On the Applicability of Features of Linguistic Complexity to the Classification of Political Speeches

Zarah Weiß
University of Tübingen
zarah-leonie.weiss@uni-tuebingen.de

Abstract

This paper evaluates the applicability of linguistic complexity features to the tasks of party and government affiliation classification in political speeches from German *Bundestag* by comparing them to word embeddings. Using a linear SVM word embeddings outperform classification performance of complexity features, which, however, perform above chance. Different complexity dimensions exhibit intriguing performance differences and give evidence, that especially morphological and lexical complexity structurally varies across parties and government vs. non-government speeches.

1 Introduction

Linguistic complexity is a common field of research in computational linguistics (Kyle, 2016; Sheehan et al., 2013; von der Brück, 2008). It is usually applied in the educational domain: Together with accuracy and fluency, complexity is one of the dominating concepts of first language (L1) and second language (L2) development and commonly used to approximate language proficiency, text readability, and writing quality (Hancke et al., 2012; Hancke, 2013; Kyle, 2016; Biber et al., 2014; Crossley et al., 2015; Sheehan et al., 2013; Lu and Ai, 2015). The benefit of using features of linguistic complexity to classify data is that their results are highly interpretable. This is mandatory for educational purposes: It is not sufficient to rate a learner's proficiency, it is necessary to provide concrete feedback on how to improve. However, using linguistically interpretable features known from complexity analyses has also received interest in related areas of stylometric analyses (Argamon et al., 2009; Dahllöf, 2012; Le et al., 2011), see Grieve (2007) for an overview.

This article investigates a similar direction and analyzes political speeches from German *Bundestag* in terms of complexity. More concretely, 210 complexity features measuring sentential & phrasal, lexical, and morphological complexity, as well as textual cohesion and deagentivation are used to a) identify the party of a speaker, and b) determine whether a speaker is member of the ruling coalition at the time of her or his speech. The underlying questions are:

1. How beneficial are complexity features for party or government affiliation classification?

2. Can complexity features give linguistic insights in the data with respect to party or government affiliation classification?

To help answering these questions, word embeddings are used for comparison as an alternative approach. They are popular for a broad variety of tasks in computational linguistics, such as part of speech (POS) tagging, dependency parsing, but also author profiling tasks (Al-Rfou et al., 2013). While they do capture semantic relatedness by assigning similar positions in the semantic space covered by them to related terms, they are less straight forward to interpret for humans than complexity features. However, they may be used corpus independent and are thus more efficient to use than complexity features.

The remainder of the paper is structured as follows: first, a brief overview over complexity analyses and other classification tasks on political speeches is given. Then, the data set used is introduced. Section 4 discusses the features used in the classification experiments reported in section 5. The article closes with concluding remarks.

2 Related Work

One common strand in the analysis of political speeches is the analysis of political affiliations of parties or individuals, allowing to estimate similarities between them with respect to certain positions. This is highly related to political science and journalism. These analyses are often relying on word patterns (Lowe et al., 2011, 124). Lowe et al. (2011) identify political affiliation on a continuous left to right scale using political categories assigned to sentences based on their content. Sim et al. (2013) are identifying the ideological position of American politicians expressed in their political texts using an ideology lexicon and a Bayesian Hidden Markov Model. Other research focuses more on the aspect of organization affiliation: Dahllöf (2012) performs a binary classification of gender, age, and political affiliation based on Swedish parliament speeches from 2003 to 2010 using binary word vectors indicating absence or presence of words. Koppel et al. (2009) identify both, ideological and organizational affiliation for Arabic texts. Party prediction is often performed based on social media data, in order to facilitate the analysis of voting behavior, see Gottipati et al. (2013). Abu-Jbara et al. (2012) use unsupervised learning methods to identify sub-group affiliation in political fora.

