Visualising Linguistic Evolution in Academic Discourse

Verena Lyding
European Academy of Bolzano-Bozen
verena.lyding@eurac.edu

Ekaterina Lapshinova-Koltunski
Saarland University
e.lapshinova@mx.uni-saarland.de

Stefania Degaetano-Ortlieb
Saarland University
s.degaetano@mx.uni-saarland.de

Henrik Dittmann
European Academy of Bolzano-Bozen
henrik.dittmann@eurac.edu

Christopher Culy
The University of Tübingen
christopher.culy@uni-tuebingen.de

Abstract

The present paper describes procedures to visualise diachronic language changes in academic discourse to support analysis. These changes are reflected in the distribution of different lexico-grammatical features according to register. Findings about register differences are relevant for both linguistic applications (e.g., discourse analysis and translation studies) and NLP tasks (notably automatic text classification).

1 Introduction

The present paper describes procedures to visualise diachronic language changes in academic discourse with the aim to facilitate analysis and interpretation of complex data. Diachronic changes are reflected by linguistic features of registers under analysis. Registers are patterns of language according to use in context, cf. (Halliday and Hasan, 1989).

To analyse register change, we extract lexico-grammatical features from a diachronic corpus of academic English, and visualise our extraction results with *Structured Parallel Coordinates* (SPC), a tool for the visualisation of structured multidimensional data, cf. (Culy *et al.*, 2011).

Our approach is based on the inspection and comparison of how different features change over time and registers. The major aim is to determine and describe tendencies of features, which might become rarer, more frequent or cluster in new ways. The amount and complexity of the interrelated data, which is obtained for nine disciplines in two time periods (see section 2) makes the analysis more difficult.

Structured Parallel Coordinates provide a tool for the compact visual presentation of complex data. The visualisation of statistical values for different linguistic features laid out over time and register supports data analysis as tendencies become apparent. Furthermore, interactive features allow for taking different views on the data and focussing on interesting aspects.

2 Data to Analyse

2.1 Features and theoretical background

When defining lexico-grammatical features, we refer to Systemic Functional Linguistics (SFL) and register theory, e.g., (Quirk, 1985), (Halliday and Hasan, 1989) and (Biber, 1995), which are concerned with linguistic variation according to contexts of use, typically distinguishing the three contextual variables of *field*, *tenor* and *mode* of discourse. Particular settings of these variables are associated with the co-occurrences of certain lexico-grammatical features, creating distinctive registers (e.g., the language of linguistics in academic discourse). We also consider investigations of recent language change, observed, e.g., by (Mair, 2006), who analyses changes in preferences of lexico-grammatical selection in English in the 1960s vs. the 1990s.

As a case study, we show an analysis of modal verbs (falling into the contextual variable of *tenor*), which we group according to (Biber, 1999) into three categories of meaning that represent three features: *obligation*, *permission* and *volition* (see Table 1).

2.2 Resources

The selected features are extracted from SciTex, cf. (Degaetano *et al.*, 2012) and (Teich and

Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 44–48,
Avignon, France, April 23 - 24 2012. ©2012 Association for Computational Linguistics

categories of meanings (feature)	**realisation**
obligation/necessity (obligaton)	*can, could, may*, etc.
permission/possibility/ability (permission)	*must, should*, etc.
volition/prediction (volition)	*will, would, shall*, etc.

Table 1: Categories of modal meanings for feature extraction

Fankhauser, 2010), an English corpus which contains full English scientific journal articles from nine disciplines (see Figure 1). The corpus covers two time periods: the 1970/early 1980s (SaSciTex) and the early 2000s (DaSciTex), and includes ca. 34 million tokens. Our focus is especially on the subcorpora representing contact registers, i.e. registers emerged out of register contact, in our case with computer science: computational linguistics (B1), bioinformatics (B2), digital construction (B3), and microelectronics (B4).

