Sentence Complexity in French: a Corpus-Based Approach

Ludovic Tanguy, Nikola Tulechki

Intelligent Information Systems 2009
ISBN 666-666-666, pages 1–14

Sentence Complexity in French: a Corpus-Based Approach

Ludovic Tanguy and Nikola Tulechki

CLLE (Cognition, Langues, Langage, Ergonomie - UMR 5263)
University of Toulouse and CNRS, France

Abstract

Language complexity is a notion widely used in a number of linguistic fields and language applications, and can be described by a number of linguistic features and practical measures. This work proposes a closer, data-oriented look at sentence complexity. Starting from a number of different studies, we selected and implemented 52 linguistic features and measured them on a corpus of varied French texts. Using statistical methods, we identify five underlying dimensions of sentence complexity. In addition to providing a better understanding of the phenomenon, these dimensions have been used in some information retrieval experiments.

Keywords: NLP, complexity, linguistics, corpus, information retrieval

1 Introduction

The measure of language complexity has been investigated by many different fields of (computational) linguistics. These studies often lead to the definition of linguistic features to quantify the morphological, lexical or syntactic sophistication of a sentence or a text.

The work presented here proposes to compare different approaches to sentence complexity, and to identify the most important features. As it is illusory to search for a single measure encompassing all aspects of complexity, we instead developed an approach to identify the different dimensions of this phenomenon. We based our work on a 2.4 million words French corpus of varied texts, and used a large set of features, selected from different fields.

We first gathered from various domains of application the different features proposed (§ 2). Starting from this first set of 52 features, we developed specific NLP programs in order to measure the corresponding values on our corpus (§ 3). We then applied correlation measures to detect redundancy between these features, and selected a smaller set of 21 features, giving precedence to simpler features in terms of required NLP techniques. The resulting 21 features, and their application to our corpus have been submitted to a principal components analysis, resulting in the identification of 5 main factors accounting for the variations across our corpus (§ 4).

In addition to providing a new insight on language, and a practical method for comparing sentences and texts in terms of complexity, the resulting dimensions

have been used in a study focusing on information retrieval. The combined features associated to these factors can actually be used for the detection of query difficulty, with higher efficiency than any individual feature (§ 5).

2 Overview of sentence complexity

In this section we will first give a short presentation of the different fields and applications interested by sentence complexity. For each of the selected research areas, we will present why and how complexity has been investigated. In a second time, we will give an overview of the different linguistic features we gathered.

2.1 Fields and applications concerned with sentence complexity

2.1.1 Readability metrics

Readability metrics is probably the field most concerned with sentence complexity. The notion of *readability* or *reading ease* is based on the fact that differences in writing style produce texts that require more or less attention, persistence and reading skill in order to be fully understood. To prevent writers from producing difficult texts, publishers have developed a number of formulae that produce a numerical score reflecting the reading ease of a given text, which is often expressed as the grade (as in the number of years in the educational system) required to fully understand the text.

The most commonly used formulae are the Automated Readability Index, Coleman-Liau Index, Flesch Reading Ease, Flesch Kincaid Grade Level, Gunning Fox index and SMOG index (see DuBay (2004) for an overview). The majority of these formulae rely on the average word length and the average sentence length to produce the score, and are thus easily calculated.

2.1.2 Psycholinguistics

Research in the fields of psycholinguistics and developmental psychology has yielded a variety of linguistically valid scales and metrics of complexity. A common way of building them is by observing language as it is acquired by young children. Assuming that simpler syntactic constructions appear earlier than more complex ones, a score of complexity can be associated with them based on order of acquisition. Examples of such scales are DSS (Lee (1974)), D-Level (Rosenberg and Abbeduto (1987)) and IPSyn (Scarborough (1990)). Estimating the complexity for a given text requires the identification and counting of particular syntactic structures and the computing of a score according to a fixed score sheet. Even though these scales were initially designed for manual analysis only, with the advance of NLP, some of them have recently been automated (Sagae *et al.* (2005); Voss (2005)).

2.1.3 Controlled Languages

Another field concerned with sentence complexity is that of controlled language design and validation. A Controlled Language (CL) is "*a language standard with*

restricted grammar, style, and vocabulary, whose general goal is to simplify or clarify the text by rendering it less ambiguous and more predictable" (Brown (2006)). Most often used in industrial contexts where information circulates between individuals with different backgrounds and/or non-native speakers, a CL standard insures that texts are written in the most understandable fashion. For English, the most widely used standard is Simplified Technical English (STE). Many tools exist for automatic CL validation (CL checkers), that can identify unauthorised language uses; some of them also provide a rewriting of non-valid segments (Mitamura *et al.* (2003)).