The approach of this paper, however, goes in another direction and stands in the tradition of complexity analyses. Although sharing some features with stylometry (Grieve, 2007; Argamon et al., 2009), linguistic complexity is especially known in the field of L1 and L2 development. Generally, complexity develops along the lines of the quantity, the depth of the internal structure, and the inter-relatedness of its constituents (Housen et al., 2012; Rescher, 1998). In linguistics, this notion of complexity is applied to separate linguistic domains such as syntax, morphology, or the lexicon and operationalized in terms of ratios, frequencies or indexes (Housen et al., 2012). These are often used for predictive purposes, for example to evaluate L1 and L2 proficiency (Hancke, 2013; Kyle, 2016; Biber et al., 2014), text readability (Hancke et al., 2012; Sheehan et al., 2013; von der Brück, 2008), and writing skills (Crossley et al., 2016; Crossley et al., 2015; Lu and Ai, 2015; Yannakoudakis et al., 2011). However, their high interpretability makes them also especially suited for descriptive analyses (Kyle, 2016), for example the investigation of developmental progression (Kyle, 2016; Crossley et al., 2015) or the analysis of differences in German schoolbooks across publishers (Bryant et al., To Appear).

3 Data

For this study, German political speeches from the 13th to 17th legislative period of German *Bundestag* were analyzed. 215,061 speeches held between February 8th, 1996 and September 3rd, 2013 were extracted from 985 protocols obtained from the PolMine-Plenardebattenkorpus (PDK) by Blättle and Berenz (2012), which contains protocols from German parliaments on state and federal level. These speeches are hybrids between spoken and formal written language, as they are usually formulated prior to the session, and held by proficient speakers, yet they are also adjusted spontaneously to reactions from the audience or the development of the current discussion. Audience comments encoded in the speeches were removed for the analysis. Also, moderation sections such as speaker announcements and attendance lists were excluded automatically, by removing speeches held by the respective *Bundestags(-Vize)Präsident(-in)*. The remaining data set contains 120,331 speeches.

The PDK contains meta information on each speaker's party, function, and name, as well as on each protocol's legislative period, protocol number, and protocol date. Participation in the governing coalition was inferred by combining a speaker's party affiliation with the speech's legislative period and added as additional meta information.[1] For this, CDU and CSU were considered to be a single party as well as PDS and Linke, although PDS and Linke only merged in June 2007. Removing 102 speeches without party information, this lead to 62,650 government and 57,579 opposition speeches. The government affiliation subcorpus (GA) was compiled from all opposition speeches and a random sample of the same amount of government speeches. For party classification only speeches from CDU, FDP, Bündnis 90/Grüne, Linke and SPD were considered in the party affiliation sub corpus (PA). Since the amount of speeches contributed by the parties is highly unbalanced, for each party a random sample of

[1] The governing coalition of each legislative period was obtained from `https://de.wikipedia.org/wiki/Liste_der_deutschen_Bundesregierungen`.

11,059 speeches was used for PA.

4 Features

Two different types of high-dimensional feature vectors were designed: on the one hand complexity features and on the other hand word embeddings as a strong base line to test these complexity features against. In order to keep the analyses feasible despite the large data set, each document was presented by a single vector averaged over the entire document, instead of allowing for multidimensional matrices.

4.1 Polyglot Word Embeddings

German word embeddings were retrieved from the Polyglot project (Al-Rfou et al., 2013), in which word embeddings for 107 languages were trained based on their respective Wikipedias. Their 50-dimensional German word embeddings were trained on 687,000,000 tokens and 9,474,000 types. They did not perform normalization with respect to case or inflection to preserve linguistic information.

Each speech was tokenized using the *nltk.tokenize* WORDPUNCTTOKENIZER and then represented with word embeddings. Preliminary experiments were performed to identify the best performing speech representation. The results are shown in table 1. In order to represent each speech in a single vector, word embeddings for each token in a speech were averaged to a single vector. This significantly outperformed concatenating word vectors of the first 20 tokens in a speech.[2] On the averaged word embeddings, further preprocessing was tested: Speeches were stemmed using the *nltk.stem* SNOWBALLSTEMMER. This did not increase performance when either using the original Polyglot embeddings or embeddings, where embeddings with the same stem were averaged. Also, two stop word lists were introduced. Since stop words are highly data dependent (Forman, 2003), two stop word lists were compiled automatically for each task sub corpus, once by considering every token a stop word that occurred at least once in 50 percent of all speeches and once for 67 percent of all speeches. The application of the lists either harmed performance or did not affect it. Thus, simple, averaged word embeddings were chosen for the embedding model.

