Topic Modeling of Research Fields: An Interdisciplinary Perspective

Michael Paul and Roxana Girju
University of Illinois at Urbana-Champaign
{mjpaul2,girju}@illinois.edu

Abstract

This paper addresses the problem of scientific research analysis. We use the topic model Latent Dirichlet Allocation [2] and a novel classifier to classify research papers based on topic and language. Moreover, we show various insightful statistics and correlations within and across three research fields: Linguistics, Computational Linguistics, and Education. In particular, we show how topics change over time within each field, what relations and influences exist between topics within and across fields, as well as what trends can be established for some of the world's natural languages. Finally, we talk about trend prediction and topic suggestion as future extensions of this research.

Keywords

topic models; scientific research analysis; statistical approaches

1 Introduction

No one can predict (at least not in detail) the stringent issues that science and society will consider in the next decades. However, if we look at some top-priority issues of today - such as health, economy, homeland security, stem-cell research, science teaching - and pressing research questions, such as how to enhance child development and learning and even how to make sense of the huge amount of information with which we deal daily, we can say that future topics will be so complex as to require insights from multiple disciplines.

Interdisciplinary research will thus facilitate the integration of information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or sources of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or field of research.

We believe that like other disciplines, Computational Linguistics (CL) will drastically benefit from an interdisciplinary perspective. This research is part of a larger project whose goal is to design a system which will help foster interdisciplinary research in order to make breakthrough predictions for future directions. The system is also intended to promote interdisciplinary collaborations by providing novel topic suggestions to professionals who would like to engage in research discussions with other parties, but who are not familiar with those areas.

This paper presents details about the current version of our system which *researches* a set of three fields: Linguistics, Computational Linguistics, and Education (we include here Educational psychology). Based on various topic models which classify research papers into topic and language categories, the system displays a series of statistics, correlations, and graphics which show the dynamics of topics and local and global trends based on the proceedings of the major conferences and journals in the three fields over many years. In particular, we show how topics change over time within each field, what relations exist between topics, what temporal correlations and topic influences can be determined across fields, as well as what trends can be established for some of the world's natural languages. Finally, we mention some future extensions of this research including suggestions for novel topics by combining research ideas across fields as well as predicting future trends from this combination.

2 Previous Work

Most of the work on the analysis of scientific research deals with citations [11]. This includes the examination of the frequency, patterns and graphs of citations in articles and books. Citation analysis uses citations in scholarly works to establish a graph with links between works and researchers. The web has had a major impact on this type of research leading to the creation of databases such as *Scopus* (www.scopus.com) and *Google Scholar* (scholar.google.com) which allow the analysis of citation patterns of academic papers.

Citation analysis, however is limited in that the citation graphs created are sparse and they do not span related fields. For example, the citation analysis literature [9] shows that 90% of papers published in academic journals are never cited. Moreover 50% of papers are never read by anyone else but their authors, referees, and journal editors.

Another approach to the analysis of scientific research relies on topic models which uncover structures used to explore text collections. In particular, they divide documents according to their topics and use the hidden structure to determine similarity between documents. Popular unsupervised topic models such as Latent Dirichlet Allocation (LDA) [2] and hierarchical models [7] have been successfully applied to various publications such as *The American Political Science Review* and *Science*. In Computational Linguistics, the only work of which we are aware is that of Hall et al. 2008 [4] who study the history of ideas using LDA and topic entropy.

In this paper we extend over the work of Hall et al. 2008 [4] by adding two related fields (Linguistics and Education) and by employing various novel topic models for scientific research analysis.

3 Approach

In this section we present the data used in this research and the topic models employed. We categorize both by topics

Field	Venue	Number of documents	Year range
LING	Language	1031	78-08
LING	Linguistics, Journal of	152	97-08
LING	Linguistic Inquiry	338	98-08
LING	Ling. & Philosophy	652	77-08
CL	ACL	1826	79-08
CL	EACL	517	83-06
CL	NAACL	543	01-07
CL	Applied NLP	262	83-00
CL	COLING	1549	63-
EDU	Education, Journal of	491	75-06
EDU	Educational Psych.	1116	90-08

Table 1: *This table presents the number of documents per field and publication venue. CL stands for Computational Linguistics, LING - Linguistics, EDU - Education.*

and by language.

