Exploring and Analyzing Differences Across Levels of Readability in Easy-to-Read Text

Wilgot Brissman
Department of Computer
and Information Science
Linköping University
and
Fodina Language Technlogy AB
Linköping, Sweden
wilbr523@student.liu.se

Arne Jönsson
Department of Computer
and Information Science
Linköping University, Linköping, Sweden
arne.jonsson@liu.se

Abstract

In this paper, we present results from investigations of text complexity using cohesion measures and their importance related to other text complexity measures. To provide additional nuance, we introduce the interrelated concepts of epistemic stance and narrativity, deepening the analysis of the statistical findings. These concepts also facilitate further discussion on complexity and cohesion as they relate to reading skills and knowledge asymmetries. We employ principal component analysis (PCA) to uncover these statistical relationships on a broader scale, while conducting more specific in-depth analyses of certain metrics. Our findings, which mostly align with existing literature, reaffirm the significance of narrativity in contextualizing cohesion. However, we unexpectedly found a clear link between higher complexity and less narrative text. Additionally, the PCA reveals a more nuanced picture of referential cohesion and the use of its constituent metrics, which varies depending on both narrativity and complexity.

1 Introduction

In writing easy-to-read (ETR) books much thought and effort goes into the process of crafting comprehensible text with understandable plot. Clearly, this entails consideration of vocabulary usage, sentence structure, and other aspects of text complexity. However, a large aspect of the overall complexity that may not always be explicitly taken into account is cohesion. Cohesion describes how interwoven a text is in various aspects, for example semantically, structurally, and conceptually (Graesser et al., 2011). These factors can have a significant impact on the overall impression and difficulty of a text (McNamara, 2013), as well as the degree to which it is interconnected in regards to aspects like causality and temporality.

To provide additional nuance to this discussion, and context for the results, the concept of epistemic stance is central. This describes how the author views the relationship between the knowledge contained in a text and the pre-existing knowledge of the target audience (McNamara, 2013). Related to the notion of epistemic stance is the dimension of narrativity. Narrativity describes the degree to which a text is informational or narrative in its purpose and style (Graesser et al., 2011). Variation in epistemic stance and narrativity have accompanying implications for the complexity and cohesion of the text which will be investigated. However, the degree to which the expected patterns will be reflected in the findings is unclear. These uncertainties are partly a result of the possible impact the ETR nature of the text could have, not to mention the fact that the extent to which the theory surrounding cohesion applies to Swedish is not well-studied.

2 Dataset and Method

In the gathering of data a corpus called Nypon-Vilja was used, created by the publisher *Nypon och Vilja*. It contains ETR books which are divided into six levels of readability based on scales of the companies' own devising. These books come from two different branches of the same company, namely *Nypon*, and *Vilja*, which focus on children's ETR and adult's ETR respectively. As a consequence, there are two parallel interpretations of these levels. In this research, the different classification systems were merged to consolidate the dataset into a smaller number (5) of distinct and meaningful readability levels.

This proved straightforward as the two systems cover relatively similar ranges of readability and both had a total of six levels. The exact extent to which the levels differ is, however, unclear.

SCREAM (Falkenjack, 2018) was used for analysing the texts in the dataset through the SAPIS API (Falhborg and Rennes, 2016) with added Coh-

Metrix measures related to cohesion.

Coh-Metrix is a suite of metrics aimed at establishing a clear picture as to the cohesiveness of a text. Only a smaller subset of these measures will be employed in this study. These include a selection of co-reference metrics, Latent Semantic Analysis (LSA), and ratios of some types of connectives (Graesser et al., 2004). Co-reference metrics and some aspects of LSA are calculated on two levels. Adjacent measurements consider each sentence in relation to the neighbouring sentences exclusively, while global measurements consider each sentence in relation to all other sentences in the same text. This means adjacent measurements are comprised of calculations based on a maximum of two sentence pairings whereas global measurements are based on all possible sentence pairs (Graesser et al., 2004).

The process of data collection resulted in a total of 199 measurements related to complexity and cohesion. An initial screening of variables was conducted, where those that lacked values or were not on a continuous scale were removed. The metrics consisted of 176 continuous variables, 156 related to complexity and 20 to cohesion. Pearson's correlation coefficient *r* was employed when determining correlations throughout the analyses.

All texts were manually classified as either informational or narrative. This was done primarily by consulting the information regarding each book as found on *Nypon och Vilja's* website. Special attention was paid towards the genres to which they belonged. Moreover, an assessment regarding the purpose of the book, and epistemic stance of the author, was a major factor. As such, the process involved a subjective component. Another potential issue is the lack of nuance inherent to a binary categorization, especially when dealing with a topic as complex as narrativity. Thus, there is reason to caution against taking the final partitioning of books as more than a rough guideline.

In order to reduce relevant variables to a more manageable number, two PCA's were conducted, c.f. Jönsson et al. (2018).

