

Pinning down text complexity

An Exploratory Study on the Registers of the Stockholm-Umeå Corpus (SUC)

Marina Santini & Arne Jönsson

RISE Research Institutes of Sweden | Linköping University

In this article, we present the results of a corpus-based study where we explore whether it is possible to automatically single out different facets of text complexity in a general-purpose corpus. To this end, we use factor analysis as applied in Biber's multi-dimensional analysis framework. We evaluate the results of the factor solution by correlating factor scores and readability scores to ascertain whether the selected factor solution matches the independent measurement of readability, which is a notion tightly linked to text complexity. The corpus used in the study is the Swedish national corpus, called *Stockholm-Umeå Corpus* or SUC. The SUC contains subject-based text varieties (e.g., hobby), press genres (e.g., editorials), and mixed categories (e.g., miscellaneous). We refer to them collectively as 'registers'. Results show that it is indeed possible to elicit and interpret facets of text complexity using factor analysis despite some caveats. We propose a tentative text complexity profiling of the SUC registers.

Keywords: text complexity, readability, corpus-based analysis, multivariate statistics, factor analysis, Multi-Dimensional Analysis (MDA), correlation tests

1. Introduction

Text complexity is an important dimension of textual variation. It is crucial to pin it down because texts can be customized to different types of audiences, according to cognitive requirements (e.g., texts for the dyslexic), social or cultural background (e.g., texts for language learners) or the text complexity that is expected in certain genres or registers (e.g., academic articles vs. popularized texts).

In this study, we explore text complexity variation in the Swedish national corpus, called *Stockholm-Umeå Corpus* or SUC (Källgren et al. 2006). The SUC

is a collection of Swedish texts and represents the Swedish language as used by native Swedish adult speakers. The SUC includes a wide variety of texts written for several types of audiences, from academics, to newspaper readers, to fiction readers.

Given the composition of the SUC, we assume that there are different levels of text complexity across SUC's text categories. The assumption underlies the rationale of the study, which is to identify how linguistic features co-occur in texts that have different levels of text complexity. Arguably, text complexity in children's books is low, while specialized professionals, such as lawyers and physicians, must be able to understand very complex texts in order to practice their professions. In between easy texts for children and the domain-specific jargon used by specialized professionals, there exist texts that present different levels of textual difficulty.

The experiments presented here are part of a larger research endeavour aimed at finding computational methods that help identify easy-to-read texts characterized by low text complexity from texts that are more difficult to read and show higher text complexity. The relation between readability and text complexity is discussed in Section 2.

The automatic identification of easy-to-read texts and of fine-grained text complexity facets plays an important role in many Language Technology (LT) applications that span from education (e.g., automatic essay correction), to consumer-oriented text simplification (e.g., the layfication of medical jargon), to linguistic tools to cope with cognitive impairments (e.g., text simplification for the dyslexic), to the facilitation of text understanding for vulnerable members in a society, such as non-native speakers, the elderly, and children.

The statistical exploration of the SUC is based on factor analysis. More specifically, we investigate whether it is possible to detect text complexity facets using factor analysis as applied in Biber's (1988) multi-dimensional analysis (henceforth MDA). The SUC contains subject-based text varieties (e.g., hobby), press genres (e.g., editorials), and mixed categories (e.g., miscellaneous). We refer to them collectively as 'registers', as defined in Biber & Conrad (2009). Multi-dimensional analysis cuts across all text varieties irrespective of their different nature (cf. Biber and Kurjian 2007; Biber & Egbert 2016), while some other techniques (like supervised machine learning) might benefit from a more stringent distinction between subject-based vs genre-based text varieties.

Since the validation of empirical results is integral part of any research, we evaluate the results of this exploratory study by correlating the factors returned by factor analysis with a readability index, called LIX (Björnsson 1968). This will help ascertain whether the selected factor solution matches an independent measurement of readability, which is a notion complementary to or overlapping text

complexity. We use the LIX readability index for the Swedish language,¹ because this index is widely used in Sweden. The motivation underlying this type of validation can be summarised as follows: if the factor scores and the LIX scores correlate significantly (i.e., $p < 0.05$), we get a statistically significant indication that the text complexity facets elicited with MDA are statistically reliable. LIX is a coarse but robust gauge of readability. Multi-dimensional analysis, on the other hand, is based on a sophisticated statistical method (factor analysis) whose results are interpreted linguistically. If these two very different approaches show a statistically significant correlation, then we have a robust indication that the computational essence of SUC's text complexity has been pinned down successfully. Conversely, if factor scores and LIX scores do not significantly correlate, this would indicate that SUC's text complexity remain undisclosed, and consequently more investigation is required. In either case, the results of the experiments are informative and pave the way to future research directions.

We introduce several adaptations to the traditional MDA approach (Biber 1988) that is by now consolidated by many studies in corpus linguistics (Sardinha & Pinto 2014). First, we propose using parallel analysis (Horn 1965) to select statistically significant factors. Second, we rely on an extrinsic validation of the factors via an external measurement (readability scores). Third, we split the factors into signed (\pm) textual dimensions and interpret the signed dimensions as text complexity facets. Fourth, we profile SUC registers using the text complexity facets in combination with LIX scores. Essentially, we dig into three ores: first we analyse SUC registers per factor, then per signed dimension, and finally per register. Each of these statistical descriptions provide a different perspective into text complexity. The ultimate goal of this study is to show the potential of the approach rather than presenting full-fledged results.

The article is organized as follows: Section 2 discusses the notions of readability and text complexity; Section 3 summarizes previous work; Section 4 describes the SUC corpus and dataset and presents MDA, together with the factor solution used in this study; in Section 5 we correlate the factor scores and LIX scores and interpret the signed dimensions; in Section 6 we propose a text complexity profiling of SUC registers; Section 7 presents a discussion of the findings; finally, in Section 8, we draw conclusions and point to future work.

1. Web interface: <<http://lix.se/>>.

2. Text complexity and readability

Broadly speaking, text complexity refers to the level of cognitive engagement a text provides to human understanding (Vega et al. 2013). If a text is difficult, it requires more cognitive effort than an easy-to-read text and vice versa.

Text complexity is a multifarious notion, since the complexity can affect the lexicon of a text, its syntax, how the narration of the text is organized, etc. For this reason, several definitions and several standards of text complexity exist.

For instance, in theoretical linguistics, Dahl (2004) puts forward an interpretation of “complexity” that is not synonymic with “difficulty” (p.2). In his view, complexity is “an objective property of a system”, i.e., a measure of the amount of information needed to describe or reconstruct it (Chapter 2). His notion of grammatical complexity is the result of historical processes often subsumed under the rubric of grammaticalization and involves what can be called mature linguistic phenomena, that is “features that take time to develop” (Chapter 3).

Another linguistic field where there is persistent interest in the study of language complexity is second language (L2) research. For instance, Pallotta (2015) notes that the notion of linguistic complexity is still poorly defined and often used with different meanings. He proposes a simple and coherent view of complexity, which is defined in a purely structural way and arises from the number of linguistic elements and their interrelationships. More recently, Housen et al. (2019) present an overview of current theoretical and methodological practices in L2 complexity research that includes five empirical studies focussing on under-explored forms of complexity spanning from the cross-linguistic perspective to novel forms of L2 complexity measurements.

In education, one of the more comprehensive text complexity models that has been devised for teaching is the CCSS – Common Core State Standards (Hiebert 2012). This model, mostly applied in the United States, is a three-part model geared towards the evaluation of text complexity gradients from three points of view: qualitative, quantitative, and by assessing the interaction between the reader and the task (see Figure 1). CCSS is one of many other models of text complexity that have been proposed for educational purposes. It is to be noted that none of them has gained universal status.

In recent years, the concept of text complexity has drawn the attention not only of linguists and educators, but also of consumer-oriented terminologists, of specialists dealing with writing and reading disorders and, more recently, of researchers working in computational linguistics. The study that we present in this article belongs to the line of research that can be framed within Computational Linguistics, Language Technology (LT), Natural Language Processing (NLP) and Information Retrieval (IR). In these research areas, text complexity is

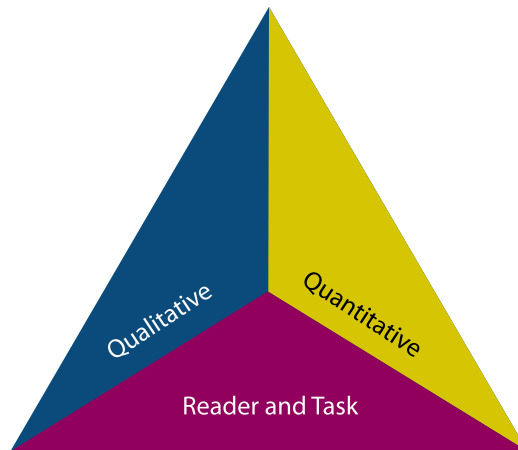


Figure 1. CCSS standard model (Common Core State Standards Initiative 2010)

tightly linked to corpus-based and data-driven analysis of textual difficulty, e.g., in second language acquisition (Lu 2010), and it is also connected to the development of LT applications, such as automatic readability assessment (Feng 2010) and automatic simplification, e.g. for those who suffer from dyslexia (Rello et al. 2013a). Text complexity can also be seen as a subfield of Text Simplification, which is a well-developed NLP task (Saggion 2017; Štajner, & Saggion 2018).

Text complexity is a concept inherently tied to the notion of readability. According to Wray and Janan (2013), readability can be redefined in terms of text complexity. As pointed out by Falkenjack (2018), readability incorporates both the actual text and a specific group of readers, such as middle school students (Dale & Chall 1949) or dyslexic people (Rello et al. 2013b), while text complexity seems to pertain to the text itself, or to the text and a generalised group of readers. Readability indices are practical and robust but coarse since they cannot account for the nature of the complexity. Critics of readability indices have also pointed out some genre-based discrepancies and the bias caused by short sentences and high frequency vocabulary on the readability scores (Hiebert 2012). It must be noted, however, that no perfect method exists to gauge text complexity and readability infallibly. Therefore, complexity and readability scores are useful, but they must be taken with a grain of salt.

3. Previous work

In this section we summarize previous approaches to detect the co-occurrence of linguistic features (Section 3.1) and to automatically determine the level of readability or text complexity (Section 3.2).

3.1 Multi-dimensional analysis

Biber (1988) describes in detail the application of factor analysis to linguistic data. Biber's multi-dimensional analysis refers to factor analysis (a bottom-up multi-variate statistical method) to uncover patterns of linguistic variation across the registers² collected in a corpus. The basic idea of MDA builds on the notion of "co-occurring linguistic features that have a functional underpinning". The co-occurrence of linguistic features across registers into factors is interpreted in terms of underlying textual dimensions (Biber 1988:121). Biber (1988) distinguishes between 'genre', i.e., text classes based on external cultural criteria, and 'text types', i.e., grouping of text that are similar in their linguistic form, irrespective of genre (Biber 1988:170). The ultimate goal is to build a universal text typology (Biber 1988:207; Biber 1989) based on the textual dimensions identified using factor analysis and cluster analysis. The term 'genre' has been replaced by the term 'register', which is now commonly used in Biberian research.

Multi-dimensional analysis is extensively applied in corpus linguistics (Sardinha & Pinto 2014) and can be applied to a wide range of investigations, from cross-linguistic comparisons (Biber 1995), to the description of discourse structure (Biber et al. 2007), from the analysis of register variation in the searchable web (Biber & Egbert 2016), to the prediction of grammatical text complexity (Biber et al. 2016). Multi-dimensional analysis has been applied not only to the English language but also to other languages, such as Spanish (Asención-Delaney & Collentine 2011), Portuguese (Sardinha et al., 2014) or Czech (Cvrček et al. 2020). To our knowledge, the study presented here is the first application of MDA to Swedish.

3.2 Readability-text complexity: Automatic approaches

Collins-Thompson (2014) presents an overview on how readability of texts can be assessed automatically. He reviews state-of-the-art algorithms for automatic modelling and for the prediction of reading difficulty and proposes new challenges and opportunities for future explorations. Napolitano et al. (2015) have developed the TextEvaluator system for providing text complexity and Common Core-aligned readability information. Detailed text complexity information is provided by eight component scores, presented in such a way as to aid the user's understanding of the overall readability metric, which is delivered as a holistic score on a scale of 100 to 2000. The user may also select a targeted US grade level and receive additional analysis relative to it.

2. In Biber (1988), "registers" are called "genres".

Several linguistic feature sets have been proposed to assess readability. Interesting findings were reported by Kate et al. (2010) who propose several feature sets to predict readability in a genre-diversified corpus. Using regression over a diverse combination of syntactic, lexical and language-model based features, they build a system that accurately predicts readability as judged by linguistically-trained expert human judges and exceeds the accuracy of naive human judges. Language-model based features were found to be most useful for this task, but syntactic and lexical features were also helpful. They also found that for a corpus consisting of documents from a diverse mix of genres, using features that are indicative of the genre significantly improve the accuracy of readability predictions. Tight relationships between readability and genres have also been detected by Dell’Orletta et al. (2013, 2014).

Pilán et al. (2016) presented a model for predicting linguistic complexity for Swedish second language learning material on a 5-point scale. Some researchers have linked the concept of notion of readability to the notion of text quality. For instance, Pitler and Nenkova (2008) combine lexical, syntactic, and discourse features to produce a highly predictive model of human readers’ judgments of text readability, and demonstrated that discourse relations are strongly associated with the perceived quality of text, while surface metrics, generally expected to be related to readability, are not very good predictors of readability judgments in the *Wall Street Journal Corpus*.

Modern models of readability analysis for classification often use classification algorithms such as Support Vector Machines (SVM) (Falkenjack et al. 2013; Feng et al. 2010; Petersen 2007), which establish whether a text is easy-to-read or not. Such models have very high accuracy. However, they do not tell us much about the levels of readability, since they are based on binary classification.

In the study that we present in this article, we resume and expand the previous work carried out on the SUC (Falkenjack et al. 2016; Jönsson et al. 2018; Santini et al. 2019). It is to be noted that while all previous computational approaches to the SUC are based on supervised classification, here we take a different approach and explore whether MDA can help elicit text complexity facets across SUC registers.

