A Component based Approach to Measuring Text Complexity

Simon Jönsson, Evelina Rennes, Johan Falkenjack, Arne Jönsson

Linköping University, Linköping, Sweden
simjo241@student.liu.se, evelina.rennes|johan.falkenjack|arne.jonsson@liu.se

Abstract

We present results from assessing text complexity based on a factorisation of text property measures into components. The components are evaluated by investigating their ability to classify texts belonging to different genres. Our results show that the text complexity components correctly classify texts belonging to specific genres, given that the genres adhere to certain textual conventions. We also propose a radar chart visualisation to communicate component based text complexity.

1. Introduction

Recent years' development of speed and accuracy of text analysis tools has made new text features available for readability assessment. For instance, phrase structure parsing has been used to find the average number of sub-clauses, verb phrases, noun phrases and average tree depth (Schwarm and Ostendorf, 2005). For Swedish, Heimann Mühlenbock (2013), Falkenjack and Jönsson (2014), and Falkenjack et al. (2013) have addressed such data driven text complexity assessment.

2. Text complexity measures

For the study presented in this paper we use the publicly available toolkit TECST (Falkenjack et al., 2017) and the text complexity analysis module SCREAM (Heimann Mühlenbock, 2013; Falkenjack et al., 2013). As of today, SCREAM calculates 119 features of text complexity that roughly can be divided into the following categories:

Shallow features are features that can be extracted after tokenisation by simply counting words and characters. Shallow features include mean word length and mean sentence length.

Lexical features are based on categorical word frequencies extracted after lemmatisation and calculated using the basic Swedish vocabulary SweVoc (Heimann Mühlenbock, 2013). They are further divided into groups such as everyday use and communication.

Morpho-syntactic features concern a morphology based analysis of the text. The analysis relies on previously part-of-speech annotated text. Measures include a number of part-of-speech tags and ratio of content words.

Syntactic features are features that can be estimated after syntactic parsing of the text. Features include a number of dependency distance measures.

Text quality metrics include measures that traditionally are used to measure readability.

Several studies have explored how text complexity measures can be combined and clustered in different ways to be more comprehensive and easier to understand, c.f. (Falkenjack et al., 2016). One way of conducting clustering is through factor analysis, allowing large amounts of variables to be combined into fewer clusters or factors. Biber (1988) conducted such analyses in order to find the factors that distinguish spoken language from written language. Through a principal factor analysis, 67 features were reduced to 7 factors.

Our study is inspired by Biber's three step analysis. The first step is to decide on a method for analysis. The method used by Biber (1988) is Principal Factor Analysis (PFA, also known as common factor analysis). Another method of factor analysis is Principal Component Analysis (PCA). A fundamental difference between PFA and PCA is that PFA does not account for all the variance, only the variance that is shared between variables (Biber, 1988). Henry (1979) and Lee et al. (2012) are two examples of studies that used PCA in terms of combining linguistic features into fewer components.

The second step is to decide on how many factors to extract. This can be done by analysing a screen plot and determine where additional factors do not contribute to the overall analysis (Biber, 1988). It is also possible to analyse a table to see how much variance each factor explains and how much the factors explain together. A third way of determining the number of factors to be extracted is through parallel analysis (O'Connor, 2000). The analysis is a way to test how many eigenvalues that are statistically significant.

The third step is to choose what type of rotation that should be used. Biber (1988) chooses an oblique structure, Promax, which allows for more correlations, even minor ones, among the factors.

3. Corpus

The text material used in our studies comprises texts from the SUC corpus (Ejerhed et al., 2006). In the experiments we want to investigate the ability to distinguish different text domains, or genres, using text complexity measures factorised into components as suggested by Biber (1988). There is a theoretical distinction between the concepts of genre and domain. Here domain refers to the shared general topic of a group of texts. For instance, "Fashion", "Leisure", "Business", "Sport", "Medicine" or "Education" are examples of broad domains. Genre is a more abstract concept. It characterises text varieties on the basis of conventionalised textual patterns. For instance, an *academic*

paper obeys to textual conventions that differ from the textual conventions of a *tweet*; a *letter* complies to conventions that are different from the conventions of an *interview*. *Academic papers*, *tweets*, *letters*, *interviews* are examples of genres. For more details see (Falkenjack et al., 2016). If we apply this distinction to the nine top genres included in the SUC, we end up with six "proper" genres, see Table 1.

Table 1: The six proper SUC genres used in our study

Genre	Size
Press Reportage (**A**)	269
Press Editorial (**B**)	70
Press Review (**C**)	127
Biographies/Essays (**G**)	27
Learning/Scientific Writing (**J**)	86
Imaginative Prose (**K**)	130

4. Procedure

Similar to Biber (1988), a factor analysis was conducted in order to group linguistic features. The method used here was a Principal component analysis (PCA). Features which either did not have any values or were already represented by other features by having one-to-one correlations were excluded from the feature set.

