Hybridization effects in literary texts

Stefania Degaetano-Ortlieb, Saarland University

Introduction

We present an analysis of subregisters, whose differentiation is still a difficult task due to their hybridity reflected in conforming to a presumed "norm" and encompassing something "new". We focus on texts at the interface between what Halliday (2002: 177) calls two opposite "cultures", literature and science (here: science fiction texts).

Texts belonging to one register will exhibit similar choices of lexico-grammatical features. Hybrid texts at the intersection between two registers will reflect a mixture of particular features (cf. Degaetano-Ortlieb et al. 2014, Biber et al. 2015, Teich et al. 2013, 2016, Underwood 2016). Consider example (1) taken from Mary Shelley's Frankenstein. While traditionally grounded as a literary text, it shows a *registerial nuance* from the influential register of science. This encompasses phrases (bold) also found in scientific articles from that period (e.g. in the Royal Society Corpus, cf. Kermes et al. 2016), verbs related to scientific endeavor (e.g. *become acquainted, examine, observe, discover*), and scientific terminology (e.g. *anatomy, decay, corruption, vertebrae, inflammable air*) packed into complex nominal phrases (underlined). Note that features marking this registerial nuance include not only lexical but also grammatical features.

(1) ***I became acquainted*** *with* *the science of anatomy*, *but this was not sufficient;* ***I must also observe*** *the natural decay and corruption of the human body. […] Now* ***I was led to examine*** *the cause and progress of this decay.* ***I succeeded in discovering*** *the cause of generation and life.* (Frankenstein, Mary Shelley, 1818/1823).

Thus, we hypothesize that hybrid registers while mainly resembling their traditional register in the use of lexico-grammatical features (H1 register resemblance), will also show particular lexico-grammatical nuances of their influential register (H2 registerial nuance). In particular, we are interested in (a) variation across registers to see which lexico-grammatical features are involved in hybridization effects and (b) intra-textual variation (e.g. across chapters) to analyze in which parts of a text hybridization effects are most prominent.

Data and methods

We take our data from the Novel450 Corpus obtained from the .txtLab[1] (Piper 2016; approx. 22 million tokens) ranging from 1771-1929 and the English scientific Royal Society Corpus obtained from the Clarin-D repository[2] (RSC v4.0; Kermes et al. 2016; approx. 34 million tokens from 1665-1869).

In pursuing answers for H1 (register resemblance) and H2 (registerial nuance), we need a measure of how much hybrid registers diverge from or converge with their traditional and influential registers. Based on the approach of Fankhauser et al. 2014, we use Kullback-Leibler divergence (KLD; cf. Equation 1), a widely applied measure for

[1] https://figshare.com/articles/txtlab_Novel450/2062002/3

[2] https://fedora.clarin-d.uni-saarland.de/rsc/

comparison between two probability distributions of given linguistic items (e.g. words, parts of speech, n-gram features, etc.) (cf. Hughes et al. 2012, Klingenstein et al. 2014, Bochkarev et al. 2014, Degaetano-Ortlieb et al. 2016, Degaetano-Ortlieb and Teich 2016, 2017, 2018, Degaetano-Ortlieb and Piper 2019).

$$D(A||B) = \sum_i p(item_i|A) \log_2 \frac{p(item_i|A)}{p(item_i|B)} \quad (1)$$

$p(item_i|A)$ is the probability of the ith item in corpus *A* and $p(item_i|A)$ of that item in corpus *B*. $D(A||B)$ gives the amount of additional bits needed to encode e.g. words distributed according to *A* by the words' distributions in corpus *B*. The higher the amount of bits, the more the two corpora diverge based on their words' distribution. Moreover, KLD accounts for asymmetry in a comparison (i.e. $D(A||B)$ is not necessarily equal to $D(B||A)$[3]). Difference in vocabulary size is controlled for by using ngram language models with Jelinek-Mercer smoothing (lambda at 0.05; cf. Zhai and Lafferty 2004).

We also inspect divergence of individual items (*pointwise* KLD ; cf. Tomokiyo 2003) to see which items contribute most to the overall divergence (cf. Equation 2). The results are interpretable, as to which features contribute how much to a distinction between a range of 1 to -1. The closer to 1, the more indicative a feature is of class *A*, while the closer to -1, the more the feature is indicative of class *B*.

$$D(A||B) = p(item_i|A) \log_2 \frac{p(item_i|A)}{p(item_i|B)} \quad (2)$$

To model hybridization effects within texts (intra-textual variation), we adapt the approach of Degaetano-Ortlieb and Teich (2018; used for longitudinal diachronic analysis) to the sequential analysis of single texts, comparing a chapter with a previous one by KLD.

Pilot study

Based on register theory (Biber et al. 1999, Halliday 2004), we consider variation in the TRANSITIVITY system by extracting the main process types (material, mental, verbal, behavioral) from three texts according to the three registers: Mrs.Dalloway (fiction: F), Frankenstein (science fiction: SF) and Fulton1924 (science: S) based on verb lists taken from Halliday (2004)[4].

