Feature Discovery for Diachronic Register Analysis: a Semi-Automatic Approach

Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski, Elke Teich

Saarland University
Universität Campus A2.2,
66123 Saarbrücken, Germany
s.degaetano, e.lapshinova, e.teich@mx.uni-saarland.de

Abstract

In this paper, we present corpus-based procedures to semi-automatically discover features relevant for the study of recent language change in scientific registers. First, linguistic features potentially adherent to recent language change are extracted from the SciTex Corpus. Second, features are assessed for their relevance for the study of recent language change in scientific registers by means of correspondence analysis. The discovered features will serve for further investigations of the linguistic evolution of newly emerged scientific registers.

Keywords: diachronic register analysis, correspondence analysis, linguistic feature evaluation

1. Introduction

The present paper describes semi-automatic procedures to discover features reflecting diachronic changes in scientific registers. Registers are patterns of language according to use in context, cf. (Halliday and Hasan, 1989). Findings about register change are relevant both for linguistic applications, e.g., discourse analysis, language pedagogy, translation studies, and NLP tasks, notably automatic text classification.

To discover candidate features for the analysis of register change, i.e., features that are distinctive diachronically, we carry out a corpus-linguistic analysis, which includes extraction and evaluation of candidate lexico-grammatical features from a diachronic corpus. The basis for our work is provided by register theory (Halliday and Hasan, 1989) and previous studies of recent language change, notably (Mair, 2006).

We address the following questions: (1) Which lexico-grammatical features are good candidates for the study of recent language change in written scientific English? (2) How can they be reliably extracted from text corpora?, and (3) Which of the candidate features are actually suitable for studying the evolution of scientific registers?

2. Data and Existing Approaches

2.1. Approaches to Register and Language Change

Register theory, e.g. (Quirk et al., 1985), (Halliday and Hasan, 1989) and (Biber, 1995), is concerned with linguistic variation according to contexts of use, which involve language use expressed in the co-occurrences of particular lexico-grammatical features, creating distinctive registers (e.g., in the scientific domain: the language of biology or linguistics).

Both contexts of use and language use change over time resulting in the evolution of registers: the existing ones become obsolete and new ones evolve (e.g., in the scientific domain: the language of bioinformatics or computational linguistics). Such changes are directly reflected in lexico-grammar, some features becoming rarer, others more frequent, and features cluster in novel ways.

One approach to the study of recent language change was developed by (Mair, 2006), who investigated the Brown corpus family for changes in preferences of lexico-grammatical selection in British and American English between the 1960's and 1990's[1].

We test the features described by (Mair, 2006) for their suitability for the present task analysing their frequencies in SciTex, a diachronic corpus of academic English, cf. (Degaetano-Ortlieb et al., 2012). Our first extraction results show that there are two groups of potential feature candidates in our corpus: (a) lexical and (b) grammatical.

2.2. Data

2.2.1. Lexical features

As the linguistic reaction to the developing scientific disciplines we observe the evolution of new scientific registers, with the most prominent changes in lexis (particularly specialized terminology). Here, we focus on the word formation process of prefixation. Table 1 gives an overview of prefixes involved in the formation of new words over time, potentially also relevant in the scientific domain.

2.3. Grammatical features

Grammatical changes unfold much more slowly than lexical changes. However, the trends of their development are detectable by inspecting frequencies of use of competing grammatical variants, e.g. noun or verb complementation patterns, the use of voice, etc. Table 2 shows some of the relevant grammatical features for recent language change used by (Mair, 2006), which we generalise for our study.

[1]The Brown corpus family consists of four parts, the Brown corpus (1961, AmE), the LOB corpus (1961, BrE), the Frown corpus (1992, AmE) and the FLOB corpus (1991, BrE), each subcorpus containing text samples from written English of 15 broader registers (e.g., scientific, religious, fiction etc).

feature	example
up+VERB	*update*
down+VERB	*download*
prefix+VERBing	*incoming, outgoing*
post+NOUN or ADJ	*post-editing, postprocessing*
hyper+	*hyperarticulation*
super+	*superscript, superset*