A CL checker consists of two main components: a dictionary of approved words and a grammar of linguistic restrictions. This second component is a list of rules intended to minimise ambiguity and complexity. Such restriction can aim at limiting the size and nature of syntactic constituents, or forbidding the use of some tenses or verbal constructs.

2.1.4 Automatic text summarisation

The goal of automatic text summarisation (ATS) is "*to take a document as input, extract information content from it, and present the most important content to the user in a condensed form in a manner sensitive to the user's or application's needs*" (Brown (2006)). A major field of NLP for quite a long time, research in ATS has produced systems based on a variety of methods. Some of them take syntactic complexity in account: the underlying idea being that complex sentences often contain non-crucial information, and that identifying and removing those parts may result in reduction of size without significant loss of informational content. Resulting from such approaches are scales of importance of the different syntactic constituents (Monod and Prince (2005)). These scales can be seen as measures of syntactic complexity, the least important constituents being those most contributing to an increased complexity.

2.1.5 Syntactic simplification

Somewhat similar to summarisation, but without the objective of size reduction, is syntactic simplification. This is the "*process of reducing the grammatical complexity of a text, while retaining its information content and meaning*" (Siddharthan (2006)). Syntactic simplifiers are useful for a number of applications: easying understanding by people suffering from language disabilities, adapting the display of a text to limited number of characters, reducing information load for readers involved in a complex task.

An automatic syntactic simplifier has first to identify the sentences that need simplification and, secondly, to generate a simplified version of those sentences, while not affecting the overall coherence of the text. Although the bulk of the research in this area is struggling with the second phase, in this study, we are only interested with the first one; a simplifiable sentence is, by definition, a complex one. Therefore, automatic identification of such sentences is based on features of sentence complexity.

2.1.6 Information retrieval

In Information Retrieval (IR), an increased complexity of the indexed documents or the queries can be associated with a decrease in the performance of the system. Accurately predicting document or query complexity in order to selectively trigger further processing of more complex zones may therefore have an impact on overall performance. One of the many aspects of complexity involved in IR is sentence complexity and research has been done to measure it automatically in an IR-improvement perspective (Mothe and Tanguy (2005)). This point will be further developed in § 5.1. Similar work has been done, but this time on the documents retrieved, in order to detect if more complex documents are considered more relevant (Karlgren (1996)).

2.2 First features collection

Our approach for collecting usable features from the previous fields has been the following:

1. identify a complex linguistic phenomenon and/or direct measure from the works cited in the previous sections;
2. select the ones that can be automatically calculated, given the processing tools available (see below);
3. adapt these features to French, if needed.

Starting from more than 100 candidate complexity measures, the resulting set contains a total of 52 features.

The simpler features are those used in basic readability metrics: *average word length*, in terms of *characters* or *syllables*, *sentence length* in terms of *words* or *characters* and *part-of-speech (POS) category counts* are examples of such features, along with the classic *readability measures* mentioned above. Following the same technical ideas, features coming from the fields of text summarisation or simplification focus on sentence structures such as *coordination*, *subordination*, or the *number of verbs* in the sentence.

More complex features are issued from the definition of controlled languages. Based on the notion of norm violation, these features count the particular syntactic constructions that are not accepted in controlled languages. Some examples are *noun phrases containing three or more nouns* and *verbs conjugated in non-basic tenses* (i.e. other than infinitive, present, imperfect, present perfect, imperative and future). Although CL features come from recommendations for English, they could easily be adapted to French.

Features issued from research in psycholinguistics are also based on counting more or less complex syntactic constructions, such as *clauses conjoined with a coordinating conjunction* and *nominalisations serving as subject of main verb.*

We also included features based on the overall syntactic structure, proposed for IR by Mothe and Tanguy (2005):

- *depth of the parse tree*: this feature represents "vertical" complexity and corresponds to the maximum number of nested syntactic constituents in the sentence;

- *syntactic link span*: representing "horizontal" complexity, this feature is the average distance between words linked by syntactic relations.