First, we assess divergence between the fiction and scientific text: $D(Fulton1924||MrsDalloway)=0.546$ bits versus $D(MrsDalloway||Fulton1924)=1.545$ bits. Divergence shows that science is better modeled with fiction than vice versa, i.e. Mrs.Dalloway exhibits a particular use of process types that is not well captured in the probability distribution of Fulton's text. Considering contribution of each process type (pointwise KLD; see Figure 1), material verbs are distinctive for Fulton, while mental,

[3] E.g., science fiction texts might be less well modeled by fiction texts due to their hybrid nature encompassing additional nuances from science (e.g. particular 'scientific' verbs might not be used by literary texts (see also example (1)). Literary text might be well captured by science fiction texts, as they encompass most of the lexico-grammatical features of literary texts.

[4] Examples of verbs used: material: *develop, form, build, make*; mental: *perceive, sense, see, notice*; verbal: *say, tell, praise, report*; behavioral: *look, watch, listen, worry*.

verbal and especially behavioral verbs are distinctive for Mrs.Dalloway. This is in line with Piper (2016) showing perception processes as particularly distinctive of fiction texts, reflected here in the distinctive use of mental and behavioral verbs (a conscious physical act involved in perception, e.g. *see, watch*, *look*; cf. Thompson 2004).

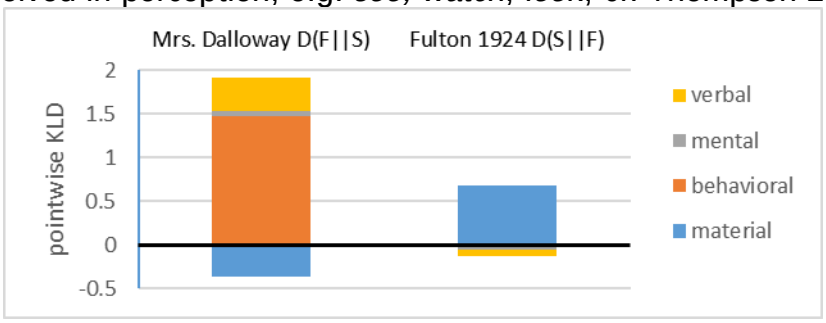

Figure 1: Contribution of process types for science (Fulton1924) and fiction (Mrs.Dalloway)

Considering register resemblance (H1), we assume science fiction to diverge less from fiction than science. We test this by using (a) features distinctive of the traditional register measuring divergence between science fiction and fiction: *D(Frankenstein||Mrs.Dalloway)=0.112* bits; and (b) features distinctive of the influential register measuring divergence between science fiction and science: *D(Frankenstein||Fulton1924)=0.644* bits, which is higher. Thus, science fiction is more similar to fiction than to science in the use of process types.

For registerial nuance (H2), we assume science fiction to better model science than literary text model science. In fact *D(Fulton1924||Frankenstein)=0.276* bits and *D(Fulton1924||Mrs.Dalloway)=0.546* bits, i.e. science fiction shares some process type use with science. Comparing process types distinctive of science fiction vs. fiction and science vs. fiction (see Figure 2), material verbs (blue) are those properties adopted from science in science fiction.

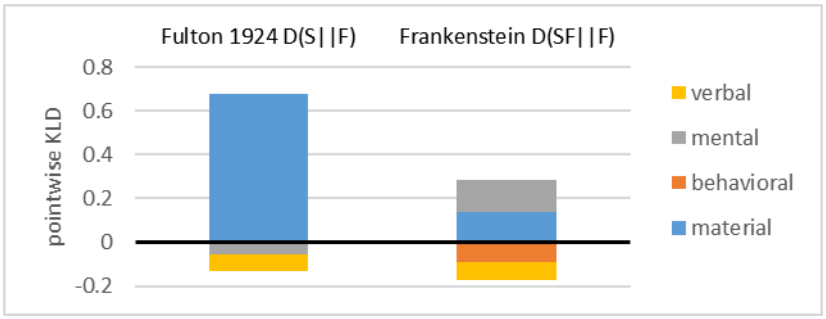

Figure 2: Contribution of process types: science (Fulton 1924) vs. science fiction (Frankenstein)

Moreover, we investigate the degree of register resemblance and registerial nuance for particular novels, in other words where hybridization effects are most prominent within a text. We assume that as a text evolves and features of science are introduced in science fiction, the lexico-grammatical setup of features will change accordingly. By considering chapters of the Frankenstein novel, we sequentially slide from one chapter to another, comparing probability distributions of process types with KLD. Here, we

focus on registerial nuance considering process types possibly adopted in science fiction from science. We limit the investigation to material and mental verbs, being the most frequent process types in science and thus possibly adopted in science fiction. Based on distinctive verbs used in science and fiction[5], we compare material and mental process types across chapters of the Frankenstein novel. A chapter is compared to its preceding chapter by KLD. Figure 3 shows that the use of scientific mental verbs (mental-S) is distinctive in Chapter IV (reddish square) when compared to its preceding chapter (see examples (2) and (3)). These mental verbs reflect scientific mental activities (e.g. *examine* in example (2) and *consider* in example (3)). In fact, Chapter IV (from which example (1) is taken) marks Victors love for natural science and his work on creating a new being, i.e. a particular point in the science fiction novel where science is woven into.