Table 1: Features for language change: prefixation

feature	example
NOUN+*for*+NP+to-inf	*the need* ***for linguists to meet***
VERB+to-inf	*begin/start* ***to do*** *smth*
VERB+(obj)+to-inf	***help*** *(smb)* ***to do*** *smth*
VERB+(obj)+bare-inf	***help*** *(smb)* ***do*** *smth*
VERB+VERBing	***consider maximizing*** *the metric*
VERB+obj+(from)+VERBing	***prevent*** *smb (from)* ***doing*** *smth*
get+passive	*buffer will* ***get chosen*** *by processor*
modal verbs	*we* ***shall*** *be able to treat*

Table 2: Features for language change: grammatical variants

3. Methods and Tools

3.1. Corpus and Overarching Research Interest

For our investigations we use the English Scientific Text Corpus (SciTex), cf. (Degaetano-Ortlieb et al., 2012), which contains full English scientific journal articles from nine disciplines. The corpus covers two time periods (each contained in an own subcorpus), the early 2000's (DaSciTex) and the 1970's/early 1980's (SaSciTex), and amounts to approx. 34 million tokens. In this study, we focus on subcorpora representing 'mixed' disciplines, i.e., the disciplines in contact with computer science which form a new discipline in the time period investigated (see figure 1: computational linguistics (B1), bioinformatics (B2), digital construction (B3), microelectronics (B4)); cf. (Teich and Holtz, 2009) and (Teich and Fankhauser, 2010).
The corpus is annotated for sentence boundaries, tokens, lemmata and parts-of-speech by means of a dedicated processing pipeline (Kermes, 2011) and can be queried with the IMS Corpus Query Processor (CQP), cf. (Evert, 2005).

3.2. Procedures for Feature Discovery

To obtain features suitable for diachronic register analysis, we need to (1) extract instances of the candidate features (feature identification), and (2) assess them in terms of their relevance for further analysis (feature evaluation).

Feature identification As candidate features, we decide for the lexico-grammatical features presented in tables 1 and 2 above. To extract them for the two time slices and all nine registers of SciTex, we elaborate CQP-based queries, which contain both string-based lexical and grammatical constraints (e.g., part-of-speech information or sentence position).

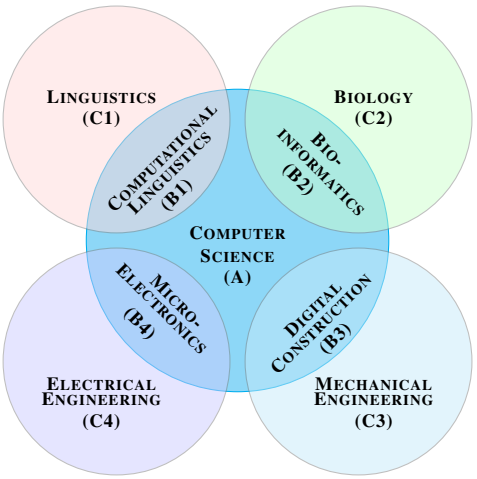

Figure 1: Scientific disciplines in the SciTex corpus

An example of a query extracting grammatical features is shown in figure 2. This query is designed to extract verbs which vary in their complementation patterns between *from*+gerund and NP+gerund, cf. examples in (1a) and (1b).

(1a) *This* ***prevents us from computing*** *the probability of random agreement.*

(1b) *The lack of a real gold standard... will still* ***prevent comparing such an algorithm*** *to...*

Feature evaluation To distinguish features relevant for language change in scientific registers, we carry out a correspondence analysis on our data. Correspondence analysis is a multivariate technique which seeks to provide a low-dimensional map of the data usually plotted in a two dimensional graph (cf. (Baayen, 2008)). This statistical method works on observed frequencies and is able to show their relations to variables in one single space (cf. (Greenacre, 2007)). This suits our purpose as we want to investigate the relations between the observed frequencies of our extracted lexico-grammatical features and the respective subcorpora (variables). To perform the analysis, the *ca* package (cf. (Nenadić and Greenacre, 2007)) for the statistical environment R (cf. (Venables and Smith, 2010)) is used. The input for the analysis is a data frame which comprises features in rows and subcorpora in columns with the respective frequencies (see Table 3).

	A.da	B1.da	C1.da	...
up+VERB	271	105	48	...
down+VERB	8	30	13	...
prefix+VERBing	119	55	62	...
...	...	...	...	...

