- New methods from complex network analysis make it possible to better describe the relations between keywords, authors and their respective evolution over time.

We report here an experiment based on the ACL Anthology. How has the field evolved? What have been the more active research areas over time? Can we predict how it could evolve? These are some of the questions we would like to investigate.

We take the ACL Anthology as an example, but the method is of course reproducible for other domains as well. The availability of large archives containing documents spanning over large periods of time makes it possible to observe the dynamics of ideas over time. It is especially the case in the scientific field, where researchers produce a prolific literature: briefs, research reports, scientific papers, etc. For a growing number of domains, large scientific archives are now available over several decades.

These scientific archives have already given birth to a large body of research. Collaboration networks have for example been automatically extracted so as to study the topology of the domain (Girvan and Newman, 2002) or its morphogenesis (Guimera et al., 2005). Referencing has also been the subject of numerous studies on inter-citation (Garfield, 1972) and co-citation (Small, 1973). Other variables can be taken into account like the nationality of the authors, the projects they are involved in or the research institutions they belong to, but it is the analysis of the textual content (mostly titles, abstracts and keywords provided with the papers) that have attracted the most part of the research in the area since the seminal work of Callon (Callon et al., 1986; Callon et al., 1991).

The ACL Anthology is a digital archive of conference and journal papers in natural language processing and computational linguistics. It has received recent attention thanks to the ACL Special Workshop "Rediscovering 50 Years of Discoveries" in 2012, that produced a few papers on the subject, among which a "history" of the field by Anderson and collaborators (Anderson et al., 2012). Since we would like to build on this work, we selected for our experiments the 12,000 articles published in the period 1985-2008 already used by Anderson and his colleagues.

Our study is complementary to the one of (Anderson et al., 2012). The similarities lie of course in the corpus but also in the method based on a joint study of the semantic network (keywords linked together in order to give a semantic description of the field) and the social network (authors publishing together). However Anderson et al. are mainly interested in extracting key facts from the data (what we call the macro-structure, e.g. to what extent US funding shaped the domain?) whereas we are more interested in mapping the domain (providing different maps to observe the domain from a static as well as a dynamic point of view). Modeling the possible evolution of the domain is also of high interest for research planning and observation.

The rest of the paper is structured as follows. We first propose a map of the domain based on the automatic extraction of relevant keywords. We then propose a technique to represent the evolution of the domain over time. Lastly,

