Coherence and Cohesion for the Assessment of Text Readability

Amalia Todirascu, Thomas François, Nuria Gala, Cédrick Fairon, Anne-Laure Ligozat, Delphine Bernhard

Coherence and Cohesion for the Assessment of Text Readability

Amalia Todirascu[1], Thomas François[2], Núria Gala[3],
Cédrick Fairon[2], Anne-Laure Ligozat[4], Delphine Bernhard[1]

[1] FDT, LiLPa, Université de Strasbourg
[2] CENTAL, UCLouvain, Place Blaise Pascal 1, 1348 Louvain-la-Neuve, Belgique
[3] LIF-CNRS, AMU, 163 av. de Luminy case 901, 13288 Marseille Cedex 9, France
[4] LIMSI-CNRS, Orsay, France
E-mail: todiras@unistra.fr, thomas.francois@uclouvain.be, nuria.gala@univ-amu.fr, cedrick.fairon@uclouvain.be, annlor@limsi.fr, dbernhard@unistra.fr

Abstract. Text readability depends on a variety of variables. While lexico-semantic and syntactic factors have been widely used in the literature, more high-level discursive and cognitive properties such as cohesion and coherence have received little attention. This paper assesses the efficiency of 41 measures of text cohesion and text coherence as predictors of text readability. We compare results manually obtained on two corpora including texts with different difficulty levels and show that some cohesive features are indeed useful predictors.

1 Introduction

Although reading is considered as a crucial skill in education, reaching a sufficient level is a complex challenge for a significant part of the population. A recent publication from the Council of the European Union reports that “on average in the EU-27 in 2009, 19.6 % of [15-year-old] students were low achievers in reading” (De Coster et al., 2011: 22). One way to sustain the growth of one's reading skills is to offer him/her opportunities for practice, whether guided or independent. Various experiments indicate that regular practice improves reading skills (Mastropieri et al., 1999). For this practice to be profitable, it is also necessary that the texts suit the level of students (O'Connor et al., 2002), which is not always the case.

To assist teachers or readers themselves to more easily find adequate texts, tools have been developed since the 1920's in the field of readability. They are called readability formulas and aim to match readers of various reading abilities with texts that are within their reach, using various textual characteristics for prediction.

Classic formulas such as Flesch's (1948) first focused on a few number of lexico-syntactic characteristics (e.g. the average number of words per sentence or the average number of syllables per word). In the 1980's, the structuro-cognitivist approach of readability stressed the importance of higher textual dimensions such as the inference load (Kintsch, 1979; Kemper, 1983), the conceptual density (Kintsch and Vipond, 1979), or organisational aspects (Meyer, 1982). However, these new dimensions were hardly investigated at that time due to the complexity of the linguistics models involved. Even since, only a few studies -that will be covered in more details in Section 2- focused on those high-level textual dimensions.

Among those high-level dimensions, the level of coherence of the texts is an important one and will be the focus of this paper. It has been shown that a higher level of coherence between a pair of related sentences decreases their reading time and improves their recall (Kintsch et al., 1975). Myers et al. (1987) focused on causal relations and compared the reading speed and the recall of four similar pairs of sentences (expressing the same cause and consequence), ranging from an incoherent version to a very coherent one. They obtained surprising results: while the reading time decreases as the coherence level increases, the recall follows a quadratic function in the shape of an inverted U. In other words, moderately connected sentences are the best remembered ones. Such sentences generally require the reader to make an inference to explicit their relationship barely sketched in the text. This inference generation process produces a higher reading time, but also a richer connection network between the representations of both sentences in memory, leading to a better recall. Mason and Just (2004) used functional magnetic resonance imaging (fMRI) to test this hypothesis with subjects reading the sentences of Myers et al. (1987). They observed activation patterns consistent with Myers et al. (1987)'s findings.