Table 3: Extract of the data frame for correspondence analysis

Correspondence analyses are performed on all A-B-C triples of subcorpora (e.g., A - computer science, B1 - computational linguistics, C1 - linguistics) for both time slices

	query building blocks	comments	extracted examples
1		pre-verbal material	*This*
2	[pos="V.*"]	verb	*prevents*
3	(	object start	
4	[pos="DT|PP|PDT"]?	one or none determiner	
5	[pos="RB.*|JJ.*|VVN|N.*"]{0,3}	up to 3 modifiers	
6	[pos="POS"]?	one or none possessive	
7	[pos="N.*|PP"]?	noun or pronoun	*us*
8	)	object end	
9	[word="from"&pos="IN"]?	optional *from*	*from*
10	[pos="V.*G"]	*ing*-verb	*computing*
11		post-verbal material	*the probability of random agreement*

Figure 2: Query for extraction of verbs with two complementation patterns

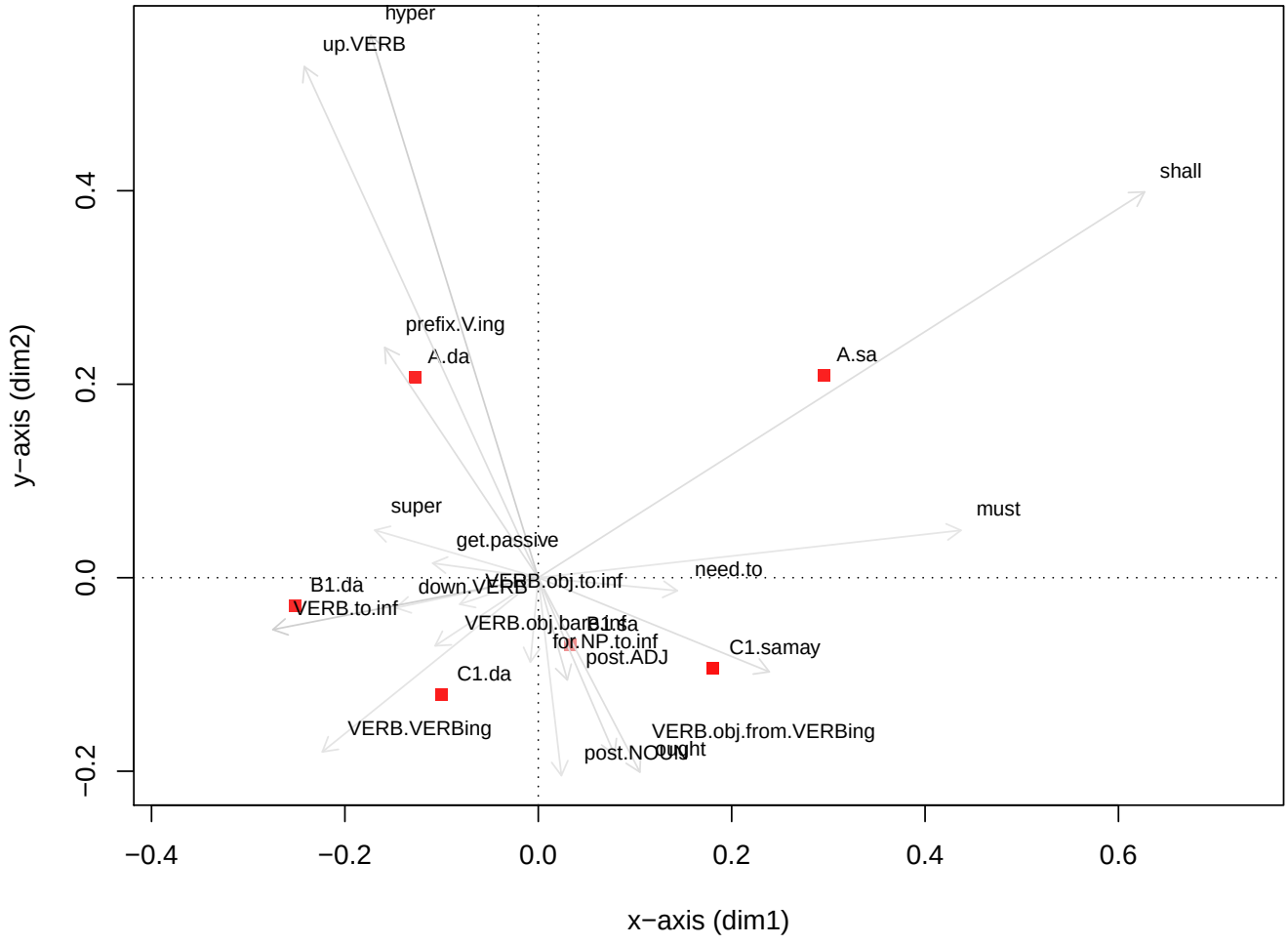

Figure 3: Graph for correspondence analysis on the A-B1-C1 triple

(e.g., 2000's: A.da, B1.da, C1.da; 1970's/80's: A.sa, B1.sa, C1.sa)[2].