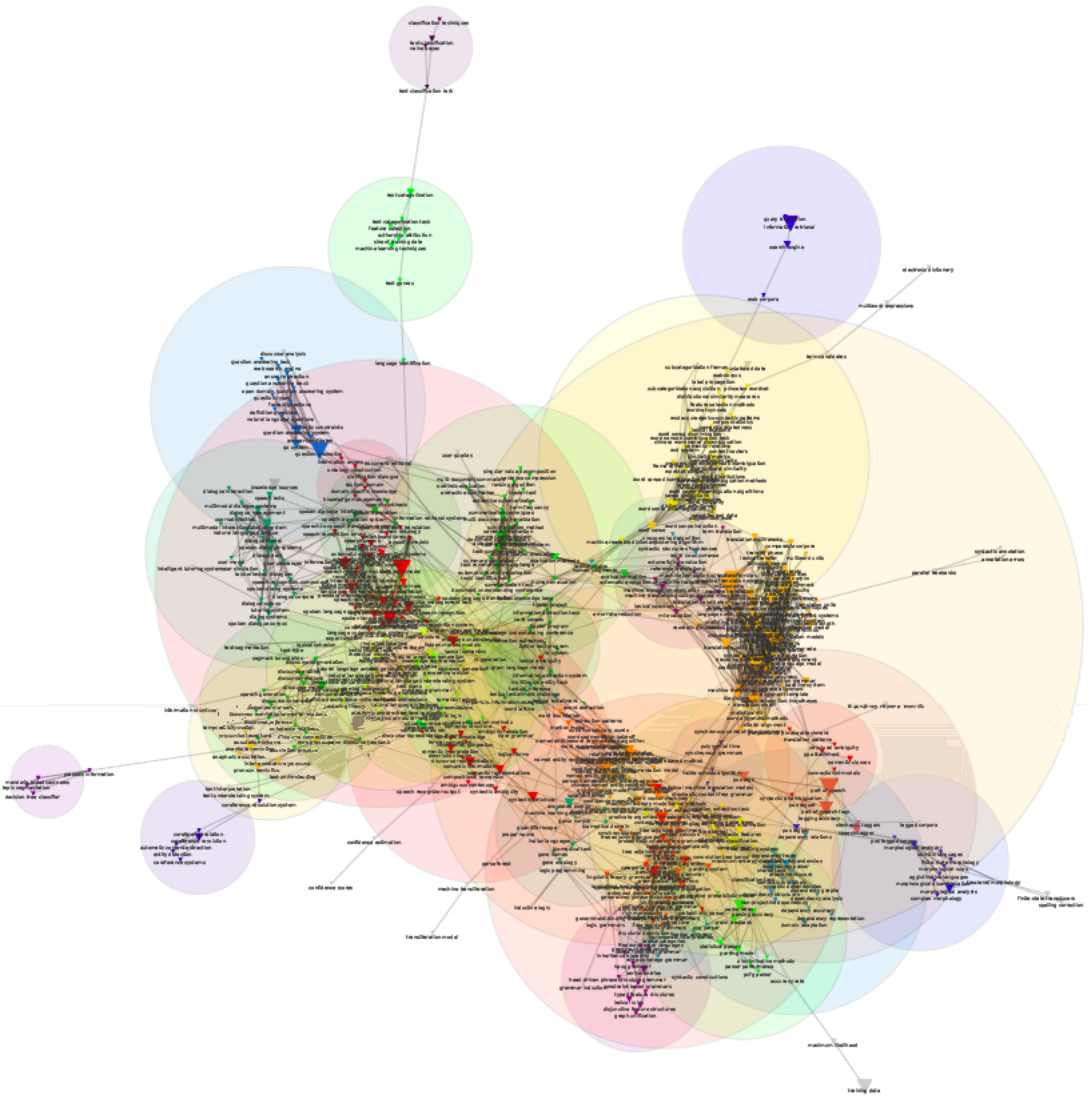

Figure 1: Semantic map obtained from the automatic analysis of the ACL Anthology.

we propose some elements for the definition of a predictive model mixing the semantic and the social network under study.

## 2. Mapping the computational linguistics domain

We first need to define a "semantic map" of the computational linguistics domain, which means divide the domain into different homogeneous subfields and identify their most specific keywords. We do this thanks to an inductive method, extracting keywords (i.e. multiword expressions) from the papers and then clustering these keywords so as to obtain a semantic map of the domain.

### 2.1. Keyword extraction

The first step consisted in automatically extracting the terms that correspond to concepts and methods of the field from titles and abstracts of the papers. To achieve this we pre-process the text with POS-tagging, chunking, normalization and stemming. Noun phrases are automatically extracted and multi-terms that convey a certain semantic unit, i.e. those with the highest "unithood", are selected following the method defined by (Van Eck and Waltman, 2011). To sort the list of candidate terms and select the most relevant ones we then combine together two statistical criteria: first we estimate the quality of each term using the c-value method (Frantzi and Ananiadou, 2000) and then we remove irrelevant multi-terms with low "termhood", i.e. terms that may occur very frequently but do not help characterize the content of the paper, such as "review of literature" or "past articles". All these techniques are implemented and available online through the Cortext platform (http://docs.cortext.net/lexical-extraction/).

Since the number of papers published every year increases, our dataset naturally contains more recent papers than older ones. In order to avoid excluding concepts that were popular a few decades ago but are not so much now (and would then be relatively infrequent in the dataset as a whole), we divided the papers set in three time slices, and extracted a list of 1000 terms from each subset (some more fine grained divides are of course possible). We then merged the three

lists (since as expected the three lists have some terms in common) and eliminated all the terms that appeared in less than five papers, obtaining a list of about 1500 terms. We then showed the list to an expert of the field who validated it manually, eliminating all the terms that were too general (like "computational linguistics") or not relevant, and producing as a result a final list of 673 terms describing the concepts and methods of the field.