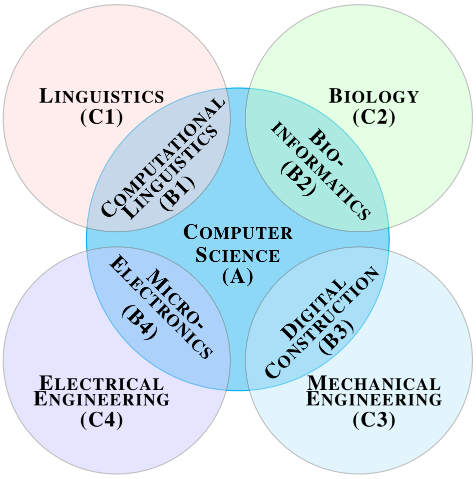
LINGUISTICS (C1)
BIOLOGY (C2)
COMPUTATIONAL LINGUISTICS (B1)
BIO-INFORMATICS (B2)
COMPUTER SCIENCE (A)
MICRO-ELECTRONICS (B4)
DIGITAL CONSTRUCTION (B3)
ELECTRICAL ENGINEERING (C4)
MECHANICAL ENGINEERING (C3)

Figure 1: Scientific disciplines in the SciTex corpus

SciTex is annotated[1] with information on token, lemma, part-of-speech and sentence boundary, as well as further information on text boundary, register information, etc., and can be queried in form of regular expressions by the Corpus Query Processor (CQP), cf. (Evert, 2005).

2.3 Feature Extraction and Analysis

To extract the above described features for the two time slices (1970/80s and 2000s) and for all nine registers of SciTex, we elaborate queries, which include both lexical (based on token and lemma information) and grammatical (based on part-of-speech or sentence boundary information) constraints.

Annotations on the register information allow us to sort the extracted material according to specific subcorpora. This enables the analysis of features possibly involved in creating distinctive registers. Comparing differences and/or commonalities in the distribution of features for A-B-C triples of subcorpora (e.g., A-computer science, B1-computational linguistics, C1-linguistics, cf. Figure 1), we analyse whether the contact disciplines (B-subcorpora) are more similar to computer science (A-subcorpus), the discipline of origin (C-subcorpus) or distinct from both (A and C). The two time periods in SciTex (70/80s vs. 2000s) enable a diachronic analysis. A more fine-grained diachronic analysis is also possible with the information on the publication year annotated in the corpus.

3 Analysing language changes with SPC

3.1 SPC visualisation

Structured Parallel Coordinates (Culy *et al.*, 2011) are a specialisation of the *Parallel Coordinates* visualisation (cf. (d'Ocagne, 1885), (Inselberg, 1985), (Inselberg, 2009)) for representing multidimensional data using a two-dimensional display. *Parallel Coordinates* place data on vertical axes, with the axes lined up horizontally. Each axis represents a separate data dimension and can hold either categorical or numerical data. Data points on different axes are related which is indicated by colored lines connecting all data items belonging to one record.

Targeted to the application to language data, SPC additionally provide for ordered characteristics of data within and across data dimensions. In the *n-grams with frequencies/KWIC*[2] implementations of SPC, ordered axes represent the linear ordering of words in text.

In our analysis of language change based on linguistic features, we are interested in two directions of changes across data sets that can be represented by ordering: changes over time and

[1] Annotations were obtained by means of a dedicated processing pipeline (Kermes, 2011).

[2] www.eurac.edu/linfovis

changes across registers, e.g., from linguistics and computer science to computational linguistics.