[2]Concatenating more tokens was not feasible given corpus size and word embedding dimensions. An embedding length was deemed not feasible if a single fold took more than five minutes to train and test.

4.2 Complexity Features

For the complexity vectors, 210 measures of linguistic complexity were used. They cover the sentential & clausal, phrasal, lexical, and morphological domain as well as the domains of textual cohesion and deagentivation. These features were designed by Hancke (2013), Galasso (2014), and Weiß (2015), but re-organized into more coherent feature sets following Kyle (2016), McNamara et al. (2014).[3]

4.2.1 Single Feature Sets

Sentential & Clausal Complexity (SEN) contains 21 features, most of which origin from the syntactic feature set by Hancke (2013). These are several clause, (complex) t-unit, non-terminal, and sentence ratios measuring either construction depth or length. Furthermore, it includes all three topological field related features by Weiß (2015), which measure syllable distances between arguments and verbs, finite verbs and their verbal complex, and the ratio of non-subject prefields. For formulae and detailed feature descriptions, please see Hancke (2013, 17–41), Weiß (2015, 30–32).

Phrasal Complexity (PHR) contains 57 features, most of which origin from the syntactic feature set by Hancke (2013). They measure ratios of phrase coordination, modifier quantities, *to*-marked infinitives, phrase lengths, dependents, and NP, VP, and PP frequencies. Furthermore, all complex NP and VP features by Weiß (2015) were included, which measure POS and position dependent modifier ratios, verb cluster lengths and types, and periphrastic tense patterns. For formulae and detailed feature descriptions, please see Hancke (2013, 17–41) and Weiß (2015, 20–30).

Lexical Complexity (LEX) contains 46 features, including lexical diversity measures such as five variants of type-token ratio, and 12 types of lexical variation features, lexical frequency features, features of lexical relatedness, such as hypernymy and synonymy, and shallow syllable based word length measures. For formulae and

[3]Feature calculation requires extensive Natural Language Processing (NLP) and is thus computationally costly. Feature calculation took in total eight days when calculating each legislative period separate on a server. The resulting feature table is available `https://github.com/zweiss/ML4CL_2016/tree/master/ml4cl_project/`.

Model	Government Affiliation			Party		
	F1	Pre	Rec	F1	Pre	Rec
Avgeraged word embeddings	0.6705	0.6864	0.6556	0.2933	0.3059	0.2997
Concatenated word embeddings	0.6317	0.6553	0.6098	0.2788	0.2811	0.2794
Stemmed text on avg. word embeddings	0.6699	0.6863	0.6543	0.2796	0.2949	0.2867
Stemmed text on avg. stem embeddings	0.6677	0.6845	0.6520	0.2919	0.3023	0.2965
Stop word t=0.50 avg. word embeddings	0.6408	0.6578	0.6251	0.2798	0.2895	0.2840
Stop word t=0.67 avg. word embeddings	0.6590	0.6701	0.6487	0.2823	0.2920	0.2863

Table 1: Performance of word embedding models on party and government affiliation classification.

detailed feature descriptions, please see Hancke (2013, 17–35).

Morphological Complexity (MOR) contains 42 features: verbal inflection (person, tense, mood), verb type (full, auxiliary, modal), derivational morphology determining latinate terms and compound ratios. For formulae and detailed feature descriptions, please see Hancke (2013, 42–45).

Textual Cohesion (COH) includes 36 features, most of which origin from Galasso (2014). They capture pronoun ratios, noun, argument, stem and content word overlap, connective counts, and features based on syntactic functions, such as subject, object, etc., as well as three features used to describe the basic properties of each text, namely total number of sentences, paragraphs and words. Furthermore, they contain all five conditional clause measures from Weiß (2015). For further details, please see Galasso (2014, 7–16) and Weiß (2015, 34–37).

Textual Deagentivation (DEA) contains 8 features implemented by Weiß (2015). They measure ratios of *man* (one), *lassen* (to let) + reflexive pronoun, (quasi) passives, participle modifiers, half modals, and infinitival constructions as well as the coverage of deagentivational patterns. For more details, please see Weiß (2015, 32-33).