3.1 The Data

Our corpus consists of approximately 4,700 papers (1965-2008) from the ACL Anthology [1], 2,300 papers from Linguistics journals (1977-2008), and 1,700 papers from Education journals (1975-2008). The exact distribution is shown in Table 1. To best represent each field, we chose top journals (Linguistics and Education) and conferences (CL) that have broad topic coverage of their respective fields. The papers were obtained from library and publisher websites. Only titles and abstracts were freely (and electronically) available for papers in the Linguistics and Education journals.

3.2 Modeling the Research Fields

3.2.1 Modeling Linguistics

Some, albeit only a small fraction, of the Linguistics papers were already categorized (with overlap) in their original publications. Specifically, *Language*, *Journal of Linguistics*, and *Linguistic Inquiry* provided 320 labels for 190 of these papers. Moreover, we manually labeled an additional 147 papers with 185 labels to increase coverage and to create more training data for underrepresented categories. We labeled these papers with categories from the original set as well as new topics that were missing, such as *typology*, *pragmatics*, and *metaphor*. In the end, there were 86 distinct categorization topics. Of the remaining 2,149 unlabeled papers, we had abstracts for 281, and only titles for the rest. The small training set and document lengths make this a difficult classification problem. To begin, we constructed a basic Naïve Bayes classifier that assigns each document D a probability of belonging to each category c_j, defined as

$$P(c_j|D) = P(c_j) \prod_{f_i \in F_D} (P(f_i|c_j)) \quad (1)$$

where F_D is the feature set of document D. The feature space consists of both words from the text, titles and abstracts of documents as well as the bigrams from these strings. Bigrams are useful here – for example, the word "languages" is not so informative, but the phrases "languages in" and "languages of" indicate discussion of languages in a certain region or family, and thus should tilt toward the *language documentation* and *typology* categories.

Since some categories are significantly under-labeled, instead of defining $P(c_j)$ as the observed probability, we assume that the categories have a uniform distribution. The probability of a feature given a class is estimated using Laplace smoothing [10]:

$$P(f_i|c_j) = \frac{n_i + 1}{n + |F|} \quad (2)$$

where n_i is the number of examples labeled c_j that have f_i as an active feature, n is the number of unique active features among all examples labeled c_j, and F is the feature set.

We want to allow a paper to be placed into multiple categories, or none, if it does not match any category. For such an any-of classification task, one would typically create a binary classifier for each class and determine membership in each class individually [8]. We do not do this here because papers were labeled with some but not all of the categories they might belong to, so we cannot assume that the absence of a label implies that a document can be used as a negative example for membership to a class.

Instead, we label a paper with some subset of categories in which $P(c_j|D)$ is significantly greater than the others. To do this, we first perform z-score normalization on the probabilities [5]. The z-score of a value p is defined as $\frac{p-\bar{P}}{\sigma}$, where $\bar{P}$ is the average over each p_j and σ is the standard deviation.

We then say that a paper D belongs to all categories such that the z-score of $P(c_j|D)$ is above some threshold. This means that the probability assigned to the category is greater than the average by some distance relative to the standard deviation of the probability values.

In an attempt to strengthen the training data, we took a semi-supervised approach and added to the training set documents that were labeled with a probability above a constant confidence threshold. This process was iteratively repeated until no new examples were added to the training set.

For an estimate of this classifier's performance, we performed 10-fold cross validation. Table 2 shows however that its initial performance was not good enough to make accurate observations.

We improved over this approach employing a model proposed by Zelik & Hirsh 2000 [12]. The idea is to use an unlabeled corpus of background knowledge to match unlabeled examples with labeled examples. Thus, if a labeled document A is similar to some document W in the background corpus and an unlabeled document B is similar to the same document W, then perhaps B should have the same label as A. This method is particularly useful when the training set is very small and when the strings to classify are short.

To create our background corpus, we grabbed the Wikipedia articles categorized under *Linguistics*, truncating the documents down to the main content part and removing the HTML tags. Each article is represented as a vector of the tf-idf measures of the words in its text, with log term frequencies and $\text{IDF}(t) = log(\frac{|d|}{|d_t|})$.

