Are Cohesive Features Relevant for Text Readability Evaluation?

Amalia Todirascu, Thomas François, Delphine Bernhard, Núria Gala, Anne-Laure Ligozat

Are Cohesive Features Relevant for Text Readability Evaluation?

Amalia Todirascu
FDT, LiLPa
Université de Strasbourg
todiras@unistra.fr

Thomas François
Post-doc FNRS
IL&C, CENTAL
UCLouvain
thomas.francois@uclouvain.be

Delphine Bernhard
FDT, LiLPa
Université de Strasbourg
dbernhard@unistra.fr

Núria Gala
LIF-CNRS, Aix-Marseille Université
nuria.gala@univ-amu.fr

Anne-Laure Ligozat
LIMSI-CNRS, Orsay
annlor@limsi.fr

Abstract

This paper investigates the effectiveness of 65 cohesion-based variables that are commonly used in the literature as predictive features to assess text readability. We evaluate the efficiency of these variables across narrative and informative texts intended for an audience of L2 French learners. In our experiments, we use a French corpus that has been both manually and automatically annotated as regards to co-reference and anaphoric chains. The efficiency of the 65 variables for readability is analyzed through a correlational analysis and some modelling experiments.

1 Introduction

Since the 1920's, various readability formulae have been designed to match texts with the reading skills of specific readers. The most famous of these formulas, such as Flesch's (1948) or Dale and Chall's (1948) are typical of what are called "classic" formulas. They rely on a few lexico-syntactic characteristics (e.g., the average number of words per sentence or the average number of syllables per word) to estimate the reading difficulty of a text. This strategy worked to some extent, but, from the late 70's onward, classic formulae have been seriously criticised. Zakaluk and Samuels (1988, 124) thus said: "A basic limitation of readability formulas is that they ignore such critical text factors as cohesiveness and macrolevel organization".

Studies in readability from this period stressed the importance of higher textual dimensions, focusing on inference load (Kemper, 1983), conceptual density (Kintsch and Vipond, 1979), or organisational aspects (Meyer, 1982). As a result, the classic lexico-syntactic features were disregarded for years. However, Miller and Kintsch (1980) soon noticed that including lexico-syntactic features in their cognitive readability formulas improved performance. Chall and Dale (1995, 111) had a more mixed opinion, arguing that variables based on higher textual dimensions "discriminate better among materials requiring greater maturity in reading ability", while classic lexico-syntactic variables work better to discriminate at lower levels of difficulty.

Recently, taking advantage of the opportunities offered by Natural Language Processing (NLP) techniques, readability studies have tried to leverage the semantic and discursive properties of texts to better model text difficulty (Pitler and Nenkova, 2008; Feng et al., 2009). Among those high-level dimensions that have attracted substantial attention are the level of cohesion and coherence of texts. Although psycholinguistic experiments have shown that a higher level of cohesion and coherence between a pair of related sentences decreases their reading time (Kintsch et al., 1975; Mason and Just, 2004), the added value of these textual dimensions for readability models (compared to traditional features) remains unclear, as it will be covered in more details in Section 2.

This is why this paper aims at further investigating the importance of cohesion aspects for the assessment of text readability, as the cohesive dimension is the one that have been investigated the most (see Section 2.2). Based on a corpus of texts for learners of French as a foreign language (L2), which has

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 987–997, Osaka, Japan, December 11-17 2016.

been manually annotated for co-reference chains, the three following research questions will be investigated: (1) are cohesive features relevant for text readability assessment? (2) what is the impact of NLP routines, which are error-prone, on the efficiency of cohesiveness features? and (3) does the genre of the texts (here narrative and informative) influence the discriminating power of cohesiveness features? The methodology applied to investigate these three questions is described in Section 3, while the results are presented in Section 4. The paper concludes with a discussion and some perspectives in Section 5.

2 Cohesion Features to Assess Text Readability

2.1 Coherence and Cohesion

Coherence is defined as a "semantic property of discourse, based on the interpretation of each individual sentence relative to the interpretation of other sentences" (Van Dijk, 1977, 93). The order of the ideas, a logical structuring of the text and coherent relations (consequence, cause-effect) between sentences facilitate the reader's understanding of a specific topic. In addition, readers might use external knowledge as regards the specific situation described in the text.

Cohesion is a property of text represented by explicit formal grammatical ties (discourse connectives) and lexical ties that signal how utterances or larger text parts are related to each other. Halliday and Hasan (1976) identify specific cohesive devices aiming to reinforce lexical ties, such as anaphoric chains or co-reference chains (Schnedecker, 1997), as well as lexical chains (sets of expressions related by hypernymy or hyponymy relations or expressions from the same domain, e.g. *patient–disease-treatment*).

