The March of Data: English Linguistics across Disciplinary Borders

Workshop organizers:
Jukka Tyrkkö (Linnaeus University)
Steven Coats (University of Oulu)
Veronika Laippala (University of Turku)

Multiple modals in the wild: A study of 24,530 multiple modal sequences in naturalistic North American speech

Steven Coats

Multiple modals are a well-known feature of dialectal speech in the Southern United States (Bernstein 2003; Di Paolo 1989; Fennell & Butters 1996; Mishoe & Montgomery 1994; Montgomery 1989, 1998), but questions remain as to their geographical distribution, a fact due not only to their rareness in spoken language, but also to the limited scope and heterogenous nature of data collection in previous research. In this study, double and triple modal sequences are considered in the Corpus of North American Spoken English (CoNASE; Coats 2019, Coats forthcoming), a 1.2 billion word corpus of Automatic Speech Recognition (ASR) transcripts from YouTube channels of local government entities in the United States and Canada, the first large corpus of geographically localized speech for North American Englishes.

After identifying potential multiple modal sequences using regular expressions, a table was created in which each of the 24,530 double- or triple modal instances in the corpus is linked to a URL for the corresponding video at the time of its utterance. A subset of sequences was then manually inspected and, based on local discourse coherence and prosodic criteria, classified as "true" double modals or as various types of false positives, such as homonyms, overlaps, disfluencies, or transcript errors. While a significant proportion of multiple modal sequences in the data represent corrections, disfluencies, or other false positives, particularly for some types, the manual verification procedure allowed the frequency and geographical extent of authentic multiple modals to be estimated in the corpus, especially for frequent types.

The contributions of the study are both methodological and empirical: First, CoNASE demonstrates that "messy" ASR data can be profitably used for the investigation of linguistic phenomena – a methodological development that will undoubtedly continue to grow in importance in coming years. The study demonstrates that data from social media or video sharing ecosystems such as YouTube can be organized in a manner that makes specific phenomena of interest immediately verifiable in their original naturalistic context for the researcher and the larger research community, in contrast to previous studies, in which the data upon which research has been conducted is mostly not verifiable. Second, the study documents and provides links to naturalistic use of a number of forms that have not previously been identified in the literature on American double modals, especially first-tier deontic or dynamic double modals such as *must may*, *must might*, *should might*, *will can*, and others. Third, while most multiple modals, such as the archetypal *might could,* are indeed most frequent in the Southeastern States of the US, they can be found in speech from across the North American continent, including in locations where they have until now not been documented, such as Midwestern, New England, Western and Northern US states, and in Canada. Fourth, the nature of the underlying data, which is more consistent than the heterogenous aggregate data employed in previous studies, allows for the mapping of relative frequencies of some of the more common double modals using spatial autocorrelation. Fifth, the relatively large proportion of disfluencies in the sample, particularly for

some modal combinations, offers new perspectives for further research in cognitive linguistics or conversation analysis, particularly for the study of corrections or “self-repair”.

Finally, the study offers a preliminary interpretation of the observed frequency results: Rather than being restricted as lexicalized forms to particular regional or social varieties, double modals may represent the manifestation of a generalized mechanism by which tentative epistemicity can be expressed in North American Englishes, alongside other grammatical and lexical options. In light of the ongoing reorganization of the semantic space of the modal verbs in American and Canadian Englishes (Dollinger 2008; Facchinetti et al. 2003; Leech et al. 2009; Myhill 1995) and the resulting lack of clarity, for many speakers, as to the appropriate choice of epistemic modal in some pragmatic contexts, double modals may be a means by which some North American speakers spontaneously realize tentative epistemicity, particularly in the context of careful discussion or negotiation.

References

Bernstein, Cynthia. 2003. Grammatical features of southern speech: Yall, might could, and fixin to. In Stephen J. Nagle & Sara L. Sanders (eds.), *English in the Southern United States*, 106–118. Cambridge: Cambridge University Press.

Coats, Steven. 2019. A Corpus of regional American language from YouTube. In Costanza Navarretta et al. (eds.), *Proceedings of the 4th Digital Humanities in the Nordic Countries Conference*, 79–91. Aachen, Germany: CEUR.

Coats, Steven. (in review). Dialect Corpora from YouTube.

Di Paolo, Marianna. 1989. Double modals as single lexical items. *American Speech* 64(3). 195–224.