### 2.2. Mapping the domain

Starting from this list we then construct a semantic map of the field through a network-based approach. The nodes of the network are the extracted terms. Two terms are connected if they co-occur in the same title or abstract at least once. The links are weighted using Mutual Information, defined as the logarithm of the ratio between the number of joint occurrences of the two terms in the same document and a measure of the expected number of co-occurrences between them. In order to have the best readable clear-cut network, we then eliminate all the links whose weight is lower than a threshold defined so as to avoid the network to split into multiple connected components (consisting of more than three nodes).

The goal is to obtain a network consisting of several densely connected components of concepts co-occurring together because they belong to the same subfield of the discipline. Through this analysis we expect to get a map in which the different subfields of natural language processing and computational linguistics naturally emerge. We thus apply an algorithm for community detection of graph: such algorithms are used to partition a network into groups of nodes which are densely connected among each other and loosely connected with the rest of the network (a technique also known as clustering). In this study we use Infomap (Rosvall and Bergstrom, 2008), which is found to be one of the best algorithm for the task (we also tried the algorithm from Louvain (Blondel et al., 2008) that obtained slightly worse results to map the domain as a whole but can be interesting to analyze the evolution of the domain over time, see next section). The semantic network obtained is shown in Figure 1. Each circle surrounds a detected "community", each representing a thematic cluster, such as word sense disambiguation or POS tagging.

### 2.3. Evaluation

Different clustering techniques (Infomap and Louvain, cf. supra) and different settings have been tried and qualitatively evaluated by an expert of the field. Additionally, for each cluster we randomly selected 10 projected articles and the expert had to evaluate whether each article fitted well in the cluster. We then computed the precision of a cluster as the fraction of relevant articles. The average precision obtained is 0.84, which is judged acceptable for this kind of task.

Below are three examples of clusters automatically obtained with the method described:

**Cluster 1:** entity detection - coreference relation - Automatic Content Extraction - coreference resolution - coreference resolution system - coreference system

**Cluster 2:** Sentence Compression - text summarization system - term frequency - Document Understanding Conference - human judgments - sentence extraction - TIPSTER Text - topic identification - automatic text summarization - Automatic Summarization - multi-document summarization - extractive summaries - ranking algorithm - evaluation methods - text summarization - summarization method - summary generation - human evaluation - summarization evaluation - Text Summarization Challenge - document summarization - summarization system - summarization techniques - evaluation metrics - summarization task - Singular Value Decomposition - extractive summarization

**Cluster 3:** natural language understanding system - semantic lexicon - lexical knowledge base - Montague grammar - temporal expressions - lexical semantics - semantics

## 3. Mapping the evolution of the domain

We now want to describe the main evolutions of the domain of computational linguistics over time, which means representing the relative importance of the different subfields over time. We would like to know what subfield has attracted the most important part of the research effort, which subfields have merged or appeared during the period, etc.

### 3.1. Method for the analysis

The approach to perform this analysis can be divided into four different steps.

1. the corpus is divided over different periods of time;
2. all the papers related to a given period are put together and keywords are extracted as explained in the previous section;
3. clustering algorithms are applied over the set of keywords so as to obtain clusters of keywords representing the different subfields of the domain;
4. lastly, the different subfields identified for each period are mapped over time.

The clustering algorithms used are Infomap (Rosvall and Bergstrom, 2008) and Louvain (Blondel et al., 2008) as said in the previous section.

The mapping of subfields over time is a challenging operation since all subfields evolve: keywords may disappear from a given cluster and new keywords may be added just because the techniques evolve. The issue is then to determine to what extent two clusters represent the same subfield or not.

Basically, two clusters are connected if they share enough common keywords. A threshold has to be defined so as to avoid connecting clusters sharing too few keywords over time. Note that this simple approach makes it possible to match one cluster *c* at a period of time *t* with one cluster *c'* at period *t+1* but also to associate one cluster *c* with two clusters *c'* and *c"* at period *t+1*: this is typically the case when one subfield gives birth to two different subfields sharing themselves few keywords together (for example we observe that the cluster corresponding to message understanding gives birth to two subfields: named entity recognition and information extraction; these are considered as