Anaphoric chains are composed of two expressions, one antecedent and one anaphora. In Figure 1, the interpretation of the definite noun phrase *the ship* (the anaphora) is dependent on its antecedent (*the RMS Titanic*). Co-reference chains are composed of at least three referring expressions corresponding to the same discourse entity (Schnedecker, 1997). In Figure 1, the expressions *Edward Smith, an English naval reserve officer, He, He* refer to the same entity, the Titanic's commander. Lexical chains are composed of associated words or expressions related by ontological relations (synonymy, hypernymy, hyponymy) or relative to the same domain (Hirst and St-Onge, 1998), such as *naval reserve officer, vessels, ship sank, voyage* (Figure 1).

Edward John Smith was *an English naval reserve officer*. *He* served as commanding officer of numerous White Star Line vessels. *He* is best known as *the captain of* ***the RMS Titanic***, perishing when **the ship** sank on the 15th April 1912. (Wikipedia)

Figure 1: Example of anaphoric and of co-reference chain.

These three devices strengthen the links between several utterances and contribute to the overall understanding of the text (Charolles, 1995). Lexical chains are effective mechanisms to find the main domain or theme of the document. Cohesive devices such as anaphora or co-reference chains correspond to one entity expressed by various linguistic expressions (so called mentions). These expressions are related by complex morpho-syntactic, syntactic or semantic constraints (Grosz et al., 1995). Mentioning the same entity several times reinforces text cohesion (Poesio et al., 2004), (Hobbs, 1979). Cohesive devices reinforce local coherence relations in some specific genres (persuasive genres) (Berzlnovich and Redeker, 2012).

An interesting characteristic of cohesive devices is that their use is dependent on the type or genre of texts (Carter-Thomas, 1994). For instance, informative texts use specific referential expressions such as definite or demonstrative noun phrases as mentions, while narrative texts contain more chains composed of proper nouns or personal pronouns (Schnedecker, 2005). The composition, the length or the choice of the first mention of the co-reference chain is also dependent on the genre. For instance, in newspapers portraits (Schnedecker, 2005), co-reference chains start with a proper noun and contain mainly definite noun phrases and personal pronouns. For example, in law and administrative texts, reference chains start with indefinite noun phrases and the mentions are mainly definite or demonstrative noun phrases (Longo and Todirascu, 2014).

In this article, we consider explicit lexical ties such as anaphoric, co-reference and lexical chains as cohesive features. We study the correlation between these cohesive features and text complexity.

2.2 Coherence and Cohesion in Readability

As both coherence and cohesion are important text properties that are known to influence the readability of texts, readability studies have attempted to exploited both dimensions. However, most studies focused on phenomena that falls inside the category of cohesion as defined in Section 2.1 which is why we decided to focus on cohesive features in this paper.

The first to investigate the issue of text cohesion in readability analysis is probably Bormuth (1969). He considered that the correct identification of anaphoric relations was a prerequisite to the correct understanding of a text and thus computed 12 variables based on various characteristics of anaphora, showing that the density of anaphora to be the best predictor with a $r = 0.532$.

More recently, text cohesion were investigated in readability with another approach that relies on latent semantic analysis (LSA) (Landauer et al., 1998). This technique projects sentences in a semantic space in which each dimension roughly corresponds to a semantic field. This makes it possible to better measure the semantic similarity between sentences, since it can capture lexical chains through lexical repetitions, even through synonyms or hyponyms. However, this method cannot detect cohesive clues such as ellipsis, pronominal anaphora, substitution, causal conjunction, etc. Folz et al. (1998) were the first to apply this technique to readability, by computing the average similarity between each pair of sentences in a text. This variable was also included in Coh-Metrix (Graesser et al., 2004), along with similar measures such as word overlap, noun overlap, stem overlap, and argument overlap. However, the efficiency of this variable for readability was not assessed before Pitler and Nenkova (2008), who measured its association with text difficult and obtained a non significant correlation ($r = -0.1$). Later, McNamara et al. (2010) reached a similar conclusion, showing that an LSA-based variable has not much of a predictive power. On the opposite, François and Fairon (2012; 2013) obtained a higher correlation ($r = 0.63$) for an L2 corpus, while Dascalu et al. (2013) got good discriminating features using both LSA and LDA (Latent Dirichlet Allocation), when classifying TASA (Touchstone Applied Science Associates) texts.

An alternative approach to LSA, Lexical Tightness (LT), was suggested by Flor et al. (2013). They define the LT of a text as the mean value of the Positive Normalized Pointwise Mutual Information for all pairs of content-word tokens in a text. It represents "the degree to which a text tends to use words that are highly inter-associated in the language". They obtained a good correlation between this new cohesive metric and the grade levels on two corpora (respectively $r = -546$ and $r = -0,441$). Interestingly, they also show that LT works better to discriminate between literary texts than informative ones.

