Visualizing Facets of Text Complexity across Registers

Marina Santini*, **Arne Jönsson**†, **Evelina Rennes**†
*RISE, Research Institutes of Sweden
Linköping, Sweden
marina.santini@ri.se
†Department of Computer and Information Science
Linköping University, Linköping, Sweden
{arne.jonsson, evelina.rennes}@liu.se

Abstract

In this paper, we propose visualizing results of a corpus-based study on text complexity using radar charts. We argue that the added value of this type of visualisation is the polygonal shape that provides an intuitive grasp of text complexity similarities across the registers of a corpus. The results that we visualize come from a study where we explored whether it is possible to automatically single out different facets of text complexity across the registers of a Swedish corpus. To this end, we used factor analysis as applied in Biber's Multi-Dimensional Analysis framework. The visualization of text complexity facets with radar charts indicates that there is correspondence between linguistic similarity and similarity of shape across registers.

Keywords: radar charts, text complexity, readability, Multi-Dimensional Analysis

1. Introduction

Data visualization refers to the graphical representation of information, data, results or findings. Graphical representations like charts, graphs, and maps, help the human brain understand and interpret trends and patterns in data. Effective data visualizations place meaning into complex information because they help disentangle complexities and unveil underlying patterns in a clear and concise way. The easiest and most common way to create a data visualization is to use bar graphs, pie charts or line graphs. These types of charts are effective and widely used. Recently, more sophisticated visualizations have been introduced, such as bullet graphs, heat maps, radial trees, radar charts or infographics.

It goes without saying that the effectiveness of the visualization depends on the purpose and on the type of data. In this paper, we ponder about the best way to "shape" the results of a corpus-based study on text complexity in order to show how different registers differ according to a number of text complexity features. The insights provided by this study may be useful to understand how to visually represent a complex notion like text complexity.

Text complexity is an important dimension of textual variation. It is crucial to pin it down because texts can be customised to different types of audiences, according to cognitive requirements (e.g. texts for the dyslectic), 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. popularised texts). Text complexity can be analysed in several ways. The approach we used is based on factor analysis as applied in Biber's Multi-Dimensional Analysis framework (Biber, 1988) (henceforth MDA). The corpus used in our analysis was the Swedish national corpus, called Stockholm-Umeå Corpus or SUC. Results are described in detail in Santini and Jönsson (2020), and indicate that it is indeed possible to elicit and interpret facets of text complexity using MDA, regardless some caveats due to the small size of the corpus.

When we tabulated the results (see Table 1) and plotted them in a bar chart (see Figure 1), we observed that tabulation and a bar chart were useful for the identification of the text complexity similarities and dissimilarities across the registers, but their interpretation required some effort and time even for linguists. At this point we were intrigued by the following question: how can we visually shape the different facets of text complexity generated by the study in an efficient and intuitive way? In this paper, we focus on this research question and we argue that the type of visualization that seems to be the most appropriate for this type of results is the radar chart because it plots a polygonal "shape" that helps emphasise similarities and dissimilarities across categories.

2. Previous Work

To our knowledge, radar charts have never been used to visualize text complexity across registers. Since there is no previous work that explores this topic, we divide this section into two separate parts, the first one focusing on text complexity, and the second one listing linguistic studies that relied on radar charts visualization.

2.1. Text Complexity

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 organised, 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 synonymous with "difficulty". Rather, in his interpretation complexity is "an objective property of a system", i.e. "a measure of the amount of information needed to describe or reconstruct it". In his view, "[g]rammatical 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".