The first PCA was performed on a total of 98 complexity metrics[1]. 21 components with an eigenvalue larger than one were found, explaining a total of 78% of the variance in the data. The biggest component explained 27% of the cumulative variance. The Kaiser criterion was employed in selecting the number of components to proceed with, meaning all 21 components were kept. Promax rotation was subsequently used in rotation to a final solution. The largest component was taken to be an adequate representation of complexity. No in-depth interpretations were made for the individual components due to the sheer number of included metrics.

A second PCA was then conducted on 15 metrics related to cohesion using the same methodology[2]. A total of four components with an eigenvalue over one were found, accounting for 78.5% of variance. The four components explain 29%, 26%, 15%, and 8.5% of variance respectively. Once again, Promax rotation was utilized. All four components were also kept, with interpretations made for each.

3 Results & Analysis

Complexity, as presented and analysed here, is taken to be the largest component of the complexity PCA, i.e. the component which explains the largest amount of variance in the data. In Figure 1 the variance of complexity can be seen for each readability level. The black bar for each level shows mean complexity. The blue rectangle indicates the range in which the middle 50% of values occur, while overall variance is illustrated by the thinner line. Dots and stars with related numbers show texts which are considered outliers.

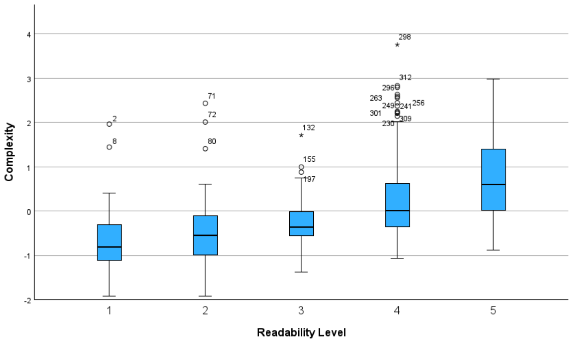

Figure 1: Variance in complexity across readability levels.

Overall, the resultant scores seem reasonable as a corresponding increase in mean complexity is found for each increase in readability level. This picture is somewhat complicated by the details seen

[1] A Kaiser-Meyer-Olkin (KMO) test was run, as well as Bartlett's test of sphericity. With a resulting meritorious KMO score of .817 (Statistics, 2015) and Bartlett's test of sphericity being statistically significant ($p < .001$) it could be concluded that conducting a PCA was a valid approach.

[2] KMO score .689 and Bartlett's test of sphericity showing statistical significance ($p < .001$) indicates a PCA could be appropriate, though the KMO score is considered mediocre.

in Figure 1, where it can be established that there is a significant amount of overlap in complexity across all levels. Most variance can be seen within level 5. The distribution of outliers is, however, disproportionately weighted towards level 4, with all but level 5 containing a minimum of two outliers. A particularly noteworthy finding given that level 5 also contains the six level 6 texts.

The PCA conducted on the metrics related to cohesion revealed four components, see Table 1.

Metric	C1	C2	C3	C4
LSA_Adj_Std	.904			
LSA_Adj_Avg	.887			
Content words_Adj	.871			
Content words_Glob	.817			
Nouns_Adj		.943		
Stems_Adj		.768	.301	
Stems_Glob		.728	.318	
LSA_Givenness	.390	-.821	.346	
Nouns_Glob		.753		-.309
Arguments_Adj			.810	
Arguments_Glob			.938	
Anaphors_Glob			.830	
Anaphors_Adj	-.337		.802	
Causal_Conn				.796
Temporal_Conn				.833

Table 1: Pattern matrix containing the loadings of the four cohesion components.

The table shows the loadings of metrics on each component, which determines the impact they have. A higher absolute value of a loading indicates larger impact. Through a cursory interpretation of the loadings we suggest the following interpretation:

C1 The first component seems to be shaped mainly by two metrics, semantic cohesion as measured by LSA, and content word co-reference. Both adjacent anaphora co-reference and giveness have a secondary role.

C2 For the second component, noun and stem co-reference are important. LSA giveness also plays a significant negative role.

C3 Argument and anaphora co-reference defines this component, with stem co-reference and givenness contributing secondary influences.

C4 The strongest effects on this component are conferred by casual and temporal connectives. There is also a small loading on global noun co-reference.

The interpretations of each component, alongside their associated relationships with complexity, see Table 2, paint an interesting picture. It can, for instance, be noted that all components have a strong correlation to complexity, positive for all but **C1**. This suggests that they all play an important role when measuring text complexity.

Component	Complexity	P-value
C1	-.458	< .001
C2	.589	< .001
C3	.464	< .001
C4	.475	< .001

Table 2: The cohesion components' respective correlations with complexity.

A further analysis through the lens of narrativity provides differences in patterns of complexity and cohesion, as seen in Figures 2 and 3.