Another approach is to detect co-reference chains and compute some of their characteristics. Barzilay and Lapata (2008) considered a text as a matrix of discourse entities 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. Dascalu et al. (2013) computed other characteristics of lexical chains and co-reference pairs (such as the number of chains, the distance between entities, the average word length of entities, etc.). However, with these features, they only reached a precision of respectively 0.367 and 0.384 for a six-class classification problem.

Todirascu et al. (2013) argued that these mixed results might be due to approximations of the NLP systems, since automatically annotating co-reference chains remains a challenge. They manually annotated co-reference chains in 20 texts and correlated various characteristics of lexical chains with the difficulty of these texts. They showed that considering the type of entities, and not only their syntactic transitions, could be valuable. However, only four features appeared to be significantly correlated with difficulty, possibly due to the limited size of their corpus.

3 Methodology

Faced with this mixed findings in the literature regarding the efficiency of cohesive features for the assessment of text readability, our goal is to further investigate this issue. In particular, we present experiments focusing on cohesive features : anaphora chains, reference chains and lexical chains (evaluating sentence similarity).

For this purpose, we followed three steps: (1) we manually annotated a corpus of 83 French texts with co-reference chains and anaphoric chains; (2) we applied RefGen (Longo and Todirascu, 2010; Longo, 2013), a tool that automatically identifies co-reference and anaphoric chains in French, on the same corpus ; and (3) we evaluated the discriminating power of 65 coherence and cohesion-based features to assess text readability, comparing the results obtained on the manual and automatic annotation.

3.1 Corpus Description and Annotation

The corpus used in this study is a subset of the corpus of FFL (French as a Foreign Language) texts gathered by François (2009), which includes 2,160 texts extracted from 28 FFL textbooks. All the textbooks 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 use a stratified sampling to select informative texts and narrative texts from the levels A2 to C1 (about 11 texts for each combination of level and genre).

In a second step, the corpus was annotated for co-reference chains (containing at least three mentions) and anaphoric chains (two mentions) by six human annotators, following an annotation guide. The annotation process was as follows: first, all mentions were detected, then we assigned an identification number to the chain containing the mention, finally the syntactic role as well as the type of the mention were annotated (see Figure 2 for an example of the annotation format). We used 16 different mention categories (e.g. proper names, indefinite NP, definite NP, personal pronouns, etc.) and 6 syntactic functions: S-subject, OD - direct object, OI - indirect object, CN - genitive, Mod - modifier, and X - other functions. Additionally, we annotated adverbs (*ici, là-bas*), resumptive anaphora or groups (the pronoun *ils* in Fig. 2 refers to the group composed of Antoine and Catherine).

Based on these guidelines, a common batch, composed of 10 randomly selected files, was annotated by all the annotators. It was used to identify annotation divergences between annotators[1] and to correct the annotation guide. We computed the overall inter-annotator agreement on this common batch using the mean Krippendorff's alpha on each text and we obtained 0.47, which corresponds to a *moderate* agreement between annotators. Such value is however not unusual in co-reference annotation (Muzerelle et al., 2014). Then, following the annotation guide, each expert annotated a batch of 12 texts from the corpus. At the end of the process, the principal annotator checked all batches against the guidelines, thus creating the reference for our experimentation.

[Antoine]_1/S/NPr/partie(3) fait la connaissance de [Catherine]_2/CN/NPr/partie(3). [Antoine]_1/S/NPr est [un beau parleur]_1/X/GNI et [la jeune fille]_2/S/GND [s']_2/X/Pronref intéresse à [lui]_1/OI/Pron. [Ils]_3/S/Pron vont au cinéma ensemble.

Figure 2: Example of annotated data : the number of the entity, the syntactic function and the category, eventually the relation with other referents : []_nb/syn/category/relation.

3.2 Automatic Annotation

For the automatic annotation of co-reference chains, we used a rule-based tool which identifies co-reference chains for French written texts, RefGen (Longo and Todirascu, 2010), (Longo, 2013). RefGen is one of the few systems available for French[2]. This tool integrates a POS-tagger adapted for French,

[1]Common problems that arose are: incorrect delimitation of mentions, wrong labelling of mention type or of the syntactic functions, wrong chain delimitation and relation categories (anaphoric vs co-reference).

[2]Other systems for French were proposed by Lassalle and Denis (2011) - but it detects only bridging anaphora - and by Desoyer et al. (2014), whose system detects coreference in oral data.

TTL[3](Ion, 2007), which provides the lexical category, the lemmas and simple chunk annotations (noun phrases, verb phrases). RefGen applies a set of preprocessing tools to identify complex noun phrases, named entities and impersonal pronouns.

Using information from the preprocessing step, RefGen identifies candidates for low accessibility mentions (proper nouns or named entities, definite noun phrases, indefinite noun phrases) (Ariel, 2001). These candidates open new co-reference chains. Anaphoric candidates with a high accessibility (personal pronouns, reflexive pronouns, demonstrative determiners or possessive determiners, demonstrative pronouns) are compared with possible antecedents. If the pair of candidates satisfies a complex set of syntactic, morpho-syntactic and semantic constraints, then the pair is included in a co-reference chain.