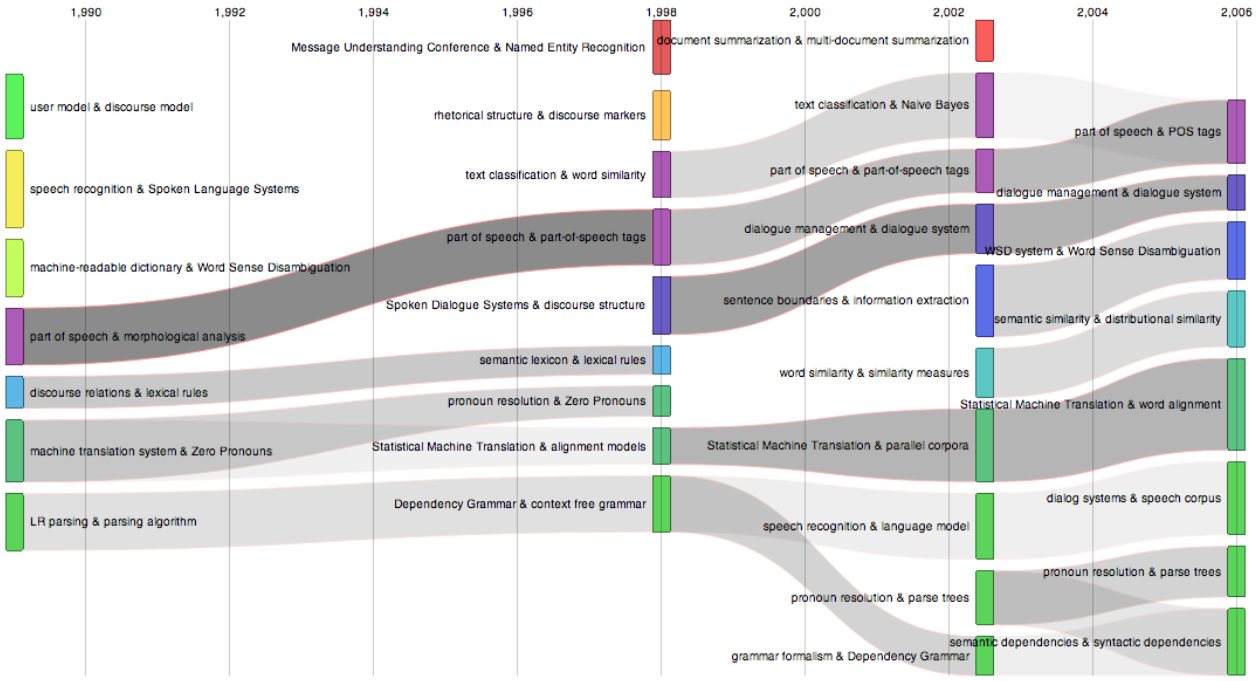


Figure 2: Observation of the evolution of the computational linguistics domain over time at the macro-scale.

two different subfields since the automatic keyword analysis reveals that they contain few keywords in common). The reverse operation can be observed when two subfields give birth to a unique new subfield merging techniques from the two previous subfields (for example statistical parsing and dependency grammar merging to give birth to the field of statistical dependency parsing). Lastly, when no correspondence can be found, the subfield is supposed not to survive in itself.

An extensive description of the techniques used can be found in (Chavalarias and Cointet, 2013), as well as all the implementation details. For our experiments, we used the Cortext platform which implements all the procedure and provides various choices for each step (the platform implements different techniques for keyword extraction, keyword clustering and cluster mapping over time). These alternative choices mean that various maps can be obtained for a same domain, providing different views over the evolution of the domain.

It must be noted that the different algorithms will provide different maps. These maps do not always show the same results, especially when looking at the details. There is no "good" or "bad" map but there are different maps, giving different views of the domain. Of course, the representation must be checked carefully and interpreted: for example if a cluster is not connected to any other cluster, it does not directly means that the subfield has disappeared. It may have largely evolve so that at period $t+1$ no cluster contains enough common keywords to be connected to the original cluster $c$. It may have merged with two different other subfields with few keywords in common overall, etc. The map should be considered as a way to kick-start the analysis, not as a definitive result per se.

Different maps should be produced to examine the "threshold effect". For example, two clusters may not be connected on one map but may be connected on another map generated with only a small variation in the parameter settings, which means that the change between the two observed periods of time is probably not as radical as one map may suggest.