Dollinger, Stefan. 2008. *New-dialect formation in Canada: Evidence from the English modal auxiliaries*. Amsterdam: John Benjamins.

Facchinetti, Roberta, Frank Palmer & Manfred Krug (eds.). 2003. *Modality in Contemporary English*. Berlin & New York: Mouton de Gruyter.

Fennell, Barbara A. & Ronald R. Butters. 1996. Historical and contemporary distribution of double modals in English. In Edgar W. Schneider (ed.), *Focus on the USA: Varieties of English around the World*, 265–88. Amsterdam: John Benjamins.

Leech, Geoffrey, Marianne Hundt, Christian Mair & Nicholas Smith. 2009. *Change in contemporary English: A grammatical study*. Cambridge: CUP.

Mishoe, Margaret & Michael Montgomery. 1994. The pragmatics of multiple modal variation in North and South Carolina. *American Speech* 69(1). 3–29.

Montgomery, Michael B. 1989. Exploring the roots of Appalachian English. *English World-Wide* 10(2). 227–278.

Montgomery, Michael B. 1998. Multiple Modals in LAGS and LAMSAS. In Michael B. Montgomery & Thomas E. Nunnally (eds.), *From the Gulf States and Beyond: The Legacy of Lee Pederson and LAGS*, 90–122. Tuscaloosa: University of Alabama Press.

Myhill, John. 1995. Change and continuity in the function of the American English modals. *Linguistics* 33. 157–211.

Using machine learning to predict keywords

Veronika Laippala, Jesse Egbert, Douglas Biber & Aki-Juhani Kyröläinen

The large-scale linguistic analysis of texts involves two general approaches: the descriptive approach, typical of corpus linguistics, describes text varieties, such as registers (Biber and Conrad, 2009), whereas the predictive approach, usually employed in natural language processing, applies machine learning (ML) to predict text categories. Both approaches have their

strengths that make them useful for scientific purposes in general and text analysis in particular. Descriptive studies offer theoretically based understanding of the data and detailed information about the text characteristics, whereas predictive models reach generally high accuracies, and thus give reliable information about the data. Furthermore, predictive models allow the empirical testing and comparison of different theoretically based hypothesis (see Breiman 2001; Schmueli 2010; Tagliamonte and Baayen 2012).

In this presentation, we compare the linguistic findings resulting from the descriptive and predictive approaches. Specifically, we: 1) apply ML to predict register differences with respect to a set of discriminative words (DWs), 2) analyze whether the predicted DWs can be used to linguistically describe the texts, 3) examine DWs generated with different parameters, and 4) compare DWs to the results of a traditional keyword analysis commonly used in corpus linguistics (Scott 1997; Baker 2004).

As data, we used 23 registers and 25,038 documents extracted from the Corpus of Online Registers of English (Biber et al., 2015). We trained a linear support vector machine (Joachims, 1998; Vapnik, 1998) with the registers as the response variable and the document words as features. The model estimated weights were used to identify the set DWs (Guyon and Elisseef, 2003).

In the analysis, we first identify the DWs that best discriminate between two registers: Sports report and Advice. We compare DWs generated with different parameters. In particular, we evaluate DWs generated with various levels of minimum and maximum word frequency and analyze the usage or not of tf-idf (term frequency—inverse document frequency; Jones, 1972) to assign more importance to words occurring frequently in a small number of documents. Finally, we compare the sets of DWs to keywords produced with frequency-based log-likelihood (Scott 1997).

We demonstrate that DWs carry linguistic meaning and they can be used to describe text characteristics. Furthermore, thanks to the predictive power of the SVM, we can compare the predictive importance of the DWs for different register classes. Finally, our analysis shows that DWs generated with tf-idf do not generalize very well to describe the register classes as a whole. Thus, despite being commonly applied in information retrieval, this weighing scheme has limitations in text description.

References

Biber, Douglas and Conrad, Susan (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, Douglas, Egbert, Jesse and Davies, Mark. (2015). Exploring the Composition of the Searchable Web: A Corpus-based Taxonomy of Web Registers. *Corpora* 10(1):11-45.

Breiman, Leo. (2001). Statistical modeling: The two cultures. *Statist. Sci.* 16: 199–215.

Egbert, Jesse, and Biber, Douglas. (2018). Incorporating text dispersion into keyword analyses. *Corpora* 14(1):77-104. https://doi.org/10.3366/cor.2019.0162

Guyon, Isabelle and Elisseeff, Andree. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research* 3:1157–1182.