Through a parallel analysis, the number of clusters to extract from the PCA was elicited (O'Connor, 2000). The method compares raw data, principal component eigenvalues that correspond to the actual data, with random data eigenvalues. If the first value, raw data, is larger than the 95th percentile, it is considered a significant eigenvalue and is included. The extracted number of significant eigenvalues is the number of components extracted through the PCA.

With a Promax, oblique structure, the PCA was done on the set of data containing the remaining linguistic features, each with a total of 1040 data points, where each data point represents results from analyses as described above.

Using the obtained components we investigate their ability to classify the SUC genres. We are using a 18×15 *softmax* neural network with linear activation function. Since SUC has the issue of uneven amount of genre representatives we sample the data as a tensor, Batches $\times$ Samples $\times$ Components (where a batch is a 10×6 matrix of sampled measures of SUC texts), to attempt solving this issue. Genre **G** has fewer data points than the rest of the genres, giving a limited training sample. Classifying a genre that is underrepresented gives a vague model and therefore genre **G** is excluded.

Two of the components obtained can be seen in Table 2. Components were obtained by quantitatively analysing correlation between features and removing features such that we obtain maximal classification. The correlation cut-off was $|0.8|$ where we found local optimum of classification rate 84.0%.

5. Results

From the parallel analysis, a total of 28 eigenvalues were elicited that were used as number of components to be extracted. A total of 93 features remained in the data set after removing 19 features, features with either a prediction of 0, no result at all, already subsumed by other features with a one to one correlation, or not having a predictability higher than 0.65 (.503 - .646).

Table 2: Example of extracted components

Comp.	Feature	Weight	Explanation
1	pos_PN	.816	Pronouns
	pos_NN	-.808	Nouns
	nrValue	-.807	Nominal ratio
	avgNoSyllables	-.730	Average number of syllables
	dep_PA	-.729	Complement of preposition
	dep_ET	-.714	Other nominal post-modifier
	dep_MS	.612	Macrosyntagm
	ratioSweVocC	.607	SweVoc lemmas fundamental for communication
	dep_IO	.573	Indirect object
	pos_AB	.572	Adverb
	dep_SS	.525	Other subject
	dep_DT	-.524	Determiner
	avgPrepComp	-.522	Average number of prepositional complements per sentence in the document
	pos_PS	.487	Possessive pronoun
	dep_NA	.473	Negation adverbial
	dep_MA	.446	Attitude adverbial
	dep_I	.425	Question mark
	pos_RG	-.407	Cardinal number
	dep_AA	.400	Other adverbial
	dep_F	.388	Coordination at main clause level
	dep_PL	.382	Verb particle
	dep_OO	.365	Direct object
	pos_HA	.322	WH-adverb
	dep_AT	-.302	Nominal (adjectival) pre-modifier
	ratioSweVocTotal	.301	Unique, per lemma, SweVoc words in the sentence.
2	pos_PM	-.858	Proper noun
	dep_HD	-.788	Head
	lexicalDensity	.710	Lexical density
	ratioSweVocTotal	.706	Unique, per lemma, SweVoc words in the sentence.
	ratioSweVocH	.573	SweVoc other highly frequent lemmas (category H)
	ratioSweVocC	.544	SweVoc lemmas fundamental for communication
	dep_SS	.429	Other subject
	dep_AN	-.393	Apposition
	ratioSweVocD	.356	SweVoc lemmas for everyday use (category D)
	ratioVerbalRoots	.347	The ratio of sentences with a verbal root
	pos_NN	.332	Noun

An analysis using the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (.595) and Bartlett's Test of Sphericity shows the validity of PCA to interpret the data set ($p < .05$)

The variables chosen for each component had a magnitude over 0.3 and under -0.3. The total variance explained by the 28 components is 60.5%, of which the first component explains 8% on its own.

The results from classification using the neural network is presented in Table 3.

Table 3: F1-Scores for the components

Genre	F1
Press Reportage (**A**)	0.814
Press Editorial (**B**)	0.793
Press Review (**C**)	0.831
Learning/Scientific Writing (**J**)	0.826
Imaginative Prose (**K**)	0.9324

We note that the F1-scores of respective genre are fairly

consistent, except genre **B**, which has a significantly lower score and genre **K** which has a significantly higher score. The former might be due to the properties of genres who has a Press origin being similar in some textual sense. Whereas Imaginative Prose, **K**, might differ from the rest of the genres in a text complexity sense, which makes it easier for the classifier to distinguish the genre. Analogously the classifier might have difficulties distinguishing Press related data points, to some extent.

Table 4: Confusion matrix for the components. Each genre has been classified 150 times.

	A	B	C	J	K
A	120	6	9	8	7
B	11	111	8	15	5
C	8	4	125	9	4
J	4	8	7	128	3
K	2	1	2	0	145

To further analyse the classification results, Table 4 presents the resulting confusion matrix. From Table 4 we note that genres **A, B, C, J** have many False Positives (FP) and many False Negatives (FN), whereas genre **K** only have strong FN, which means that the other genres are misclassified as genre **K** but **K** seldomly is misclassified as any other genre, this implies that **K** is more separated from other genres in our feature space. Also one can deduce that the other genres have more interlacement in our feature space.