The same tf-idf representation is used for our labeled and unlabeled research papers. The cosine measure is used to

calculate the similarity $sim(D_i, W_j)$ between a paper and a Wikipedia article. Let $v_{D,c}$ be the score assigned to a document D for a category c based on this matching through Wikipedia. We define this as

$$v_{D,c} = \sum_{A_i \in W_D} \sum_{L_j \in X_{Ai}} I(c \in L_j) sim(L_j, A_i)^2 sim(A_i, D) \quad (3)$$

where W_D is the set of Wikipedia articles that are similar to D above a threshold cosine score λ_c, X_{Ai} is the set of labeled papers with similarity scores to an article A_i greater than λ_c, and I is the indicator function. λ_c is defined as some constant k standard deviations above the mean similarity measurement between a category c and each article. This assigns a larger v to the categories with the highest similarity to the Wikipedia articles which are highly similar to D (the paper we are attempting to label).

We can now classify a document by augmenting the original Naïve Bayes probability $P(c|D)$ with this new v score as follows.

During testing, we noticed that $P(c|D)$ was a more accurate weight than $v_{D,c}$ (or vice versa) for certain categories. For example, Naïve Bayes alone could correctly label a paper as *historical linguistics* in the presence of a word such as "history", but matching through Wikipedia would usually point the classifier to an irrelevant class.

To compensate for this problem, we introduce a bias factor α, defined as the mean value of $P(c|D)$ or $v_{D,c}$ for each document D in the training set that is labeled as c, where both $P(c|D)$ and $v_{D,c}$ have been normalized to the same range. We calculate α values during a run of the cross-validation test, then rerun the classifier with these values.

Finally, to label a document, we assign each document a weight toward a class c, defined as

$$w_{D,c} = log(1+(frac_{NB,c}P(c|D))) + log(1+(frac_{W,c}v_{D,c})) \quad (4)$$

where $frac_{Z,c} = \frac{\alpha_{Z,c}}{\alpha_{NB,c}+\alpha_{W,c}}$.

$frac$ distributes the weights between the methods (Naïve Bayes or Wikipedia matching) according to how they usually perform on the class c.

$w_{D,c}$ increases with the size of class c, so we adjust the confidence threshold according to this. Then, a paper D is labeled as the category c if the z-score of $w_{D,c}$ is above the variable threshold δ_c, which is defined as the prior probability $P(c)$ normalized to fit the range $[\delta_L, \delta_U]$ for some constant lower/upper threshold bounds.

This classification performance is listed in Table 2 (Combined NB + Wikipedia). This performance is actually slightly worse than the semi-supervised Naïve Bayes classifier. However, it was able to correctly classify papers that Naïve Bayes could not.

We then took an additional step of semi-supervision and extracted the top results (with a score above some threshold) of our combined Naïve Bayes with Wikipedia classifier and added them to the training set. We re-ran our original semi-supervised Naïve Bayes classifier with this new training set to get our final results (last row in Table 2).

3.2.2 Modeling Computational Linguistics

Most of the papers in the ACL Anthology are not categorized, so unsupervised methods were needed to label them.

Model	P	R	F
Supervised NB	0.59	0.68	0.63
Semi-supervised NB	0.89	0.68	0.77
Combined NB + Wikipedia	0.85	0.67	0.75
Semi-super. w/ new labels	0.91	0.78	0.84

Table 2: *Classification performance for the Linguistics data. NB stands for Naïve Bayes.*

We chose to use the generative model Latent Dirichlet Allocation [2], which represents documents as random mixtures over latent topics, where each topic is characterized by a multinomial distribution of words.

After removing a standard list of stop words, we ran LDA to induce 100 topics on the text of the papers and saved 72 topics that were relevant. A sample of these topics is shown in Table 3.