RefGen identifies almost all of the manually annotated categories, with the exception of resumptive anaphora. Concerning demonstrative NPs, the tool identifies only simple cases of antecedence (those with the same lexical head *le chien - ce chien*). Another significant drawback of this tool is that it is not able to handle complex referents such as groups or collections of objects. Adverbs are not considered as potential mentions by the tool.

3.3 Features

We replicated most features introduced in the literature described in section 2 and added new variables: the proportion of deictic pronouns, of resomptive anaphora and of adverbs, as well as the probability that a specific type of mention might open a co-reference chain in a given text. We ended up with 67 variables, divided in six classes :

1. **POS tag-based variables**: Pronouns and articles are crucial elements of cohesion. We computed 10 variables based on these parts-of-speech, namely the ratio of pronouns and nouns (1); 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); the average proportion of definite articles per sentence (8), per word (9), and the ratio of definite articles with respect to the total number of articles (10). We also computed the ratio of proper names per word (11).

2. **Lexical cohesion measures**: We replicated several methods aimed at measuring lexical cohesion in a text as the average cosine similarity between adjacent sentences. These sentences were projected either in a word space, transformed with tf-idf (term frequency-inverse document frequency) only, or in a concept space, which was obtained with LSA. We defined 6 features, taking into account various linguistic entities: the inflected forms in the texts (word overlap) (12); the lemmas (13); only the nouns, proper names, and pronouns, either through their lemmas (14), or their inflected forms (15); a token-based LSA (16) and a lemma-based LSA (17).

3. **Entity cohesion**: Mentions of co-reference chains are often found in adjacent sentences and they often have the same syntactic function as the antecedent found in the previous sentence. For example, a proper noun is the subject of sentence *n* and the anaphoric pronoun referring to it is often the subject of sentence *n+1* (" Subject to Subject" transition). However, the syntactic functions of mentions might change across sentences : the object of the sentence *n* becomes the subject of the next sentence. Drawing from Pitler and Nenkova (2008)'s work, we replicated several variables evaluating the relative frequency of the possible transitions between the four syntactic functions played by the entity in sentences *n* and *n+1* : subject (S), object (O), other complements (X), and (N) when the entity is absent in the next sentence (variables 18 to 29), but also the number of transitions (30).

4. **Entity density**: We computed the average proportion of referring entities included in co-reference chains (simple and complex noun phrases, pronouns, etc.) per document normalized by the number of words (31), the proportion of the number of entities per document normalized by the number of words (32), the proportion of unique entities per document normalized by the number of words (33), and the average number of words per entity normalized by the number of words (34).

[3]Tokenizing, Tagging and Lemmatizing free running texts

5. **Co-reference chain properties**. We included several properties of co-reference chains: the proportion of various types of mentions (variables 35–46): indefinite NP, definite NP, proper names, personal pronouns, possessive determiners, demonstrative determiners, reflexive pronouns, relative pronouns, NPs without a determiner, indefinite pronouns, demonstrative pronouns, the average length of reference chains. The proportion of the opening mentions of the co-reference chains are also computed (variables 47–57): indefinite NPs, definite NPs, proper names, NPs without a determiner, demonstrative NPs but also pronouns (personal, demonstrative, indefinite, relative), possessives. As we mentioned in section 2.2., the composition and the structure of the co-reference chains are influenced by the genres or the type of the texts. These variables are used to evaluate the correlation between text types and the various types of mentions. Additionally, for the manually annotated corpus, we count additional features such as the proportion of specific deictic pronouns (such as *en,y*) (58), the proportion of adverbs as mentions (59), the resumptive pronouns (60), complex mentions (including groups or collections) (61). We compute also the proportion of these categories being used to open a new chain (variables 62–65).

6. **Classic features** : Finally, we replicated two efficient features from the readability literature as a baseline: the mean number of word per sentence or NMP (66), which provides an indication of the syntactic complexity, and a unigram model (67), estimating the vocabulary difficulty.

4 Results

We assessed the efficiency of our cohesive features through three devices. First, we computed their Spearman correlation with the CEFR levels of the texts in our corpus (see Table 1) in order to evaluate their informativeness when considered in isolation. Second, we computed a semi-partial correlation ($sr_{k(66,67)}$) (Kerlinger and Pedhazur, 1973, 92) between the target variable and the text CEFR levels, while controlling for the two classic variables (NMP and unigram). The reason for this analysis had been put forward by Boyer (1992) who said "it is conceivable that there are relations between the surface features of the text measured by [classic] readability formula and text characteristics of higher level". Therefore, semi-partial correlation will help determine whether our variables contribute to text readability prediction with additional information besides sentence length and word frequency. Third, all significant variables as regards the semi-partial correlation have been combined within a readability model and compared with a classic readability formula. In this section, we will first discuss the efficiency of the variables on the manually annotated corpus, then on the one automatically annotated with RefGen, then modelling experiments are discussed.