These studies confirm the idea that the more coherent a sequence of sentences, the better these are understood. From these findings, it appears that readability models should benefit from taking into account high-level textual dimensions such as coherence or cohesion. However, as detailed in Section 2.2, current explorations of the issue have failed to achieve a consensus on their efficiency, as they are based on automatic parameterization procedures, prone to errors. In this paper, we propose to investigate whether various measures of text cohesion and coherence are useful to assess the readability of texts, when their parameterization is manually performed. Section 2 further discusses the concepts of coherence and cohesion, and summarizes previous approaches of those dimensions in the readability literature. Section 3 presents the methodology applied in the paper to assess the usefulness of several measures of coherence and cohesion for the prediction of text readability. We also describe the tools and the corpora used in our tests. Section 4 reports a preliminary experiment exploring how features based on cohesion device such as reference chains vary between a normal and a simplified version of the same texts. Based on these results, Section 5 investigates a larger set of variables measuring text coherence and text cohesion, assessing their efficiency to predict French as a foreign language (FFL) text difficulty.

2 Coherence and Cohesion in Readability

2.1 Coherence and Cohesion

Coherence and cohesion are two important properties of texts. Text coherence is considered as a “semantic property of discourses, based on the interpretation of each individual sentence relative to the interpretation of other sentences” (Van Dijk, 1977: 93). A text is realised as a sequence of related utterances. Some theories describe coherence relations by the existence of explicit linguistic markers reinforcing cohesion (Charolles, 1997; Hobbs, 1979). However, cohesive markers are not mandatory elements to obtain coherent texts, although they contribute to the overall text interpretation (Charolles, 1997).

Halliday and Hasan (1976) identified several cohesive devices helpful for the semantic interpretation of the whole text: coreference relations (various expressions referring to the same entity), discourse connectives, lexical relations such as synonymy, hypernymy, hyponymy, meronymy, and thematic progressions. Among these cohesive devices, coreference relations are expressed via anaphoric chains (Kleiber, 1994) or reference chains (Schnedecker, 1997). Anaphoric chains consist of two elements: the anaphor (an expression semantically related to a discourse entity already introduced in the text) and its antecedent (the referred or related entity). The anaphor and its antecedent might be related by various semantic relations such as referential identity or meronymy. Reference chains contain at least three referring expressions, related to the same entity (Schnedecker, 1997). The following example (fig.1) contains two reference chains (un lion étranger ‘a foreign lion’/s'/l'intrus ‘the intruder’; le chef de la tribu ‘the chief of the tribe’/il ‘it’/le dominant ‘the dominant male’), but one anaphoric chain (un combat ‘a fight’/s').

Fig.1. An example of reference and anaphoric chains

Lorsqu'[un lion étranger]_1 au groupe [s']_1 approche, [un combat]_2 [s']_2 engage (parfois jusqu'à la mort) entre [le chef de la tribu]_3 et [l'intrus]_1. S'[il]_3 gagne, [le dominant]_3 reste dans le groupe.

'When a foreign lion approaches the group, a mortal fight involves the chief of the tribe and the intruder. If he wins, the dominant male stays within the group'.

Referring expressions introducing new entities are proper names, indefinite noun phrases, or definite noun phrases. Anaphors referring to known entities are mainly represented by personal pronouns, reflexive pronouns, possessive determiners, demonstrative determiners.

The use of anaphoric or reference chains reinforces the presence of the same entity along the text (Hobbs, 1979). More recent studies such as the Centering theory (Grosz, *et al,*, 1995) claim that some entities used in an utterance are more important than others (centre). This theory proves that local coherence is influenced by the centering properties of the utterance and by the selection of various referring expressions. Referring expressions should verify complex morpho-syntactic (gender and number agreement) and syntactic constrains (syntactic parallelism) to remain the centre of the discourse unit. Along with lexical repetition, such chains contribute to preserve the main topic of the paragraph or of the document.