6. Visualising text complexity

Each of the components derived from the factor analysis comprises several individual text complexity features that depict different aspects about the analysed texts, as can be seen in Table 2. The components can not easily be labelled in a meaningful way. Instead we propose to visualise them in a radar diagram, c.f. Branco et al. (2014), Figure 1.

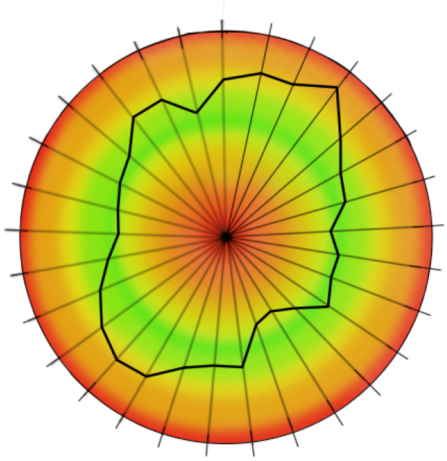

Figure 1: Visualisation of components.

The pattern in the radar chart, resulting from a text analysis, communicates something about a text's complexity, the inner line in the radar chart in Figure 1. Different texts provide different patterns and it may be possible to use such patterns to characterise a text's complexity and also compare its complexity with other texts' complexity.

The components are, thus, visualised in an intuitive way, where the pattern communicates text complexity. However, the components should also have justified names and definitions. A remaining issue is the domain-specific terminology concerning text complexity, as the meaning of the components has to be communicated along with the assessment. A huge endeavour as the components comprise features that reflect different, and sometimes opposing, qualities of a text.

The components should also be sorted in a way in which related components are closer to one another. Making use of the interactivity of a digital tool, the visualisation could be revised even further. By combining the extracted components into overall categories that may be presented at first, revealing the most important components from each category by selecting the corresponding section in a radar chart, the radar chart may become more comprehensible. This final visualisation with the components therein needs to be evaluated to properly see if it is more intuitive and if the components give users an understanding of a text's complexity.

7. Conclusions

We have shown that a component based text complexity analysis can be used to classify texts in genres. Assuming that genres have different text properties the components, thus, also say something about texts' complexity. The results are based on measuring complexity of Swedish, but very few measures are specific for Swedish. Further research should study how to define genres such that the text complexity feature space is more separated, thus leading to "stronger" genres, i.e. more distinguished genres - in a textual complexity sense.

We have also suggested that component based text complexity measures can be visualised in a radar diagram. Further research on visualisation includes conducting studies on users' understanding of text complexity using radar charts and on finding meaningful ways to reorganise the components.

Acknowledgements

This research is financed by Vinnova and RISE SICS East. We are indebted to Jakob Säll and Simon Cavedoni for initial research on PCA for visualisation.

References

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge: Cambridge Univ. Press, 1988.

António Branco, João Rodrigues, Francisco Costa, Joao Silva, and Rui Vaz. 2014. Rolling out text categorization for language learning assessment supported by language technology. In *International Conference on Computational Processing of the Portuguese Language*, pages 256–261. Springer.

Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 2006. Stockholm Umeå Corpus version 2.0.

Johan Falkenjack and Arne Jönsson. 2014. Classifying easy-to-read texts without parsing. In *The 3rd Workshop*

on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014), Göteborg, Sweden.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa-2013), Oslo, Norway*, number 085 in NEALT Proceedings Series 16, pages 27–40. Linköping University Electronic Press.

Johan Falkenjack, Marina Santini, and Arne Jönsson. 2016. An Exploratory Study on Genre Classification using Readability Features. In *Proceedings of the The Sixth Swedish Language Technology Conference (SLTC 2016), Umeå, Sweden*.

Johan Falkenjack, Evelina Rennes, Daniel Fahlborg, Vida Johansson, and Arne Jönsson. 2017. Services for text simplification and analysis. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, Gothemburg, Sweden*.

Katarina Heimann Mühlenbock. 2013. *I see what you mean. Assessing readability for specific target groups*. Dissertation, Språkbanken, Dept of Swedish, University of Gothenburg.

G. Henry. 1979. The relation between linguistic factors identified by a principal components analysis of written style and reading comprehension as measured by cloze tests. *Journal of Research in Reading*, 2(2):120–128.

Yi-Shian Lee, Hou-Chiang Tseng, Ju-Ling Chen, Chun-Yi Peng, Tao-Hsing Chang, and Yao-Ting Sung. 2012. Constructing a novel chinese readability classification model using principal component analysis and genetic programming. In *Proceedings of the 12th IEEE International Conference on Advanced Learning Technologies, ICALT 2012*, pages 164–166, 07.

Brian P O'Connor. 2000. SPSS and BAS programs for determining the number of components using parallel analysis and velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32(3):396–402.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.