Another linguistic field where there is a persistent interest in the study of language complexity is second language (L2) research. For instance, Pallotti (2015) notes that the notion of linguistic complexity is still poorly defined and often used with different meanings. He proposes a simple, coherent view of the construct, which is defined in a purely structural way, i.e. the complexity directly arising 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 and describe five empirical studies that investigate under-explored forms of complexity from a cross-linguistic perspective or that propose 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-parts 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. Its benefits and drawbacks have been analysed by Fang (2016). Many other models of text complexity have been proposed for educational purposes, but 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 also of researchers working in computational and language technology (LT). In LT, text complexity is tightly linked to corpus-based and data-driven analysis of textual difficulty, e.g. in second language acquisition (Lu, 2010) and to the development of LT applications, such as automatic readability assessment (Feng, 2010) or the automatic text simplification for those who have dyslexia (Rello et al., 2013a). Text complexity can also be seen as a sub-field of Text Simplification, which is currently a well-developed LT research area (Saggion, 2017).

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 and Chall, 1949) or dyslectic people (Rello et al., 2013b), while text complexity seems to pertain to the text itself, or the text and a generalised group of readers. Readability indices are practical and robust but coarse since they cannot provide 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 date to gauge text complexity and readability infallibly. Therefore, complexity and readability scores are useful, although they must be taken with a grain of salt.

2.2. Radar Charts

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 axis starts from the centre. All axes are arranged radially, with equal distances between each other and with the same scale. Grid lines that connect from axis-to-axis are often used as a guide (Jelen, 2013). Radar charts have already been used to display linguistic data, but not text complexity across registers. For instance, Branco et al. (2014) used a radar chart for their tool that "supports human experts in their task of classifying text excerpts suitable to be used in quizzes for learning materials and as items of exams that are aimed at assessing and certifying the language level of students taking courses of Portuguese as a second language". In their tool, the arms of the radar chart are the reference scales obtained from 125 texts. When a new text is fed into the tool, its values are mapped into the radar chart to visualize its linguistic profile. Egbert and Biber (2018) plotted six radar charts to profile linguistic variation across registers. Each register has five pairs of textual dimensions. One member of the pair has been obtained with MDA, the other one with Canonical Discriminant Analysis (CDA). The purpose was to show the extent of the overlap between the two statistical methods when analysing linguistic data. Jönsson et al. (2018) used a radar chart to display text complexity analysed with Principal Component Analysis (PCA). Their radar chart displays the principal components (not registers) and how text complexity varies across them.

3. MDA and Text Complexity

In this section, we summarise the main findings of our study on text complexity variation in the SUC. Full details can be found in Santini and Jönsson (2020). Below, we briefly describe the SUC corpus and dataset, and present MDA, together with the 3-factor solution used in the study.

3.1. SUC Corpus and Dataset

The SUC (Gustafson-Capková and Hartmann, 2006) is a collection of Swedish texts and represents the Swedish language as used by native Swedish adult speakers in the 90s. The SUC includes a wide variety of texts written for several types of audiences, from academics, to newspapers' readers, to fictions' readers and contains subject-based text varieties (e.g. Hobby), press genres (e.g. Editorials), and mixed categories (e.g. Miscellaneous). We call them collectively "registers", as defined in Biber and Conrad (2009). Given the composition of the SUC, we assume the presence of different levels of text complexity across SUC registers. This 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 specialised professionals, such as lawyers and physicians, must be able to understand very complex texts in order to practise their professions. In between easy texts for children and the domain-specific jargon used by specialised professionals, there exist texts that present different levels of textual difficulty.
From the SUC, a text complexity dataset has been extracted via SAPIS (Fahlborg and Rennes, 2016), an API Service for

SUC Registers	Number of texts per SUC register	Mean of normalised LIX scores	Mean of normalised Dim1+ scores	Mean of normalised Dim1- scores	Mean of normalised Dim2+ scores	Mean of normalised Dim3+ scores	Mean of normalised Dim3- scores
		Readabilty level	Pronominal-Adverbial (Spoken-Emotional) Facet	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	27.60	37.25	80.25
k_imaginative_prose_genre	130	50.50	52.55	0	33.58	35.21	71.20
Total	**1040**						

Table 1: Summary table of all the facets and readability level across the SUC registers.