2.2 The Use of Coherence and Cohesion Measures for Readability

As mentioned above, the level of coherence and cohesion of texts impacts the understanding of readers. However, these aspects were initially not considered in classic readability models (Flesch, 1948), which were limited to lexical and syntactic characteristics. Bormuth (1969) is probably the first to explore the issue. For him, resolving anaphoric relations correctly is a prerequisite to a good understanding of a text. Therefore, he defined 10 classes of anaphora and computed their proportion, as well as the density of anaphora in the text and the mean distance between each anaphora and its antecedent. These two latter features appeared to be the best predictors of text readability among the 12, with respectively a correlation r = 0.532 for density and r = 0.392 for the mean distance. Later, Kintsch (1979) analysed the impact of inferences on understanding and found out that the mean number of inferences required in a text was not well correlated with text difficulty.

Another approach of coherence in readability is based on the latent semantic analysis (LSA) developed by Landauer et al. (1998). This method projects sentences in a semantic space in which each dimension roughly corresponds to a semantic field. Therefore, it better allows assessing the semantic similarity between sentences, since it can capture lexical repetitions, even through synonyms or hyponyms. However, this method is not sensitive to cohesive clues such as ellipsis, pronominal anaphora, substitution, causal conjunction, etc. The application of this technique to readability was first investigated by Folz et al. (1998), who computed the average similarity between each pair of sentences in a text as a proxy of the text overall coherence. This variable was also included in Coh-Metrix (Graesser et al., 2004), along with variations such as word overlap, noun overlap, stem overlap, and argument overlap. However, the efficiency of this variable was not assessed before Pitler and Nenkova (2008), who measured its association with text difficulty and obtained a non-significant $r=-0.1$. Later, McNamara et al. (2010) reached a similar conclusion, showing that an LSA-based variable has not much predictive power. François and Fairon (2012) obtained a higher correlation for French (r = 0.63), but it was due to some specificities of their corpus. They used FFL (French as a Foreign Language) texts from textbooks, including some texts from beginner's textbooks that were merely a list of disconnected sentences. Therefore, the LSA-based feature tended to consider disconnected texts as easy ones, increasing the strength of the correlation and inverting its direction.

An alternative approach to LSA was suggested by Barzilay and Lapata (2008), who view a text as a matrix of the discourse entities[1] present in each sentence. The cohesive level of a text is then computed based on the transitions between those entities. Pitler and Nenkova (2008) implemented this model through 17 readability variables, but none was significantly correlated with difficulty. Feng et al. (2009) also replicated this technique, without getting more efficient features.

Finally, Pitler and Nenkova (2008) drew from statistical language models to propose a cohesion model in which texts are viewed as a bag of discourse relations (temporal, comparison, etc.). These relations are either explicit (when marked) or implicit. The authors computed the likelihood of a text based on its discourse relations, having trained their model on the Penn Discourse Treebank. They obtained interesting correlations for this variable (r = 0.48), which is their best feature.

To conclude, we see that only a few studies focused on using coherence and cohesion measures as predictive variables for readability purposes and mostly for English. It also appears that most variables experimented in the literature were not found significantly correlated with text difficulty. Our study aims to further investigate this issue, focusing on French and taking advantage of (a) several linguistic studies about specific cohesion devices such as reference chains (Schnedecker, 2005) and (b) the availability of RefGen (Longo and Todirascu, 2010), a tool that can help us to capture cohesion and coherence information for French texts.

[1] They define a "discourse entity" as nominal phrases being part of a co-reference relation and having a function (subject, object, etc.).

3 Methodology to Assess whether Coherence and Cohesion Correlate with Readability

Our goal is to investigate the use of several cohesive and coherence properties to evaluate the difficulty of French texts. To this aim, we built two annotated corpora to be used in our experiments. Then, drawing on the literature reported in Section 2, we defined 41 variables aiming at measuring text coherence and cohesion (see Section 4). Although we intended to annotate all of them manually, some of them were eventually computed with RefGen (a tool that we introduce in Section 3.2), when the error rate of their annotation process was deemed low enough. Finally, the efficiency of these variables as predictors of text difficulty was assessed on the corpora (see Sections 4 and 5).