4.2.2 Feature Set Combinations

Aside from the single feature sets, the following feature set combinations were used in separate models: all feature sets combined (ALL), all feature sets except for the worst performing feature set for the respective task (ALL-???), and an automatically selected feature set (AUTO), which was derived automatically from ALL by using importance weights assigned by *sklearn.feature_selection*'s SELECTFROMMODEL. As estimator a linear SVM with an error term penalty of 0.01 on L1 penalty norm assuming a primal optimization problem lead to the best results, reducing feature dimensionality from 210 to 66 for party classification and to 68 for government affiliation identification.

5 Experiment

Two classification experiments were performed: On the one hand, government affiliation classification (GAC) as a binary task: government vs. opposition, on the other hand, party affiliation to either CDU, FDP, Bündnis 90/Grüne, Linke and SPD. It should be noted, that party affiliation classification (PAC) and GAC are necessarily dependent on each other, and that it is beyond the limit of the experiments to account for that.

5.1 Set-Up

For both tasks a Support Vector Machine (SVM) with a linear kernel was chosen as classification algorithm as recommended by Dahllöf (2012), Forman (2003). The experiments rely on the LINEARSVC implementation from the SKLEARN package. Classification was treated as a primal optimization problem for all experiments, because the LINEARSVC documentation suggests this, if sample size exceeds feature size.[4] With this comes the requirement of using the squared hinge loss function to penalize the model instead of the simple hinge loss function. Aside from

[4] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC.

these adjustments, the default setup was used, because parameter changes either decreased performance (especially for increases in the stopping criterion tolerance) or did not affect it significantly. Due to the known sensitivity of SVMs to feature scaling,[5] embeddings as well as all complexity features were scaled to the range $[-1; 1]$ using the MAXABSSCALER from the *sklearn.preprocessing* module. This scaling range was preferred over a $[0; 1]$ scaling to preserve information on zero values, which may possibly occur in complexity feature vectors.

Training and testing was performed using stratified 10-folds cross validation. For GAC averaged binary F1-score, precision, and recall were chosen and averaged weighted F1-score, precision, and recall for PAC. To establish whether the performance differences are beyond chance, two-sampled significance testing for related samples was performed using the *scipy.stats* implementation and assuming a conservative significance level $\alpha \leq 0.01$. As a lower base line on chance level the DUMMYCLASSIFIER from the *sklearn.dummy* module was used with a stratified classification strategy. On the balanced sub corpora, this performs equivalent to a uniformly random classification strategy.

5.2 Results & Discussion

Table 2 shows f1 score, precision, and recall for GAC and PAC for each model as well as model complexity in terms of feature vector length. Most models outperform the random baseline, except for sentential complexity and textual coherence for GAC, which do not show a significant difference, and textual deagentivation for both tasks, which performs even significantly worse than the random baseline in terms of f1 score and recall for GAC. Also, all models in both classification tasks perform significantly worse than word embeddings: Even the best performing model, which is for both tasks the one containing all feature sets except for textual deagentivation (ALL-DEA), performs for each measure more than 4% worse than word embeddings for GAC. For PAC, the difference between word embeddings and ALL-DEA is with 1% less drastic, yet still significant. Second best feature combination for both tasks is using ALL. However, the difference between ALL and ALL-DEA is not significant for either task. Also AUTO did not improve performance for either task. Concerning the first question of this study, namely how beneficial complexity features are for PAC and GAC, these results lead to the following conclusion: While complexity features allow to distinguish speeches held by members of the respective current ruling coalition and speeches held by members of different parties, easier obtainable methods, such as word embeddings, are clearly better suited for this task.