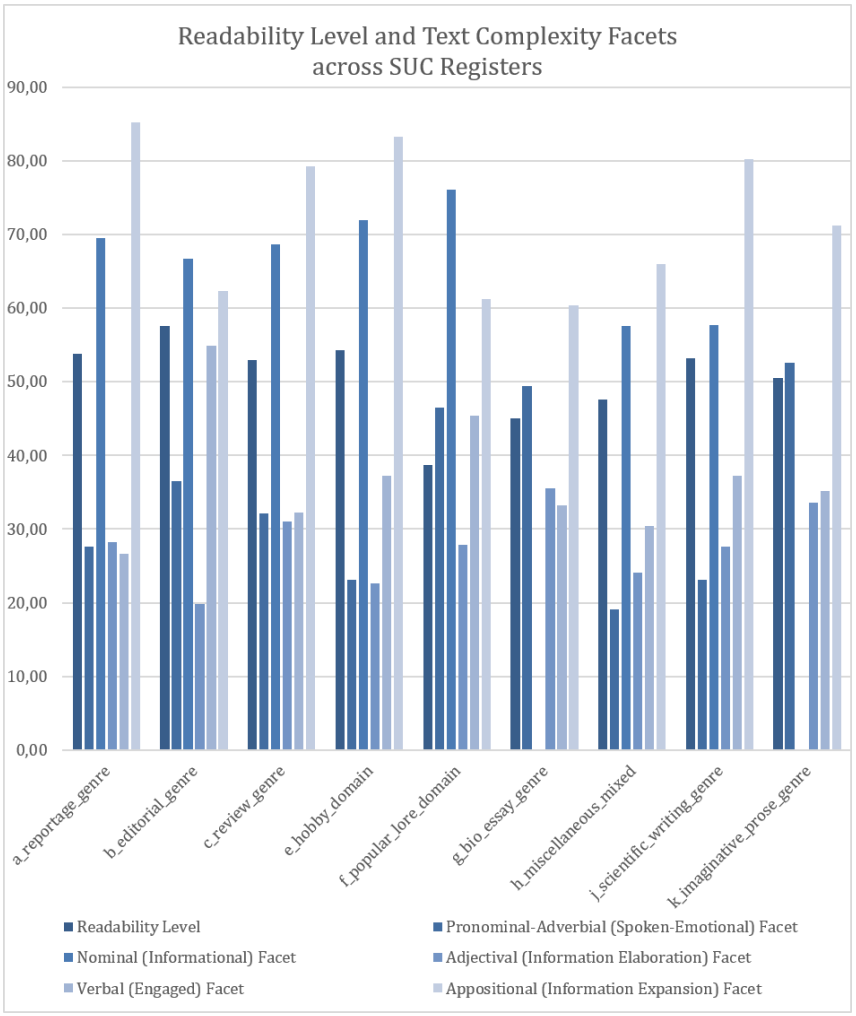

Figure 1: Summary chart of all the facets and readability levels across SUC registers.

Text Analysis and Simplification of Swedish text. The SUC dataset returned by SAPIS contains 120 linguistic features described in Falkenjack et al. (2013). This dataset is the source dataset used in the study.

3.2. MDA

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 multivariate statistical method) to uncover patterns of linguistic variation across the registers collected in a corpus. The basic idea of MDA builds on the notion of "co-occurring linguistic features that have a functional underpinning" (Biber, 1988, p. 121). The co-occurrence of linguistic features across registers into factors is interpreted in terms of underlying textual dimensions.

There are three main steps in MDA, variable screening, running MDA proper, and the interpretation of the factors.

3.2.1. Variable Screening

We started off from the SUC dataset extracted from the SUC corpus via SAPIS. The dataset contains 1,040 records and 120 features. 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. This redundancy is detrimental for MDA because it causes multicollinearity, a statistical phenomenon that may lead to distorted results. We ditched out multicollinear features and ended up with 45 linguistic features that are listed in the Appendix.

