# Modeling the socio-semantic dynamics of scientific communities

Elisa Omodei

ENS

# Thèse de Doctorat

**En vue de l'obtention du grade de**

# DOCTEUR
# DE L'ÉCOLE NORMALE SUPÉRIEURE

## École doctorale

Transdisciplinaire Lettres/Sciences

**Discipline ou spécialité :**

Mathématiques appliquées aux sciences sociales

---

**Présentée et soutenue par :**

Elisa OMODEI

**le** 19 décembre 2014

**Titre**

## Modeling the socio-semantic dynamics of scientific communities

---

**Unité de recherche** Laboratoire LaTTiCe (UMR 8094)

**Thèse dirigée par** Thierry POIBEAU

**Membres du jury**

Clémence MAGNIEN

Emmanuel LAZEGA

Thierry POIBEAU

Roger GUIMERA

Jean-Pierre NADAL

Jean-Philippe COINTET

Numéro identifiant de la Thèse :

# Abstract

How are the social and semantic structures of a scientific community driving future research dynamics? In this thesis we combine natural language processing techniques and network theory methods to analyze a very large dataset of scientific publications in the field of computational linguistics, *i.e.* the ACL Anthology. Ultimately, our goal is to understand the role of collaborations among researchers in building and shaping the landscape of scientific knowledge, and, symmetrically, to understand how the configuration of this landscape influences individual trajectories of researchers and their interactions. We use natural language processing tools to extract the terms corresponding to scientific concepts from the texts of the publications. Then we reconstruct a socio-semantic network connecting researchers and scientific concepts, and model the dynamics of its evolution at different scales. To achieve this, we first build a statistical model, based on multivariate logistic regression, that quantifies the role that social and semantic features play in the evolution of the socio-semantic network, namely in the emergence of new links. Then, we reconstruct the evolution of the field through different visualizations of the knowledge produced therein, and of the flow of researchers across the different subfields of the domain. To summarize, we have shown through our work that the combination of natural language processing techniques with complex network analysis makes it possible to investigate in a novel way the evolution of scientific fields.



*Keywords:* socio-semantic dynamics, co-authorship networks, semantic networks, automatic term extraction, statistical modeling, computational linguistics

# Résumé

Comment les structures sociales et sémantiques d'une communauté scientifique guident-elles les dynamiques de collaboration à venir ? Dans cette thèse, nous combinons des techniques de traitement automatique des langues et des méthodes provenant de l'analyse de réseaux complexes pour analyser une base de données de publications scientifiques dans le domaine de la linguistique computationnelle : l'ACL Anthology. Notre objectif est de comprendre le rôle des collaborations entre les chercheurs dans la construction du paysage sémantique du domaine, et, symétriquement, de saisir combien ce même paysage influence les trajectoires individuelles des chercheurs et leurs interactions. Nous employons des outils d'analyse du contenu textuel pour extraire des textes des publications les termes correspondant à des concepts scientifiques. Ces termes sont ensuite connectés aux chercheurs pour former un réseau socio-sémantique, dont nous modélisons la dynamique à différentes échelles. Nous construisons d'abord un modèle statistique, à base de régressions logistiques multivariées, qui permet de quantifier le rôle respectif des propriétés sociales et sémantiques de la communauté sur la dynamique microscopique du réseau socio-sémantique. Nous reconstruisons par la suite l'évolution du champ de la linguistique computationelle en créant différentes cartographies du réseau sémantique, représentant les connaissances produites dans le domaine, mais aussi le flux d'auteurs entre les différents champs de recherche du domaine. En résumé, nos travaux ont montré que la combinaison des méthodes issues du traitement automatique des langues et de l'analyse des réseaux complexes permet d'étudier d'une manière nouvelle l'évolution des domaines scientifiques.

*Mots-clés:* dynamiques socio-sémantiques, réseaux de collaboration, réseaux sémantiques, extraction lexicale, modélisation statistique, linguistique computationelle

**Alla mia mamma**

# Acknowledgments

First of all I would like to thank my two supervisors, Thierry Poibeau and Jean-Philippe Cointet, for their scientific guidance and support, but also for their humanity and kindness.

I would also like to thank Jean-Pierre Natal and Emmanuel Lazega for having served as my committee members, and especially Clémence Magnien and Roger Guimerà for their thorough and constructive reviews of my thesis.

Thanks a lot to all my colleagues and friends of the *Institut des Systèmes Complexes de Paris Île-de-France*, which feels quite like a home, of the LaTTice lab, always very welcoming, and of the Complex Systems community, whose enthusiasm and motivation constantly inspire me.

A special thanks to all my friends. Some I have known for many years, some have entered my life more recently. Some are geographically far but always present, some others have become my Parisian family and will be very hard to leave.

Last but not least, I would like to thank my loving and caring family, whose continuous support has allowed me to get this far.

# Contents

# Modélisation des dynamiques socio-sémantiques dans les communautés scientifiques

## 1 Remarque liminaire

Ce document constitue une présentation en français d'une partie essentielle des travaux contenus dans la thèse. Cette présentation n'est pas exhaustive : elle est au contraire limitée à certains résultats qui nous semblent les plus intéressants et les plus caractéristiques parmi ceux obtenus. Afin de produire un document homogène, nous avons aussi fait le choix de ne présenter que des résultats liés au corpus ACL Anthology. Le manuscrit en anglais est évidemment plus complet et rapporte l'essentiel des expériences effectuées pendant la thèse sur les corpus APS et ACL Anthology, ainsi qu'un état de l'art, d'autres expériences et une réflexion générale sur l'apport des techniques explorées. Nous espérons cependant que cette présentation rapide permettra au lecteur francophone qui ne pourrait lire la version anglaise d'avoir une vision synthétique et précise des travaux effectués.

## 2 L'analyse automatique de la littérature scientifique

L'analyse des masses de données (en anglais *big data*) est un thème de recherche porteur aujourd'hui. Les masses de données permettent en effet de mettre au jour des phénomènes difficilement observables sans méthodes automatiques, et la numérisation de tous les secteurs de la société permet aujourd'hui d'avoir accès à des données en grandes quantités pour un grand nombre de domaines.

La science est un des domaines qui produit ainsi de nombreuses données informatisées (littérature scientifique, mais aussi données brutes sous forme de textes, d'images, de chiffres, etc.) et la numérisation des données passées

permet aujourd'hui d'avoir accès, pour plusieurs domaines très variées, à des collections d'articles de recherche s'étendant sur plusieurs dizaines d'années. Le monde de la linguistique informatique n'est pas en reste et l'ACL Anthology met aujourd'hui à la disposition des chercheurs plus de 24500 articles au format PDF. Les plus anciens articles datent de 1965 (première édition de la conférence COLING) mais ce n'est qu'à partir des années 1980 qu'on commence à avoir des données relativement conséquentes, le volume allant grandissant chaque année depuis lors (il y a donc une très grande disparité dans les volumes de données disponibles suivant les années considérées). Il existe des bases de données similaires pour la biologie et le domaine biomédical (par ex. Medline), les systèmes complexes ou la physique (par ex. APS data set de la *American Physical Society*) pour citer quelques bases ayant fait l'objet d'enquêtes diverses.

Ces données ont fait l'objet de nombreux travaux : elles sont ainsi souvent utilisées pour extraire des réseaux de collaboration que l'on construit en liant les auteurs selon des liens de co-publication (Girvan and Newman, 2002) pour mieux comprendre les processus de morphogenèse sous-jacents (Guimera et al., 2005). La structure du réseau de références est au cœur du projet « scientométrique » et a généré de très nombreux développement depuis les premiers travaux sur les réseaux d'inter-citation (Garfield, 1972) et de co-citation (Small, 1973). Encore bien d'autres dimensions d'analyse sont susceptibles d'être employées pour dresser des cartes de domaines scientifiques : données géographiques associées aux publications, institutions de rattachement des auteurs, ou encore projet structurant le travail de recherche. Mais c'est bien le contenu textuel (qu'il provienne des titres, résumés ou des termes utilisés par les auteurs pour étiqueter leurs articles) qui a suscité, avec l'analyse des références, le plus grand nombre de travaux depuis ceux, séminaux, de Callon (Callon et al., 1986, 1991). Dans nos expériences, nous portons une attention particulière aux dynamiques cognitives et donc à l'analyse du contenu textuel. Cette analyse est néanmoins couplée aux trajectoires individuelles des chercheurs dans cet espace conceptuel, ce qui nous permet d'interroger, de façon empirique et à grande échelle, les dynamiques d'innovation dans le champ de la linguistique computationnelle.

L'ACL Anthology a reçu un intérêt particulier en 2012 pour les 50 ans de l'*Association for Computational Linguistics*. Un atelier s'intitulant « *Rediscovering 50 Years of Discoveries* » a été organisé cette année-là (Banchs,

2012) : il s'agissait pour l'association de jeter un regard sur l'évolution du domaine depuis 50 ans. Au-delà de ces circonstances particulières, cet événement a été l'occasion d'analyser les données accumulées depuis 50 ans (mais pour les raisons données plus haut, la plupart des études portent sur les articles produits depuis 1980) avec les outils modernes issus à la fois du traitement des langues et des systèmes complexes, afin d'analyser l'évolution du domaine. L'analyse de ce type de données passe en général par l'extraction d'informations pertinentes (auteurs, termes utilisés, etc.) puis par leur mise en relation : on obtient alors des graphes et les algorithmes développés pour l'analyse des réseaux sociaux peuvent être sollicités. Les relations évoluent au cours du temps : c'est alors à l'algorithmique des graphes évolutifs qu'il faut faire appel. Ces techniques sont mises en œuvre pour répondre à des questions liées à l'histoire des sciences ou, du moins, à l'histoire des différents domaines scientifiques pour lesquels on dispose d'archives conséquentes : on voit donc ici une alliance possible entre le traitement automatique des langues et les systèmes complexes, pour permettre de voir sous un jour nouveau de grandes masses de données qui sont difficiles à analyser sans outils. Les outils permettent de mettre au jour des faits chiffrés et quantifiés, et de vérifier ainsi certaines hypothèses sur l'histoire et l'évolution du domaine, mais aussi sur les techniques utilisées, la mobilité des chercheurs entre différentes thématiques, etc.

L'article « Towards a computational History of the ACL : 1980-2008 » (Anderson et al., 2012) est de ce point de vue très riche. Les auteurs essaient de déterminer les grands domaines de recherche au sein du traitement automatique des langues (TAL) depuis une trentaine d'années. Ils montrent aussi des résultats moins prévisibles, comme l'effet de concentration de la recherche dû aux sources de financement américaines : quand une agence américaine sponsorise des recherches sur un thème donné, celui-ci devient dominant et fédérateur ; à l'inverse, pendant les époques avec moins de financement et sans campagne d'évaluation sur un thème privilégié, la communauté est plus dispersée. Ces résultats peuvent sembler logiques mais il est malgré tout remarquable de pouvoir les observer directement, suite à une modélisation du domaine : il n'était pas du tout évident que les campagnes d'évaluation américaines aient un effet aussi visible sur un corpus aussi vaste que l'ACL Anthology. Ce résultat montre également bien le poids de la recherche américaine dans ce corpus sur la période 1980-1990.

L'étude d'Anderson *et al.* présente des résultats et une méthode d'analyse importante dont on s'inspire largement dans ce document. Nous souhaitons pour notre part pouvoir catégoriser automatiquement les termes suivant le type d'information qu'ils véhiculent. Nous proposons donc de combiner l'analyse des termes avec la reconnaissance automatique de la structure argumentative des textes analysés (ce que les anglo-saxons appellent '*argumentative zoning*' ou '*text zoning*' (Teufel, 1999)), ce qui permet de typer les termes en fonction du type de phrase dans lequel ils apparaissent.

Ce document est organisé comme suit. Nous présentons dans un premier temps la technique d'extraction de termes utilisée et une série de cartes visant à donner une représentation exploitable du domaine, à partir de l'-analyse des relations entre termes utilisés dans les articles. Nous poursuivons avec la technique permettant le marquage de la fonction argumentative des phrases (*text zoning*) mise en œuvre pour catégoriser les termes repérées dans les textes. Nous présentons ensuite différents résultats de l'application de cette technique au corpus ACL Anthology, afin d'en faire ressortir certains faits remarquables. Nous concluons par un résumé et quelques perspectives.

# 3 Cartographier le domaine de la linguistique informatique

Nous voulons tout d'abord dresser une carte sémantique du domaine, à partir des listes de termes repérés dans les titres et les résumés.

## 3.1 Méthode d'analyse

La première étape consiste à extraire les termes caractéristiques du domaine à partir de l'analyse des titres et des résumés d'articles.

On a donc recours à des outils d'extraction terminologique, qui visent précisément à identifier de façon automatique les termes pertinents dans un corpus en utilisant des méthodes de traitement automatique des langues. L'ensemble des termes repérés permet de proposer une modélisation conceptuelle d'un domaine. L'approche classique pour extraire des termes d'un corpus peut être décomposée en deux parties. Dans une première phase, des outils d'analyse linguistique sont utilisés pour construire une liste de candidats possibles qui sont filtrés dans une secondes phase.

La construction des termes candidats consiste classiquement (Bourigault and Jacquemin, 1999) à appliquer au texte un étiqueteur morphosyntaxique (« POS-tagging ») puis à utiliser les informations grammaticales associées à chaque mot pour effectuer une analyse syntaxique de surface (appelée « chunking » en anglais) qui permette d'identifier les groupes nominaux dans le texte : les candidats termes extraits constituant ainsi des candidats grammaticalement valides. Dans une deuxième phase, les termes sont filtrés, soit en faisant appel à des ressources extérieures, soit en fonction de scores associés tels que leur fréquence ou leur spécificité (Frantzi and Ananiadou, 2000).

Dans cette étude, nous nous sommes attaché aux contenu présent dans les résumés du corpus. Pratiquement, nous avons utilisés le module NLTK de traitement automatique des langues : une fois une liste de candidats-termes obtenus, nous avons sélectionné les 1000 termes ayant les meilleurs scores de fréquence et spécificité et apparaissant dans au moins 5 articles différents. La liste a ensuite été filtrée et validée par un expert du domaine.

Nous construisons ensuite, à partir des termes extraits du texte, une carte sémantique du champ scientifique considéré sous forme d'un réseau. Les nœuds du réseau correspondent aux termes extraits et deux termes sont liés dans le réseau s'ils sont apparu au moins une fois ensemble dans un titre ou un résumé. Les liens sont pondérés grâce à la mesure d'information mutuelle, qui mesure la dépendance statistique entre les deux termes considérés. Afin d'améliorer la lisibilité du réseau, les liens inférieur à un seuil donné sont supprimés.

L'objectif est d'obtenir un réseau constitué de plusieurs composantes densément connectées décrivant des sous-domaines bien identifiés à l'intérieur du domaine scientifique considéré. Un algorithme de détection de communautés (dit aussi algorithme de clustering) est alors appliqué au réseau ainsi obtenu : ce type d'algorithmes permet en effet de partitionner un réseau en groupes de noeuds densément connectés (clusters) et reliés de manière lâche avec le reste du réseau.

Dans cette étude, nous utilisons Infomap (Rosvall and Bergstrom, 2008), qui se trouve être l'un des meilleurs algorithme pour la tâche (l'algorithme de Louvain a également été testé (Blondel et al., 2008) mais celui-ci a obtenu des résultats jugés légèrement moins bons pour cartographier l'intégralité du domaine (cet algorithme est toutefois intéressant pour analyser l'évolution

du domaine au cours du temps). Sur la figure 1, chaque communauté ainsi identifiée est entourée d'un cercle, chacun représentant en fait un groupe thématique ou un thème de recherche particulier, comme la « désambiguïsation lexicale » ou « analyse morphosyntaxique » (« part of speech tagging »).

## 3.2 Evaluation

Différentes techniques de détection de communautés (Infomap et de Louvain) et différents paramètres ont été évalués par des experts du domaine. De plus, pour chaque groupement de termes obtenu, 10 articles ont été choisis au hasard puis projetés sur la carte en fonction des termes identifiés. L'expert devait ensuite dire si le groupement de termes identifié correspondait bien au thème majeur de l'article. La précision moyenne obtenue est de 0,84, ce qui est jugé acceptable pour ce genre de tâche.

Voici trois exemples de groupements (ou « cluster ») obtenus automatiquement avec la méthode décrite :

**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 of natural language - situation semantics - intensional logic - knowledge base - Generative Lexicon - artificial intelligence - knowledge representation - meaning representations

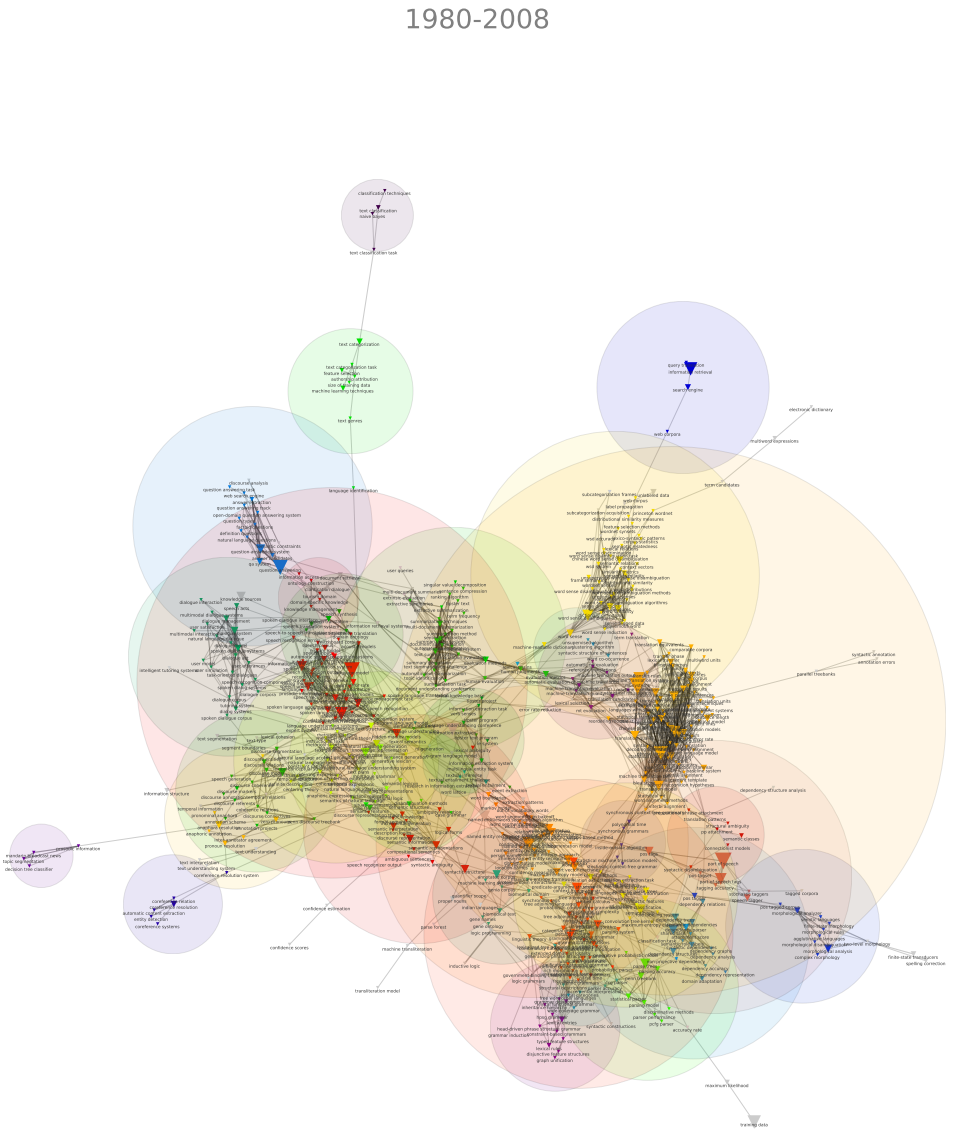



FIGURE 1 – Carte représentant le domaine de la linguistique informatique.

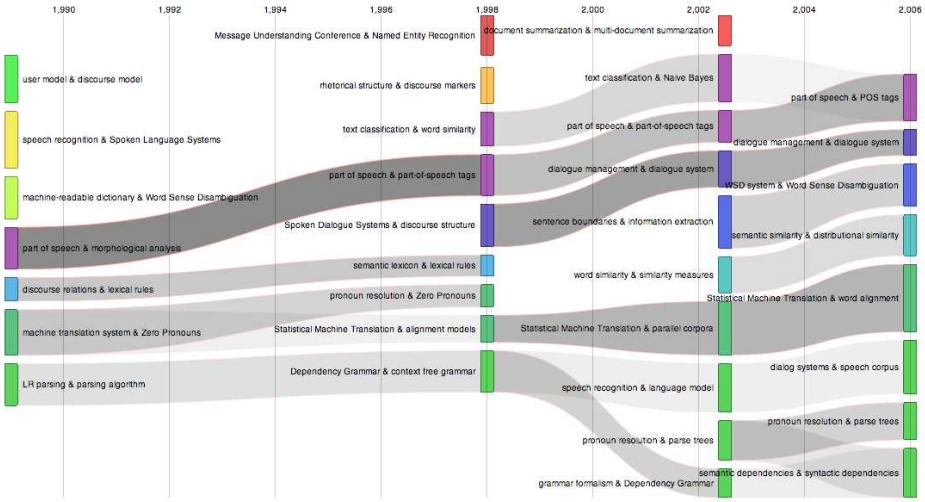



FIGURE 2 – Observation de l'évolution du domaine de la linguistique informatique dans le temps au niveau macroscopique.

# 4 Cartographier l'évolution du domaine

Nous voulons ensuite décrire les principales évolutions du domaine considéré (la linguistique informatique) au fil du temps, afin de suivre par exemple l'évolution de l'importance respective des différents sous-domaines identifiés (quel thème de recherche a émergé, ou au contraire a disparu ou s'est transformé du fait de l'évolution des techniques du domaine).

## 4.1 Méthode d'analyse

Schématiquement, l'approche comporte quatre étapes :

1. division du corpus en plusieurs tranches correspondant à des périodes de temps différentes ;
2. extraction des termes représentatifs de chaque tranche du corpus ;
3. application de l'algorithme de détection de communautés sur le graphe composé à partir de l'ensemble des termes identifiés ;
4. enfin, modélisation des liens temporels, en reliant entre eux les différents sous-domaines qui partagent un ensemble de termes communs suffisant au cours d'un intervalle temporel prédéfini.

Les algorithmes de classification utilisés sont les mêmes que précédemment (Infomap et Louvain).

Définir une stratégie pour lier deux ensembles de termes à des intervalles de temps différents est un problème difficile. Des termes apparaissent et disparaissent simplement parce que les techniques évoluent. Il faut alors déterminer quel degré de similarité ou de divergence doit être pris en compte.

L'approche adoptée ici reste toutefois relativement simple : comme nous l'avons vu, nous partons du principe que deux ensembles de termes sont connectés s'ils partagent suffisamment d'éléments communs (en fonction d'un seuil prédéfini). On peut noter que cette approche simple permet de lier un groupe de concepts *c* à une période de temps *t* avec un groupe *c'* à la période *t+1*, mais aussi d'associer un groupe de concepts *c* avec deux groupes *c'* et *c"* à la période *t+1* : c'est par exemple le cas quand un thème de recherche donne naissance à deux thèmes différents (par exemple, on observe que le groupement de concepts correspondant à la notion de « compréhension » donne naissance à deux nouveaux thèmes de recherche : la « reconnaissance d'entités nommées » et l'« extraction de l'information ». Ces thèmes sont considérés comme deux objets d'étude différents dans la mesure où ils partagent un nombre très faible de termes communs à la période *t+1*. Deux thèmes de recherche peuvent aussi fusionner pour produire un thème unique (par exemple on observe que l' « analyse statistique » et la « grammaire de dépendance » fusionnent pour donner naissance à un nouveau thème de recherche « analyse statistique en dépendances »). Enfin, si aucune correspondance ne peut être trouvée à *t+1*, le thème de recherche disparaît de la carte.

Pour nos expériences, nous avons utilisé la plate-forme CorText qui implémente l'ensemble des algorithmes, permet l'élaboration de l'ensemble de la procédure et fournit différents paramétrages pour chaque étape (la plate-forme met notamment en œuvre différentes techniques d'extraction de termes, de regroupement des termes en classes homogènes, ainsi que la cartographie des résultats ainsi obtenus). Différents paramétrages permettent d'obtenir différentes vues de l'évolution du domaine.

Comme il y a plusieurs représentations possibles, il faut bien garder en tête qu'il n'y a pas ici de « bonne » ou de « mauvaise carte », mais il y a des cartes différentes, donnant des vues différentes du domaine. La représentation ainsi obtenue doit être contrôlée avec soin et doit en outre donner lieu à une interprétation. Par exemple, si un ensemble de termes n'est plus connecté à *t+1*, il ne faut en déduire mécaniquement que ce thème de recherche

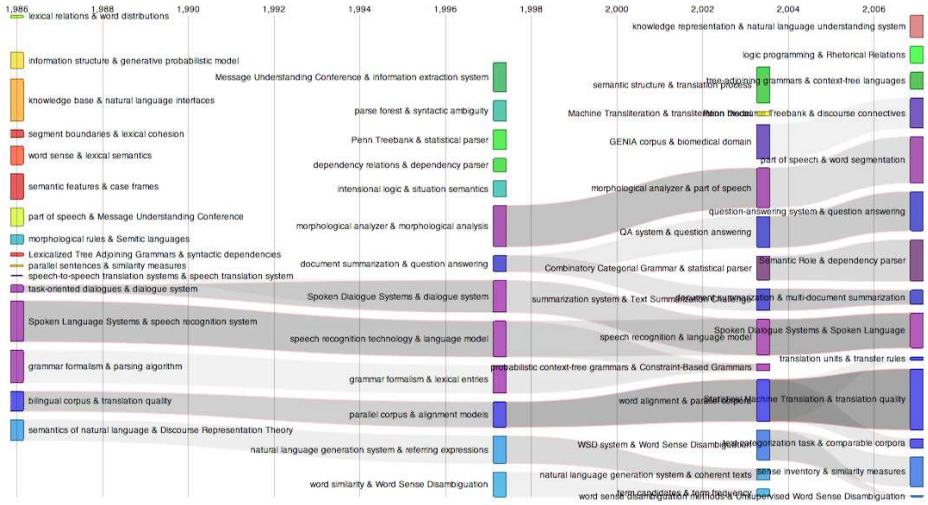



FIGURE 3 – Observation de l'évolution du domaine de la linguistique informatique dans le temps au niveau mesoscopique.

a simplement disparu. Ce thème peut à l'inverse avoir largement évolué, de sorte qu'à *t+1* aucun regroupement de termes ne contienne suffisamment de termes communs avec le regroupement d'origine *c*. Ce thème de recherche peut aussi avoir fusionné avec un autre. Bref, les cartes produites avec ce type de techniques doivent être considérées comme un moyen de relancer l'analyse, et non comme un résultat définitif en soi.

Enfin, idéalement, différentes cartes devraient être produites pour pouvoir examiner comment la représentation évolue, au niveau des valeurs limite notamment. Les effets de seuil doivent faire l'objet d'une attention toute particulière dans la mesure où certains regroupements peuvent apparaître avec certains paramétrages mais pas avec d'autres, particulièrement quand les valeurs sont proches des seuils définis.

## 4.2 Résultats

Nous fournissons ici trois cartes montrant l'évolution du domaine de la linguistique informatique depuis la fin des années 1980 jusqu'à nos jours.

La figure 2 montre les grandes tendances de l'évolution du domaine. Chaque période est constituée d'environ 8 à 12 groupements de termes montrant l'évolution des principaux sous-domaines de recherche au fil du temps (il faut noter que le nombre de regroupements est lié au jeu de paramètres

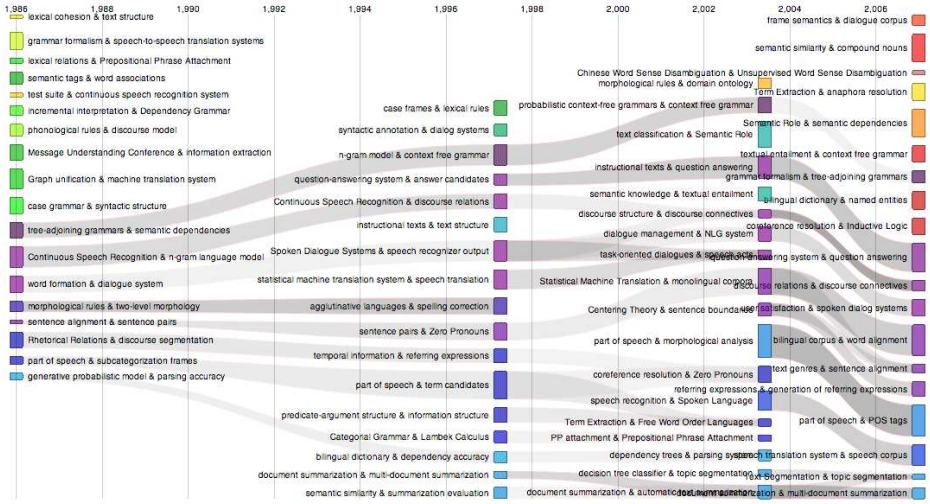



FIGURE 4 – Observation de l'évolution du domaine de la linguistique informatique dans le temps au niveau microscopique.

utilisé : en particulier, il n'est pas possible de définir directement le nombre de regroupements visé à un instant *t*. Seuls les thèmes de recherche partageant un nombre relativement important de termes sont reliés par des tubes couleur grise. On peut observer sur cette carte que le domaine le plus populaire (ou, du moins, celui donnant lieu aux recherches les plus nombreuses) est maintenant la traduction automatique : l'importance de ce domaine n'a cessé d'augmenter depuis la fin des années 1980. Nous pouvons également observer le développement de la tâche de « question-répondeur » depuis la fin des années 1990 : ce domaine a été particulièrement populaire à l'époque grâce à plusieurs campagnes d'évaluation mettant cette tâche en avant à la fin des années 1990.

Sur la carte, plusieurs domaines semblent ne pas devoir se poursuivre à *t+1*. Par exemple, l'extraction d'information à partir de dictionnaires électroniques a été très populaire dans les années 1980 mais aujourd'hui les équipes ont recours aux grands corpus et à l'apprentissage artificiel pour en extraire de l'information. La notion de « compréhension de textes » semble disparaître aussi, mais en fait ce domaine correspond aujourd'hui à une tâche d'« extraction d'information » (l'évolution sémantique est à cet égard intéressante). Entre « compréhension de textes » et « extraction d'information », il s'agit essentiellement d'une re-dénomination de la tâche. Ces deux regroupements ne sont pas liés entre eux probablement du fait de l'évolution

des termes employés. De ce point de vue, le renommage de la tâche reflète aussi la très grande évolution des techniques qui s'opère très rapidement, en quelques années. L'intérêt continu pour la « désambiguïsation » n'apparaît pas directement non plus, probablement parce que les approches par apprentissage automatique ont considérablement renouvelé ce sous-domaine : on voit bien la transition entre des systèmes essentiellement fondés sur des données symboliques (grammaire, dictionnaire) et les techniques d'apprentissage artificiel couramment employées depuis la fin des années 1990.

Les figures 3 et 4 donnent un aperçu plus précis du domaine. Ces cartes permettent d'observer les résultats à un niveau plus global et donnent une vue d'ensemble des grandes évolution.

# 5 Annotation de la structure argumentative des articles

Nous souhaitons à présent analyser de manière plus fine le contenu même des articles considérés. Dans ce cadre, la reconnaissance et l'annotation de la structure discursive des articles scientifiques sont devenus des enjeux importants pour la communauté du traitement des langues. Ce type de techniques peut en effet permettre de savoir si une section d'un article scientifique donné concerne, par exemple, le protocole expérimental employé, les données d'expériences ou la discussion et la comparaison avec les travaux antérieurs. Ce type d'analyse donne des résultats de plus en plus précis et commence à intéresser les grandes maisons d'éditions scientifiques, dans la mesure où on peut ainsi enrichir les bases de connaissances existantes et proposer de nouveaux parcours de lecture.

## 5.1 Etat de l'art

Les premiers travaux d'importance dans le domaine sont certainement ceux de Simone Teufel (Teufel, 1999) qui a proposé de catégoriser les phrases d'articles de traitement automatique des langues suivant sept étiquettes différentes : BKG (arrière-plan scientifique), OTH (description neutre de travaux antérieurs), OWN (description neutre du travail de l'auteur), AIM (objectifs de l'article), TXT (annonce de l'organisation de l'article), CTR (comparaison avec des travaux antérieurs) et BAS (description des travaux

antérieurs sur lesquels s'appuie l'article).

La tâche est appelée « rhetorical zoning » ou « argumentative zoning » par l'auteur, dans la mesure où le balisage doit permettre d'identifier la fonction rhétorique ou argumentative de chaque phrase du texte.

Le travail initial de S. Teufel (Teufel, 1999) est fondé sur l'annotation manuelle de 200 articles représentatifs du domaine issus des conférences de l'ACL et de la revue *Computational Linguistics*. Un classifieur (c'est-à-dire un algorithme permettant de classer automatiquement des objets, ici des phrases suivant des catégories discursives prédéfinies) est ensuite entraîné sur cette base : il permet d'obtenir une annotation automatique de nouveaux textes donnés en entrée à l'analyseur. L'auteur rapporte que le système automatique donne le bon résultat dans 70% des cas, comparé à un accord de 88% entre humains. Le classifieur repose sur un modèle bayesien naïf car les méthodes plus sophistiquées testées par l'auteur ne semblent pas permettre d'obtenir de meilleurs résultats.

Teufel montre dans une publication ultérieure (Teufel and Moens, 2002) comment cette technique peut être utilisée pour générer des résumés automatiques de qualité. Les techniques de résumé traditionnelles sont fondées sur la sélection de phrases en fonction de leur intérêt informatif supposé, essentiellement sur la base des noms et des verbes qui la compose (les mots les plus centraux, souvent appelés centroïdes (Radev et al., 2004)), ce qui pose problème pour générer des textes tenant compte de la variété du texte de départ. Le repérage de la structure argumentative répond partiellement à ce problème dans la mesure où il est dès lors possible de générer des résumés reflétant les différentes zones repérées ou, au contraire, privilégiant une zone donnée suivant les besoins informationnels du lecteur.

Teufel a enfin montré (Teufel et al., 2006) comment le marquage argumentatif peut être couplé avec les références scientifiques. Les articles scientifiques sont en effet fondées sur des citations des travaux antérieurs mais ces citations peuvent avoir différents statuts : simple mention de travaux antérieurs donnant l'arrière-plan de la recherche en cours, travaux précis auxquels s'oppose la publication en cours, référence à des travaux utilisant le même protocole expérimental, etc. Coupler repérage de références et balisage argumentatif permet de typer les citations, toujours dans le but de faciliter la lecture en fonction des besoins informationnels du lecteur.

Les travaux de S. Teufel ont depuis donné lieu à différents types d'é-

tudes, d'une part pour affiner la méthode d'annotation, d'autre part pour vérifier son applicabilité à différents domaines scientifiques. Pour le premier point, les recherches ont porté sur les traits pertinents pour la classification, l'évaluation de différents algorithmes pour la tâche et surtout la diminution de la quantité de texte à annoter pour obtenir un système fonctionnel. Pour le second, c'est surtout le domaine de la biomédecine et de la biologie qui ont montré le plus d'intérêt pour ce type de techniques, du fait de la quantité d'articles disponible dans ce domaine et de la nécessité d'accéder de manière transversale à cette littérature (les biologistes peuvent par exemple avoir besoin d'accéder à tous les protocoles expérimentaux pour un problème donné) (Mizuta et al., 2006; Tbahriti et al., 2006).

Les travaux de Y. Guo (Guo et al., 2011, 2013) reprennent l'analyse de la structure argumentative en complétant les travaux initiaux de S. Teufel sur un certain nombre de points : recours à une vaste liste de critères pour déterminer la classification des phrases, évaluation de plusieurs algorithmes d'apprentissage et diminution de la quantité de données annotées à fournir au système pour l'entraînement.

Y. Guo *et al.* (2011) proposent en particulier d'avoir recours à l'apprentissage actif (*active learning*) pour entraîner leur système. On sait en effet que l'apprentissage actif permet de réduire la quantité de données annotées en utilisant parallèlement une grande masse de données non annotées : cette méthode est bien indiquée dans notre cas dans la mesure où les corpus à analyser en traitement des langues (et particulièrement l'ACL Anthology) sont souvent d'assez grande taille mais il ne sont évidemment pas annotés. Les traits utilisés pour l'apprentissage sont de trois types : *i*) positionnels (localisation de la phrase au sein de l'article), *ii*) lexicaux (mots, classes de mots, bigrammes, etc. sont pris en considération) et *iii*) syntaxiques (les différentes relations syntaxiques, ainsi que les classes de noms en position sujet et les classes de noms en position objet sont pris en considération). L'analyse est donc considérablement plus riche que celle de Teufel mais nécessite en contrepartie un analyseur syntaxique.

## 5.2 Application de l'analyse argumentative au corpus de l'ACL

La méthode développée par Y. Guo et ses collègues (2011) semble particulièrement bien adaptée à notre problème. Nous souhaitons en effet caté-

goriser les termes repérés à l'étape précédente afin notamment d'identifier les méthodes mentionnées dans le corpus ACL Anthology et pouvoir ainsi analyser, par exemple, leur évolution dans le temps. Les termes apparaissant dans des phrases se rapportant au protocole expérimental employé sont donc susceptibles de particulièrement nous intéresser. Il faut noter à ce propos qu'il n'y a pas de frontière étanche entre thèmes et méthodes de recherche dans la mesure où le traitement automatique des langues s'appuie sur ses propres résultats pour concevoir des systèmes en couches empilées : ainsi, un analyseur sémantique reposera fréquemment sur un analyseur syntaxique employé comme outil (et apparaissant donc dans la section méthodologique de l'article).

L'annotation ne porte que sur les résumés des articles. On fait en effet l'hypothèse que les résumés contiennent assez d'information et sont assez redondants pour observer l'évolution du domaine. A l'inverse, aborder l'étude en utilisant le texte complet des articles entraînerait probablement du bruit et complexifierait inutilement les traitements.

Le jeu d'annotation initialement adopté comporte sept catégories différentes et une catégorie AUTRE pour les phrases ne pouvant pas être catégorisées par les étiquettes définies. Ces étiquettes sont les suivantes :

- OBJECTIF : décrit les objectifs de l'article ;
- METHODE : méthodes employées par l'article ;
- RESULTATS : résultats obtenus ;
- CONCLUSION : conclusion de l'article ;
- ARRIERE-PLAN : contexte scientifique ;
- TRAVAUX LIES : positionnement par rapport à des travaux directement liés à ceux présentés ;
- AUTRES TRAVAUX : positionnement par rapport à d'autres travaux.

Ces catégories sont reprises des travaux précédents, notamment (Mizuta et al., 2006; Guo et al., 2011, 2013). Il nous a semblé important de reprendre un jeu de catégories existantes dans la mesure où ces catégories, avec de légères variations, se sont globalement imposées depuis les premiers travaux de S. Teufel. Certaines catégories sont malgré tout peu présentes dans les résumés de l'ACL Anthology, et finalement quatre catégories transparaissent principalement : les catégories OBJECTIF, ARRIERE-PLAN, RESULTATS et METHODE. Il est rare de trouver des comparaisons avec d'autres travaux dans les résumés de l'ACL Anthology (alors qu'on en trouve fréquemment

dans les résumés en biologie par exemple).

Une centaine de résumés d'article issus de l'ACL Anthology ont ensuite été annotés manuellement avec ces catégories (environ 500 phrases, les résumés de l'ACL Anthology étant souvent très courts dans la mesure où il s'agit en grande majorité de résumés d'articles de conférence). Les articles annotés ont été choisis aléatoirement, en prenant soin toutefois qu'ils couvrent différentes périodes et qu'ils contiennent des termes variés. L'annotation a été faite en suivant le guide d'annotation mis au point par Y. Guo, notamment en ce qui concerne les phrases complexes, se rapportant potentiellement à plus d'une catégorie définie (un jeu de préférences est défini pour résoudre ces cas difficiles).

L'algorithme de (Guo et al., 2011) est ensuite repris et adapté à notre cas de figure. L'analyse se fonde en particulier sur les traits positionnels, lexicaux et syntaxiques comme expliqué dans la section précédente. Pour l'analyse syntaxique, l'analyseur C&C est utilisé (Curran et al., 2007) et pour la classification, on a recours à l'implémentation des SVM linéaires de Weka. Comme résultat, pour chaque phrase du corpus, l'algorithme associe une étiquette choisie parmi les étiquettes possibles.

## 5.3 Résultats et discussion

Pour valider les résultats obtenus, un ensemble de résumés est choisi aléatoirement. Les quatre catégories principales sont bien représentées mais inégalement réparties : 18,05 % des phrases sont catégorisées comme ARRIERE-PLAN, 14,35 % comme OBJECTIF, 14,81 % comme RESULTAT et 52,77 % comme METHODE. On voit bien, à la lecture de ces chiffres, l'importance de la dimension méthodologique dans le domaine.

On observe ensuite, pour chaque catégorie possible, le pourcentage de phrases correspondant effectivement à cette étiquette, ce qui permet de mesurer les performances du système en terme de précision. Les résultats obtenus sont présentés dans le tableau 1.

Ces résultats sont conformes à l'état de l'art (si on les compare avec ceux de (Guo et al., 2011) par exemple). On voit que les résultats sont globalement satisfaisants, particulièrement en regard du peu de phrases annotées pour l'entraînement. La richesse des traits pris en compte et la stratégie d'apprentissage actif permettent en outre d'avoir des résultats portables d'un domaine à l'autre sans tâche d'annotation lourde. Les résultats sont légère-

TABLE 1 – Résultat de l'analyse argumentative (en précision)

| Catégorie | Précision |
|---|---|
| Objectif | 83,87 % |
| Arrière-plan | 81,25 % |
| Méthode | 71,05 % |
| Résultats | 82,05 % |

ment moins bons pour la catégorie METHODE car celle-ci est sans doute plus diversifiée que les autres et donc moins facile à cerner.

L'exemple montré en figure 5 est un texte annoté suite à l'analyse du système (il s'agit de l'article de (Lee et al., 2002), choisi au hasard parmi ceux qui présentent une bonne diversité dans les catégories utilisées). La catégorisation s'effectue au niveau des phrases, ce qui n'est pas sans poser problème : par exemple, dans ce résumé, le fait qu'une méthode hybride est utilisée est indiqué dans une phrase étiquetée OBJECTIF par le système. Les phrases marquées METHODE contiennent toutefois des termes précieux, comme *lexical pattern* ou *tri-gram estimation*, ce qui peut permettre d'inférer le fait qu'il s'agit d'un système hybride. On aperçoit au passage des problèmes de numérisation, qui sont typiques du corpus étudié : l'ACL Anthology comprend des textes convertis automatiquement à partir de fichiers PDF de conférences passées, ce qui entraîne parfois des problèmes de qualité.

# 6 Application : contribution à l'étude de l'évolution du traitement automatique des langues d'après l'ACL Anthology

Comme nous l'avons dit dans l'introduction, nous nous situons dans la lignée des travaux de (Anderson et al., 2012). L'ACL Anthology est utilisé ici comme un cas d'étude typique : le corpus s'étendant sur une période de temps conséquente (plus de 30 ans si on retient les articles depuis 1980), il peut être intéressant d'en étudier les grandes évolutions.

FIGURE 5 – **Exemple :** *Un résumé annoté avec l'analyseur de la structure argumentative. Les catégories ajoutées au texte sont indiquées en gras.*

Most of errors in Korean morphological analysis and POS (Part-of-Speech) tagging are caused by unknown morphemes. **BACKGROUND**
This paper presents a generalized unknown morpheme handling method with POSTAG (POStech TAGger) which is a statistical/rule based hybrid POS tagging system. **OBJECTIVE**
The generalized unknown morpheme guessing is based on a combination of a morpheme pattern dictionary which encodes general lexical patterns of Korean morphemes with a posteriori syllable tri-gram estimation. **METHOD**
The syllable tri-grams help to calculate lexical probabilities of the unknown morphemes and are utilized to search the best tagging result. **METHOD**
In our scheme , we can guess the POS's of unknown morphemes regardless of their numbers and positions in an eojeol , which was not possible before in Korean tagging systems. **RESULTS**
In a series of experiments using three different domain corpora , we can achieve 97% tagging accuracy regardless of many unknown morphemes in test corpora. **RESULTS**

## 6.1 Repérage des termes spécifiques

Il nous a semblé particulièrement intéressant de nous intéresser à l'évolution des méthodes employées en traitement automatique des langues. Pour cela, il est nécessaire d'identifier les termes particulièrement présents dans les phrases étiquetées METHODE.

Les termes caractéristiques du corpus sont extraits avec NLTK, comme indiqué supra (cf. section 3.1).