For example, it can be desirable to consider smaller or larger periods of time. The domain can also be divided into a smaller or larger number of subfields, depending on the granularity that one wants to observe in the end. Broad representations (maps considering fewer clusters and fewer periods of time) will highlight the main tendencies of one field while detailed descriptions will allow one to reconstruct the precise phylogeny of a domain.

### 3.2. Results

We provide here three maps showing the evolution of the computational domain from the late 1980s to nowadays.

Figure 2 shows the major trends in the evolution of the domain. Each period consists in approximately 8-12 clusters showing the evolution of the main research subfields over time (but note that the number of clusters is the result of the parameter settings but one cannot directly define the number of clusters per time using the clustering techniques implemented for this study). Only clusters sharing a relatively large number of keywords are connected through grey tubes. We observe that the main field is now machine translation: this field has continuously increased since the late 1980s. We can also observe the development of the Question answering task since the late 1990s: this field has been especially popular at the time thanks to the QA evaluation tracks at TREC for example.

Even if they have been reincarnated in a way or another,

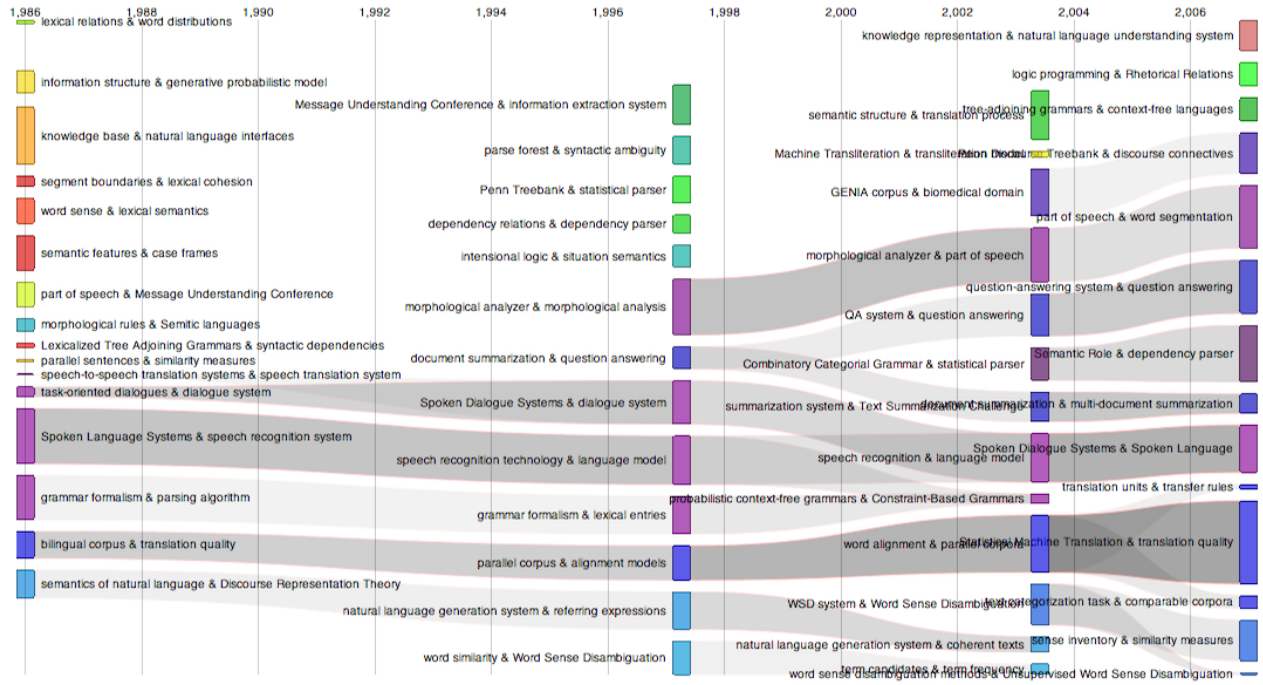


Figure 3: Observation of the evolution of the computational linguistics domain over time at the meso-scale.

