Register impact on change in language use across sciences

Stefania Degaetano-Ortlieb, Saarland University

Studies on written scientific language have shown a diachronic tendency towards structural compression (Biber & Gray 2013) accompanied by informationally dense phrasal constructions (e.g. Halliday & Martin 1993, Mair 2006, Degaetano-Ortlieb & Teich 2016a/b, 2018). Besides general trends in scientific writing, Biber & Gray (2013) have shown how different disciplinary fields differ, e.g. in their use of nominal features reflecting particular compression strategies, indicating the importance of investigating trends in subregisters.

Following this line of research, we analyze all texts (33,479) from the Proceedings of the Royal Society of London, considering their mathematical and biological sciences (Series A and B). We investigate whether over time these fields show differing or common changes in their use of grammatical structures. Common changes will indicate trends of the scientific register in general. Discipline-specific changes will indicate a particular need within a discipline that calls for a particular change. In addition, change related to the scientific register in general might have developed in one of the two fields first and spread to the other.

Methodologically, we take a text-linguistic exploratory perspective, investigating change of use of *n*gram-based parts-of-speech (POS) sequences approximating grammatical structure (ensuring reproducibility and broad coverage of features). By calculating relative entropy *D* (Kullback & Leibler 1951, Fankhauser et al. 2014) between probability distributions of POS sequences (here trigrams) diachronically, we inspect *when* changes occur and *which* structures contribute to change. Specifically, we compare preceding (past 10) and following (future 10) years of a given year, sliding over the time line (see Equation 1) based on the periodization procedure by Degaetano-Ortlieb & Teich (2018). Intuitively, relative entropy allows us to measures how well the future of a given year can be modeled by the past. A rise in relative entropy indicates a period of change, where the future is distinctively different from the past.

$$D(future|past) = \sum_{i=0}^{n} p(trigram_i|future) \log_2 \frac{p(trigram_i|future)}{p(trigram_i|past)} \qquad (1)$$

Ranking trigrams by their contribution to a difference (increase in relative entropy), we obtain those trigrams used most distinctively in the future compared to the past. The highest-ranking trigrams are then mapped to grammatical structures.

Preliminary results (see Figure 1) show that indeed scientific fields (here mathematical and biological) are subject to discipline-specific changes in use of grammatical structures, not reflected by an overall model (gray vs. colored lines). A major change occurs around the mid-1920s. Figure 2 and 3 show complex nominal structures to be involved in this change, which seems to begin in the biological series but having a higher impact in the mathematical ones. Around the mid-1950s, biology shows a second major change – distinctive use of passive in past tense (e.g. *mice <were immunized by> the transplantation*) born from the need to report experiments – not reflected in mathematics.

We will further investigate which other structures are involved in change, their lexical realizations and functional properties. Our work contributes to register analysis, especially of subregisters, as well as data mining techniques, such as register classification (cf. Atkinson 1992, Argamon et al. 2008, Eisenstein et al. 2011, Degaetano-Ortlieb et al. 2014, Teich et al. 2013, 2016, Clarke & Grieve 2017).

References

Argamon, Solomon, Jeff Dodick & Paul Chase. 2008. Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. *Scientometrics 75*(2). 203–238.

Atkinson, Dwight (1992). The evolution of medical research writing from 1735 to 1985: The case of the Edinburgh Medical Journal. *Applied Linguistics 13*(4). 337–374.

Biber, Douglas & Bethany Gray. 2013. Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* 41. 104–134.

Clarke, Isobelle & Jack Grieve. 2017. Dimensions of abusive language on twitter. In *Proceedings of the First Workshop on Abusive Language Online.* Vancouver, Canada. 1–10. ACL

Degaetano-Ortlieb, Stefania, Peter Fankhauser, Hannah Kermes, Ekaterina Lapshinova-Koltunski, Noam Ordan & Elke Teich. 2014. Data mining with shallow vs. linguistic features to study diversification of scientific registers. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, Iceland. 1327–1334. ELRA.

Degaetano-Ortlieb, Stefania & Elke Teich. 2016a. Information-based modeling of diachronic linguistic change: From typicality to productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Berlin, Germany. 165–173. ACL.

Degaetano-Ortlieb, Stefania, Hannah Kermes, Ashraf Khamis & Elke Teich. 2016b. An information-theoretic approach to modeling diachronic change in scientific English. In Suhr, Carla, Terttu Nevalainen & Irma Taavitsainen (eds). *From Data to Evidence in English Language Research*. Leiden: Brill.

Degaetano-Ortlieb, Stefania & Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* at COLING2018. Santa Fe, NM. 22–33. ACL.

Eisenstein, Jacob, Noah A. Smith & Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* Portland, Oregon. 1365–1374. ACL.

Peter Fankhauser, Jörg Knappen & Elke Teich. 2014. Exploring and visualizing variation in language resources. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, Iceland. 4125–4128. ELRA.

Halliday, M.A.K. & J.R. Martin. 1993/1996. *Writing science: Literacy and discursive power*. London: Falmer Press.

Kullback, Solomon & Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1). 79–86.

Mair, Christian. 2006. *Twentieth-century English: History, variation and standardization*. Cambridge: CUP.