4. Method

In this section, we describe the SUC corpus and dataset and present MDA, together with the factor solution proposed in this study.

4.1 The SUC corpus and dataset

The SUC is a collection of Swedish texts that represent the Swedish language of the 1990's. The corpus consists of about 1 million tokens with manually checked base forms, part-of-speech tags and morphological information.³ The SUC follows the general layout of the Brown corpus and the LOB corpus, with 500 samples of texts around 2,000 words each.

It is worth stressing that the SUC was created to represent the Swedish language, and not to represent the different nature of text varieties. As a matter of fact,

the texts of the SUC can be excerpts from longer texts (as from books), single whole texts, or composed of several short texts. The latter is often the case with articles from newspapers and magazines, which rarely are as long 2.000 [sic] words in themselves. (Källgren et al., 2006)

To date, the SUC is still the only general-purpose corpus that officially represents the Swedish language at one point in time. The new *Koala Multi-genre Annotated Swedish Corpus* (Adesam et al. 2018) is still in its beta version.

The SUC is the only Swedish corpus from which a text complexity dataset has been extracted though SAPIS (Fahlborg & Rennes 2016), an API Service for Text Analysis and Simplification of Swedish text. The SUC dataset returned by SAPIS contains 120 linguistic features described in Falkenjack et al. (2013).⁴ This dataset is the source dataset used in the study we describe in this article. Technically speaking, the SUC is divided into 1,040 bibliographically distinct text chunks, each assigned to a 'genre' and 'subgenre'.

4.2 Multi-dimensional analysis: Technicalities

In this section, we describe and motivate the technical choices made when carrying out the MDA. It is worth noting that factor solutions remain tentative until confirmed by an extrinsic evaluation that shows the functional and practical validity of the elicited factors. All the statistical procedures described in this study have been run in R.⁵

3. The SUC corpus can be downloaded from <<https://spraakbanken.gu.se/eng/resources/suc>>.

4. The dataset is available on the companion website.

5. The R code is available on the companion website.

4.2.1 Variable screening

A main tenet of MDA is to identify the linguistic features that can then be explained functionally once the factors have been extracted. Multi-dimensional analysis, when applied to English register analysis, is normally based on 67 linguistic variables that have been comprehensively described in Biber (1988) and successive work. Since the SUC is in Swedish and since our ultimate purpose is to identify text complexity facets rather than analysing registers, we relied on a feature set that has been created for the Swedish language and for text complexity and readability (Falkenjack et al. 2013, Section 2).

We started off from the SUC dataset extracted from the SUC corpus via SAPIS. The dataset contains 1,040 records and 120 features. We set apart lexical and grammatical features from readability scores, since we want to use readability measurements independently from the other features in order to perform an extrinsic evaluation of the factors.

We noticed that some of the linguistic features in the dataset were somewhat redundant. For example, both *pos_det* and *dep_det* refer to the number of the determiners; similarly, the feature *lexicalDensity* was overlapping with other variables like nouns or determiners. This redundancy is detrimental for MDA because it causes *multicollinearity*, a statistical phenomenon that may lead to distorted results. Collinearity can be detected in several ways. We identified the multicollinear variables using `eigen()` (R base function), which returns the magnitude of the variables. If the magnitude is not uniform and some values are very high, then multicollinearity is affecting the dataset and measures must be taken. We ditched out multicollinear features using `lm()` (R base function) and ended up with 45 linguistic features listed in the Appendix.⁶

4.2.2 Running multi-dimensional analysis

After having screened the variables, we carried out MDA by building a correlation matrix, checking the determinant, assessing the sample adequacy and finally by determining the number of factors.

The correlation matrix was built using the Kendall correlation coefficient. Kendall is non-parametric coefficient, and we chose it because we did not want to make any assumptions about the underlying population.

The determinant of a correlation matrix is difficult to define in simple words, but it is a very useful number. The determinant of the correlation matrix will equal 1.0 only if all correlations equal 0, otherwise the determinant will be less than 1. The determinant is related to the volume of the space occupied by the bulk

6. This dataset is available on the companion website.

of data points represented by standard scores on the measures involved. When the measures are uncorrelated, this space is a sphere with a volume of 1. When the measures are correlated, the space occupied becomes an ellipsoid whose volume is less than 1. The threshold for acceptability is normally a determinant greater than 0.00001, which also indicates the absence of multicollinearity (Field 2000: 445).⁷ Additionally, when the determinant of a correlation matrix is smaller than 0.00001, the correlation matrix is ‘not positive definite’ and factor analysis cannot be run. The determinant `det()` of our correlation matrix is 3.170923e-05, which is greater than 0.00001. Hence, the value of determinant indicates that the correlation matrix is suitable to run factor analysis.

Additional measures can be used to assess the suitability of a sample for factor analysis. We used two measures, namely Kaiser-Meyer-Olkin (KMO) and Bartlett’s test of sphericity. The KMO index `KMO()` indicates the Measure of Sampling Adequacy (MSA) of factor analytic data matrices. KMO returns values between 0 and 1. The rule of thumb for interpreting the statistic is as follows: KMO values between 0.8 and 1 indicate the sampling is adequate; KMO values less than 0.5 indicate the sampling is not adequate. The overall MSA for our correlation matrix is 0.87 which means that a factor analysis may be useful with our data.

Bartlett’s test of sphericity `cortest.bartlett()` tests the hypothesis that a correlation matrix is an identity matrix, which would indicate that variables are unrelated and therefore unsuitable for factor structure detection. Small values of the significance level ($p < 0.05$) indicate that a factor analysis may be useful. The significance level for our correlation matrix is 0 (zero) which indicates that our data is suitable for factor analysis.

Since the sample and the correlation matrix are adequate for investigating the factorial structure underlying the 45 variables, we must determine the best number of factors that account for the latent structure. The key concept of factor analysis is that multiple observed variables have similar patterns of responses because they are all associated with a latent (i.e. not directly measured) ‘factor’. Deciding the number of factors is not straightforward. Traditionally, decisions are made looking at the scree plot (Cattell 1966). However, it has been pointed out that the interpretation of scree plots can be subjective and arbitrary (Ledesma et al. 2015). More recently, it has been shown that parallel analysis (Hayton et al. 2004) can help identify the most suitable number of factors. Parallel analysis compares the raw data eigenvalues to the percentile. When the percentile becomes larger

7. See also other public discussions of how to define a determinant of a correlation matrix, e.g. <<https://www.quora.com/What-does-the-determinant-of-the-correlation-matrix-represent>>, retrieved 13 January 2020.

than the eigenvalue, then the factor is not statistically significant. We ran an R implementation of Horn's original parallel analysis (Horn 1965) called `paran()` on the correlation matrix to identify the number of factors that best represent the data. The result of this parallel analysis suggests retaining three factors (see scree plot shown in Figure 2). The magnitude of these three factors varies (see adjusted eigenvalues in Figure 3). The first factor is very large (10.58), while the other two are considerably smaller (both around 1.0).

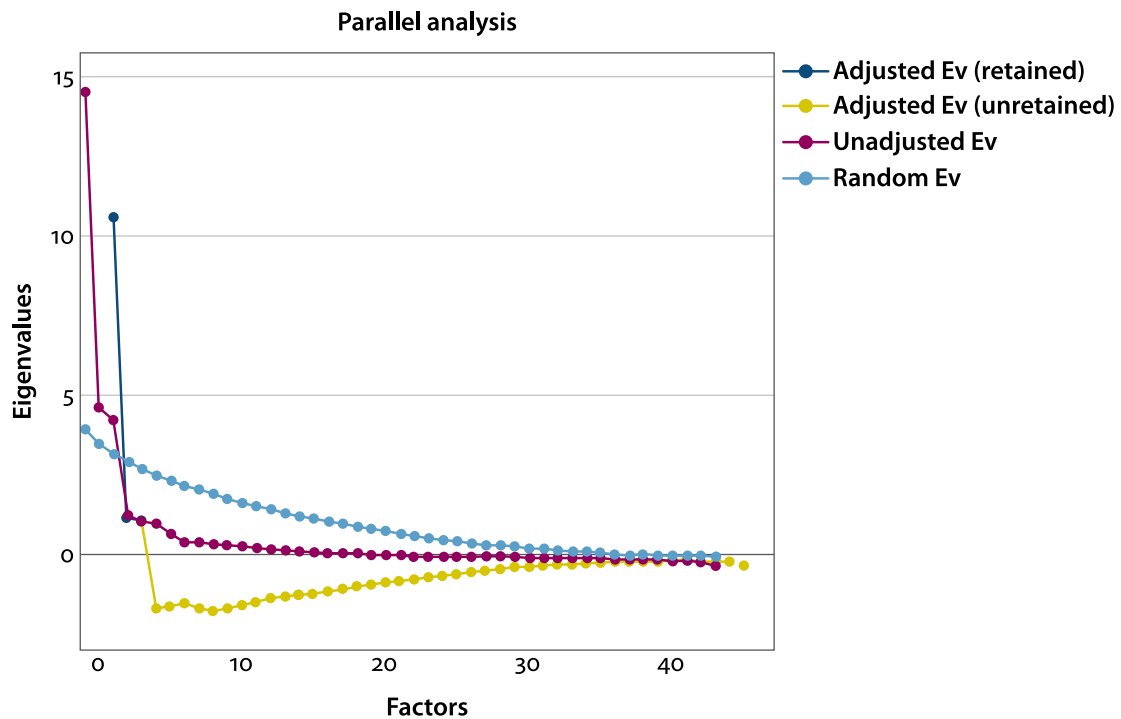


Figure 2. Parallel Analysis: 45 variables

4.2.3 Three-Factors solution

We extracted three factors from the correlation matrix using `factanal()` (R base function), which by default performs maximum likelihood estimation. We apply the oblique rotation called 'promax', as recommended in Biber (1988). We ditched out the loadings smaller than 0.30 (a common practice). Loadings are correlations with the unobserved factors. Normally, each of the identified factors should have at least three variables with high factor loadings, and each variable should load highly on only one factor. More precisely, a 0.30 loading translates to approximately 10 percent variance explanation, and a 0.50 loading denotes that 25 percent of the variance is accounted for by the factor. The loading must exceed 0.70 for the factor to account for 50 percent of the variance of a variable. The loadings of the 3-factors solution are shown in Table 1.

Figure 3. Snippet of the output: Eigenvalue magnitude

Result of Horn's Parallel Analysis for factor retention 5000 iterations, using the 95 centile estimate			
Factor	Adjusted Eigenvalue	Unadjusted Eigenvalue	Estimated Bias
1	10.586225	14.536998	3.950773
2	1.178404	4.664057	3.485652
3	1.076877	4.245909	3.169031
4	-1.665968	1.262513	2.928481
5	-1.626151	1.078842	2.704993
6	-1.531345	0.991574	2.522920
7	-1.672954	0.673953	2.346907
8	-1.745216	0.444235	2.189452
9	-1.648947	0.393657	2.042604

In Figure 4, ‘SS loadings’ indicates the sum of squared loadings. This value is sometimes used to determine the value of a particular factor. Normally, a factor is worth keeping if the SS loading is greater than 1. All loadings are greater than 1. ‘Proportion Var’ is simply the proportion of variance explained by each factor. ‘Cumulative Var’ tells us the cumulative proportion of variance explained and ranges from 0 to 1. For this factor solution, the explained variance is 0.22.

Admittedly, the three factor solution accounts for a relatively small proportion of the overall variance. Normally, when undertaking a factor analysis, a researcher must make several interrelated methodological decisions. One decision is how to determine the appropriate amount of variance to factor analyze. Different rules of thumb exist in different fields. For instance, in the natural sciences, factors should account for at least 95 percent of the variance. In contrast, in the social sciences, where information is often more elusive, it is common to consider a solution that accounts for 60 percent of the total variance or even less.⁸ Natural language data and linguistic features are often more slippery than social science data. The linguistic data that we find in the texts of a corpus can be very idiosyncratic and ambiguous. This elusiveness is reflected in the factor solution.

Even if the factor solution produced for the SUC shows quite low percentage of explained variance, it consists of a “strong” factor (10,58) and two “weak” factors (see Figure 3). In practice, one good factor can be scientifically quite useful

8. See also other public discussion on this topic, e.g. <https://www.researchgate.net/post/If_i_have_good_result_with_low_variance_explanation_in_exploratory_factor_analysis_is_there_any_problem_to_be_proceed>, retrieved 13 January 2020.

Table 1. Loadings on each of the three factors

Linguistic Features	Factor1	Factor2	Factor3
pos_JJ.adjective		0.62	
pos_DT.determiner		0.44	
pos_NN.noun	-0.64		
pos_VB.verb			0.63
pos_IE.infinitivalMarker			0.39
pos_IN.interjection	0.56		
pos_SN.subordinatingConj			0.53
pos_PM.properNoun			-0.43
pos_PN.pronoun	0.66		
pos_AB.adverb	0.57		
pos_PP.preposition	-0.65		
pos_PS.possessivePronoun	0.37		
pos_PC.participle		0.36	
<i>ratioSweVocC</i>	0.45		0.49
dep_AN.apposition			-0.47
dep_AT.premodifier		0.76	
dep_I.questionMark	0.52		
dep_IK.comma			-0.39
dep_IP.period		-0.46	
dep_IU.exclamationMark	0.44		
dep_KA.comparativeAdverbial	0.36		
dep_MA.attitudeAdverbial	0.55		
dep_NA.NegationAdverbial	0.41		
avgSentenceDepth		0.49	
avgVerbalArity			-0.42
avgNominalPremodifiers		0.87	
<i>avgNominalPostmodifiers</i>	(-0.36)	0.45	

to get some insights into the data. For this reason, we proceed further with a 3-factors solution and in the next section, we will assess how informative this 3-factor solution is.

Figure 4. Loadings and Variance

	Factor1	Factor2	Factor3
SS loadings	4.19	3.43	2.41
Proportion Var	0.09	0.08	0.05
Cumulative Var	0.09	0.17	0.22

5. Meaningful factors? Evaluation and interpretation

It is important that the rotated factors make theoretical sense. If the variables that are loading on the same factor make sense together and it is possible to name the concept or the function they represent, then this is an indication that the factor solution is a reasonable one.