Nous calculons ensuite la spécificité de chaque terme par rapport aux catégories définies pour l'analyse discursive. La spécificité est calculée grâce au test de Kolmogorov-Smirnov, qui quantifie une distance entre les fonctions de répartition empiriques de deux échantillons :

$$D = \max_x |S_{N_1}(x) - S_{N_2}(x)| \tag{1}$$

où $S_{N_1}(x)$ et $S_{N_2}(x)$ sont les fonctions de répartition empiriques des deux échantillons (ce qui correspond dans notre cas au nombre d'occurrences du terme dans chaque zone identifiée par la marquage argumentatif, et au nombre total d'occurrences (en considérant tous les termes) dans chaque zone) (Press et al., 2007). Une valeur importante de $D$ pour un terme donné signifie donc que le terme est très spécifique d'une zone. A l'inverse, une valeur faible indique que le terme est éparpillé dans toutes les zones et est donc peu spécifique.

La liste est ensuite triée par mesure de spécificité et les cent cinquante premiers termes sont catégorisés par un expert du domaine. On obtient ainsi le tableau 2 : celui-ci ne contient pas tous les éléments a priori pertinents (c'est-à-dire toutes les méthodes utilisées en traitement automatique des langues) mais il contient les termes les plus spécifiques d'après la méthode précédente. Il ne faut donc pas s'étonner de trouver une liste incomplète par rapport à l'ensemble des méthodes utilisées dans le domaine.

## 6.2 Evolution des méthodes dans le temps

L'analyse automatisée du corpus permet avant tout de tracer l'évolution des différentes tendances dans le temps. Pendant la période considérée, les méthodes utilisées ont beaucoup changé, le principal fait marquant étant peut-être le recours massif à l'apprentissage artificiel (une technique informatique permettant d'inférer des connaissances à partir de grandes masses

TABLE 2 – Classement des termes les plus spécifiques trouvés dans les phrases étiquetées METHODE

| Methods | | |
|---|---|---|
| Category | Method | N-grams |
| Machine learning | Bayesian methods | baesyan |
| | Vector Space model | space model, vector space, cosine |
| | Genetic algorithms | genetic algorithms |
| | HMM | hidden markov models, markov model |
| | CRF | conditional random fields |
| | SVM | support vector machines |
| | MaxEnt | maximum entropy model, maximum entropy approach, maximum entropy |
| | Clustering | clustering algorithm, clustering method, word clusters, classification problem |
| Speech & Mach. Trans. | Language models | large-vocabulary, n-gram language model, Viterbi |
| | Parallel Corpora | parallel corpus, bilingual corpus, phrase pairs, source and target languages, sentence pairs, word pairs, source sentence |
| | Alignment | phrase alignment, alignment algorithm, alignment models, ibm model, phrase translation, translation candidates, sentence alignment |
| NLP Methods | POS tagging | part-of-speech tagger, part-of-speech tags |
| | Morphology | two-level morphology, morphological analyzer, morphological rules |
| | FST | finite-state transducers, regular expressions, state automata, rule-based approach |
| | Syntax | syntactic categories, syntactic patterns, extraction patterns |
| | Dependency parsing | dependency parser, dependency graphs, prague dependency, dependency treebank, derivation trees, parse trees |
| | Parsing | grammar rules, parser output, parsing process, parsed sentences, transfer rules |
| | Semantics | logical forms, inference rules, generative lexicon, lexical rules, lexico-syntactic, predicate argument |
| Applications | IE and IR | entity recognition, answer candidates, temporal information, web search, query expansion, google, user queries, keywords, query terms, term recognition |
| | Discourse | generation component, dialogue acts, centering theory, lexical chains, resolution algorithm, generation process, discourse model, lexical choice |
| | Segmentation | machine transliteration, phonological rules, segmentation algorithm, word boundaries |
| Words and Resources | Lexical knowledge bases | lexical knowledge base, semantic network, machine readable dictionaries, eurowordnet, lexical entries, dictionary entries, lexical units, representation structures, lookup |
| | Word similarity | word associations, mutual information, semantic relationships, word similarity, semantic similarity, semeval-2007, word co-occurrence, synonymy |
| | Corpora | brown corpus, dialogue corpus, annotation scheme, tagged corpus |
| Evaluation | Evaluation | score, gold standard, evaluation measures, estimation method |
| Calculation & complexity | Software | tool development, polynomial time, software tools, series of experiments, system architecture, runtime, programming language |
| | Constraints | relaxation, constraint satisfaction, semantic constraints |

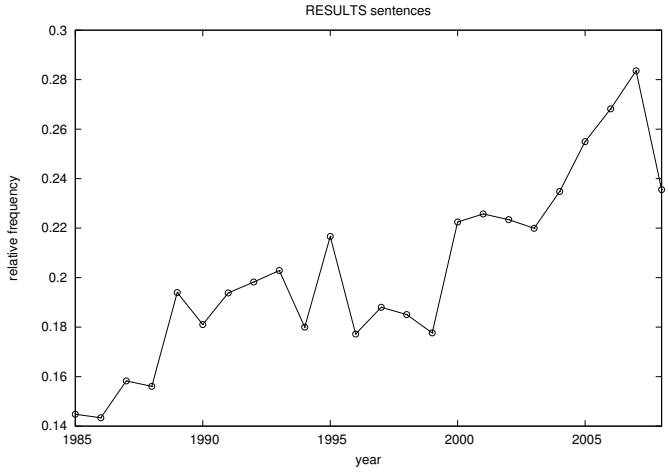



FIGURE 6 – Évolution dans le temps de la proportion de phrases catégorisées par l'outil d'analyse discursive comme étant des phrases concernant des résultats (par rapport au nombre total des phrases contenues dans les articles publiés dans l'année correspondante).

de données représentatives) depuis la fin des années 1990. Cette tendance est marquée par un recours quasi systématique dans les articles actuels à des expérimentations donnant lieu à des résultats chiffrés.

Pour confirmer de façon quantitative cette hypothèse, nous nous intéressons à l'évolution dans le temps de la proportion de phrases étiquetées RESULTAT. Sur la figure 6, nous pouvons ainsi observer que la courbe correspondante croît de façon quasi linéaire du début des années 1980 jusqu'à la fin des années 2000.

Il est ensuite possible de faire des traitements plus fins pour suivre dans le temps l'évolution des différents groupes de méthodes identifiés. Les résultats sont visibles sur la figure 7. Les méthodes à base de règles et de ressources linguistiques élaborées manuellement se maintiennent ou baissent légèrement, tandis que les méthodes à base d'apprentissage connaissent un succès de plus en plus grand à partir des années 1990. Ceci n'est pas en soi surprenant : on sait que des systèmes à base de règles continuent d'être utilisés tandis que l'apprentissage s'est généralisé. La figure indique toutefois un constat plus équilibré qu'on pourrait le penser : les deux types de méthodes coexistent. Les méthodes d'apprentissage sont probablement souvent employées en collaboration avec des méthodes fondées sur l'apprentissage et

les deux paradigmes se complètent sans doute plus qu'ils ne s'opposent.

Le détail montre des tendances qu'il faudrait confirmer par une étude plus approfondie. On voit toutefois le succès de l'analyse en dépendance à la fin des années 1980 (probablement grâce au succès des grammaires d'arbres adjoints à cette époque) qui connaît à nouveau un certain succès depuis les années 2000 grâce au développement des techniques d'apprentissage et des corpus étiquetés en dépendances (ce qui a par exemple donné lieu à plusieurs tâches partagées (*shared tasks*) lors des conférences CONLL de 2006 à 2009).

Les méthodes d'apprentissage se succèdent par vagues mais chaque méthode continue par la suite d'être employée, perfectionnée et appliquée à de nouvelles tâches. Les HMM et les n-grammes connaissent un pic très net dans les années 1990, probablement suite aux expériences initiales de Jelinek et ses collègues inaugurant l'ère de la traduction automatique statistique (Brown et al., 1990). Les SVM et les CRF ont eu un succès plus récent comme on le sait.

Nous nous sommes aussi intéressés à la distribution des méthodes entre les articles et entre les auteurs. La figure 8 montre le nombre moyen de termes apparaissant dans la section METHODE du résumé des articles au cours du temps. On peut observer que le nombre d'éléments méthodologique augmente, surtout dans les années 1980, montrant peut-être un accroissement de la complexité des systèmes développés.

## 6.3 La dynamique des auteurs dans l'espace des méthodes

Les éléments observés jusqu'ici confirment des résultats en partie déjà connus. La méthode proposée peut toutefois permettre d'aller plus loin : on peut essayer d'observer les dynamiques à l'œuvre dans l'évolution du domaine. Comment les nouvelles méthodes d'analyse sont-elles introduites dans le domaine ? Sont-elles plutôt amenées par des chercheurs débutant ou sont-ce plutôt des chercheurs confirmés du domaine qui inventent ou importent des méthodes nouvelles depuis des domaines connexes ? Les spécialistes du TAL sont-ils en général spécialistes d'une méthode ou d'un domaine étroit de spécialité ou ont-ils plutôt une expertise large et diversifiée ?

Il s'agit évidemment de questions complexes et chaque individu a une trajectoire particulière. Les méthodes automatiques peuvent toutefois donner des indicateurs, surtout dans la durée. Comme nous l'avons déjà vu, Anderson et al. (2012) montrent ainsi que les conférences d'évaluation ont

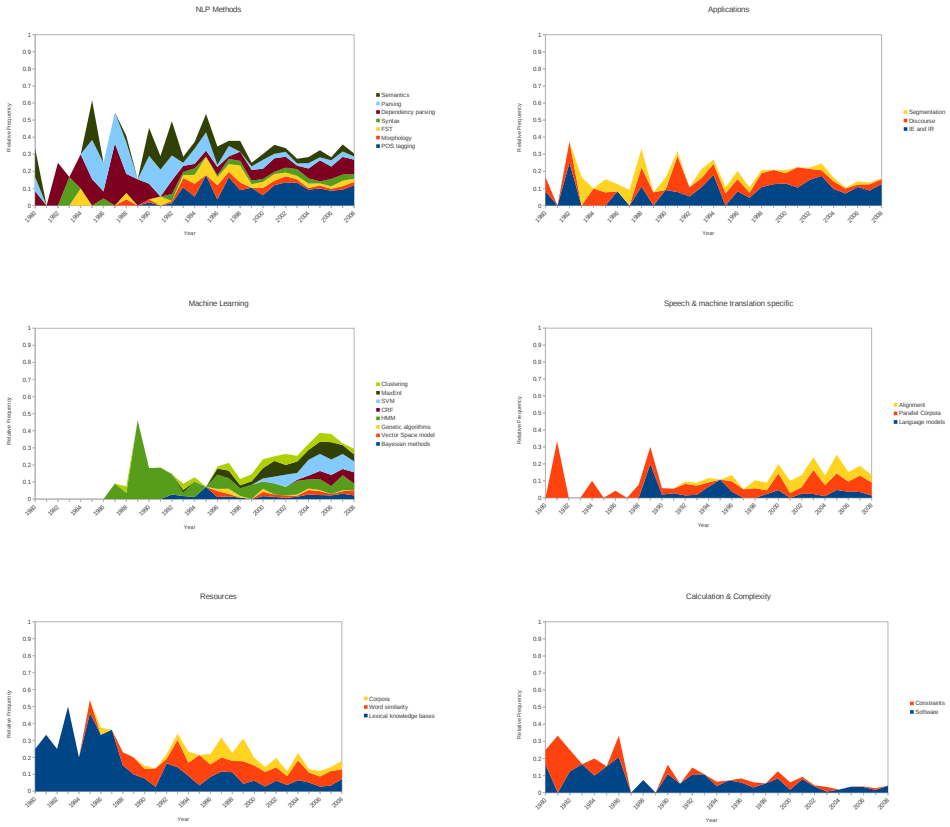



FIGURE 7 – Évolution dans le temps de la fréquence relative des différents groupes de méthodes identifiés.

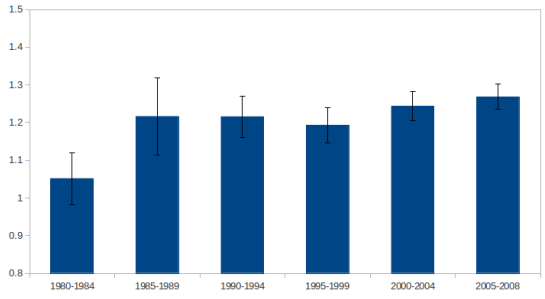



FIGURE 8 – Évolution du nombre de méthodes par article dans le temps.

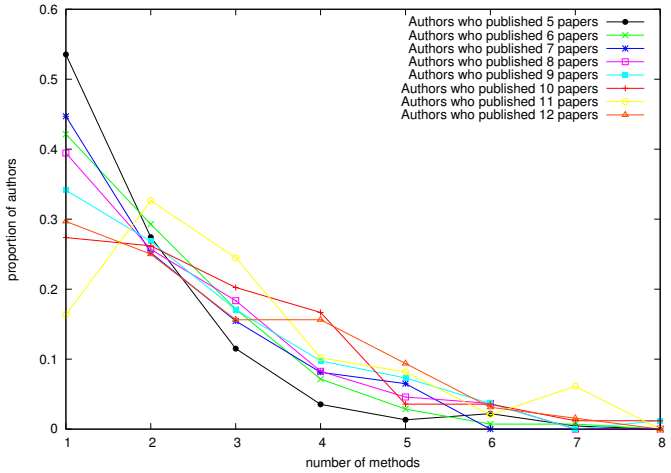



FIGURE 9 – Proportion d'auteurs experts d'un nombre de méthodes donné, pour différents catégories de chercheurs.

eu un impact sur le domaine, en limitant la diversité des recherches à certaines périodes clés, ce qui ne veut pas dire qu'il n'y avait pas à ces époques aussi des recherches originales en dehors de ces campagnes. Il s'agit donc d'essayer de mettre au jour certaines tendances spécifiques d'un domaine scientifique, qui pourraient par exemple amener à des comparaisons avec d'autres domaines scientifiques. Les outils fournissent avant tout des hypothèses : ils poussent le chercheur à aller voir plus loin mais il ne s'agit évidemment que d'outils d'aide à l'analyse. Nous ne prétendons pas donner une vue exacte et absolument objective du domaine.

Pour mener à bien notre enquête, nous ne prenons en compte que les auteurs qui ont produit au moins cinq articles dans l'ACL Anthology, afin de ne prendre en compte que les auteurs ayant contribué au domaine pendant un temps assez long pour la pertinence de l'étude.

La figure 9 montre le nombre d'auteurs spécialistes d'une ou plusieurs méthodes données. On constate que la plupart des auteurs font référence à une seule méthode. Logiquement, les courbes sont décroissantes : il y a finalement peu d'auteurs utilisant une très large gamme de méthodes différentes. Ces résultats mériteraient évidemment d'être confirmés par une étude de plus grande ampleur prenant en compte une plus grande diversité de termes regroupés par famille. Il nous semble malgré tout que cette expérimentation montre des tendances intéressantes pour ce corpus.

Nous nous concentrons ensuite sur les « pionniers », que nous définissons comme étant les premiers auteurs ayant publié un article où le terme référant à une méthode donnée apparaît (par exemple, les premiers articles où le terme 'support vector machine' ou 'SVM' apparaît). Parmi l'ensemble des articles mentionnant une méthode, seuls les articles correspondant aux 16 premiers centiles (autrement, les 16% d'articles publiés en premier) sont considérés comme pionniers : cette valeur a été choisie en se fondant sur les travaux de Rogers sur la diffusion des innovations (Rogers, 1962), qui montrent l'importance du role joué par les innovateurs (qui constituent le premier 2.5%) et des adopteurs précoces (qui constituent les $13,5\%$ suivants). Ces deux populations ensemble peuvent être considérées comme formant l'ensemble des pionniers.

Nous essayons de déterminer à quelle moment de leur carrière les chercheurs utilisent des méthodes novatrices. Pratiquement, nous examinons à quelle étape de leur carrière les auteurs que nous avons considérés comme « pionniers » ont publié les articles ayant permis de les classer ainsi (par exemple, si un auteur est un des premiers à avoir utilisé les SVM, l'a-t-il fait lors de ses premières publications ou plus tard au cours de sa carrière ?). Le résultat est visible sur la figure 10, où on compare la fraction d'articles publié par les « pionniers » avant d'introduire une nouvelle méthode (par rapport à leur production totale), et le même type de données pour les autres chercheurs (c'est à dire la fraction d'articles publié avant de commencer à utiliser une méthode nouvelle pour eux mais pas pour le domaine). Nous observons que 50% des « pionniers » n'avait jamais publié dans le domaine avant d'introduire la nouvelle méthode en question (contre 40% seulement en ce qui concerne les autres chercheurs). Ces valeurs montrent que les nouvelles méthodes semblent émaner assez largement de nouveau venus, probablement de chercheurs ayant déjà éprouvé la méthode sur un autre domaine (de fait, l'équipe de Jelinek, qui a joué un rôle essentiel dans l'essor des chaînes de Markov cachées à partir des années 1990 (Brown et al., 1990), avait surtout été active en reconnaissance de la parole jusque là et n'avait quasiment pas publié d'articles faisant partie du corpus ACL, même s'il s'agissait bien évidemment de chercheurs confirmés).

La figure 10 révèle aussi que 70% des « pionniers » ont publié moins du tiers de leur production totale au moment où ils utilisent une nouvelle méthode. On observe donc un regroupement partiels entre ces pionniers et les

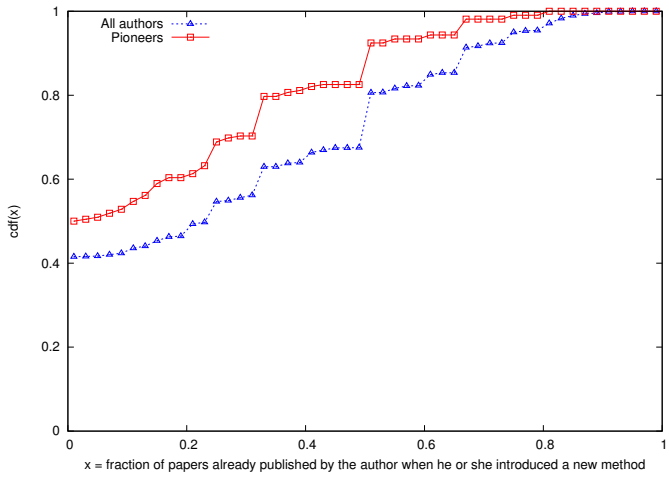



FIGURE 10 – Fonction de répartition de la proportion d'articles que les « pionniers » avaient déjà publié au moment où il sont publié leur premier article sur une nouvelle méthode, par rapport à la production totale de leur carrière.

jeunes chercheurs du domaine ou, comme on l'a vu dans le paragraphe précédent, entre ces pionniers et des chercheurs ayant jusque là publié dans des communautés proches mais néanmoins différentes. Il faudrait donc étudier en parallèle d'autres corpus (en informatique, en linguistique, en sciences cognitives, etc.) pour pouvoir affiner la description, mais la tâche est dès lors difficile.

On peut ensuite se poser la question de la diversité de méthodes employées par les auteurs du domaine, en particulier par le groupe que nous avons appelé « pionniers ». La figure 11 montre le nombre de méthodes détectées par article pour les pionniers d'une part (en rouge) et pour l'ensemble des auteurs d'autre part (en bleu). On voit chez les pionniers (en prenant en compte l'intégralité de leur production scientifique dans la collection ACL Anthology) une nette sous-représentation de chercheurs utilisant une seule méthode, et une sur-représentation (statistiquement significative) du nombre d'auteurs utilisant quatre méthodes ou plus. Le groupe que nous appelons « pionniers » a donc une tendance marquée à utiliser plus de méthodes (et aussi à aborder davantage de sous-domaines du traitement automatique des langues) que l'ensemble des auteurs pris globalement.

Finalement, nous essayons de mesurer les flux entre méthodes : un chercheur ayant travaillé sur une méthode donnée a-t-il plus de chances de

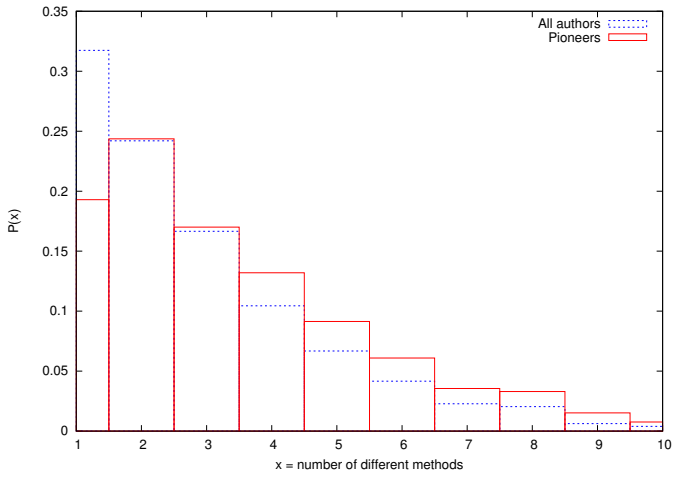



FIGURE 11 – Proportion de « pionniers » experts d'un nombre donné de méthodes et comparaison avec ce même indicateur pour l'ensemble des auteurs du corpus.

travailler ensuite sur telle ou telle autre méthode (par exemple, un chercheur ayant utilisé les HMM a-t-il plus de chances de se tourner vers les SVM ou les CRF si les deux méthodes sont populaires en même temps ?). Nous mesurons ces flux en analysant les évolutions de méthodes d'une période à l'autre (articles d'un auteur donné ayant utilisé une méthode pendant une période puis une autre méthode la période suivante par exemple). Les flux sont ensuite normalisés en prenant en compte le nombre total d'auteurs concernés. Les figures 12, 13 et 14 montrent les résultats ainsi obtenus.

Nous pouvons observer que le flux d'auteurs des années 1980 aux années 1990 concerne principalement les méthodes de TAL, les techniques d'apprentissage automatique n'étant pas encore utilisées, à l'exception des modèles de Markov cachés, qui sont devenus populaires à partir des années 1990 (figure 12). Des années 1990 à la première moitié des années 2000 les méthodes employées concernent davantage l'apprentissage automatique comme, par exemple, les *Support Vector Machines*, devenus très populaires (figure 13). De la première à la seconde moitié des années 2000, les chercheurs se concentrent davantage sur les *Conditional Random Field* (une technique d'apprentissage automatique pour le traitement du langage naturel), et sur un domaine spécifique de la syntaxe : l'analyse en dépendances (*Dependency Parsing* en anglais), qui a fait l'objet de plusieurs campagnes d'évaluation

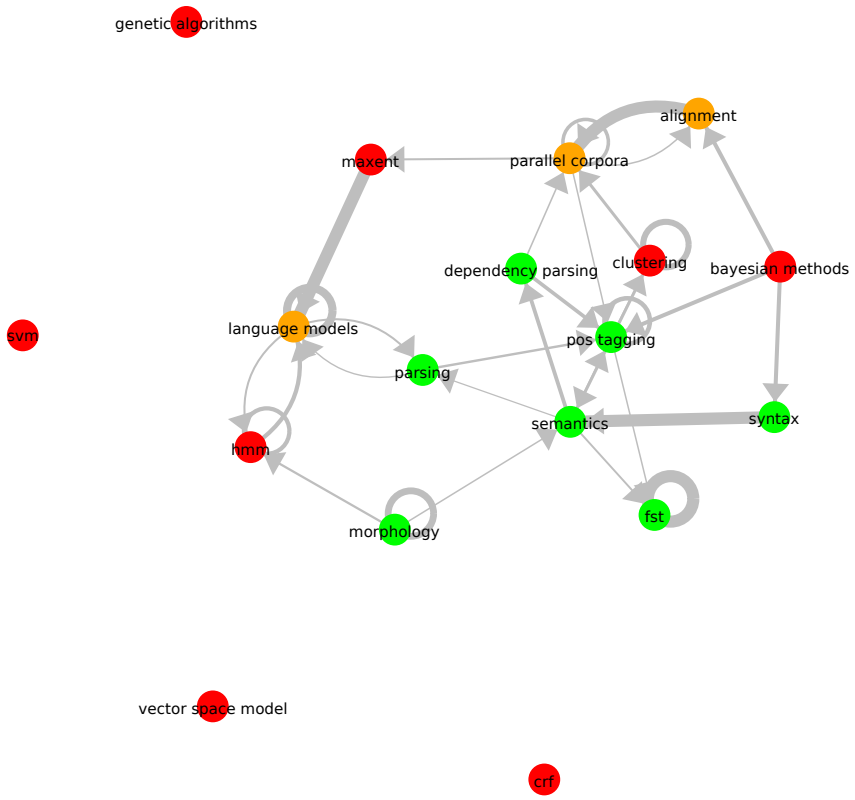



FIGURE 12 – Réseau des flots d'auteurs entre méthodes. Pour chaque couple de méthodes, nous avons calculé le flot de l'une à l'autre en comptant le nombre d'auteurs qui ont publié successivement un article employant la première dans les années 1980 puis la deuxième méthode considérée dans les années 1990. Chaque flot est normalisé en fonction du nombre total d'auteurs concernés (toutes les flots inférieurs à 10% sont supprimés).

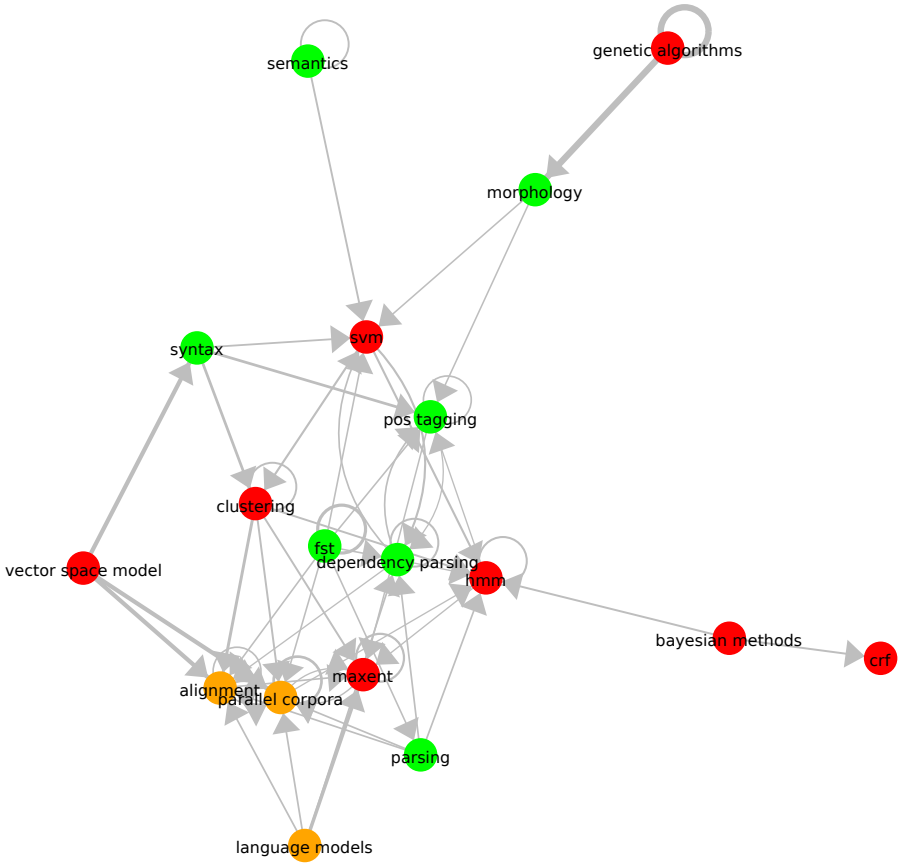



FIGURE 13 – Réseau des flots d'auteurs entre méthodes des années 1990 à la première moitié des années 2000.

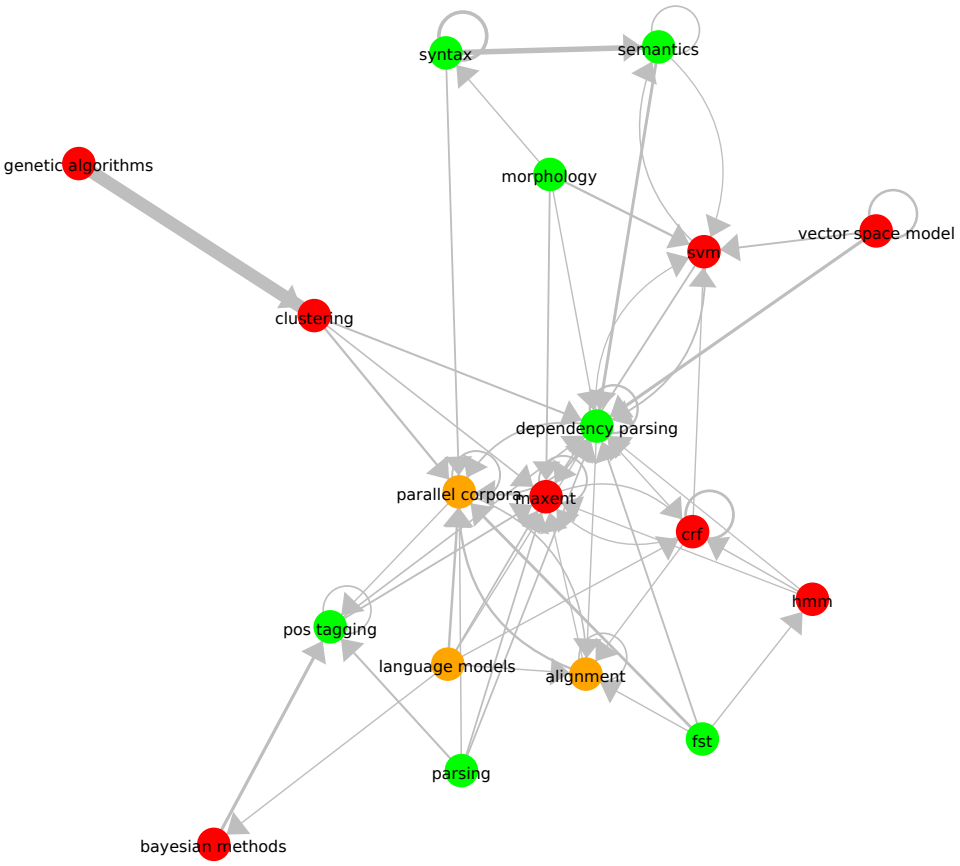



FIGURE 14 – Réseau des flots d'auteurs entre méthodes de la première à la deuxième moitié des années 2000.

dans les années 2000, en particulier au cours des conférences CoNLL de 2006 à 2009 (figure 14). Nous observons aussi que l'analyse morphosyntaxique (*POS tagging*) a toujours occupé un rôle important, ce qui est probablement dû au fait que cette technique est quasi systématiquement utilisée en linguistique informatique comme prétraitement. Enfin, nous remarquons que l'alignement et les corpus parallèles sont devenus majeurs depuis les années 2000, ce qui reflète la popularité de la traduction automatique depuis plus d'une décennie.

## 7 Conclusion

Nous avons présenté une analyse du corpus ACL Anthology visant à en faire ressortir certaines caractéristiques remarquables. Notre analyse se fonde d'une part sur une méthode classique d'extraction de termes, d'autre part sur l'analyse de la structure argumentative des textes considérés afin de catégoriser les termes en fonction de leur contexte et de leur contenu informationnel. Nous avons montré que ce type de technique contribue à affiner la description de la dynamique du domaine.

Il s'agit encore une fois de simples observations. Les outil mettent en avant certains phénomènes qu'il faut ensuite expliquer par un retour aux textes, voire par une enquête de terrain. Cette recherche par nature pluridisciplinaire nous amène maintenant à nous tourner vers des spécialistes d'histoire des sciences pour poursuivre ce travail en collaboration. Les outils et l'infrastructure mise en place sont toutefois d'ores et déjà utilisables et seront appliqués à d'autres corpus, comme le corpus APS présenté dans l'introduction.

## Références

Ashton Anderson, Dan Jurafsky, and Daniel A. McFarland. 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, Corée, 2012. Association for Computational Linguistics.

Rafael E. Banchs, editor. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Association for Computational Linguistics, Jeju Island, Corée, 2012.

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

Didier Bourigault and Christian Jacquemin. Term extraction + term clustering : An integrated platform for computer-aided terminology. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 15–22, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 16(2) :79–85, June 1990.

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

Michel Callon, Jean-Pierre Courtial, and Françoise Laville. 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, 1991.

James Curran, Stephen Clark, and Johan Bos. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Meeting of the Association for Computation Linguistics (ACL)*, pages 33–36, 2007.

K Frantzi and S Ananiadou. Automatic recognition of multi-word terms :. the C-value/NC-value method. *International Journal on Digital . . .*, 2000.

E Garfield. Citation analysis as a tool in journal evaluation. *Science (New York, NY)*, 178(4060) :471–479, November 1972.

Michelle Girvan and Mark E J Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99 :7821–7826, 2002.

R Guimera, Brian Uzzi, Jarrett Spiro, and L A N Amaral. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science (New York, NY)*, 2005.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

Yufan Guo, Roi Reichart, and Anna Korhonen. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 928–937, 2013.

Gary Geunbae Lee, Jong-Hyeok Lee, and Jeongwon Cha. Syllable-pattern-based unknown-morpheme segmentation and estimation for hybrid part-of-speech tagging of korean. *Computational Linguistics*, 28(1) :53–70, March 2002.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75(6) :468–487, 2006.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition : The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007.

Dragomir R. Radev, Hongyan Jing, Malgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6) :919–938, November 2004.

Everett M. Rogers. *Diffusion of Innovations, 1st Edition*. Simon and Schuster, 1962.

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

Henry G Small. 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, 1973.

Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. Using argumentation to retrieve articles with similar citations : An inquiry into improving related articles search in the medline digital library. *I. J. Medical Informatics*, 75(6) :488–495, 2006.

Simone Teufel. *Argumentative Zoning : Information Extraction from Scientific Articles*. University of Edinburgh, 1999.

Simone Teufel and Marc Moens. Summarizing scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics*, 4 (28) :409–445, 2002.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of Empirical Methods in Natural language Processing (EMNLP)*, pages 103–110. Association for Computational Linguistics, 2006.

# Introduction

As a result of the exponential growth of Internet based communication technologies in the last decades, social interactions and knowledge exchange nowadays increasingly take place on digital platforms. People interact on social media like Facebook and Twitter, sharing and discussing links to various digital content. Everyday bloggers and newspapers publish directly on the web thousands of articles that may receive comments and be shared in the social networks. Scientists were among the first to experiment the digitization of part of their daily activity: scientific knowledge is massively available online, stored in large collections of digitized papers. This dissertation aims at seizing this opportunity to analyze the dynamics of scientific communities at various scales – from individual trajectories to the emergence of the structure of scientific fields. The availability of these data allow us to address questions like: How do scientific fields evolve? What is the role of social structures in scientific activity?

The study of scientific field evolution has been an active area of research since the second half of the 20th century. Two main approaches can be identified. On the one hand, philosophers and sociologists of scientific knowledge like Thomas Kuhn (Kuhn, 1962) or Karl Popper (Popper, 1959) have been carrying out qualitative studies and developed theories about scientific endeavor. On the other hand, since the seminal works of Derek J. de Solla Price (Price, 1963) and Eugene Garfield (Garfield, 1972), quantitative studies have given birth to the fields of bibliometrics and scientometrics.

Qualitative research has been focusing on building theories on how scientific knowledge is elaborated, and on the role that social and cultural factors play

in this process. Although qualitative, these theories are nevertheless usually validated through the analysis of empirical case studies.

Nowadays, due to the growing availability of large digital repositories, science evolution can also be investigated by performing extensive quantitative analysis of scientific paper archives. Examples of widely used datasets are PubMed[1] (which is a collection of the literature in the biomedical domain), the American Physics Society dataset[2] (which contains data about over a century of publications in physics), or the Thomson Reuters Web of Science[3], a subscription-based citation indexing service covering sciences, social sciences, arts, humanities, and multidisciplinary research.

Quantitative research investigates science evolution via the statistical and mathematical analysis of empirical data. In the last few years, the growing production and availability on the Internet of huge amounts of data of different kinds (from Wikipedia, news and scientific articles, to people interactions on social networks, for example), have given birth to the so-called "Big Data" science. This trend has lead some thinkers to go as far as to claim that theory and the scientific method are not needed anymore because the automatic analysis of all these data will be enough to understand the world (Anderson, 2008). However, we think this vision is clearly too extreme. Even when performing quantitative studies based on empirical data, scientists still need to formulate specific research questions, and define what objects they want to study. Moreover, they need to choose an appropriate methodology among the different available statistical, mathematical and computational techniques, so as to design new approaches.

Data can be of different kinds. In particular, scientific archives generally contain collections of metadata about a representative set of publications in a given field. For each paper in the collection, the following data are usually available: title, author names and their institutional affiliations, journal in which it was published, publication year, citations made and sometimes received, possibly a few keywords. Because of the development

[1] `http://www.ncbi.nlm.nih.gov/pubmed`
[2] `http://journals.aps.org/datasets`
[3] `http://thomsonreuters.com/thomson-reuters-web-of-science/`

of natural language processing methods in the last few decades, we can nowadays analyze in an automated (at least to some extent) fashion the content of these texts.

However, we know that automatically extracting information from text is not straightforward. Textual data carry complex information and meaning that cannot be directly interpreted by a machine. Methods for automatic information extraction are nowadays available, but we first need to define what kind of information has to be obtained. Common tasks concerning the semantics of texts are, for example, terminology extraction (Kageura and Umino, 1996), named-entity recognition (Finkel et al., 2005), and topic modeling (Blei et al., 2003). Natural language processing tools are now mature enough to reliably extract essential information from texts, and support the exploration of large textual datasets. However, the detection of relations between different pieces of information, and the interpretation of this information, are challenges still far from being solved and constitute the focus of a very active area of research. An important application in the biomedical domain is, for example, the development of natural language processing methods that would make it possible to create comprehensive databases of protein–protein interactions by means of automatic information extraction from the relevant biological literature (Ono et al., 2001).

Another important feature of scientific publication archives is that they contain data spanning over several decades, with a publication date associated with each paper. This is a fundamental information that allows the investigation of the evolution of scientific fields over time. In particular, we can explore the emergence of new research areas, their growth or decline, or any kind of transformations. Moreover, we can explore the individual trajectories of researchers across the different areas of their field, or the dynamics of the whole community of researchers working in a given field.

The analysis of these complex data requires the use of advanced mathematical, statistical and computational methods. Modeling relations between different kinds of objects, and, moreover, modeling how these relations evolve over time, is a non trivial task that lies at the root of complex systems science. This relatively new discipline studies how

interactions among the different components of a system give rise to the emergence of collective behaviors. Examples of complex systems are societies (emerging from people and their interactions) and the brain, formed out of neurons that are constantly exchanging information through their synapses. The study of complex systems can be carried out thanks to network theory, namely the science of representing complex systems as graphs illustrating relations between discrete objects. This active field of research constitutes a proper framework for our study, therefore it is at the heart of the methodology developed in this thesis.

The goal is to provide a novel approach for the exploration of large scientific archives and allow their interpretation thanks to network theory and advanced natural language processing techniques, making it possible to automatically access the content of scientific papers.

In particular, I focus on the social and semantic dimensions of scientific research. I am interested in understanding the role of collaborations among scientists in building and shaping the landscape of scientific knowledge. Symmetrically, I investigate how the configuration of this landscape influences individual trajectories of scientists and their interactions. Ultimately, my goal is to understand how the social and the semantic dimensions co-evolve over time, leading to the evolution of scientific fields. I carry out this research project by studying a specific field of research, *i.e.* computational linguistics, through the analysis of the ACL Anthology[4] dataset, that constitutes a representative collection of publications in the field. A second dataset, concerning physics (APS Dataset[5]), is also used for some of the analyses, in particular to check the generality of our approach.

Scientific archives have already given birth to a large body of research. For what concerns the social dimension, a number of works reconstructed and analyzed the structural properties of co-authorship networks representing scientific communities in different fields, such as physics (Newman, 2001b) and mathematics (Grossman, 2002). Since the seminal work of Callon (Callon et al., 1991), researchers have also reconstructed networks representing the structure of the knowledge produced in different fields,

[4] `http://www.aclweb.org/anthology-new/`
[5] `https://publish.aps.org/datasets`

and its evolution over time.

Most studies have focused on only the social or only the semantic dimension at a time. A notable exception is (Roth, 2005), in which the author showed that interactions between actors in knowledge communities are driven by both their social and semantic similarity. Building on this work, in this thesis I would like to explore the relation between the social and the semantic network: what is the influence of the social structure on research collaboration? On the other hand, how does the knowledge landscape influence exchanges and collaborations? Thanks to the availability of these large datasets spanning over several decades, it is now possible to propose new methods so as to automatically explore the time-evolution of keywords, authors, and their interactions, and therefore try to explore how these different dimensions co-evolve over time.

More specifically, the scientific contribution of this thesis relies on the three following points.

1. My contribution to the field of computational linguistics is a new method for the automatic characterization of terms extracted from the abstracts of scientific papers. This method is based on a fine-grained categorization of terms depending on their context of use, performed by means of a technique called argumentative zoning.

2. My second contribution concerns the field of complex network analysis. I perform a systematic investigation of the role of social and semantic features in the dynamics of socio-semantic networks modeling scientific research. This investigation leads to the formulation of a statistical model based on temporal data, and relying on multivariate logistic regression.

3. Thirdly, and more generally, this thesis presents a multi-scale description of the evolution of the field of computational linguistics, with a particular focus on its methods and techniques, performed by combining quantitative methods from different disciplines, and qualitative interpretation by experts of the field.

Even if most of the analyses have been performed on the ACL Anthology

dataset, the approach and the techniques described in the thesis are more general and could be applied to other corpora as well. For example, the analyses of the APS dataset have been performed to confirm this, so we are confident that our results are valid beyond the ACL Anthology corpus.

The rest of the dissertation is structured as follows. In the first part I review the literature on the subject (Chapter 1), discuss the methodological foundations of my work, and introduce the data on which our analyses are performed (Chapter 2).

In the second part I present the modeling framework. Firstly, in Chapter 3 I explain the methodology used to extract pertinent phrases from titles and abstracts, and in particular to characterize them in order to identify the terms corresponding to methods and techniques. Then I present how the social and semantic landscapes of scientific fields can be modeled as complex networks connecting researchers and scientific concepts (Chapter 4). Lastly, in Chapter 5 I introduce a representation of these networks as time-evolving systems.

The third part is dedicated to the investigation of the dynamics of the socio-semantic landscape of scientific fields at different scales. In Chapter 6 I present a microscopic level analysis of the dynamics of socio-semantic networks, and investigate the role played by social and semantic factors. Then in Chapter 7 I focus on the meso-level structure, namely I explore the evolution of computational linguistics by mapping the birth, growth, split and merge of the different research areas within the field. Finally, in Chapter 8, I investigate the individual trajectories of scientists across the different research areas, and in particular across the different methods and techniques, in an effort to bridge the micro and meso description levels.

# Part I

# Methodological foundations

# Chapter 1

# State of the art

**Contents**



The study of the evolution of scientific fields, and of the scientific activity in general, constitutes an active area of research since the second half of the 20th century. As already mentioned in the introduction, we can identify two main approaches to address this question. On the one hand various theories about science dynamics have been blooming based on qualitative studies carried out by philosophers, historians or sociologists of science. On the other hand, quantitative studies have given birth to the fields of bibliometrics and scientometrics. We will present and discuss these two approaches in this chapter and we will examine if they can be reconciled through quali-quantitative studies.

## 1.1 From early qualitative studies to the advent of big data

### Early qualitative studies

On the qualitative side, we recall the seminal work of Thomas Kuhn. In his book *The Structure of Scientific Revolutions* (Kuhn, 1962), Kuhn claims that science evolves through shifts from old to new "paradigms". In his vision, the acceptance or rejection of a particular paradigm is not only a logical process, but a social process too. Following a scientific discovery, widespread collaboration is necessary to establish a new framework.

In the 1970s, Nicholas C. Mullins published his paper *The Development of a Scientific Specialty: The Phage Group and the Origins of Molecular Biology* (Mullins, 1972). Therein he demonstrates that the birth of a new discipline cannot be explained only by means of the competitive position and relative status of each of the specialties from which it is formed, but intellectual and social activities need to be taken into account too. Specifically, he investigates in detail the development of molecular biology from the American academic group studying the bacteriophage (a virus which infests bacteria) in the mid-20th century. Mullins shows that the development of this new discipline was possible thanks to the successful growth of the network of communications, co-authorship, colleagueship, and apprenticeship.

### The "actor-network theory"

In the 1980s, French scholars Michel Callon and Bruno Latour, and collaborators, developed the so-called "actor-network theory" (ANT) (Latour, 1987), in response to the need of a new social theory adjusted to science and technology studies.