3.2.2. Running MDA

After having screened the variables, we carried out MDA by building a correlation matrix, checking the determinant, assessing the sample adequacy and finally determining the number of factors. 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 easy. Traditionally, the decision is made by looking at the scree plot. More recently, it has been shown that parallel analysis (Hayton et al., 2004) can help identify the most suitable number of factors. We then ran parallel analysis that suggested three significant factors. We extracted three factors from the correlation matrix and applied 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 two or three variables with high factor loadings, and each variable should load highly only on one factor.
The 3-factor solution explained 0.22 variance, which is, admittedly, a relatively small proportion of the overall variance. However, this in not uncommon with natural language data, because the linguistic data that we find in texts can be very idiosyncratic and ambiguous and this elusiveness is reflected in the factor solution.

3.2.3. Grammatical Breakdown of the Factor Solution

The results of the 3-factor solution was interpreted grammatically and functionally in terms of textual dimensions (Biber, 1988). The functional interpretation of the textual dimensions is described in Santini and Jönsson (2020). Here we list the grammatical makeup of each dimension. Since each dimension has a positive (+) and a negative side (-), that normally are mutually exclusive, we interpreted each side of each dimension as a facet characterising an aspect of text complexity (we ditched out Dim2- because its loadings were below 0.30).
Dim1+ represents the Pronominal-Adverbial Facet. Features that tend to co-occur in Dim1+ are: pronouns, adverbs, interjections, attitude adverbials, question marks, common Swedish words, exclamation marks, negation adverbials, possessive pronouns and comparative adverbials.
Dim1- represents the Nominal Facet. This dimension has two loadings, both quite high, namely on prepositions and nouns, that both indicate the nominal character of the dimension.
Dim2+ represents the Adjectival Facet. This dimension has an adjectival nature since premodifiers, postmodifier and adjectives have the highest loading on this dimension. They are all grammatical devices that elaborate and specify the exact nature of nominals and nouns.
Dim3+ represents the Verbal Facet. The features that characterise Dim3+ are verbs, subordinators and infinitival markers and basic vocabulary.
Dim3- represents the Appositional Facet. The features that characterise this facet are appositions, the verb arity and commas. Appositions are "a maximally abbreviated form of postmodifier, and they include no verbs" (Biber et al., 1999). Commas are a common punctuation device to specify apposition. Verb arity indicates the number of arguments a verb may have. A high average indicates that a high amount of nominal information is glued to verbs.