3.1. The Corpora

Two corpora were collected for our experiments, both being annotated in terms of text difficulty. The first one is a corpus of comparable texts from Wikipedia and Vikidia[2] (a simplified encyclopaedia targeted at children between 8 and 13 years old). We collected 13 informative texts from Wikipedia, describing animals or geographic areas (7,597 tokens) and selected texts on the same subject from Vikidia (5,308 tokens). This corpus was used as a way of detecting interesting features for the rest of the analysis and to gather significant differences between simplified and original texts. To analyse significant features for readability, we manually annotated the corpus' reference chains.

The second is a subset of the corpus of FFL texts gathered by François (2009). This corpus consists of 2,160 texts, selected from 28 FFL textbooks, as long as they are related to a reading comprehension task. All textbooks considered comply with the Common European Framework of Reference for Languages (CEFR), a standard scale for foreign language education in Europe that uses 6 levels (A1 to C2). Therefore, each text was assigned the level of the textbook it came from. In this study, we only used texts from levels A2 to C1 and selected only informative texts to control for the genre of the texts across both experiments. A1 texts were rejected because several of them were just a collection of unconnected sentences. C2 texts were not considered either because there were not enough informative texts for this level in François (2009)'s corpus.

3.2 Annotation of Discourse Entities and of Reference Chains

The computation of our variables for both corpora would require a large amount of manual work, which led us to consider the automation of some tasks (e.g. POS-tagging or detection of entities), provided that their error rate remains low. Few tools are available for coreference resolution in French (Victorri, 2005; Popescu-Belis, *et al*, 1999) but most of them focus on specific anaphora type (Lassalle and Denis, 2011) or specific domains or tasks (human-machine dialogue systems (Salmon-Alt, 2001)). RefGen is a rule-based system for French which performs the automatic annotation of reference chains (Longo and Todirascu, 2010b), but also entity detection and-POS tagging. RefGen tags and lemmatizes the texts using TTL (Ion, 2007) and it annotates potential referring expressions such as: complex noun phrases (simple NP modified by several PP or relative clauses), named entities (persons and organisations), definite or indefinite noun phrases. In addition, the tool applies several heuristics to label syntactic functions (subject, object, and others). After deleting impersonal occurrences of the 3rd person singular pronoun ('il'), the tool identifies a set of referring expressions as possible starters of a reference chain. Then, RefGen computes a set of antecedent and anaphor pairs by checking several morpho-syntactic and semantic features. Finally, the system groups the candidate pairs into reference chains.

Longo and Todirascu (2010) evaluated RefGen by using a corpus of 7,230 tokens and obtained good results for the entity annotation module (for the module identifying complex noun phrases: recall = 0.87 and precision = 0.91, for the named entity recognition: recall = 0.85 and precision = 0.91) and promising results for the reference chain identification module (recall = 0.58 and precision = 0.70). Reference chain identification is known to be a difficult task, which explains the lower results obtained for this second task. As a consequence, we decided to use this tool to identify discourse entities, but we manually annotated the relations between the referring expressions as well as their syntactic functions.

[2] This corpus was build and annotate by Ratiba Khobzi, University of Strasbourg.

4 Reference Chains in Wikipedia and Vikidia

As a first investigation of the usefulness of coherence and cohesion variables for text readability prediction, we studied the behaviour of reference chains in a corpus of original texts and their simpler version. It should be mentioned that reference chains have a specific behaviour according to text types or genres. Schnedecker (2005) and Schnedecker and Longo (2012) identify specific properties of reference chains in newspapers genres, such as portrays and news. These studies investigated properties such as the length (the number of referring expressions composing the reference chain), the distance (the number of sentences separating the expressions composing the same chains), the types of referring expressions, and the type of the first element starting a chain. The same properties have been studied in several text genres: law texts, editorials, novels, public reports (Longo et Todirascu, 2013). The study shows significant variations in these properties: longer chains characterize novels, newspaper articles contain medium-sized chains, while law texts contain very short ones. The types of referring expressions composing reference chains also differ from one genre to another: news contain more proper names and personal pronouns, while law texts and public reports contain more indefinite and definite noun phrases. To control as much as possible for this variation across genres, we restricted our analysis to one genre: informative texts.