Joachims, Thorsten. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning,* 137–142, London: Springer-Verlag.

Jones, K.S. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1), 11–21.

Scott, Michael. (1997). PC analysis of key words - and key words. *System* 25(2): 233-45.

Shmueli, Galit. (2010). To Explain or to Predict?. *Statist. Sci*., 25(3): 289--310.

Tagliamonte, Sali & Baayen, Harald. (2012). Models, forests and trees of York English: Was/were variation as a case study for statistical practice. Language Variation and Change. 24(2).

Vapnik, Vladimir. (1998). *Statistical learning theory*. New York: Wiley Interscience.

The visualisation and evaluation of semantic and conceptual maps

Gerold Schneider

Like other social sciences, linguistics is partly turning into a subdiscipline of data science and statistics, and data-driven approaches are used increasingly. They can be used to illustrate semantic and conceptual changes across time, which mirror lexical linguistic changes as well as historic and social developments for Digital Humanities applications. We particularly address the question of how semantic relations can be visualized onto maps. In this, we rely on and compare different machine learning approaches, thus addressing thematic fields 2 to 4 of the workshop.

The paper investigates linguistic, historical and social developments in the COHA corpus. We show a conceptual map generated via Kernel Density Estimation, visualized in *gephi*[1], in Figure 1. We compare Document Classification (e.g. Jurafsky and Martin 2009), Topic Modeling (Blei 2012), Distributional Semantics (Firth 1957, Sahlgren 2006, Mikolov et al. 2018) and Kernel Density Estimation[2].

Systematic evaluations are sought for, but difficult to attain. While the baseline method of document classification is straightforward to evaluate, for example by measuring the accuracy of predicting the 50-year period of a text, evaluations of Topic Modeling exist, but they are more contested. They measure semantic similarity (Röder et al. 2015), while we rather want to discover associations (Fitzmaurice et al. 2007). Distributional Semantics and Kernel Density Estimation are even harder to evaluate automatically, but as their visualisations are similar, they can be compared.

Preliminary results indicate that each method brings different aspects to the surface, but that they are also sensitive to pre-processing, for example the size of the pseudo-documents has a large influence. Also the different visualization methods and programs (e.g. *gephi* vs. *visone*[3]) can lead to considerable differences. Document Classification delivers meaningful lexical features, both linguistic and semantic, but very long lists need to be sifted, and the level of false positives is high. Topic Modeling manages to abstract from words to concepts, is easier to interpret, but the random component of the frequent LDA method needs to be assessed critically. While societal trends emerge, linguistic changes are hardly visible. Distributional Semantics discovers synonyms only if very large amounts of data are available, otherwise it tends to get sidetracked by individual documents (which is a problem) or reports associations (which turns out to a plus for the analysis of social and historical trends). Kernel Density Estimation performs reasonably well with small amounts of data, and clusters by associations rather than synonymity which is very apt for the investigation of social trends, but it tends to invite overinterpretation, and is less meaningful in the central parts of the map.

References

Blei, David. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Firth, John Rupert. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, 1-32.

[1] https://gephi.org

[2] https://github.com/davidmcclure/textplot

[3] https://visone.info

Fitzmaurice, Susan, Justyna Robinson, Marc Alexander, Iona Hine, Seth Mehl, and Fraser Dallachy. 2017. Linguistic DNA:
Investigating Conceptual Change in Early Modern English Discourse, *Studia Neophilologica* 89:sup1, 21-38.
Jurafsky, Daniel, and James H. Martin. 2009. *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. 2nd edn. Upper Saddle River, NJ: Prentice-Hall.
Röder, Michael, Andreas Both and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. Proceedings of *WSDM'15,* February 2–6, 2015, Shanghai, China.
Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional Analysis to represent syntagmatic and paradigmatic relations between words in high- dimensional vector spaces.* Ph.D. thesis, Stockholm University.
Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in Pre-Training Distributed Word Representations. https://arxiv.org/abs/1712.09405