Their approach differs from traditional sociology since they claim that there exists no such thing as a 'social context' to explain the features that economics, psychology, linguistics, and other sciences cannot account for. Latour defines the 'social' as "a trail of *associations* between heterogeneous

elements", and 'sociology' as "the *tracing of associations* between things that are not themselves social" (Latour, 2005). Therefore in this context the 'social' is what emerges from the associations between different actors, and not a distinct domain of reality defined *a priori*.

ANT is based on the assumption that sociologists should track not only human actors, but also all the other non-human elements involved in the process of innovation and creation of knowledge in science and technology. To give an example, Callon explains that the history of the American electrical industry is not reducible to its inventors and their relations. To understand it we need to also take into account intellectual property, patent regulation, and the electric technologies themselves, and build a network that traces the associations between all these human and non-human actors (Callon and Ferrary, 2006).

In this context, to study the evolution of science, we should track not only researchers but also the traces they disseminate, especially their publications. The texts and the ideas therein play a central role and, moreover, in the ANT framework, these are put on the same level as human actors.

ANT received several critics, mainly because of the role given to non-humans, which are not capable of intentionality and should therefore not be put at the same level as human actors, according to (Winner, 1993). However, this methodology is still actively used today and we think that its founding principles are inspirational. Therefore, in this thesis we also consider both humans (researchers) and non-human actors (scientific concepts). Our approach is inspired by ANT, but also presents some differences that we will detail in Section 1.3.

### Scientometrics

While the "actor-network theory" was developed by Callon and Latour, quantitative analyses of scientific activity also started to be carried out, giving birth to the field of scientometrics. Pioneers in this field were Eugene Garfield, who created the first scientific citation index (Garfield, 1979), and Derek John de Solla Price, who analyzed the growth of science (Price,

1963), and proposed the first model of growth of networks of citations between scientific papers (Price, 1965). A dedicated academic journal, Scientometrics, was created in 1978.

As interestingly described in (Leydesdorff and Milojević, 2012), whereas in the 1980s sociology of science started to increasingly address micro-level analysis focusing on the behavior of scientists in laboratories (Latour and Woolgar, 1979), scientometrics focused on the quantitative analysis of scientific literature at the macro scale, often considering a whole discipline. Therefore, since then, the field of science and technology studies increasingly bifurcated into two streams of research: on the one hand the qualitative sociology of scientific knowledge, and, on the other hand quantitative studies of scientometrics and science indicators, which soon involved evaluation and policy issues too.

### Network oriented studies

During the first decade of this century, the increasing availability of scientific publication archives, and the development of network science, led to large scale studies of co-authorship and citation networks, since the seminal work of Newman (Newman, 2001d)[1]. The framework of network theory allows new kinds of studies, based on the relationships between authors and papers, such as the investigation of the heterogeneity in the number of collaborators, the transitivity of collaborations, and the emergence of community structure, in which authors and papers are clustered in different groups, often corresponding to expertise in different subfields of science (Girvan and Newman, 2002). Moreover, new network visualization techniques allow to study science and its different disciplines through maps representing the landscape of scientific knowledge (Börner et al., 2003). These new interdisciplinary exchanges between scientometrics, computer science and physics has lead to an impressive growth of scientometrics studies, making the discipline a very active area of research (Leydesdorff and Milojević, 2012).

[1] The idea of studying co-authorship patterns was firstly introduced in (Mullins, 1972), but Newman's work represents the first detailed reconstruction of an actual large-scale collaboration network.

## 1.2 False ideas about big data: "qualitative vs quantitative"

The debate on qualitative versus quantitative research is very active today. The availability of large datasets is in fact not restricted to scientific archives but, thanks to the exponential growth of the Internet and related technologies, a huge amount of data is now produced online and partly available, like for example interactions between people on social networks such as Twitter, Facebook, etc. This has lead to the so-called "Big Data" science. A few years ago Chris Anderson, at that time editor-in-chief of *Wired*, wrote:

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. [...] The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists. The models are then tested, and experiments confirm or falsify theoretical models of how the world works. This is the way science has worked for hundreds of years. [...] But faced with massive data, this approach to science - hypothesize, model, test - is becoming obsolete. [...] We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot. (Anderson, 2008)

We think that this vision is too extreme and even wrong for certain aspects. The size of the data and the available computational power are of course contributing to revolutionize the way research in social sciences is done, but it is misleading to think that using digital traces is a straightforward process that makes all social theories obsolete. The ideas defended by Anderson are

based on a series of wrong assumptions, as extensively discussed in (Boyd and Crawford, 2011).

Firstly, it is not because a corpus of data is large that it is representative. Therefore, the documents to study should be chosen with care. They should contain representative information and be diversified so that sampling does not introduce a bias and does not affect the observed dynamics. Moreover, digital traces are rarely directly usable: they must be cleaned so as to make sense. While it is true that digital traces give a more direct access to phenomena that were hard or even impossible to observe before the digital era, these traces are rarely directly usable. They must be selected, cleaned and organized. One should also keep in mind that digital traces are generally not produced for social science scholars (they are rather produced for observation, surveillance or information sharing purposes). Thus, they reflect a specific point of view or interest towards a given phenomenon and this point of view can be radically different from that of the social scientists. It is thus important to keep in mind what are the data, and possibly what is missing among these data.

Scientific archives are particularly good datasets with respects to this point. The two datasets that we analyze in this thesis do not contain all the publications in the concerned discipline, but they should be considered representative since they contain all the publications in several journals and conferences of the respective disciplines. Moreover they were produced for scholars for research purposes, as stated by their providers[2].

A second and more fundamental issue concerns the definition of the objects under study. Social sciences often operate with complex notions (science fields, sociological categories, etc.) which are hard to define formally and do not have precise boundaries (how to define the frontiers of a research area?). Digital traces provide very little help for defining relevant categories

[2]Concerning the APS dataset: "Over the years, APS has made available to researchers data based on our publications for use in research about networks and the social aspects of science. In order to further facilitate the use of our data sets in this type of research, researchers may now request access to this data by filling out a simple web form." [`http://journals.aps.org/datasets`]
Concerning the ACL dataset: "This is the home page of the ACL Anthology Reference Corpus, a corpus of scholarly publications about Computational Linguistics. [...] We hope this corpus will be used for benchmarking applications for scholarly and bibliometric data processing." [`http://acl-arc.comp.nus.edu.sg/`]

and formalize them so that one can track their evolution in a longitudinal corpus for example.

Moreover, automation is not straightforward. By this observation, we mean that even if the data are clean and representative, they must be organized. Numerical methods are not neutral since any computing method involves some choices. What kinds of calculations are applied to the data? Even if the computer can automate some calculations, it does not give any insight on what kind of measure or modeling should be done. This is of course far from neutral: there are different ways to compute the similarity between two concepts, or the influence of the context, for instance.

**Relationality and temporality**

Another fundamental characteristics of big data, underlined by (Boyd and Crawford, 2011), is the following:

> Big Data is notable not because of its size, but because of its relationality to other data. Due to efforts to mine and aggregate data, Big Data is fundamentally networked. Its value comes from the patterns that can be derived by making connections between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself.

The characteristic relationality of data constitutes a new challenge that can be addressed in the framework of network theory, namely the science providing the mathematical tools to analyze and model complex systems as graphs illustrating relations between discrete objects. Network science allows us to uncover the properties of complex networked systems at different scales: from patterns of centrality and of similarity between objects, to the emergence of aggregates and connections among them.

Moreover, recent studies have focused on the analysis of the temporal aspect of networks (see (Holme and Saramäki, 2012) for a review on the subject). Thanks to the availability of data spanning over several

decades, we can nowadays empirically investigate the evolution of social systems and scientific activity with the help of new methods, still under development, to model time-evolving objects and relations.

We should underline that the use of network theory is not a straightforward process either. Network analysis provides a suitable framework and useful mathematical and computational tools, but some more fundamental questions in its applications to science studies remain: what objects do we want to study and use as nodes of our network, and how do we extract them from the data? How do we identify and quantify the strength of a relation between two objects? What measures on the resulting network are interesting for our study and would be useful to answer questions on the evolution of scientific fields?

### Networks and social science theory

The methodology we follow in this thesis, based on the notion of network, does not rely on traditional social science theories. As very well explained in (Callon and Ferrary, 2006), the network-based approach has a series of advantages.

In particular, this approach let us avoid making use of sociological categories, and of a strict distinction between micro and macro structures. In our methodology, 'groups' (also called 'communities' in network terminology) are defined as emerging structures in the network, namely as sets of nodes highly connected among each other, and loosely connected with the rest of the network. This is what we call the 'mesoscopic' level, and this is what we will use to model groups and communities, instead of accepting the traditional sociological definitions of these notions, which we think are too subjective.

## 1.3 Towards socio-semantic networks representing scientific research

This thesis constitutes a contribution to scientometrics studies. More precisely, we want to apply network theory to the study of the evolution of scientific fields. Many aspects can be studied, such as citations and their impact (Hirsch, 2005), collaborations (Newman, 2004), geographical distribution of laboratories and publications (Frenken et al., 2009), and research funding (Boyack and Börner, 2003). In this thesis we focus in particular on the social dimension of scientific production, and on the distribution of the resulting knowledge.

### Co-authorship networks

Examining collaborations among researchers can capture the social dimension of science. This kind of information can be extracted directly from publications by tracking co-authorship. From this information we can reconstruct networks of collaborations in different disciplines. In the last few decades a number of works reconstructed large-scale co-authorship networks representing scientific communities in different fields, such as physics (Newman, 2001b,c), mathematics (Grossman, 2002), neuroscience (Barabási et al., 2002), biomedical research, and computer science (Newman, 2001d).

These structures reveal interesting features of scientific communities. It has been shown that all fields seem to have a heterogeneous distribution of the number of collaborators per author, with most researchers having only a few collaborators, and a few having hundreds or in some extreme cases even thousands of them. Moreover, any researcher in the network can be easily reached from any other author in a small number of steps (moving from collaborator to collaborator)[3]. Relations also tend to be transitive: if two researchers both collaborated with a third researcher, chances are that the former are also co-authors. Lastly, these networks also appear to have a well

[3] as long as they belong to the same connected component, whose precise definition will be given in Chapter 4.

defined community structure: researchers actually tend to group together so as to form scientific communities working on the same research topic or methodology (Newman, 2004).

Scientific collaboration networks have also been explored from a temporal perspective. (Newman, 2001a) shows that the probability that a researcher has new collaborators increases with the number of her/his past collaborators, and that the likeliness that two researchers initiate a new collaboration increases with the number of collaborators they share. (Barabási et al., 2002) then proposed a model for the evolution of co-authorship networks based on preferential attachment, i.e. on the idea that the more collaborators a researcher already has, the higher the probability that she/he will collaborate with even more scholars in the future. Since then, other works have explored the role of preferential attachment in the time-evolution of other empirical co-authorship networks using, for example, the Web of Science database (Wagner and Leydesdorff, 2005; Tomassini and Luthi, 2007).

(Guimerà et al., 2005b) investigate instead the mechanisms that lead to the formation of teams of creative agents, and how the structure of collaboration networks is determined by these mechanisms. Team organization and functioning has been widely investigated also by (Monge and Contractor, 2003). (Lazega et al., 2008) explore the interdependencies between collaboration networks and inter-organizational networks connecting the scientific laboratories in which researchers work. Other works have explored topological transitions in the structure of co-authorship networks as the corresponding scientific field develops (Bettencourt et al., 2009), or the emergence of disciplines from splitting and merging of social communities (Sun et al., 2013).

In this thesis we propose a model of growth of co-authorship networks which is based not only on preferential attachment mechanisms and social features, such as the number of common collaborators, but also on researcher similarity, as expressed through knowledge production and investigation.

**Co-word networks**

We elaborate on the idea of Callon and Latour that the process of creation of knowledge can be understood only by tracking human and non-human actor traces, and analyze not only collaboration structures but also the semantic content of scientific publications. We do not directly consider the papers as nodes of our networks: instead we base our analysis on the relations between researchers and concepts extracted from the text and/or the metadata. Therefore the networks we build are composed of both humans (researchers) and non-human actors (scientific concepts), but we make a distinction between the two, and call 'social' the associations between researchers (that we trace through co-authorship), and 'semantic' the associations between concepts (that we trace through co-occurrences in texts). We thus acknowledge that equal importance should be given to the social and the semantic dimensions, but still assume that the two types of links are not equivalent, as they may support different processes.

As already said, the analysis of texts is not straightforward. Firstly, we need to define what kind of information we want to extract from them, then we need to find or develop the methods to do it, which, for large datasets, should be automated tools. Finally, we need to understand how we can connect together the different pieces of information. These issues are at the core of nowadays sociology research, as explained by (Venturini et al., 2012):

> Quantitative data can have many different forms (from a video recording to the very memory of the researcher), but they are often stored in a textual format (i.e. interviews transcriptions, field notes or archive documents...). The question therefore becomes: how can texts be explored quail-quantitatively? Or, more pragmatically, how can texts be turned into networks?

In this thesis, we try to extract, from the titles and abstracts of the papers, the terms that correspond to scientific concepts, in order to reconstruct the landscape of knowledge distribution of scientific fields (Callon et al., 1986). Moreover, we introduce an original method to automatically classify these terms in order to extract the ones corresponding to techniques, so that we can study more fine-grained facts about the evolution of scientific

fields. On a methodological level, we therefore combine a network theory approach with computational linguistics methods that make it possible to automatically extract information directly from the publication content[4].

Relevant maps of scientific domains can be built using network theory: in this context, nodes of the network correspond to terms extracted from texts and two nodes are connected if the corresponding terms co-occur together in different papers or abstracts (Eck, 2011). For example, (Cambrosio et al., 2006) use inter-citation and co-word analysis to map clinical cancer research.

The study of the time-evolution of the structure of different scientific fields through co-word network representations is a prolific area of research, and different disciplines have been analyzed, such as chemistry (Boyack et al., 2009) physics (Herrera et al., 2010), (Pan et al., 2012), and biology (Chavalarias and Cointet, 2013). In this thesis we mainly focus on the evolution of the field of computational linguistics, and, to some extent, also to the evolution of physics research.

### Socio-semantic networks

The originality of this thesis relies in the fact that we consider the social and the semantic dimension of science *at the same time*, and we try to uncover how these two dimensions co-evolve over time. We rely on the work of (Roth, 2005), which was among the first to consider interactions in knowledge communities in both their social and semantic dimensions. Roth analyzes the community of biologists studying the zebrafish, and shows that collaborations are driven both by social distance and by semantic proximity between researchers. However his approach focuses on only one variable at a time, and ignores the simultaneous effect of parameters with respect to each other.

The original contribution of this thesis, with respect to the work of Roth, is to build a more holistic statistical model that takes into account all features at the same time. Moreover, we explore the evolution of collaboration networks, but also the evolution of co-word networks representing scientific

[4]Natural language processing methods for term extraction are reviewed in Chapter 3.

knowledge, and build a comprehensive model based on social and semantic features.

Our work is also largely inspired by the previous study of (Anderson et al., 2012). The field of computational linguistics has been the subject of several scientometric studies in 2012, for the 50 years of the Association for Computational Linguistics (ACL). More specifically, a workshop called "Rediscovering 50 Years of Discoveries" was organized to examine 50 years of research in natural language processing (NLP). This workshop was also an opportunity to study a large scientific collection with recent NLP techniques and see how these techniques can be applied to study the dynamics of a scientific domain. The paper "Towards a computational History of the ACL: 1980-2008", published in the proceedings of this workshop, is very relevant from this point of view. The authors propose a methodology for describing the evolution of the main sub-domains of research within NLP since the 1980s. They demonstrate, for instance, the influence of the American evaluation campaigns on the domain: when a US agency sponsors a sub-domain of NLP, one can observe a rapid concentration effect since a wide number of research groups suddenly concentrate their efforts on the topic; when no evaluation campaign is organized, research is much more widespread across the different sub-domains of NLP.

Similarly, we propose to study the evolution of the field of computational linguistics, but we also make a technical contribution to the field itself, as we introduce a new method to automatically categorize keywords according to the information they carry. Among all the terms relevant in the domain, we are especially interested in terms referring to methods and techniques since these terms make it possible to trace the technical evolution of the field.

# Chapter 2

# Methods and data

**Contents**



## 2.1 Overview of the methodology

In this thesis we would like to understand the influence of scientific collaborations on the configuration of the landscape of scientific knowledge. Symmetrically, we would like to explore how the configuration of this landscape influences individual trajectories of researchers and their interactions. Ultimately, our goal is to uncover how the social and the semantic dimensions co-evolve over time, leading to the evolution of scientific fields.

To reach this goal, we perform a study based on the empirical analysis of a scientific publication archive. We carry out a quali-quantitative study based

on mathematical and computational analyses of metadata and texts, and on the validation and interpretation of the results of these analyses by experts of the corresponding field.

Several archives collecting most of the publications in a given discipline over several decades are nowadays available in digital format, and some of them are freely downloadable for research purposes. This is the case for example of the ACL Anthology, which contains publications in the field of computational linguistics. In particular, the ACL Anthology supplies the abstracts of the papers[1], providing a direct access to the textual content of publications. This is the main dataset analyzed in this thesis.

To test the robustness of some specific results (*i.e.* the statistical modeling of the dynamics of the social and semantic networks), we also used a second dataset, namely the American Physical Society (APS) dataset, which contains metadata about over a century of publications in physics.

A detailed description of the characteristics of both datasets is given in the second part of this chapter (Section 2.2). It is important to underline that even though we analyze two particular corpora, our approach is not bounded to these specific datasets and could easily be generalized to any other scientific archive with the same characteristics.

### 2.1.1 Modeling the socio-semantic space of scientific fields

What is the role of social structures in scientific activity? To address this question we first focus on the **researchers** in a given field, and on their interactions. In our study the set of researchers of a discipline simply consists of the authors of the publications in the corresponding dataset.

What knowledge is produced by the scientific endeavor and how does it evolve over time? To address this second question, we focus on the knowledge produced in the field, that we trace by analyzing the textual content of each publication. We use terms extracted from titles and abstracts, which we assume to correspond to the **scientific concepts** and

[1] Full text is also available but is the result of an automatic OCR conversion which leaves too much noise and formatting issues to make a reliable automated analysis possible.

methods specific to the field.

Contrary to author names which are generally provided in publication archives as metadata and are therefore directly extractable, it is not always straightforward to identify the most salient concepts addressed in a paper. We thus need a method to extract key information representing scientific concepts and methods, which is known to be a difficult task. We could ask an expert of the field to make a list of the concepts characterizing the discipline, but this would be biased by subjectivity and probably suffer from incompleteness. Alternatively, these concepts can be retrieved through an automatic analysis of the textual content of the papers. Through natural language processing methods we can automatically retrieve sequences of words (n-grams) corresponding to terms (precisely noun phrases) that represent scientific concepts. The obtained list is then filtered and validated by experts of the field. This is the approach that we follow in this thesis. The corresponding work is presented in the first part of Chapter 3.

The terms specific to a scientific field can be of different nature: they can for example correspond to tasks, methods, or evaluation procedures. Application-oriented disciplines like computational linguistics are in fact characterized by tasks (such as machine translation or word sense disambiguation, for example) and by different methods developed to achieve such tasks (such as the different machine learning techniques used in the field for instance). The correspondence between tasks and methods is not univocal, because different methods can be used to perform the same task, and at the same time one method can be used for different tasks. To investigate fine-grained facts in the dynamics of scientific production, it would thus be useful not only to extract the relevant terms, but also to identify which ones correspond to methods. Therefore, we introduce an original method that combines automatic term recognition (Kageura and Umino, 1996) with argumentative text zoning (Teufel, 1999). This new approach relies on the idea that terms can be categorized using the information given by the context in which they are used. This work is presented in the second part of Chapter 3.

To summarize, we create two lists of terms. The first list contains all the

scientific terms characterizing the field of computational linguistics. The second list only contains the terms corresponding to methods and techniques used in the field. In the rest of the thesis, when investigating specific questions regarding the methods and techniques of the discipline, we use the second list, whereas, for the other analysis, we use the first, which is more relevant when studying the whole field.

In the physics dataset, no textual content is available[2]. Therefore we use the keywords provided by the authors, but it is not possible to directly infer a list of terms referring to techniques, using the above methodology.

Once the objects of study are identified (researchers and scientific concepts), we want to analyze the **relations** between them. As already said, a suitable framework to study entities and relations between entities is network theory, that we introduce in Chapter 4.

In this chapter, we propose to represent the social space by building a **social network** in which the nodes are the researchers of the field. To this end, we have to define a way to identify the presence of a connection between two researchers, and to measure its strength. Checking if two researchers belong to the same institution could be an option, but most institutions are composed of many people working in different areas (even within the same laboratory), who do not have real connections with respect to scientific research. Moreover, during the course of their career, people move from institution to institution, and therefore tracking relationships on the basis of institutional affiliations across several decades constitutes a tricky task. A both simpler and more accurate indicator is co-authorship. Newman, who makes the same choice, gives the following justification (Newman, 2001d):

> I study networks of scientists in which two scientists are considered connected if they have coauthored a paper. This seems a reasonable definition of scientific acquaintance: most people who have written a paper together will know one another quite well. It is a moderately stringent definition, since there are many scientists who know one another to some

[2]Except for the titles, but we consider them too little information to derive a satisfying description of a paper.

> degree but have never collaborated on the writing of a paper. Stringency, however, is not inherently a bad thing. A stringent condition of acquaintance is perfectly acceptable, provided, as in this case, that it can be applied consistently.

Therefore in our social network two researchers are connected if they co-authored a paper, and the strength of the connection is directly proportional to the number of papers they co-authored.

Secondly, we try to represent the knowledge landscape by building a **semantic network** in which the nodes are the scientific concepts specific to the field. When should two concepts be linked? Relatedness between concepts is typically measured using co-occurrence relations between the corresponding terms (*i.e.* between linguistic sequences referring to the different concepts of the field). Therefore in our network two concepts are connected if they are employed by researchers in the same context, namely if the corresponding terms are found together in titles or abstracts, or if the corresponding keywords are both provided by the authors to describe the content of a paper. The strength of these relations can then be determined through the frequency of these co-occurrences.

Lastly, to unveil the relations between researchers and concepts, and to eventually investigate the co-evolution of the social and the semantic space, we also build a **socio-semantic network** composed of two types of nodes: researchers and concepts. In addition to the links between researchers and those between concepts previously defined, this network also includes links between researchers and concepts. We consider that a researcher and a concept are connected if the former has used the term corresponding to the concept in the title or abstract of his/her papers, or has listed the corresponding keyword to describe the content of some of his/her papers.

In our analysis, time is a fundamental element since we want to investigate the evolution of social and semantic relations, and their evolution. Therefore the networks that we build are **temporal networks**, with a time granularity of one year. We start by considering the first year in a given interval (from 1980 to 2008, as we will explain in the next section), and build the networks described above considering all the papers published that year. Then, for

every subsequent year in the interval, we build a new network that contains all the nodes and links present in the network of the previous year, and then enrich it by looking at the papers published during the year under consideration and adding the new authors and concepts that appear, and all the new relations. In this way, we can trace how social and semantic relations change over time, in particular we can trace when new relations emerge, and when and to what extent existing relations are strengthened.

### 2.1.2 Investigating the socio-semantic dynamics of scientific fields at different scales

In this thesis we want to investigate whether the likelihood that two reseachers collaborate is purely a function of 'social' factors (*i.e.* the number of common collaborators), or if 'semantic' factors (*i.e.* the number of concepts they both already investigated) also impacts this process. Symmetrically, we want to understand if the likelihood that two concepts become connected is purely a function of their 'semantic' proximity (*i.e.* the number of other concepts they both co-occur with), or if their 'social' similarity (*i.e.* the number of researchers that already addressed these two concepts) plays a role too. Finally, we want to explore whether both the social and the semantic dimensions play a role in the likelihood that a researcher will investigate a given concept she/he has never worked on before.

To this end we look at the evolution over time of the networks defined in the previous section, and build a statistical model based on multivariate logistic regression. This approach allows to understand how the probability of creation of a new social, semantic, or socio-semantic link is dependent on a set of social and semantic variables. In the course of the thesis, we will show that our results confirm our hypothesis that both the social and the semantic dimensions play a significant role in the evolution of scientific collaboration and knowledge production.

Our model focuses on the emergence of interactions between the actors of the system, and by the role played by social and semantic features characterizing these actors. This is the 'microscopic' level of analysis.

The analyses of the social and semantic networks that we present in Chapter 4 show that these networks are characterized by a well defined community structure, meaning that in both networks there are groups of nodes highly connected among each other, and more loosely connected with the rest of the network. The emergence of these social and semantic aggregates led us to focus then on what we call the the 'mesoscopic' level, an intermediate level between the individual actors and the totality of the network.

In particular, we make the hypothesis that the aggregates of concepts emerging in the semantic network might correspond to the different sub-areas of the research field under study. We therefore analyze the semantic network at the meso-level by applying community detection algorithms, which implement techniques dedicated to unveil the different highly connected clusters of nodes in a network. By combining community detection and network visualization tools, we explore the meso-level structure of the semantic network of computational linguistics, and its evolution over time. A thorough evaluation of our results has been done by experts in the field. They confirmed that the results reflect the different trends in the field and their evolution. Network analysis and its methods are therefore able to provide a representation of the characteristics of scientific fields and of their evolution over time based on the automatic analysis of the textual content of publications, and not on a theoretical *a priori* reconstruction by a human expert.

Lastly, we investigate researcher individual trajectories across the different research areas of computational linguistics. This last analysis is an attempt to uncover the interaction between the microscopic and the mesoscopic level of description. We analyze individual trajectories (microscopic level) across the semantic aggregates representing the different research areas and classes of methods (mesoscopic level). We analyze in particular the characteristics of the researchers that introduce methodological or thematic innovation in the field. Moreover, we find a positive correlation between the size of the flow of researchers moving from one research area to another, and the strength of the semantic links crossing the two corresponding areas. The presence of this correlation is a further confirmation of the high degree of interrelatedness between the two dimensions through which we propose to

analyze science: the semantic dimension giving an overview of the knowledge landscape emerging from scientific publications, and the human dimension based on the researchers working in the field.

## 2.2 Data description

As already said, our investigation of the evolution of scientific fields is mainly based on the analysis of the ACL Anthology dataset. This corpus has been chosen among the existing publicly available datasets for two main reasons: *i)* we had access to different experts who could provide valuable comments on the interest and interpretation of our analyses, and *ii)* this domain has never been analyzed to such extent before.

We also perform some of the analyses presented in this thesis on a second publicly available corpus, *i.e.* the APS dataset.

These two corpora cover two very different domains: physics belongs to natural sciences, whereas computational linguistics is closely related to computer science but also involves models and theories developed in linguistics or even cognitive science. The second difference is the size of the datasets: the computational linguistics archive contains about twenty thousands papers, whereas the physics corpus contains a few hundreds of thousands. We chose to study these two rather different corpora so as to observe their differences and, perhaps more interestingly, their similarities. We hope to observe some common behavioral patterns that could give some hints for the study of other corpora of this kind.

It is important to remark that these datasets do not cover the totality of the publications in the given field. They are nevertheless the largest collections available in each field, and include different journals and conference proceedings. They can therefore be considered a good sample of the respective scientific areas. The lack of exhaustiveness is nevertheless something to keep in mind when interpreting the results obtained, and one should be cautious that the results presented in this thesis are based on data that are not fully comprehensive, but as far as possible, representative. It is anyway impossible to be fully comprehensive and studies always concern specific archives or sets of archives. It should also be noted that what we call a “scientific field” or “domain” is a useful abstraction but cannot be formally defined and has no clear boundaries.

### 2.2.1 ACL dataset

The first dataset is the ACL Anthology[3], which is a digital archive of conference and journal papers in natural language processing and computational linguistics. This corpus is mostly based on the journal *Computational Linguistics*, and on the proceedings of the conferences organized by the Association for Computational Linguistics (such as ACL, NAACL, and EACL for example). The proceedings of the COLING conference and of some other newer conferences like LREC are also included.

The ACL Anthology has received recent attention thanks to the 2012 ACL Special Workshop "Rediscovering 50 Years of Discoveries"[4], that produced a few papers on the topic, among which a "history" of the field by Jurafsky and collaborators (Anderson et al., 2012). Since we would like our analyses to be comparable with theirs, we also restrained our analysis to the articles published during the period 1980-2008 already used by Jurafsky and his colleagues. In order to extract concepts from this dataset, we used natural language processing methods able to recognize the most relevant keywords from titles and abstracts. The analysis combines together linguistic and statistical features. The resulting dataset consists of 10128 papers, 8725 authors and 665 concepts.

### 2.2.2 APS dataset

The APS dataset[5] contains metadata about over 450000 articles published in the journals *Physical Review Letters*, *Physical Review*, and *Reviews of Modern Physics* from 1893 to 2009.

Part of these metadata consists in the PACS codes characterizing each article. The PACS system is "a hierarchical subject classification scheme designed to classify and categorize the literature of physics and astronomy"[6]. These codes are provided by the authors in order to characterize the content

[3] `http://www.aclweb.org/anthology/`
[4] `http://translit.i2r.a-star.edu.sg/r50/`
[5] `https://publish.aps.org/datasets`
[6] `http://www.aip.org/pacs/`

of their papers. Each code thus indicates a physical concept addressed in the scientific paper under consideration, like for example "neutrino interactions" or "solid-liquid transitions". This classification system was introduced in 1970, but it is only from 1985 on that the majority of the articles in the dataset are assigned such codes. Since concepts constitute a key feature in our analysis, we restrict our analysis to the articles published in the years 1985-2009, with PACS codes.

We have also filtered this dataset by eliminating all the articles with 10 or more authors, as suggested in (Martin et al., 2013), in order to get rid of the publications in the experimental particle physics domain, which are signed by often hundreds or even thousands of authors. This happens because all the people working in the corresponding consortium are included in the list, even though there was probably no real direct collaboration among all of them. Therefore excluding those articles makes it possible to avoid topological artifacts in the co-authorship network. Moreover, in this dataset, author names are not uniquely identified. As a consequence there may be an issue for polysemous names[7], especially concerning very common Asian names. Several name disambiguation techniques have been proposed, but there is no consensus yet on what would constitute a really effective method. Therefore in order to minimize this problem, we decided to restrict our analysis to a corpus constituted by the subset of papers published in European institutions. This is surely a drastic filtering strategy, but it makes it possible to minimize name ambiguity issues, which would introduce non negligible artifacts in the network. Therefore, the analysis performed on the resulting dataset (which consists of 98404 papers, 95043 authors and 5078 concepts) describes the evolution of European physics only.

[7] In the ACL dataset this issue is not as relevant because all the papers in the collection provide full names encoded in the same format.

# Part II

# Modeling the socio-semantic space of scientific research

# Introduction

In this second part of the thesis we introduce the modeling framework of our study.

In Chapter 3 we present our method to "model" *texts* to extract information needed to define the semantic space of scientific research in a given domain. We want to extract, from the titles and abstracts of the scientific publications of the field under study, the terms corresponding to scientific concepts characterizing this scientific field. We perform this task by using state-of-the-art automatic terminology extraction tools. We then introduce a new method to identify terms corresponding to methods and techniques used in the field. This categorization makes it possible to study more fine-grained facts in the evolution of computational linguistics.

In Chapter 4 we present a model for the *relations* between researchers and between scientific concepts. In particular, we introduce network theory and explain how the socio-semantic space of scientific research can be modeled as complex networks connecting scientists and scientific concepts. Moreover, we analyze the characteristics of the resulting networks and discuss their properties.

Finally, in Chapter 5, we introduce a representation taking into consideration the *temporal* aspect of our data, which span across several years. We define the time-varying version of the socio-semantic networks introduced in Chapter 4, and present the evolution of their main characteristics over time.

# Chapter 3

# Modeling the textual content of scientific publications

**Contents**



In this chapter we focus on the main textual production of researchers, namely their scientific publications. These publications make it possible to study the circulation of "ideas" produced by scientific activity. To achieve this goal, we need to define an appropriate model of the texts. A possible way to do this is to try to identify the key terms characterizing their content. This surely constitutes a drastic reduction of the information embedded in texts, but we think this strategy is relevant since our goal is to identify the concepts investigated in each paper.

Advanced natural language processing methods are nowadays available to perform this kind of operation. In Section 3.1 we review the literature on

existing methods in automatic term recognition, and present our approach to perform this task on scientific publication archives (which generally provide at least titles and abstracts of the papers present in the collection). As a result, we create a list of terms characterizing the scientific field under study, which are used in the following chapters to reconstruct the semantic space of the scientific domain we are interested in, *i.e.* computational linguistics.

Texts can be exploited to extract more fine grained information that goes beyond keyword extraction. We would like for example to label terms with categories reflecting their information content, such as 'method' or 'task'. To this end, in Section 3.2 we present a new method that combines term extraction with text zoning so as to categorize terms depending on their context of use. The approach is based on the assumption that terms that repeatedly appear in the part of the abstract describing the methodology used are more likely to be terms describing methods or techniques. This may sound obvious but it should be noted that the method section (of the papers published in the ACL Anthology) also contains a wide variety of terms that need to be filtered. As a result, our goal is to produce a reduced list of terms containing only those describing methods and techniques: this will be especially useful to investigate the methodological evolution of the field under study.

## 3.1 Term extraction

To identify the terms representing the specific concepts of a scientific domain we could rely on a list made by an expert of the field, but, as already said, this would be biased by subjectivity and suffer from incompleteness. Alternatively, these concepts can be retrieved through an automatic analysis of the articles published in the field. Through natural language processing methods we can automatically retrieve sequences of words (n-grams) referring to terms that represent scientific concepts (Manning and Schütze, 1999).

We can define the notion of term more precisely thanks to two notions firstly introduced in (Kageura and Umino, 1996). The first notion is *unithood*, that is defined as the degree to which a phrase constitutes a semantic unit, i.e. a phrase consisting of words that are conventionally used together. We can also relate the notion of *unithood* to the notion of collocation, i.e. an expression of two or more words that co-occur more often than would be expected by chance (Manning and Schütze, 1999). The second notion is *termhood*, that is defined as the degree to which a semantic unit or a collocation represents a concept specific to a particular domain. Then, following (van Eck et al., 2010), we can define a *term* as a semantic unit with a high degree of *termhood*. (van Eck et al., 2010) also provide a good illustrative example of these notions:

> "To illustrate the notions of unithood and termhood, suppose that we are interested in statistical terms. Consider the phrases *many countries*, *United States*, and *probability density function*. Clearly, *United States* and *probability density function* are semantic units, while *many countries* is not. Hence, the *unithood* of *United States* and *probability density function* is high, while the *unithood* of *many countries* is low. Because *United States* does not represent a statistical concept, it has a low termhood. *probability density function*, on the other hand, does represent a statistical concept and therefore has a high termhood. From this it follows that *probability density function* is a statistical term."

In the next section we present an overview of the literature on automatic term extraction methods, which is based on (Pazienza et al., 2005; van Eck et al., 2010).

### 3.1.1 Literature overview

Several methods have been developed in the area of computational terminology to recognize and extract terms from texts in an automatic fashion, using both supervised and unsupervised techniques. In particular, two kinds of approaches can be distinguished: a linguistic approach and a statistical one. The former seeks to identify terms using pure linguistic filtering techniques, that is to say the identification of phrases that correspond to specific syntactic patterns. The latter aims at finding appropriate measures of the *unithood* and *thermhood* of phrases in order to identify terms by filtering the ones with high values of such measures. Finally, hybrid approaches combining linguistic and statistical features have been proposed. Usually in this case a linguistic analysis is used to identify the candidate n-grams and then statistical filters are applied to keep only the n-grams that correspond to real terms.

#### Linguistic approaches

Linguistic approaches rely on the hypothesis that most terms have the syntactic form of a noun phrase. This means an expression centered around a noun that may also contain adjectives, prepositions and possibly other nouns. Scientific terms usually have this kind of structure. Examples in computational linguistics are "syntactic structure of sentences" and "word sense disambiguation", in physics "renormalization-group theory" and "equations of state of nuclear matter"[1]. These approaches are based on filters that look for sequences of words that correspond to specified syntactic patterns, such as for example n-grams consisting of nouns only or nouns and adjectives only. Several syntactic patterns have been proposed in different studies, as for example in (Bourigault, 1992; Daille et al.,

[1]PACS codes 64.60.ae and 21.65.Mn, respectively.

1994). An extended study has been carried out in (Daille et al., 1996) to identify the most common syntactic structures of terms in English.

Linguistic methods for automatic term extraction generally adopt the following four steps:

**Part-of-speech tagging.** The first step consists in performing part-of-speech (PoS) tagging (Brill, 1994) in order to identify nouns, adjectives, prepositions, verbs and other parts of speech in the text.

**Linguistic filter.** A filter is then applied, usually using regular expressions (Kleene, 1951), to extract from the text all the n-grams corresponding to predefined patterns (such as an adjective followed by a noun, or a noun followed by a preposition and then by another noun).

**Variation identification.** This step aims at putting together n-grams that convey the same meaning and can therefore be considered as variations of the same term (Daille et al., 1996). Variation identification is thus done by assembling together n-grams that differ only by a stop-word (i.e. the too generic words of the language, such as for example "and", "this", "of") (Fox, 1989), in the number (singular or plural), or in the conjugation (e.g. presence or omission of the "-ing"). One of the most common types of variation is the use or not of the preposition *of*: for example "distribution of wealth" and "wealth distribution" are two variants of the same term. Another example of n-grams representing the same term is: "natural language dialogues", "natural language dialogue", "dialogues in natural language". However, it should be noted that the procedure is not specifically designed to detect synonyms: a thorough analysis would be needed to evaluate the approach, taking this limitation into account.

The described approach produces a list of candidate terms that requires a final step of validation, since it does contain irrelevant sequences of words that do not correspond to terms. Examples of n-grams in scientific texts that should finally be discarded are "proposed method" and "experimental

results". Pure linguistic approaches rely on human expert manual validation. A more widely used option is instead the use of statistical measures, through which one can compute the *thermhood* of the candidate terms and filter the list based on this score. In this case the linguistic approach is enriched with a statistical assessment.

### Statistical approaches

Pure statistical approaches aim at finding terms by the sole use of frequency count and other statistical measures, without taking into account the linguistic characterization of the words in the text. Statistical approaches are more powerful when used in combination with linguistic ones and therefore also take into account the syntactic-semantic features of words and their sequences. This leads to what are called hybrid approaches.

The most well known pure statistical approach is the TF-IDF measure (Salton et al., 1975). This method is based on a combination of word frequency and document specificity. The underlying idea is that relevant terms are n-grams that are frequent enough but at the same time specific of certain documents only. N-grams that appear in all documents are more likely to correspond to very generic phrases with low termhood. The tf-idf score of a candidate term is given by

$$\text{tf-idf}(t) = \text{tf}(t) \times \text{idf}(t, D) \tag{3.1}$$

where $\text{tf}(t)$ is simply the frequency of the term $t$ in the corpus, and $\text{idf}(t, D)$ is the inverse document frequency, defined as:

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{3.2}$$

where $N$ is the total number of documents in the corpus, and $D$ the set of these documents. $\{d \in D : t \in d\}$ represents therefore the number of documents in which the term $t$ appears. This second factor penalizes the terms that appear in a large fraction of documents: the higher the number of documents a term appears in, the closer to 1 is the ratio inside the logarithm,

which lowers the idf and brings the tf-idf closer to 0.

The simplest approach to measure *unithood* is the frequency of occurrences (Justeson and Katz, 1995). More advances approaches are based on measures of mutual information (Church and Hanks, 1990a), likelihood ratio (Dunning, 1993), or the more recent C-value (Frantzi and Ananiadou, 2000), whose definition will be detailed in the next section.

*Termhood* can be measured based on co-occurrence distributions (Matsuo and Ishizuka, 2004; van Eck et al., 2010), which will also be described in more detail in the next section. Other proposed approaches worth citing are the NC-value (Frantzi and Ananiadou, 2000) and the SNC-value (Maynard and Ananiadou, 2000), which are extensions of the C-value by integrating into the measure the notion of *thermhood* in addition to *unithood*.

### 3.1.2 Term extraction from the ACL Anthology corpus

We now describe the procedure we followed to automatically extract n-grams corresponding to scientific concepts in computational linguistics from the titles and abstract of the papers in the ACL Anthology.

Concerning the APS corpus, related to physics, we do not need to perform any automatic term extraction since all the papers are already associated with some keywords (called PACS codes)[2], as explained in Chapter 2.2.

The technique we use follows the procedure introduced and implemented by researchers working on the CorTexT platform[3], and relies on a hybrid linguistic and statistical approach.

We first pre-process the text with PoS tagging, using the dedicated NLTK module (Bird et al., 2009). Then we build the set of possible terms by extracting all the n-grams that match the syntactic patterns defined by the following regular expressions:

- `<JJ.*>*<NN.*|>+`

[2] Also abstracts are not provided in this corpus, therefore the task would not be feasible.
[3] `http://docs.cortext.net/lexical-extraction/`

- `<JJ.*>*<NN.*|>+<CC>*<NN.*|>*`
- `<JJ.*>*<NN.*|>+<IN>?<PRP|DT>?<JJ.*>*<NN.*|>+`
- `<JJ.*>*<VBN>*<VBG>*<NN.*|>+`
- `<JJ.*>*<VBN>*<VBG>*<NN.*|>+<CC>*<NN.*|>*`
- `<JJ.*>*<VBN>*<VBG>*<NN.*|>+<IN>?<PRP|DT>?<JJ.*>*<VBN>*`

  `<VBG>*<NN.*|>+`
- `<JJ.*>*<VBN>*<VBG>*<NN.*|>+<IN>?<PRP|DT>?<JJ.*>*<VBN>*`

  `<VBG>*<NN.*|>+<IN>?<PRP|DT>?<JJ.*>*<VBN>*<VBG>*<NN.*|>+`

where JJ indicates an adjective, VBN the part participle of a verb, VBG the gerund or present participle of a verb, NN a noun, IN a preposition or subordinating conjunction, PRP a personal pronoun, and DT a determiner (Santorini, 1990).

The linguistic process then has to provide normalization and stemming, from which we can build classes of equivalent candidate terms. Normalization means removing capitalization differences and correcting spelling differences between n-grams, usually arising from the presence or absence of hyphens. For example we consider that “multi-word expression” and “multiword expression” belong to the same class. Stemming is performed to automatically put in the same class the n-grams that share the same stems, like for example the singular and the plural version of the same term. Moreover, stop-words are removed and the remaining words are ranged alphabetically. N-grams becoming identical after this operation are put in the same class, like for example “information extraction” and “extraction of information”.

At the end of this process we obtain a list of candidate terms, grouped in equivalence classes. We then filter this list using statistical methods that measure the *unithood* and *termhood* of each n-gram. This second processing phase consists of four steps:

- Firstly, we enumerate every n-gram in the list throughout the whole corpus, to obtain their frequency. If two candidate terms are nested, every time we find the larger n-gram we increment only its frequency and not the one of the other one. For example if "syntax-based machine translation" is found in an abstract, we increment the frequency of the n-gram "syntax-based machine translation", but not the one of the smaller n-gram "machine translation".

- Secondly, we measure *unithood* using the C-value method proposed in (Frantzi and Ananiadou, 2000). We define *unithood* as

$$u(i) = \log(n+1) f_i \tag{3.3}$$

  where n is the number of words that constitute the n-gram $i$, and $f_i$ the n-gram frequency in the corpus. This measure is proportional to the number of occurrences of the term, since this is a natural measure of the stability of a phrase. Concurrently, the first factor favors longer n-grams, since they carry more information and are therefore more likely to be real terms.

- The list of candidate terms is then sorted according to the *unithood* of the selected sequences of text and the list is pruned to the n-grams with the highest C-value. This step removes less frequent n-grams, and more importantly it filters the list before the final step.

- The last step consists in measuring the *termhood* of the candidate terms in order to obtain a final list of the $N$ n-grams with the higher *unithood* and *termhood*. To do this we adopt a similar approach to that proposed in (van Eck et al., 2010). Low *termhood* word chains are n-grams that do not help characterizing the content of the text, although they might still occur very frequently, and thus have a high value of *unithood*. Examples of such n-grams are "past articles" and "experimental results". The rationale behind the method we use is that such irrelevant n-grams appear in any paper in the corpus, whereas the real terms appear only in a subset of them. Therefore irrelevant n-grams should have an unbiased distribution compared to the other terms in the list. We compute the co-occurrence matrix $M$ between each item in the list of candidate terms, i.e. the matrix whose

entry $(i, j)$ corresponds to the number of times the candidate term $i$ is found in the same title or in the same abstract together with the candidate term $j$. We then define the *termhood* of the candidate term $i$ as:

$$\theta(i) = \sum_{j \neq i} \frac{(M_{ij} - M_i M_j)^2}{M_i M_j} \tag{3.4}$$

where $M_i = \sum_j M_{ij}$. A high value of this sum means that the term co-occurs more frequently with only a subset of the other terms in the list and is therefore more specific.