Finally, we added a few simple features from our own intuition, such as the number of *punctuation marks*, the *ratio of function words*, and number of *numerals* (the latter being used as a more *stylistic* feature, able to identify some kind of specific sentences, such as titles, or containing formulas, addresses or similar data).

2.3 Related work

Research involving large sets of heterogeneous linguistic features has been done for a variety of purposes.

2.3.1 Improving readability metrics

Coh-Metrix[1] (Graesser *et al.* (2004)) is a project aimed at designing a valid and operational readability metric, replacing the over-simplistic, but widely used classic metrics. Difficulty can be seen as partly resulting from the lack of *cohesion* and *coherence* in texts. These are multifaceted phenomena operating both locally (at the syntactic level) and globally (at the text level). Furthermore a number of different conceptual categories can be distinguished. Referential, temporal, locational, causal and structural coherence involve various linguistic cues, ranging from pronoun and determiner type, through different types of adverbials and syntactic structures to semantic similarity between words. Automatically assessing these phenomena thus involves a large number of features and measures. Correlating them with the results of experiments on actual human-comprehension allows to establish weights for individual features and thus design a formula producing a single readability score.

2.3.2 Variation in language

Douglas Biber and his colleagues (Biber *et al.* (1998); Conrad and Biber (2001)) have developed corpus-driven methods in their studies of *variation*. Variation results from a number of both conscious and unconscious choices made by speakers and writers on the form of the texts they produce. Accounting of variation in language, Biber argues, is only possible by empirical analysis of large collections of natural texts, while keeping track of a number of contextual factors.

One way to explain variation is to focus on a large number of simple linguistic features and to identify patterns of their distribution across texts in the collection based on their individual frequencies. Due to the size of the collection and the number of features, manual analysis is impossible and one must turn to statistical methods developed for data analysis, such as Factor Analysis or PCA (§ 4.1). Biber has developed a step-by-step methodology for performing such studies extending from feature selection, through corpus design, to the interpretation of the results. Our study on sentence complexity is based on this methodology, but focuses on different phenomena.

[1] `http://cohmetrix.memphis.edu/cohmetrixpr/index.html`

3 Corpus and processing

The next phase of our work was to apply the 52 selected features to a large corpus, and to observe their variation and correlation.

3.1 Corpus overview

Due to lack of a readily available general corpus for the French language, we had to design one. By including texts from as many different registers as possible, differing both with regard to their intended public and their purpose, our aim was to maximise internal variation. This resulted in a 2.4 million word corpus composed as follows (the parts being of approximately equal sizes):

- Fiction: three novels[2] and a collection of 70 tales by the Grimm brothers;
- News: articles from daily newspapers *Le Monde* and *L'Est Républicain*;
- Recipes: cooking recipes;
- Technical: reports of medical interventions and law articles;
- Pedagogical: two handbooks of mechanics and geology;
- Comments: users' comments on TV series collected from dedicated websites.

Overall, this corpus contains 121,636 sentences. Each of these sentences has been pre-processed by state-of-the-art NLP tools before being used as a base for calculating our selected linguistic features.

3.2 Processing tools

In order to count features we first analysed the corpus with two generic NLP tools: a part-of-speech tagger and an syntactic parser.

For part-of-speech tagging we used TreeTagger[3], a tool which provides a single morphological category for each word in the text, based on decision trees.

For syntactic analysis, we used Syntex (Bourigault (2007)), a robust and efficient syntactic parser for French. Syntex is a dependency parser, which means that as a result the words are linked together with a set of syntactic relations. It can be used to identify syntactic constituents (phrases, clauses), build a syntactic tree, or simply identify the function of a word in a sentence.

Both word- and sentence-level tokenisation are done by the parser, thus our notion of word and sentence is dependant on these treatments. Of course, none of these tools is error-free, and their performance has an impact on the results. However, as we will discuss below, we tried to limit these implications by favouring more reliable methods when given a choice.

Finally, the nominalisations were extracted from the Verbaction[4] morphological database.

[2] *Les Noces Barbares* by Yann Queffélec, *Du coté de chez Swann* by Marcel Proust, *Le Mystère Frontenac* by François Mauriac

[3] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[4] `http://w3.erss.univ-tlse2.fr/verbaction/main.html`

3.3 First result: smaller set of features

Based on the automatic annotations provided by TreeTagger and Syntex, we computed the values of our initial 52 features for each of the 121,636 sentences in the corpus. Our first goal here was to detect redundancy between these features, in order to reduce their number and to come out with a set of uncorrelated variables for a statistical analysis. This was done in three steps.