3.2 Adjustments to SPC

For the analysis of linguistic features with SPC, we start off with the *n-grams with frequencies* implementation. In analyzing just two time dimensions the ordered aspect of SPC is not as crucial and a similar analysis could have been done with Parallel Coordinates. However, the setup of *n-grams with frequencies* conveniently provides us with the combination of categorical and numerical data dimensions in one display but separated visually. For our diachronic register analysis, we create a *subcorpus comparison* application where the feature under analysis as well as some of the corpus data are placed on the unordered categorical axes, and frequencies for the two time periods are placed on ordered axes with numerical scales. As shown in Figure 2 below, unordered dimensions are followed by ordered dimensions, the inverse situation to *n-grams with frequencies*. To visually support the categorical nature of data on the first three axes, SPC was adjusted to display the connecting lines in discrete colors instead of the default color scale shading from red to blue. To improve the comparability of values on numerical axes, a function for switching between comparable and individual scales was added that applies to all axes right of the separating red line. Figure 2 and 3 present numerical values as percentages on comparable scales scaled to 100.

3.3 Interactive features for analysis

SPC provide a number of interactive features that support data analysis. To highlight and accentuate selected parts of the data, an axis can be put into focus and parts of axes can be selected. Lines are colored according to the axis under focus, and filters apply to the selected portions of axes, with the other data rendered in gray. Users can switch between discrete colors and scaled coloring of connecting lines. The scales of numerical axes can be adjusted interactively, as described above. Hovering over a determined connecting line brings it out as a slightly wider line and gives a written summary of the values of that record.

4 Interpreting Visualisation Results

Visualised structures provided by SPC supply us with information on development tendencies, and thus, deliver valuable material for further interpretation of language variation across registers and time.

To analyse the frequencies of modal meanings (see Table 1) for A-B-C triples of subcorpora, we use the *subcorpus comparison* option of SPC. The interactive functionality of SPC allows us to focus on different aspects and provides us with dynamically updated versions of the visualisation.

First, by setting focus on the axis of modal meanings, the visualisation in Figure 2 shows diachronic changes of the modal meanings from the 1970/80s to the early 2000s. In both time periods the *permission* (blue) meaning is most prominent and has considerably increased over time. The *volition* (green) and *obligation* (orange) meanings are less prominent and we can observe a decrease of *volition* and a very slight decrease of *obligation.*

Second, by setting the axis of the registers into focus and selecting the disciplines one by one, we can explore whether there are changes in the use of modal meanings between the A register, the contact registers (B), and the respective C registers. In Figure 3, for example, computer science and biology have been selected (gray shaded) on the 'disciplines' axis. For this selection, the structures starting from the 'registers' axis represent (1) computer science (blue) being the A register, (2) biology (green) from the C registers, and (3) bioinformatics (orange) from the B registers as the corresponding contact register. In terms of register changes, Figure 3 shows that bioinformatics differs in the development tendencies (a) of *permission* from biology and computer science (less increase than the former, more increase than the latter) and (b) of *obligation* from biology (decrease for biology, whereas nearly stable for bioinformatics and computer science).

5 Conclusion and Future Work

The results described above show that *Structured Parallel Coordinates* provides us with a means for the interactive inspection of complex data sets facilitating our diachronic register analysis. The visualisation allows to gain an overview and detect tendencies by accomodating a complex set of data in one display (nine registers over two time periods for three meanings).

The interactive features of SPC give the possibility to put different aspects of the data into fo-