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 rare in the dataset as a whole), we divide the papers set in three even time slices, and extract a list of 1000 terms from each subset, following the procedure just described. Some more fine grained divides are of course possible, but we consider this periodization reasonable enough. We then merge the three lists (since as expected the three lists have some terms in common) and eliminate all the terms that appear in less than 5 papers, obtaining a list of about 1500 terms. The choice $N = 1000$ has been made because this number is a high enough to capture all the most salient terms of this very specific field of science, but at the same time it is not too high with respect to computational time (the co-occurrence matrix $M$ is of order $O(N^2)$). The choice is also justified *a posteriori* by the fact that in the obtained list (before eliminating the terms with low frequency) there are also n-grams with frequency lower than 5, which we consider a reasonable lower bound for a candidate n-gram to represent a field specific term.

The final step is to show the result to an expert in the field who validates the list of terms manually, eliminating all terms that are 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. Manual validation is a sensitive issue because it introduces subjectivity in the approach. However fully automatic methods always produce noisy results, and manual validation injects additional knowledge in the system, which is

necessary to filter out all the unwanted n-grams.

The complete list of terms is reported in Appendix A. This list contains scientific terms specific of the domain of computational linguistics. Terms can be of different nature though: they can represent concepts, tasks, or methods for example. Being able to distinguish among these different classes of terms is fundamental in order to investigate fine-grained phenomena in the evolution of scientific fields. To achieve this, in the next section we propose a new method that combines standard automatic term recognition with argumentative text zoning. This new approach relies on the idea that we can categorize terms using the information carried by the context in which they are used.

## 3.2 Term categorization

Argumentative text zoning is the automatic analysis of the argumentative structure of scientific papers. It categorizes sentences of scientific papers according to different labels, such as aims, methods, and results. We make the hypothesis that we can gain additional information about terms extracted from scientific papers by looking at their location in the abstract, namely in which sentence category (as given by the text zoning analysis) they usually appear in. This hypothesis is based on the assumption that abstracts have, for the most part, a common structure: a few sentences providing an introduction to the field and the background of the work, one or two sentences stating the objective, then a brief description of the methodology used, and finally the results. A term is for example more likely to refer to a concept if it often appears at the beginning of the abstract, where authors state what is the objective of the paper. For the same reason, a term is more likely to refer to a method if it appears in the part of the abstract where the methods used are explained. Building on this intuition, we annotate the ACL Anthology with a text zoning analyzer and use the results to categorize the related terms.

### 3.2.1 Literature overview on text zoning

The first important contributions to text zoning are probably the experiments by S. Teufel who proposed to categorize sentences in scientific papers (and more specifically, in the Natural Language Processing domain) according to different categories (Teufel, 1999) like BKG: 'General scientific background', AIM: 'Statements of the particular aim of the current paper', or CTR: 'Contrastive or comparative statements about other work'. The task is called rhetorical zoning or argumentative zoning since the goal is to identify the rhetoric or argumentative role of each sentence of the text.

The initial work of Teufel was based on the manual annotation of 80 papers representing the different areas of NLP (the corpus was made of papers published within the ACL conferences and the journal Computational Linguistics). A classifier was then trained on this manually annotated corpus. The author reports interesting results despite "a 20% diference

between [the] system and human performance" (Teufel and Moens, 2002). The learning method uses a Naive Bayesian model since more sophisticated methods tested by the author did not obtain better results. Teufel in subsequent publications shows that the technique can be used to produce high quality summaries (Teufel and Moens, 2002) or precisely characterize the different citations in a paper (Ritchie et al., 2008).

The seminal work of Teufel has since then given rise to different kinds of works, on the one hand some researchers try to refine the annotation method, while on the other hand they check its applicability to different scientific domains. Concerning the first point, research has focused on the identification of relevant features for classification, on the evaluation of different learning algorithms for the task and more importantly on the reduction of the volume of text to be annotated. Concerning the second point, biological and bio-medical domains have attracted much of the attention, since scientists in these domains often have to access the literature "vertically" (i.e. experts may need to have access to all the methods and protocols that have been used in a specific domain) (Mizuta et al., 2006; Tbahriti et al., 2006).

Guo has developed a similar trend of research to extend the initial work of Teufel (Guo et al., 2011, 2013): she has tested a large list of features to analyze the zones, evaluated different learning algorithms for the task and proposed new methods to decrease the number of texts to be manually annotated. The features used for learning belong to three categories:

*i*) positional: location of the sentence inside the paper

*ii*) lexical: words, classes of words, bigrams, etc. are taken into consideration

*iii*) syntactic: the different syntactic relations as well as the class of words appearing in subject or object positions are also considered.

The analysis is thus based on more features than in Teufel's initial work and requires a specific parser.

### 3.2.2 Text zoning analysis of the ACL Anthology corpus

The method developed by Guo and colleagues (Guo et al., 2011) seems particularly well suited to our problem. We want to categorize terms used in the ACL Anthology so as to identify the main methods used in the domain and study their evolution over time.

In our experiment, we only use the abstracts of the papers. Our hypothesis is that abstracts contain enough information and are redundant enough to study the evolution of the domain. Taking into consideration the full text would probably give too many details and thus introduce noise in the analysis.

The annotation scheme includes five different categories, which are the following: OBJECTIVE (objectives of the paper), METHOD (methods used in the paper), RESULTS (main results), CONCLUSION (conclusion of the paper), BACKGROUND (general context), as in (Reichart and Korhonen, 2012). These categories are also close to those of (Mizuta et al., 2006; Guo et al., 2011, 2013) and have been adapted to abstracts (as opposed to full text[4]). It seems relevant to take into consideration an annotation scheme that has already been used by various authors so that the results are easy to assess.

Around one hundred abstracts from the ACL Anthology have then been manually annotated using this scheme (∼500 sentences; ACL abstracts are generally quite short since most of them are related to conference papers). The selection of the abstracts has been done using stratified sampling over time and journals, so as to obtain a representative corpus (papers must be related to different periods of time and different sub-areas of the domain). The annotation has been done according to the annotation guideline defined by Guo, especially for long sentences when more than one category could be applied (preferences are defined to solve complex cases[5]).

[4]The categories used in (Teufel, 1999) were not relevant since this model focused on full text papers, with a special emphasis on the novelty of the author's work and the attitude towards other people's work, which is not the case here.

[5]The task is to assign a single category to each sentence. The choice of the category should be made according to the following priority list: Conclusion > Objective > Result > Method > Background.

The algorithm defined by (Guo et al., 2011) is then adapted to our corpus. The analysis is based on positional, lexical and syntactic features, as explained above. No domain specific information is added, which makes the whole process easy to reproduce. As for parsing, we use the C&C parser (Curran et al., 2007). All the implementation details can be found in (Guo et al., 2011), especially concerning annotation and the learning algorithm. As a result, each sentence is associated with a tag corresponding to one of the zones defined in the annotation scheme.

**Results and discussion**

In order to evaluate the text zoning task, a number of abstracts are chosen randomly (∼300 sentences that do not overlap with the training set). CONCLUSION represents less than 3% of the sentences and is therefore dropped (together with the sentences therein classified) for the rest of the analysis. The four remaining zones are unequally represented: 18.05 % of the sentences refer to BACKGROUND, 14.35% to OBJECTIVE, 14.81 % to RESULT and 52.77 % to METHOD. Just by looking at these numbers, one can see how methodological issues are important for the domain.

We then calculate for each of the categories the percentage of sentences that received the right label (as assessed by an expert), which allows us to calculate precision. The results are given in Table 3.1. These results are similar to the state of the art (Guo et al., 2011), which is positive taking into consideration the small number of annotated sentences used for training. The diversity of the features used for learning makes it easy to transfer the technique from one domain to another without any new heavy annotation phase. Results are slightly worse for the METHOD category, probably because this category is more diverse and thus more difficult to recognize. The fact that NLP terms can refer either to objectives or to methods also contributes to the fuzziness of this category (most NLP systems are made of different layers and require various NLP techniques; for example, a semantic analyzer may use a part-of-speech tagger and a parser, which means NLP tools can appear as part of the method).

Figure 3.1 shows an abstract annotated by the text zoning module.

Table 3.1: Text zoning analysis results (precision).

| Category | Precision |
|---|---|
| Objective | 83,87 % |
| Background | 81,25 % |
| Method | 71,05 % |
| Results | 82,05 % |

Figure 3.1: An abstract annotated with text zoning information. Categories are indicated in bold face. The paper is (Lee et al., 2002): it has been chosen randomly between those containing the different types of zones.

Most of errors in Korean morphological analysis and POS (Part-of-Speech) tagging are caused by unknown morphemes. **BACKGROUND**
This paper presents a generalized unknown morpheme handling method with POSTAG (POStech TAGger) which is a statistical/rule based hybrid POS tagging system. **OBJECTIVE**
The generalized unknown morpheme guessing is based on a combination of a morpheme pattern dictionary which encodes general lexical patterns of Korean morphemes with a posteriori syllable tri-gram estimation. **METHOD**
The syllable tri-grams help to calculate lexical probabilities of the unknown morphemes and are utilized to search the best tagging result. **METHOD**
In our scheme , we can guess the POS's of unknown morphemes regardless of their numbers and positions in an eojeol , which was not possible before in Korean tagging systems. **RESULTS**
In a series of experiments using three different domain corpora , we can achieve 97% tagging accuracy regardless of many unknown morphemes in test corpora. **RESULTS**

### 3.2.3 Term categorization in the ACL Anthology corpus

Our ultimate goal is to identify the terms referring specifically to methodological issues (*e.g* different machine learning techniques). From this perspective, terms appearing in the METHOD sentences are thus particularly interesting for us.

Here, we voluntarily use a minimal approach for term extraction and filtering since we want to keep most of the information for the subsequent text zoning phase. We thus perform only the linguistic filtering described in Section 3.1.2, but do not apply the statistical filters. Only the noun phrases appearing in more than 10 papers are kept for subsequent processing.

For each term in the list, we enumerate the number of sentences of each zone category it appears in. Terms are then ranked per zone, according to their degree of specificity (the zone they are the most specific of). We use the Kolmogorov-Smirnov (KS) test to measure the specificity. The KS test computes the distance between the empirical distribution functions of two samples. It is calculated as follows (Press et al., 2007):

$$D = \max_x |S_{N_1}(x) - S_{N_2}(x)| \tag{3.5}$$

where $S_{N_1}(x)$ et $S_{N_2}(x)$ are the empirical distribution function of the two samples (that correspond in our case to the number of occurrences of the term in a given zone, and to the total number of occurrences of all the terms in the same zone, respectively). A high value of $D$ for a given term means that it is specific of the considered zone. At the opposite, a low value means that the term is spread over the different zones and not specific to any zone.

Finally, an expert of the domain manually examined and filtered the top 150 specific terms in the METHOD category, and divided them into clusters corresponding to the different kinds of methods used in computational linguistics. Methods were also grouped by the expert into broader categories that will help us explore the methodological evolution of the field over time. The results are shown in Table 3.2. Logically, given our approach, the table does not contain all the terms relevant for the computational linguistics domain, but it contains the most specific ones according to the above approach. One should thus not be surprised not to see all the terms used in the domain.

Table 3.2: Most specific terms found in the METHOD sentences.

| Methods | | |
|---|---|---|
| Category | Method | N-grams |
| Machine learning | Bayesian methods | baesyan |
| | Vector Space model | space model, vector space, cosine |
| | Genetic algorithms | genetic algorithms |
| | HMM | hidden markov models, markov model |
| | CRF | conditional random fields |
| | SVM | support vector machines |
| | MaxEnt | maximum entropy model, maximum entropy approach, maximum entropy |
| | Clustering | clustering algorithm, clustering method, word clusters, classification problem |
| Speech & Machine Trans. | Language models | large-vocabulary, n-gram language model, Viterbi |
| | Parallel Corpora | parallel corpus, bilingual corpus, phrase pairs, source and target languages, sentence pairs, word pairs, source sentence |
| | Alignment | phrase alignment, alignment algorithm, alignment models, ibm model, phrase translation, translation candidates, sentence alignment |
| NLP Methods | POS tagging | part-of-speech tagger, part-of-speech tags |
| | Morphology | two-level morphology, morphological analyzer, morphological rules |
| | FST | finite-state transducers, regular expressions, state automata, rule-based approach |
| | Syntax | syntactic categories, syntactic patterns, extraction patterns |
| | Dependency parsing | dependency parser, dependency graphs, prague dependency, dependency treebank, derivation trees, parse trees |
| | Parsing | grammar rules, parser output, parsing process, parsed sentences, transfer rules |
| | Semantics | logical forms, inference rules, generative lexicon, lexical rules, lexico-syntactic, predicate argument |
| Applications | IE and IR | entity recognition, answer candidates, temporal information, web search, query expansion, google, user queries, terms, query terms, term recognition |
| | Discourse | generation component, dialogue acts, centering theory, lexical chains, resolution algorithm, generation process, discourse model, lexical choice |
| | Segmentation | machine transliteration, phonological rules, segmentation algorithm, word boundaries |
| Words and Resource | Lexical knowledge bases | lexical knowledge base, semantic network, machine readable dictionaries, eurowordnet, lexical entries, dictionary entries, lexical units, representation structures, lookup |
| | Word similarity | word associations, mutual information, semantic relationships, word similarity, semantic similarity, semeval-2007, word co-occurrence, synonymy |
| | Corpora | brown corpus, dialogue corpus, annotation scheme, tagged corpus |
| Evaluation | Evaluation | score, gold standard, evaluation measures, estimation method |
| Calculation & complexity | Software | tool development, polynomial time, software tools, series of experiments, system architecture, runtime, programming language |
| | Constraints | relaxation, constraint satisfaction, semantic constraints |

## 3.3 Conclusions

In this chapter we have presented the method that we followed to automatically extract the relevant terms from texts of scientific publications. Our method is based on the analysis of the titles and abstracts of the publications of the field under study, and relies on a hybrid linguistic and statistical approach. We have introduced a novel strategy for the categorization of scientific terms based on the analysis of the argumentative structure of the abstracts they appear in. We have successfully applied the proposed methods to the ACL Anthology, so as to automatically extract a list of terms related to the field of computational linguistics. We are also able to identify the terms that refer to the techniques used in this domain. This simply follows the logical hypothesis that these terms are the ones that appear specifically in the methodology section of abstracts.

# Chapter 4

# Modeling scientific research as a socio-semantic network

**Contents**

Our goal is to investigate scientific collaboration and knowledge production, therefore the *relations* between researchers and between concepts are fundamental to our study. A suitable framework for this kind of analysis is *network theory*, namely the science of representing complex systems as graphs illustrating relations between various objects. In this chapter we present how the structure of scientific collaborations and the structure of the knowledge produced in a given field can be modeled as networks. This sets the basis for the investigation of the dynamics of scientific research in a given field that is carried out in Part III.

This chapter is structured as follows. In Section 4.1 we present an introduction to network theory, mainly based on (Barrat et al., 2008; Newman, 2010). This is not meant to be a comprehensive review, but only an outline of the concepts that are used throughout this thesis. In Section 4.2 we give an overview on the applications of network theory to the study of scientific production. Then in Section 4.3 we define precisely the empirical social and semantic networks that we study, and how they can be built from the data. Lastly, in Section 4.4 we present the main features of these networks and discuss their interpretation.

## 4.1 An introduction to network theory

A *network*, also called *graph* in mathematics, is a collection of nodes connected by links.

Nodes and links can also be called actors and ties in social network analysis – a sociological field aiming at formalizing social systems as graphs to understand their functioning. In this context actors are defined as "discrete individual, corporate, or collective social units", and the defining feature of a relational tie is that "it establishes a linkage between a pair of actors". Having defined actors and relations, "a social network consists of a finite set or sets of actors and the relation or relations defined on them" (Wasserman and Faust, 1994).

In particular, a graph representing a system that is self-organized in a growing structure that exhibits non-trivial topological features, is called a *complex network* (Barrat et al., 2008). Networks representing real-world systems are often of this kind. Complex network theory is an active area of interdisciplinary scientific research carried on mainly by computer scientists, statistical physicists, and mathematicians, but with applications in a variety of other disciplines, such as biology, sociology and economics.

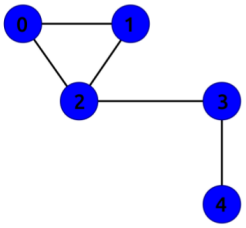
### 4.1.1 Mathematical foundations

Mathematically, a graph is an ordered pair $G = (N, E)$ comprising a set $N$ of nodes, together with a set $E$ of links connecting pairs of elements of $N$. The most popular mathematical representation of a network is the adjacency matrix. The adjacency matrix $A$ of a graph is the matrix with elements $A_{ij}$ such that:

$$A_{ij} = \begin{cases} 1 \text{ if there is a link between } i \text{ and } j \\ 0 \text{ otherwise.} \end{cases} \tag{4.1}$$

The nodes connected to a given node $i$ are often referred to as its *neighbors*, and the set comprising them is called the node *neighborhood*.

A network can be *directed*, meaning that each link points from one node to




$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Figure 4.1: An example of simple undirected network (left), and the corresponding adjacency matrix (right).

another, and relations need not to be reciprocated. Moreover, some networks can allow multiple links between pairs of nodes, and links to a node to itself. The networks we study in this thesis do not have such characteristics, therefore from now on, when not specified otherwise, the networks we refer to will be undirected, with single links, and no self-loops.

A very simple example of network and the corresponding adjacency matrix is given in Figure 4.1.

### 4.1.2 Weighted networks

In some cases, it is useful to represent links in association with some numerical information. This information, which is usually called *weight*, measures the intensity or the capacity of the interaction between two nodes (Barrat et al., 2004).

In social networks weights can for example indicate the frequency of contact between actors, whereas in transportation networks they can represent the amount of traffic between two locations.

Networks characterized by these additional values are called *weighted networks*. Their mathematical representation is the same as for the unweighted case, with the only difference that the adjacency matrix does not contain only 0 and 1 anymore, but for each link connecting $i$ and $j$ the corresponding entry of the adjacency matrix is equal to the value of the corresponding weight $w_{ij}$.

### 4.1.3 Bipartite networks

There is a particular type of network that will be used in the next sections of the chapter, namely the *bipartite network*, also called two-mode network in sociology. In this type of network there are two kinds of nodes, and links only connect nodes of different kinds.

A well-known example is the movie actor network, based on the Internet Movie Database (Watts and Strogatz, 1998). In this network one kind of nodes refers to actors, and the other to movies. Each actor is connected by a link to each movie he or she appeared in.

Although bipartite networks provide the closest to reality representation, it is often convenient to consider networks containing only one kind of nodes and having connections between them directly. We can achieve this by creating a one-mode projection of the original bipartite network.

For example we can create the one-mode projection of the movie actor network by constructing a network in which the nodes represent the actors and which features links between two actors when they played in the same movie. To retain some of the information we loose when performing this projection, we can also build a weighted network in which the weight of the link connecting two actors has a value proportional to the number of movies they co-starred.

An example of bipartite network and its two one-mode projections is shown in Figure 4.2.

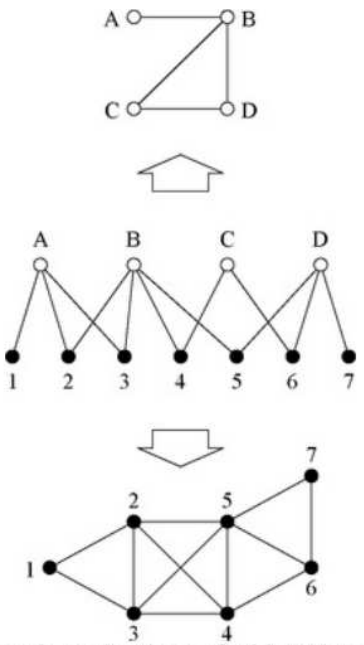



**Figure 6.5: The two one-mode projections of a bipartite network.** The central portion of this figure shows a bipartite network with four vertices of one type (open circles labeled A to D) and seven of another (filled circles, 1 to 7). At the top and bottom we show the one-mode projections of the network onto the two sets of vertices.

Figure 4.2: An example of bipartite network and its two one-mode projections (Newman, 2010).

### 4.1.4 Network characterization

We now present the main measures used to characterize the topology and the properties of a network.

#### Degree and strength

Firstly, we introduce a measure that characterizes the importance of nodes in a network. Various measures have been defined, among which the simplest and most commonly used is the *degree centrality*.

The degree $k_i$ of a node $i$ is defined as the number of links in the network incident on the node $i$. This measure has a straightforward interpretation: it quantifies how well a node is connected to other nodes in the network.

For weighted networks, we can define the *strength* $s_i$ of a node $i$ as the sum of the weights attached to the links incident on node $i$ (Barrat et al., 2004). The strength of a node can be considered as a natural generalization of the degree since it integrates, on top of the number of connections, information about their importance.

The degree (and the strength) not only gives us information about single nodes, but can be used to gain important insights about the structure of the whole network. One of the fundamental properties of most real-world complex networks lies in the heterogeneous distribution of node degrees, which means that within the same network different nodes can have very different degree values. We define the *degree distribution* $P(k)$ of a given network as the fraction of nodes in the network having degree $k$. We can interpret this function also as a probability distribution: it gives the probability that a randomly chosen node in the network has degree $k$.

For details about other measures of centrality such as the closeness or the betweenness centrality, we refer the reader to (Newman, 2010).

### Paths

Another important issue in the structure of networks is the reachability of nodes. Can we get from one node to another following the connections given by the links? And if we do, can we identify the shortest way to do so? A network in which every node is reachable from any other node is called connected. Many real-world networks are not connected, but consist of more than one connected component, each being composed of nodes reachable by all the other nodes in the component. In many cases the analysis of empirical networks is restricted to the giant component, which is the largest connected component of the network.

While networks usually lack a metric, a natural measure of distance between two nodes is defined as the number of links traversed by the shortest path $l_{ij}$ connecting them. This metric can be used to define the linear size of a network by introducing the *average shortest path length*, which is the average number of steps along the shortest path between all the possible pairs of nodes in the network:

$$l = \frac{1}{N(N-1)} \sum_{i,j} l_{ij} \ . \tag{4.2}$$

### Clustering

The last of the three most robust and important measures of the structure of a network is *clustering*, also called *transitivity* in sociology (Wasserman and Faust, 1994). This concept refers to the tendency observed in many real-world networks that, if a node $i$ is connected to node $j$, and at the same time node $j$ is connected to node $h$, then with a high probability $i$ is also connected to $h$. This property can be quantitatively measured by means of the *clustering coefficient*. In particular, we can define the local clustering coefficient of node $i$ as:

$$C_i = \frac{\text{number of couples of neighbors of } i \text{ that are connected}}{\text{number of couples of neighbors of } i}. \tag{4.3}$$

Then the average clustering coefficient of a network is simply given by

$$C = \frac{1}{N} \sum_i C_i. \tag{4.4}$$

**Community structure**

The last measure we present is *modularity*, which has been introduced in (Newman and Girvan, 2004) to characterize the structure of networks in terms of emerging communities, *i.e.* groups of nodes forming dense sub-graphs with few inter-group links. This is a common property of many real-world networks. Social networks may include communities based on common location or interests for example. Metabolic networks are characterized by communities based on functional groupings.

Several methods and algorithms have been proposed to detect community structure in networks, among which we recall the stochastic blockmodel (Holland et al., 1983), the Girvan and Newman algorithm (Newman and Girvan, 2004), based on modularity optimization, clique percolation (Palla et al., 2005), Infomap (Rosvall and Bergstrom, 2008), Louvain (Blondel et al., 2008), and OSLOM (Lancichinetti et al., 2011). For a comprehensive review see (Fortunato, 2010).

Given a network and a partition of its nodes into some communities, the modularity $Q$ reflects the concentration of links within communities compared with a random distribution of links between all nodes regardless of communities. It is defined as

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j) \tag{4.5}$$

where $m$ is the number of links in the network, $\delta(c_i, c_j)$ is equal to one if $i$ and $j$ belong to the same community and zero otherwise, and $\frac{k_i k_j}{2m}$ represents the probability that $i$ and $j$ are connected regardless of the community structure of the network. Modularity is strictly less than one and it takes positive values if there are more links between nodes belonging to the same community than it would be expected by chance, and negative values if there are less.

### 4.1.5 Real-world network characteristics

The informatics revolution of recent years has made it possible to gather and analyze data sets on several large-scale networks. Co-authorship and co-words networks, which constitute the focus of this dissertations, are examples of real-world networks built from data. Other examples are: metabolic networks (Jeong et al., 2000), protein interaction networks (Barabási and Oltvai, 2004), the Internet (Pastor-Satorras and Vespignani, 2007), the World Wide Web (Albert et al., 1999), social networks (Wasserman and Faust, 1994), transportation networks (Guimerà et al., 2005a).

These data have many fundamental differences, starting from the nature of the elements that compose them, but their large size makes it possible to characterize their structural and functional properties in statistical terms. This allowed network scientists to uncover common properties and patterns that could lead to a classification of empirical networks.

Let us first consider the distance among nodes in a network, that we can measure through the average shortest path length, as we have seen in the previous section. Most empirical network exhibit the so-called "small-world phenomenon", which means that it is possible to go from one node in the network to any other one through a very small number of intermediate nodes. In more precise mathematical terms, the small-world property refers to networks in which the average shortest path length scales logarithmically with or more slowly than the number of nodes in the network. This property is known in sociology as the "six degrees of separation" phenomenon. In 1967 Milgram performed an experiment through which he showed that six acquaintances is on average enough to connect any two randomly chosen people in the United States (Milgram, 1967).

At first sight this appears to be a very peculiar feature, but it can actually be explained by the presence of randomness only. Let us consider a simple model of network in which the presence of a link between two nodes is a random event occurring with the same probability for any couple of nodes. This is called a random network. (Bollobás, 1981) rigorously demonstrates that in this kind of network the average shortest path length

approximately scales with the logarithm of the number nodes. This explains the omnipresence of this property in real-world networks, since in all natural systems the presence of some level of "shortcuts" dramatically diminishes the diameter of a network.

A more interesting characteristic of many real-world networks is that the small-world effect goes along with a high level of clustering (Watts and Strogatz, 1998). This feature cannot be explained by the sole presence of randomness, and in fact in large random networks the clustering coefficient is very small, namely of the order of magnitude of the inverse of the number of nodes in the network. Regular grid networks (*i.e.* graphs forming a grid-like structure) tend to be clustered but they are not small-words, whereas random networks are small-world but do not feature high clustering. The famous small-world network model of Watts and Strogatz addresses this problem and produces networks that are characterized by both features at the same time.

Another evidence for the presence of some structural organization in real-world networks is given by the statistical analysis of the degree centrality. The functional form of the degree distribution of large-scale networks defines two broad network classes: statistically homogeneous and heterogeneous networks. The former refers to networks with degree distributions whose form can be approximated by Poisson distributions, or more generally by fast decaying tails. The latter concerns networks whose degree distribution is, on the contrary, skewed and heavy-tailed. This is the case for many real-world networks, in which most nodes have only a few connections, and a few nodes are linked to hundreds or sometimes even thousands of nodes (these nodes are called "hubs"). This feature is easily seen for example in the airport network, where a few airports attract most connections, and in the World Wide Web, where some websites are very popular and receive a large number of hyperlinks, whereas most pages are hardly linked. This feature often leads to a degree distribution that can be approximated by a power-law distribution $P(k) \sim k^{-\gamma}$, which results in a linear shape on a log-log scale (Barabási and Albert, 1999). We define the networks belonging to this second class scale-free, because the average degree is not meaningful and therefore does not define a characteristic scale for the network. Barabasi and Albert developed the famous model of

growing network that is able to reproduce this kind of degree distribution. The model is based on the so-called "preferential attachment" mechanism, which is founded on the fact that incoming nodes create connections with higher probability with already highly connected nodes. This is also called the "rich-get-richer phenomenon", which had been formerly introduced by (Price, 1965), among others.

Lastly, another common characteristic of many real-world networks is community structure. As seen in the previous section this refers to the emergence of groups of nodes that are more densely connected to each other than with the rest of the nodes in the network, leading to an emerging "natural" partition of the network. Empirical analysis of real-world networks have shown that optimal modularity values of network characterized by community structure typically fall in the range from about 0.3 to 0.7, and higher values are rare (Newman and Girvan, 2004).

As a final note, it is worth noting that the Barabasi-Albert model produces networks with scale-free degree distributions, but fails to produce the high level of clustering typical of real-world network. At the same time, the Watts and Strogatz model, which successfully produces high level of clustering, produces homogeneous degree distributions of Poisson type, and not heterogeneous ones typical of many real-world networks too. Lastly, neither of these models produces networks with a well defined community structure.

## 4.2 Modeling scientific production in different fields with networks

The representation and modeling of scientific production as complex networks began with the 1965 study by Price on the network of citations between academic papers (Price, 1965). In his model nodes correspond to scientific articles and an article is linked to another one if the former includes a citation of the latter. The most cited papers tend to attract most of the citations in a given domain (the "rich get richer" effect).

Since then, a large amount of studies have analyzed and modeled science using this framework. Most of this work has been developed in the last decades though, because it relies on the recent availability of large databases of bibliometric data in a digitalized form, that make it possible to perform significant quantitative analysis. In this context, three main types of networks have been built and analyzed: citation, co-authorship and co-word networks. In this thesis we focus on scientific collaboration and knowledge production, therefore we study networks of the last two types. In the rest of the section we give an overview of the main works in the network literature that focus on these kinds of networks. A review of the works focusing on the evolution of such networks is presented in Chapter 6.

### 4.2.1 Co-authorship networks

A co-authorship network is a graph connecting researchers that co-authored one or more paper together. It is therefore a representation that documents scientific collaborations between authors, and the structure of this network reveals interesting features of scientific communities. As already discussed in Chapter 1, in the last few decades a number of studies have reconstructed large-scale co-authorship networks representing scientific communities in different fields.

In particular it has been shown that the distribution of the number of collaborators per author is heterogeneous, with most researchers collaborating with only a few other authors, and a few having hundreds or

in some cases even thousands of collaborators. Moreover, these networks have a very small average distance between nodes and a high clustering coefficient, and can therefore be considered small-world. Lastly, they also appear to have a well defined community structure: researchers tend to form communities working on the same research topic or methodology (Newman, 2004).

### 4.2.2 Co-word networks

A complementary way to uncover the structure of scientific fields is to focus on their prominent concepts. This can be done by shifting the focus of the analysis on the texts and the words therein. In Chapter 3 we explained how we can automatically extract relevant terms corresponding to scientific concepts from scientific papers. These techniques can be coupled with network theory to build co-word networks characterizing a given scientific field. A co-word network is indeed a graph connecting words or terms that occur together in a text. These networks are meant to capture the relationships between ideas and unveil the structure of scientific knowledge (Callon et al., 1986).

Co-word network analysis has mainly been used as a tool for mapping and visualizing scientific production. We focus on this issue in Chapter 7. Here, we build and analyze the co-word networks of the two fields under study (computational linguistics and physics) in the exact same way in which we analyze the corresponding co-authorship networks, namely focusing on the characteristics of their structure, rather than on their visualization.

## 4.3 Social and semantic network definition

From each of the scientific archives we want to study, we can build complex networks that represent the structure of the connections between researchers, between scientific concepts, and between the former and the latter too. These networks are defined as follows.

Let us first consider all the papers in a given dataset, and, for each paper, its authors. We can build a bipartite network (defined in 4.1.3) in which papers on the one hand and researchers on the other constitute two different classes of nodes. A link is created between a paper and a researcher if she/he is the author or one of the authors of the given paper. It is then possible to build the one-mode projection of this bipartite network on the author set. The resulting graph constitutes a co-authorship network, and in the rest of the thesis we will refer to it as the *social network*. Its mathematical definition is the following.

> **Social network.** $G^{soc} = (A, E^{soc})$ is the undirected weighted graph in which $A$ is the set of authors in the considered field, and $E^{soc}$ is the set of links connecting elements in $A$. Two authors $(a_i, a_j)$ are connected in $G^{soc}$ if they co-authored a paper belonging to the dataset, and the weight $w(a_i, a_j)$ of the corresponding link has a value equal to the number of papers they co-authored.

Then we consider the semantic information pertaining to the content of the papers in the given corpus. In the case of the APS dataset, we use the PACS codes designated by the authors of the paper as the concepts characterizing it, as explained in Section 2.2. In the case of the ACL dataset, we use the list of terms obtained in Chapter 3. In this context we are interested in all types of terms, and not only in the methods, so we use the complete list created as described in Section 3.1.2 and reported in Appendix A. For sake of simplicity, from now on we will refer to PACS codes and technical terms as *concepts*.

Let us consider the concepts characterizing a given field, and the papers in the corresponding collection. We then build another bipartite network

in which the two classes of nodes are constituted by the papers and the concepts. A concept is linked to a paper if

- the corresponding term appears in the title or in the abstract of the paper, in the ACL dataset case;
- the corresponding PACS code was designated by the authors to characterize the content of their paper, in the APS dataset case.

Then, it is possible to build the one-mode projection of this bipartite network on the concept set. The resulting graph constitutes a co-word network, which, in the rest of this thesis, we will call *semantic network*, and define in mathematical terms as:

> **Semantic network.** $G^{sem} = (C, E^{sem})$ is the undirected weighted graph in which $C$ is the set of concepts in the considered field, and $E^{sem}$ is the set of links connecting elements in $C$. Two concepts $(c_i, c_j)$ are connected in $G^{sem}$ if they co-occur in a paper: in the case of the APS dataset this means that the two corresponding PACS codes have been used to characterize the same paper; in the case of the ACL dataset it means that the two terms have been used either in the same title or abstract. The weight $w(c_i, c_j)$ of the corresponding link has a value equal to the number of papers in which they co-occur.

Moreover, since we are interested in uncovering also the relations between authors and concepts, we define the two following networks, that will then be used in Chapter 6.

> **Socio-semantic network.** $G^{soc-sem} = ((A, C), E^{soc-sem})$ is the bipartite undirected weighted graph whose nodes are the authors $A$ and the concepts $C$, and $E^{soc-sem}$ is the set of links connecting elements of $A$ to elements of $C$. An author $a_i \in A$ is connected to a concept $c_j \in C$ if $a_i$ uses $c_j$ in one of her/his papers, and the corresponding weight $w(a_i, c_j)$ has a value equal to the number of such publications.

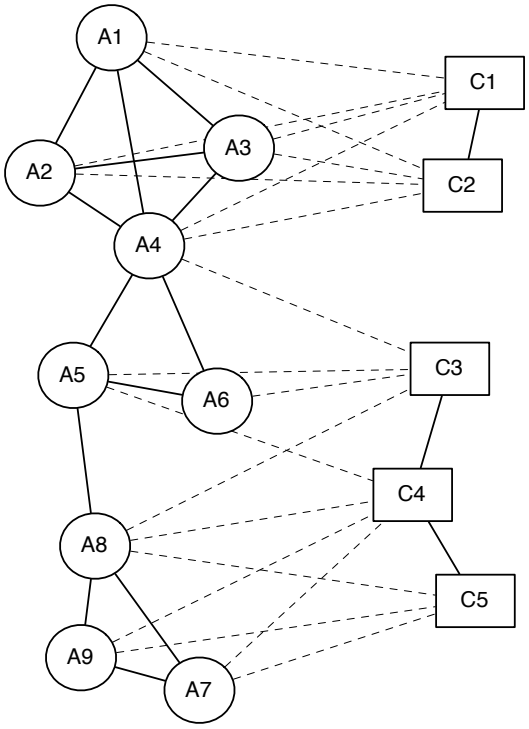



Figure 4.3: This schematic network is produced from the analysis of 4 articles. The first one gathered authors A1, A2, A3, A4 along with concepts C1 and C2, the second one A4, A5, A6 along with concept C3. Authors A5 and A8 then published an article using concepts C3 and C4. Last, A7, A8 and A9 co-authored an article about C4 and C5.

The **complete socio-semantic network** is then defined as the graph whose node set is the union of the set of authors $A$ and the set of concepts $C$ as in the socio-semantic network, and the link set is the union of the three sets $E^{soc}$,$E^{sem}$ and $E^{soc-sem}$. Therefore in this global network authors are connected to all their co-authors and to all the concepts they addressed in their publications. Symmetrically, concepts are connected to all the other concepts they co-occur with and to all the authors who use them in their publications. A schematic representation is shown in Figure 4.3.

## 4.4 Network characteristics

The social network defined above represents the structure of scientific collaborations in a given field. The semantic network is a representation of the knowledge produced in the field, even if we are aware that restricting our representation to keyword analysis is a strong limitation and does not reflect all the subtleties of a scientific field.

To gain insights about these structures we analyze the relevant features that have been introduced and discussed in Section 4.1.

### 4.4.1 Social network

The main features of the social networks are reported in Table 4.1.

**Degree distribution.** The degree distributions of the two networks built from the two corresponding data sets are shown in Figure 4.4. We observe that in both cases it is an heterogeneous distribution ranging over two orders of magnitude, which means that in both fields the vast majority of researchers only have a few co-authors, whereas a few have hundreds.

**Average shortest path length.** The value of the average shortest path length is in both cases strictly lower than the logarithm of the number of nodes in the giant connected component, as shown in Table 4.1, indicating that the networks exhibit the small-world feature.

**Clustering coefficient.** Both networks are highly clustered. The values of the average clustering coefficient are in fact orders of magnitude higher than the inverse of the number of nodes, which is

| | $\|A\|$ | $\|E^{soc}\|$ | $l$ | $\log N_{GC}$ | $C$ | $N^{-1}$ | $Q$ |
|---|---|---|---|---|---|---|---|
| ACL | 8725 | 22955 | 5.88 | 8.76 | 0.61 | $10^{-4}$ | 0.87 |
| APS | 95043 | 353495 | 6.79 | 11.33 | 0.66 | $10^{-5}$ | 0.83 |

Table 4.1: Social network characteristics.

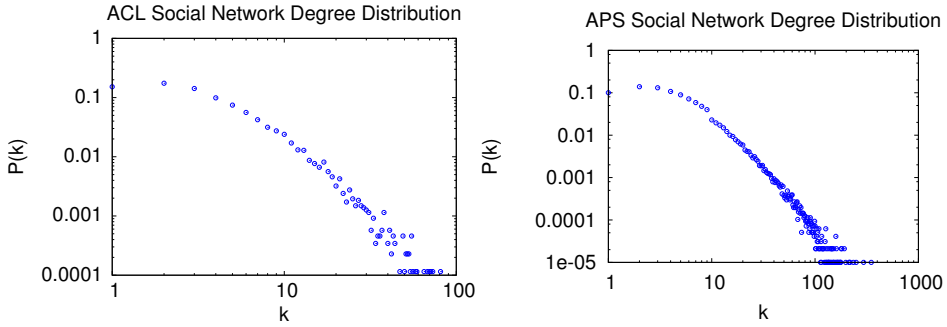



Figure 4.4: Social Network degree distribution, for the ACL (left) and the APS (right) data sets. Note that both axes are in logarithmic scale.

the order of magnitude of the value of the clustering coefficient in large random graphs, as discussed in Section 4.1.5.

**Optimal modularity.** The networks also present a very well defined community structure. The optimal modularity values (obtained using the algorithm proposed by (Blondel et al., 2008)) are indeed very high for both datasets.

To summarize, the empirical social networks representing co-authorship in the computational linguistics and in the physics communities exhibit low average short path lengths and high levels of clustering, like the small-world networks studied and modeled by (Watts and Strogatz, 1998). At the same time though, they also exhibit scale-free heterogeneous degree distribution. Moreover, they are characterized by a high modularity value, indicating a strong community structure. This means that the two domains under study are structured around different sub-communities. These are made of few researchers that publish a large amount of papers with several different collaborators, and of a majority of researchers that only collaborate with a few colleagues. These results are consistent with previous works on co-authorship networks in physics and other fields of science (Newman, 2004).

| | $\|A\|$ | $\|E^{soc}\|$ | $l$ | $\log N_{GC}$ | $C$ | $N^{-1}$ | $Q$ |
|---|---|---|---|---|---|---|---|
| ACL | 665 | 16267 | 2.04 | 6.50 | 0.41 | $10^{-3}$ | 0.35 |
| APS | 5078 | 113585 | 2.91 | 8.52 | 0.42 | $10^{-4}$ | 0.54 |

Table 4.2: Semantic network characteristics.

### 4.4.2 Semantic network

Following the same procedure, we also explore the structural features of the semantic networks. The degree distributions are shown in Figure 4.5. We observe that they are also heterogeneous ranging over two orders of magnitude. In the case of the APS dataset, the degree distribution seem to have approximatively the shape of a power law, whereas in the ACL case the distribution would probably be better approximated by a log-normal distribution. In both cases the distributions are skewed and heavy-tailed though, indicating degree heterogeneity in both networks. This means that a few frequent concepts are used in association with a lot of different concepts, whereas the majority of them co-occur only with a few others. In Table 4.2 we report the values of the other features analyzed. We observe that the semantic networks also exhibit the small-world feature and a strong community structure: they both have a small average shortest path length, a high clustering coefficient, and a optimal modularity value. It is interesting to notice that all the characteristics of the social networks hold also for the semantic networks. In this case the emergence of a community structure seem to indicate that concepts form thematic clusters that could correspond to the different research areas or topics of the given field, as will be shown in Chapter 7.

### 4.4.3 Frequency distributions

Figure 4.6 shows two more distributions: on the left side, we display the distribution of the number of publications per researcher (that corresponds to a measure of their activity); on the right side, the distribution of concept frequencies.

Both distributions are clearly heterogeneous, in particular the number of publications per researcher (on the left) shows that most authors publish

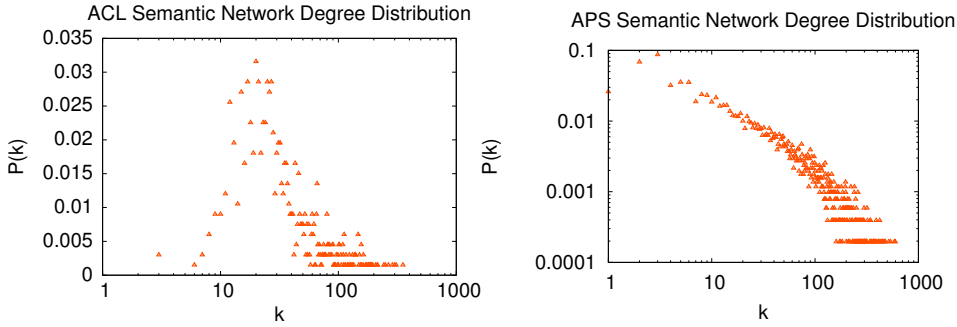



Figure 4.5: Semantic Network degree distribution, for the ACL (left) and the APS (right) data sets. Note that in the right figure both axes are in logarithmic scale, whereas in the left figure only the horizontal axis is.

very little, whereas a few authors publish a large number of papers. This phenomenon is known as Pareto's principle (or the 80/20 rule), which states that roughly 80% of the consequences derives from 20% of the causes (Pareto, 1964). Similarly, (Zipf, 1949) showed that the distribution of words is also skewed, more precisely he showed that the frequency of any word is inversely proportional to its rank in the frequency table. Since then, Zipf's idea has been applied to different areas and different kinds of human activities, like for example the number of edits of Wikipedia contributors (Muchnik et al., 2013). In particular, Newman already showed that it applies to scientific production too (Newman, 2001d).

Moreover, the distribution of concept frequencies (on the right in Figure 4.6) shows that most concepts appear in scientific publications only a few times, whereas a few concepts are very frequent. Zipf's distributions have indeed been found firstly in language, and one of the most well know cases is the frequency distribution of words in the Brown corpus[1] (Manning and Schütze, 1999, Chapter 1).

These facts shed light on the shape of the degree distributions in the social and semantic networks. The more publications a researcher has, the more likely it is that she/he has many collaborators too. Therefore the heterogeneity in the number of collaborators presented in the previous

[1]The Brown corpus is a dataset that "contains 500 samples of English-language text, totaling roughly one million words, compiled from works published in the United States in 1961" [`http://en.wikipedia.org/wiki/Brown_Corpus`]

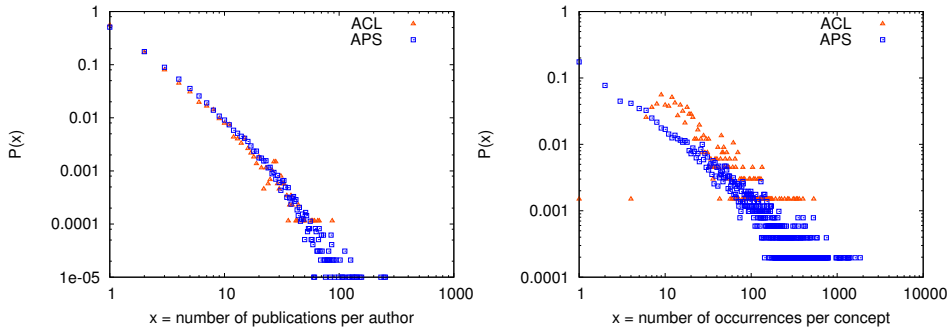



Figure 4.6: Distribution of the number of publications per author (left), and of the number of occurrences per concept (right).

section probably follows from the heterogeneity in scientific production. In the same way, the more frequently a concept appears, the more likely it is that it co-occurs with a large number of other concepts. Therefore the heterogeneous distribution of the number of semantic associations per concept probably follows from this heterogeneity in concept frequency.