The properties highlighted by the above studies were manually annotated in our first corpus (Vikidia and Wikipedia). In addition, we compared the number of reference chains, the syntactic functions of the referring expressions composing the chains and the relation between the reference chains and the text topic.

We noticed that the number of reference chains was slightly more important in simple texts (49) than in the original (44). For most of the texts, the average length of the reference chains found in simple texts is shorter than the length of the chains in the original texts. To give an example, 'Le lion' is the main referent in both of the following excerpts; the Wikipedia text contains four expressions referring to it while the Vikidia one has only two (pronouns) (fig.2):

Fig.2. An example of annotated reference chains in Wikipedia et Vikidia texts.

[Le lion]_1 (Panthera leo) est un mammifère carnivore de la famille des félidés du genre Panthera (félins). [Il]_1 est surnommé" [le roi des animaux]_1" car [sa]_1 crinière [lui]_1 donne un aspect semblable au Soleil, qui apparaît comme " le roi des astres ". (Wikipedia)
'The lion is a carnivore mammal, member of the family Felidae, in the genus Panthera (felins). It is named «king of animals » due to its mane, which gives it the aspect similar to the Sun, which is « the king of asters »'

[Le lion]_1 est un mammifère carnivore ressemblant au chat. [Il]_1 fait partie, comme lui, des félins. [Son]_1 nom scientifique est Panthera leo. (Vikidia)
'The lion is a carnivore mammal similar to a cat. It is member of the felins. Its scientific name is Panthera leo'

The distribution of the referential expressions types shows that while the relative frequencies of indefinite (0.2 for Wikipedia and 0.35 for Vikidia) or definite noun phrases (2.59 vs 2.83) are similar in both corpora, several categories are more frequent in simple texts than in the original: proper names (0.07 for Wikipedia; 0.26 for Vikidia), personal pronouns (2.23 for Wikipedia; 3.69 for Vikidia) and demonstrative pronouns (0.04 for Wikipedia; 0.2 for Vikidia) .

It should also be noticed that the first element opening a reference chain is more likely to be a definite noun phrase or a NP without determiner for Wikipedia texts, while we observed a preference for indefinite noun phrases in simple texts. In both cases, however, the entity referred to within the longest reference chain is generally the global topic of the document.

Finally, we studied the syntactic function of the referring expressions contained in the chains. We investigated the subject, object and other syntactic functions of the mentions contained in chains. We counted all the transitions (subject-object, subject-subject; subject-other function etc.) between two consecutive sentences containing mentions of the same entity (e.g. part of the same reference chain). The most interesting cases are those with the same syntactic function kept in two consecutive sentences. We observed that this happens more frequently in complex texts than in simple ones. The number of subject pronouns is also more important in simple texts than in Wikipedia texts. In other words, we noticed several variations between the behaviour of reference chains between simple texts and their Wikipedia counterparts. To confirm these trends, we then performed a more quantitative investigation, described in the next section.

5 Cohesion and Coherence for the Readability of FFL Texts

5.1 Variables of Text Coherence and Cohesion

At the end of our preliminary study on Wikipedia and Vikidia texts, several characteristics of text coherence and cohesion appeared to be valuable for readability prediction. Therefore, based on the literature in readability and the work of Schnedecker (1997, 2005), we defined 41 variables, divided up within five classes as follows:

1. *P.O.S. tag-based variables*: Pronouns and articles are crucial elements of coherence and cohesion. We computed 9 variables based on these part-of-speeches, namely (1) the ratio between pronouns and nouns; the average proportion of pronouns per sentence (2) and per word (3); the average proportion of personal pronouns per sentence (4) and per word (5); the average proportion of possessive pronouns per sentence (6) and per word (7); and the average proportion of definite articles per sentence (8) and per word (9). We also computed the ratio of proper names per word (10).
2. *Lexical coherence measures*: We also replicated several methods based on lexical cohesion, namely (11) the average similarity – measured with cosinus – between adjacent sentences projected in a LSA space, (12) the word overlap (number of common words in two consecutive sentences), (13) the lemma overlap, and the noun and pronouns overlap, based either on lemmas (14), or inflected forms (15). More precisely, every text from the corpus was transformed in a list of bag-of-words vectors (one per sentence), before these vectors were weighted. In the case of the various "overlap" variables, $tfidf$(term frequency-inverse document frequency) was used for the weighting, while we applied a singular value decomposition (SVD) for LSA[3].
3. *Entity coherence*: consecutive sentences can share similar arguments (the subject of the sentence *n* is also the subject of the sentence *n+1*, the object of the sentence *n* becomes the subject of the sentence *n+1*, etc.). We followed Pitler and Nenkova (2008) by counting the relative frequency of the possible transitions between the four syntactic functions played by the entity in sentence *n+1*: subject (S), object (O), other complements (C), and (N) when the entity is absent (variables 16 to 28).
4. *Entity density*: we computed the average proportion of entities (simple and complex noun phrases, pronouns, etc.) per document (29), the average number of entities per sentence (30), the average proportion of unique entities per document (31), and the average number of words per entity (32). These features were obtained with the automatic annotation provided by RefGen.
5. The last class gathers features corresponding to various properties of the reference chains: the proportion of the various types of expressions included in a reference chain : indefinite NP (33), definite NP (34), personal pronouns (35), possessive determiners (36), demonstrative determiners (37), demonstrative pronouns (38), reflexive pronouns (39), or proper nouns (40); the average length of reference chains (41).

5.2 Analysis of the Variable Efficiency

We saw that findings in the literature about the efficiency of coherence and cohesion-based variables for readability are not consistent: some of the studies report non-significant correlations, while other show significant correlations. An explanation for this variation could be the fact that most of those studies rely on an automatic approach of coherence and cohesion, which are notoriously difficult to automatize.

To better control for this aspect, we opted, in this study, for a manual approach of all variables whose automatic annotation would have been impaired by a significant error rate. These experiments were performed on the FFL corpus, that includes texts with a larger spectrum of difficulty. However, since manual annotation requires a much larger amount of resources, we restricted the experiment to 5 texts per

[3] To compute the $tfidf$and the LSA, we used a large amount of texts from the François (2009)'s corpus that were not used for this study. For the LSA, we compared various sizes for the reduced space with a cross-validation procedure that led us to retain a small 15-dimensional space.

level, for a total of 20 texts. We manually annotated the reference chains and their syntactic functions, and then computed all variables described in Section 5.1. Their efficiency as readability predictors was then assessed through Spearman correlations[4] between each variable and the levels of the texts. Table 1 reports the most significant correlations.

Table 1. The most significant correlations obtained from the manually annotated corpus. The numbers preceding the variables refer to numbers used in Section 5.2

Variable	Corr. and p-value	Variable	Corr. and p-value
35. PRON	-0.59 (p = 0.005)	3. Pers. Pro./S	-0.41 (p = 0.07)
33. Indef NP	-0.50 (p= 0.02)	10. Names/W	-0.4 (p = 0.08)
18. S → O	0.46 (p = 0.04)	9. # def. art./W	0.38 (p = 0.1)
22. O → O	-0.44 (p= 0.048)	17. S → S	-0.36 (p = 0.12)

Interestingly, two variables based on reference chains are significant: the proportion of transitions of the type subject (S) to object (O) between sentences, as well as the proportion of object (O) to object (O). S-O transitions seem to appear more frequently in harder texts, while the O-O (and also S-S) are typical of easier texts[5]. This finding is interesting, since neither Pitler and Nenkova (2008) nor Feng and al. (2009) were able to show the efficiency of the class of variables for readability, using an automatic approach.

Considering the type of referring expressions used in the chains also seems promising. Our two best features are indeed the proportion of personal pronouns and indefinite NP in the chains. Both types of phrases tend to be more present in easier texts. As regards the average length of the chains, it was surprising to notice that long chains are represented similarly in simple texts and in complex ones.