To answer the second question, namely if complexity features give linguistic insights in the data that word embeddings cannot give, a closer look at the single feature sets is required. In fact, feature sets widely differ in terms of their performance: The best performing feature set for GAC is morphological complexity (MOR), which significantly outperforms lexical complexity (LEX), the second best feature set. These two sets are among the best performing models for PAC, too, together with phrasal complexity (PHR). Neither significantly differs from the others. The worst performing feature sets are textual deagentivation (DEA) and sentential complexity (SEN), again for both tasks. As for general differences between GAC and PAC: Aside from the across all models higher performance for GAC, which is to be expected given the lower complexity of binary tasks, two differences seem remarkable: First, while precision and recall are very similar for all models for PAC, the same holds only for the random baseline and LEX for GAC. For all othere models, including the one using word embeddings, precision is higher than recall i.e. the models overestimate affiliation with the government. Second, while for GAC there are clear performance differences between the single feature sets, the same does not hold for PAC, where several feature sets perform alike. For the second question these results lead to the following conclusion: Some complexity features detect differences between speeches held by members of a government and an opposition. This holds especially for lexical and morphological features. These features have proven to be highly beneficial in previous studies on complexity, too (Vajjala and Meurers, 2012; Hancke et al., 2012). However, some features, like textual coherence and deagentivation or sentential complexity seem to be less relevant. For DEA this low performance is probably influenced by the small set size. Yet, for SEN

[5] http://scikit-learn.org/stable/modules/svm.html.

Model	Dim	Government Affiliation			Party Affiliation		
		F1	Pre	Rec	F1	Pre	Rec
Base Line	0	0.4986	0.4985	0.4987	0.2004	0.2005	0.2004
Embeddings	64	0.6705^{*}	0.6864^{*}	0.6556^{*}	0.2933^{*}	0.2997^{*}	0.3059^{*}
ALL	210	$0.6154^{*,-}$	$0.6327^{*,-}$	$0.5995^{*,-}$	$0.2773^{*,-}$	$0.2834^{*,-}$	$0.2896^{*,-}$
ALL-DEA	202	$0.6231^{*,-}$	$0.6428^{*,-}$	$0.6051^{*,-}$	$0.2801^{*,-}$	$0.2858^{*,-}$	$0.2920^{*,-}$
AUTO	68; 66	$0.6140^{*,-}$	$0.6298^{*,-}$	$0.5995^{*,-}$	$0.2655^{*,-}$	$0.2745^{*,-}$	$0.2805^{*,-}$
SEN	21	0.5041 $^{-}$	$0.5484^{*,-}$	0.4682 $^{-}$	$0.2257^{*,-}$	$0.2363^{*,-}$	$0.2394^{*,-}$
PHR	57	$0.5361^{*,-}$	$0.5656^{*,-}$	0.5104 $^{-}$	$0.2463^{*,-}$	$0.2544^{*,-}$	$0.2611^{*,-}$
LEX	46	$0.5667^{*,-}$	$0.5689^{*,-}$	$0.5656^{*,-}$	$0.2430^{*,-}$	$0.2509^{*,-}$	$0.2586^{*,-}$
MOR	42	$0.5959^{*,-}$	$0.6054^{*,-}$	$0.5869^{*,-}$	$0.2451^{*,-}$	$0.2536^{*,-}$	$0.2605^{*,-}$
COH	36	0.5257 $^{-}$	$0.6013^{*,-}$	0.4677 $^{-}$	$0.2327^{*,-}$	$0.2454^{*,-}$	$0.2506^{*,-}$
DEA	8	$0.4604^{v,-}$	$0.5021^{*,-}$	$0.4259^{v,-}$	0.2086 $^{-}$	$0.2223^{*,-}$	$0.2236^{*,-}$

Table 2: F1 score, precision, and recall of government and party affiliation classification given varying models. Significant difference for $\alpha \leq 0.01$ compared to base line indicated by * (better) and *v* (worse), to word embeddings by + (better) and – (worse).

and COH the results indicate that neither sentential complexity nor the use of coherence markers are indicative for government affiliation. As for PAC, the differences between feature sets are less telling, except for the very small DEA feature set. However, since the features perform significantly better than the random baseline and feature set combination leads to a significant performance increase, this does not seem to indicate that there are no differences in the style of speeches held by members of different parties. Instead, this seems to indicate that there are relevant differences in each of these feature sets.

Finally, it should be noted that there is the strong possibility of government and party affiliation influencing each other, since the ruling coalition does not consist of the same parties each legislative period. Further in depth investigations would be necessary, to identify whether and to which extend these two factors influence each other.