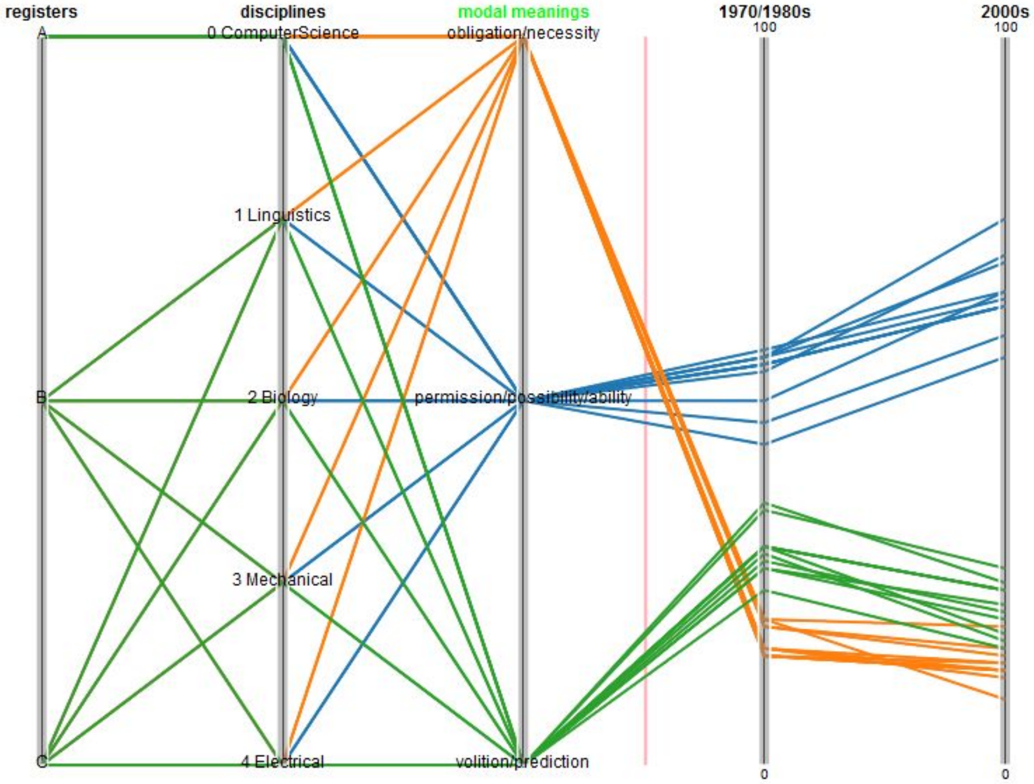

Figure 2: Modal meanings in SciTex in the 1970/80s and 2000s

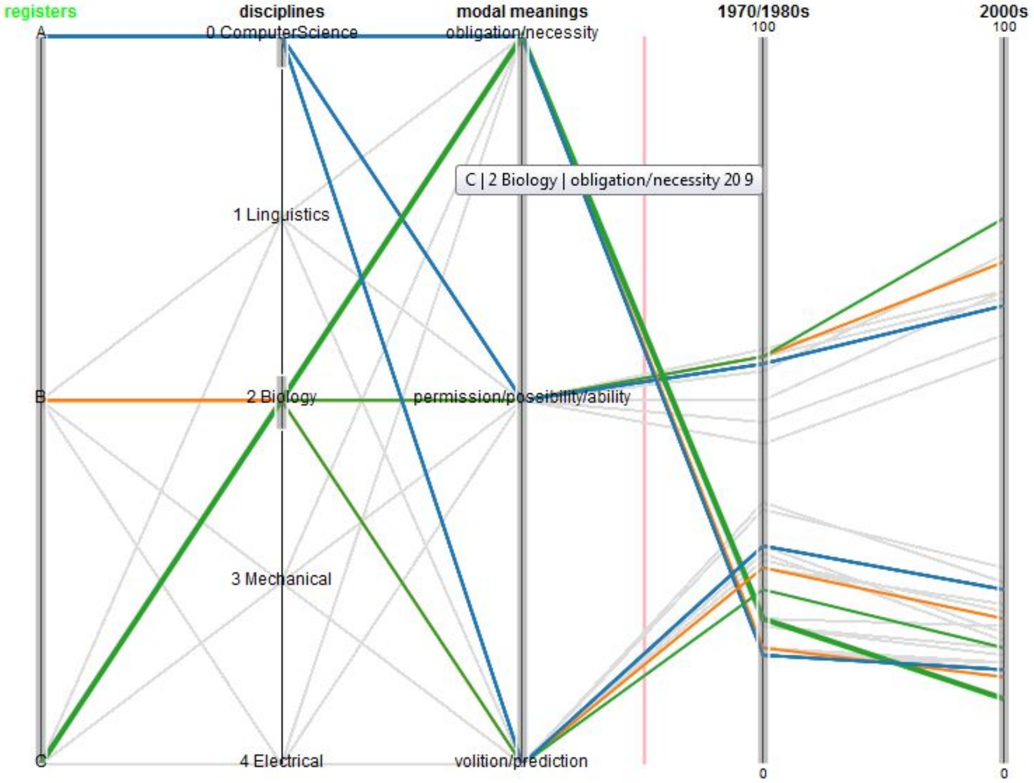

Figure 3: Modal meanings in computer science (A-subcorpus; blue), bioinformatics (from B-subcorpus; orange) and biology (from C-subcorpus; green)

cus, and thus to successively zoom into specific subsets of the data for detailed analyses. In this way, we can determine general tendencies (e.g., increase of *permission* over time) or provide detailed analyses for certain linguistic features and registers by selecting subparts of the data and by highlighting different data dimensions (e.g., comparing changes between different registers).

Future work comprises to use the data obtained from the corpus to feed several different SPC visualisations. For example, the data presented in Figure 2 can also be layed out to place values for registers instead of values for time periods on the numerical axes.

Future analyses will focus on inspecting further tendencies in the feature development for the three contextual variables mentioned in 2.1, e.g., verb valency patterns for *field* or conjunctive relations expressing cohesion for *mode*. We also aim at analysing several linguistic features at the same time to possibly detect feature sets involved in register variation of contact registers. Additionally, a more fine-grained diachronic analysis according to the publication years, which are annotated in the corpus, might also prove to be useful.

From a technical point of view, the issue with fully overlapping lines being displayed in one color only will be tackled by experimenting with semi-transparent or stacked lines. Furthermore, SPC should in the future be expanded by a function for restructuring the underlying data to create different layouts. This could also include the merging of axes with categorical values (e.g., axes *registers* and *disciplines* in Figure 2 above). Furthermore on each data dimension a 'summary' category could be introduced that would represent the sum of all individual values, and would provide an extra point of reference for the analysis. For interactive data analysis, support could be provided to select data items based on crossings or declination of their connecting lines.