### 4.4.4 Null model comparison

As discussed in Section 4.1.5, many real-world networks exhibit heterogeneous degree distributions coupled with small average shortest path lengths and high clustering. Preferential attachment models successfully reproduce the first feature but not the last two, whereas Watts and Strogats model succeeds in producing the last two but not the first one. However, the networks we initially built from the data are bipartite graphs, that we subsequently projected into one-mode network representing social and semantic interactions. (Guillaume and Latapy, 2004) show that some properties of one-mode projection networks, especially high clustering, may be a consequence of the projection process rather than a feature of the underlying data themselves. Therefore, to test the significance of the presented features, we compare our networks to a random model for bipartite network with a given degree sequence.

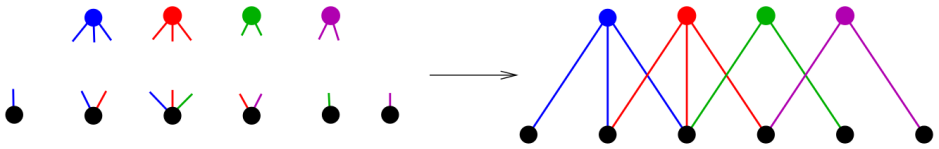


Figure 4.7: (Guillaume and Latapy, 2003): "Construction of a random bipartite network with prescribed degree distribution: first top and bottom nodes are drawn and each node is assigned a degree with respect to the given distributions, then links are chosen randomly between the two sets."

### The random bipartite model

The random bipartite model proposed by (Guillaume and Latapy, 2003) consists in a random uniform sampling of bipartite networks with *a priori* given "top" and "bottom" degree distributions. Its construction entails the following steps (see Figure 4.7 for an illustration):

1. both top and bottom nodes are generated and each is assigned a degree drawn from the respective distributions,
2. each node is assigned as many connection points as its degree,
3. top and bottom connection points are then connected randomly.

This model is able to account for the degree distribution, average distance and clustering coefficient of one-mode real world networks built as projection of bipartite networks.

### Results

For each dataset, we build a realization of the relative author-paper and concept-paper network using the random bipartite model algorithm just described, starting from the degree sequences of the relative empirical bipartite networks. We then compute the relative one-mode projections on the set of authors and on the set of concepts. Next, we compute on the resulting random versions of the social and the semantic networks the measures discussed in the previous section.

| | Social Network | | | |
|---|---|---|---|---|
| | APS | | ACL | |
| | real | random | real | random |
| $l$ | 6.79 | 4.25 | 5.88 | 4.07 |
| $C$ | 0.66 | 0.57 | 0.61 | 0.52 |
| $m$ | 0.83 | 0.39 | 0.87 | 0.52 |

Table 4.3: Real and random social network comparison.

| | Semantic Network | | | |
|---|---|---|---|---|
| | APS | | ACL | |
| | real | random | real | random |
| $l$ | 2.91 | 2.46 | 2.04 | 1.93 |
| $C$ | 0.42 | 0.47 | 0.41 | 0.41 |
| $m$ | 0.54 | 0.05 | 0.35 | 0.08 |

Table 4.4: Real and random semantic network comparison.

The random model indeed produces networks with the same heterogeneous degree distributions, and with similar values of average shortest path length and clustering coefficient as the empirical networks, as shown in Tables 4.3 and 4.4, for the social and the semantic networks respectively. The relevant difference between the real networks and their random version then mainly relies in the optimal modularity value. Concerning the social networks, Louvain-obtained modularity in the real network is about the double than in the corresponding random network. As for the semantic network, it is one order of magnitude higher in the real network in the APS dataset case, and five times higher in the ACL case. These striking results indicate that both the social and the semantic networks have indeed a strongly defined community structure that is not typical of random networks with the same degree distributions. Moreover, this suggests that groups of highly connected authors within the social networks, and, symmetrically, groups of highly connected concepts in the semantic networks, emerge in a clear-cut fashion from local interactions and are well-defined aggregates that define a different meaningful scale in the system with respect to the microscopic scale of individual authors and concepts. We will resume this point in Chapter 7.

## 4.5 Conclusions

In this chapter we have introduced network theory in order to represent the structure of scientific collaboration and knowledge production in a given field as a complex network. We have defined on the one hand the social network modeling collaborations between researchers, and on the other hand the semantic network capturing the relations between the different scientific concepts of the field. We have shown that both networks feature special topological properties such as a heterogeneous degree distribution and a well defined community structure.

This means that the majority of the researchers only have a few collaborators, whereas a central few have a large number of collaborators. Moreover, researchers tend to form different sub-communities in which most collaborations happen within the same sub-community while there are only a few collaborations with researchers belonging to other sub-communities. Similarly, a few central concepts are used in association with many different concepts, whereas most concepts co-occur only with a few others. The interactions among these concepts in the semantic network also lead to the emergence of clear-cut communities that correspond, as we will see in Chapter 7, to the different research areas of the field under study.

In Part III we investigate the possible micro-level dynamics that may account for these specific properties. Moreover, we map the communities emerging in the semantic network, describe their evolution in time, and analyze researchers trajectories in this space. The first step to achieve these goals is to provide a simple formalisms for describing social and semantic network dynamics, which is the main objective of Chapter 5.

# Chapter 5

# Modeling the time evolution of scientific research

**Contents**



In the previous chapter we have defined a social network representing the structure of collaborations among researchers in a given scientific field, and a semantic network representing the knowledge produced by these researchers through their publications. These networks are the result of a process of aggregation of all the data in the datasets under study. In other words, time has not been taken into consideration so far. However, our data are time-stamped: each paper is associated with a publication year. Thanks to this information we can investigate the time evolution of the fields under study through the evolution of the corresponding networks. Therefore in this chapter we define the time-evolving version of the networks introduced in the previous chapter. This sets the basis for the investigation of the dynamics of scientific fields carried out in Part III.

## 5.1 Time-dependent Network Definition

The first step to investigate network dynamics is to define the time-dependent versions of the networks introduced in Chapter 4.3.

The time-dependent *social network* $G^{soc}(t) = (A(t), E(t))$ is an undirected weighted graph in which $A(t)$ is the set of authors having published no later than time $t$ a paper belonging to the given data set. $E(t)$ is the set of links connecting pairs of authors in $A(t)$ that co-authored a paper published no later than time $t$. The weight $w(a_i, a_j)$ of the corresponding link is equal to the number of papers they co-authored up to time $t$. The definition of the time-dependent *semantic network* $G^{sem}(t) = (C(t), E^{sem}(t))$, and of the socio-semantic and complete socio-semantic networks, directly follow from this one. A schematic representation is shown in Figure 5.1.

Let us underline that in this framework time is modeled as a discrete process, and $t$ corresponds to a given year. The temporal information we have is in fact the paper publication year, and this is therefore the temporal granularity of our analyses.

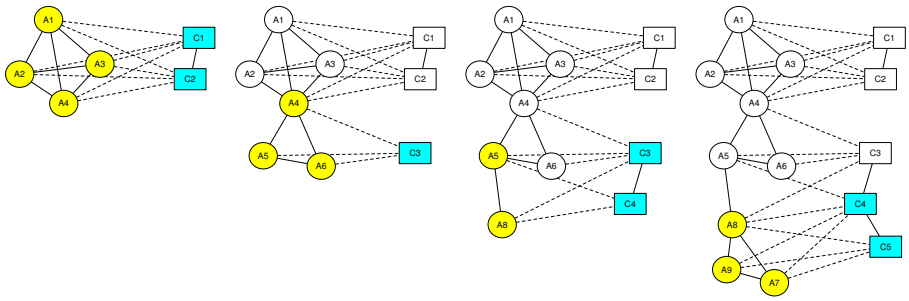



Figure 5.1: Time-evolving version of the simple example of socio-semantic network represented in Figure 4.3.

## 5.2 Social and semantic network evolution

In this section we give a general overview of the evolution of the social and semantic networks derived from the ACL and the APS datasets, and show that the time component is indeed fundamental since the two networks considerably grow over time. We present the evolution of the number of authors, concepts, and links of the three types (social, semantic, and socio-semantic) over the 24 years of overlap of the two datasets (1985-2008).

Figure 5.2 shows three different curves. For every year we plot: *i)* the number of active researchers, *i.e.* the number of authors who published at least one paper that year (blue squares), *ii)* the number of new researchers, *i.e.* the number of authors who published their first paper in the dataset during the considered year (orange triangles), and *iii)* the overall number of authors that have entered the field up to that year (black circles, the relative $y$ scale is on the right). We observe that the number of new authors per year increases over time, leading to a supra-linear growth of the total number of researchers over the years.

Figure 5.3 shows the evolution of the number of concepts. In this case, the number of newly introduced concepts (orange triangles) is relatively stable over time (a part for an initial decrease, which is due to boundary effects, since we consider all concepts present in papers published in 1985 as new). Concepts thus exhibit a very different temporal evolution than authors. It might be natural to expect that knowledge grows at a slower pace than "population", but it is nevertheless surprisingly to observe that the number of concepts introduced every year is relatively stable over time, since one could also hypothesizes an exponential growth. However we should be aware that this result might be biased by the procedures needed to identify concepts (automatic term extraction and the PACS code classification system).

Figure 5.4 shows the growth of the number of social links over time, and Figure 5.5 and 5.6 the growth of the number of semantic and socio-semantic links, respectively. Blue squares represent the number of links created during the corresponding year, and orange triangles represent the number of new links only (i.e. links that were formed for the first time during that year,

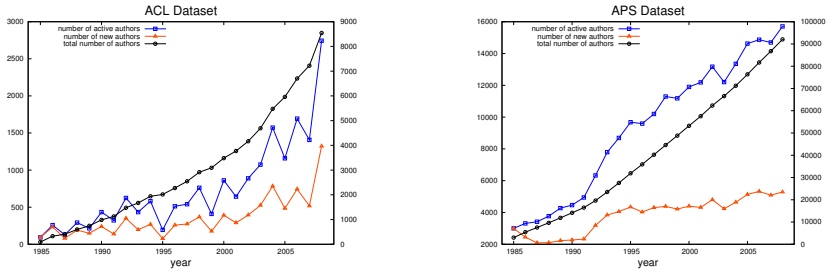



Figure 5.2: Evolution of the number of authors over time. The scale for the black dots (i.e. the total number of authors) is on the right $y$ axis, whereas the scale for blue squares and orange triangles is on the left.

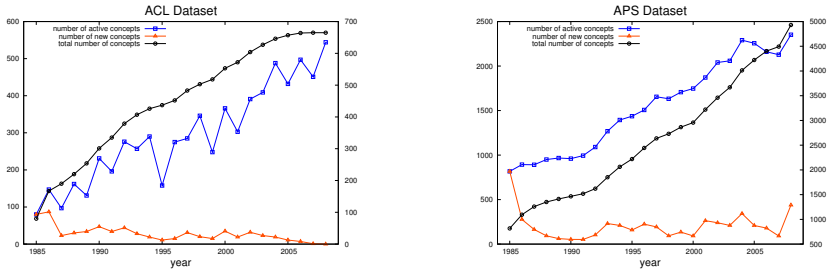



Figure 5.3: Evolution of the number of concepts over time. The scale for the black dots (i.e. the total number of concepts) is on the right $y$ axis, whereas the scale for blue squares and orange triangles is on the left.

and are not a repetition of a previously observed connection). Since blue squares represent the total number of links created or reinforced during the corresponding year, and orange triangles the cardinality of the subset of such links that are newly created that year, then the difference between the two numbers indicates the number of links that were reinforced. The fact that this difference is significantly higher than zero indicates that, other than creating new links, there is also a tendency to reinforce collaborations and existing semantic and socio-semantic links. This means that, naturally: *i)* researchers who have co-authored a paper tend to keep collaborating, *ii)* concepts which have been addressed together once are likely to keep being used together, and *iii)* a researcher that has addressed a concept in one of her/his publications tends to keep working on that concept. Lastly, black circles represent the total number of links in the network, taking into account all new and old links created in the network up to the considered year. As expected, in all three cases (social, semantic, and socio-semantic) a growth of the number of new links over time is observed.

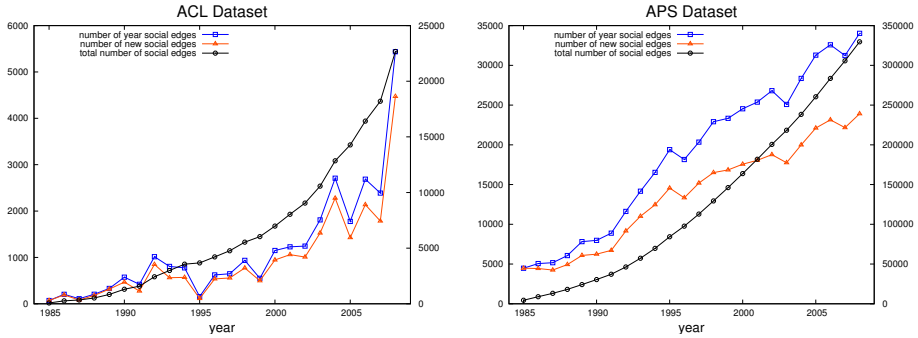



Figure 5.4: Evolution of the number of social links over time. The scale for the black dots (i.e. the total number of links) is on the right $y$ axis, whereas the scale for blue squares and orange triangles is on the left.

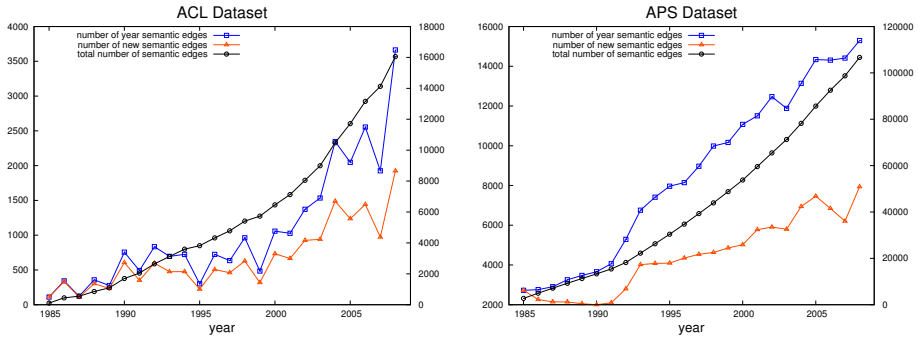



Figure 5.5: Evolution of the number of semantic links over time. The scale for the black dots (i.e. the total number of links) is on the right $y$ axis, whereas the scale for blue squares and orange triangles is on the left.

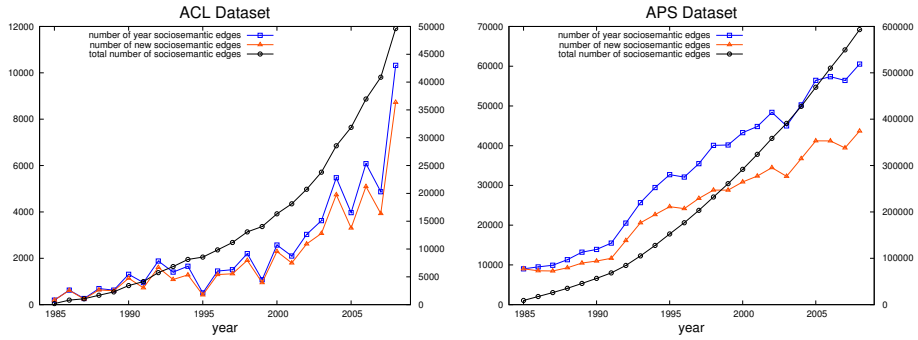



Figure 5.6: Evolution of the number of socio-semantic links over time. The scale for the black dots (i.e. the total number of links) is on the right $y$ axis, whereas the scale for blue squares and orange triangles is on the left.

## 5.3 Method and technique evolution

In the previous section we have analyzed the growth over time of the number of researchers and scientific concepts, and of the number of connections between them, in the fields of computational linguistics and physics. In this section we get back to the list of methods and techniques used in computational linguistics that has been introduced in Chapter 3.2. As already discussed, the main focus of this thesis is on the field of computational linguistics, therefore the analysis of the evolution of methods and techniques will pertain to this discipline only. Moreover, the APS dataset does not provide the abstracts of the papers, therefore, we could not have applied our method for term characterization (which is based on the argumentative analysis of the abstracts) to retrieve the methods and techniques used in physics.

During the last 30 years, the methods used in computational linguistics have changed to a large extent, the most notable shift being probably the generalization of machine learning methods since the late 1990s. This is outlined by the fact that papers in the domain nowadays nearly always include a section that describes an experiment and some results. Before the introduction of machine learning methods, the field of computational linguistics was in fact mostly concerned with language formalization through formal logic. The abstract of recent publications testing and validating new methods is always provided with a few sentences describing the results, whereas publications presenting mathematical descriptions of language are more concerned with the presentation of the formal model, which does not need validation and result discussion. To confirm this hypothesis, we observe the relative frequency over time of sentences tagged as RESULTS by the text zoning analysis of the ACL Anthology corpus presented in Chapter 3.2.2 (the reader should keep in mind that these sentences are tagged thanks to the automatic analysis presented in Chapter 3 but were of course not explicitly categorized as RESULTS in the raw corpus). In Figure 5.7, we see that the curve linearly increases from the 1980s until the late 2000s.

It is also possible to make more fine-grained observations, for example to follow over time the different kinds of methods under study. The results

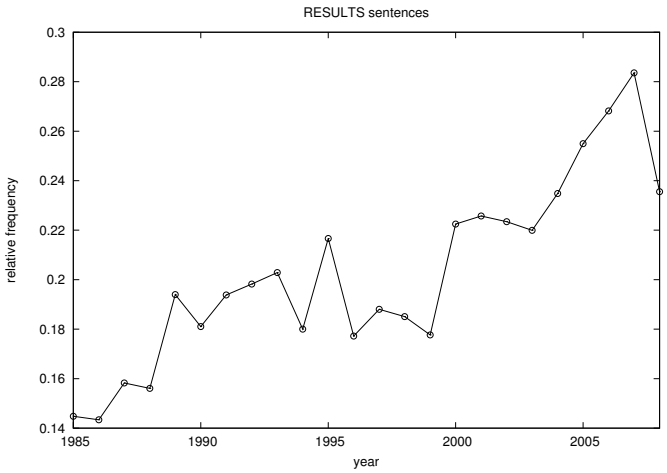



Figure 5.7: Evolution of the relative frequency of sentences tagged as RESULTS in the abstracts of the papers in the ACL Anthology.

are shown in Figures 5.8 to 5.13. Rule based methods and manually crafted resources are used all over the period, while machine learning based methods are more and more successful after the late 1990s. This is not surprising since we know that machine learning is now highly popular within the field. However, symbolic methods are still used, sometimes in conjunction with learning methods. The two kinds of methods are thus more complementary than in competition.

One could observe details that should be checked through a more thorough and qualitative study. We observe for example the success of dependency parsing in the end of the 1980s (probably due to the success of the Tree Adjoining Grammars (Joshi and Schabes, 1997) at the time) and the new popularity of this area of research in the early 2000s (dependency parsing has been the subject of several evaluation campaigns in the 2000s, see for example the Conference on Computational Natural Language Learning (CONLL) shared tasks from 2006 to 2009).

Different machine learning methods have been popular over time but each of them continues to be used after a first wave corresponding to their initial success. Hidden Markov Models (HMM) and n-grams are highly popular in the 1990s, probably thanks to the experiments made by Jelinek and his colleagues, which have opened the field of statistical machine

translation (Brown et al., 1990). More recently Support Vector Machines (SVM) (Vapnik, 1998) and Conditional Random Fields (CRF) (Lafferty et al., 2001) have received much attention in the field.

Lastly, we are interested in the distribution of these methods between papers and authors. Figure 5.14 shows the average number of terms appearing in the METHOD section of the papers over time. We see that this number increases over time, especially during the 1980s, possibly showing a gradually increasing complexity of the developed systems described in the publications.

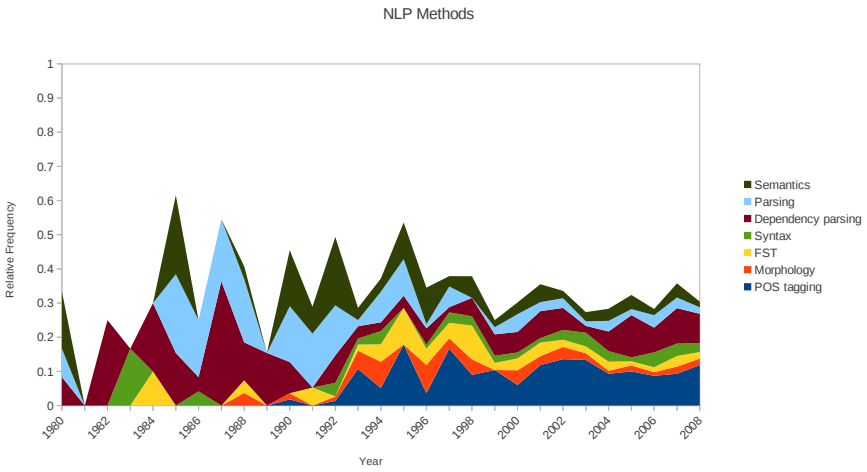



Figure 5.8: Evolution of the relative frequency of NLP method terms over time.

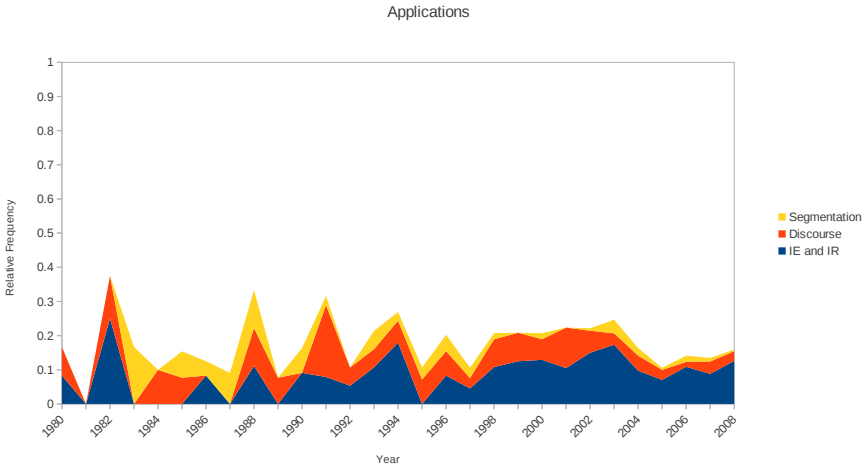



Figure 5.9: Evolution of the relative frequency of terms about applications over time.

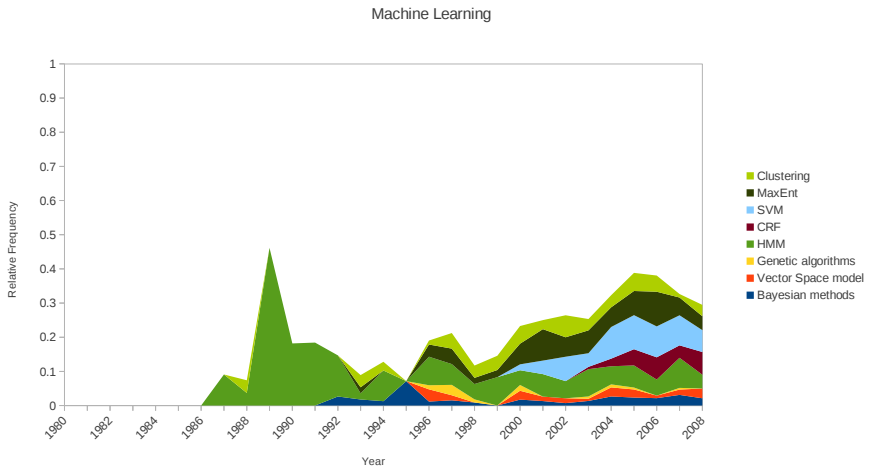



Figure 5.10: Evolution of the relative frequency of machine learning terms over time.

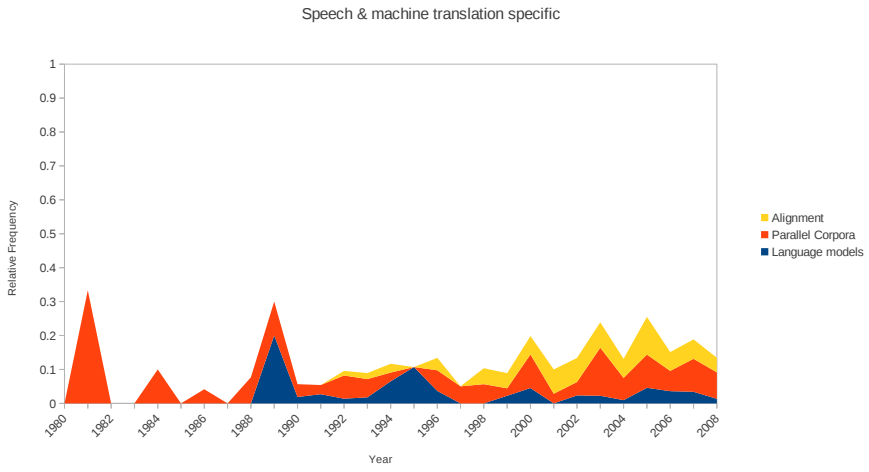



Figure 5.11: Evolution of the relative frequency of speech and machine learning terms over time.

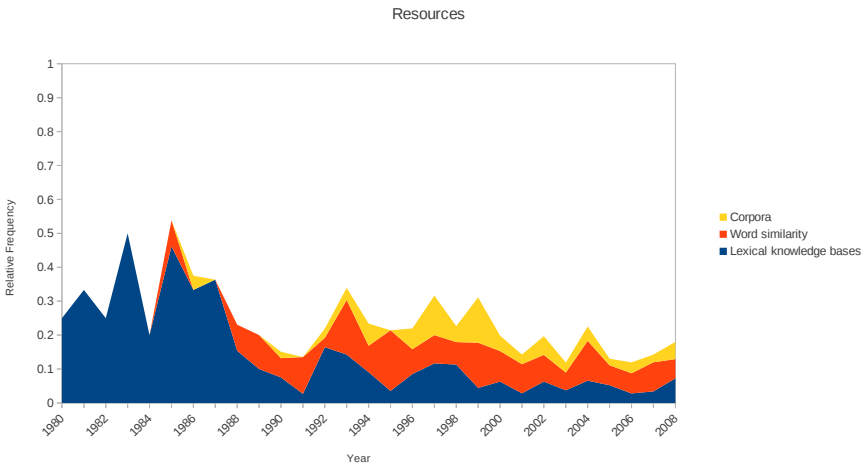



Figure 5.12: Evolution of the relative frequency of term about words and resources over time.

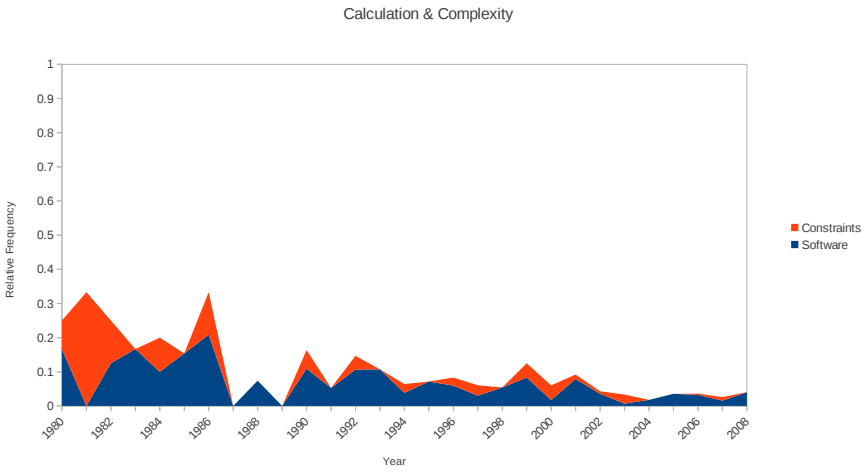



Figure 5.13: Evolution of the relative frequency of terms about calculation and complexity over time.

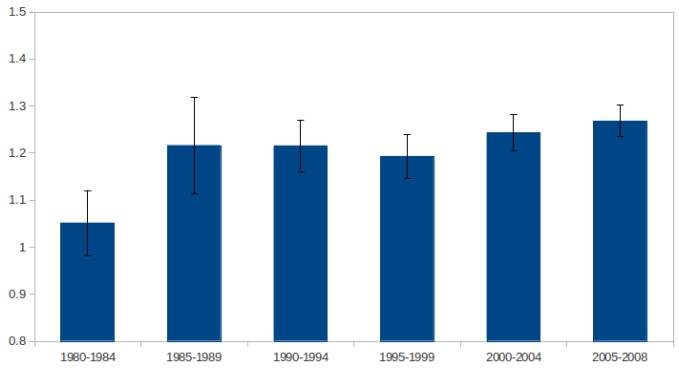



Figure 5.14: Evolution of the number of method terms per paper over time.

## 5.4 Conclusions

In this chapter we have introduced a time-evolving representation of the social and semantic networks introduced in Chapter 3. We have analyzed the growth of the number of researchers and scientific concepts, but also the evolution of their connections, in the fields of computational linguistics and physics. Lastly, we have analyzed the evolution over time of the different methods and techniques used in computational linguistics.

This chapter, together with the other two chapters composing Part II, sets the basis for the investigation of more fine-grained properties in the evolution of a scientific field. We are aware that any representation is only partial and cannot fully capture the complexity of the scientific endeavor. In particular, our representation focuses on dyadic relations. These might not fully capture the reality of the process underlying the production of a paper, which could for example be modeled also with hypergraphs (a generalization of networks in which a link can connect any number of nodes), connecting with only one link all the co-authors and concepts of a given paper.

However we have been able to show that the social and semantic networks defined are both characterized by a heterogeneous degree distribution and a well defined community structure, which has already been widely shown for collaboration networks but not systematically for semantic networks connecting scientific concepts. We have also shown that while the number of new authors always increases over time, the number of new concepts is relatively stable. This last result is interesting since one could have hypothesized a growth in the number of new concepts introduced by researchers every year, and this evolution had not been put forward in the previous studies on this field.

Moreover, to overcome some of the limitations raised above, we also decided to focus on a mesoscopic analysis of these networks. This analysis focuses on aggregates of highly connected nodes, as we will see in Chapter 7. Overall, our representation makes it nevertheless possible to at least investigate some precise questions and hopefully capture some interesting regularities in the evolution of scientific fields.

# Part III

# Investigating the socio-semantic dynamics of scientific research at different scales

# Introduction

In the third and final part of this thesis we investigate the social and semantic dynamics of scientific production.

In Chapter 6 we investigate the probability of emergence of new links in the complete socio-semantic network built from the ACL and APS corpora. We analyze what is the role played by the local neighborhood of researchers and concepts in the complete socio-semantic network, and build a statistical model based on multivariate logistic regression that quantifies their relative contribution. We consider that this level of analysis is "microscopic" since it focuses on individual actors and their interactions.

In Chapter 7 we analyze the semantic network of computational linguistics, built from the ACL corpus, at a higher scale, that we call the "mesoscopic" level. In this case the focus is on aggregates of concepts emerging from their co-occurrences. We show that the structure of the semantic network reveals groups of highly connected concepts, each corresponding to a specific area of research in the field.

Finally, in Chapter 8 we try to bridge these two levels of analysis by investigating researcher individual trajectories in two different spaces which are characteristic of the field of computational linguistics: the semantic space (built from the extracted terms and their co-occurrences), and the method space. We reconstruct the flow of researchers across the different areas forming these spaces. Lastly, we try to establish the specificities of researchers introducing innovations in the field.

# Chapter 6

# Investigating socio-semantic dynamics at the micro-level

**Contents**



In Part II we have built two dynamic *social networks* representing the structure of scientific collaboration in computational linguistics and physics, based on the analysis of co-authorship in the ACL and APS datasets, respectively. Moreover, we have built two corresponding dynamic *semantic networks* representing the structure of knowledge production within the two fields under study by connecting concepts that are addressed together in the papers of the corresponding datasets. Lastly, from the same datasets, we have built two dynamic *socio-semantic networks*, one for each of the two datasets, connecting researchers to the concepts they address in their publications.

In this chapter we investigate the dynamics of these networks. Our goal is to understand what are the fundamental mechanisms playing a role in the evolution of these networks and in particular in the creation of new links. This means that we would like to uncover what leads to *i)* the initiation of a new scientific collaboration, *ii)* the creation of a connection between two different scientific concepts, and *iii)* the adoption of a new concept by a researcher. Our hypothesis is that the social and the semantic networks are co-evolving structures, and therefore each of these processes can only be understood by taking into account the whole socio-semantic structure. We propose three statistical models based on multivariate logistic regression that successfully account for these three processes.

The chapter is structured as follows. In Section 6.1 we review the literature aiming at modeling the evolution of co-authorship and co-word networks (that was briefly introduced in Chapter 1), and illustrate the novelty of our work. Then in Section 6.2 we present our statistical models representing the emergence of new links in the social and in the semantic networks. Finally, the model for the emergence of new links in the socio-semantic network is introduced in Section 6.3.

## 6.1 Literature overview

The diffusion of knowledge and information on the web is an active area of research nowadays. (Gruhl et al., 2004), for example, investigate the role played in this process by the structure of the underlying social network. Moreover, the co-evolution of the social network structure and content has been the focus of recent studies. (Teng et al., 2012) analyze three datasets of online interactions (Twitter, the online virtual world SecondLife, and the Enron email corpus[1]) and show that there is a correlation between the diversity and novelty of the information being communicated, and the structure of the underlying social network.

In these works the network of social relations is reconstructed from the data and analyzed, but content is studied in terms of distinct topics: the relations between these topics are not taken into account to actually reconstruct a semantic network representing the structure of the produced knowledge. A notable exception is the work of (Wang and Groth, 2010). In this paper the authors reconstruct both a social and a semantic network from papers published in the World Wide Web conferences as well as from a Dutch political forum. For each author they measure her/his degree and clustering coefficient, and the degree and betweenness centrality of the topics she/he addresses in her/his publications. They then use time-series autoregressive models to investigate the dynamic influence of each property on the other ones.

In this thesis we want to explore the co-evolution of these social and semantic structures in terms of emergence of new links between researchers and between concepts. As already discussed in Chapter 1, we build on the work of (Roth, 2005). Roth shows that the interaction propensity between two researchers (in the context of the community of biologists studying the zebrafish) is correlated to their degree in the social network, their semantic capital (*i.e.* the number of concepts they have investigated), and their social and semantic distance (which are measured in terms of number of common collaborators and number of common concepts, respectively).

[1] "The Enron Corpus is a large database of over 600000 emails generated by 158 employees of the Enron Corporation." [`http://en.wikipedia.org/wiki/Enron_Corpus`]

Compared to this previous study, the novelty of the work presented in this chapter relies on the following points. Firstly, Roth focuses on only one variable at a time, whereas we build a statistical model that takes into account all the features under consideration at the same time, to understand the relative contribution of each of them. Secondly, we also build a model for the evolution of the semantic network which is based on the whole socio-semantic structure, whereas Roth only investigates the evolution of the social network. Lastly, we explore the propensity of a researcher to address a new concept, which is a feature that had not attracted much attention so far.

## 6.2 Modeling the dynamics of the social and the semantic networks

What leads to a new scientific collaboration? What are the factors that may help us predict that two researchers who never worked together beforehand will co-author a paper in the future? Moreover, can we identify predictors for events such as two concepts being addressed together by researchers at some point although they were only referred to separately in the past?

In the network analysis framework this can be formalized as a link prediction problem, *i.e.* the task of finding out which links are missing or will be created in the future (Liben-Nowell and Kleinberg, 2007; Lü and Zhou, 2011). However, and contrary to this kind of studies, our goal is to assess what are the key factors that play a role in the creation of a new link. Our actual objective is essentially to identify key determinants of the link production process and to quantify the respective importance of these determinants.

We focus on the exploration of factors pertaining to the whole socio-semantic structures built to represent the fields under study. A complete picture would of course need the inclusion of more features, such as institutional affiliation, geography, grant availability, etc. We plan on extending our analysis and include more factors in the future, but the idea of this thesis is to study the interplay of the social and semantic dimensions and understand what role those specific features play.

In Chapter 4 we have shown that both the social and the semantic networks built from the ACL and the APS datasets present the following characteristics: *i)* a heterogeneous degree distribution, and *ii)* a well defined community structure. In the social network case, this means that *i)* a few researchers have many collaborators, whereas the majority has only a few, and *ii)* researchers are clustered in sub-communities made of researchers with many connections among each others (in terms of co-authorship), and few relations with other researchers from other sub-communities in the network. In the semantic network case, the two characteristics recalled above mean that: *i)* a few concepts are connected to many others, whereas the majority is connected with only a few other

concepts, and *ii)* concepts are clustered in well defined communities that correspond, as we will see in Chapter 7, to the different research areas of the field under study.

The heterogeneity of the degree distribution suggests that one mechanism taking place in the dynamics of these networks might be "preferential attachment", *i.e.* the tendency to connect to nodes that are already highly connected. Moreover, the well defined community structure suggests that at the same time homophily mechanisms may also play a role. Researchers might for example preferably connect to researchers with whom they already have common collaborators, or who work on the same topics. Similarly, a concept is more likely to become connected to concepts that already belong to the same thematic area.

Therefore, we make the hypothesis that there are three fundamental variables at stake in the emergence of a new collaboration between a pair of researchers: *i)* the number and the strength of the connections they already have with other researchers, *ii)* the number of collaborators they already have in common, and *iii)* the thematic or methodological similarity of their previous works. Similarly, we propose the three following variables for the emergence of a new connection between two concepts: *i)* the number and the strength of the connections they already have with other concepts, *ii)* the number of common concepts they are both already related to, and *iii)* the number of researchers who have already explored both concepts in their previous works in separate contexts[2].

The different variables just introduced can be measured in several ways. Therefore, the first step of our analysis is to identify the best way to measure each variable. We do this by computing the predictive power of different possible measures, and then pick the most predictive one, as will be shown in Section 6.2.1.

To test whether the identified variables play a role in the evolution of a given scientific field, and then their relative contribution, we build a statistical model based on multivariate logistic regression. In statistics, regression is

[2]If they had been addressed together in the same context they would already be connected in the semantic network, and therefore they would not pertain to this investigation since the focus is on *new* links.

an approach for modeling the relationship between a dependent variable and one or more explanatory variables. In particular logistic regression estimates the probability of an event occurring. More formally, the probability of an event $Y$ occurring, as a function of three explanatory variables $\vec{x} = (x_1, x_2, x_3)$, is given by

$$P(Y|\vec{x}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}. \tag{6.1}$$

Regression makes it possible to directly estimate from the data the value of the corresponding coefficients $\vec{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$, that represent the relative contribution of each variable to the probability of the event occurring. This is usually done through maximum likelihood estimation (McCullagh and Nelder, 1989).

In our model, the event under consideration is the formation of a new link between two nodes at time $t+1$. The three variables we want to investigate are measured on the network at time $t$. So, at every year $t$ for which there are publications in our data sets, we measure the value of the three variables for every pair of researchers who are not connected in the social network, that is to say researchers who have never co-authored any paper up to $t$. We then look at the structure of the network at year $t + 1$ and check if a link connects these two researchers, *i.e.* if they did co-author a paper that year. The same approach is used to extract the same information from the semantic network.

In the rest of the chapter we will detail the definition of the proposed explanatory variables, present our results, and discuss their interpretation and significativity.

### 6.2.1 Defining and selecting the measures

We now present, for each of the three variables introduced, the different ways in which they could be measured, and how we test and select the best definition for each variable.

#### Node degree

The first explanatory variable introduced, namely the number of collaborators a researcher already has, and the number of other concepts a given concept is already connected to, can be computed "in network terms" as a function of the node degree $k$. In the social network the degree of a researcher is equal to the number of her/his co-authors, and in the semantic network the degree of a concept is equal to the number of other concepts it co-occurred with. Alternatively, related relevant information is carried by the node strength $s$, which takes into account the frequency of co-authorships and co-occurrences by summing over the weights of the links incident on a node instead of just enumerating them. Therefore the first variable should be a function of the degrees or of the strengths of the two nodes under consideration.

This is indeed what preferential attachment models are built on too. Usually the selected function is simply the product of the two degrees (Barabási et al., 2002). This number becomes very large when computing it for two hubs. Degree distribution being heterogeneous (as shown in Figure 4.4 and 4.5), using a product leads to an even broader distribution of values. For this reason, we also test the square root of the degree product.

To summarize, we have four candidate measures for the connectedness of a pair of nodes $(i, j)$:

i) degree product: $k_i k_j$

ii) strength product: $s_i s_j$

iii) degree product square root: $\sqrt{k_i k_j}$

iv) strength product square root: $\sqrt{s_i s_j}$

To assess which measure is the one that contains the most information about the probability of creation of a new link, we test each option as the explanatory variable in a univariate logistic regression model. We then compare the different measures through the predictive power of the respective models. Let us underline that the four measures have first been normalized to lie in the range $[0, 1]$, so that regression results can be compared straightforwardly.

To evaluate the relative quality of the models we use the Akaike Information Criterion (AIC) (Akaike, 1974). This index is based on information theory and offers a relative estimation of the amount of information that is lost when a given model is used to represent the process generating the data. It gives a trade-off between the goodness of fit of the model and its complexity. It is defined as follows

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{6.2}$$

where $k$ is the number of parameters in the model, and $L$ its maximum likelihood. The AIC value has no meaning in itself but becomes interesting when compared to the value of other models: given a set of candidate models for the data at hand, the best one is the model that has the lowest AIC value. Other tests, like the likelihood-ratio test for example, can only be used to compare nested models. AIC, on the contrary, does not have this limitation, and can therefore be used in our case.

Results are reported in Table 6.1. We observe that in all cases but one the best model is the one using the square root of the degree product. For the only case in which this measure does not lead to the minimum information loss, the difference with the AIC values of the other competing models is very small compared to the other cases. Therefore we select the square root of the degree product as measure of the node connectedness for the global model, since it is the index that better predicts the formation of a new link.

#### Similarity between two nodes

The second explanatory variable is the number of common collaborators between two researchers, and the number of concepts to which two given concepts are both connected. These measures also have a straightforward

| | APS Dataset | | ACL Dataset | |
|---|---|---|---|---|
| Index | Authors | Concepts | Authors | Concepts |
| $k_i k_j$ | 144168 | 270616 | 12173 | 43520 |
| $s_i s_j$ | 144653 | 272941 | 12175 | 43850 |
| $\sqrt{k_i k_j}$ | 142695 | 269147 | 12177 | 43204 |
| $\sqrt{s_i s_j}$ | 143454 | 270876 | 12172 | 43297 |

Table 6.1: AIC values for the logistic regression models using the different proposed measures of node connectedness. The cells highlighted in gray correspond to the minimum values.

interpretation in network terms as the number of common neighbors that two nodes have. In the network literature, measures based on this notion are called measures of proximity or similarity, and several indexes have been proposed. For a comprehensive review see (Lü and Zhou, 2011). We will therefore call this kind of measure *social similarity between two researchers* when measured in the social network, and *semantic similarity between two concepts* in the semantic network case. An illustration of this idea is shown in Figure 6.1. (Liben-Nowell and Kleinberg, 2007) tested a number of these measures for link prediction and showed that there is no single winner among them, but all "significantly outperform the random predictor, suggesting that there is indeed useful information contained in the network topology alone". We test two of the proposed measures, the Jaccard (Jaccard, 1912) and the Adamic-Adar (Adamic and Adar, 2003) indexes, but also their weighted counterparts, defined in (Lü and Zhou, 2010), in order to check whether the inclusion of the additional information of the link strength leads to better predictions. Hence four measures of similarity between two nodes $i$ and $j$ are defined as follows.

**Jaccard Similarity**

$$J(i,j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \tag{6.3}$$

where $\Gamma(i)$ is the set of neighbors of node $i$.

**Weighted Jaccard Similarity**

$$WJ(i,j) = \frac{\sum_{h \in \Gamma(i) \cap \Gamma(j)} w(i,h) + w(j,h)}{\sum_{h \in \Gamma(i) \cup \Gamma(j)} w(i,h) + w(j,h)} \tag{6.4}$$

where $w(i,h)$ is the weight of the link connecting nodes $i$ and $h$.