5.1 Evaluation: Correlating LIX scores & factor scores

As motivated earlier, the evaluation of the factor solution is carried out by correlating the factor scores of the three factors against LIX scores, which have been kept separated from any calculation of MDA factor solution.

In the following paragraphs we briefly describe how to compute and interpret LIX scores, explain what factor scores are and what they represent and finally we discuss what the correlation between these two types of scores indicates.

Readability for the Swedish language has a rather long tradition. One of the most popular, easy-to-compute formulas is LIX (that stands for “Läsbarhetsindex”, en: ‘Readability Index’) proposed in Björnsson (1968) and equipped with a public online calculator.⁹ LIX is formulated in a way similar to Flesch metric (Flesch 1948). LIX formula (Formula 1) accounts for the average amount of words per sentence added to the percentage of long words (more than 6 characters) divided by the total amount of words:

$$\frac{A}{B} + \frac{C \cdot 100}{A}$$

where:

A is the number of words

B is the number of periods (defined by period, colon or capital first letter)

C is the number of long words (more than 6 letters)

Formula 1. “Läsbarhetsindex”, en: ‘Readability Index’ as reported in Siteimprove (2020)

9. See <https://www.lix.se/index.php> website.

LIX scores give an indication of reading levels. A possible interpretation of the levels of readability are shown in Table 2.

Table 2. LIX scores and readability levels

LIX Scores	Readability Level	Examples of Text Varieties
< 20	Very easy	Very simple texts, e.g. children's books
20–30	Easy	Simple texts: e.g. easy-to-read texts, simplified Wikipedia etc.
31–40	Average	Normal texts: fiction, reportage, everyday conversation, etc.
41–50	Difficult	Normal texts (more complex texts): argumentative texts, editorial, factual texts, technical texts, academic articles, etc.
> 50	Very Difficult	Specialized texts (very complex texts): bureaucratic texts, legal texts, etc.

For each of the texts in the SUC, a LIX readability score has been calculated separately from the MDA. These LIX scores will be correlated to the factor scores of the three factors returned by MDA for the purpose of evaluation.

Factor scores are numerical values that provide information about an individual text's placement on the factor(s) (DiStefano et al. 2009). Since factors are latent variables that underly the observed variables, the interpretation of each of these factors is based on the content of the original variables. Factor scores, therefore, indicate the score of each text on the underlying latent variable. Factor scores can be computed using several mathematical methods (DiStefano et al. 2009). In this study, we generated regression factor scores for each of the 1040 SUC texts.

Correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables becomes weaker. The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a - sign indicates a negative relationship.

There exist several formulas to compute correlation coefficients (e.g. Pearson's, Kendall's, Spearman's, etc.). In this study, we use the non-parametric Kendall correlation because we do not wish to make any assumption about the distribution of the population. Normally, Kendall's correlation coefficient is called "*tau*". However, when the Kendall's correlation is plotted in a scatter plot using R, the "*tau*" is called "*R*" (see top-right corner of Figures 5, 6 and 7) and statistical

significance is indicated by “ p ” (which stands for p value). In the discussion below, we use “ R ” (instead of τ) to be consistent with the legend in the figures.

Usually, a note of caution is needed when using statistical indicators, like p values, that help determine how certain we are that the results observed did not arise by chance. Normally, when small sample sizes are used, the risk is high that observations will be due to chance. However, studies with larger sample sizes can detect tiny or small associations that might not be important or relevant. It is indeed important to know the statistical significance of a result, since without it there is a danger of drawing firm conclusions from studies where the sample is too small to justify such confidence. However, statistical significance does not tell us the full story. One way to overcome confusion is to report the effect size. Effect size is a statistical concept that measures the strength of the relationship between two variables on a numeric scale. The effect size is usually measured in three ways: (1) standardized mean difference, (2) odd ratio, or (3) correlation coefficient (Field et al. 2012). Here we use Kendall’s correlation coefficient. By using this approach, we can get a good indication of the reliability of our findings based on the value of the correlation coefficient together with its statistical significance.

5.1.1 *Factor 1 scores & LIX scores*

We observe a moderate negative correlation between Factor₁ and LIX scores (Figure 5). R is -0.32 , the p value of the test is $< 2.2e-16$, which is less than the significance level $\alpha = 0.05$. We can conclude that LIX scores and Factor₁ scores are significantly correlated.

5.1.2 *Factor 2 scores & LIX scores*

LIX scores and Factor 2 scores show a moderate positive correlation (Figure 6). R is -0.33 , the p value of the test is $< 2.2e-16$. We can conclude that LIX scores and Factor 2 scores are significantly correlated.

5.1.3 *Factor 3 scores & LIX scores*

Factor 3 scores and LIX scores show a weak negative correlation (Figure 7), R is -0.013 , the p value of the test is 0.52 . We can conclude that LIX scores and Factor 3 scores are NOT significantly correlated since the p value is greater than 0.05 .

5.1.4 *Summary*

In summary, for Factor 1 and Factor 2, we can reject the null hypothesis that MDA and LIX produce completely different results. From a statistical evaluation perspective, this means that the two factors make sense and they are not an ungrounded outcome of MDA. As emphasized previously, LIX is not a perfect measure, but it gives a reasonable assessment of readability. We cannot however

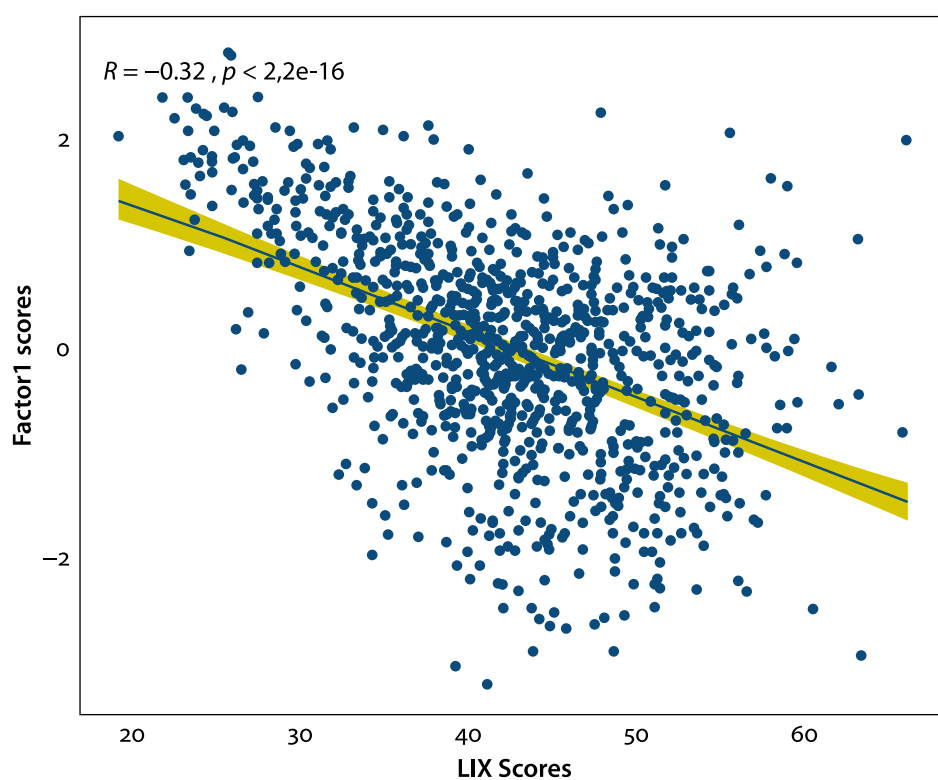


Figure 5. Scatter plot, Factor1-LIX

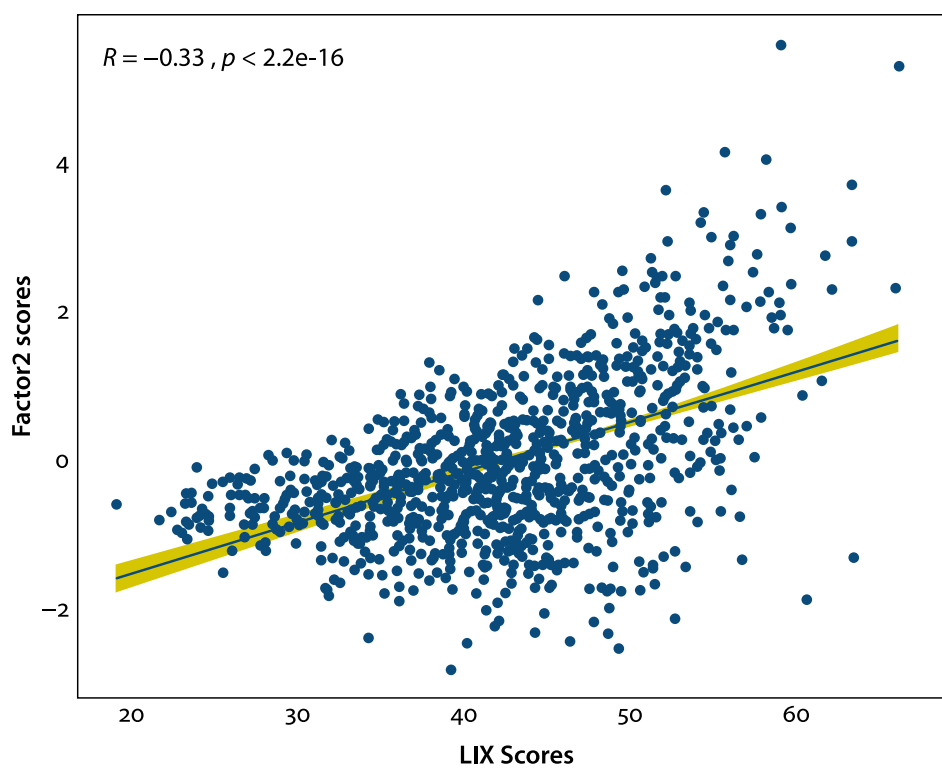


Figure 6. Scatter plot, Factor2-LIX

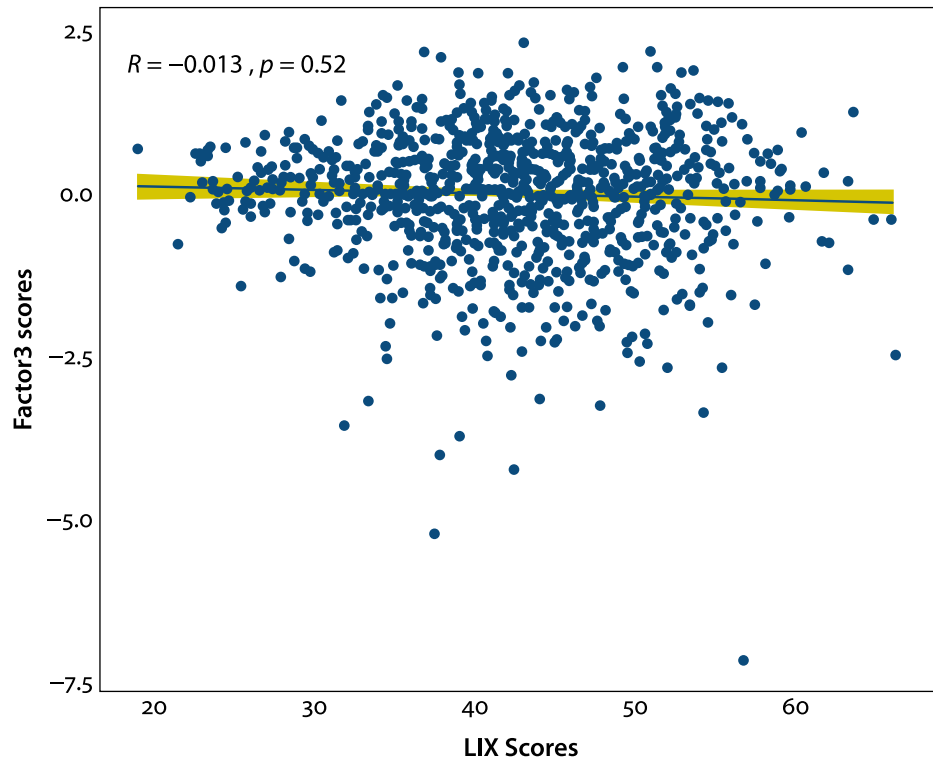


Figure 7. Scatter plot, Factor 3-LIX

state the same for Factor 3 because its correlation with LIX is weak and statistically not significant.

In Figure 8, the three factors are plotted together with LIX scores to show the overall characterization of the SUC. Figure 8 is indicative, but not accurate enough. In the next section, we explore the makeup of the factors and their composition more in-depth and propose a functional interpretation.

5.2 Interpretation: Signed dimensions & text complexity facets

Each factor has a positive and a negative side, which means that the features collected on one side of a factor tend to be mutually exclusive with the features grouped on the opposite side. According to the MDA framework (Biber 1988:101ff), when interpreted functionally, a factor becomes a ‘textual dimension’ that can be explained functionally and linguistically, thus helping shed light on the co-occurrence of certain linguistic features in a certain group of texts. In this study, we isolate each side of a factor, and refer to it as a ‘signed dimension’, where the ‘sign’ (i.e. the symbols + and –) indicates the positive side and the negative side of a factor. We then attempt to interpret each signed dimension as a text complexity facet.