The output of the correspondence analysis is plotted into a two dimensional graph with arrows representing the observed frequencies of a feature and points representing the subcorpora, allowing the inspection of the arrow length and point position in relation to the arrows. The length of the arrows indicates how pronounced a particular feature is, see

[2]We use .da for the DaSciTex subcorpus and .sa for the SaSciTex subcorpus.

(Jenset and McGillivray, 2012) for details. The position of the points in relation to the arrows indicates the relative importance of a feature for a subcorpus. The arrows pointing in the direction of an axis indicate a high contribution to the respective dimension. Figure 3 shows the graph for the A-B1-C1 triple and the two time slices.

To assess how well our data is represented in the graph, we calculate the eigenvalues for each dimension with R (see figure 4, for the triple A-B1-C1 for both time slices). Good results are obtained by a relatively high cumulative value by the first two dimensions, as they are the ones used to plot the two-dimensional graph.

dim	eigenvalue	%	cumulative value %	contribution to the graph
1	0.033370	56.2	56.2	**********
2	0.017905	30.1	86.3	**********
3	0.006138	10.3	96.7	****
4	0.001494	2.5	99.2	*
5	0.000493	0.8	100.0	
Total:	0.059399	100.0		

Figure 4: Contributions of dimensions (A.sa, A.da, B1.sa, B1.da, C1.sa, C1.da)

4. Results and Interpretation

Our extraction tool delivers information on feature frequencies for all candidate features described in section 2 in our corpus, as shown in Table 4. Detailed CQP-based queries, e.g., the one in figure 2 above, allow to reliably detect feature instances in the corpus.

The numerical results obtained for all candidate features provide the basis for the correspondence analysis. In the case of the A-B1-C1 triple, the cumulative value for the first two dimensions is 86.3% (see figure 4), i.e. the first two dimensions represent almost 90% of the data which indicates that our data is well represented in the graph. Similar values are obtained for the other A-B-C triples for the two time slices. Considering the *y* axis in figure 3, there is a clear separation according to discipline, with computer science (A.da, A.sa) in the top and computational linguistics (B1.da, B1.sa) and linguistics (C1.da, C1.sa) in the bottom. The strongest feature here is prefixation, notably the prefixes *hyper* and *up*. Considering the *x* axis, there is a clear diachronic cut between the data from the 2000's and those from the 1970's/80's, with A.da, B1.da and C1.da on the left and A.sa, B1.sa, C1.sa on the right. The features whose arrows are situated in the direction of the *x* axis, e.g., *get*+passive and VERB+to-inf, contribute to the diachronic division. The feature *shall*, which is very pronounced, contributes to both the diachronic and the disciplinary division.

We carried out correspondence analyses for the other A-B-C triples for the two time slices in the same way. Taken together, we get the same tendencies for all triples of A-B-C, the first two dimensions always involved in the diachronic and discipline-related tendencies. Finally, to determine a cut-off between relevant and irrelevant features, we employ the chi-square p-value of <0.05. Table 5 shows which of the features tested are relevant for diachronic analysis (marked by checks).

We thus obtain a clear idea about which of the candidate features are relevant for the study of language change in our corpus. In addition, the analysis allows us to see how pronounced the diachronic trends are according to discipline (e.g., use of *shall* in A.sa).

5. Conclusions and Future Work

In this paper, we have shown how features relevant for diachronic register analysis can be discovered semi-automatically. Our procedures enable us to extract and evaluate candidate features involved in register change, thus providing a basis for the analysis of linguistic evolution of scientific registers.

We will combine the features used in the present study with additional features we have used for register analysis before ((Degaetano-Ortlieb et al., 2012) and employ them to build a model of linguistic evolution of newly emerged scientific registers.