4.1 Results on the Manually-annotated Corpus

First, simple variables measuring the use of pronouns and articles based on POS-tagged information are correlated with text readability (e.g. nb. of pronouns per sentence: $\rho = 0.24$; nb. of definite articles per sentence: $\rho = 0.22$). This effect was also found by Todirascu et al. (2013), but it is likely to be due to sentence length because the semi-partial correlations – when controlling for sentence length – are not significant neither for the number of pronouns per sentence ($sr = 0.14$) nor for the number of definite articles per sentence ($sr = -0.11$). Besides, the correlations for the number of pronouns ($\rho = 0.04$) and of definite articles ($\rho = 0.01$) are nonsignificant when normalized at the word level. The situation is the same on narrative and informative texts.

Interestingly, semi-partial correlation are significant for the number of pronouns per word ($sr = 0.25$) and for the number of personal pronouns ($sr = 0.23$), on all texts. The more difficult a text is, the more pronouns are used. Pronoun resolution requires background knowledge and high reading proficiency, which explains their higher frequency in difficult texts, even when text length is controlled. For comparison, Pitler and Nenkova (2008) found no effect for both variables.

There is a very interesting pattern of results for lexical coherence measures. As regards the correlation, there is a clear distinction between the four features based on word overlap (and their variation) – none of which are significant –, and the two LSA-based features, which are significant. The LSA-based feature using lemma is the second best feature after NMP on the whole corpus, while the token variant is the

very best feature for the informative texts. Such efficiency is in line with previous results (François and Fairon, 2012; Dascalu et al., 2013), but the semi-partial correlation offers a more nuanced figure, since the features based on LSA are not efficient when word frequency and sentence length are controlled. On the other hand, a more naive approach such as word overlap appears to provide more specific information as shown by the semi-partial correlations computed on informative texts ($sr = -0.41$ for lemma overlap and $sr = -0.4$ for NP word overlap).

Another interesting feature is the number of chains, which is negatively correlated with text complexity for all texts ($\rho = -0.22$) and narrative texts ($\rho = -0.35$): the lower the number of chains is (which means less referents but longer chains), the more difficult a text is. Besides, the ratio of unique entities is a valuable feature for all texts ($\rho = -0.26$) as well as for narrative texts ($\rho = -0.38$). More difficult narrative texts have a lower number of unique entities, probably because they include longer descriptions of the same elements, psychological introspection, or repetitions of the same mention. However, semi-partial correlations show that these variables are redundant with sentence length and word frequency, whereas the average word length of entity then becomes significant ($sr = -0.28$).

On the contrary, the proportion of the various syntactic transition types in a text hardly conveys information about text difficulty. Out of the 12 types of transitions, only "Object to None" is significant for all texts ($\rho = 0.24$) and for informative texts ($\rho = 0.42$). This feature means that the distance between two consecutive mentions of the same chain is larger than one sentence, a phenomenon that often occurs in informative texts where the same referent may be repeated across the text, even after several sentences. It should also be mentioned that the "Object to Object" transition was found significant ($\rho = 0.41$ and $sr = 0.40$) exclusively in narrative texts. On the whole, we are much in line with the negative results of Pitler and Nenkova (2008) as regards this category of variables.

Finally, Todirascu et al. (2013) suggested to consider the proportion of the different types of the entities and found both the proportion of pronouns and indefinite NP to be useful features. Globally, variables in this category show a poor correlation in our experiment. The type of entities that emerged as noticeable is the proportion of demonstrative NP ($\rho = 0.22$) in all texts, which nevertheless loses significance on the two sub-genre corpora as well as when sentence length and word frequency are controlled ($sr = -0.06$). It is also interesting to note that the proportion of the first mention of a chain being specific deictic pronouns is significant for all texts ($\rho = 0.22$), and even stronger when the two classic variables are controlled ($sr = 0.24$). A summary of the correlations for the most interesting features is available in Table 1.

4.2 Results on the Automatically-annotated Corpus

When comparing the manual and the automatic annotations, when relevant,[4] we find some features in which the two approaches converge such as the number of transitions, the proportion of mentions being a pronoun or a proper noun, etc. These are cases corresponding to easier phenomena to detect automatically. Conversely, some variables demonstrate large discrepancies in effectiveness between their manual and automatic versions, such as the average word length of entities, the proportion of "Object to Object" transitions, the proportion of definite mention, or the proportion of the first mention being a definite or a proper noun.