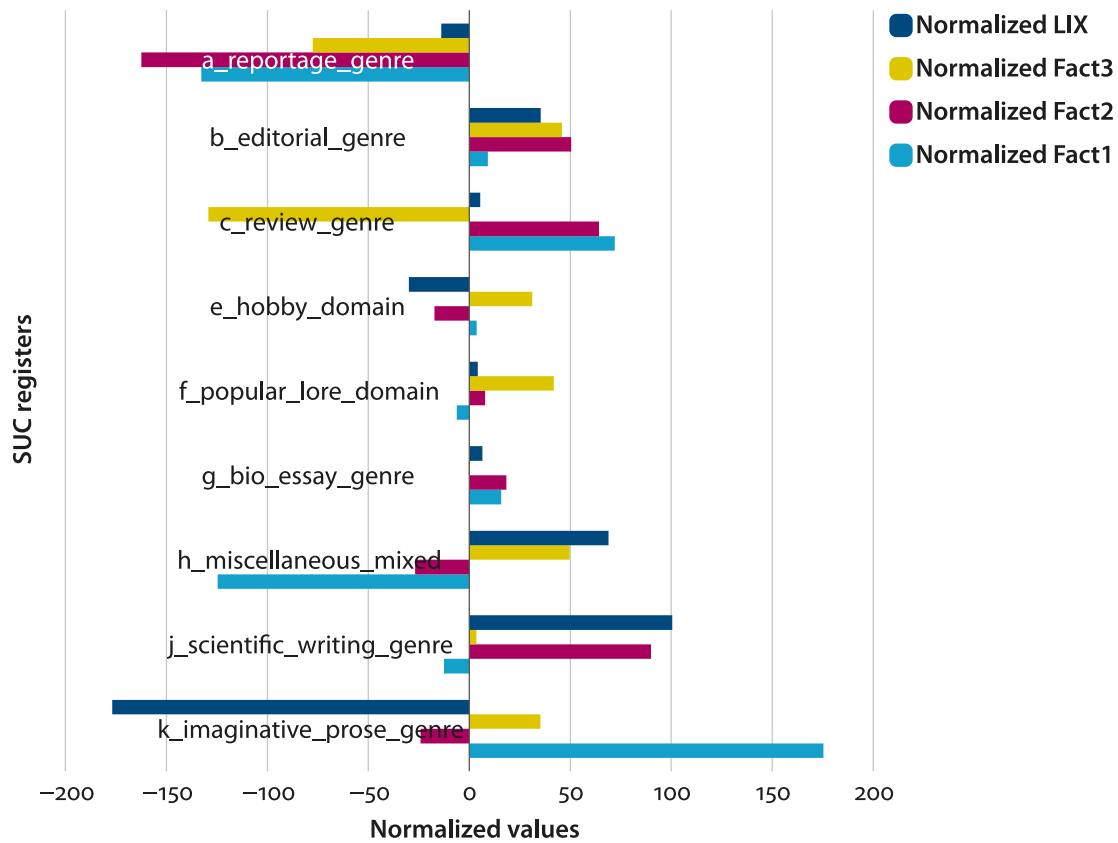


Figure 8. LIX and factors across SUC registers, normalized scores

5.2.1 Factor1: Dim1+ & Dim1–

The features gathered on the two sides of Factor 1 show the well-explored divide between spoken language (positive side) and written language (negative side). The positive side collects linguistic features that are more frequent in the spoken language like exclamations, questions, pronouns etc., while in the negative side, nominal devices are grouped together (Table 3).

5.2.2 Dim1+: Pronominal-Adverbial (spoken-emotional) facet – Average readability

Most of the features that tend to co-occur in Dim1+ (namely, *pronouns*, *adverbs*, *interjections*, *attitude adverbials*, *question marks*, *ratioSweVocC*, *exclamation marks*, *negation adverbials*, *possessive pronouns* and *comparative adverbials*) show the spoken and emotional nature of the dimension.

The largest loading is on pronouns. Functionally speaking, “pronouns replace fully specified noun phrases and can be regarded as an economy device.” (Biber et al. 1999). Essentially, pronouns mark “a relatively low informational load, a lesser precision in referential identification, or a less formal style” (Biber 1988:225). Attitude (a.k.a. stance) adverbials are more frequent in conversation

Table 3. Factor 1, Dim1+ (top) and Dim1– (bottom)

LinguisticFeatures	Factor1
pos_PN.pronoun	0.66
pos_AB.adverb	0.57
pos_IN.interjection	0.56
dep_MA.attitudeAdverbial	0.55
dep_I.questionMark	0.52
ratioSweVocC	0.45
dep_IU.exclamationMark	0.44
dep_NA.NegationAdverbial	0.41
pos_PS.possessivePronoun	0.37
dep_KA.comparativeAdverbial	0.36
pos_PP.preposition	–0.65
pos_NN.noun	–0.64

than in any other types of texts (Biber et al. 1999: 765–766): “Speakers use stance adverbials to convey their judgments and attitudes, to claim the factual nature of what they are saying, and to mark exactly how they mean their utterances to be understood”. Negations are more common in the spoken language than in the written language. The higher frequency distribution in the spoken is attributed to “the greater frequency of repetitions, denials, rejections, questions and verbs in speech” (Biber 1988: 245). Questions indicate a concern with “interpersonal functions and involvement with the addressee” (Biber 1988: 227), while interjections are expressive of the speaker’s emotion (Biber et al., 1999: 1083). Interpersonal involvement and emotion find an easier way to verbal expression in the spontaneous speech. A high ratio of SweVoc words (Mühlenbock 2013) indicate a more easy-to-read text. In particular, the SweVocC lemmas are those fundamental for communication. This indicates that the vocabulary is common and fast like the lexicon normally used in everyday conversations. Possessive pronouns and comparative adverbials are also more frequent in conversation than in the written language.

All these features characterize spoken language and emotional communicative interaction, and in particular they suggest spontaneous communication. This type of language is used not only in real-world communication but also in those types of texts that rely on the imitation of the spoken and emotional language, for example imaginative prose.

In order to get an indication of how reliable this interpretation is, we correlate all the Dim1+ scores that characterize 545 SUC texts to the LIX scores calculated for these texts. All the results are shown in Table 6, row 1.

There exists a moderate statistically significant negative correlation between Dim1+ and the matching LIX scores. The mean of the LIX scores that match Dim1+ is 39.57 with a standard error (SE) of 0.36 (see Table 6). The standard error reflects the reliability of the mean. A small SE is an indication that the sample mean is a more accurate reflection of the actual population mean. The SE of 0.36, being relatively small, gives us an indication that the LIX mean is reliable. A mean of 39.57 indicates that the readability level is *average* (cf. Table 2).

It appears that k_imaginative_prose_genre (129 texts), c_review_genre (109 texts) and a_reportage_genre (87 texts) are highly connotated by this facet and this level of readability, while f_popular_lore_domain is the least characterized by them (Figure 9).

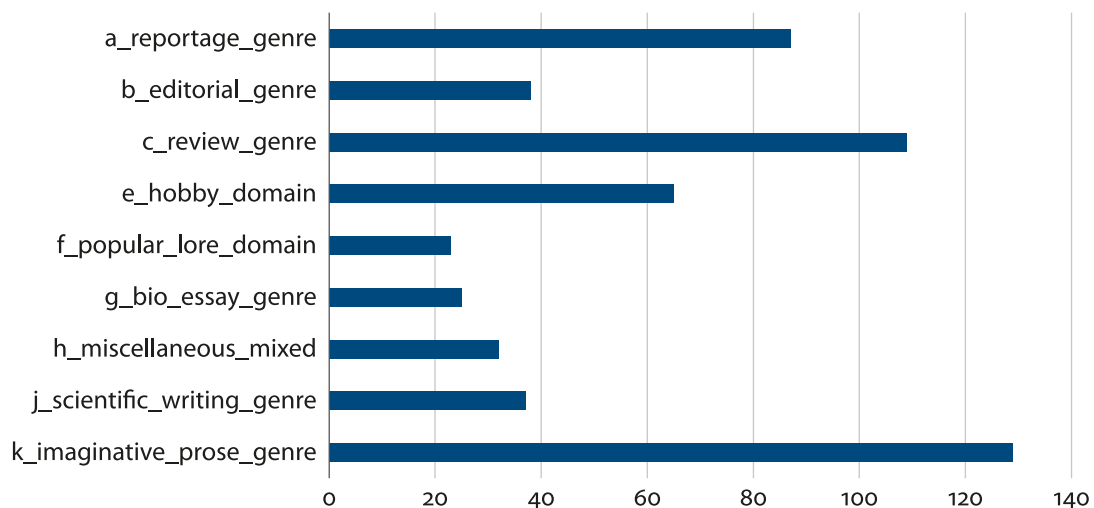


Figure 9. Distribution of SUC registers along Dim1+

The facet elicited by Dim1+ has a *pronominal-adverbial* linguistic nature and a *spoken-emotional* functional focus. It is characterized by an average level of readability, i.e., the texts that belong to this facet are relatively unproblematic to read for the ‘ordinary’ Swedish reader. To show how this linguistic characterization is reflected in the actual texts, we show the content (in Swedish and English) of a text that has a high Dim1+ score (Box 1). The style is simple and spontaneous. It is connotated by short sentences and many personal pronouns that indicate emotional involvement.

Box 1. An excerpt of a SUC text connotated by the pronominal-adverbial facet with a spoken-emotional focus

Dim1+ score	LIX score	Filename	SUC Register
2.82	2.60	kl10.xml	k_imaginative_prose_genre
Swedish		English translation	
Men det är mer en journal.		But it's more of a journal.	
Och inte fick jag laga maskinen heller.		And I couldn't fix the machine either.	
Då blev han dyster igen.		Then he became gloomy again.	
Men att dom är sorgsna är alldeles klart.		But that they are sad is perfectly clear.	
Allt du behöver göra för att vinna henne tillbaka är		All you have to do to win her back is to	
att visa att du älskar henne, mer än du älskar		show that you love her more than you love	
hundarna.		the dogs.	

5.2.3 Dim1–: Nominal (informational) facet – Difficult readability

This dimension has two loadings, both quite high, namely on *prepositions* and *nouns*, that both indicate the nominal character of this dimension. More specifically, “prepositions are an important device for packing high amounts of information in a text and can be described as a device that is used to expand and elaborate on the idea unit expressed by nouns.” (Biber 1988: 237) and “nouns are the primary bearers of referential meaning in a text and a high frequency of nouns indicates a great density of information” (Biber 1988: 104).

In order to get an indication of how reliable this interpretation is, we correlate all the Dim1– scores that characterize 495 SUC texts to the LIX scores calculated for these texts. All the results are shown in Table 6, row 2.

There exists a statistically significant, but weak, negative correlation between Dim1– and the matching LIX scores. The mean of the LIX scores matching Dim1– is 45.19 with a standard error of 0.29. A mean of 45.19 indicates that the readability level is *difficult* (cf. Table 2).

From this analysis, it appears that a_reportage_genre (182 texts), h_miscellaneous_mixed (132), and e_hobby_domain (59) are highly characterized by this facet and readability level, while k_imaginative_prose_genre is the least characterized by them (Figure 10).

The facet elicited by this dimension has a *nominal* linguistic nature and an *informational* functional focus. As shown in the excerpt reported in Box 2, the style is heavily informational and characterized by nouns and nominalizations. Personal pronouns are not used. Sentences are much longer than those shown in the excerpt in Box 1. The level of readability that combines with this facet is more demanding and requires the knowledge of more advanced lexicon that expresses complex cultural or social situations (e.g. *flyktingförläggningen*, en: *refugee camp*)

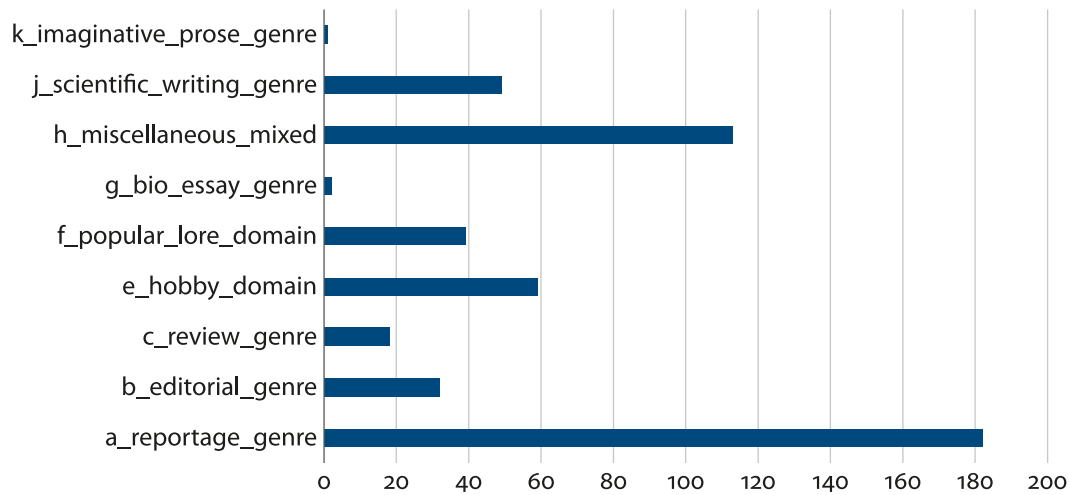


Figure 10. Distribution of SUC registers along Dim1–

Box 2. An excerpt of a SUC text connotated by the nominal facet with an informational focus

Dim1– score	LIX score	Filename	SUC Register
3.20	41.08	afo6j.xml	a_reportage_genre
Swedish		English translation	
Man i Älvkarleö anhållen för hot		Man in Älvkarleö arrested for threats	
En 33-årig man vid flyktingförläggningen i Älvkarleö greps på måndagskvällen av Tierpspolisen. Mannen är misstänkt för olaga hot och misshandel av sin hustru.		A 33-year-old man at the refugee camp in Älvkarleö was arrested by the Tierp's police on Monday evening. The man is suspected of unlawful threats and mistreatment of his wife.	

5.2.4 Factor 2: Dim2+

As shown in Table 4, Factor 2 has only the positive side (since there is only a negligible loading on the negative side).

5.2.5 Dim2+: Adjectival (information elaboration) facet – Difficult readability

This dimension has an adjectival nature. As shown in Table 4 (top), *premodifiers*, *postmodifiers*, and *adjectives* have the highest loading on this dimension, and they are all grammatical devices that elaborate and specify the exact nature of the nominal referents (Biber 1988:140). The greater frequency of adjectives in the written registers reflect the reliance on noun phrases to present information, as pointed out in Biber et al. (1999: 506).