3.2.3 Modeling Education

We used a similar process for the field of Education. We ran LDA on the Education papers using the words from the titles and abstracts, as the full text was not available. We used these data to induce 30 topics and chose 18 that were relevant. Two of these 18 topics were specifically relevant to language – *reading/language comprehension* and *reading/language instruction*. We repeated the LDA process on this subset of language-related papers and grouped them into 8 additional topics. Samples of these topics are shown in Table 3.

3.3 Categorizing by Language

In addition to labeling papers by topic, we noted which languages were discussed in papers. Thus, we simply labeled a paper with the languages that appear in its text above a certain frequency threshold. Intuitively, this should work except for a few cases and with a few languages. English, for example, is not always explicitly mentioned in papers that focus on English, and Greek returns false positives in Education because of Greek culture studies. Otherwise, empirically this works quite well – if a language is mentioned at least a few times in a paper, then we would like this paper to be labeled as such.

4 Data Analysis

In the following subsections we present insightful observations on the data classification and discuss potential trends.

4.1 Changes Over Time

To measure a topic's prominence over time, we look at the fraction of papers within that topic dated in a given year out of all papers from that year. We perform least squares linear regression on the temporal data points for each topic to see if and by how much a topic has a general upward/downward trend.

Within Computational Linguistics, our findings are similar to those presented by Hall et al. 2008 [4]. Thus, *text classification* has the largest upward trend. *Natural language interfaces* and *speech act interpretation* are

Topic	Keywords
Linguistics	
Pragmatics	pragmatics attitudes meaning semantics pragmatic inference communication
Prosody	accent intonation initial prosodic prosody contour fall stress phonological
Psycholinguistics	mental psychological language processing psychology representations triggers
Quantifiers	quantifier quantification quantifiers existential scope generalized polyadic
Semantics	semantic semantics meaning lexical content pragmatics meanings conceptual
Computational Linguistics	
Morphology	morphological word morphology lexical level forms form lexicon stem words
MT Evaluation	evaluation score human scores sentence automatic quality reference metrics
Multimodal NLP	multimodal speech gesture user language input figure spoken based systems
Named Entities	entity names named entities ne information person location muc extraction
Optimality Theory	constraints constraint dominance theory language phonological structure stress
Education	
Race/Ethnicity Issues	american students african teachers ethnic minority stereotypes Educational
Reading Instruction	reading children phonological instruction awareness grade spelling skills
Reading Comprehension	reading language comprehension english children vocabulary word readers
Self Concept/Efficacy	self concept efficacy academic model relations skill domain ability
Teaching Effectiveness	learning multimedia students evaluations teaching effectiveness factor

Table 3: *Slice of topics and their top keywords in Linguistics, Computational Linguistics, and Education.*

among the strongest-declining topics. In general, *formal semantics* and similar theoretical topics have taken a nose-dive since the 1980s, while *statistical/probabilistic methods* have strongly increased. Within the area of *semantics*, there are some topics on the rise, such as *word sense disambiguation*, *semantic role labeling*, and *event/temporal semantics*.

In Linguistics, no topic showed a strong rise in prominence, at least not among topics that were large enough to give accurate trends over time. For the most part, topics fluctuate year-to-year, but do not have an overall trend. There were, however some topics with a noticeable decline. Thus, an interesting observation is that while in Computational Linguistics *formal semantics* took a significant plunge, it has declined less dramatically in the field of Linguistics while still remaining relatively prominent (see Table 1). The statistics indicate that *language documentation*, *historical linguistics*, and *pragmatics* show the most marked decline in Linguistics. *Discourse* shows a sudden decline in the late 1990s – to compare, *discourse segmentation* has a steady rise in CL, but *discourse Centering Theory* has a steady decline. Interestingly, the prevalence of computational Linguistics papers in the linguistics journals peaked in the late-80s and early-90s and has since declined. Moreover, *language acquisition* has declined in the Linguistics field in the past decade, whereas it has risen in the Education field in the same time period.

Morphology, *prosody*, and *quantifiers* have a steady decline in CL, whereas they stayed fairly consistent (but small) in Linguistics.

In Education, there is a markedly strong rise in prominence of topics about *language and reading*. *Student performance* is another topic with a strong increase, while *epistemology* has slightly declined. These are shown in Figure 2.