References

Douglas Biber. 1995. *Dimensions of Register Variation. A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Douglas Biber. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson ESL.

Chris Culy, Verena Lyding, and Henrik Dittmann. 2011. Structured Parallel Coordinates: a visualization for analyzing structured language data. In *Proceedings of the 3rd International Conference on Corpus Linguistics, CILC-11*, April 6-9, 2011, Valencia, Spain, 485–493.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ekaterina Lapshinova-Koltunski and Elke Teich. 2012. SciTex – A Diachronic Corpus for Analyzing the Development of Scientific Registers. In: Paul Bennett, Martin Durrell, Silke Scheible & Richard J. Whitt (eds.), *New Methods in Historical Corpus Linguistics*. CLIP, Vol. 2, Narr: Tübingen.

Stefan Evert. 2005. The CQP Query Language Tutorial. IMS, Universität Stuttgart.

M.A.K. Halliday and Ruqaiya Hasan. 1989. Language, context and text: Aspects of language in a social semiotic perspective. OUP.

Alfred Inselberg. 2009. *Parallel Coordinates: VISUAL Multidimensional Geometry and its Applications*. New York: Springer.

Alfred Inselberg. 1985. The plane with parallel coordinates. *The Visual Computer* 1(2), pp. 69–91.

Hannah Kermes. 2011. Automatic corpus creation. Manual. Institute of Applied Linguistics, Translation and Interpreting, Universität des Saarlandes, Saarbrücken.

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

Maurice d'Ocagne. 1885. *Coordonnées Parallèles et Axiales: Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèlles*. Paris: Gauthier-Villars.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Harlow: Longman

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