Teich, Elke, Stefania Degaetano-Ortlieb, Hannah Kermes & Ekaterina Lapshinova-Koltunski. 2013. Scientific registers and disciplinary diversification: A comparable corpus approach. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora.* Sofia, Bulgaria. 59–68. ACL.

Teich, Elke, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes & Ekaterina Lapshinova-Koltunski. 2016. The linguistic construal of disciplinarity: A data mining approach using register features. *Journal of the Association for Information Science and Technology (JASIST).* 67(7). 1668–1678.

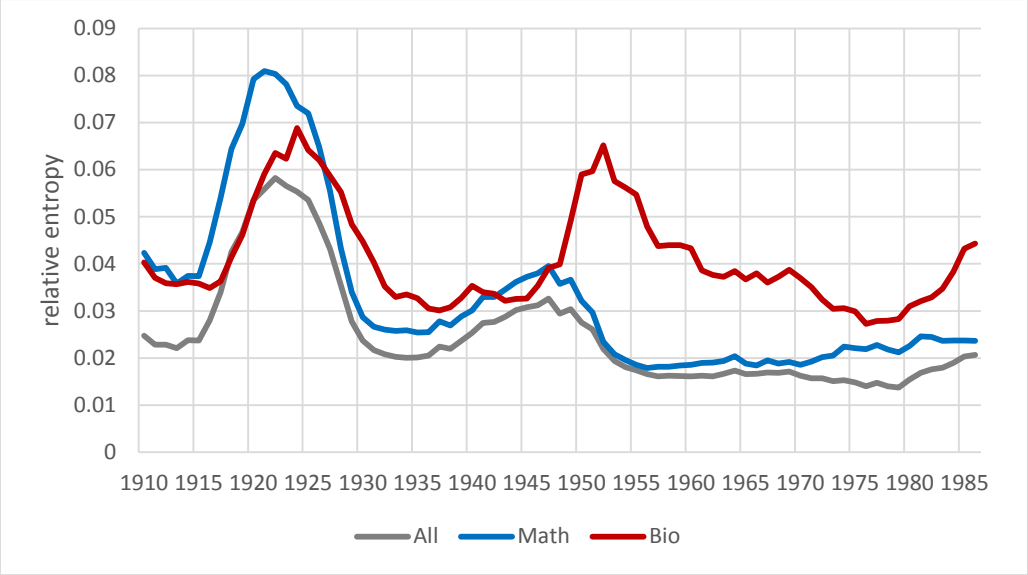

Figure 1: Relative entropy comparing at each year the previous and following 10 years.

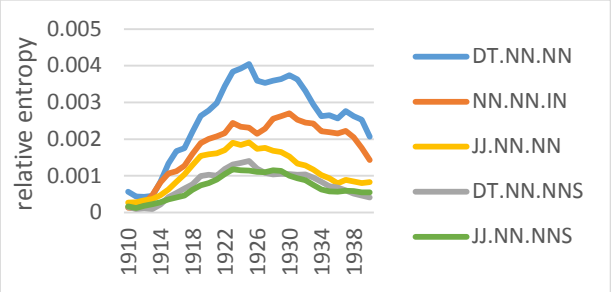

Figure 2: Top 5 distinctive POS trigrams math (1920s)

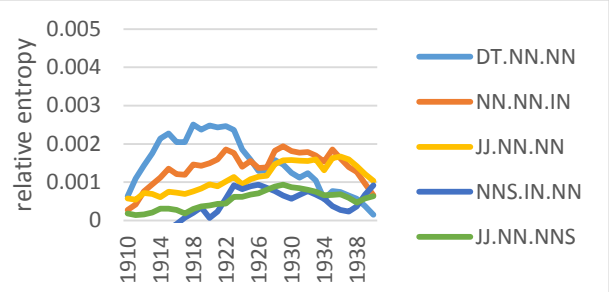

Figure 3: Top 5 distinctive POS trigrams biology (1920s)

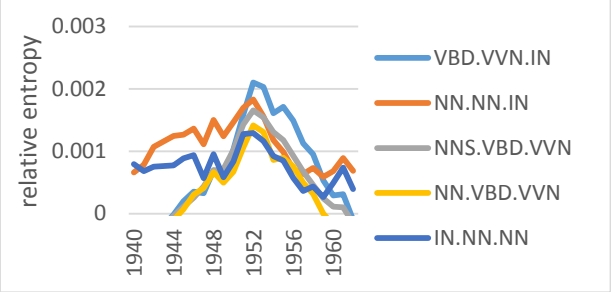

Figure 4: Top 5 distinctive POS trigrams biology (1950s)

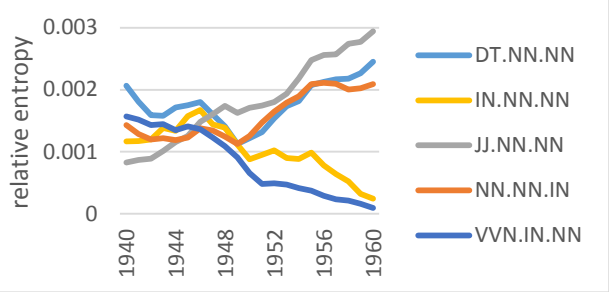

Figure 5: Top 5 distinctive POS trigrams math (1950s)