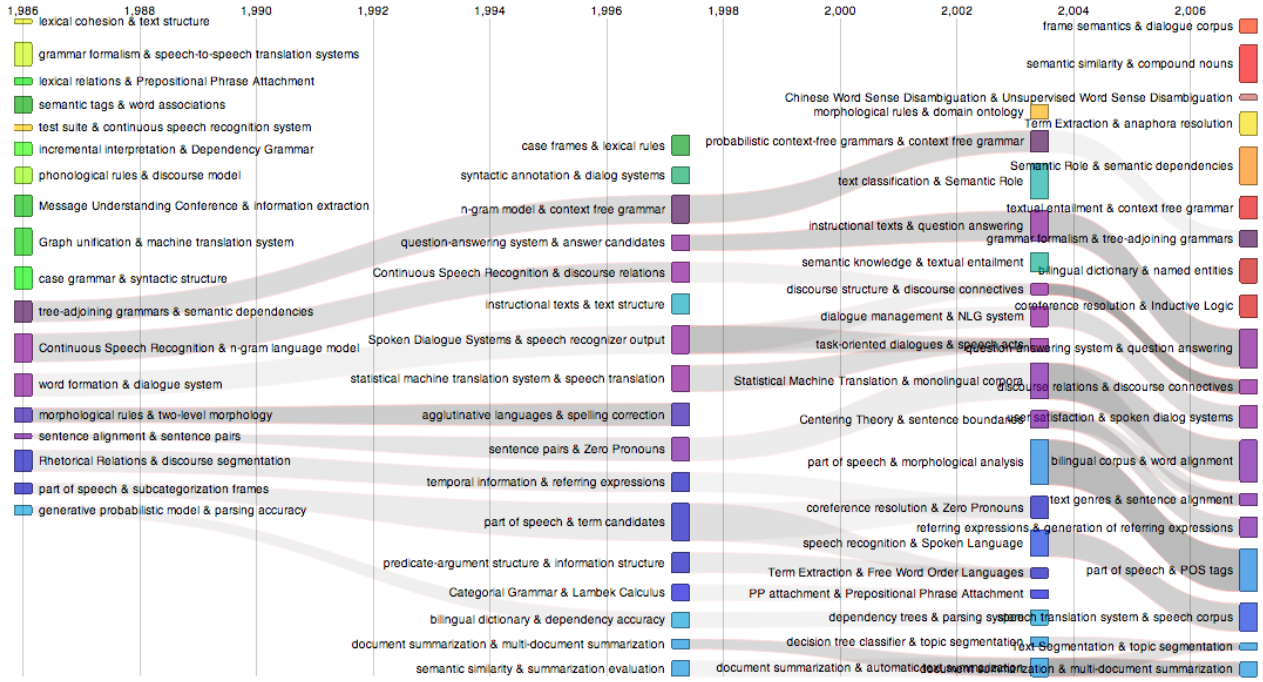


Figure 4: Observation of the evolution of the computational linguistics domain over time at the micro-scale

we can see different isolated subfields. Machine readable dictionary was a popular research field in the 1980s and has since then been outdated by the rise of corpus-based studies. Message understanding is shown as being typical of the 1980s and 1990s (the field is now known as Information extraction and the techniques used are quite different, hence the lack of continuity on this map). The continuous interest in word sense disambiguation does not directly appear since machine learning approaches have considerably renewed the approach: we observe a discontinuity between the rule-based approach largely used in the 1980s and 1990s and the machine learning techniques used since the late 1990s.

Figure 3 and figure 4 are much more precise overviews of the domain. We can observe for example the fact that spoken dialogue merged with statistical machine translation to give birth to a new field of research combining the two approaches for task-oriented dialogue interfaces. Speech also

merged with the discourse subfield at the end of the 1990s which shows a new interest in the management of dialogue structures, etc.

## 4. Toward a predictive model mixing the semantic and the social network

The goal of our study is to understand how the social and the semantic structures of a research community are driving future research dynamics. Previous studies have mainly focused on one of these two dimensions, but we claim that both are fundamental in the evolution of scientific research communities. The goal of the following analysis is to try to quantify the contribution of each.

In order to answer such questions, we first need to characterize the "social features of the domain". We do this by introducing a second graph called "social network", in which the nodes are the researchers of the field, i.e. authors of the ACL Anthology papers. Two researchers are connected in this network if they co-authored at least one paper.

We first begin by investigating how the social network evolves over time, *i)* depending only on its endogenous characteristics (which represent the social features), and then *ii)* depending on the semantic similarity between researchers. Then, using a similar approach, we study the evolution of the semantic network *i)* depending only on its endogenous characteristics and then *ii)* also depending on the social similarity between concepts.