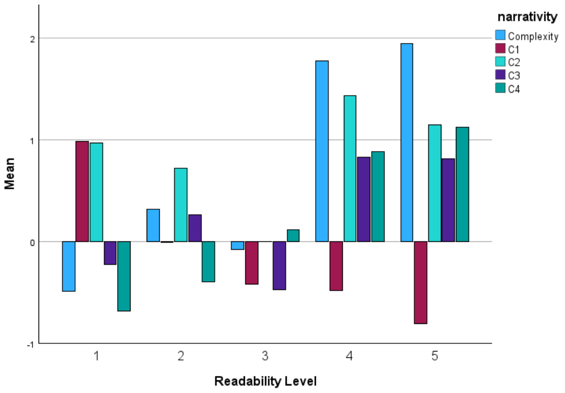

Figure 2: Complexity and cohesion for informational texts.

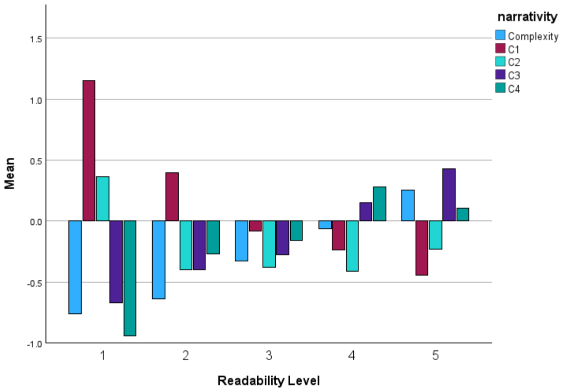

Figure 3: Complexity and cohesion for narrative texts.

The challenges in mapping of the relationships between complexity and cohesion components was largely a product of narrativity being consistently present as a confounding variable. Consequently, it was difficult to distinguish whether a correlation was due to the nature of increasing complexity or the narrativity of the text.

C1 A clear negative relationship with complexity can be seen in the case of C1, which remains in both narrative as well as informational texts. Arguably, narrativity may play a role in the strength of this relationship.

C2 This component has a closer relationship with narrativity, where consistently high levels are seen in informational texts with less dependence on complexity. Narrative texts, in contrast, show persistently low levels of this type of cohesion.

C3 The component tends to track complexity fairly closely in narrative texts. In the case of high-complexity informational texts, the relationship is significantly less strong. Whether this is related to narrativity or a product of high-enough levels of complexity is unclear.

C4 Like with C3, complexity appears to be the determining factor. Due to the high complexity of some informational texts, it is still difficult to establish the exact role narrativity has.

The overall trend of complexity is that it increases as readability level increases. However, the introduction of narrativity into the analysis presents a more unexpected picture. The significant difference lies in the much higher complexity of informational as opposed to narrative text, especially among the highest readability levels. This is noteworthy since lower complexity is presented as a common way of compensating for the increased difficulty of a larger knowledge gap in informational texts (McNamara, 2013). Hence, the expected pattern would be the reverse. As it is, informational texts seemingly pose a greater challenge to reading skill, along with the challenges associated with its greater asymmetry in knowledge. This could be a problematic situation depending on the goal of a text, since a text meant to communicate knowledge might be less effective in its purpose if the higher complexity on its own is too great a challenge for the reader's reading skill. On the other hand a narrative text, virtually by definition, lacks a significant knowledge gap. If it also does not tax the reader's reading skill, it seems to have little value as an educational tool.

4 Conclusion

We report results from the interplay of complexity and cohesion, specifically as it occurs in the ETR books contained in a dataset from *Nypon och Vilja*.

We show that both complexity and cohesion generally increase along with the readability level. However, the interaction of cohesion and complexity with narrativity, partly through its relationship with epistemic stance, proved essential to account for. While cohesion, with the exception of C1, adhered to the predictions of epistemic stance, complexity is another matter. The expectation of lower complexity accompanying less narrative text was overturned. A possible consequence of this was a close apparent correlation between cohesion and complexity.

An additional finding was that of the cohesion components produced by the PCA. It can be established that the linguistic features under the header of referential cohesion can be separated in terms of their use, as all are not related to narrativity and complexity in the same ways.

The findings work to strengthen the theory surrounding the relationship between epistemic stance and cohesion. However, clear and unexpected variation in use of cohesive devices suggest that more research is needed. Results related to complexity also justify further investigation, especially regarding its relationship to narrativity.

Acknowledgments

This research is part of the project Text Adaptation for Increased Reading Comprehension, funded by The Swedish Research Council.

References

D. Falhborg and E. Rennes. 2016. Introducing SAPIS – an API service for text analysis and simplification. *The second National Swe-Clarin workshop: Research collaborations for the digital age*.

J. Falkenjack. 2018. *Towards a Model of General Text Complexity for Swedish*. Licentiate thesis, Linköping University.

A. C. Graesser, D. S. McNamara, and J. M. Kulikowich. 2011. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5):223–234.

A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behaviour Research Methods, Instruments, & Computers*, 36(2):193–202.

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

D. S. McNamara. 2013. The epistemic stance between the author and reader: A driving force in the cohesion of text and writing. *Discourse Studies*, 15(5):579–595.

Laerd Statistics. 2015. Principal components analysis (PCA) using SPSS statistics. https://statistics.laerd.com/premium/spss/pca/pca-in-spss.php. [Accessed: 10 Apr 2024].