The procedures described in this paper will be integrated into dedicated processing pipelines for the semi-automatic analysis of register change, which can be applied in both linguistic studies and for NLP tasks. The results of studies of recent change in language will be valuable not only to languages-for-specfic-purposes (LSP) and historical linguistics, but also to automatic text classification, where the search for relevant features is a core issue.

6. Acknowledgements

The project *Register im Kontakt: Zur Genese spezialisierter wissenschaftlicher Diskurse* (Registers in contact: linguistic evolution of specialized scientific registers) is supported by a grant from Deutsche Forschungsgemeinschaft (DFG). We are especially grateful to Hannah Kermes for providing the necessary corpus processing pipeline. Also, we wish to thank the anonymous reviewers for their suggestions for improving our paper. All remaining errors remain ours.

7. References

R. Harald Baayen. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Douglas Biber. 1995. *Dimensions of Register Variation. A cross linguistic comparison*. Cambridge University Press, Cambridge.

Stefania Degaetano-Ortlieb, Kermes Hannah, Ekaterina Lapshinova-Koltunski, and Teich Elke. 2012. Scitex – a diachronic corpus for analyzing the development of scientific registers. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics*, volume Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP, Vol. 2. Narr.

Stefan Evert, 2005. *The CQP Query Language Tutorial*. IMS Stuttgart. CWB version 2.2.b90.

Michael Greenacre. 2007. Tying up the loose ends in simple, multiple and joint correspondence analysis. In A. Rizzi and M. Vichi, editors, *COMPSTAT 2006: Proceedings in Computational Statistics*. Heidelberg: Springer-Verlag.

candidate features	frequencies in SciTex			
	2000's		1970's/80's	
	pm	pc	pm	pc
lexical				
up+VERB	765363	76.54	234637	23.46
down+VERB	973881	97.39	26119	2.61
prefix+VERBing	743304	74.33	256696	25.67
post+NOUN or ADJ	423208	42.32	576792	57.68
hyper+	624113	62.41	375887	37.59
super+	760421	76.04	239579	23.96
grammatical				
NOUN+*for*+NP+to-inf	610052	61.01	389948	38.99
VERB+to-inf	603188	60.32	396812	39.68
VERB+(obj)+to-inf	735849	73.58	264151	26.42
VERB+(obj)+bare-inf	764925	76.49	235075	23.51
VERB+VERBing	636364	63.64	363636	36.36
VERB+obj+(from)+VERBing	529577	52.96	470423	47.04
get+passive	1000000	100.00	0	0.00
modal verbs	580658	58.07	419342	41.93

Table 4: Candidate feature frequencies in SciTex

general categories	features	relevance
word formation	prefix+VERB(ing)	√
	prefix+NOUN/ADJ	√
noun postmodification	NOUN+*for*+NP+to-inf	
verb complementation	VERB+to-inf	√
	VERB+(obj)+to-inf	
	VERB-(obj)+bare-inf	
	VERB+VERBing	√
	VERB+(obj)+(from)+VERBing	
voice	*get*+passive	√
modal verbs	*shall, need to, ought*	√
	must, may	√

Table 5: Features for diachronic register analysis

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1989. *Language, context and text: Aspects of language in a social-semiotic perspective*. Oxford University Press, Oxford.

Gard B. Jenset and Barbara McGillivray. 2012. Multivariate analyses of affix productivity in translated english. In Michael P. Oakes and Meng Ji, editors, *Quantitative Methods in Corpus-Based Translation Studies*, pages 301–324. John Benjamins.

Hannah Kermes, 2011. *Automatic corpus creation*. Universitt des Saarlandes, Saarbrcken.

Christian Mair. 2006. *Twentieth-Century English: History, Variation and Standardization*. Cambridge University Press, Cambridge.

Oleg Nenadić and Michael Greenacre. 2007. Correspondence analysis in r, with two- and three-dimensional graphics: The *ca* package. *Journal of Statistical Software*, 20(3):1–13.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Elke Teich and Peter Fankhauser. 2010. Exploring a corpus of scientific texts using data mining. In S. Gries, S. Wulff, and M. Davies, editors, *Corpus-linguistic applications: Current studies, new directions*, pages 233–247. Rodopi, Amsterdam and New York.

Elke Teich and Mônica Holtz. 2009. Scientific registers in contact. an exploration of the lexicogrammatical properties of interdisciplinary discourses. *International Journal of Corpus Linguistics*, 14(4):524–548.

William N. Venables and David M. Smith. 2010. *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics*.