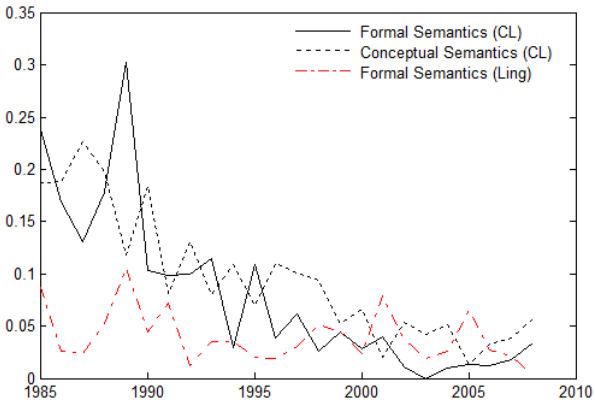

Fig. 1: Semantics *in Computational Linguistics and Linguistics over time.*

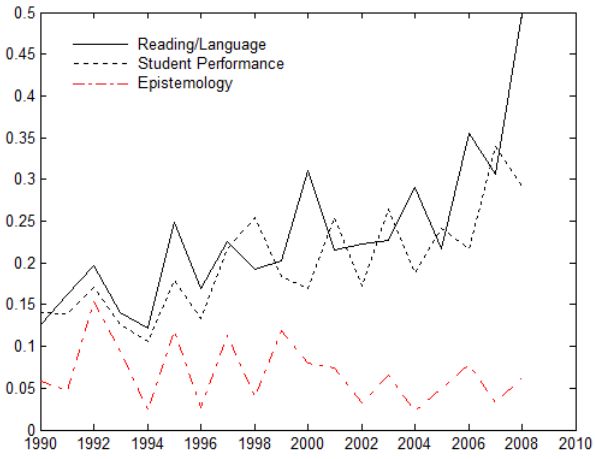

Fig. 2: *Most prominent upward/downward trends in Education.*

4.2 Relations Between Topics

We can see how different research areas are related by allowing papers to be assigned to multiple topics. For example, within computational Linguistics we found that the *dialogue systems* topic overlaps with *natural language interfaces* and *speech recognition* - some percentage of papers labeled as *dialogue systems* have also been labeled with these topics.

Of course, we also want to see how topics relate across

fields. Since we only modeled topics within each distinct field, we must determine which topics of different fields are similar. Thus, we create a topic meta-document for each topic by concatenating the words of every document within the class. We can then represent each topic as a vector of words from these documents (again weighted by their tf-idf value) and compute the similarity of these topic vectors by their cosine.

Figure 3 highlights the interdisciplinary nature of these fields. The links in the diagram show a sample of the highest-scoring similarity matches, where line thickness indicates increasing similarity value.