(2) *I paused, examining and analysing all the minutiæ of causation, as exemplified in the change from life to death [...]* (Chapter IV).

(3) *yet , when I considered the improvement which every day takes place in science and mechanics [...]* (Chapter IV).

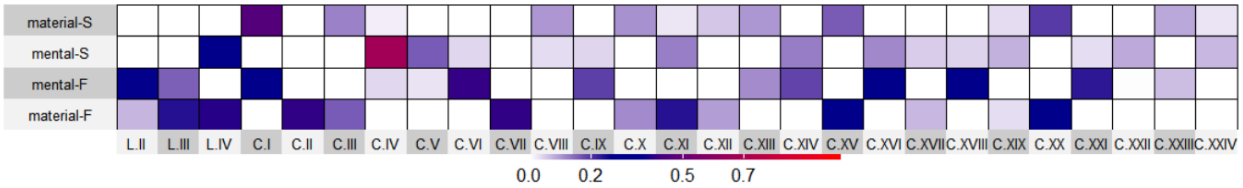

Figure 3: Pointwise KLD of process types across chapters in Frankenstein (red: high contribution)

Summary and future work

We have shown how hybrid registers are affected in linguistic terms by the influential register being able to model inter- as well as intra-textual variation using an asymmetric measure of divergence and considering particular lexico-grammatical features. While exemplified here on process types, we aim to adopt this methodology for the analysis of hybrid registers considering a wide range of distinctive lexico-grammatical features of the traditional and influential registers. As a theoretical framework to draw features from, we use besides register theory (Biber et al. 1999) also Systemic Functional Linguistics (Halliday 2004), particularly well-suited for register analysis. Moreover, we are experimenting with the use of word embeddings to enhance register-adaptation of lexical and semantic features.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman, Harlow, UK.

Biber, D., Egbert, J. and Davies, M. (2015). Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora* 10.1: 11-45.

[5] Here verb lists are based on the most frequent material and mental verbs in the RSC corpus and the COCA fiction subcorpus.

Bochkarev, V. Solovyev, V.D. and Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface*, 11(101).

Degaetano-Ortlieb, S., Fankhauser, P., Kermes, H., Lapshinova-Koltunski, E., Ordan, N. and Teich, E. (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, pages 1327-1334.

Degaetano-Ortlieb, S., Kermes, H., Khamis, A. and Teich, E. (2016). An information-theoretic approach to modeling diachronic change in scientific English. In Suhr, C., Nevalainen, T. and Taavitsainen, I. (eds). *From Data to Evidence in English Language Research*. Leiden: Brill. 258–281.

Degaetano-Ortlieb, S. and Teich, E. (2016). 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, pages, 165-173. ACL.

Degaetano-Ortlieb, S. and Teich, E. (2017). Modeling intra-textual variation with entropy and surprisal. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at ACL 2017*, Vancouver, Canada, pages 68-77. ACL.

Degaetano-Ortlieb, S. and Teich, E. (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 COLING*, pages 22–33, Santa Fe, NM, USA. ACL.

Degaetano-Ortlieb, S. and Piper, A. (2019). The scientization of literary study. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at NAACL,* Minneapolis, MN. ACL.

Fankhauser, P., Knappen, J., and Teich, E. (2014). Exploring and visualizing variation in language re-sources. In *Proceedings of the 9th LREC*, pages 4125–4128, Reykjavik. ELRA.

Halliday, M.A.K. (2002). *Linguistic Studies of Text and Discourse*. Continuum, New York.

Hughes, J.M., Foti, N.J., Krakauer, D.C and Rockmore, D.N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. In *Proceedings of the National Academy of Sciences*,109(20):7682–7686.

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J., and Teich, E. (2016). The Royal Society Corpus: from uncharted data to corpus. In *Proceedings of the LREC 2016*. Portoroz, Slovenia, pages 1928-1931. ELRA.

Klingenstein, S., Hitchcock, T. and DeDeo, S. (2014). The civilizing process in London's Old Bailey.In *Proceedings of the National Academy of Sciences*, 111(26):9419–9424.

Piper, A. (2016). Fictionality. *Journal of Cultural Analytics*. Dec 20.

Teich, E., Degaetano-Ortlieb, S., Kermes, H. and Lapshinova-Koltunski, E. (2013). Scientific registers and disciplinary diversification: A comparable corpus approach. In *Proceedings of 6th Workshop on Building and Using Comparable Corpora (BUCC)*. Sofia, Bulgaria, August, pages 59-68. ACL.

Teich, E., Degaetano-Ortlieb, S., Fankhauser, P., Kermes, H. and Lapshinova-Koltunski, E. (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.

Thompson, G. (2004). *Introducing Functional Grammar*. Hodder Arnold, London, 2nd edition.

Tomokiyo, T. and Hurst, M. (2003). A language model approach to keyphrase extraction. In *Proceedings of the ACL MWE '03 Workshop*, pages 33–40, Stroudsburg, PA, USA. ACL.

Underwood, T. (2016). The life cycles of genres. *Journal of Cultural Analytics*. May 23.

Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214.