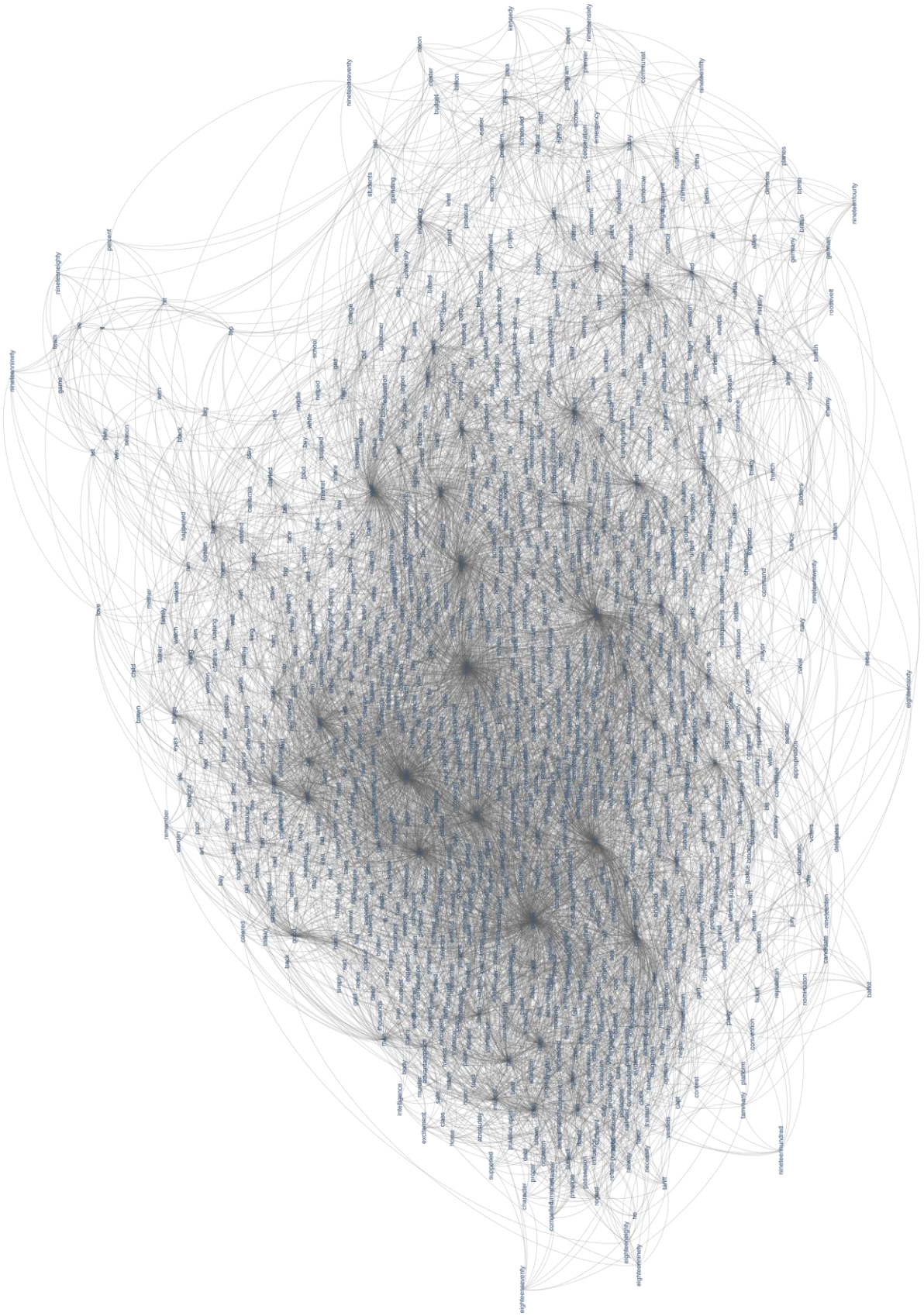

Figure 1. Conceptual map generated via Kernel Density Estimation, visualized in gephi

Size matters: An algorithm-based approach to social networks

Masoud Fatemi & Mikko Laitinen

This presentation focuses on social networks and explores network size as a key determinant in the network theory in sociolinguistics. Social networks are characterized by ties of varying strength (strong or weak), and they bear relevance to language change. Weak-ties promote innovation diffusion, whereas strong ties lead to norm-enforcing communities that resist change (Milroy 1987). Most sociolinguistic studies are, however, based on small ego networks of 30–50 individuals (Milroy & Milroy 1992: 5), and it has been argued that the concept of a network "cannot be easily operationalized in situations where the population is socially and/or geographically mobile" (Milroy 1992: 177). Empirical evidence suggests however that an average human maintains networks well over one hundred nodes (McCarthy et al. 2001), which means that previous sociolinguistic research has only covered a fraction of possible network sizes.