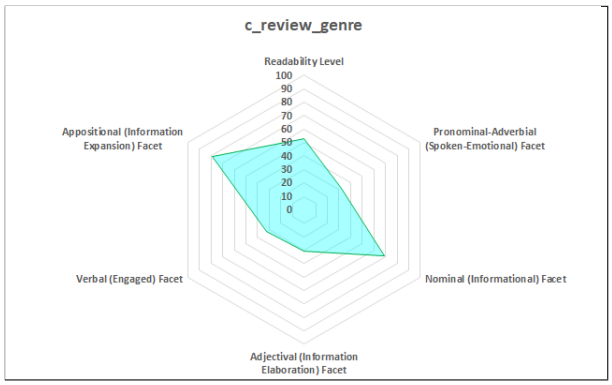

Figure 2: Review

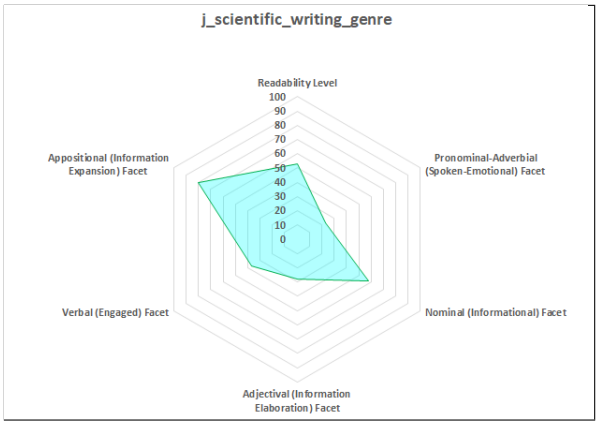

Figure 3: Scientific writing

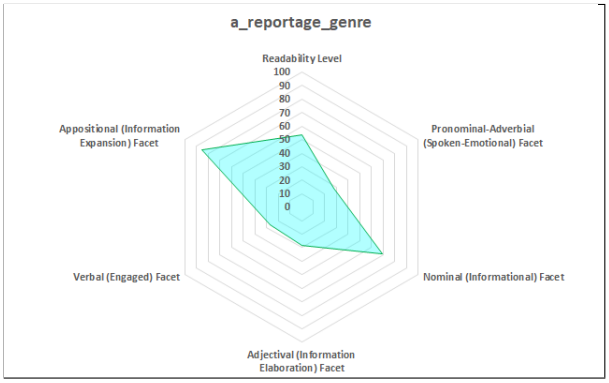

Figure 4: Reportage

3.2.4. Table and Bar Chart

We normalised the positive and negative values of the dimensions on a 0-100 scale in order to have a more accurate picture of how the text complexity facets and readability levels (Björnsson, 1968) vary across the SUC registers. Table 1 shows the SUC registers with normalised values plotted in Figure 1. The chart in Figure 1 is neat and provide interesting insights. For instance, we can observe that the readability level is rather uniform across the registers. When we map these readability values with those in Table 1, we can see that six SUC registers (the majority) have a readability level >50 (Very difficult), two registers are between 41 and 50 (difficult). Therefore all the registers in the SUC are rather difficult with the exception of popular lore (38.7), which appears to be easier to read than other registers. We can also observe that the nominal facet is often strong when also the appositional facet is pronounced.

We realised that the interpretation of the results with this type of visualization was indeed possible but required some cognitive effort and time, even for specialised people like linguists.

4. Visualizing Text Complexity in "Shapes"

To get a more intuitive understanding of the differences and similarities across the registers, we plotted each register as a radar chart and analyzed the a polygonal shape.

We could then observe that the faceted makeup of reviews (Figure 2), scientific writing (Figure 3) and reportage (Figure 4) is very similar. These three registers have a strong nominal facet associated with a pronounced appositional facet. The pronominal-adverbial facet is very flat, and the verbal and adjectival facets are weak. These characteristics are exemplified in the excerpts shown in Tables 2, 3 and 4.

Bio-essay and imaginative prose have similar shapes (see Figures 5 and 6). The bio-essay and imaginative prose registers are characterized by strong pronominal-adverbial, adjectival and appositional facets. These characteristics are exemplified in the excerpts shown in Tables 5 and 6.

The hobby and miscellaneous registers (see Figures 7 and 8) are strong on the nominal-appositional facet (a similarities with the reportage, review and scientific writing registers) but they are also characterised by some prominence of the verbal facet, while the pronominal-adverbial facet and the adjectival facet are rather flat. These characteristics are exemplified in the excerpts shown in Tables 7 and 8.

The editorial and popular lore registers are two singletons (see Figures 9 and 10). They have a shape that is not similar to other registers in the SUC. Editorials have a strong nominal facet, but quite weak appositional facet. The texts in this register are difficult to read and they show a pronounced verbal facet that arguably implies more complex syntax. The adjectival facet is weak, so is the pronominal-adverbial facet. These characteristics are exemplified in the excerpts shown in Tables 9 and 10.

5. Discussion

We used radar charts to profile the registers of the SUC corpus with five text complexity facets and with readability levels. Figures 2-10 visually show the shape of the similarities and dissimilarities across the registers. 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? Essentially, a text complexity facet breaks down the linguistic nature

Excerpt	**LIX score**	**Text ID**	**SUC Register**
	39.05	cb05i	c_review_genre
Swedish		**English Translation**	
Revoltens år, omstörtande och chockerande för många äldre, en optimistisk kamp för framtiden för de unga.		The year of the revolt, destructive and shocking for many elderly people, an optimistic struggle for the future of the young.	

Table 2: Excerpt from a review

Excerpt	**LIX score**	**Text ID**	**SUC Register**
	54.24	ja05	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).	