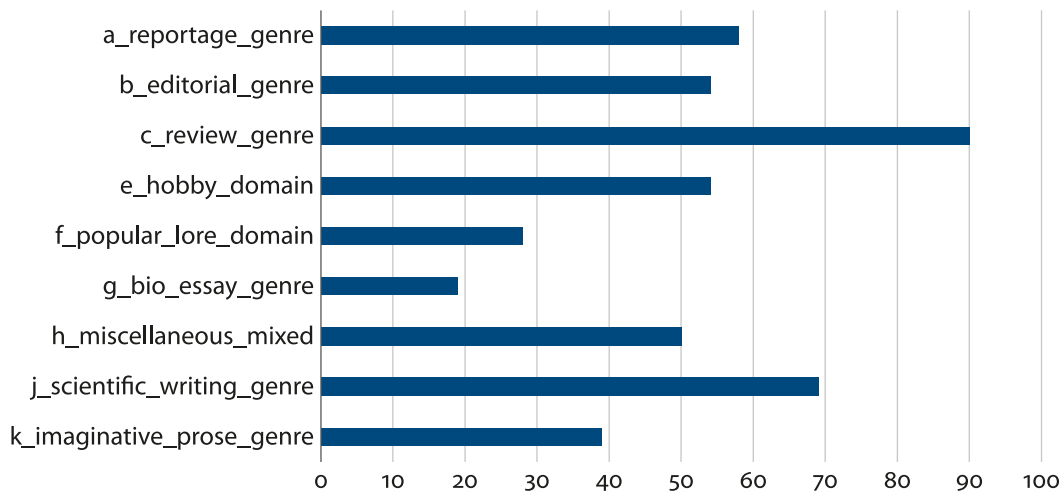
In order to get an indication of how reliable this interpretation is, we correlate all the Dim2+ scores that characterize 461 SUC texts to the LIX scores calculated for these texts. All the results are shown in Table 6, row 3.

Table 4. Factor2, Dim2+ (top), Dim2– (bottom)

LinguisticFeatures	Factor2
avgNominalPremodifiers	0.87
dep_AT.premodifier	0.76
pos_JJ.adjective	0.62
avgSentenceDepth	0.49
avgNominalPostmodifiers	0.45
pos_DT.determiner	0.44
pos_PC.participle	0.36
dep_IP.period	(−0.46)

There exists a moderate statistically significant positive correlation between Dim2+ and the matching LIX scores. The mean of the LIX scores matching Dim2+ is 46.38 with a standard error of 0.31. A mean of 46.38 indicates that the readability level is *difficult* (cf. Table 2).

SUC registers that are characterized by this facet and the corresponding level of readability are c_review_genre (90 texts), j_scientific_writing_genre (69), and a_reportage_genre (58), while g_bio_essay_genre is the least characterized by them (Figure 11).

**Figure 11.** Distribution of SUC registers along Dim2+

The facet elicited by this dimension has an *adjectival* linguistic nature and the *elaboration of information* as functional focus. As shown in Box 3, texts connotated by this facet contain long sentences and many adjectives and past participles. The information given to the reader is elaborated and enriched by the use of these

modifiers. The level of readability associated with this facet is rather difficult for the ‘ordinary’ reader.

Box 3. An excerpt of a SUC text connotated by the adjectival facet with elaboration of information focus

Dim2+ score	LIX score	Filename	SUC Register
5.57	59.07	ba05d.xml	b_editorial_genre
Swedish		English translation	
Detta har förstärkt de farhågor som vuxit fram på den franska sidan av den omskrivna samarbetsaxeln för att man skall få en obunden tysk stormakt som svårhanterlig granne.		This has reinforced the fears that have emerged on the French side of the rewritten axis of cooperation in order to gain an unbounded German great power as a difficult-to-manage neighbor.	

5.2.6 Factor 3: Dim3+ & Dim3–

The features gathered on the two sides of Factor 3 are shown in Table 5. The positive side collects mostly *verbs*, *infinitival markers*, *subordinating conjunctions* and *easy vocabulary*. These grammatical features often co-occur in the spoken language. On the negative side, we observe a nominal characterization, with *appositions* and *commas* (a punctuation device that is often used together with appositions to indicate the nominal expansion of information) and the arguments associated with verbs.

Table 5. Factor3, Dim3+ (top), Dim3– (bottom)

LinguisticFeatures	Factor3
pos_VB.verb	0.63
pos_SN.subordinatingConj	0.53
ratioSweVocC	0.49
pos_IE.infinitavalMarker	0.39
dep_AN.apposition	–0.47
avgVerbalArity	–0.42
dep_IK.comma	–0.39

5.2.7 Dim3+: Verbal (engaged) facet – Difficult readability

The features that characterize Dim3+ are *verbs*, *subordinators* and *infinitival markers* and *easy communication vocabulary*. A high number of subordinating conjunctions “seem to be associated with the expression of information under real

time production constraints, where there is little opportunity to elaborate through a precise lexical choices” (Biber 1988: 107).

In order to get an indication of how reliable this interpretation is, we correlate all the Dim3+ scores that characterize 599 SUC texts to the LIX scores calculated for these texts. The results are shown in Table 6, row 4.

There exists a weak statistically significant positive correlation between Dim3+ and the matching LIX scores. The mean of the LIX scores matching Dim3+ is 42.02 with a standard error of 0.34. A mean of 42.029 indicates that the readability level is slightly *difficult*, but less difficult than Dim2+ (see Table 2).

SUC registers that are characterized by this facet and the corresponding level of readability are a_reportage_genre (126 texts), h_miscellaneous_mixed (104) k_imaginative_prose_genre (100), while g_bio_essay_genre and c_review_genre are the least characterized by them (Figure 12).

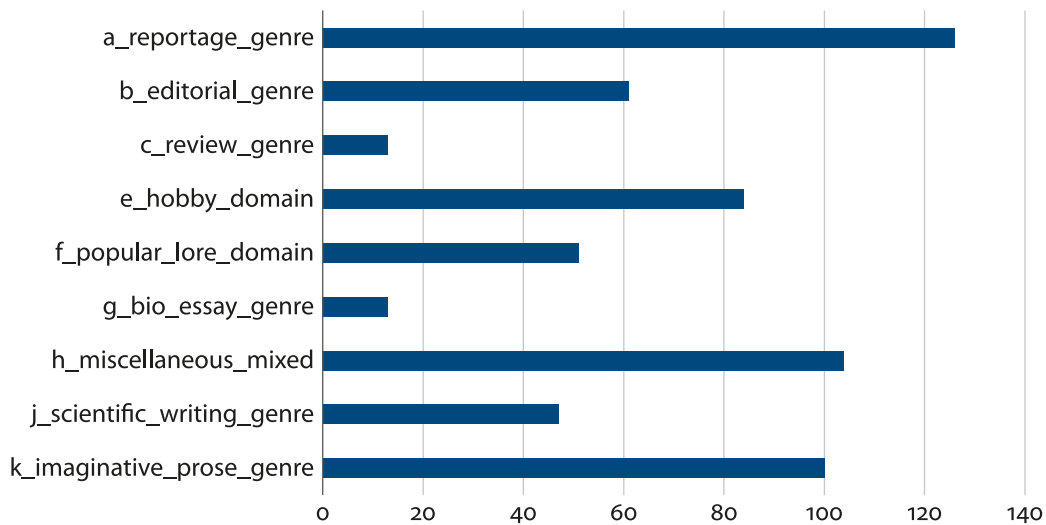


Figure 12. Distribution of SUC registers along Dim3+

The facet elicited by this dimension has a *verbal* linguistic nature and some level of communicative *engagement* as functional focus. The verbal nature of the facet is characterized by an interactive or conversational style, but at the same time, the high frequency of subordinating conjunctions also indicates an effort to articulate the discourse more accurately. As shown in Box 4, texts connoted by this facet contain long sentences, many verbs and subordinators.

5.2.8 Dim3–: *Appositional (information expansion) facet – Difficult readability*

Appositions are “a maximally abbreviated form of postmodifier, and they include no verbs. Appositions are thus favoured in the type of texts with high informational density.” (Biber et al. 1999: 639). Commas are common punctuation device

Box 4. An excerpt of a SUC text connotated by the verbal facet with an engaged focus

Dim3+ score	LIX score	Filename	SUC Register
2.32	43.09	he09c.xml	h_miscellaneous_mixed
Swedish		English translation	
När journaler överförs per telefax finns risk för att obehöriga kan ta del av dem, inte minst om den som faxar råkar knappa in fel nummer.		When journals are transmitted by fax, there is a risk that unauthorized persons can access them, not least if the person who faxes accidentally dials the wrong number.	

to specify apposition. Verb arity indicates the number of arguments a verb may have. The higher the average the more nominal information can be glued to verbs.

In order to get an indication of how reliable this interpretation is, we correlate all the Dim3– scores that characterize 442 SUC texts to the LIX scores calculated for these texts. The results are shown in Table 6, row 5.

There exists a weak statistically significant negative correlation between Dim3– and the matching LIX scores. The mean of the LIX scores matching Dim3– is 42.55 with a standard error of 0.36. A mean of 42.55 indicates that the readability level is slightly *difficult*, but less difficult than Dim2+ (see Table 2).

SUC registers that are characterized by this facet and the corresponding level of readability are a_reportage_genre (143 texts) and review (114) which are highly characterized by this facet and readability level, while b_editorial_genre is the least characterized by them (Figure 13).

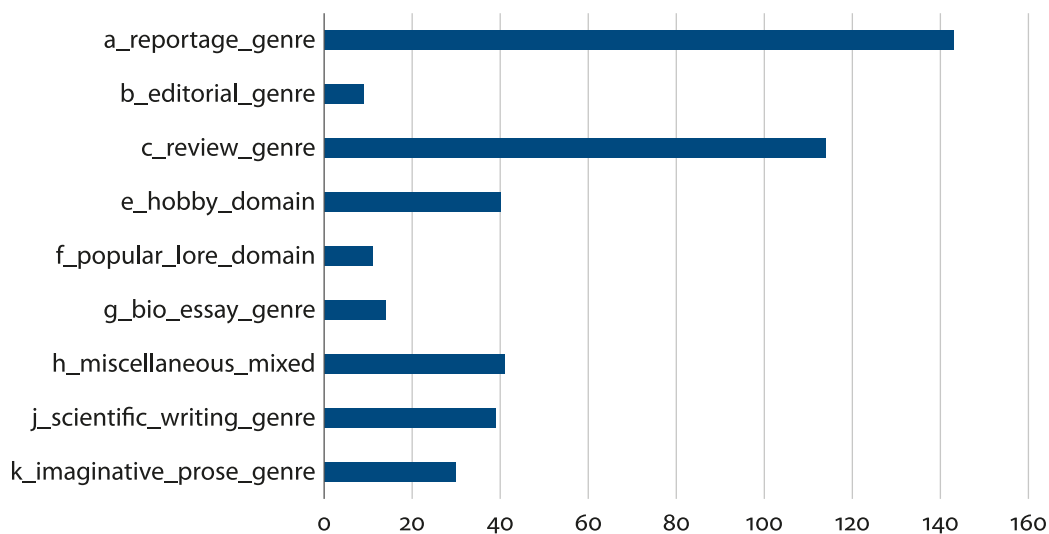


Figure 13. Distribution of SUC registers along Dim3–

Linguistically, the facet elicited by this dimension has an *appositional* nature and is geared towards the *expansion of nominal information* as functional focus.

As shown in Box 5, texts connotated by this facet contain information expansions through the use of genitives and dates.

Box 5. An excerpt of a SUC text connotated by the appositional facet to expand nominals

Dim3– score	LIX score	Filename	SUC Register
–3.34	54.24	jao5.xml	j_scientific_writing_genre
Swedish		English translation	
Om kungamaktens tillbakagång under perioden 1906–1918 se Axel Brusewitz’ klassiska Kungamakt, herremakt, folkmakt (1951).		On the decline of the king’s power during the period 1906–1918, see Axel Brusewitz’s seminal book “Kungamakt, herremakt, folkmakt” (1951).	

5.2.9 Summary

Table 6 summarizes the findings presented in this section. We divided the factors into signed textual dimension. Dim2– was not included because of insufficient loadings. We ended up with five textual dimensions, each of which was interpreted linguistically and functionally. We then proposed five text complexity facets.

The complexity facets elicited in this analysis are composite, since they are characterized by linguistic traits that are interpreted functionally in a communicative situation and associated with readability levels. These findings show that the Adjectival (Information Elaboration) facet is frequent in difficult-to-read texts, while the Pronominal-Adverbial (Spoken-Emotional) facet characterizes easier-to-read texts. The other facets – namely Nominal (Informational), Appositional (Information Expansion) and Verbal (Engaged) – are placed between these two extremes. This profiling is a coarse picture of the SUC as a whole, where each register is viewed with respect to individual facets and readability levels. In the next section, we will provide a more comprehensive profiling of individual SUC registers.

6. Profiling SUC registers

Since the dimension scores and LIX scores are spread on different intervals, we normalized them all on 0–100 scale in order to have a more accurate picture of how the text complexity facets and readability levels vary across SUC registers. Table 7 shows the normalized scores.

In Figure 14, SUC registers are profiled with normalized values. These pictorial profiles are neat and provide interesting insights. For instance, we can observe

Table 6. Summary table: LIX scores & dimension scores across all SUC categories

	# texts	LIX Mean	Median	SD	SE	Kendall's coff.	Sig.
LIX scores matching Dim1+	545	39.57	39.65	8.32	0.36	−32	$p < 2.2e-16$
LIX scores matching Dim1−	495	45.19	44.52	6.45	0.29	−14	$p = 1.556e-06$
LIX scores matching Dim2+	461	46.38	46.36	6.72	0.31	+0.41	$p < 2.2e-16$
LIX scores matching Dim3+	599	42.02	42.08	8.28	0.34	+0.069	$p = 0.012$
LIX scores matching Dim3−	441	42.55	42.54	7.6	0.36	−0.12	$p = 0.0002263$

that the readability level is rather uniform across the registers, with the exception of popular lore, which appears to be easier to read than other registers. We can also observe that the nominal (informational) facet is often strong when also the appositional facet is pronounced.