6 Conclusion

This paper investigated the applicability of complexity features to a domain different from language acquisition. Applicability was operationalized in terms of a) general benefit for a text classification task, and b) achieved additional linguistic insights in the data. As text classification tasks, a speaker's party and government affiliation were chosen. The results showed that while complexity features could not compete with word embeddings, they did find enough structural differences to outperform the uninformed baseline. Furthermore, the results showed that government affiliation seems to influence lexical and morphological complexity, while sentential complexity and textual cohesion are not, and party affiliation seems to be influenced by a series of complexity dimensions. To investigate these differences further, a more detailed analysis of the single features within each feature set to identify the most informative features would be necessary, to obtain more detailed information on how government and party affiliation influence speech style. Also, a careful analysis on how party and government affiliation influence each other as dependent factors would be necessary in future work. The here presented study successfully showed that complexity analyses are of relevance outside the domain of language development. It also lays the foundations for such future work by making a table of the PDK with 210 complexity features publicly available.

References

Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi, and Dragomir Radev. 2012. Subgroup detection in ideological discussions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 399–409. Association for Computational Linguistics.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Douglas Biber, Bethany Gray, and Shelley Staples. 2014. Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, pages 1–31.

Andreas Blättle and Silvia Berenz. 2012. Polmine – korpusunterstützte politikforschung – dokumentation.

Doreen Bryant, Karin Berendes, Detmar Meurers, and Zarah Weiß. To Appear. Schulbuchtexte der sekundarstufe auf dem linguistischen prüfstand. analyse der bildungs- und fachsprachlichen komplexität in abhängigkeit von schultyp und jahrgangsstufe. In Mathilde Hennig, editor, *Linguistische Komplexität – ein Phantom?* Stauffenburg Verlag, Tübingen, Germany.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2015. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. In *Behavior research methods*, pages 1–11. Springer.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The development and use of cohesive devices in l2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32:1–16.

Mats Dahllöf. 2012. Automatic prediction of gender, political affiliation, and age in swedish politicians from the wording of their speeches – a comparative study of classifiability. *Literary and Linguistic Computing*, 27(2):139–153.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Sabrina Galasso. 2014. Exploring textual cohesion characteristics for german readability classification. B.A. Thesis, August.

Swapna Gottipati, Minghui Qiu, Liu Yang, Feida Zhu, and Jing Jiang. 2013. Predicting user's political party using ideological stances. In *International Conference on Social Informatics*, pages 177–191. Springer.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic and morphological features. In *Proceedings of COLING*, pages 1063–1080, Mumbai, December.

Julia Hancke. 2013. Automatic prediction of cerf proficiency levels based on linguistic features of learner language. Master's thesis, Eberhard Karls Universität Tübingen, April.

Alex Housen, Ineke Vedder, and Folkert Kuiken, 2012. *Document Viewing Options: Title: Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA*, volume 32 of *Language Learning & Language Teaching*, chapter 1–2. John Benjamins Publishing, Amsterdam, Philadelphia.

Moshe Koppel, Navot Akiva, Eli Alshech, and Kfir Bar. 2009. Automatically classifying documents by ideological and organizational affiliation. In *ISI*, pages 176–178.

Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.D. thesis, Georgia State University.

Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Literary and Linguistic Computing*, 26(4):435–461.

Will Lowe, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. Scaling policy preferences from coded political texts. *Legislative studies quarterly*, 36(1):123–155.

Xiaofei Lu and Haiyang Ai. 2015. Syntactic complexity in college-level english writing: Differences among writers with diverse l1 backgrounds. *Journal of Second Language Writing*, 29:16–27.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Camebridge University Press.

Nicholas Rescher. 1998. *Complexity: A philosophical overview*. Transaction Publishers.

Kathleen M Sheehan, Michael Flor, and Diane Napolitano. 2013. A two-stage approach for generating

unbiased estimates of text complexity. In *Proceedings of the 2th Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 49–58, Atlanta, Georgia. Association for Computational Linguistics.

Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 91–101.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*, volume 7, pages 163–173, Montréal, Canada, June. Association for Computational Linguistics.

Tim von der Brück. 2008. A readability checker with supervised learning using deep indicators. *Informatica*, 32:429–435.

Zarah Leonie Weiß. 2015. More linguistically motivated features of language complexity in readability classification of german textbooks: Implementation and evaluation. B.A. Thesis, September.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, volume 49, pages 180–189, Portland, Oregon. Association for Computational Linguistics.