Table 3: Excerpt of scientific writing

Excerpt	**LIX score**	**Text ID**	**SUC Register**
	41.08	af06j	a_reportage_genre
Swedish		**English Translation**	
Man i Älvkarleö anhållen för hot En 33-årig man vid flyktingförläggningen i Älvkarleö greps på måndagskvällen av Tierpspolisen. Mannen är misstänkt för olaga hot och misshandel av sin hustru.		Man in Älvkarleö arrested for threats 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.	

Table 4: Excerpt from a reportage

Excerpt	**LIX score**	**Text ID**	**SUC Register**
	43.03	gb02	g_bio_essay_genre
Swedish		**English Translation**	
Vi anpassade oss till omständigheterna och valde en läsart, vilken koncentrerade sig på karaktärerna och deras utveckling snarare än på scentekniska mirakel.		We adapted to the circumstances and chose a type of reader, which focused on the characters and their development rather than on the technical miracles.	

Table 5: Excerpt from a text in the Bio-Essay register

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

Table 6: Excerpt from a text of imaginative prose

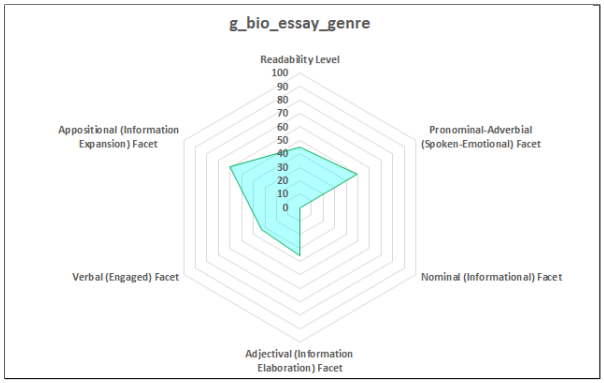

Figure 5: Bio-Essay

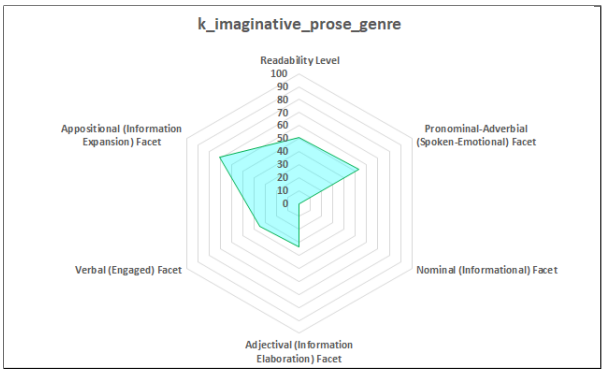

Figure 6: Imaginative prose

Excerpt	LIX score	Text ID	SUC Register
	50.06	ec02d	e_hobby_domain
Swedish		**English Translation**	
Samtidigt varnar han för att Tyskland kan utmålas som syndabock om man inför den europeiska rymdorganisationen Esas ministermöte i november förklarar att landet ensidigt skall dra ned på sitt engagemang.		At the same time, he warns that Germany could be painted as a scapegoat if faced with the European Space Agency ESA's ministerial meeting in November declares that the country should unilaterally reduce its commitment.	

Table 7: Excerpt from a text of the Hobby register

Excerpt	LIX score	Text ID	SUC Register
	43.09	he09c	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.	