We have two research questions. First, we test the extent to which computational data mining and machine learning of complex metadata in Twitter make it possible to extract large networks of mobile individuals and to assign strength labels to these networks. Second, we specifically concentrate on the effect of network size on the validity of the social network theory by investigating the diffusion of a linguistic innovation (semi-modal *NEED to* + V) in large networks of varying strength. The objective is to test if the difference between weak and strong ties disappears once the network size grows to become large.

Our algorithm-based approach uses mutual interaction parameters in Twitter (obtained through the Twitter API), and the empirical part illustrates how these metadata can be used to assign labels (weak or strong) to networks and to model information flow within networks. The algorithms are widely used in other fields such as graph theory, set theory, and machine learning, but they have not so far been applied in variationist sociolinguistics. The methods make use of centrality measures of betweenness (Freeman 1977) and closeness (Perez & Germon 2016), and we also use Jaccard Similarity Coefficient, which is a symmetric measure that calculates the similarity between two sets. Moreover, we introduce two novel purpose-built methods based on weights of account activity and a method of disjointness, which enables us to estimate how well the social network nodes are connected if the ego node is removed.

Our results show that the methods enable us to extract large digital networks with differing qualities (both weak and strong-ties). These networks can not only be visually confirmed (Figure 1 below), but also quantitatively verified. The methods scale up and using data from over 500 real networks, we illustrate how the distinction between weak and strong-tie networks levels for large networks (>100 nodes).

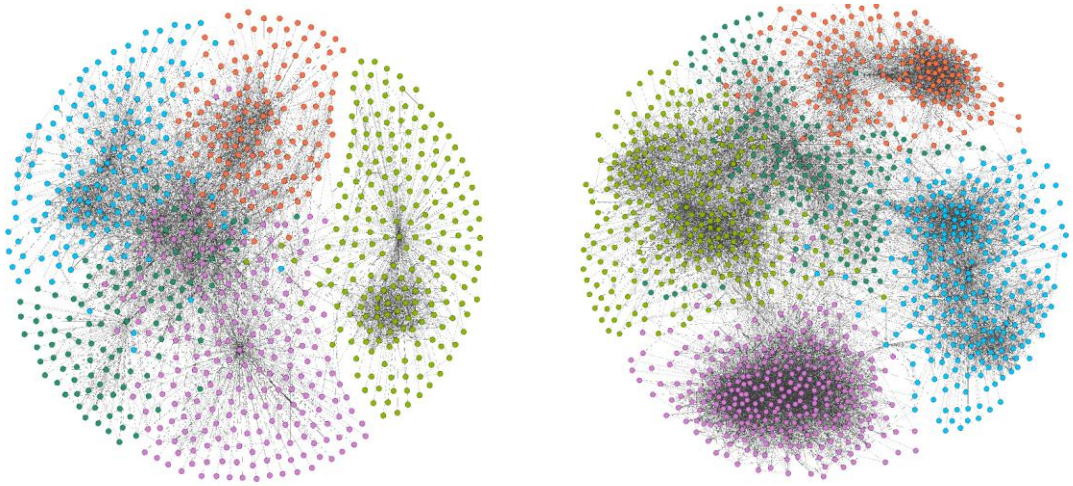

Figure 1. Visualizing weak (left) and strong-tie (right) networks extracted through the algorithms

References

Freeman, Linton C. 1977. A set of measures of centrality based on betweenness. *Sociometry*, 40, 35–41.

Milroy, James. 1992. *Linguistic Variation and Change*. Oxford: Blackwell.

Milroy, Lesley & James Milroy. 1992. Social network and social class: Toward an integrated sociolinguistic model. *Language in Society* 21, 1–26.

Milroy, Lesley. 1987. *Language Change and Social Networks*. 2nd edition. Oxford: Blackwell.

Perez, Charles & Rony Germon. 2016. Graph Creation and Analysis for Linking Actors: Application to Social Data. *Automating Open Source Intelligence*. Chapter 7, 103–129. https://doi.org/10.1016/B978-0-12-802916-9.00007-5.