6 Conclusion and Future Work

The experiments in this paper demonstrated that some variables of text coherence and text cohesion are interesting predictors of text readability. We showed that variables based on syntactic transitions present a different profile in simple and complex texts, with more transitions keeping the same function from one sentence to the next one in simpler texts. This is already an interesting finding, since previous approaches of the issue, based on automatic modelling, obtained non-significant correlations. Furthermore, based on the work of Schnedecker (2005) and Schnedecker and Longo (2012), we suggested new features for readability, like the proportion of the type of referring expressions in the chains. Our most interesting finding is that among those features, two of them (PRON, Indef NP) appeared to be good variables, actually our best ones. Therefore, it is useful not only to consider the function of the referring expressions, but also their type. Simpler texts from our corpus indeed tend to use more pronouns and indefinite NPs.

Our manual approach confirmed the interest to consider textual dimensions, such as coherence and cohesion, to assess the readability of informative texts. Several of our variables indeed were able to discriminate between L1 texts (Wikipedia and Vikidia) and FFL texts of various levels. Since, previous work, based on an automatic analysis, were more mitigated on this issue, especially regarding variables based on the syntactic transitions, our findings could be interpreted in two ways: (1) either the significant correlations we observed are due to some specificities of our corpora (genre of the texts, small amount of observations, etc.), or (2) the fact that previous work had trouble to demonstrate the efficiency of coherence and cohesion variables for readability is mostly due to errors in the annotation procedure, performed automatically.

[4] Spearman correlation formula is described among others in Howell (2008). We did not use the Pearson correlation here, since readability variables often do not have a linear relationship with difficulty.

[5] This second feature is also rarely observed and it is not obvious that its efficiency would scale to a larger set of data.

To decide between these two conclusions, our analysis should be replicated on a larger corpus, on one side, but should also be performed via an automatic annotation procedure. This would allow to check whether our best variables remain efficient once they are extracted via an automatic system such as RefGen. In further experiments, we plan to investigate if the use of automatic annotations of reference chains, and the inherent annotation errors would impact the efficiency of our coherence and cohesion variables. A last step to our investigation would be to test whether coherence and cohesion dimensions really bring new information to a readability model, as regards to information already contained in lexico-syntactic features.

Acknowledgements

This project is partly financed by the Programme Hubert Curien (PHC) Tournesol 2013 (France and Fédération Wallonie-Bruxelles). We would like to also acknowledge the help of Bernadette Dehottay during the collection of the FFL corpus and Ratiba Khobzi for the annotation of the Vikidia and Wikipedia corpora. RefGen has been developed during Laurence Longo's Ph.D.thesis ("Vers de moteurs de recherche sémantiques"), University of Strasbourg.

References

Barzilay, R., Lapata, M. (2008) Modeling Local Coherence: An Entity-based Approach, *Computational Linguistics*, 34(1):1-34.

Bormuth, J. (1969) *Development of Readability Analysis*. Rapport technique, Projet n°7 – 0052, U.S. Office of Education, Bureau of Research, Department of Health, Education and Welfare, Washington, DC.

Charolles, M. (1995) Cohesion, coherence et pertinence de discours, *Travaux de Linguistique*, 29:125-151.

De Coster, I., Baidak, N., Motiejunaite, A., and Noorani, S. (2011) Teaching Reading in Europe: Contexts, Policies and Practices. *Education, Audiovisual and Culture Executive Agency, European Commission*.

Feng, L., Elhadad, N. et Huenerfauth, M. (2009) Cognitively motivated features for readability assessment. *Proceedings of EACL 2009*, pp. 229-237.

Flesch, R. (1948) A new readability yardstick. *Journal of Applied Psychology*, 32(3):221-233.

Foltz, P., Kintsch, W. and Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2):285-307.

François, T. (2009) Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. *Proceedings of EACL 2009: Student Research Workshop*, pp. 19-27.

François, T. and Fairon, C. (2012) An "AI readability" formula for French as a foreign language. *Proceedings of EMNLP 2012*, Jeju, pp. 466-477.

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

Grosz, B., Joshi, A., Weinstein, S. (1995) Centering: A Framework for Modeling the Local Coherence of Discourse.

Halliday, M.A.K. and Hasan, R. (1976) *Cohesion in English*. London: Longman.