**Adamic-Adar Similarity**

$$AA(i,j) = \sum_{h \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(h)|} \tag{6.5}$$

**Weighted Adamic-Adar Similarity**

$$WAA(i,j) = \sum_{h \in \Gamma(i) \cap \Gamma(j)} \frac{w(i,h) + w(j,h)}{\log(1 + s(h))} \tag{6.6}$$

where $s(h)$ is the strength of $h$, *i.e.* the sum of the weights of all the links incident in $h$.

The Jaccard index is simply the number of common neighbors over the cardinality of the union of all the neighbors of the two authors. The Adamic-Adar index considers instead only the common neighbors and weights them inversely proportionally to the logarithm of their degree. This penalizes the neighbors with high degree, which are connected to many other nodes, and therefore the connection with the two nodes under consideration is not very peculiar. As previously done, the different measures are normalized to lie in the range $[0, 1]$.

We test the four different measures using the same methodology as for the first variable. The resulting AIC scores are shown in Table 6.2. In half of the cases the best model is the one using the weighted Jaccard index, whereas in the other half it is the model using the Adamic-Adar measure. The difference between the AIC scores corresponding to the weighted Jaccard and the Adamic-Adar indexes in the case of the models in which the Adamic-Adar index performs better is lower than the difference between the two scores in the case in which the weighted Jaccard index performs better. Moreover, in the next section we show that, for what concerns the third variable, the weighted Jaccard index results in the best models. Therefore, we choose the weighted Jaccard index as a measure of node similarity for the global model.

| | APS Dataset | | ACL Dataset | |
|---|---|---|---|---|
| Index | Authors | Concepts | Authors | Concepts |
| J | 138034 | 245552 | 11715 | 42621 |
| WJ | 134642 | 244032 | 11414 | 42428 |
| AA | 134411 | 255055 | 11402 | 42614 |
| WAA | 138523 | 260462 | 11526 | 42736 |

Table 6.2: AIC values for the logistic regression models using the different proposed measures of similarity between nodes. The cells highlighted in gray correspond to the minimum values.

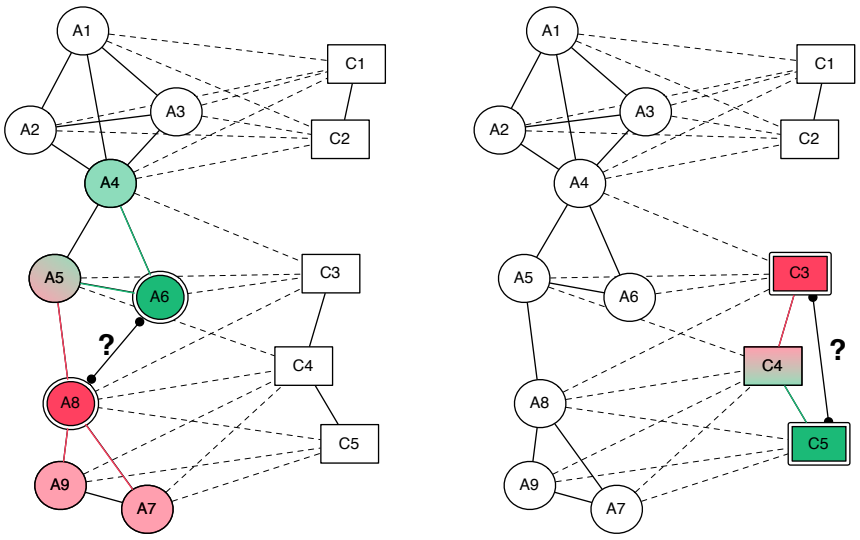



Figure 6.1: An illustration of node similarity in the social (left) and semantic (right) network.

| | APS Dataset | | ACL Dataset | |
|---|---|---|---|---|
| Index | Authors | Concepts | Authors | Concepts |
| J | 135502 | 263115 | 11857 | 43693 |
| WJ | 130114 | 263036 | 11741 | 43522 |
| AA | 137543 | 263510 | 11995 | 42931 |
| WAA | 141111 | 265807 | 12008 | 42973 |

Table 6.3: AIC values for the logistic regression models using the different proposed measures of similarity between nodes in the socio-semantic network. The cells highlighted in gray correspond to the minimum values.

**Similarity between two nodes in the socio-semantic network**

The first two explanatory variables proposed are both measures of the local structure of the network under study. The third variable we propose, on the contrary, is meant to capture the role of scientific concepts in the evolution of the social network, and, symmetrically, the role of researchers in the evolution of the semantic network. Therefore to define this third measure we have to consider the structure of the socio-semantic network. Let us recall that in this network researchers are only linked to concepts, and vice-versa. The neighborhood of a node representing a researcher is thus composed of the concepts that she/he has investigated in her/his papers. Symmetrically, the neighborhood of a node representing a concept is composed of the researchers who have worked on it. The notion of common neighbors then makes it possible to define a *semantic similarity between two researchers*, based on the number of concepts they both have already worked on, and a *social similarity between two concepts*, based on the number of researchers who have already worked on both concepts, even though in separate contexts. An illustration of this idea is shown in Figure 6.2. The four indexes defined in the previous section can be transposed in this context just by substituting the notion of social or semantic network neighborhood with the notion of neighborhood in the socio-semantic network.

Once again, we test the four different measures and select for the global model the weighted Jaccard index, since the majority of the models based on this index have the lowest AIC value, as shown in Table 6.3.

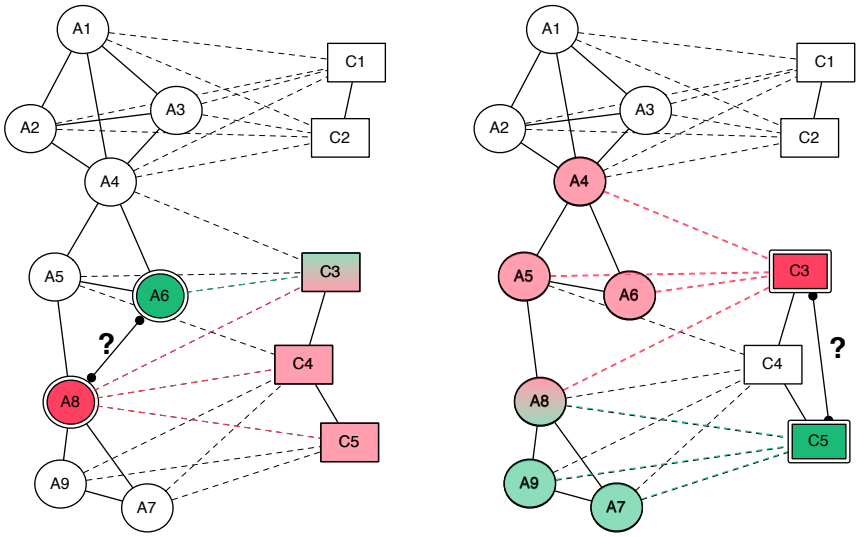



Figure 6.2: An illustration of node similarity in the socio-semantic network for researchers (left) and concepts (right).

### 6.2.2 Results

#### Social Dynamics

Table 6.4 shows the regression coefficients obtained by performing the maximum likelihood estimation for our model of emergence of new collaborations between researchers in function of the three explanatory variables *i)* square root of degree product, *ii)* social weighted Jaccard similarity, and *iii)* semantic weighted Jaccard similarity.

Since in logistic regression the function linking the explanatory variables to the probability of the event is not linear, the values of the different coefficients do not correspond in a linear way to the contribution of the respective variables, *e.g.* we cannot say that finding a coefficient $\beta_i = 5$ means that the odds of the event increase of 5 for every unit increase of the corresponding variable $x_i$. However, if $\beta_i > \beta_j$, we can say that a given variation of $x_i$ improves the odds of the event occurring in a larger way than the same variation of $x_j$. More importantly, if $\beta_i \neq 0$ (even taking into account its standard error), we can conclude that $x_i$ plays a significant role in the model. A detailed discussion of how to perform the quantitative interpretation of the coefficients is carried out in the appendix to this chapter

(Section 6.4). Here we rather want to focus on a more general interpretation meant to uncover the significance of the role played by each factor.

Firstly, we can observe that although the values of the coefficients are different for the two data sets, their relative weight is consistent, and they are all greater than 0. The largest coefficient is $\beta_2$, which means that what counts the most in the probability that two researchers will become collaborators is being "socially close", *i.e.* having many common collaborators. Semantic similarity, which measures to what degree two researchers have worked on the same topics, also plays an important role (cf. coefficient $\beta_3$). This is of course something we expected and to some extent something that is trivial, but, at the same time, the role of this variable had not been systematically taken into account and quantified in previous studies. Lastly, we see that the degree of the researchers plays the smallest role.

From these indicators we can infer two main results. The first one concerns the different roles played by the social and the semantic similarities between two scientists. One could think that the key factor for two researchers to collaborate would be to have common research interests (i.e. to have worked on a large number of same concepts). Our results show that, even if the semantic similarity plays a role, having a large number of collaborators in common is even more crucial. One could argue that two researchers who share many common collaborators are of course probably also very similar in terms of thematic expertise, but the regression coefficients represent the contribution of each variable *ceteris paribus*, *i.e.* "all other things being equal". This means that, given two pairs of researchers with the same degree of semantic similarity, if one pair has more common collaborators then their odds of collaborating in the future are much higher. On the other hand, given two pairs of researchers with the same number of common collaborators, if one pair is more similar in terms of thematic expertise, then their chances to collaborate are also higher with respect to the other pair, but their increment is smaller compared to the one in the former case.

This result is probably at least partly explained by the fact that having many common collaborators could imply geographic proximity between two researchers, or even the same institutional affiliation, which would

make a new collaboration more likely with respect to a collaboration with a researcher that belongs to another institution and/or lives in another country. In the future it would be interesting to introduced new variables, such as institutional affiliation and geographic proximity, to uncover the probable correlations and decouple the different contributions.

The second and more interesting result is the role played by the degree, which becomes of secondary importance, contrary to what previous studies seem to show. Our results show that researchers tend to create new collaborations because they already have many common collaborators, rather than on the basis of "popularity" (that corresponds to the notion of degree in our model). However, this introduces the following issue. If the degree does not play a such fundamental role, then why do we observe an heterogeneous degree distribution in the social network? This can be explained if we also take into account the connections made by the new researchers coming in the field every year. So far we have only considered researchers already in the field, since the model only accounts for new links between researchers that have already published during the previous year (since otherwise we could not define a social or a semantic similarity). Previous studies have shown that new incoming researchers preferentially connect with researchers with high degree, and this explains the heterogeneity of the degree distribution (Barabási et al., 2002). One possible reason is that newcomers are for a large part PhD students who publish their first papers together with their supervisors, who are likely to be established professors who have cumulated a large number of co-authors during the years. We can therefore conclude that newcomers tend to connect to highly connected researchers, whereas researchers who are already in the field rather tend to create new collaborations with researchers with whom they already share common collaborators and research interests.

### Semantic Dynamics

Table 6.5 shows the regression coefficients obtained by performing the maximum likelihood estimation for our model of emergence of new semantic connections between concepts in function of the three explanatory

variables *i)* square root of degree product, *ii)* weighted Jaccard similarity, and *iii)* social weighted Jaccard similarity.

We observe that, at least regarding physics, the results are symmetric to the social dynamics case. The biggest contribution is given by the semantic similarity in the semantic network, and by the social similarity in the case of the social network. The similarity with respect to the exogenous dimension (i.e. the social one in this case, the semantic one in the other case), follows. And the lowest (even though still significant) contribution is given by the degree. In the case of computational linguistics the value of the social similarity coefficient is higher than the value of the semantic similarity coefficient.

Despite this difference, these results show that, as expected, the emergence of new connections between concepts is significantly influenced by both the social and the semantic similarity. Two different concepts that have been investigated with many other common concepts in the past are more likely to be addressed together in the future, with respect to concepts that are "far" in the semantic network. This phenomenon is indeed quite intuitive since we in fact expect concepts to become more connected within the same subfield. Links within the node neighborhood are in fact likely to be inter-community links, and we will see in the next chapter that the communities emerging in the semantic network correspond to the different subfields of the discipline under study. In the future it would be interesting to investigate also what leads to the formation of the "weak ties", *i.e.* those edges that bridge different subfields.

The second result is the role played by the 'social similarity' between two concepts, which has not been taken into account by previous studies. Let us consider two concepts with a given semantic proximity and a given degree. The odds that researchers will start addressing them together significantly increase when a large number of researchers have already used both concepts in the past, but in distinct papers[3]. This results shows the tendency of researchers who work on different concepts to eventually create bridges among them.

[3]As already explained, if they had already been addressed together in the same paper they would already be connected in the semantic network, and therefore they would not pertain to this investigation since the focus is on *new* links.

| | APS Dataset | | | ACL Dataset | | |
|---|---|---|---|---|---|---|
| Coeff. | Est. | Std Err | p-value | Est. | Std Err | p-value |
| $\beta_0$ | -9.56 | 0.02 | <2e-16 | -7.11 | 0.06 | <2e-16 |
| $\beta_1$ | 5.01 | 0.14 | <2e-16 | 0.75 | 0.34 | 0.0323 |
| $\beta_2$ | 10.61 | 0.11 | <2e-16 | 8.34 | 0.24 | <2e-16 |
| $\beta_3$ | 7.37 | 0.06 | <2e-16 | 5.66 | 0.26 | <2e-16 |

Table 6.4: Social dynamics model resulting coefficients.

| | APS Dataset | | | ACL Dataset | | |
|---|---|---|---|---|---|---|
| Coeff. | Est. | Std Err | p-value | Est. | Std Err | p-value |
| $\beta_0$ | -6.39 | 0.01 | <2e-16 | -5.09 | 0.03 | <2e-16 |
| $\beta_1$ | 0.54 | 0.07 | 1.78e-15 | 1.51 | 0.14 | <2e-16 |
| $\beta_2$ | 7.15 | 0.05 | <2e-16 | 3.41 | 0.13 | <2e-16 |
| $\beta_3$ | 2.99 | 0.23 | <2e-16 | 5.38 | 0.33 | <2e-16 |

Table 6.5: Semantic dynamics model resulting coefficients.

Lastly, it is noteworthy to observe that, again, the degree of the nodes seem to play a smaller role in the dynamics of the semantic network. This means that concepts that co-occur with many other concepts generally do tend to become associated with even more concepts, but this effect is smaller compared to the role played by the semantic and social similarity.

**Assessing the model significativity**

The results discussed above already show that all three identified variables play a significant role in the dynamics of link creation. However, adding variables to a regression model always increases (or at least leaves unchanged) its likelihood. The scientific method relies on the claim that we should prefer the simplest model that is able to explain a phenomenon, and add complexity to the model only if it adds a significant improvement to its predictive power (Occam's razor). Therefore a valid method to assess the model goodness is to explore nested models. This means starting with a model that contains only one explanatory variable and then progressively adding the other variables one by one and check if their introduction leads to a significant increase in the likelihood of the model. As already said, to evaluate the models significance we use the Akaike Information Criterion (AIC). This index gives a trade-off between the goodness of fit of the

model and its complexity, discouraging overfitting (increasing the number of parameters in a model always improves the goodness of the fit). The simplest way to compare models with AIC is to compute their $\Delta_i$, *i.e.* the difference between a given model $i$ and the minimum AIC value, i.e. the one corresponding to the best model. Given an alternative model $i$ that we want to compare to the best one, $\Delta_i < 2$ suggests that the alternative is significant, values between 3 and 7 indicate that it has considerably less significance, whereas a $\Delta_i > 10$ indicates that the model is very unlikely (Burnham and Anderson, 2002). The results of this analysis for the social dynamics model and all its combinations of nested models are reported in Table 6.6, and the results for the semantic dynamics models in Table 6.7. In both cases, the model including all three variables is the one with the lowest AIC, confirming that it is the most informative one. Moreover we observe that all the $\Delta_i$ but one are higher than 10, and for the most part orders of magnitude higher (the only exception is the ACL social network model, in which the inclusion of the researcher degree improves the AIC value of 2 points only). Therefore we conclude that to obtain the best prediction we do need to take into account all three features.

| | APS Dataset | | ACL Dataset | |
|---|---|---|---|---|
| Model | AIC | $\Delta_i$ | AIC | $\Delta_i$ |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ | 123647 | | 11055 | |
| $\log \frac{P}{1-P} = \beta_0 + \beta_2 x_2 + \beta_3 x_3$ | 124581 | 934 | 11057 | 2 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_3 x_3$ | 129321 | 5674 | 11742 | 687 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ | 132134 | 8487 | 11377 | 322 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_3 x_3$ | 130114 | 6467 | 11741 | 686 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_2 x_2$ | 134642 | 10995 | 11414 | 359 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1$ | 142695 | 19048 | 12177 | 1122 |

Table 6.6: AIC values of the social dynamics alternative models.

| | APS Dataset | | ACL Dataset | |
|---|---|---|---|---|
| Model | AIC | $\Delta_i$ | AIC | $\Delta_i$ |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ | 243824 | | 42073 | |
| $\log \frac{P}{1-P} = \beta_0 + \beta_2 x_2 + \beta_3 x_3$ | 243884 | 60 | 42189 | 116 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_3 x_3$ | 259365 | 15541 | 42647 | 574 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ | 243981 | 157 | 42289 | 216 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_3 x_3$ | 263036 | 19212 | 43522 | 1449 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_2 x_2$ | 244032 | 208 | 42428 | 355 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1$ | 269147 | 25323 | 43204 | 1131 |

Table 6.7: AIC values of the semantic dynamics alternative models.

## 6.3 Modeling the dynamics of the socio-semantic network

So far we have investigated the factors that lead to the emergence of new collaborations between researchers and of new connections between concepts. We may also investigate the determinants that lead to a new socio-semantic link, in other words to the exploration of a new concept by a given researcher.

We proceed in a similar way as before. Firstly, a feature we want to investigate is the degree of the researcher and of the concept under study in the complete socio-semantic network (in which both researchers and concepts are connected to both concepts and researchers), *i.e.* the number of other concepts and researchers they are connected to. As we did for the two previous analysis, we want to explore the "rich get richer" phenomenon and see if nodes that are more connected tend to become even more so. To be consistent with previous analysis, we use the square root of the degree product.

| | APS Dataset | | | ACL Dataset | | |
|---|---|---|---|---|---|---|
| Coeff. | Est. | Std Err | p-value | Est. | Std Err | p-value |
| $\beta_0$ | -7.40 | 0.01 | <2e-16 | -5.50 | 0.02 | <2e-16 |
| $\beta_1$ | 3.14 | 0.05 | <2e-16 | 2.63 | 0.11 | <2e-16 |
| $\beta_2$ | 9.04 | 0.19 | <2e-16 | 5.46 | 0.28 | <2e-16 |
| $\beta_3$ | 9.92 | 0.04 | <2e-16 | 4.65 | 0.12 | <2e-16 |

Table 6.8: Socio-semantic dynamics model resulting coefficients.

| | APS Dataset | | ACL Dataset | |
|---|---|---|---|---|
| Model | AIC | $\Delta_i$ | AIC | $\Delta_i$ |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ | 603548 | | 75466 | |
| $\log \frac{P}{1-P} = \beta_0 + \beta_2 x_2 + \beta_3 x_3$ | 606429 | 2881 | 75955 | 489 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_3 x_3$ | 605369 | 1821 | 75733 | 267 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ | 650052 | 46504 | 76619 | 1153 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_3 x_3$ | 608353 | 4805 | 76212 | 746 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_2 x_2$ | 662605 | 59057 | 78147 | 2681 |
| $\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1$ | 660908 | 57360 | 77141 | 1675 |

Table 6.9: AIC values of the the alternative models for the socio-semantic dynamics.

Secondly, we define a measure of "social similarity" between a researcher and a concept, which assesses by what fraction of the researcher's collaborators the considered concept has already been addressed. We do this using again the weighted Jaccard index:

$$WJ^{soc}(a_i, c_j) = \frac{\sum_{a_h \in \Gamma^{soc}(a_i) \cap \Gamma^{soc}(c_j)} w(a_i, a_h) + w(c_j, a_h)}{\sum_{a_h \in \Gamma^{soc}(a_i) \cup \Gamma^{soc}(c_j)} w(a_i, a_h) + w(c_j, a_h)} \tag{6.7}$$

where $\Gamma^{soc}(a_i)$ is the set of researchers $a_i$'s co-authors (i.e. $a_i$'s neighbors in the social network), and $\Gamma^{soc}(c_j)$ is the set of authors that investigated $c_j$ in their paper (i.e. $c_j$'s neighbors in the socio-semantic network).

Lastly, we define a measure of "semantic similarity". Through this measure we want to explore the degree of thematic proximity between a researcher and a concept: has the author already investigated the concepts to which $c_j$ is connected to? Again, we measure this by means of the weighted Jaccard index:

$$WJ^{sem}(a_i, c_j) = \frac{\sum_{c_h \in \Gamma^{sem}(a_i) \cap \Gamma^{sem}(c_j)} w(a_i, c_h) + w(c_j, c_h)}{\sum_{c_h \in \Gamma^{sem}(a_i) \cup \Gamma^{sem}(c_j)} w(a_i, c_h) + w(c_j, c_h)} \tag{6.8}$$

where $\Gamma^{sem}(a_i)$ is the set of concepts author $a_i$ has already addressed (i.e. $a_i$'s neighbors in the socio-semantic network), and $\Gamma^{sem}(c_j)$ is the set of concepts to which $c_j$ is already connected (i.e. $c_j$'s neighbors in the semantic network).

We then built another model based on multivariate logistic regression that tests the probability of creation of a new link between a researcher and a concept in function of the three variables just defined: *i)* square root of the degree product, *ii)* social similarity, and *iii)* semantic similarity. The results are reported in Table 6.8. As for the previous analyses, the results are consistent for the two datasets. The social and the semantic similarities play similar roles, followed by the degree.

This means that a researcher is more likely to address a concept in the future if, on the one hand, her/his co-authors have already worked on that concept. This is also the case if, on the other hand, the author herself/himself has already addressed the concepts that are neighbors of the given concept in the semantic network. These results are quite intuitive, but the novelty relies in

the fact that thanks to our methodology we were able to quantify the role played by each of these features. This shows the importance of the social dimension in the choice of the concepts that researchers decide to tackle. This means that not only they collaborate with certain people because they work on the similar topics, but researchers actually also work on certain topics because their collaborators do.

The significativity of the model is again attested by using the AIC score, comparing the value of the AIC for the complete model with the value of it for all the other possible nested models. Results are reported in Table 6.9 and show that, as expected, the complete model has the lowest AIC and the other simpler models all have very high differences with the minimum value, meaning they are much less informative.

## 6.4 Appendix: How to interpret logistic regression coefficients?

Since in logistic regression the function linking the explanatory variables to the probability of the event is not linear, the contribution of the different variables cannot be directly interpreted. The role played by each variable can be understood computing the so-called *odds ratio increment*. The odds of the event occurring (i.e. in this analysis the creation of a new link) is defined as $\frac{P}{1-P}$. If we rewrite the probability in logarithmic terms we have

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{6.9}$$

which indeed corresponds to the logarithm of the odds. Let us consider an increment $\Delta x = x^{''} - x^{'}$ in one of our variables, for example $x_1$, keeping the other two constant. Then we have that the difference in the logarithm of the odds ratio for $x^{''}$ and for $x^{'}$ is given by

$$\begin{aligned} &\log \tfrac{P}{1-P}(x_1 = x^{''}) - \log \tfrac{P}{1-P}(x_1 = x^{'}) \\ &= (\beta_0 + \beta_1 x^{''} + \beta_2 x_2 + \beta_3 x_3) - (\beta_0 + \beta_1 x^{'} + \beta_2 x_2 + \beta_3 x_3) \\ &= \beta_1 (x^{''} - x^{'}) = \beta_1 \Delta x_1 \end{aligned} \tag{6.10}$$

If take the exponential of both sides of the formula and exploit the fact that the difference of two logarithms is equal to the logarithm of the ratio of the two arguments, we obtain

$$\frac{\frac{P}{1-P}(x_1 = x^{''})}{\frac{P}{1-P}(x_1 = x^{'})} = \exp(\beta_1 \Delta x_1) \tag{6.11}$$

The first terms is defined as the odds ratio, *i.e.* the increment of the odds of the event occurring, for an increment $\Delta x$ of a given variable. Leaving all the other explanatory variables unchanged, and fixing an increment of a given variable $x_i$, we can therefore compute the odds ratio as a function of the corresponding coefficient. In particular an increment of one unit of a given variable $x_i$ corresponds to the increment of $\exp(\beta_i)$ in the odds ratio.

For example, given a pair of researchers that never co-authored a paper before, and another such pair having the same social similarity (i.e. the

| Variable | Couple I | Couple II | Diff. | Coeff. (cf. Table 6.4) |
|---|---|---|---|---|
| degree | 0.10 | 0.15 | 0.05 | 5.01 |
| Social similarity | 0.30 | 0.30 | 0.00 | 10.61 |
| Semantic similarity | 0.20 | 0.20 | 0.00 | 7.37 |
| $\rightarrow$ Odds ratio $= \exp(5.01 * 0.05) = 1.28$ | | | | |

Table 6.10: Example of regression coefficient interpretation in the case of the probability of a new scientific collaboration.

same proportion of common collaborators), as well as the same semantic similarity (i.e. the same proportion of common concepts), if the second pair has a normalized square root of the degree product for example 0.05[4] higher than the first pair, then the odds of the second pair to co-author a paper in a APS journal in the future are 128% ($\exp(5.01 * 0.05) = 1.28$) the odds of the first pair.

[4]Our three variables are all distributed between 0 and 1 by definition, since we scaled them to have consistent results and be able to compare the different coefficients. Therefore considering a one unit increase does not make sense, and that is why we take for example 0.05.

## 6.5 Conclusions

In this chapter we have presented three novel statistical models based on multivariate logistic regression that account for the following processes: *i)* the initiation of a new scientific collaboration, *ii)* the creation of a connection between two different scientific concepts, and *iii)* the adoption of a new concept by a researcher. Our results show that in order to fully understand the dynamics of these processes we need to take into account the whole socio-semantic structure representing a given scientific field, as introduced in Part II.

The statistical models presented here aim at understanding under which circumstances new links connecting researchers and concepts are more likely to be created. In the next chapter we investigate what kind of 'meso-level' structures emerge as a consequence of these processes, focusing in particular on the semantic network since, as we will see, its community structure accurately reflects the different research areas that populate a scientific field.

# Chapter 7

# Investigating semantic dynamics at the meso-level

**Contents**



In Chapter 4 we have introduced a network representation of the socio-semantic structure of the fields of computational linguistics and physics, built from the ACL and APS corpora, respectively. In particular, we have defined, among others, a *semantic network* connecting scientific concepts on the basis of their co-occurrences in scientific publications. We have shown that this network is characterized by a well defined community structure. This indicates the emergence of a structure in which concepts form groups whose elements are highly connected among each other, and loosely connected with the rest of the network. These aggregates define a different scale of description compared to the individuals observed during the analysis of micro-level interactions in the previous chapter. In this

chapter we focus on the analysis of the semantic network at this scale, that will be called the meso-level. This analysis relies on the intuition that these emerging aggregates actually define the different research areas of a scientific field. Therefore investigating the network modular structure should give a good overview of the corresponding scientific landscape.

Science mapping is a field of research that has seen an important development in the last twenty years or so, thanks to the growing availability of digital repositories of scientific publications, and the exponential increase in computer computational power (Shiffrin and Börner, 2004). Recent studies have even proposed to analyze the totality of scientific production to design an exhaustive "atlas of science" (Börner, 2010). Many works in this area are based on citation analysis, either co-citation (Small, 1973) or bibliographical coupling (Kessler, 1963). They uncover the emergence of disciplinary and sub-disciplinary structures by aggregating papers over common citations (Leydesdorff and Rafols, 2009; Grauwin et al., 2012). Another line of work is based on co-word analysis. In this case the microscopic units forming the maps are words or noun phrases representing scientific concepts, and their aggregation is based on co-occurrence in titles, keywords, and/or abstracts of scientific publications (Cambrosio et al., 2006; Eck, 2011). Since the focus of this thesis is on scientific concepts, we will follow this route as well.

Classically these maps provide a snapshot of scientific production at a given period. However, in recent years, new methods have been proposed to produce maps that capture the evolution of scientific fields over time. Again, we can distinguish between methods based on citation pattern analysis (Rosvall and Bergstrom, 2010) and methods based on term co-occurrences (Chavalarias and Cointet, 2013). In particular, (Herrera et al., 2010) have analyzed and mapped the evolution of physics by using the APS dataset and PACS code co-occurrences.

In this and the next chapter we focus only on the ACL Anthology dataset and the field of computational linguistics. As already mentioned in Chapter 2.2, this choice has been made for two main reasons: *i)* we had access to different experts who could provide valuable comments on the maps, their interest and especially their interpretation, and *ii)* this domain has never been

represented in such a way before. A pool of 5 computational linguists has evaluated and given feedback over the different maps. They are all trained professional in the field, with several years of experience, from different nationalities (French, British, Finnish, Spanish and Russian).

In the rest of the chapter we present a mapping of the field of computational linguistics based on the publications in the ACL Anthology. We first introduce an aggregated map (Section 7.1), before proposing an illustration of the evolution of the domain (Section 7.2). The methodology used is based on the state of the art techniques cited above, but the work is original in its object of study. To our knowledge this is in fact the first large scale visualization of the structure and evolution of the field of computational linguistics.

## 7.1 Mapping the domain

### 7.1.1 Methodology

The construction of the "semantic map" of computational linguistics is done is five steps:

*i*) definition of the fundamental units constituting the map

*ii*) measurement of the proximity between those fundamental units

*iii*) application of a community detection algorithm to uncover emerging subfields

*iv*) map visualization

*v*) map evaluation

The fundamental units of our map are the terms representing scientific concepts and methods of the field, extracted directly from the publications in the ACL Anthology. We use the list produced as described in Chapter 3, and reported in Appendix A, as we did for the semantic network defined in Chapter 4.

We then build the semantic network by connecting two terms if they co-occur in the same title or abstract at least once. Before obtaining a good visualization of the computational linguistics field, we first need to define a suitable measure of distance between terms, based on some normalization of co-occurrences. This will then be used to assign a significant weight to edges in our network.

Several measures of co-occurrences normalization have been proposed. They can be classified in two classes: direct and indirect measures. Direct measures only take into account the raw co-occurrence number between two objects, and adjust this number for the total number of occurrences or co-occurrences of each of the objects, while indirect measures account for the global distribution of co-occurrences of the two target objects with all the other objects. Popular direct measures are for example the

cosine (normalized dot product) and the Jaccard index. Some works have performed systematic analyses of different measures to identify the most accurate (Leydesdorff, 2008; Eck and Waltman, 2009), but no consensus on a unique measure has emerged yet.

Therefore we tested different measures and found that we could reach the clearest visualization by using the indirect version of the mutual information measure defined in (Church and Hanks, 1990b). Let $c_{ij}$ be the number of joint occurrences of $i$ and $j$ in the same title or abstract, $s_i = \sum_{j,j\neq i} c_{ij}$ the total number of co-occurrences of $i$, and $N = \sum_i s_i$ the total number of co-occurrences. Then the mutual information of $i$ and $j$ is defined as

$$I(i,j) = \log \frac{c_{ij}}{s_i s_j / N} \ . \tag{7.1}$$

It compares the probability of observing $i$ and $j$ together (which is given by $c_{ij}/N$), with the probability of observing $i$ and $j$ independently ($s_i/N$ and $s_j/N$). Its indirect version, which takes into account the global distribution of co-occurrences, is:

$$MI(i,j) = \frac{\sum_{k\neq i,j;I(i,k)>0} \min(I(i,k), I(j,k))}{\sum_{k\neq i,j;I(i,k)>0} I(i,k)} \ . \tag{7.2}$$

In order to have the most readable and 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 in graphs: 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 was demonstrated to be one of the best community detection

algorithms (Lancichinetti and Fortunato, 2009). We also tested the Louvain algorithm (Blondel et al., 2008), which provided slightly worse results to map the domain as a whole. The evaluation of the corresponding clusters by experts of the fields (detailed below) shows that the results of this algorithm are in fact less convincing. This algorithm can however be useful to analyze the evolution of the domain over time, as we will see in the next section.

### 7.1.2 Results

CorText Manager[1] was used to draw a final representation of the network, which is shown in Figure 7.1. Each circle surrounds a detected "community", representing a thematic cluster, such as word sense disambiguation or part-of-speech tagging.

#### Evaluation

Different clustering techniques (Infomap and Louvain, cf. *supra*) and different settings have been tested and qualitatively assessed by a pool of experts of the field. Additionally, for each cluster we randomly selected 10 projected papers and an expert had to evaluate whether each article fitted well in the cluster.

Papers are projected to clusters in the following way. Each cluster is characterized by a vector of length equal to the total number of terms used to create the map. Each element of the vector is equal to zero if the term does not belong to the cluster, and otherwise equal to the centrality of the term in the cluster, computed as the weighted percentage of links in the semantic map connecting the term to other terms also belonging to the cluster. Each paper is also characterized by a vector of the same length. In this case, each element corresponding to a term that appears in the paper is equal to the relative frequency of the term in the paper, multiplied by the logarithm of the absolute frequency of the term in the corpus. For each paper, we compute the cosine similarity of its characteristic vector with every vector characterizing the different clusters. We then assign the paper

[1] `http://www.cortext.net/projects/cortext-manager.html`

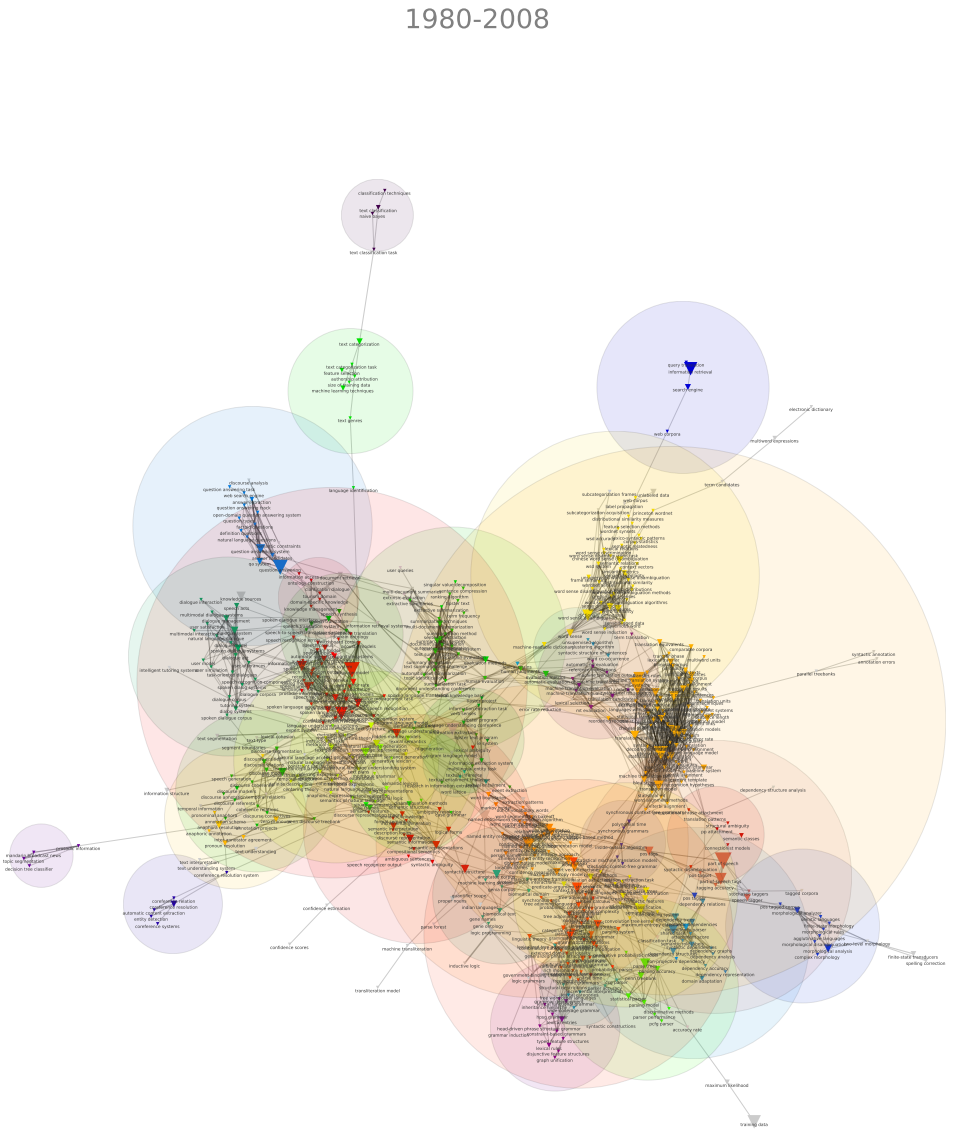



Figure 7.1: Semantic map of the ACL Anthology corpus.

to the cluster for which the similarity is the highest, provided that its value is above a given threshold.

We then computed the precision of a cluster as the fraction of papers that the expert considered relevant with respect to the cluster. The average precision obtained is 0.84, which we judge acceptable for this kind of task.

Three examples of clusters automatically obtained with the method described are provided below as an illustration. The full list is reported in Appendix B.

**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 of natural language - situation semantics - intensional logic - knowledge base - Generative Lexicon - artificial intelligence - knowledge representation - meaning representations

## 7.2 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 which subfield has attracted the most part of research effort, which subfields have emerged or transformed during the different periods.

### 7.2.1 Methodology

We base our analysis on the four following steps.

1. The corpus is divided into different periods of time. We choose 4 periods containing about the same number of publications, which results in the following intervals: 1980s, 1990s, the first and second half of the 2000s.
2. All the papers related to a given period are put together, term co-occurrences are computed for each subset of papers, and for each period a semantic network weighted through mutual information is created as described in the previous section.
3. Clustering algorithms are applied on each semantic network so as to obtain clusters of terms representing the different subfields of the domain for the different periods.
4. Lastly, the different subfields identified for each period are inter-temporally re-connected.

The clustering algorithms used are Infomap (Rosvall and Bergstrom, 2010) 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: terms may disappear from a given cluster and new terms 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 terms. A threshold has to be defined so as to avoid connecting clusters sharing too few terms 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 terms 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 two different subfields since the automatic term analysis reveals that they contain few terms 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.

As already mentioned, for our experiments, we use CorText Manager, which implements all the procedure and provides various choices for each step (the platform implements different techniques for term extraction, term clustering and cluster mapping over time). These alternative choices mean that various maps can be obtained for a same domain, providing different views of the evolution of the domain.

It must be noted that 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. These maps should be considered as broad overviews of the evolution 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. A cluster may seem to disappear because the terms constituting it are not among the most frequent during the following period. Alternatively, it may have largely evolve so that at period *t+1* no cluster contains enough common terms to be connected to the original cluster *c*. It may have merged with two different other subfields

with few terms in common overall. These 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 trends of one field while detailed descriptions will allow one to reconstruct the precise phylogeny of a domain.

### 7.2.2 Results

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

Figure 7.2 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 (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 terms are connected. We observe that the main field is now machine translation: this field has continuously grown since the late 1980s. We can also see the development of the 'question answering' (QA) task since the late 1990s: this field has been especially popular at the time thanks to the QA evaluation tracks at the Text REtrieval Conference (TREC) for example.

We can observe 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

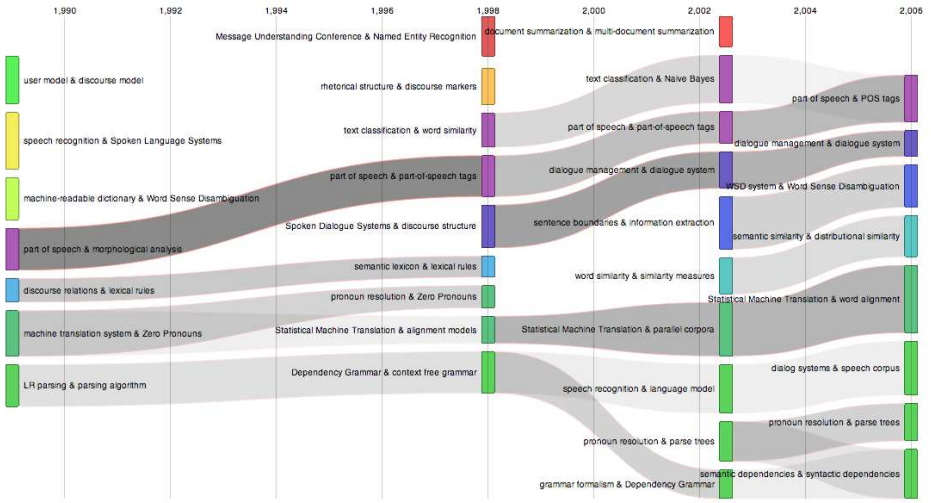



Figure 7.2: Observation of the evolution of the computational linguistics domain over time at the macro-scale.

'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 7.3 and Figure 7.4 are much more precise overviews of the domain. We can observe 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 for instance. Speech also merged with the discourse subfield at the end of the 1990s which shows a new interest in the management of dialogue structures.

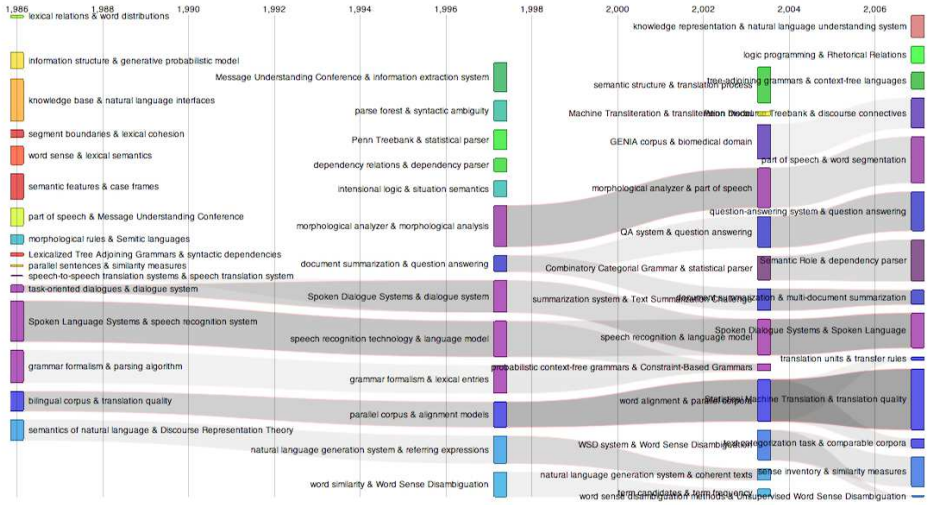



Figure 7.3: Observation of the evolution of the computational linguistics domain over time at the meso-scale.

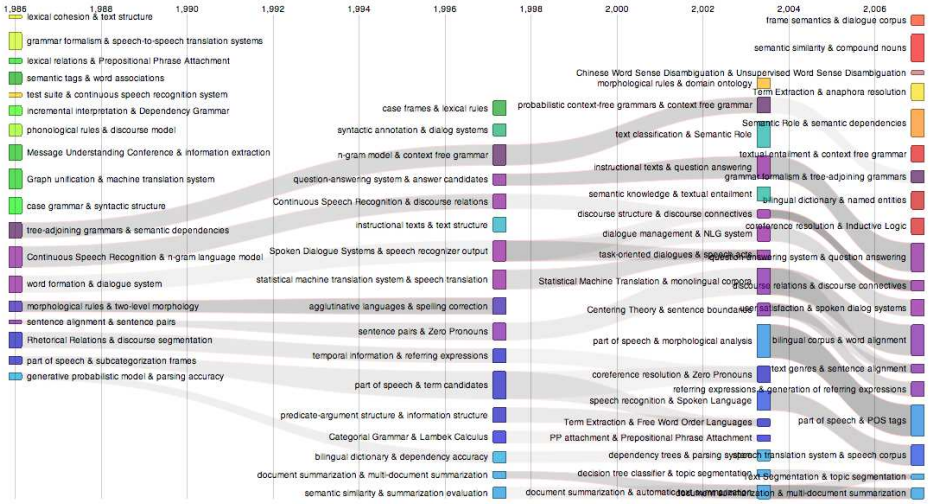



Figure 7.4: Observation of the evolution of the computational linguistics domain over time at the micro-scale

## 7.3 Conclusions

In this chapter, we have tried to analyze the evolution of the domain of computational linguistics between 1980 and 2008.

Starting from a list of terms characterizing the domain, we built a "semantic map" of computational linguistics using term co-occurrences and graph community detection methods to highlight the different "semantic communities" structuring 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 opens new avenues to historians and sociologists of science for the exploration of the main trends driving computational linguistics history (how new subfields have emerged, how some subfields have merged together or even disappeared, etc.).

In the next chapter we discuss the interplay between the micro-level analysis presented in Chapter 6, and the meso-level analysis presented in the present chapter, in order to provide an integrated description of the computational linguistics field.