Sexual dimorphism in language, and the gender shift hypothesis of homosexuality

Severi Luoto

Psychological sex differences have been studied scientifically for more than a century, yet linguists still debate about the existence, magnitude, and causes of such differences in language use. Sex difference research is conducted at increasing rates in psychology, behavioural science, neuroscience, genetics, and biomedical science, but the way in which psychological sex differences may be reflected in language use has received surprisingly limited attention. The aim of my doctoral research has been to connect and synthesise contemporary research in psychology, cognitive science, developmental and behavioural neuroscience, evolutionary science, and computational linguistics by analysing sex differences and sexual orientation differences in a corpus of 694 novels, and comparing the findings with existing research from those fields. I sought to find out whether psychological sex differences and sexual orientation differences—as reported in psychological research on contemporary people—replicate when examining the psycholinguistic outputs of authors of literary fiction living decades and centuries ago. I made predictions for psycholinguistic sex differences and sexual orientation differences based on prior research, both theoretical and empirical, conducted in the abovementioned fields.

I collected a sample of 694 novels written mainly by British, American, and Irish authors between 1800 and 2018, comprising more than 66.9 million words, and used Linguistic Inquiry and Word Count (LIWC 2015) to extract quantitative psycholinguistic data from the novels. I analysed these data using multilevel hierarchical models to account for non-independence of novels by the same author and to control for variation in publication year and authors' age. Significant sex differences were found for 17 out of 24 psycholinguistic categories. The results indicate the existence of psycholinguistic sexual dimorphism in a sample of canonical and prize-winning English-language novels written by heterosexual authors in a way that is consistent with the theoretical predictions which arose from the multidisciplinary research synthesis. The psycholinguistic sex differences in this sample of novels largely align with known psychological sex differences, such as empathising–systemising, and people–things orientation. A few unexpected results include the higher use of verbs and cognitive process words by female authors, which was not predicted by theory. The current findings on 304 novels provide further challenges to the gender similarities hypothesis proposed by some psychologists whilst supporting the sex differences hypothesis arising from and supported by evolutionary science. Furthermore, consistent with predictions from cognitive neuroscience, novels (n = 158) by lesbian authors showed minor signs of psycholinguistic masculinisation, while novels (n = 167) by homosexual men had a female-typical psycholinguistic pattern, supporting the gender shift hypothesis of homosexuality. This study extends prior psychological research to canonical, prizewinning, and contemporary literary art. The added benefit of this multidisciplinary approach is that it connects research from corpus linguistics, computerised text analysis, and biocultural literary theory with psychological science, evolutionary science, and developmental, cognitive, and behavioural neuroscience, thus providing a broader, more convincing, and more evidence-based research synthesis than is possible with any unidisciplinary approach. The findings on this large corpus of 66.9 million words indicate how psychological group differences based on sex and sexual orientation manifest in language use in two centuries of literary art.

Exploring the potential uses of sentiment analysis in historical linguistics

Jukka Tyrkkö

Over the last two decades, various research fields engaged with the computational analyses of text and language have proceeded in separate, yet variously intertwined paths. The methods developed and employed in corpus linguistics, computational linguistics, data science, and digital humanities overlap in various ways, but they also often differ in terms of research objectives, principles of knowledge production, and the core skillsets required of the practitioners. However, as the approaches and tools have matured, there is a growing need for cross-disciplinary exposure of methods and techniques, and of critical analyses of their applicability to various discipline-specific research tasks.

The present study is linked to the Erasmus+ funded project Digital Methods Platform for Arts and Humanities (DiMPAH), which is ongoing at six European universities in collaboration with the iSchools organisation (https://lnu.se/en/dimpah). DiMPAH develops Open Educational Resources in Digital Humanities for the dariahTeach platform (https://teach.dariah.eu), within which framework the present paper belongs to the OER "Linguistics meets data science". In this presentation, I will discuss the usability and potential benefits of modular data science environments in historical linguistics, a research field that has traditionally made relatively little use of such tools. As a case study, I will discuss the freely-available DS platform Knime (https://www.knime.com) and the well-established task of sentiment analysis. I will explain the general design philosophy and functionalities of Knime, present a worked example of a sentiment

analysis workflow using the platform and a small corpus of historical medical texts, and evaluate the quality and usability of the output from the standpoint of historical linguistics. The workflow will be available for download after the presentation. Along the way, I will also comment on the advantages and disadvantages of data science and business intelligence platforms in comparison to scripted solutions.