Table 8: Excerpt from a text in the Miscellaneous register

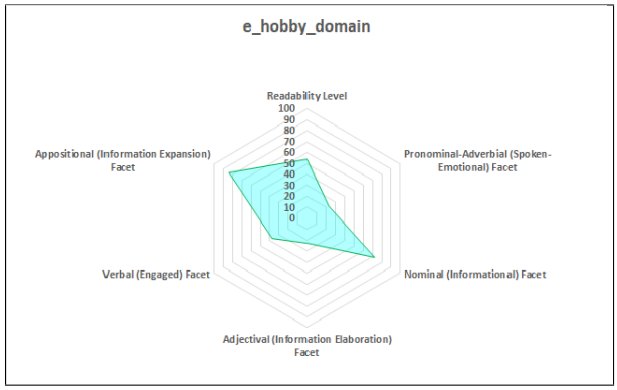

Figure 7: Hobby

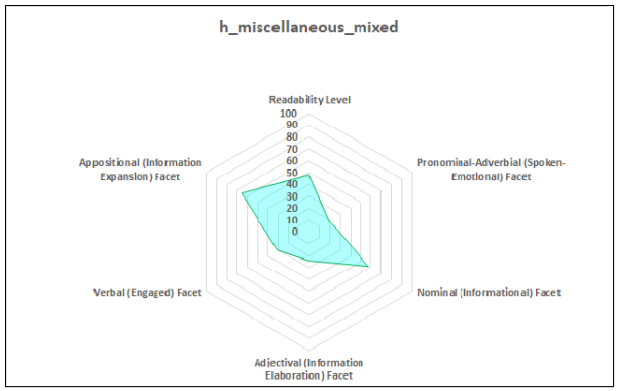

Figure 8: Miscellaneous

Excerpt	LIX score	Text ID	SUC Register
	59.07	ba05d	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.	

Table 9: Excerpt from an editorial

Excerpt	LIX score	Text ID	SUC Register
	46.53	fh03b	f_popular_lore_domain
Swedish		**English Translation**	
Metoden gör att man på ett enkelt sätt kan minska risken för uppkomst av sprickor, förhindra tillväxt av defekter och ge skydd mot plötsliga rörbrott.		The method allows you to easily reduce the risk of cracking, prevent the growth of defects and provide protection against sudden pipe failure.	

Table 10: Excerpt from a text in the Popular lore register

of text complexity and show how influential that facet is with respect to other facets that have a different linguistic makeup. It is, however, the combination of text complexity facets, and not the single facet, that gives us the characterisation of the texts in a register.

6. Conclusion and Future Work

In this paper, we argue that radar charts give an added value to the visualization of the results of MDA by producing "shapes" that help pin down more intuitively linguistic similarities across registers. In the study, we visualized the results of MDA applied to text complexity. From a 3-factor solution, we derived five text complexity facets. These facets highlight combinations of several linguistic aspects. The visualization of text complexity facets with radar charts indicates that there is correspondence between linguistic similarity and similarity of shape across registers. This is the main take away of this paper and it opens up new directions for future research. For instance, it could be possible to automatically compute shape similarity or poly-

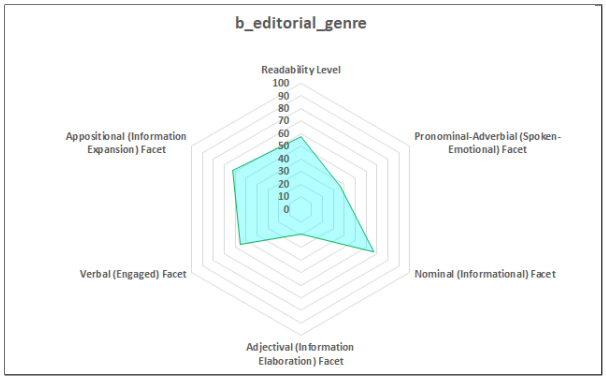

Figure 9: Editorial

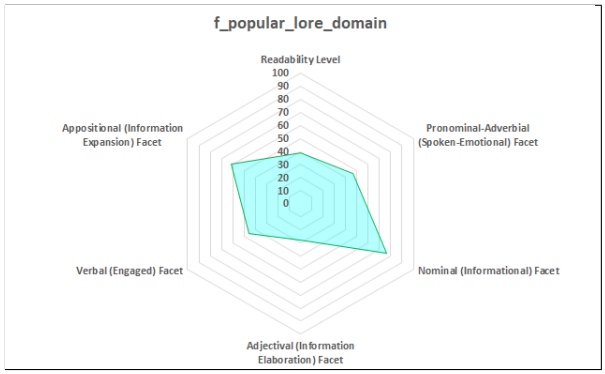

Figure 10: Popular lore

gon matching, which have a long tradition in geometry, to classify text complexity. What is more, the visualization of text complexity in different shapes could help people with cognitive impairments, such as people with dyslexia who have difficulties in detecting words (especially small function words) but have strong visual and spatial reasoning skills. Last but not least, shapes generated by automatic linguistic analysis could be used to as a "hallmark" of the different levels of text complexity and readability and used to guide the reader.