# Chapter 8

# Investigating the micro-meso bridge

**Contents**



In the previous chapters we have studied the dynamics of socio-semantic networks, representing scientific collaborations and knowledge production, at two scales.

i) In Chapter 6 we investigated the probability of emergence of new links in the complete socio-semantic network. These new social, semantic, and socio-semantic links represent, respectively, collaborations between researchers, connections between scientific concepts, and a researcher addressing a given concept in her/his publications. We analyzed the probabilities of these events in function of the local neighborhood of researchers and concepts in the complete socio-semantic network. We consider that this level of analysis is microscopic since it focuses on individual actors and their dyadic interactions.

ii) In Chapter 7 we then analyzed the semantic network of computational linguistics at a higher scale, that we call the mesoscopic level. In this case the focus is on aggregates of concepts emerging from their co-occurrences. We showed that the structure of the semantic network reveals groups of highly connected concepts, each group corresponding to a specific area of research in the field under study, such as, for example, machine translation or word sense disambiguation in computational linguistics.

We now want to explore the following questions. How do researchers move in this landscape of scientific knowledge, characterized by several "valleys" representing the different research areas in the field? Is there a correlation between the semantic links connecting different areas (that emerge from papers at the frontiers between two areas, and in which we therefore find concepts belonging to both) and the flow of researchers from a research area to another during the course of their scientific carrier?

In this final chapter we investigate researcher individual trajectories in two different spaces that characterize the field of computational linguistics. Firstly, in Section 8.1, we consider the methods identified in Chapter 3.2. Like the research areas emerging from the semantic network, the different methods and techniques also constitute a meso-level description of the knowledge landscape. Each method that we have identified corresponds to the grouping of different terms used by researchers describing a given technique in the abstracts of their papers. Secondly, in Section 8.2, we explore researcher trajectories in the more general semantic landscape that we have reconstructed in Chapter 7.

## 8.1 The dynamics of researchers in the method space

In this first section we analyze the dynamics of researchers in the landscape defined by the methods used in computational linguistics, and we especially focus on the researchers that have introduced new techniques in the field. We will call these individuals "pioneers".

Since the seminal work of Everett Rogers (Rogers, 1962), several studies have shown the importance of the role played by innovators and early adopters in the diffusion of innovations. (Coleman et al., 1966) investigated the diffusion of a new medical drug among physicians, and were the first to empirically track the diffusion of an innovation through an interpersonal network. This has been shown to be true also in the context of the diffusion of opinions, in which opinion leaders play a stronger role with respect to mass media (Katz and Lazarsfeld, 1970).

Rogers defines five categories of adopters on the basis of their innovativeness: innovators, early adopters, early majority, late majority, laggards (Rogers, 1962). Later marketing studies have shown that there is a chasm between the early adopters and the remaining categories, and the most difficult step in the diffusion of a new product is the transition from the adoption by the "visionaries" (composed by innovators and early adopters) to the adoption by the rest of the population (Moore, 2002).

In this chapter we explore the characteristics of the researchers that were the first to introduce a new technique in the field of computational linguistics. Following the works cited above, we define as "pioneers" the first 16% of researchers who have published papers in which a given new method was firstly introduced (for example, the first researchers writing papers where the terms 'support vector machine' or 'SVM' appear). This percentage is given by the sum of innovators (2.5%) and early adopters (13.5%), as shown in Figure 8.1. For statistical reasons, we limit ourselves to researchers who have published at least 5 papers in the ACL Anthology, in order to take into consideration researchers who have contributed to the domain during a period of time relevant for the study.

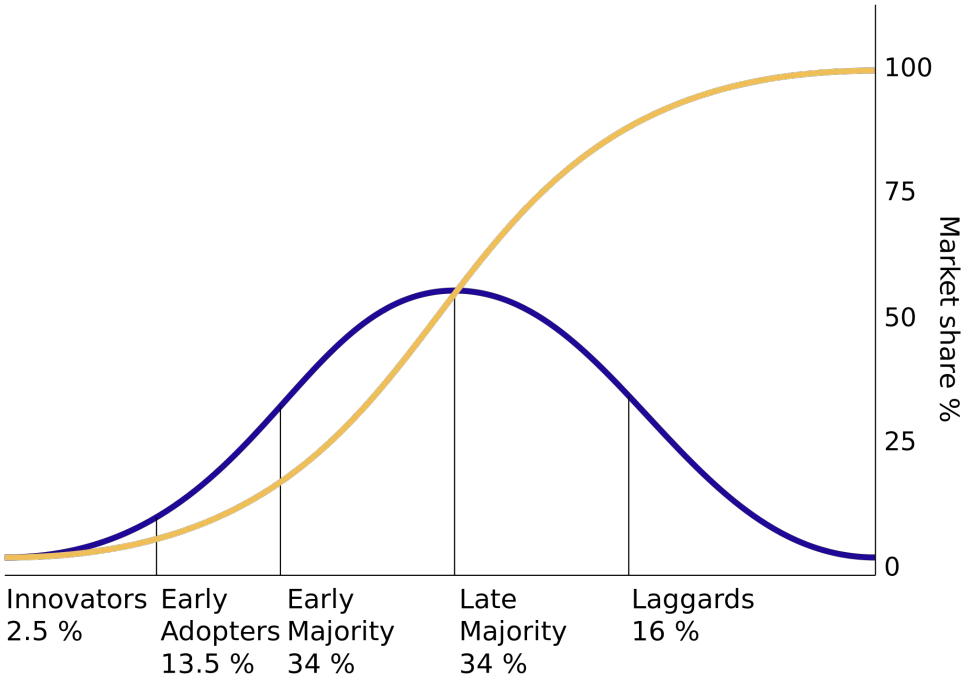



Figure 8.1: Diffusion of innovation according to Rogers (Rogers, 1962).

How are new methods introduced in the field? Are they mainly brought by young researchers or is it mainly confirmed researchers who develop new techniques (or import them from related fields)? Are natural language processing (NLP) experts specialized in one method or in a wide variety of different methods?

These questions are of course quite complex. Each individual has her/his own expertise and her/his own history but we think that automatic methods can provide some interesting trends over time. For example, (Anderson et al., 2012) show that evaluation campaigns have played a central role at certain periods of time (which does not mean of course that there was no independent research outside these campaigns at the time). Our goal is thus to exhibit some structural features that could give birth to hypothesis about the dynamics of computational linguistics or even make it possible to compare the evolution of this field with other fields. Out tools provide some hypotheses that must of course be confirmed by further observations and analyses. We do not claim that they provide a precise view of the domain.

Our results show that the "pioneers" have two interesting characteristics. Firstly, they are the most diverse researchers, diversity being simply measured by counting the number of methods that these researchers use. Secondly, these innovators are more likely to be researchers who are new to the particular field under study.

Figure 8.2 shows the distribution of researchers in terms of number of methods used throughout their publications. We measure it by counting the number of distinct methods that each researcher has used in all her/his publications. Most of the researchers use only one method and, as expected, the number of researchers decreases with the number of methods, which means that there are few researchers who are really specialists of many methods, whereas most researchers are specialists of only one or two. Researchers with many publications are probably more likely to be characterized by a higher number of techniques used. Therefore to isolate this bias from our measure, we have traced different curves for different categories of researchers, grouped according to the number of papers they have published. We observe the same decreasing trend for each category, which confirms that our result is valid independently of the researcher productivity.

One may then want to observe the diversity of methods employed in the domain especially by the set of people called pioneers in our study. Figure 8.3 shows in red (solid line) the distribution of the number of methods used by pioneers, which can be compared with the blue boxes (dashed line) featuring the distribution of the number of methods used by all the researchers (the latter being the same curve as in Figure 8.2, but taking into account in a single curve all the researchers with at least 5 publications, instead of grouping them by number of publications). We see that pioneers, when taking into consideration the whole set of papers in the ACL Anthology, are using a larger number of methods. They are over represented among researchers using 4 methods or more. This is interesting because it indicates that researchers who are pioneers in one technique are also more likely to explore different methodologies in their studies.

We then try to determine when, during their career, researchers introduce innovative methods. Practically, we examine at which point of their career

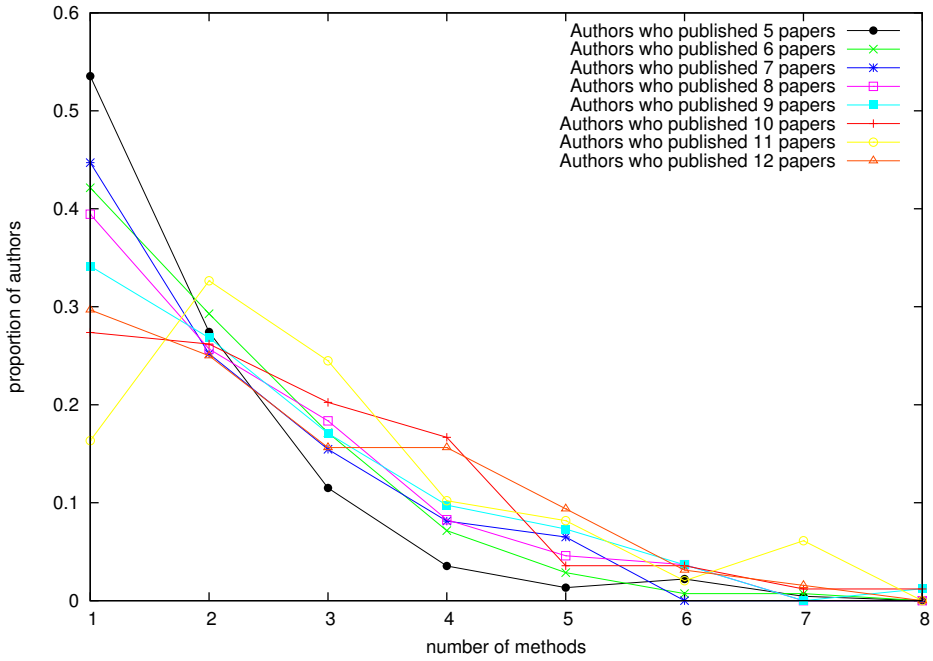



Figure 8.2: Proportion of researchers using a given number of methods, for groups of researchers having published the same number of papers.

the researchers that we have defined as "pioneers" (which refers to the first researchers using a new method) have published the paper precisely introducing this innovative method. For example, if a researcher is one of the first who employed the term 'SVM', has she/he published about this subject at the beginning of her/his career or later on?

The result is shown Figure 8.4, in which we plot the cumulative distribution of the number of researchers that have introduced a methodological innovation at a given point of their career (red squares). For deriving those curves, we first enumerate the papers a researcher has already published before the paper in which she/he introduces for the first time a new method, and then normalize this number over the total number of papers she/he published. We compare this distribution with the same measure applied to the whole population, non pioneers included (blue triangles). In this case we consider the fraction of papers that a researcher has published before using for the first time a given new method in their work. Therefore

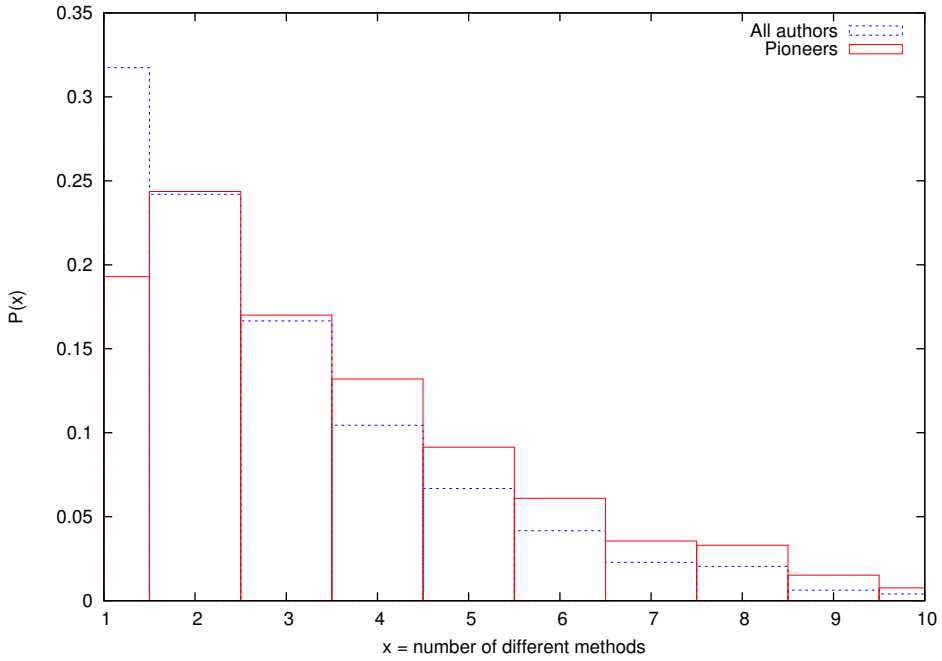



Figure 8.3: Proportion of "pioneers" using a given number of methods $x$, compared with the general distribution all the other researchers in the corpus.

the blue dashed line corresponds to the cumulative distribution of the "age" (measured in terms of fraction of published articles over her/his whole production) at which a researcher uses for the first time a method that is new for her/him because she/he has never used it before, even if the method had been already introduced in the field. The red solid line corresponds instead to the cumulative distribution of the "age" at which pioneers introduce a method which had never been used in the field before.

To make sure to avoid biases due to the fact that the years we consider constitute a limited set, we take into account only methods introduced at least 10 years after the first considered year. This should guarantee that if the researchers were already in the field in the years preceding the publication under consideration (about a new method) we are indeed able to track them, and do not actually consider as newcomer the researchers who have already published in the field but prior to the first year in the dataset.

We observe that about 50% of pioneers, when they introduce an innovation, are also entering the field for the first time. The same is true for about 40% of all the researchers. This discrepancy seems to indicate that pioneers are more likely to be newcomers in the field of computational linguistics. Newcomers can include both young researchers but also established researchers coming from different fields, in particular other areas of computer science not represented in the ACL Anthology. We know that for example Hidden Markov Models have been highly popular in 1990s after Jelinek's team introduced this technique in the field of machine translation. The technique was not so much used before in computational linguistics, but was already highly popular in speech recognition, the initial field of expertise of Jelinek's team before the 1990s (Jelinek and his colleagues were confirmed and even highly established researchers already at the beginning of the 1990s but speech processing was very poorly represented in the ACL database before the 1990s, and this is still a quite different scientific field, with its own associations and publications).

The figure also shows another interesting fact, namely that pioneers tend to introduce innovations in early stages of their careers. We observe that 70% of pioneers had already published less than a third of their scientific production when they introduced a new method (whereas for researchers in general we find a value of about 50% for the same publication fraction). This seems to indicate that innovations are mainly introduced by young researchers, or, as discussed above, by researchers that have recently arrived from close but different scientific communities.

Lastly, we measure the flow of researchers between methods over time: given a researcher who has worked on a given method, what are her or his chances to then work on this or that other method later on during her or his career? For example, is a researcher who used Hidden Markov Models more likely to move to Support Vector Machines or to Conditional Random Fields, given that both methods are popular at the considered time?

We measure these flows by analyzing which methods researchers used at different time periods. For this purpose, we use the same four time intervals defined in Chapter 7: the 1980s, the 1990s, the first half of the 2000s, and the second half. For each couple of methods, we count the number of

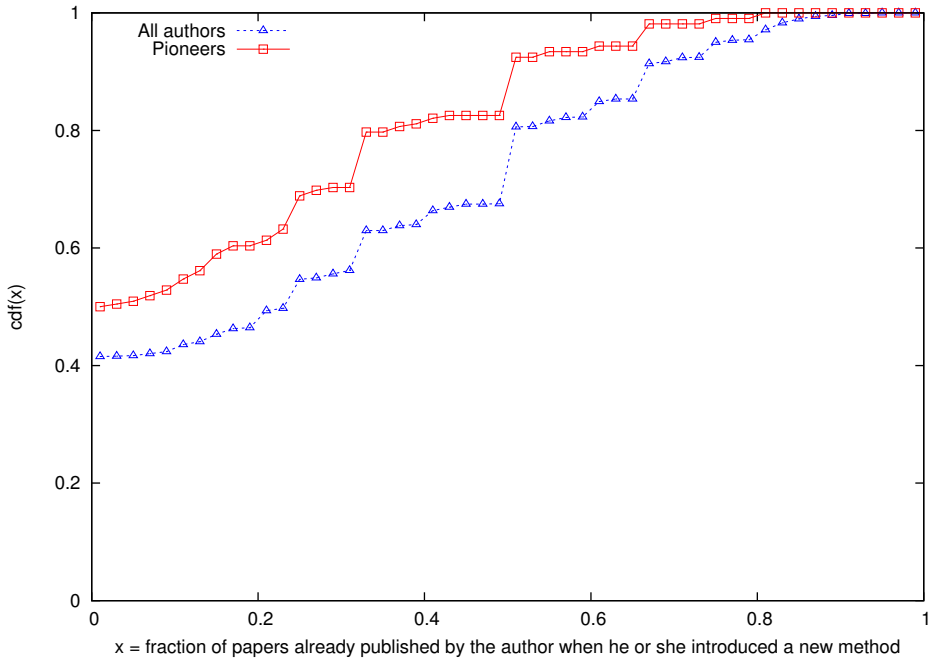



Figure 8.4: Cumulative distribution function of the number of papers already published by "pioneers" (red squares), and by all researchers (blue triangles), when they have published their paper on the new method, compared to the total production of their career.

researchers who have worked on one method in one period and then on the other in the following period. Flows are then normalized by taking into account the total number of researchers involved. Only flows that contain over 10% of outgoing researchers are conserved. Figures from 8.5 to 8.7 show a visualization of the obtained flows.

We observe that the flow from the 1980s to the 1990s mostly concerns NLP methods, machine learning techniques not being used yet, apart from Hidden Markov Models, that had been popular since the 1990s (Figure 8.5). From the 1990s to the first half of the 2000s researchers move to Machine Learning, and, for example, Support Vector Machines become very popular (Figure 8.6). From the first to the second half of the 2000s researchers focus more on Conditional Random Field (a popular machine learning technique for natural language processing), and on a specific

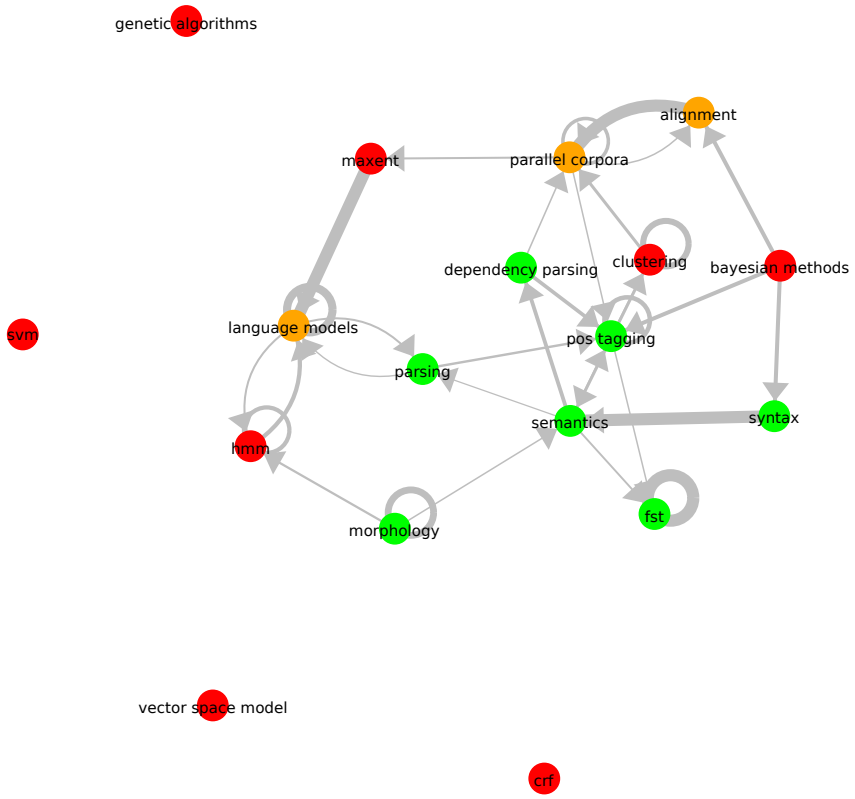



Figure 8.5: Network of researcher flows between methods from the first (1980s) to the second (1990s) time period. Colors represent the different classes of methods: green for NLP methods, red for Machine Learning methods, and orange for Speech and Machine Translation specific methods.

domain of syntax called Dependency Parsing, which was in fact the object of several evaluation campaigns in the 2000s, in particular during the Conferences on Computational Natural Language Learning (CoNLL) from 2006 to 2009 (Figure 8.7). Moreover, we observe that Part-of-speech tagging has remained very central over time, which is probably a consequence of the fact that this method is systematically used as a pre-processing task in computational linguistics. Lastly, we notice that 'alignment' and 'parallel corpora' are very central since the 2000s, which reflects the popularity of Machine Translation during the last decade.

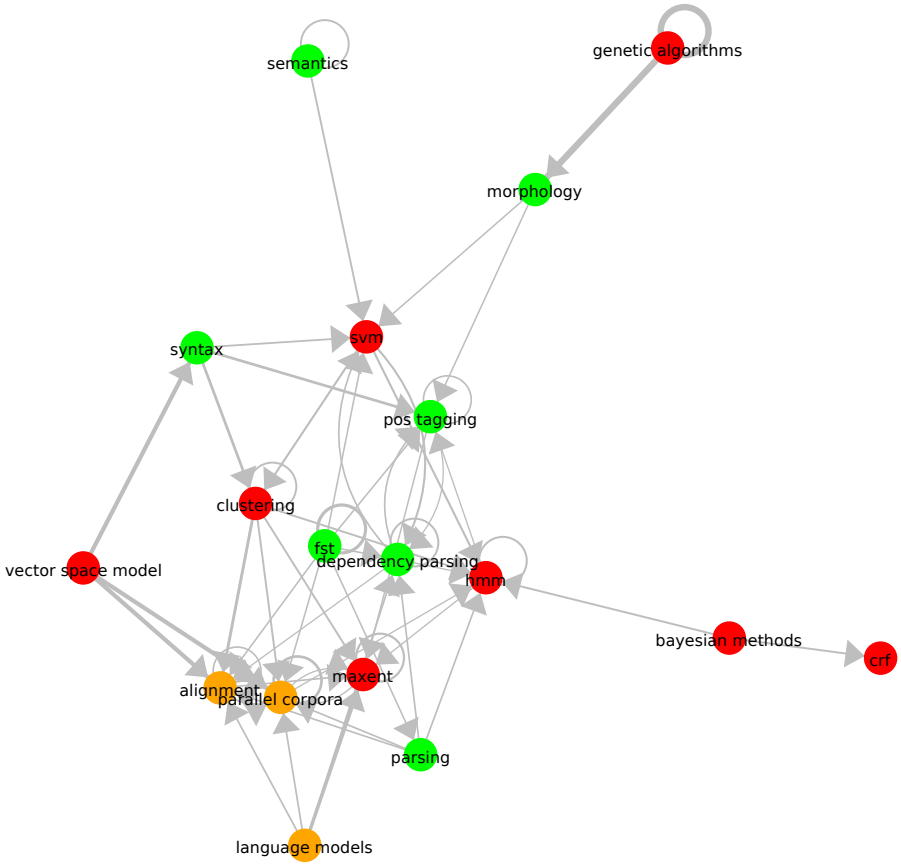



Figure 8.6: Network of researcher flows between methods from the second (1990s) to the third (first half of 2000s) time period.

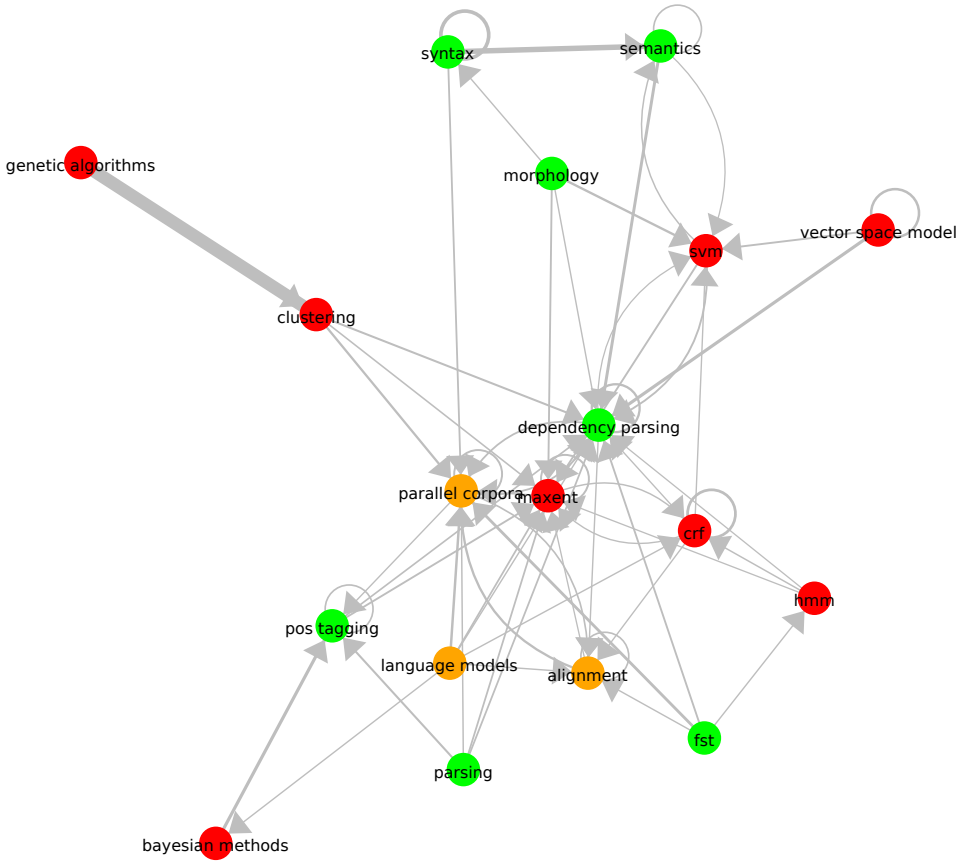



Figure 8.7: Network of researcher flows between methods from the third (first half of 2000s) to the fourth (second half of 2000s) time period.

## 8.2 The dynamics of researchers in the semantic space

After focusing on the methods and techniques used in computational linguists, in this section we consider the whole semantic space. This is composed of all the different kinds of terms we have extracted from titles and abstracts, as described in Chapter 3. In Chapter 7.1 we have shown how we can identify the different research areas by detecting the clusters of highly connected nodes in this network.

In this section we explore the characteristics of the "pioneers" in the different areas, that we define exactly in the same way as in the previous section, *i.e.* the first 16%[1] of researchers publishing in a given new area. Moreover, we explore the flow of researchers from one research area to another over time.

First, we explore the distribution of researchers having published papers in a given number of different research areas. We measure this by projecting each paper in a given semantic cluster according to the terms contained in its title and abstract (as described in Chapter 7.1) and then, for each researcher, we check how many different clusters have been assigned to her/his publications.

Figure 8.8 shows the distribution for i) the pioneers and ii) the whole set of researchers who have published at least 5 papers. We observe, like for the methods, that pioneers tend to be more thematically 'diverse' with respect to researchers in general: the proportion of pioneers connected to a low number of subfields is lower than for the proportion of researchers in general. On the other hand, the number of pioneers connected to a large number of subfields is high, which indicates that pioneers are more likely to publish in different research areas.

Secondly, we observe the cumulative distribution of the proportion of papers that pioneers in a given subfield have already published (with respect to their whole production) when they publish their first paper in a new area. We compare this distribution with the one that takes into account the whole set of researchers (with at least 5 publications). In this case we consider when researchers address a new subfield, even if not as pioneers.

[1]This number is given by the sum of innovators (2.5%) and early adopters (13.5%).

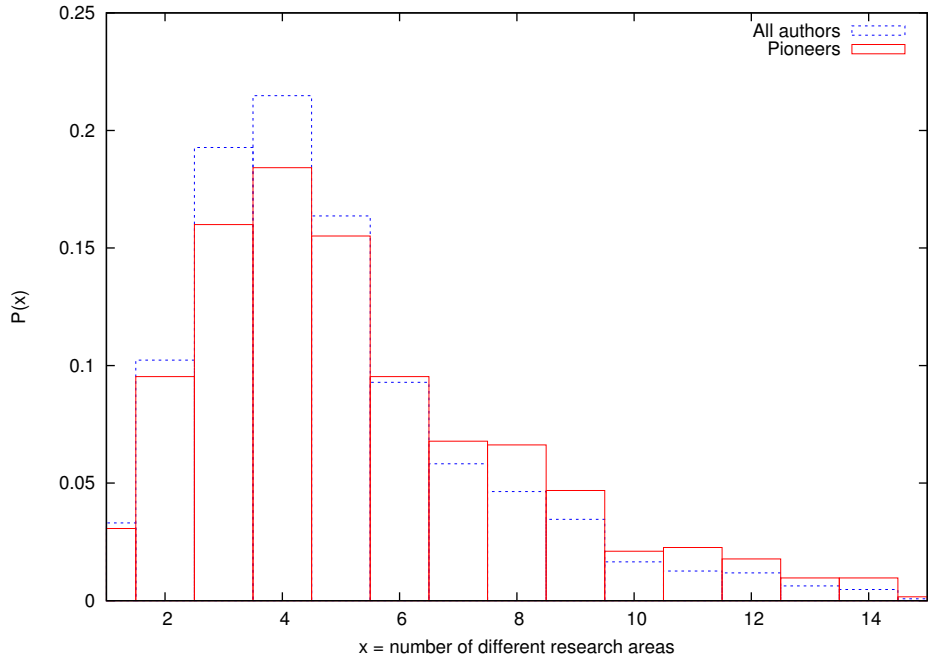



Figure 8.8: Proportion of "pioneers" experts in a given number of subfields compared to all the other researchers in the corpus.

We observe that about 25% of pioneers have just entered the field of computational linguistics when they introduce an innovation. The corresponding percentage for researchers entering a new subfield in general is about 15%. As already observed for the method space, this difference indicates a larger fraction of pioneers tends to be constituted by newcomers in the global thematic space with respect to researchers in general, which can both mean either that they are young researchers, or that innovations are rather introduced by researchers coming from different fields (who do not have earlier publications in the ACL Anthology).

Moreover, we also notice that pioneers tend to be innovative in early stages of their careers: about 60% of them had published only a third of their total production when they introduced an innovation, whereas 'only' 40% of researchers in general were at that stage when they published for the first time in a same given new field.

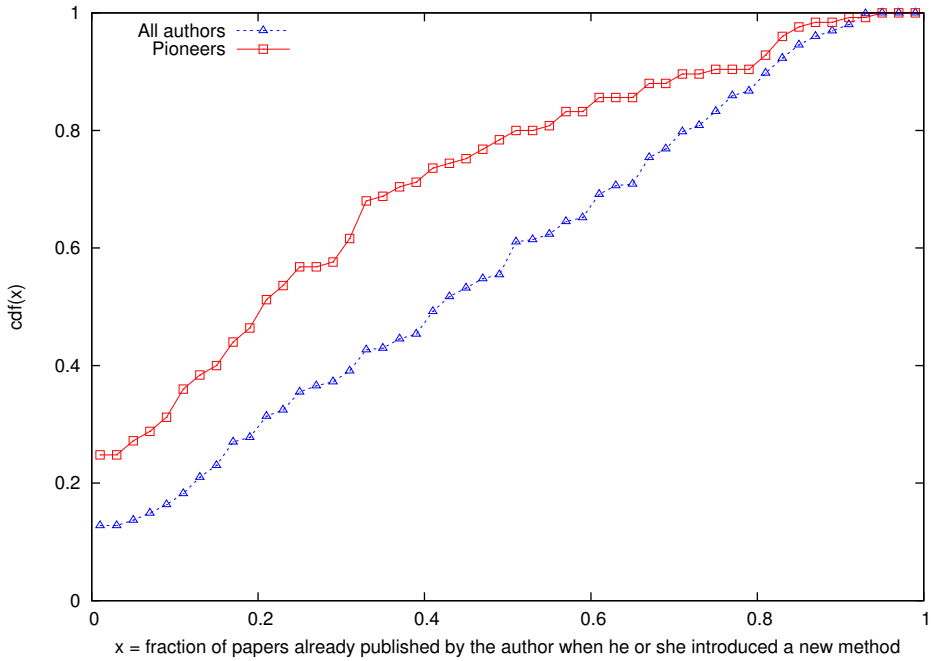



Figure 8.9: Cumulative distribution of the number of papers already published by "pioneers" (red squares), and by all researchers (blue triangles), when publishing their paper in the new subfield, compared to the total production during their career.

We note that the results obtained in this section for the dynamics of researchers in the semantic space are consistent with the results obtained in the previous section for the dynamics in the method space. This seems to indicate that our results are quite robust.

Finally, we explore the flow of researchers from one research area to another over time, as we did for the methods. Figures from 8.10 to 8.12 show a high-level representation of the semantic network shown in Figure 7.1. In this representation every node corresponds to one of the communities of concepts detected with the Infomap algorithm, as described in Chapter 7.1. The node label is given by the two most central terms of each community. Directed links between nodes represent the flow of researchers from an area of research to another, as described in the previous section for the methods.

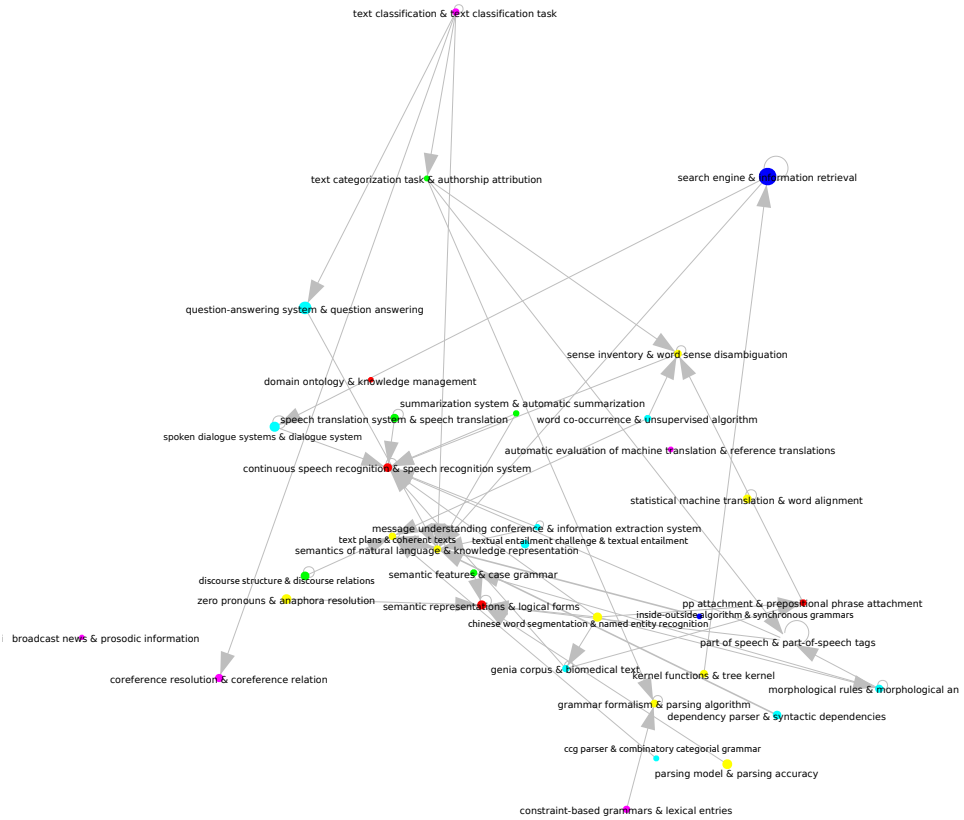



Figure 8.10: Network of researcher flows between research areas from the 1980s to the 1990s.

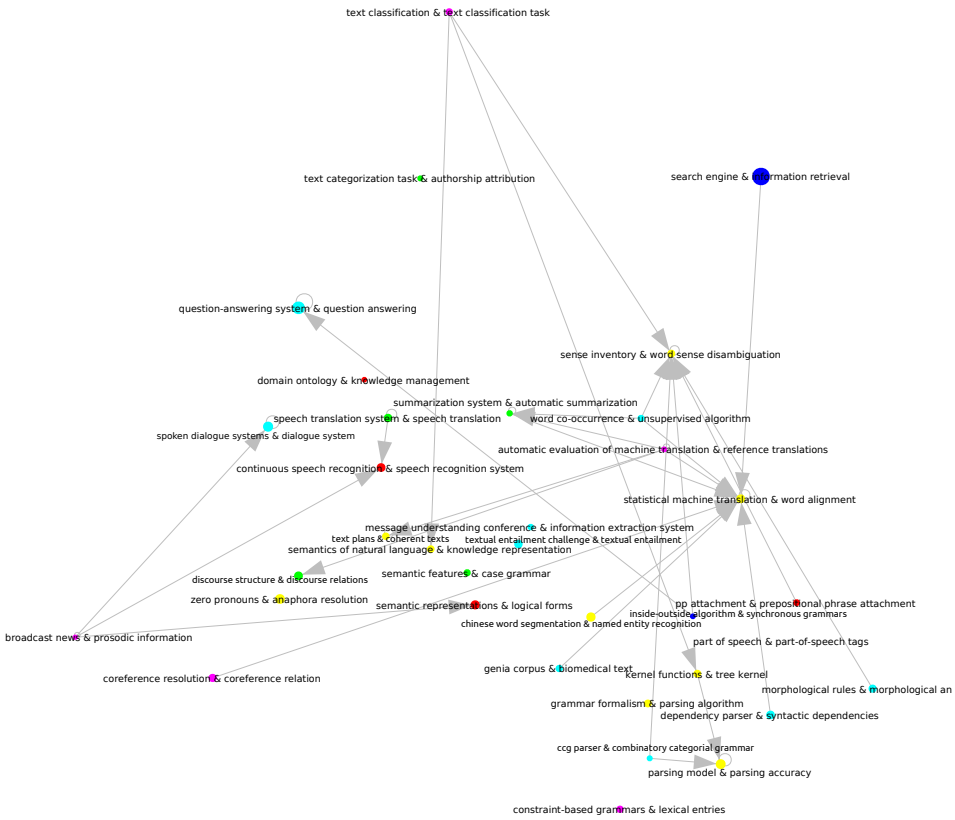



Figure 8.11: Network of researcher flows between research areas. Same as Figure 8.10, except that the period of time considered is from the 1990s to the first half of the 2000s.

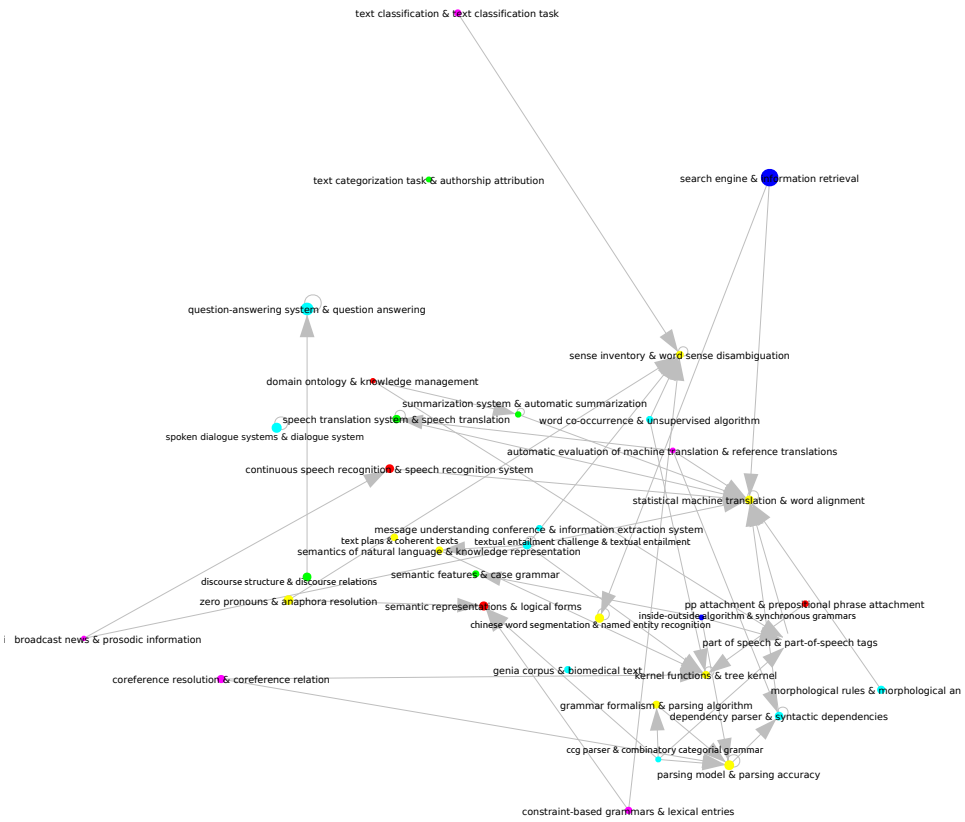



Figure 8.12: Network of researcher flows between research areas. Same as Figure 8.10 and 8.11, except that the period of time considered is from the first to the second half of the 2000s.

We observe that during the 1990s many researchers moved to Word Sense Disambiguation, to Speech Recognition, and to Semantics and Knowledge representation (Figure 8.10). From the 1990s to the first half of the 2000s, the flow towards Word Sense Disambiguation is still important, but researchers massively move to the area of Machine Translation too (Figure 8.11). From the first to the second half of the 2000s researchers keep moving to this very active area of research, but to Parsing too, which is also a strongly expanding field when considering the flow structure in the method space, as shown in the previous section (Figure 8.12).

Lastly, we investigate if there is a correlation between these flows of researchers from an area of research to another, and the strength of the semantic links connecting the two areas. The presence of a semantic link between two subfields indicates that there are publications in which concepts belonging to both areas have been used in the title or in the abstract. Therefore these links represent the presence of work at the frontier of two subfields. The flow of researchers represents instead the number of researchers that during the course of their carrier moved from one research area to another. To track this flow, we assign every paper its main subfield, by checking to which semantic community belong most of the terms of its title and abstract (as detailed in Chapter 7). We count the number of researchers that published papers in the research area $A$ at time $t$ and in the research area $B$ at time $t + \Delta t$ (where $\Delta t$ is not fixed, but depends on the interval between the publications of each researcher). We then compute the Pearson's $r$ correlation coefficient between the following two measures. For every pair of research areas $A$ and $B$, we compute the following scores:

i) the total strength of the semantic links between them, defined as:

$$s_{AB} = \sum_{(i,j) \in E^{(A,B)}} w_{ij} \tag{8.1}$$

where $E^{(A,B)}$ is the set of links connecting a concept $i \in A$ with a concept $j \in B$ in the semantic network, and $w_{ij}$ is the weight of the link between $i$ and $j$, namely the number of publications in which $i$ and $j$ co-occur;

ii) the total flow between the two areas, namely the sum of the number of researchers who published in $A$ at time $t$ and in $B$ at time $t + \Delta t$, and the number of researchers who did the opposite (i.e. we consider the flow in the two directions, since the semantic links are undirected).

We then compute the Pearson's $r$ coefficient between the vectors $S$ and $F$ containing, for each pair of research areas $A$ and $B$, the strength $s_{AB}$ and

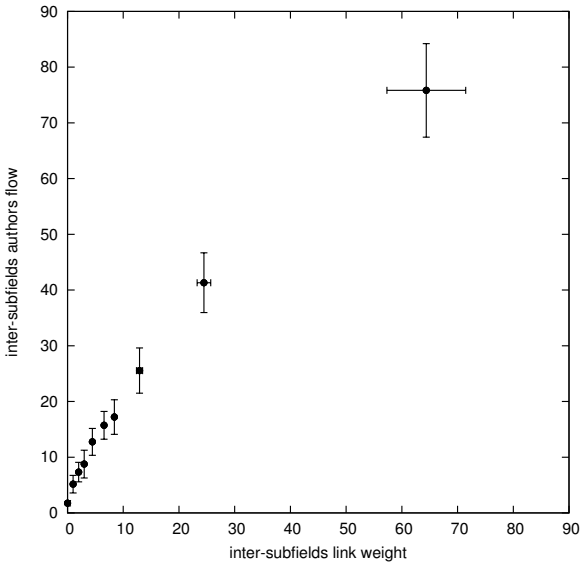



Figure 8.13: Correlation between the strength of inter-community semantic links and the flow of researchers across the corresponding communities.

the flow $f_{AB}$, respectively:

$$r = \frac{\sum_{i=1}^{n}(S_i - \bar{S})(F_i - \bar{F})}{\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}\sqrt{\sum_{i=1}^{n}(F_i - \bar{F})^2}} \tag{8.2}$$

As expected, we find a positive correlation, namely $r = 0.81$, with p-value $\ll 0.01$, which indicates that the correlation is indeed statistically significant. Figure 8.13 shows visually this correlation through a scatter plot of the two vectors, in which we have grouped the values into equally populated quantiles to make it more readable.