In such cases, especially in narrative texts, the automatic version appears to be more efficient, even when the semi-partial correlation is concerned. This is probably a side effect of annotation errors by RefGen, but as a result, more variables appear significant with the automatic annotation. Text complexity in narrative texts is thus correlated with the proportion of definite articles ($\rho = 0.37$) and proper nouns ($\rho = -0.39$), the proportion of chains starting with definite articles ($\rho = 0.52$) or proper noun ($\rho = -0.38$), the average word length of entities ($\rho = -0.48$) as well as with the proportion of syntactic transition "O to O" ($\rho = 0.41$). For informative texts, text difficulty is positively correlated with the proportion of transitions "O to N" ($\rho = 0.32$) and the proportion of first mention being a proper noun ($sr = 0.32$), but negatively correlated with average word length of entities ($\rho = -0.31$).

[4]Several features – those from the first, second, and sixth class in Section 3.3–, were only computed automatically. As a consequence, Table 1 provides only one value per subcorpus.

Variables	corpus (all)		corpus (narr.)		corpus (inf.)	
	manual	auto	manual	auto	manual	auto
Pronoun/sent.	0.24* / 0.14		0.32* / 0.24		0.38* / 0.16	
Pronoun/word	0.04 / 0.25*		0.22 / 0.26		0.03 / 0.17	
Pers. pron./word	-0.04 / 0.23*		0.07 / 0.23		-0.13 / 0.14	
LSA_Token	0.32** / 0.12		0.13 / 0.04		0,52*** / 0.23	
LSA_Lemma	0.28** / 0.14		0.20 / 0.01		0,43** / 0.23	
coRef_Lemma	-0.15 / -0.16		0.06 / 0		-0,4** / -0.41*	
coRef_NP_Lemma	0 / -0.11		0.25 / 0.07		0.31* / -0.4*	
nb. transitions	-0.15 / 0.12	0.10 / 0.18	-0.15 / 0.14	0.14 / 0.16	-0.17 / -0.10	0.07 / 0.08
X to N	0.12 / -0.07	0.20 / 0.21	0.26 / -0.03	0.27 / 0.3	-0.02 / -0.07	0.13 / 0.16
O to O	0.06 / 0.06	0.21 / 0.12	-0.09 / -0.06	0,41** / 0.40**	0.23 / 0.1	0,04 / -0.04
Nb. chains/words	-0.22 / -0.21	0.11 / 0.10	-0.35* / -0.33*	0.30 / 0.20	-0.11 / -0.1	-0.03 / -0.1
Nb. unique entity	-0.26* / -0.15	0.12 / 0.10	-0.38* / -0.24	0.32 / 0.21	-0.17 / -0.11	-0.04 / -0.12
Av. length of entity	-0.14 / -0.28*	-0.34** / -0.26*	-0.15 / -0.10	-0.48* / -0.36*	-0.20 / -0.31*	-0.31* / -0.23
Definite	0 / -0.26*	0.18 / -0.01	0.12 / -0.06	0.37* / 0.04	-0,20 / -0.3	0.04 / -0.05
Dem	0.22* / -0.06	NA / NA	0.21 / 0.07	NA / NA	0.2 / -0.07	NA / NA
Indefinite	0.04 / -0.05	-0.12 / -0.26*	-0.14 / -0.20	-0.11 / -0.14	0.21 / 0	-0.14 / -0.27
Pron	0.10 / 0.28*	0.11 / 0.21	0.19 / 0.25	0.18 / 0.21	0.18 / 0.28	0.07 / 0.15
Proper	-0.12 / -0.07	-0.05 / 0	-0.30 / -0.21	-0,39* / -0.37*	0 / 0.11	0.22 / 0.25
1st Definite	0.09 / -0.07	0.23* / 0.15	0.28 / 0.24	0.52*** / 0.37*	-0.17 / -0.20	0.03 / -0.05
1st PRONSPEC	0.22* / 0.24*	NA / NA	0.33* / 0.31	NA / NA	NA / NA	NA / NA
1st Proper Noun	0.02 / 0.07	-0.05 / 0	-0.07 / -0.09	-0.38* / -0.32*	0.07 / 0.25	0.24 / 0.32*
NMP	0,35** / NA		0,27 / NA		0,50** / NA	
unigram	-0,25* / NA		-0,32* / NA		-0,28 / NA	

Table 1: Spearman correlation / semi-partial correlation computed for each variable and difficulty. Significance of the correlation coefficient is indicated as follows: $^{*}: < 0.05$; $^{**}: < 0.01$; and $^{***}: < 0.001$.