Acknowledgements

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 if fully reproducible. Datasets, radar charts and R code are available here: `http://santini.se/registerstudies2020/`

Appendix: 45 Linguistic Features

3 lexical features
Namely: ratioSweVocC, ratioSweVocD, ratioSweVocH
SweVocC: lemmas fundamental for communication.
SweVocD: lemmas for everyday use.
SweVocH: other highly frequent lemmas.
A high ratio of SweVoc words should indicate a more easy-to-read text.

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 (subodinatingConj), 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).

4 Averages
Namely: 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.
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
avgNominalPremodifiers. The average number of nominal pre-modifiers per sentence.
avgNominalPostmodifiers: The average number of nominal post-modifiers per sentence.

7. Bibliographical References

Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). Longman grammar of written and spoken english. *Harlow: Longman*.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge Univ. Press, 1988.

Björnsson, C. H. (1968). *Läsbarhet*. Liber.

Branco, A., Rodrigues, J., Costa, F., Silva, J., and Vaz, R. (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.

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*, volume 71. John Benjamins Publishing.

Dale, E. and Chall, J. S. (1949). The concept of readability. *Elementary English*, 26(23).

Egbert, J. and Biber, D. (2018). Do all roads lead to rome?: Modeling register variation with factor analysis and discriminant analysis. *Corpus Linguistics and Linguistic Theory*, 14(2):233–273.

Fahlborg, D. and 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., Heimann Mühlenbock, K., and Jönsson, A. (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.

Falkenjack, J. (2018). *Towards a model of general text complexity for swedish*. Ph.D. thesis, Linköping University Electronic Press.

Fang, Z. (2016). Text complexity in the us common core state standards: A linguistic critique. *Australian Journal of Language & Literacy*, 39(3).

Feng, L. (2010). *Automatic Readability Assessment*. Ph.D. thesis, City University of New York.

Gustafson-Capková, S. and Hartmann, B. (2006). Manual of the stockholm umeå corpus version 2.0. Technical report, Stockholm University.

Hayton, J. C., Allen, D. G., and Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, 7(2):191–205.

Heilman, M. J., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Proceedings of NAACL HLT 2007*, pages 460–467.

Hiebert, E. H. (2012). Readability and the common core's staircase of text complexity. *Santa Cruz, CA: TextProject Inc*.

Housen, A., De Clercq, B., Kuiken, F., and Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research*, 35(1):3–21.

Jelen, B. (2013). *Excel 2013 charts and graphs*. Que Publishing Company.

Jönsson, S., Rennes, E., Falkenjack, J., and Jönsson, A. (2018). A component based approach to measuring text complexity. In *Proceedings of The Seventh Swedish Language Technology Conference 2018 (SLTC-18)*.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Nivre, J., Hall, J., and Nilsson, J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, pages 2216–2219, May.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134.

Petersen, S. (2007). *Natural language processing tools for reading level assessment and text simplification for bilingual education*. Ph.D. thesis, University of Washington, Seattle, WA.

Rello, L., Baeza-Yates, R., Bott, S., and 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*, pages 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*, pages 203–219. Springer.

Saggion, H. (2017). Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Santini, M. and Jönsson, A. (2020). Readability revisited? the implications of text complexity. *Register Studies*, 2:2.

Vega, B., Feng, S., Lehman, B., Graesser, A., and D'Mello, S. (2013). Reading into the text: Investigating the influence of text complexity on cognitive engagement. In *Educational Data Mining 2013*.

Wray, D. and Janan, D. (2013). Readability revisited? the implications of text complexity. *The Curriculum Journal*.