This result shows that there is a connection between the two levels of description that we have adopted in this thesis: the microscopic (researcher individual trajectories) and the mesoscopic level (semantic aggregates representing the different research areas). Most importantly, it represents a further confirmation of the high degree of interrelatedness between the two dimensions through which in this thesis we analyze science: the semantic dimension concerning knowledge emerging from scientific publications, and the human dimension concerning the researchers working in the field and producing such knowledge.

## 8.3 Conclusions

In this chapter we have investigated individual trajectories of researchers across the different areas of two distinct spaces that characterize the field of computational linguistics: the semantic space (built from every pertinent term we extracted), and the method space (restraining the term list to only methods). In particular, we have reconstructed the researcher flows between the different areas, which provide a bridge between individual dynamics (Chapter 6) and meso-level semantic dynamics (Chapter 7).

Moreover, we have explored the characteristics of "pioneers", defined as researchers introducing new techniques in the field or being the first to explore a new area of research within the field. Our results show two interesting facts. Firstly, pioneers are also the most eclectic researchers using the most diverse set of methods or belonging to a large number of research areas. Secondly, pioneers are oftentimes newcomers in the field.

Lastly, we have found a strong correlation between the flow of researchers from one area of research to another and the strength of the semantic links connecting the two corresponding areas. This final result means that the semantic space that we have reconstructed reflects actual research flows, since it shows that researchers tend to follow these semantic links when exploring new areas within their field.

# Conclusions

This thesis constitutes an attempt to describe in a novel way the evolution of scientific fields by combining methods coming from different disciplines, namely computational linguistics and network science. Computational linguistics makes it possible to extract knowledge directly from the texts of scientific publications, which contain the knowledge produced by scientific research. Network science can then be used to investigate how the different concepts present in these texts are interconnected. Moreover, in this thesis we also focus on human actors producing this knowledge, *i.e.* researchers and their connections. Ultimately, our goal was to model the socio-semantic landscape of scientific fields. To do this we focused on the empirical analysis of a particular scientific domain: the field of computational linguistics. The analysis is based on the ACL Anthology corpus, which is the largest available collection of publications in the field.

### Modeling the socio-semantic structure of scientific production

We first used natural language processing tools to uncover the knowledge produced in the field of computational linguistics by extracting terms describing scientific concepts from titles and abstract of papers in the ACL Anthology corpus. We then modeled the evolving social and semantic structure of the field by building two dynamic networks: *i)* a social network connecting researchers who have co-authored at least one paper, and *ii)* a semantic network connecting concepts (expressed by the extracted terms) that co-occur in the same titles or abstracts. We then analyzed the characteristics of these networks and found that they are both characterized by a heterogeneous distribution of the degree of the nodes and a well defined community structure.

**Uncovering the mechanisms underlying the socio-semantic dynamics of scientific production**

The characteristics of the structure of the social and semantic networks highlighted above seem to indicate that there are at least two mechanisms at play in the evolution of these network. On the one hand the heterogeneity of the degree distributions suggests that researchers (and concepts) tend to connect to other researchers (and concepts, respectively) that have high degree, a mechanism known as "preferential attachment". On the other hand, the well defined community structure suggest that researchers and concepts also tend to connect locally, *i.e.* to other researchers and concepts that are "close" in the network.

To confirm these hypothesis we investigated the evolution of the social and of the semantic network over time. To do this, we built a statistical model based on multivariate logistic regression so as to quantify the role of social and semantic factors in the emergence of new links between active researchers and between concepts. On the one hand we tested the role of the degree and of the similarity with other nodes in the network, measured as the fraction of common neighbors (which is a way of measuring how "close" two nodes are). On the other hand we also tested whether the evolution of the social network is significantly influenced by the structure of the semantic network, and vice-versa.

Our results show that, for the creation of a new social link, the three following factors play a statistically significant role: the social similarity of researchers (*i.e.* the number of co-authors they have in common), their semantic similarity (*i.e.* the number of concepts they have both addressed in the past), and their degree in the social network (*i.e.* the number of co-authors they already have). The role of each of these three factors had already been investigate in previous studies, but to our knowledge this is the first attempt at building a global model that takes them all into account at the same time. In particular we were able to show that the knowledge dimension plays a significant role in the emergence of new social links representing collaborations among scientists. Let us consider for example two researchers who have never collaborated before, who have a given degree in the social network and a

given number of common collaborators, and two other researchers with the same characteristics. Our results show that if the first two researchers have a higher semantic similarity with respect to the other two researchers, then the chances that the two first researchers will collaborate in the future are significantly higher. This means that by introducing the semantic dimension in the model we can improve the prediction on the evolution of collaboration networks, since, given pairs of researchers with the exact same characteristics in the social network, we are able to uncover which pair is more likely to collaborate in the future.

Symmetrically, we found that, in the probability of creation of a new link between two concepts, a significant role is played by their semantic similarity (*i.e.* the number of other concepts they both co-occurred with), their social similarity (*i.e.* the number of researcher that already worked on both concepts), and their degree in the semantic network (*i.e.* the number of concepts they already co-occurred with). To our knowledge, this is the first attempt to model the microscopic dynamics of a semantic network representing scientific knowledge and its evolution.

Finally, we found that also the probability that a researcher starts working on a new concept is affected by her/his degree, and by both social and semantic factors. In other words, the probability that a researcher addresses a new concept is affected by the number of her/his co-authors who have already addressed this concept in the past, as well as by the number of concepts (co-occurring with the new concept) that have been addressed by this researcher in the past.

To test the robustness of these results, we also performed these analysis on another another case study. The second scientific field analyzed is physics, for which we used the APS corpus, *i.e.* the largest collection of publications in this discipline. The results found on this field are completely consistent with the case of computational linguistics. Therefore in our opinion this constitutes an indication that they could be generalized to other scientific fields, even if this remains to be proven by other analyses.

**Characterizing a scientific field through its methods and techniques**

Computational linguistics is a branch of computer science that aims at automatically analyzing the content of text for different tasks (information extraction, automatic summarization, etc.). Therefore the methods and techniques developed for the different tasks play a fundamental role in the evolution of the field. In order to perform more fine grained analysis of this field, we developed a novel approach to automatically identify the terms referring to techniques from the abstracts. Our method combines state of the art automatic term extractions techniques with argumentative text zoning, *i.e.* the analysis of the rhetorical goal of sentences in a text. Our hypothesis is that it is possible to gain information about the semantic contributions of the terms by looking at the rhetorical value of the sentences in which they occur. In particular, terms referring to methods and techniques are most likely to appear in the part of the abstract in which the methodology is introduced.

Once we extracted the terms corresponding to methods and techniques, we investigated the characteristics of some researchers supposed to be particularly innovative, *i.e.* researchers who are among the firsts to introduce a new technique in the field, called "pioneers" in this study. We found that "pioneers" are more diverse than researchers according to the number techniques they use during their scientific production. Moreover, we found that innovations tend to be mainly brought by researchers at an early stage of their careers, or by researchers coming from close but different fields in which the method was developed and already exploited.

We also tested this hypothesis when new areas of research are created within the field. We analyzed the characteristics of the researchers that are among the first to publish in these new areas, and we obtained the same results as for the pioneers introducing new methods and techniques. This is an indication that our results seem robust.

**Visualizing the socio-semantic dynamics of scientific production**

In this thesis we also provided different visualizations of the field of computational linguistics and its evolution. Firstly, by applying network community detection algorithms and appropriate spatialization tools, we visualized the semantic network built by aggregating over all the papers in the corpus (regardless of their publication year), and highlighted the emerging communities in the network, *i.e.* the different groups of densely connected nodes. Several experts in field were contacted to evaluate the results. Their conclusions were that the identified communities generally correspond to well established sub-field of research within the field of computational linguistics.

We also built a temporal visualization of the field, which shows how the different research areas evolve over time: some gain importance, others shrinks, some are transformed, and new areas emerge, because new approaches are introduced for example. These representations show that, using automatic term extraction and co-occurrence based networks, it is possible to produce visual descriptions of scientific fields. These visualizations constitute useful tools for historians of science so as to help them characterize the different disciplines and their evolution.

We also built a second kind of map to visualize the flow of researchers from one research area to another along different time periods. Lastly, we made the same kind of map for the flow of researchers going from a technique to another. This is an alternative way of representing the evolution of a scientific field which focuses on the human actors, and its goal is to unveil and represent the dynamics of the actors in the knowledge landscape of scientific production.

**Uncovering the interplay between the social and the semantic dimensions of scientific production at different scales**

Our last analysis was to test if there is a correlation between the two dimensions investigated in this thesis: scientific concepts and researchers. The analysis of the evolution of the social and semantic networks already

showed that they are intertwined and somehow co-evolve over time. We confirmed this connection by finding a strong statistical correlation between the flow of researchers across the different areas of computational linguistics, and the semantic links connecting these areas. This result also constitutes a confirmation that the semantic space we have reconstructed reflects actual research flows, since it shows that researchers tend to follow these semantic links when exploring new areas of their field.

The different analyses performed in this thesis investigate the dynamics of scientific production at different scales. We firstly modeled the emergence of new links between researchers and between concepts by looking at the role of their local neighborhood. This analysis focused therefore on a microscopic level of description. We then analyzed the emergence of different aggregates of concepts that represent the different research areas of the field. This constitutes a mesoscopic level of description. Finally we navigated through the two levels by analyzing the dynamics of individual researchers across the different semantic aggregates, and we found a statistical correlation between this flow and the semantic links.

### Perspectives

We have seen that new natural language processing tools are needed to extract relevant information from raw textual data contained in large digital scientific archives. This made it possible to explore the different dimensions at stake, instead of focusing only on human actors and on the number of papers they have published. Moreover, once the objects of study were identified, we had to introduce mathematical models to account for their properties. To uncover in particular the properties of the relations between the objects under study at different levels we have drawn upon the network framework.

This shows that the availability of "big data" does not imply the end of models, since formalizing objects is not a straightforward task and, moreover, once these objects have been identified, simply enumerating them does not lead to any new insight, but models are needed to uncover the properties of the dynamics of their relations. As a consequence of

this effort of formalization and modeling we can investigate fine-grained facts about the evolution of scientific fields. These investigations are probably not approaching the level of subtlety of works led by historians and sociology of science yet, but the scale of the analysis, made possible thanks to automatic tools, entails the possibility to discover new facts, and new connections between facts. The role of human experts is also important in interpreting the data: it remains crucial to examine carefully the correlations exhibited by automatic analyses, extract the most meaningful ones so as to give birth to new models and new interpretations of the evolution of scientific domains.

In Chapter 8 we have for example shown that researchers who introduce methodological innovations in the field of computational linguistics are also the most diverse in terms of variety of methods used in their works, and they are often new to the field or at early stages of their career when they introduce an innovation. This is probably an interesting result *per se* but, in order to better understand the phenomenon and get accurate interpretations, a collaboration with historians and sociologists of science would be necessary.

Overall, we proposed a methodology for the investigation of scientific fields that could be used in the future to study any other discipline, provided that a representative collection of publications is available. The results of this investigation should then be taken into account by historians and sociologists of science who could provide useful interpretations and even propose further analyses.

# Appendix A

# ACL Anthology term list

Accuracy rate, acoustic models, agglutinative languages, Air Travel Information Service, Air Travel Information System, alignment algorithm, alignment error rate, alignment links, alignment method, alignment models, alignment quality, alignment system, alignment techniques, alignment template, ambiguity resolution, ambiguous sentences, anaphora resolution, Anaphoric Annotation, anaphoric expressions, annotated corpus, annotation errors, annotation projects, annotation schema, answer candidates, answer extraction, argument classification, artificial intelligence, ATIS Data, attachment score, authorship attribution, Automatic Content Extraction, Automatic Evaluation Of Machine Translation, Automatic MT Evaluation, automatic speech recognition, automatic speech recognition systems, Automatic Summarization, automatic text summarization, baseline system, Bayesian classifier, Bayesian model, bilingual corpora, bilingual corpus, bilingual dictionary, bilingual resources, bilingual texts, biomedical domain, biomedical text, BLEU score, broadcast news, Brown corpus, case frames, case grammar, Categorial Grammar, ccg parser, Centering Theory, chart parser, chart parsing, Chinese Named Entity Recognition, Chinese Word Segmentation, Chinese word segmentation system, Chinese Word Sense Disambiguation, Chinese-to-English translation, clarification dialogue, classification task, classification techniques, Clause Grammar, clustering algorithm, coherence relations, coherent texts, Collocation Extraction, Combinatory Categorial Grammar, comparable corpora, complex morphology, compositional semantics, compound nominals, compound nouns, Computational Complexity, conditional probabilities, Conditional Random Fields, confidence estimation, confidence measures, confidence scores, confusion network, connectionist models, constraint propagation, constraint satisfaction,

Constraint-Based Grammars, Construction Grammar, context free grammar, context vectors, context-free grammars, context-free languages, Continuous Speech Recognition, continuous speech recognition system, continuous speech recognizer, conversational systems, convolution tree kernel, coreference relation, coreference resolution, coreference resolution system, coreference systems, corpus statistics, corpus-based method, data representation, data sparseness, data sparseness problem, data-driven approach, database management, decision tree, decision tree classifier, decoding algorithm, Definite Clause Grammars, definite descriptions, definition questions, dependency accuracy, dependency analysis, Dependency Grammar, dependency graphs, dependency parser, dependency relations, dependency representation, dependency structure analysis, dependency structures, dependency trees, derivation trees, Description Logics, dialog management, dialog systems, dialogue act, dialogue corpora, dialogue corpus, dialogue interaction, dialogue management, dialogue model, dialogue system, dictionary definitions, dictionary lookup, disambiguation methods, discourse analysis, discourse annotation, discourse coherence, discourse connectives, discourse entities, discourse markers, discourse model, discourse referents, discourse relations, Discourse Representation, Discourse Representation Theory, discourse segmentation, discourse structure, discriminative methods, discriminative model, Disjunctive Feature Structures, distributional similarity, distributional similarity measures, document retrieval, document summarization, Document Understanding Conference, domain adaptation, Domain Models, domain ontology, domain-specific knowledge, electronic dictionary, ellipsis resolution, empirical methods, Encyclopedic Knowledge, entity detection, error rate, error rate reduction, evaluation methods, evaluation metrics, event detection, event extraction, Example-Based Machine Translation, expert system, Extended Domain Of Locality, extraction patterns, extractive summaries, extractive summarization, extrinsic evaluation, factoid questions, feature selection, feature selection methods, finite-state morphology, finite-state transducers, Formal Semantics, frame semantics, free word order, Free Word Order Languages, gene names, Gene Ontology, Generalized Phrase Structure Grammar, generation of referring expressions, Generative Lexicon, generative model, generative probabilistic model, GENIA corpus, Government-Binding Theory, grammar development, grammar formalism, grammar induction, grammatical analysis, grammatical errors, grammatical formalisms, Graph unification, Head-driven Phrase Structure Grammar, Hidden Markov Models, HPSG grammar, human evaluation, human judgments, human-computer interaction, IBM Model, IE system, incremental generation, incremental interpretation, incremental parser, Indian languages, Inductive Logic, inference rules, information access, information extraction, information extraction system, information extraction task, information retrieval, information retrieval systems, information retrieval

techniques, information structure, inheritance hierarchy, Inside-Outside algorithm, instructional texts, Intelligent Tutoring Systems, intensional logic, inter-annotator agreement, Inversion Transduction Grammar, kernel functions, kernel methods, knowledge base, Knowledge Extraction, knowledge management, knowledge representation, knowledge sources, Knowledge-Based Machine Translation, label propagation, Lambek Calculus, language identification, language model, language resources, language understanding, language understanding systems, languages with scarce resources, large vocabulary continuous speech recognition, large-vocabulary continuous speech recognition, lexical ambiguity, lexical categories, Lexical Chains, lexical cohesion, Lexical Conceptual Structure, lexical entries, lexical features, Lexical Functional Grammar, lexical knowledge base, lexical relations, lexical representations, lexical rules, lexical selection, lexical semantics, lexical transfer, Lexicalized Tree Adjoining Grammars, lexico-syntactic patterns, lexicon model, linear order, linguistic knowledge, linguistic patterns, linguistic theory, location names, Logic Grammars, logic programming, logical forms, LR parsing, machine learning approach, machine learning system, machine learning techniques, machine readable dictionaries, machine translation, Machine Translation Evaluation, machine translation models, machine translation output, machine translation project, machine translation quality, machine translation system, machine translation task, Machine Transliteration, machine-readable dictionary, Mandarin Broadcast News, Markov Model, Maximum Entropy, Maximum Entropy Approach, maximum entropy classifier, Maximum Entropy Framework, maximum entropy model, Maximum Likelihood, meaning representations, Message Understanding, Message Understanding Conference, Minimum Description Length, monolingual corpora, Montague grammar, morphological analysis, morphological analyzer, morphological disambiguation, morphological rules, MT evaluation, MT quality, MT systems, multi-document summaries, multi-document summarization, multi-word expressions, Multilingual Entity Task, multimodal dialogue systems, multimodal interaction, Multiword Expressions, multiword units, mutual information, n-gram language model, n-gram model, Naive Bayes, named entities, Named Entity Recognition, named entity recognition system, Named Entity Recognizer, named-entity recognition, Natural Language Access, natural language dialogue, Natural Language Generation, natural language generation system, natural language grammars, natural language interfaces, natural language questions, natural language understanding system, NE recognition, NER system, neural network, NL generation, NLG system, non-projective dependency, normal form, ontology construction, Ontology Population, open-domain question answering system, out-of-vocabulary words, parallel corpora, parallel corpus, parallel sentences, parallel texts, parallel treebanks, paraphrase acquisition, parse forest, parse selection, parse time, parse trees, parser accuracy, parser evaluation, parser

performance, parsing accuracy, parsing algorithm, parsing model, parsing process, parsing strategy, parsing system, part of speech, part-of-speech tags, PCFG parser, Penn Chinese Treebank, Penn Discourse Treebank, Penn Tree-bank, Penn Treebank, person names, phonological rules, phrase alignment, phrase-based SMT, phrase-based SMT system, pitch accent, pivot language, polynomial time, polysemous word, POS tagged corpus, POS tagger, POS tags, POS tagset, PP attachment, predicate-argument structure, prediction accuracy, Predominant Senses, Prepositional Phrase Attachment, Princeton WordNet, probabilistic context-free grammars, probabilistic parser, probability estimates, pronominal anaphora, pronoun resolution, proper nouns, prosodic information, prosodie information, protein-protein interactions, QA system, quantifier scope, query translation, question answering, question answering system, question answering task, Question Answering track, question types, question-answering system, ranking algorithm, recognition errors, reference translations, referring expressions, relation extraction, relation extraction system, relation extraction task, reordering models, research in information extraction, Resource Management task, Rhetorical Relations, rhetorical structure, Rhetorical Structure Theory, rich morphology, robust speech recognition, role labeling system, Romance languages, search engine, segment boundaries, segmentation model, semantic analysis, semantic classes, semantic constraints, semantic dependencies, semantic features, Semantic Inference, semantic information, semantic interpretation, semantic knowledge, semantic labels, semantic lexicon, semantic network, semantic relatedness, semantic relatedness measures, semantic relations, semantic representations, Semantic Role, semantic role assignment, semantic role labeling system, semantic role labels, semantic similarity, semantic structure, semantic tags, semantics of natural language, Semitic languages, sense distinctions, sense inventory, sense-tagged data, sentence alignment, sentence boundaries, Sentence Compression, sentence extraction, sentence generation, sentence length, sentence pairs, sentence realization, shared task, similarity measures, similarity metrics, Singular Value Decomposition, situation semantics, size of training data, SMT system, source language, speaker adaptation, speech acts, speech corpus, Speech Generation, speech input, speech recognition, speech recognition component, speech recognition errors, speech recognition hypotheses, speech recognition output, speech recognition system, speech recognition technology, speech recognizer output, speech synthesis, speech tagger, speech translation, speech translation system, speech understanding system, speech-to-speech translation, speech-to-speech translation systems, spelling correction, spoken dialog systems, spoken dialogue corpus, Spoken Dialogue Interface, Spoken Dialogue Systems, Spoken Language, Spoken Language Systems, Spoken Language Translation, spoken language understanding systems, SRL system, statistical language models, Statistical Machine Translation, statistical machine translation models, statistical

machine translation system, statistical model, Statistical MT, statistical MT systems, statistical parser, statistical parsing model, statistical translation models, statistical word alignment, stochastic context-free grammar, stochastic language models, stochastic taggers, structural ambiguity, structural descriptions, Subcategorization Acquisition, subcategorization frames, summarization evaluation, summarization method, summarization system, summarization task, summarization techniques, summary generation, Support Vector Machines, surface realization, Switchboard corpus, synchronous context-free grammars, synchronous grammars, Synchronous TAGs, syntactic ambiguity, syntactic analyzer, syntactic annotation, syntactic constructions, syntactic dependencies, syntactic disambiguation, syntactic features, syntactic information, syntactic parse trees, syntactic parser, syntactic structure, syntactic structure of sentences, syntactic trees, syntax-based machine translation, systemic grammars, tagged corpora, tagging accuracy, target language model, task-oriented dialogues, temporal expressions, temporal information, temporal relations, temporal structure, term candidates, Term Extraction, term frequency, term translation, test suite, text categorization, text categorization task, text classification, text classification task, text generation, text generation system, text genres, text interpretation, text planner, text plans, Text Segmentation, text structure, text summarization, Text Summarization Challenge, text summarization system, text type, text understanding, Text Understanding System, text-to-speech systems, textual entailment, Textual Entailment Challenge, textual inference, TIPSTER Program, Tipster project, TIPSTER Text, TIPSTER Text Program, topic identification, topic information, topic segmentation, tourism domain, training data, transfer phase, transfer rules, transformation rules, translation accuracy, translation candidates, translation equivalents, translation model, translation patterns, translation process, translation quality, translation results, translation task, translation units, transliteration model, Tree Adjoining Grammar, Tree Adjoining Languages, tree kernel, tree-adjoining grammars, trigram language model, tutoring system, two-level morphology, typed feature structures, Unification Categorial Grammar, unification grammars, unification-based formalisms, unlabeled data, unsupervised algorithm, unsupervised learning method, Unsupervised Word Sense Disambiguation, user interface, user model, user queries, user satisfaction, user simulation, user utterances, verb senses, Viterbi alignment, Wall Street Journal corpus, Web Corpora, Web corpus, web search engine, wide-coverage grammar, word alignment, word alignment methods, word alignment models, word alignment systems, word associations, word boundaries, word co-occurrence, word distributions, word error rate, word formation, Word Identification, word lattice, word meanings, word order, word segmentation, word segmentation algorithm, Word Segmentation Bakeoff, word segmentation performance, word segmentation system, word

sense, Word Sense Disambiguation, word sense disambiguation algorithms, word sense disambiguation methods, word sense disambiguation system, word sense disambiguation task, Word Sense Discrimination, Word Sense Induction, word similarity, word-sense disambiguation, WordNet senses, WordNet synsets, WSD accuracy, WSD system, Zero Pronouns

# Appendix B

# ACL Anthology semantic clusters

Multilingual Entity Task - TIPSTER Text Program - event extraction - Message Understanding - TIPSTER Program - research in information extraction - information extraction system - Message Understanding Conference - Tipster project - IE system - information extraction task

natural language questions - open-domain question answering system - Question Answering track - question-answering system - answer candidates - question answering - factoid questions - web search engine - answer extraction - definition questions - question types - QA system - question answering task

alignment template - languages with scarce resources - SMT system - parallel texts - phrase alignment - lexical transfer - translation results - translation model - translation accuracy - bilingual dictionary - word alignment methods - word alignment - statistical machine translation system - synchronous context-free grammars - source language - machine translation task - monolingual corpora - parallel corpora - bilingual corpora - Chinese-to-English translation - alignment quality - BLEU score - word alignment models - parallel corpus - machine translation models - bilingual texts - comparable corpora - pivot language - bilingual resources - alignment links - confusion network - sentence alignment - decoding algorithm - transfer phase - phrase-based SMT system - Inversion Transduction Grammar - alignment error rate - translation process - sentence length - translation task - target language model - IBM Model - Statistical Machine Translation - transfer rules - multiword units - translation units - baseline system - statistical word alignment - syntax-based machine translation -

statistical model - bilingual corpus - Viterbi alignment - sentence pairs - alignment algorithm - Example-Based Machine Translation - MT quality - alignment system - translation quality - machine translation - translation equivalents - lexicon model - Statistical MT - word alignment systems - alignment models - statistical translation models - parallel sentences - MT systems - alignment techniques - phrase-based SMT - alignment method - machine translation system - reordering models - speech recognition hypotheses

pronominal anaphora - inter-annotator agreement - Zero Pronouns - Anaphoric Annotation - pronoun resolution - discourse entities - anaphora resolution

unsupervised algorithm - word co-occurrence - clustering algorithm - Lexical Chains

Semantic Role - SRL system - role labeling system - convolution tree kernel - syntactic features - predicate-argument structure - maximum entropy classifier - kernel methods - semantic role labeling system - syntactic information - syntactic parse trees - kernel functions - tree kernel - relation extraction - relation extraction task - argument classification

semantic representations - Discourse Representation - Discourse Representation Theory - semantic information - semantic interpretation - incremental generation - ambiguity resolution - ambiguous sentences - compositional semantics - lexical ambiguity - semantic knowledge - logical forms - Formal Semantics - syntactic structure - syntactic ambiguity - Description Logics

information access - tourism domain - knowledge management - domain-specific knowledge - ontology construction - domain ontology

stochastic taggers - tagging accuracy - POS tagset - speech tagger - part-of-speech tags - POS tags - part of speech

coherent texts - sentence generation - text generation - Lexical Conceptual Structure - text plans - constraint satisfaction - text generation system - temporal structure - instructional texts - Centering Theory - Natural Language Generation - text planner - surface realization - definite descriptions - NLG system - rhetorical structure - Rhetorical Structure Theory - referring expressions - natural language generation system - text structure - anaphoric expressions - Rhetorical Relations - generation of referring expressions

text classification - classification techniques - Naive Bayes - text classification task

entity detection - coreference relation - Automatic Content Extraction - coreference resolution - coreference resolution system - coreference systems

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

dependency graphs - domain adaptation - dependency trees - dependency relations - dependency representation - dependency analysis - semantic role labels - dependency accuracy - shared task - semantic dependencies - dependency structures - attachment score - non-projective dependency - syntactic dependencies - dependency parser

natural language understanding system - semantic lexicon - lexical knowledge base - Montague grammar - temporal expressions - lexical semantics - semantics of natural language - situation semantics - intensional logic - knowledge base - Generative Lexicon - artificial intelligence - knowledge representation - meaning representations

synchronous grammars - statistical machine translation models - polynomial time - Inside-Outside algorithm - stochastic context-free grammar

textual inference - Semantic Inference - textual entailment - Textual Entailment Challenge

Penn Treebank - parsing model - statistical parser - parser performance - generative probabilistic model - parsing accuracy - parse trees - PCFG parser - discriminative methods

search engine - query translation - information retrieval - Web Corpora

term translation - machine translation quality - Automatic MT Evaluation - MT evaluation - Automatic Evaluation Of Machine Translation - Machine Translation Evaluation - translation candidates - machine translation output - reference translations - statistical MT systems

Switchboard corpus - speech input - acoustic models - n-gram model - database management - Air Travel Information Service - Spoken Language - speech recognition - speaker adaptation - continuous speech recognition system - spoken language understanding systems - automatic speech recognition - speech recognition technology - Spoken Language Systems - large vocabulary continuous

speech recognition - speech recognition output - robust speech recognition - Hidden Markov Models - Continuous Speech Recognition - word error rate - information retrieval techniques - Air Travel Information System - language understanding - language understanding systems - speech recognition errors - speech recognition system - language model - ATIS Data - recognition errors - continuous speech recognizer - error rate - speech corpus - speech understanding system - stochastic language models - broadcast news - Resource Management task - dialog management - prosodie information - automatic speech recognition systems

morphological analysis - complex morphology - finite-state morphology - POS tagged corpus - morphological disambiguation - morphological rules - agglutinative languages - morphological analyzer - Semitic languages - POS tagger

Gene Ontology - annotated corpus - protein-protein interactions - GENIA corpus - biomedical text - gene names - machine learning system - biomedical domain

machine learning techniques - text categorization - language identification - text categorization task - text genres - feature selection - size of training data - authorship attribution

Spoken Language Translation - speech synthesis - speech translation system - speech-to-speech translation systems - speech-to-speech translation - speech translation

person names - NER system - word segmentation performance - Support Vector Machines - Markov Model - word segmentation system - confidence measures - named entity recognition system - word segmentation algorithm - Maximum Entropy Framework - Word Segmentation Bakeoff - Named Entity Recognizer - out-of-vocabulary words - Conditional Random Fields - Chinese word segmentation system - segmentation model - NE recognition - Chinese Word Segmentation - word segmentation - maximum entropy model - discriminative model - location names - Chinese Named Entity Recognition - Maximum Entropy Approach - Named Entity Recognition - Maximum Entropy - Word Identification - word boundaries

Speech Generation - Penn Discourse Treebank - discourse annotation - discourse relations - discourse structure - annotation schema - coherence relations - discourse model - discourse coherence - discourse segmentation - discourse referents - discourse connectives - annotation projects - discourse markers

decision tree classifier - topic segmentation - prosodic information - Mandarin Broadcast News

structural descriptions - parser accuracy - wide-coverage grammar - ccg parser - lexical categories - Free Word Order Languages - Combinatory Categorial Grammar

sense inventory - word sense - semantic relatedness - semantic relations - similarity metrics - word sense disambiguation system - sense-tagged data - similarity measures - sense distinctions - Unsupervised Word Sense Disambiguation - Web corpus - word sense disambiguation algorithms - Predominant Senses - Word Sense Disambiguation - word sense disambiguation task - Princeton WordNet - corpus statistics - WSD system - dictionary definitions - lexico-syntactic patterns - distributional similarity measures - WordNet senses - distributional similarity - word distributions - machine-readable dictionary - frame semantics - word sense disambiguation methods - Word Sense Discrimination - context vectors - feature selection methods - Chinese Word Sense Disambiguation - word similarity - Word Sense Induction - semantic similarity - polysemous word - lexical relations - WordNet synsets

typed feature structures - Lexical Functional Grammar - Constraint-Based Grammars - grammar development - Disjunctive Feature Structures - lexical entries - Graph unification - Head-driven Phrase Structure Grammar - lexical rules - HPSG grammar - inheritance hierarchy

verb senses - semantic features - semantic structure - case grammar - case frames - disambiguation methods

Clause Grammar - Dependency Grammar - unification-based formalisms - chart parser - parsing system - unification grammars - normal form - Unification Categorial Grammar - Computational Complexity - Extended Domain Of Locality - natural language grammars - linguistic theory - Lexicalized Tree Adjoining Grammars - systemic grammars - Synchronous TAGs - tree-adjoining grammars - parsing algorithm - constraint propagation - Generalized Phrase Structure Grammar - Tree Adjoining Languages - context free grammar - rich morphology - chart parsing - parsing strategy - Categorial Grammar - LR parsing - derivation trees - probabilistic context-free grammars - free word order - parsing process - probabilistic parser - Definite Clause Grammars - context-free languages - word order - Tree Adjoining Grammar - grammar formalism - grammatical formalisms - Lambek Calculus - parse time

structural ambiguity - syntactic disambiguation - connectionist models - Prepositional Phrase Attachment - semantic classes - PP attachment

dialog systems - speech acts - dialogue interaction - tutoring system - multimodal dialogue systems - user utterances - user simulation - user satisfaction - dialogue

management - dialogue corpus - dialogue model - dialogue system - dialogue act - multimodal interaction - dialogue corpora - natural language dialogue - task-oriented dialogues - speech recognition component - spoken dialogue corpus - Spoken Dialogue Systems - user model - spoken dialog systems

# Bibliography

Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.

Ashton Anderson, Dan Jurafsky, and Daniel A. McFarland. 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, South Korea, 2012. Association for Computational Linguistics.

Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete, 2008.

A. L Barabási, H Jeong, Z Néda, E Ravasz, A Schubert, and T Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3–4):590–614, 2002.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2): 101–113, 2004.

A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proc Natl Acad Sci U S A*, 101(11):3747–3752, 2004.

Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

Luís M. A. Bettencourt, David I. Kaiser, and Jasleen Kaur. Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, 3(3):210–221, 2009.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. "O'Reilly Media, Inc.", 2009.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008(10):P10008, 2008.

Béla Bollobás. Degree sequences of random graphs. *Discrete Mathematics*, 33(1):1–19, 1981.

Didier Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 3*, COLING '92, pages 977–981, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

Kevin W. Boyack and Katy Börner. Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers. *J. Am. Soc. Inf. Sci.*, 54(5):447–461, 2003.

Kevin W. Boyack, Katy Börner, and Richard Klavans. Mapping the structure and evolution of chemistry research. *Scientometrics*, 79(1): 45–60, 2009.

Danah Boyd and Kate Crawford. Six provocations for big data. SSRN Scholarly Paper ID 1926431, Social Science Research Network, Rochester, NY, 2011.

Eric Brill. Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, AAAI '94, pages 722–727, Menlo Park, CA, USA, 1994. American Association for Artificial Intelligence.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media, 2002.

Katy Börner. *Atlas of Science: Visualizing What We Know*. The MIT Press, Cambridge, Mass, 2010.

Katy Börner, Chaomei Chen, and Kevin W. Boyack. Visualizing knowledge domains. *Ann. Rev. Info. Sci. Tech.*, 37(1):179–255, 2003.

Michel Callon and Michel Ferrary. Les réseaux sociaux à l'aune de la théorie de l'acteur-réseau. *Sociologies pratiques*, 13(2):37–44, 2006.

Michel Callon, John Law, and Arie Rip. *Mapping the dynamics of science and technology: sociology of science in the real world*. Macmillan, 1986.

Michel Callon, Jean-Pierre Courtial, and Françoise Laville. 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, 1991.

Alberto Cambrosio, Peter Keating, Simon Mercier, Grant Lewison, and Andrei Mogoutov. Mapping the emergence and development of translational cancer research. *European Journal of Cancer*, 42(18): 3140–3148, 2006.

David Chavalarias and Jean-Philippe Cointet. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PLoS ONE*, 8(2): e54847, 2013.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, 1990a.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, 1990b.

James Samuel Coleman, Elihu Katz, Herbert Menzel, and Columbia University Bureau of Applied Social Research. *Medical innovation: a diffusion study*. Bobbs-Merrill Co., 1966.

James Curran, Stephen Clark, and Johan Bos. Linguistically motivated large-scale NLP with c&c and boxer. In *Proceedings of the 45th Meeting of the Association for Computation Linguistics (ACL)*, pages 33–36, 2007.

Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING '94, pages 515–521, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

Béatrice Daille, Benoît Habert, Christian Jacquemin, and Jean Royauté. Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–257, 1996.

Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74, 1993.

Nees Jan van Eck. *Methodological Advances in Bibliometric Mapping of Science*. PhD thesis, Erasmus Research Institute of Management (ERIM), 2011.

Nees Jan van Eck and Ludo Waltman. How to normalize cooccurrence data? an analysis of some well-known similarity measures. *J. Am. Soc. Inf. Sci.*, 60(8):1635–1651, 2009.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486 (3–5):75–174, 2010.

Christopher Fox. A stop list for general text. *SIGIR Forum*, 24(1-2):19–21, 1989.

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

Koen Frenken, Sjoerd Hardeman, and Jarno Hoekman. Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, 3(3):222–232, 2009.

Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.

Eugene Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Isi Press, 1979.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–7826, 2002.

Sebastian Grauwin, Guillaume Beslon, Éric Fleury, Sara Franceschelli, Celine Robardet, Jean-Baptiste Rouquier, and Pablo Jensen. Complex systems science: Dreams of universality, interdisciplinarity reality. *J Am Soc Inf Sci Tec*, 63(7):1327–1338, 2012.

Jerrold W. Grossman. The evolution of the mathematical research collaboration graph. *Congressus Numerantium*, pages 201–212, 2002.

Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 491–501, New York, NY, USA, 2004. ACM.

Jean-Loup Guillaume and Matthieu Latapy. A realistic model for complex networks. *arXiv:cond-mat/0307095*, 2003. arXiv: cond-mat/0307095.

Jean-Loup Guillaume and Matthieu Latapy. Bipartite structure of all complex networks. *Information Processing Letters*, 90(5):215–221, 2004.

R. Guimerà, S. Mossa, A. Turtschi, and L. a. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *PNAS*, 102(22):7794–7799, 2005a.

Roger Guimerà, Brian Uzzi, Jarrett Spiro, and Luís A. Nunes Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005b.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283, Edinburgh, 2011.

Yufan Guo, Roi Reichart, and Anna Korhonen. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 928–937, 2013.

Mark Herrera, David C. Roberts, and Natali Gulbahce. Mapping the evolution of scientific fields. *PLoS ONE*, 5(5):e10355, 2010.

J. E. Hirsch. An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569–16572, 2005.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519 (3):97–125, 2012.

Paul Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50, 1912.

H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804): 651–654, 2000.

Aravind K. Joshi and Yves Schabes. Tree-adjoining grammars. In Prof Dr Grzegorz Rozenberg and Prof Dr Arto Salomaa, editors, *Handbook of Formal Languages*, pages 69–123. Springer Berlin Heidelberg, 1997.

John S. Justeson and Slava M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(01):9–27, 1995.

Kyo Kageura and Bin Umino. Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289, 1996.

Elihu Katz and Paul Felix Lazarsfeld. *Personal Influence, the Part Played by People in the Flow of Mass Communications*. Transaction Publishers, 1970.

M. M. Kessler. Bibliographic coupling between scientific papers. *Amer. Doc.*, 14(1):10–25, 1963.

S. C. Kleene. Representation of events in nerve nets and finite automata. Technical report, 1951.

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

John Lafferty, Andrew Mccallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80(5):056117, 2009.

Andrea Lancichinetti, Filippo Radicchi, José J. Ramasco, and Santo Fortunato. Finding statistically significant communities in networks. *PLoS ONE*, 6(4):e18961, 2011.

Bruno Latour. *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press, 1987.

Bruno Latour. Reassembling the social - an introduction to actor-network-theory. *Oxford University Press*, -1, 2005.

Bruno Latour and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, 1979.

Emmanuel Lazega, Marie-Thérèse Jourda, Lise Mounier, and Rafaël Stofer. Catching up with big fish in the big pond? multi-level network analysis through linked design. *Social Networks*, 30(2):159–176, 2008.

Gary Geunbae Lee, Jong-Hyeok Lee, and Jeongwon Cha. Syllable-pattern-based unknown-morpheme segmentation and estimation for hybrid part-of-speech tagging of korean. *Computational Linguistics*, 28(1):53–70, 2002.

Loet Leydesdorff. On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *J. Am. Soc. Inf. Sci.*, 59(1):77–85, 2008.

Loet Leydesdorff and Staša Milojević. Scientometrics. *arXiv:1208.4566 [cs]*, 2012. arXiv: 1208.4566.

Loet Leydesdorff and Ismael Rafols. A global map of science based on the ISI subject categories. *J. Am. Soc. Inf. Sci.*, 60(2):348–362, 2009.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci.*, 58(7):1019–1031, 2007.

Linyuan Lü and Tao Zhou. Link prediction in weighted networks: The role of weak ties. *EPL*, 89(1):18001, 2010.

Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Travis Martin, Brian Ball, Brian Karrer, and M. E. J. Newman. Coauthorship and citation patterns in the physical review. *Phys. Rev. E*, 88(1):012814, 2013.

Y. Matsuo and M. Ishizuka. KEYWORD EXTRACTION FROM a SINGLE DOCUMENT USING WORD CO-OCCURRENCE STATISTICAL INFORMATION. *Int. J. Artif. Intell. Tools*, 13(01):157–169, 2004.

Diana Maynard and Sophia Ananiadou. Identifying terms by their family and friends. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 530–536, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability 37. Chapman & Hall/CRC, 1989.

Stanley Milgram. The small world problem. *Psychology Today*, 1:61–67, 1967.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75(6):468–487, 2006.

Peter R. Monge and Noshir S. Contractor. *Theories of Communication Networks*. Oxford University Press, 2003.

Geoffrey A. Moore. *Crossing the Chasm: Marketing and Selling Disruptive Products to Mainstream Customers*. HarperCollins, 2002.

Lev Muchnik, Sen Pei, Lucas C. Parra, Saulo D. S. Reis, José S. Andrade Jr, Shlomo Havlin, and Hernán A. Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Sci. Rep.*, 3, 2013.

Nicholas C. Mullins. The development of a scientific specialty: The phage group and the origins of molecular biology. *Minerva*, 10(1):51–82, 1972.

M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64(2):025102, 2001a.

M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64(1):016131, 2001b.

M. E. J. Newman. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1):016132, 2001c.

M. E. J. Newman. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, 2001d.

M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101(suppl 1):5200–5205, 2004.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.

Mark Newman. *Networks: An Introduction*. Oxford University Press, Oxford ; New York, 1 edition edition, 2010.

Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami, and Toshihisa Takagi. Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161, 2001.

Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

Raj Kumar Pan, Sitabhra Sinha, Kimmo Kaski, and Jari Saramäki. The evolution of interdisciplinarity in physics research. *Sci. Rep.*, 2, 2012.

Vilfredo Pareto. *Cours d'Économie politique*. Librairie Droz, 1964.

Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, 2007.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. Terminology extraction: An analysis of linguistic and statistical approaches. In Dr Spiros Sirmakessis, editor, *Knowledge Mining*, number 185 in Studies in Fuzziness and Soft Computing, pages 255–279. Springer Berlin Heidelberg, 2005.

Karl Popper. *The Logic of Scientific Discovery*. Routledge, 1959.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007.

Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683): 510–515, 1965.

Derek John de Solla Price. *Little Science, Big Science– and Beyond*. Columbia University Press, 1963.

Roi Reichart and Anna Korhonen. Document and corpus level inference for unsupervised and transductive learning of information structure

of scientific documents. In *Proceedings of COLING (Posters)*, pages 995–1006, Mumbai, 2012.

Anna Ritchie, Stephen Robertson, and Simone Teufel. Comparing citation contexts for information retrieval. In *Proeedings of the 17th Conference on Information and Knowledge Management (CIKM)*, pages 213–222, Napa Valley, 2008.

Everett M. Rogers. *Diffusion of Innovations, 1st Edition*. Simon and Schuster, 1962.

Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105(4):1118–1123, 2008.

Martin Rosvall and Carl T. Bergstrom. Mapping change in large networks. *PLoS ONE*, 5(1):e8694, 2010.

Camille Roth. Co-evolution in epistemic networks – reconstructing social complex systems. *Structure and Dynamics*, 1(3), 2005.

G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *J. Am. Soc. Inf. Sci.*, 26(1):33–44, 1975.

Beatrice Santorini. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). *Technical Reports (CIS)*, 1990.

Richard M. Shiffrin and Katy Börner. Mapping knowledge domains. *PNAS*, 101(suppl 1):5183–5185, 2004.

Henry G Small. 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, 1973.

Xiaoling Sun, Jasleen Kaur, Staša Milojević, Alessandro Flammini, and Filippo Menczer. Social dynamics of science. *Sci. Rep.*, 3, 2013.

Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the MEDLINE digital library. *I. J. Medical Informatics*, 75(6):488–495, 2006.

Chun-Yuen Teng, Liuling Gong, Avishay Livne Eecs, Celso Brunetti, and Lada Adamic. Coevolution of network structure and content. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 288–297, New York, NY, USA, 2012. ACM.

Simone Teufel. *Argumentative Zoning: Information Extraction from Scientific Articles*. University of Edinburgh, 1999.

Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

Marco Tomassini and Leslie Luthi. Empirical analysis of the evolution of a scientific collaboration network. *Physica A: Statistical Mechanics and its Applications*, 385(2):750–764, 2007.

Nees Jan van Eck, Ludo Waltman, Ed C. M. Noyons, and Reindert K. Buter. Automatic term identification for bibliometric mapping. *Scientometrics*, 82(3):581–596, 2010.

Vladimir Vapnik. *Statistical learning theory [...] [...]*. Wiley, 1998.

Tommaso Venturini, Daniele Guido, and Tommaso Venturini. Once upon a text: an ANT tale in text analysis. *S*, (3/2012), 2012.

Caroline S. Wagner and Loet Leydesdorff. Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10):1608–1618, 2005.

Shenghui Wang and Paul Groth. Measuring the dynamic bi-directional influence between content and social networks. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *The Semantic Web – ISWC 2010*, number 6496 in Lecture Notes in Computer Science, pages 814–829. Springer Berlin Heidelberg, 2010.

Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

Langdon Winner. Upon opening the black box and finding it empty: Social constructivism and the philosophy of technology. *Science, Technology, & Human Values*, 18(3):362–378, 1993.

George Kingsley Zipf. *Human behavior and the principle of least effort*, volume xi. Addison-Wesley Press, Oxford, England, 1949.