In order to get a better grip of the differences and similarities across the registers, we plot each register as a radar chart (see Figures 15–19). A radar chart is a type of 2D chart presenting multivariate data where each variable is given an axis and the data are plotted as a polygonal shape over all axes. Each variable is provided with an axis that starts from the centre. All axes have the same scale and are arranged radially, with equal distances between each other. Grid lines that connect from axis-to-axis are often used as a guide (Jelen 2013).

We can then observe that the faceted makeup of reviews, scientific writing (Figure 15) and reportage (Figure 16) are very similar. For these three registers, readability is rather difficult, with a strong informational facet associated with a pronounced appositional facet. The pronominal-adverbial facet is very flat, and the verbal and adjectival facets are weak.

Unsurprisingly, the nominal profile of reportage, review and academic writing is associated with more difficult readability levels, while the bio-essay and imaginative prose (Figure 17) are easier to read. These two registers are characterized by strong pronominal-adverbial, adjectival and appositional facets. We could interpret these traits as an emphasis on the expansion and elaboration of events or sentiments provided in a spoken or emotional communication setting, rather than on the objective report of factual information. Intuitively, nominal information tends to be dense, a trait that is typical of reporting and academic writing which are both based on “evidence”, while emotions and verbal interactions are

Table 7. Summary table across SUC registers: Text complexity facets and readability level

SUC Registers	Number of texts per SUC register	Mean of normalized LIX scores	Mean of normalized Dim1 + scores	Mean of normalized Dim1 – scores		Mean of normalized Dim2+ scores	Mean of normalized Dim3+ scores		Mean of normalized Dim3– scores
				Nominal (Informational) Facet	Adjectival (Information Elaboration) Facet		Verbal (Engaged) Facet	Appositional (Information Expansion) Facet	
a_reportage_genre	269	53.82	27.62	69.47	28.25	26.61	85.25		
b_editorial_genre	70	57.56	36.56	66.76	19.82	54.94	62.30		
c_review_genre	127	52.91	32.11	68.71	31.07	32.24	79.29		
e_hobby_domain	124	54.25	23.09	72.00	22.58	37.28	83.34		
f_popular_lore_domain	62	38.72	46.54	76.06	27.81	45.38	61.24		
g_bio_essay_genre	27	44.99	49.44	0	35.52	33.17	60.35		
h_miscellaneous_mixed	145	47.58	19.14	57.56	24.07	30.44	66.00		
j_scientific_writing_genre	86	53.16	23.12	57.72	27.6	37.25	80.25		
k_imaginative_prose_genre	130	50.50	52.55	0	33.58	35.21	71.2		
Total	1040								

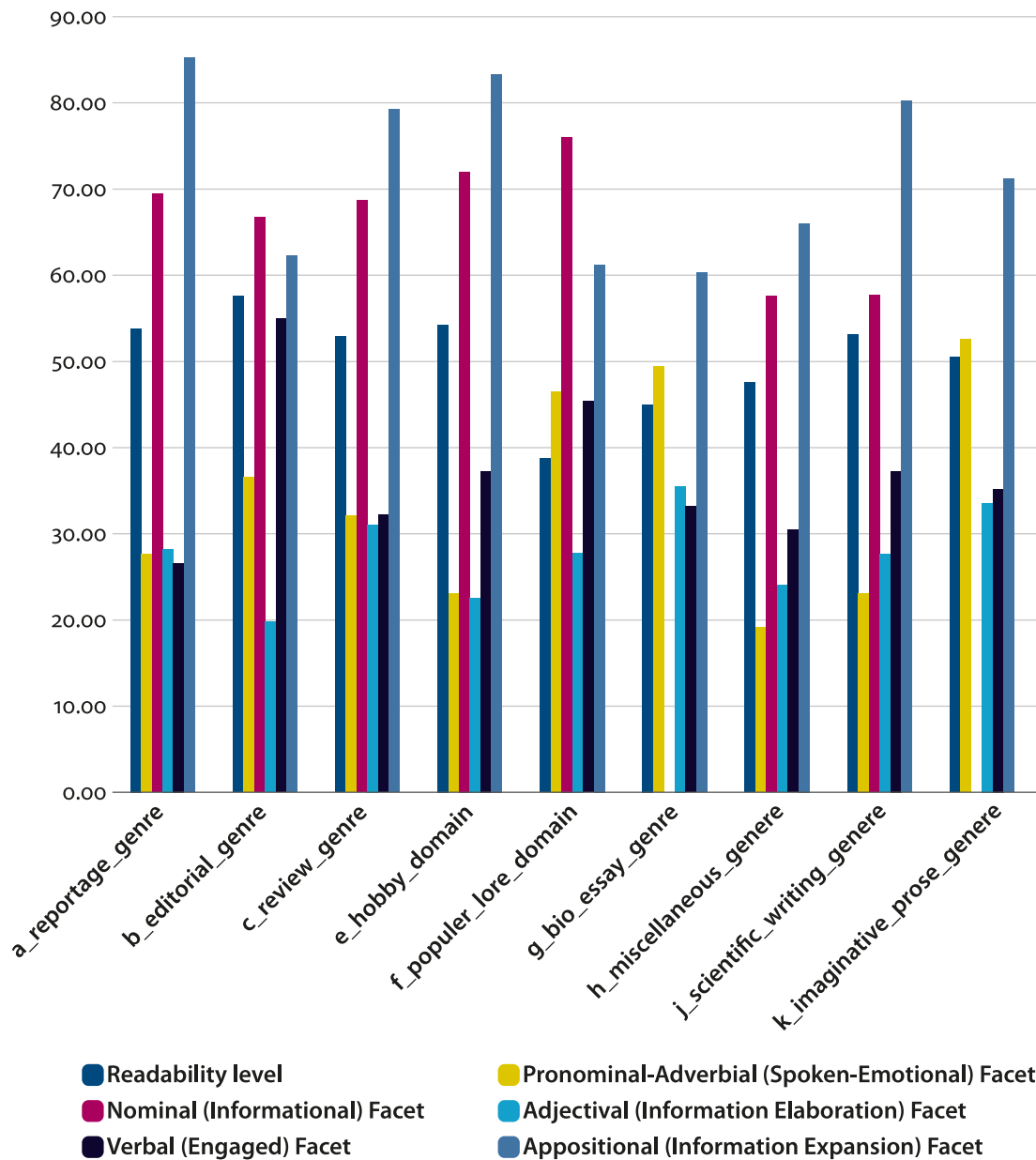
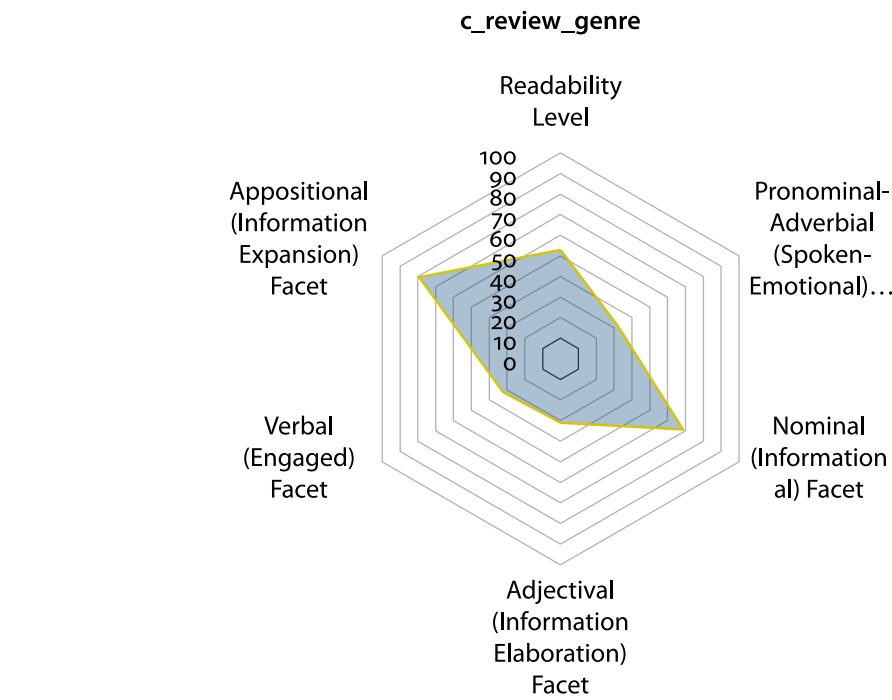


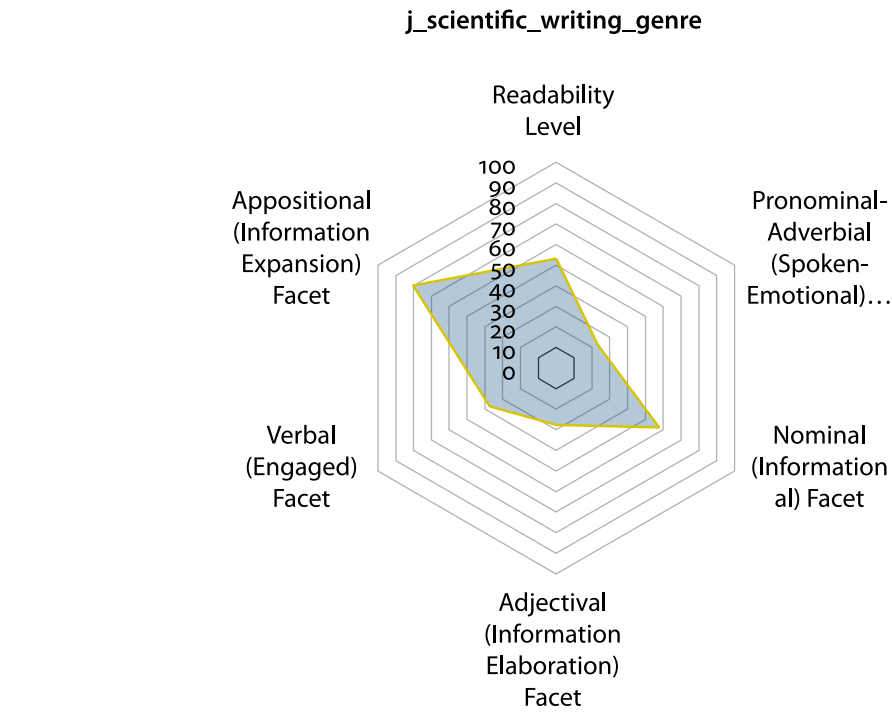
Figure 14. Summary chart of all the facets and readability level across SUC registers (normalized scores)

more associated with a point of view, which is typical of the narration unfurled in biographies and imaginative prose.

The hobby and miscellaneous registers (Figure 18) show an axis nominal-appositional (a similarity with the reportage, review and scientific writing registers) but they are also characterized by some prominence of the verbal facets, while the pronominal-adverbial facet and the adjectival facet are rather flat. The emphasis of these registers seems to be on the nominal-factual (or possibly domain-specific) information accompanied by articulated syntax, rather than on-the-fly quick communication.



(a)



(b)

Figure 15. Radar chart profiling: Review and scientific writing

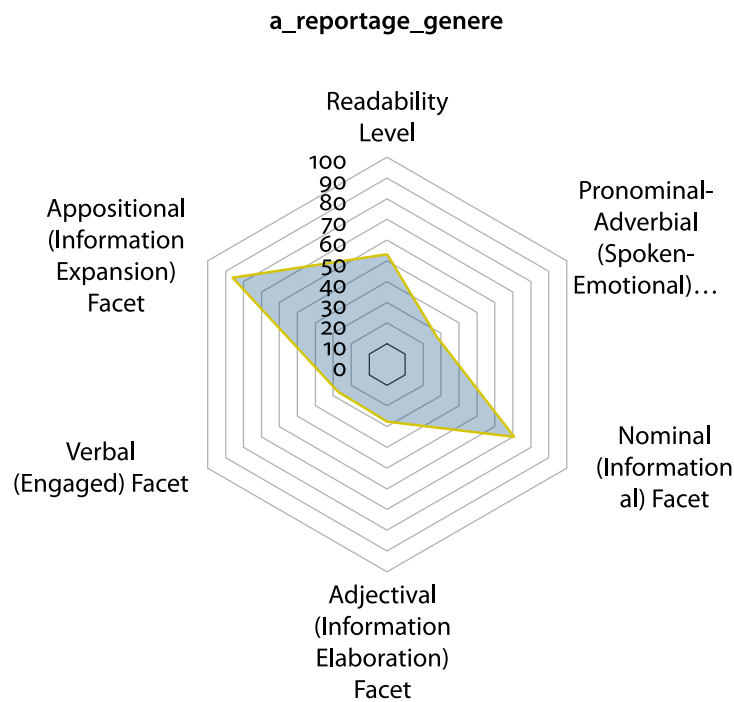


Figure 16. Radar chart profiling: Reportage

The editorial and popular lore registers are two singletons (Figure 19) in that they have a shape that is not similar to other register in the SUC. Editorials have a strong nominal facet, but a quite weak appositional facet. The texts in this register are difficult to read and they show a pronounced verbal facet that implies more engaged syntactic articulation. The adjectival facet is weak, so is the pronominal-adverbial facet. This profile seems to be compliant with the argumentative nature of editorials, which normally report both facts and claims supported by the articulated syntax of an argumentation.

The popular lore register is the easiest-to-read register in the SUC. It has a pronounced nominal aspect, flat adjectival facets, and moderate verbal, appositional, and pronominal-adverbial facets.