Hobbs, J. (1979) Coherence and Coreference. *Cognitive Science*, 3(1):67-90.

Howell, D. (2008). *Méthodes statistiques en sciences humaines*, 6ème édition. De Boeck, Bruxelles.

Ion, R. (2007) *Metode de dezambiguizare semantică automată. Aplicații pentru limbile engleză şi română*, Teză de doctorat, Bucureşti: Academia Română.

Kemper, S. (1983) Measuring the inference load of a text. *Journal of Educational Psychology*, 75(3):391-401.

Kintsch, W. (1979) On modeling comprehension. *Educational Psychologist*, 14(1):3-14.

Kintsch, W., Kozminsky, E., Streby, W., McKoon, G. et Keenan, J. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, 14(2):196-214.

Kintsch, W. and Vipond, D. (1979) Reading comprehension and readability in educational practice and psychological theory, In: Nilsson, L. (ed) *Perspectives on Memory Research*, Hillsdale, NJ: Lawrence Erlbaum, pp.329-365.

Kleiber, G. (1994) *Anaphores et Pronoms*. Louvain-la-Neuve: Duculot.

Landauer, T., Foltz, P. et Laham, D. (1998) An introduction to latent semantic analysis. *Discourse processes*, 25(2):259-284.

Lassalle, E and Denis, P. (2011). Leveraging different meronym discovery methods for bridging resolution in French In Anaphora Processing and Applications, Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Selected papers. LNAI, Springer.

Longo, L. and Todirascu, A. (2010) Genre-based Reference Chains Identification for French, *Investigationes Linguisticae,* 21:57-75.

Longo, L., Todirascu, A. (2013) Une étude de corpus pour la détection automatique de thèmes. *Actes des 6e journées de linguistique de corpus (JLC 09)*, pp. 143-155.

McNamara, D., Louwerse, M., McCarthy, P. et Graesser, A. (2010) Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes, 47(4):292-330.

Mason, R. and Just, M. (2004). How the brain processes causal inferences in text. *Psychological Science*, 15(1):1-7.

Mastropieri, M. A., Leinart, A., and Scruggs, T. E. (1999) Strategies to increase reading fluency. *Intervention in School and Clinic*, 34:278-283.

Meyer, B. (1982). Reading research and the composition teacher : The importance of plans. *College composition and communication*, 33(1):37-49.

Myers, J., Shinjo, M. and Duffy, S. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language*, 26(4):453-465.

O'Connor, R. E., Bell, K. M., and Harty, K. R. (2002). Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty. *Journal of Educational Psychology*, 94: 474-485.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. *Proceedings of EMNLP 2008*, pp. 186-195.

Popescu-Belis A., Robba I. & Sabah G. (1998) Reference Resolution Beyond Coreference: a Conceptual Frame and its Application. *Proceedings of Coling-ACL'98 (International Conference on Computational Linguistics - Meeting of the Association for Computational Linguistics)*, Montreal, Canada, p.1046-1052.

Salmon-Alt, S. (2001) *Référence et dialogue finalisé : de la linguistique à un modèle opérationnel*, Thèse de doctorat, Université Henri Poincaré, mai 2001

Schnedecker C. (1997) *Nom propre et chaînes de référence*. Recherches Linguistiques 21. Klincksieck, Paris.

Schnedecker, C. (2005) Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux de Linguistique*, 2: 85-133.

Schnedecker, C., Longo, L. (2012) Impact des genres sur la composition des chaînes de référence : le cas des faits divers, In : Neveu, F., Muni Toke, V., Blumenthal, P., Klingler, T., Ligas, P. Prévost, S., and Teston-Bonnard, S. (Eds.) *3e Congrès Mondial de Linguistique Française*, Lyon, France, July 2012, pp. 1957-1972.

Van Dijk, T. (1977) T*ext and Context: Exploration in the Semantics and Pragmatics of Discourse* . London: Longman.

Victorri, B. (2005) Le calcul de la référence, *Sémantique et traitement automatique du langage naturel*, Patrice Enjalbert (Ed.) (2005) 133-172