We first investigated to what extent the probability of a new collaboration (between two researchers who never co-authored a paper before) is affected by their *social proximity*, which is captured by the Jaccard index of common neighbors in the social (co-authorship) network:

$$J(i,j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \tag{1}$$

where $\Gamma(i)$ denotes the set of neighbors of node $i$ in the given network. The index takes values between zero and one, and is equal to zero when two authors have no collaborator in common, and one when they collaborated with the same set of researchers. The Jaccard index is a widely used statistics to measure the similarity of two sets, and for link prediction (Lu and Zhou, 2011).

To quantify the contribution of this variable to the evolution of the social network, we performed a logistic regression in which our input variable is the Jaccard index between two authors in the network (authors who are not directly connected for a given year), and the dependent variable is the presence or not of a link between them (i.e. whether they co-authored a paper or not the following year). The choice of logistic regression is based on the fact that the response variable is binary (a link is either created or not). More precisely, the model takes the following form:

$$\ln(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x \tag{2}$$

where $x$ is the Jaccard index and $p(x)$ the probability of link creation.

The obtained results indicate that the selected variable is significative (p-value $< 2e-16$) and the response is highly correlated with it. We find in fact that the coefficient is equal to $\beta_1 = 10.85 \pm 0.64$, which means that for a tenth-unit increase in the social proximity score, we expect to see the odds of the two authors becoming co-authors in the future increase of about three times.

We also want to investigate whether these odds are correlated with the notion of *semantic proximity*, which we measure in the same way, i.e. as the Jaccard index between the set of concepts used by one author and the set of concepts used by the other one. Therefore we used logistic regression with multiple explanatory variables. The model in this case is defined in the following way:

$$\ln(\frac{p(x_1,x_2)}{1-p(x_1,x_2)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{3}$$

where $x_1$ is the social proximity and $x_2$ the semantic one.

We found a coefficient for this second variable equal to $\beta_2 = 5.71 \pm 0.53$ (p-value $< 2e-16$). This means that, holding social proximity at a fixed value, for a tenth-unit increase in the semantic proximity score, we expect to see the odds of the two authors becoming co-authors in the future increase of almost twice. Moreover, we performed the likelihood ratio test to compare the first model (which considers only the social dimension) with the second one (which takes into account both the social and the semantic dimensions) and found a significant chi-square value of 67 (p-value $= 2.78e-16$), indicating that taking into account also the semantic dimension to study the social evolution of the system does improve significantly the prediction.

In a completely symmetric way, we studied the probability of two concepts not connected before to become connected (i.e. found in the same paper), at first depending on their *semantic proximity* only, and then also of their *social proximity*. In this case the semantic proximity is given by the Jaccard index of common neighbors between two concepts in the previously defined semantic network, and the social proximity by the Jaccard index between the set of authors using one concept and the set of authors using the other one. Also in this case we found significative values for the coefficients (all p-values $< 2e-16$). The coefficient corresponding to the semantic proximity is $12.70 \pm 0.33$, indicating that for a tenth-unit increase in the semantic proximity score, we expect to see the odds of the two concepts becoming connected in the future increase of about three and a half times. The coefficient for the social proximity in the second model is $9.99 \pm 0.78$, indicating an increase in the odds of more than two and half times. Also in this case the likelihood ratio test indicates a significant improvement in the prediction when considering also the second dimension: the chi-squared is 137 and the p-value $< 1.17e-31$. We can therefore conclude that also the semantic evolution of the system is driven by both networks.

The results of all the regression models are summarized in Table 1 and 2. We have shown that, as expected, two researchers are more likely to collaborate in the future if their previous work is in a related area or on the same topic. More precisely we showed that the social aspect has a stronger role, increasing the odds of a new link by three times, whereas the semantic similarity of about twice, which is a lower but still very significant value. Two con-

Table 1: Social evolution logistic regression results.

| model | coefficient | estimate | std error | p-value |
|---|---|---|---|---|
| model I | $\beta_0$ | -6.18425 | 0.04908 | <2e-16 |
| | $\beta_1$ | 10.85418 | 0.63686 | <2e-16 |
| model II | $\beta_0$ | -6.32120 | 0.05272 | <2e-16 |
| | $\beta_1$ | 10.45014 | 0.63149 | <2e-16 |
| | $\beta_2$ | 5.71085 | 0.53517 | <2e-16 |