7. Discussion

Using MDA and 45 text complexity features, we elicited text complexity facets from the SUC corpus. We explored the interpretability of a 3-factor solution and validated the factors by correlating factors scores and readability scores. In this phase, two out of three factors were statistically significant. This means that at least two of the three factors are statistically grounded. Since factors are complex constructs, we split them into signed (\pm) textual dimensions and interpreted each of the signed dimensions in term of text complexity facets. We elicited five facets

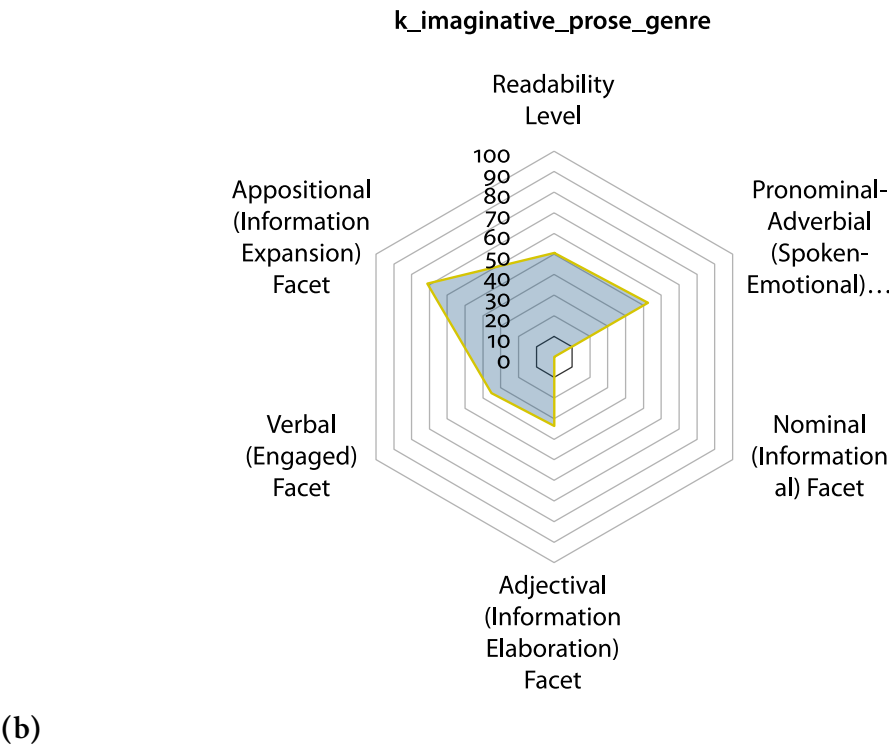
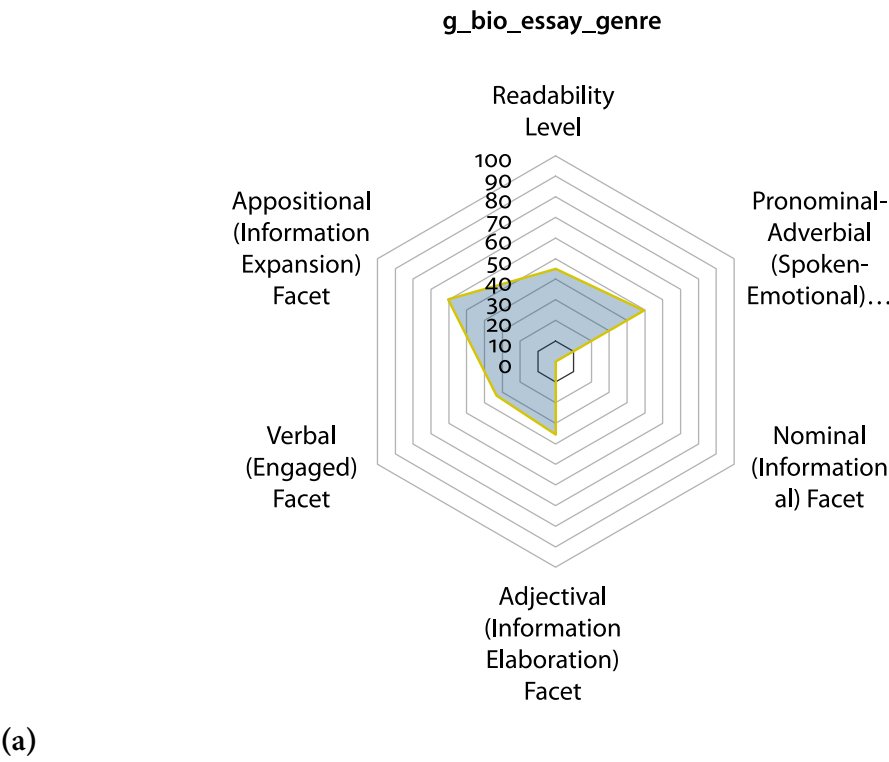


Figure 17. Radar chart profiling: Bio_Essay and imaginative prose

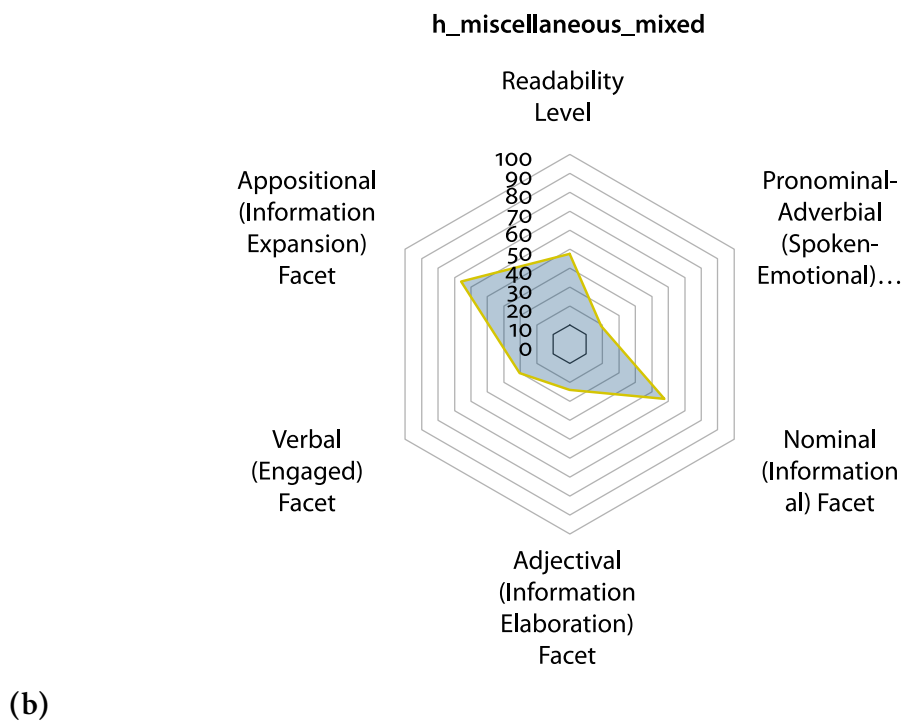
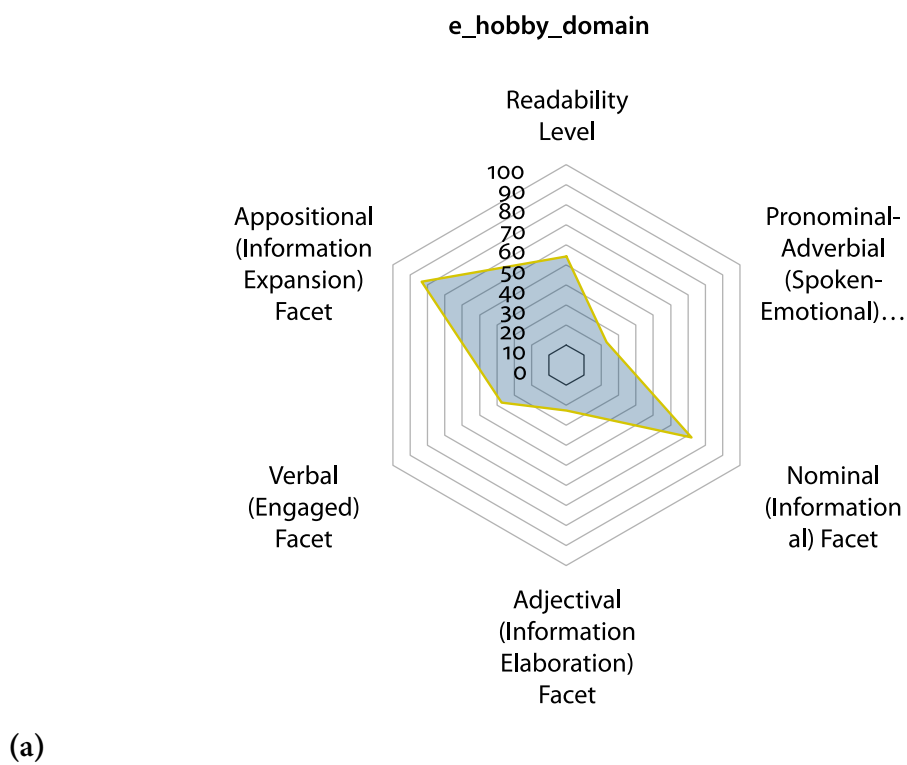


Figure 18. Radar chart profiling: Hobby and miscellaneous

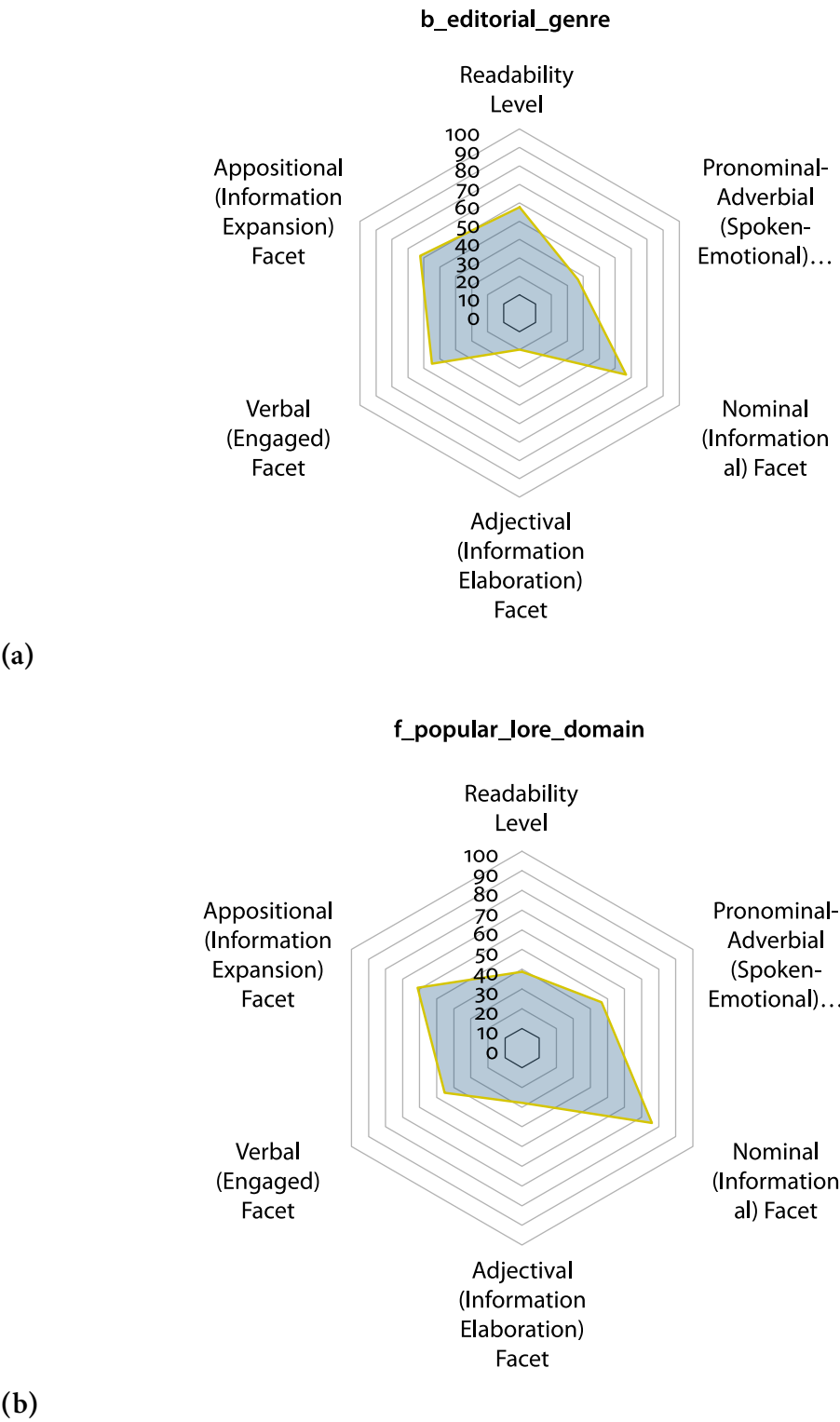


Figure 19. Radar chart profiling: Editorial and popular lore

that have both a linguistic and functional characterization, namely: Pronominal-Adverbial (Spoken-Emotional), Nominal (Informational), Adjectival (Information Elaboration), Verbal (Engaged) and Appositional (Information Expansion). We profiled the SUC registers using the text complexity facets in combination with LIX scores.

First, we analysed the whole SUC corpus per factor. Figure 8 shows how SUC registers are characterized by the three factors: some registers have only negative loadings (e.g. reportage), some only positive loadings (e.g. editorial), and some both negative and positive (e.g. hobby). This characterization is certainly indicative, but a higher degree on insight is possible. We observed that the texts of a register do not belong necessarily to either the positive or negative side of a factor. For instance, the editorial register has the majority of the scores on Dim3+, but at the same time, some editorials have scores on Dim3-.

Then, to have a better view on this register-internal variation, we split up positive and negative sides of the factors and interpreted them independently as signed textual dimensions. We ended up with five textual dimensions that we interpreted in terms of linguistic and functional facets. Essentially, each of the five signed textual dimension is a facet. We correlated each dimension with readability scores. Results show that all the signed dimensions are statistically significant when correlated to readability scores (see Table 6). This is encouraging because it means that the factor solution is grounded.

Finally, we profiled each register with the five facets and with readability levels. Table 7 and Figures 15–19 show the similarities and dissimilarities across SUC registers with respect to facets and readability. The similarity between bio-essay and imaginative writing is striking and also quite intuitive if we think of the shared narration techniques that are normally used in these two registers. Similarly, the commonalities between reportage, review and academic writing is also unsurprising given the factual nature of these registers. Editorials and popular lore stick out for their dissimilarity with the other registers.