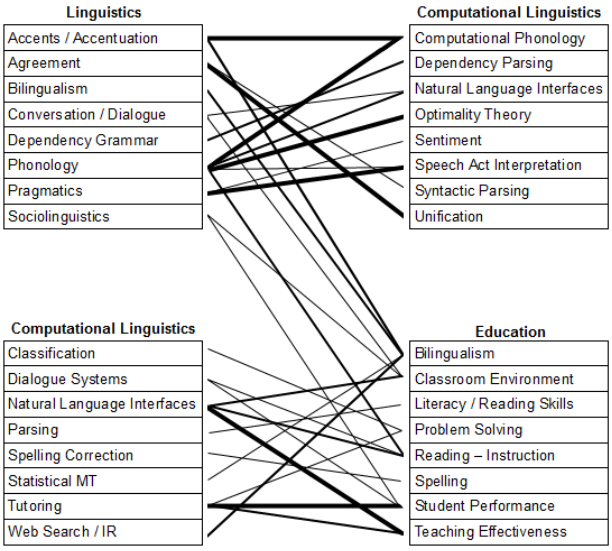

Fig. 3: *Similarity of topics across fields.*

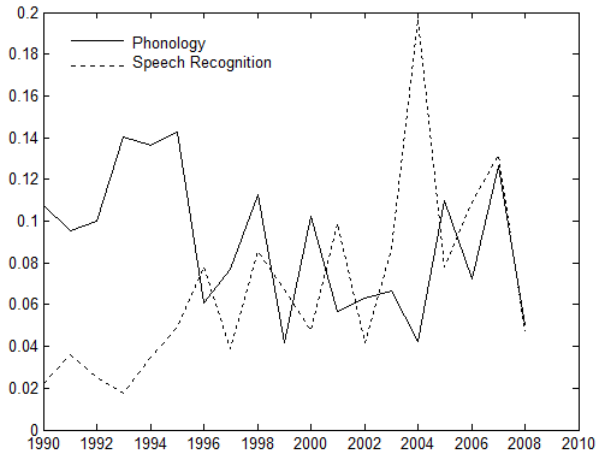

Fig. 4: *Similarity of topics across the speech and phonology fields.*

4.3 Language Trends

The most prominent languages (after English) within computational Linguistics are Japanese, German, French, Chinese, and Spanish. Within each language, we can look at the distribution of topics – although they were not strikingly different for the most part, there were some differences. The most prominent topic for Chinese, for example, is *word segmentation*, which did not receive the same attention in the other languages.

The Education field differs slightly in that its most prominent languages are Chinese, Spanish, German, and Korean in this order. Chinese and German have a pretty general topic distribution, while Spanish- and Korean-related papers are predominately about *bilingualism* and *language learning*. Japanese, German, Spanish, and French were found as the most prominent within Linguistics.

In Computational Linguistics, English and Japanese have remained consistently prominent throughout the years. Chinese and Arabic show strong increases, while Russian and Italian have a slight downward trend. French, German, and Spanish all rose through the late 80s and 90s and have since slightly declined.

Chinese and Spanish are on the rise in Education and Linguistics seems to be taking an increasing interest in Japanese.

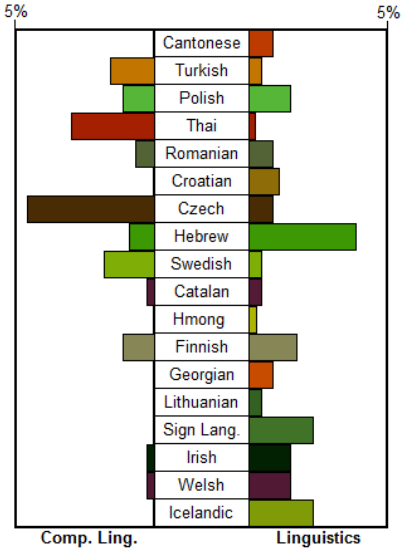

Fig. 5: *Comparison of less-spoken languages in Computational Linguistics and Linguistics.*

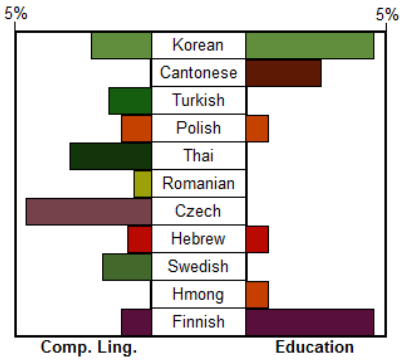

Fig. 6: *Comparison of less-spoken languages in Computational Linguistics and Education.*

It is also important to look at less-spoken but still prominent languages (Figures 5 and 6). There are some differences in the languages being discussed in Computational Linguistics compared to Linguistics and Education. For

example, Czech, Thai, and Swedish are prominent in CL, Hebrew, Icelandic, sign languages, Irish, Welsh in Linguistics, and Korean, Finnish and Cantonese in Education[1].

We also compared specific topics over time across languages. For example, all Arabic papers in CL were in the *morphology* topic. The other top languages were mostly the same: *statistical MT* and *parsing* at the top.

5 Discussion

The statistics and observations presented in this paper have an important significance for the research community at large. Besides the potential of identifying novel topics, this information is useful for assessing which areas are important and which areas might currently be overlooked. Moreover, we showed that such a system can indicate interesting correlations among related fields, correlations which can be put to work in various ways. For example, these data can be very useful to computational linguists who can get ideas of novel topics or theoretical models from Linguistics, build sophisticated systems and apply them to Education. Another possibility is to use large-scale empirical Computational Linguistics models to help identify and develop new theories in Linguistics.