4.3 Cohesion and Coherence Variables in Readability Models

In this section, the efficiency of our cohesive and coherence features for readability is tested in the context of actual readability models. On the corpus of 83 texts, we defined 4 sets of features to be used either for a classification task (with SVM classifier) or a regression task (with ϵ-SVR). The first set, that serves as a baseline, includes only sentence length (NMP) and a unigram model (ML1)[5] model have been trained. The second set includes NMP, ML1, and all variables that have been detected as significant by the correlation on the manual corpus (parsimonious_manu) and was trained on the manually annotated version of the data. The third set includes NMP, ML1 and all variables that have been detected significant by the correlation on the automatic corpus (parsimonious_auto) and was trained on the automatically annotated version of the data. The last set includes all variables (full model) and was trained on the manually annotated data, as we are interested to get the best performance possible. The optimal kernel and associated meta-parameters for all models (see Table 2) were selected via a grid-search conducted using a 10-fold cross-validation process. Once the best parameters were known, the performance of each model were then measured with two metrics – accuracy and mean average error (MAE).

Feature set	nb. variables	Model	kernel	C	Others param.	accuracy	MAE
baseline	2	SVM	RBF	5	$\gamma = 0.5$	43.6	0.89
baseline	2	SVR	polynomial	5	deg = 2; $\epsilon = 0.5$	/	1.03
parsimonious_manu	10	SVM	linear	0.1	/	43.4	0.81
parsimonious_manu	10	SVR	RBF	100	$\epsilon = 0.1$; $\gamma = 0.0001$		0.91
parsimonious_auto	8	SVM	RBF	500	$\gamma = 0.1$	41.5	0.85
parsimonious_auto	8	SVR	RBF	100	$\epsilon = 0.5$; $\gamma = 0.0001$	/	0.94
full model	67	SVM	polynomial	5	deg. = 2	40.6	0.89
full model	67	SVR	RBF	5	$\epsilon = 1$; $\gamma = 0.01$	/	0.93

Table 2: Accuracy and values of meta-parameters for the 4 models.

First, all classification models perform better than their regression counterparts. However, even for the former, no model using coherence or cohesive features is able to overcome a simple model based

[5] As those variables have been automatically computed, the results are the same for both versions of the corpus (manually and automatically annotated).

on sentence length and word frequency. The one that performs better is the SVM parsimonious model based on the manual annotation (MAE = 0.81), but compared to the SVM baseline (MAE = 0.89), the difference is not significant using a paired T-test ($t = -1.43; p = 0.19$). It is also interesting to note that using automatically detected features seems to slightly degrade performance compared to the manual annotation, although such difference is clearly not significant with a paired T-test ($t = -0.78; p = 0.45$). This is the case even though automatically-computed variables were characterized by slightly better correlations as found in Section 4.2.

5 Discussion and Conclusion

To conclude, we have performed a detailed analysis of 65 cohesive features commonly used in the readability literature. The parameterization of these variables requires heavy NLP processing and is prone to errors. We showed that nevertheless they do not seem to contribute much to the prediction of text readability when compared with simple predictors such as word frequency and sentence length. On the one hand, 6 features only were found to be significant by semi-partial correlation (when sentence length and word frequency were controlled for). On the other hand, integrating the best cohesive features in a readability model did not bring significant improvement over a simple baseline on our French data. The first lesson learned is that such kind of features, although quite popular in the literature, have an efficiency that is subject to caution, at least in the context of readability prediction, as it was already reported by some of the previous studies.

Another interesting insight of our analysis is the use of semi-partial correlation to analyze the efficiency of variables for readability. Previously, some authors (Pitler and Nenkova, 2008; François and Fairon, 2012; Todirascu et al., 2013) only used Pearson or Spearman correlations to identify and quantify the effect of a text characteristic on readability and we showed that, as was suggested by Boyer (1992), higher textual dimensions can be much correlated with lexical or syntactic features. A good example in this regard was the impact of LSA-based features. Similar to previous studies (François and Fairon, 2012; Dascalu et al., 2013), we found a large effect for this predictor, which completely vanished once word frequency and sentence length were controlled for. This allowed us to reconcile to some extent contradictory findings in this regard.

Our experiments also showed large differences between the manual and automatic annotation of lexical chain properties, which seems to lead to a loss of performance when such predictors are included into a full readability model. This should however be replicated using different co-reference extraction tools, as some of the errors are typical of the RefGen tool that we used.

Finally, the third question that we planned to investigate was whether the behavior of lexical and co-reference chains differs in narrative and informative texts, in relation to text difficulty. We noticed that the variables significantly correlated with difficulty vary depending on the genre of texts. On narrative texts, the number of chains, the number of unique entities or the ratio of first mention being a specific deictic pronoun were relevant, whereas the average word length of entities, the LSA-based features and the word overlap features were relevant for informative texts.

However, there are some limitations to our study and further investigation would be necessary before discarding co-reference chain-based features for readability. First, we have experimented on an L2 corpus, while the cohesive aspects might be more relevant for L1 texts. Moreover, the study was performed on French and the results might vary from one language to another (although our findings are mostly in line with results on English). Finally, it is not excluded that some properties of the lexical and co-reference chains that we did not consider (e.g. mean distance in words between the various entities of a chain) could demonstrate a stronger discriminative power.