But what does a text complexity facet tell us about? A facet breaks down the linguistic and functional nature of text complexity. It is the combination of text complexity facets, and not the single facet, that gives us an indication of how composite and complex the texts in a register are.

It is worth noting that the ultimate goal of this study is to show the potential of the approach rather than presenting full-fledged results because of the caveats related to the corpus and linguistic annotation. First, factor analysis has many statistical prerequisites that should be taken into account when drawing final conclusions. A corpus like the SUC, based on the Brown corpus, whose design dates back to the 1960's, is not the most adequate sample for statistical approaches, which normally impose constraints on different aspects of the sampled population. Here

we draw tentative conclusions since this is a preliminary investigation and since it is the approach rather than the results that we are presenting. Second, the text complexity features used in this study are too coarse-grained to fully benefit from the MDA approach. For example, the annotation of verbs in the SUC dataset does not include information about tenses, which are an important feature to detect rhetorical devices such as narration, often characterized by a higher degree of text complexity. The same is true for passivation and other fine-grained syntactic constructs that are not included in the SUC's tagset.

Having said that, we argue that text complexity analysis with MDA can provide useful insights also on corpora like the SUC, that are not explicitly designed for Biber's multidimensional approach.

8. Conclusion and future work

In this article, we have presented the results of a corpus-based study where we explored whether it is possible to automatically single out different facets of text complexity in a general-purpose corpus. To this end, we used MDA. We evaluated the results by correlating the factors returned by the analysis with a readability index to ascertain whether the selected factor solution matches an independent measurement of readability, a notion that goes hand in hand with text complexity. Results show that it is indeed possible to elicit and interpret facets of text complexity using MDA.

The rationale of the study is to understand the nature of text complexity across SUC registers. Findings are informative because MDA helps pin down five facets that provide a plausible grammatical and functional interpretations of text complexity across the SUC.

The overall conclusion that we draw from the findings of this preliminary investigation is that text complexity facets return a multifarious profile of registers that complement readability scores. Facets highlight combinations of different linguistic and functional aspects that help us understand the complexity and the makeup of registers. Here, we offer preliminary results based on a small corpus, tagged with coarse morphological and syntactic tagset. We plan future studies on larger corpora tagged with more fine-grained text complexity features to show the full potential of the approach.

Funding

This research was supported by E-care@home, a “SIDUS – Strong Distributed Research Environment” project, funded by the Swedish Knowledge Foundation and RISE, Research Institutes of Sweden AB.

Companion website

The study described in this paper is fully reproducible. Datasets, radar charts and R code are available here: <<http://santini.se/registerstudies2020>>.

References

- Adesam, Y., Bouma, G. and Johansson, R. (2018). The Koala part-of-speech and morphological tagset for Swedish. *SLTC*.
- Asención-Delaney, Y., & Collentine, J. (2011). A multidimensional analysis of a written L2 Spanish corpus. *Applied linguistics*, 32(3), 299–322.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511621024>
- Biber, D. (1989). A typology of English texts. *Linguistics*, 27(1), 3–44. <https://doi.org/10.1515/ling.1989.27.1.3>
- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511519871>
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.
- Biber, D., & Kurjian, J. (2007). Towards a taxonomy of web registers and text types: A multi-dimensional analysis. In *Corpus Linguistics and the Web* (pp. 109–131). https://doi.org/10.1163/9789401203791_008
- Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511814358>
- Biber, D., & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, 44(2), 95–137. <https://doi.org/10.1177/0075424216628955>
- Björnsson, C.H. (1968). Läsbarhet. Liber.
- Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10
- Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2), 97–135. <https://doi.org/10.1075/itl.165.2.01col>
- Common Core State Standards Initiative. (2010). Common Core State Standards for English Language Arts & Literacy InHistory/Social Studies, Science, and Technical Subjects. Appendix A: Research Supporting Key Elements of the Standards, Glossary of Key Terms.

- Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., Zasina, A. J., & Benko, V. (2020). Comparing web-crawled and traditional corpora. *Language Resources and Evaluation*, 1–33.
- Dahl, Ö. (2004). The growth and maintenance of linguistic complexity (Vol. 71). John Benjamins Publishing. <https://doi.org/10.1075/slcs.71>
- Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, 26(1), 19–26.
- Dell’Orletta, F., Montemagni, S., & Venturi, G. (2013), September). Linguistic profiling of texts across textual genres and readability levels. An exploratory study on Italian fictional prose. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013* (pp. 189–197).
- Dell’Orletta, F., Montemagni, S., & Venturi, G. (2014). Assessing document and sentence readability in less resourced languages and across textual genres. *ITL-International Journal of Applied Linguistics*, 165(2), 163–193. <https://doi.org/10.1075/itl.165.2.03del>
- DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14(20), 1–11.
- Fahlborg, D., & Rennes, E. (2016). Introducing SAPIs—an API service for text analysis and simplification. In *the second national Swe-Clarin workshop: Research collaborations for the digital age*, Umeå, Sweden.
- Falkenjack, J. (2018). Towards a model of general text complexity for Swedish (Doctoral dissertation, Linköping University Electronic Press).
- Falkenjack, J., Mühlenbock, K. H., & Jönsson, A. (2013), May). Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)* (pp. 27–40).
- Falkenjack, J., Santini, M., & Jönsson, A. (2016). An exploratory study on genre classification using readability features. In *Proceedings of the Sixth Swedish Language Technology Conference (SLTC 2016)*, Umeå, Sweden.
- Feng, L. (2010). Automatic readability assessment (Doctoral dissertation, CUNY Academic Works).
- Field, A. (2000). *Discovering statistics using SPSS for Windows*. Londra: Sage Publication.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–23. <https://doi.org/10.1037/h0057532>
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications.
- Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, 7(2), 191–205. <https://doi.org/10.1177/1094428104263675>
- Hiebert, E. H. (2012). Readability and the common core’s staircase of text complexity. *Text Matters*, 1.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185. <https://doi.org/10.1007/BF02289447>
- Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research*, 35(1), 3–21. <https://doi.org/10.1177/0267658318809765>
- Jelen, B. (2013). Excel 2013 charts and graphs. Que Publishing Company.
- Jönsson, S., Rennes, E., Falkenjack, J., & Jönsson, A. (2018). A component based approach to measuring text complexity. In *Proceedings of The Seventh Swedish Language Technology Conference 2018 (SLTC-18)*.

- Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., & Welty, C. (2010), August). Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 546–554). Association for Computational Linguistics.
- Källgren, G., Gustafson-Capková, S., & Hartmann, B. (2006). Manual of the Stockholm Umeå Corpus version 2.0. Department of Linguistics, Stockholm University, December. Sofia Gustafson-Capková and Britt Hartmann (eds.).
- Ledesma, R. D., Valero-Mora, P., & Macbeth, G. (2015). The scree test and the number of factors: a dynamic graphics approach. *The Spanish journal of psychology*, 18. <https://doi.org/10.1017/sjp.2015.13>
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496. <https://doi.org/10.1075/ijcl.15.4.02lu>
- Mühlenbock, K. H. (2013). I see what you mean: Assessing readability for specific target groups. (Doctoral dissertation, University of Gothenburg, Gothenburg, Sweden).
- Napolitano, D., Sheehan, K. M., & Mundkowsky, R. (2015), June). Online readability and text complexity analysis with Text Evaluator. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 96–100).
- Nenkova, A., Chae, J., Louis, A., & Pitler, E. (2010). Structural features for predicting the linguistic quality of text. In *Empirical methods in natural language generation* (pp. 222–241). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15573-4_12
- Nivre, J. (2006). *Inductive dependency parsing* (pp. 87–120). Springer Netherlands. https://doi.org/10.1007/1-4020-4889-0_4
- Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, 31(1), 117–134. <https://doi.org/10.1177/0267658314536435>
- Petersen, S. (2007). Natural language processing tools for reading level assessment and text simplification for bilingual education. (Doctoral dissertation, University of Washington, Seattle, WA, USA).
- Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1), 89–106. <https://doi.org/10.1016/j.csl.2008.04.003>
- Pilán, I., Vajjala, S., & Volodina, E. (2016). A readable read: Automatic assessment of language learning materials based on linguistic complexity. *arXiv preprint arXiv:1603.08868*.
- Pitler, E., & Nenkova, A. (2008), October). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 186–195).
- Rello, L., Baeza-Yates, R., Bott, S., & Saggion, H. (2013a). Simplify or help? Text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (pp. 1–10).
- Rello, L., Baeza-Yates, R., Dempere-Marco, L., and Saggion, H. (2013b). Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction* (pp. 203–219. Springer.
- Saggion, H. (2017). Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1), 1–137. <https://doi.org/10.2200/Soo700ED1Vo1Y201602HLT032>

- Santini, M., Danielsson, B., & Jönsson, A. (2019), August). Introducing the Notion of ‘Contrast’ Features for Language Technology. In *International Conference on Database and Expert Systems Applications* (pp. 189–198). Springer, Cham.
https://doi.org/10.1007/978-3-030-27684-3_24
- Sardinha, T. B., Kauffmann, C., & Acunzo, C. M. (2014). A multi-dimensional analysis of register variation in Brazilian Portuguese. *Corpora*, 9(2), 239–271.
<https://doi.org/10.3366/cor.2014.0059>
- Sardinha, T. B., & Pinto, M. V. (Eds.). (2014). *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber* (Vol. 60). John Benjamins Publishing Company.
<https://doi.org/10.1075/scl.60>
- Štajner, S., & Saggion, H. (2018), August). Data-Driven Text Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts* (pp. 19–23).
- Vega, B., Feng, S., Lehman, B., Graesser, A., & D’Mello, S. (2013), July). Reading into the text: Investigating the influence of text complexity on cognitive engagement. In *Educational Data Mining 2013*.
- Wray, D., & Janan, D. (2013). Readability revisited? The implications of text complexity
Published in *The Curriculum Journal*, 2013. <https://doi.org/10.1080/09585176.2013.828631>

Appendix. Linguistic Features used in the Study

Text Complexity Features for the Swedish Language (slightly adapted from Falkenjack et al. 2013). We use the following 45 features.

3 lexical features. Namely, ratioSweVocC, ratioSweVocD, ratioSweVocH

Lexical features are based on categorical word frequencies. The word frequencies are extracted after lemmatization and are calculated using the basic Swedish vocabulary SweVoc (Mühlenbock, 2013). SweVoc is comparable to the list used in the classic Dale-Chall formula (Dale and Chall, 1949) for English and developed for similar purposes, however special sub-categories have been added (of which three are specifically considered). The following frequencies are calculated, based on different categories in SweVoc:

1. **SweVocC.** SweVoc lemmas fundamental for communication (category C).
2. **SweVocD.** SweVoc lemmas for everyday use (category D).
3. **SweVocH.** SweVoc other highly frequent lemmas (category H).

A high ratio of SweVoc words should indicate a more easy-to-read text. The Dale-Chall metric (Chall and Dale, 1995) has been used as a similar feature in a number of machine learning based studies of text readability for English (Feng, 2010; Pitler and Nenkova, 2008).

20 Morpho-syntactic features. Namely, pos_JJ (adjective), pos_DT (determiner), pos_HS (whPossessive), pos_HP (whPronoun), pos_RO (ordinalNum), pos_NN (noun), pos_VB (verb), pos_IE (infinitavalMarker), pos_HD (whDeterminer), pos_IN (interjection), pos_UO (foreignWord), pos_KN (coordinatingConj), pos_HA (whAdverb), pos_SN (subordinating-Conj), pos_PM (properNoun), pos_PN (pronoun), pos_AB (adverb), pos_PP (preposition), pos_PS (possessivePronoun) and pos_PC (participle).

Unigram probabilities for 20 different parts-of-speech in the document, that is, the ratio of each part-of-speech, on a per token basis, as individual attributes. Such a unigram language

model based on part-of-speech, and similar metrics, has shown to be a relevant feature for readability assessment for English (Heilman et al., 2007; Petersen, 2007).

18 Syntactic features. Namely, dep_AN (apposition), dep_AT (premodifier), dep_CA (contrastiveAdverbial), dep_EF (relativeClauseCleft), dep_I? (questionMark), dep_IK (comma), dep_IP (period), dep_IQ (colon), dep_IS (semicolon), dep_IU (exclamationMark), dep_KA (comparativeAdverbial), dep_MA (attitudeAdverbial), dep_NA (negationAdverbial), dep_PT (predicativeAttribute), dep_RA (placeAdverbial), dep_TA (timeAdverbial), dep_XA (sotospeak), dep_XT (socalled).

The presence of syntactic features is the most evident proof of textual complexity. The more syntactically complex a text is, the more difficult to read. These features are estimable after syntactic parsing of the text. The syntactic feature set is extracted after dependency parsing using the Maltparser (Nivre et al., 2006). Unigram probabilities for the 16 dependency types resulting from the dependency parsing, on a per token basis. These features are comparable to the part-of-speech unigram probabilities and to the phrase type rate based on phrase grammar parsing used in earlier research (Nenkova et al., 2010).

4 ratio features. avgSentenceDepth, avgVerbalArity, avgNominalPremodifiers, avgNominalPostmodifiers.


avgSentenceDepth. The average sentence depth. Sentences with deeper dependency trees could be indicative of a more complex text in the same way as phrase grammar trees has been shown to be (Petersen and Ostendorf, 2009).

avgVerbalArity. Arity indicates number of arguments of a verb. The average arity of verbs in the document, calculated as the average number of dependents per verb (Dell’Orletta et al., 2011).

avgNominalPremodifiers. The average number of nominal pre-modifiers per sentence.

avgNominalPostmodifiers. The average number of nominal post-modifiers per sentence.

Address for correspondence

Marina Santini
RISE Research Institutes of Sweden AB
Varvsgatan 25
Stockholm 11729
Sweden
marinasantini.ms@gmail.com
 <https://orcid.org/0000-0002-5737-8149>

Co-author information

Arne Jönsson
Linköping University
arne.jonsson@liu.se
 <https://orcid.org/0000-0001-9852-5531>

Publication history

Date received: 26 January 2019

Date accepted: 2 June 2020

Published online: 13 August 2020