Most importantly this kind of research will hopefully foster collaboration among related fields. Moreover, tools such as the one presented here will be beneficial to young researchers who start their graduate studies looking for research topics within their field, but also in an interdisciplinary context.

In the next subsections we provide some detailed suggestions.

5.1 Suggestions for Research Directions

Although languages such as Arabic, Russian, and Korean have been studied in Computational Linguistics, they seem to be somewhat under-represented in the field relative to their importance in the world and other fields. Spanish is one such example – in spite of being one of the most-spoken world's languages and in spite of its rising importance in Education research, it continues to be underrepresented in Computational Linguistics. Cantonese, Hmong, Finnish, and Hebrew are significantly more prominent in Linguistics and Education than in CL.

Another general area that seems under-represented in Computational Linguistics is that of *dialects* and *dialectology*. While this has certainly been covered in Computational Linguistics, its prominence is small compared to that in Linguistics, and studies have mostly focused on a small subset of languages. Inconsistencies in natural language processing created by different dialects of a language is a known problem in the community [3], so this should be studied further.

Language evolution is a topic of interest in Linguistics, but very little has been done to study it using computational methods, at least among the papers in our collection. We did find a few papers on computational phylogeny in the Linguistics and Computational Linguistics corpora, but this appears to be a topic that could use more research.

[1] It seems that the Education field has focused mostly on languages spoken in the western education systems.

5.2 Trend Prediction and Topic Suggestion

We can go even further with the analysis of such correlations among related fields. For example, we have already shown in Section 4.3 that some fields seem to influence others on some particular topics, such as *phonology*, *speech recognition*, and *dialog systems*. Such possible influences indicate that we can go a step further towards trend prediction. Moreover, another goal for our research is to be able to suggest novel and interesting topics. For example, a closer look at our data collection, statistics, and trends indicates a potential research topic: *automatic note-taking* (i.e., how to build a system which takes notes automatically say, in an academic environment). To our knowledge the topic is novel in Computational Linguistics and has direct implications in Linguistics and Education. While in Linguistics it has not been studied[2], the topic has been well researched in Education (in particular from a learning and knowledge retention perspective). However, in oder to make trend prediction and topic suggestion possible, a much deeper analysis is needed on much larger text collections. This is left for future research.

6 Conclusions

In this paper we presented various novel topic models which classify research papers based on topic and language. Moreover, we gave various insightful statistics and correlations within and across three research fields: Linguistics, Computational Linguistics, and Education. In particular, we showed a number of trends in each field along with relations between topics, temporal correlations and topic influences across fields, as well as language trends.

References

[1] S. Bird. Association for Computational Linguistics Anthology. In *http://www.aclweb.org/anthology-index/*, 2008.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] N. Habash. Arabic Natural Language Processing for Machine Translation. In ACL, editor, *Tutorial at the 8th Conference of the Association for Machine Translation in the Americas (AMTA)*. MIT Press, 2008.

[4] D. Hall, D. Jurafsky, and C. Manning. Studying the history of ideas using topic models. In *Empirical Natural Language Processing Conference*, 2008.

[5] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. 2nd Edition. Morgan Kaufmann, San Francisco, 2006.

[6] R. Janda. Note-taking english as a simplified register. *Discourse Processes*, 8:437–454, 1985.

[7] W. Li and A. McCallum. Dag-structured mixture models of topic correlations. In *International Conference on Machine Learning*, 2006.

[8] C. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[9] L. Meho. The Rise and Rise of citation analysis. *Physics World*, to appear.

[10] T. Mitchell. *Machine Learning*. McGraw Hill, New York, 1997.

[11] R. Rubin. *Foundations of Library and Information Science*. 2nd ed. New York: Neal-Schuman, 2004.

[12] S. Zelikovitz and H. Hirsh. Improving short text classification using unlabeled background knowledge to assess document similarity. In *The 17th International Conference on Machine Learning (ICML)*, 2000.

[2] Actually, note-taking has been studied in Linguistics by Richard Janda [6], but the paper was published in a different journal.