References

M. Ariel. 2001. Accessibility theory: An overview. In T. Sanders, J. Schliperoord, and Spooren W., editors, *Text representation. Human cognitive processing series*, pages 29–87. John Benjamins.

R. Barzilay and M. Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

I. Berzlnovich and G. Redeker. 2012. Genre-dependent interaction of coherence and lexical cohesion in written discourse. *Corpus Linguistics and Linguistic Theory*, (8):183–208.

J.R. Bormuth. 1969. *Development of Readability Analysis*. Technical report, Projet n.7-0052, U.S. Office of Education, Bureau of Research, Department of Health, Education and Welfare, Washington, DC.

J.Y. Boyer. 1992. La lisibilité. *Revue française de pédagogie*, 99(1):5–14.

S. Carter-Thomas. 1994. Langue de spécialité : cohésion, culture et cohrence. *ASp*, 5-6:61–67.

J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

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

E. Dale and J.S. Chall. 1948. A formula for predicting readability. *Educational research bulletin*, 27(1):11–28.

M. Dascalu, P. Dessus, Ş. Trausan-Matu, M. Bianco, and A. Nardy. 2013. Readerbench, an environment for analyzing text complexity and reading strategies. In *Artificial Intelligence in Education*, pages 379–388. Springer.

A. Desoyer, F. Landragin, I. Tellier, A. Lefeuvre, and J.-Y. Antoine. 2014. Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ancor. *TAL*, 55:97–121.

L. Feng, N. Elhadad, and M. Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237.

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

M. Flor, B. Beigman Klebanov, and K. M. Sheehan. 2013. Lexical tightness and text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 29–38.

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

T. François and C. Fairon. 2013. Les apports du TAL à la lisibilité du français langue étrangère. *Traitement Automatique des Langues (TAL),*, 54(1):171–202.

T. François. 2009. Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. In *Proceedings of the 12th Conference of the EACL : Student Research Workshop*, pages 19–27.

T. François and C. Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*, pages 466–477.

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

B. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

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

G. Hirst and D. St-Onge. 1998. Lexical chains as representation of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet:An Electronic Lexical Database and Some of its Applications*. The MIT Press, Cambridge, MA.

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

R. Ion. 2007. *Word Sense Disambiguation Methods Applied to English and Romanian*. Ph.D. thesis, Romanian Academy, Bucharest.

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

F.N. Kerlinger and E.J. Pedhazur. 1973. *Multiple regression in behavioral research*. Holt, Rinehart and Winston, New York.

W. Kintsch and D. Vipond. 1979. Reading comprehension and readability in educational practice and psychological theory. In L.G. Nilsson, editor, *Perspectives on Memory Research*, pages 329–365. Lawrence Erlbaum, Hillsdale, NJ.

W. Kintsch, E. Kozminsky, W.J. Streby, G. McKoon, and J.M. Keenan. 1975. Comprehension and recall of text as a function of content variables1. *Journal of Verbal Learning and Verbal Behavior*, 14(2):196–214.

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

E. Lassalle and P. Denis. 2011. Leveraging different meronym discovery methods for bridging resolutionin french. In *Anaphora Processing and Applications, Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*.

L. Longo and A. Todirascu. 2010. Genre-based reference chains identification for french. *Investigationes Linguisticae*, 21:57–75.

L. Longo and A. Todirascu. 2014. Vers une typologie des chaînes de référence dans des textes administratifs et juridiques. *Langages. Les chanes de référence*, pages 79–98.

L. Longo. 2013. *Vers des moteurs de recherche intelligents : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence*. Ph.D. thesis, University of Strasbourg.

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

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

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

J.R. Miller and W. Kintsch. 1980. Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory*, 6(4):335–354.

J. Muzerelle, A. Lefeuvre, E. Schang, J.-Y. Antoine, A. Pelletier, D. Maurel, I. Eshkol, and J. Villaneau. 2014. Ancor_centre, a large free spoken french coreference corpus: Description of the resource and reliability measures.

E. Pitler and A. Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.

M. Poesio, R. Stevenson, B. DiEugenio, and J. Hitzeman. 2004. Centering : A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

C. Schnedecker. 1997. Nom propre et chaînes de référence. *Recherches Linguistiques*, 21.

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

A. Todirascu, T. François, N. Gala, C. Fairon, A.-L. Ligozat, and D. Bernhard. 2013. Coherence and cohesion for the assessment of text readability. *Natural Language Processing and Cognitive Science*, pages 11–19.

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

B.L. Zakaluk and S.J. Samuels. 1988. Toward a New Approach to Predicting Text Comprehensibility. In B.L. Zakaluk and S.J. Samuels, editors, *Readability: Its Past, Present and Future*, pages 121–144. International Reading Association, Newark, Delaware.