Table 2: Semantic evolution logistic regression results.

| model | coefficient | estimate | std error | p-value |
|---|---|---|---|---|
| model I | $\beta_0$ | -4.42195 | 0.02543 | <2e-16 |
| | $\beta_1$ | 12.70563 | 0.33345 | <2e-16 |
| model II | $\beta_0$ | -4.45146 | 0.02562 | <2e-16 |
| | $\beta_1$ | 11.82264 | 0.34216 | <2e-16 |
| | $\beta_2$ | 9.99627 | 0.78264 | <2e-16 |

cepts are more likely to become connected if they are already semantically similar, but also depending on the number of researchers who have already worked on both even if in separate publications. Taking these two symmetric results into account, we can suggest that the social and the semantic landscape of computational linguistics mutually influence each other over time so as to produce a kind of socio-semantic co-evolutionary system. Note that we also had recently obtained similar results for the domain of physics (Omodei et al., 2013).

## 5. Conclusions

In this paper, we have tried to analyse the evolution of the domain of computational linguistics between 1988 and 2012. We first extracted from the corpus a list of keywords characterizing the domain, thanks to a mixed approach combining linguistic and statistical information. We then built a "semantic map" of the domain using keyword co-occurrences and graph community detection methods to highlight the different "semantic communities" of the domain. The evaluation by domain experts has shown that we obtained a good representation of the different sub-domains of the field.

We have then explored how the domain evolves over time, through the creation of time-wise semantic maps that show the emergence and evolution of the different areas of research in the field. This analysis gives avenues for the exploration of the main trends in the history of computational linguistics (how new subfields have emerged, how some subfields have merged or nearly disappeared, etc.).

In the last section of the paper we added a social dimension to the semantic analysis of the domain. We showed that the emergence of new social links is strongly influenced by the proximity of the authors both in the social and in the semantic spaces. We observed a two way influence: the emergence of new semantic links is influenced by the proximity in the social network and conversely new social links are closely related to the proximity of the authors in the semantic space. We can thus observe the co-evolution of the social and of the semantic space.


## Aknowledgements

Elisa Omodei is partially supported by a PhD grant provided by the Région Ile-de-France.


## 6. References


Ashton Anderson, Dan Jurafsky, and Daniel A. McFarland. 2012. Towards a computational history of the acl: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21, Jeju Island, Core. Association for Computational Linguistics.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. In *Journal of Statistical Mechanics: Theory and Experiment*.

Michel Callon, John Law, and Arie Rip. 1986. *Mapping the dynamics of science and technology*. McMillan, London.

Michel Callon, Jean-Pierre Courtial, and Françoise Laville. 1991. Co-word analysis as a tool for describing the network of interaction between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22(1):155–205.

David Chavalarias and Jean-Philippe Cointet. 2013. Phylomemetic Patterns in Science Evolution?The Rise and Fall of Scientific Fields. *PLOS One*.

Katarina Frantzi and Sophia Ananiadou. 2000. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Eugene Garfield. 1972. Citation Analysis as a Tool in Journal Evaluation. *Science*, 178(4060):471–479.

Michelle Girvan and Mark E J Newman. 2002. Community structure in social and biological networks. *PNAS*, 99:7821–7826.

Roger Guimera, Brian Uzzi, Jarrett Spiro, and Luis A. Nunes Amaral. 2005. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308(5722):697–702.

Linyuan Lu and Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390.6:1150–1170.

Elisa Omodei, Thierry Poibeau, and Jean-Philippe Cointet. 2013. A Symmetric Approach to Understand the Dynamics of Scientific Collaborations and Knowledge Production. In *Proceedings of the 4th French Conf. on "Modèle & Analyse de réseaux : Approches mathématique & informatiques (MARAMI 2013)"*, Saint-Etienne.

Martin Rosvall and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. In *Proc. Of the National Academy of Sciences*.

Henry G Small. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of American Society for Information Science*, 24(4):265–269.

Nees Jan Van Eck and Ludo Waltman. 2011. Text mining and visualization using vosviewer. *CoRR*, abs/1109.2058.