1. As could be foreseen, many of the selected features have a high correlation with sentence length. Such is the case of most readability metrics, and for all variations of this notion of length (words, characters, syllables, etc.). Our first move was to reject all features highly correlated to the number of words in a sentence, with a linear correlation coefficient (Pearson's ρ) greater than 0.8.
2. In a second time, we isolated clusters of highly correlated features, which occur when several features are simply different means of measuring the same phenomenon. In this case, when correlation between two features was greater than 0.7, we simply rejected the most difficult one to compute. This allowed us to put aside more complicated syntactical features that relied heavily on the parser's results, and were more error-prone than their simpler counterparts.
3. Finally, we rejected unproductive features, which occurred too rarely to be taken into account in a statistical approach. Features such as "verbal coordination" or "infinitive tense in subject position" focused on uncommon phenomena and yielded positive scores for less than 15% of sentences.

The resulting 21 features are listed and described in table 1, along with sample corresponding features that have been filtered out because of a high correlation.

This set of features is a first step in the identification of the underlying structures of sentence complexity. They can be used by themselves to describe or compare sentences, but remain too numerous to be easily interpreted in conjunction. The next step in our study is to apply a dimensionality reduction technique in order to get an even smaller set of descriptors.

4 Dimensions of complexity

Using the 21 features previously described, we applied a Principal Components Analysis (PCA) in order to identify the main dimensions of sentence complexity.

4.1 Principal components analysis

The general goal of this data analysis method is to represent vectors initially represented in a space of N dimensions into a smaller space (see for example Baayen (2008) for further details). Principal Component Analysis reduces data dimensionality into spaces which are the most important as determined by the eigen values of the covariance matrix. The eigen vectors are then known to be the most useful to visualise the maximum of information. Moreover, the most specific information will be the first displayed.

Our input is a 121,636x21 matrix, containing the raw scores obtained by each linguistic feature for each sentence. Values were centered and reduced (i.e linearly

Feature	**Description**	**Redundant features**
WORD_LENGTH	average number of characters per word	number of syllables, CLI readability index
NUMERALS	proportion[5] of numerals in the sentence	
SENT_LENGTH	number of words in the sentence	other readability measures, number of different POS, number of syntactic links
FUNCTION_WORDS	proportion of function words	
NODE_RATIO	proportion of identified syntactic links	
COORDINATIONS	number of coordinating conjunctions	adjectival or nominal coordinations
PREP_A-DE	proportion of "*à*" and "*de*" prepositions	
PRONOUN_INCIDENCE	incidence[6] of personal pronouns	
SYNT_SPAN	average distance (number of words) between two syntactically linked words	
OBJECT_LENGTH	average length of the verb object(s)	
SUBORDINATIONS	number of subordinating conjunctions	
COMPLEX_TENSE	number of "complex" tenses (for French, other than present, present perfect, imperfect, future, and infinitive)	number of different tenses
VERB_SATURATION	average number of constituents syntactically linked to a verb	
SYNT_DEPTH	depth of the parse tree	
PREP_OTHER	proportion of prepositions other than "*à*" and "*de*"	
SUBJECT_LENGTH	average length (number of words) of the verbal subject	
COMMAS	proportion of commas	
NOMINALISATION	number of nominalisations	nominalised subjects
NP_INCIDENCE	incidence of noun phrases	average NP size, number of nouns per NP
INFINITIVE	proportion of verbs in the infinitive mood	
RELATIVE_PRONOUN	number of relative pronouns	

Table 1: Selected 21 Features

modified in order for each variable to have a mean of zero and a standard deviation of 1).

From the different results provided by a PCA, we focused on the description of the eigen vectors, or factors. Each resulting factor is described through coordinates in the initial variable space (i.e. features). As all resulting factors are orthogonal, they are uncorrelated to each other, and can therefore be considered as the main *dimensions* of our data space. Factors are also naturally ordered: they account for a decreasing part of the total variance. Hence, only the first few factors are significant. The main interpretation of each factor will be, at first, to identify which combinations of features are best suited to explain the variations in our corpus. The following section will present the factors in detail.

[5] The proportion of X is the ratio between the number of X and the number of words.

[6] The incidence of X is the ratio between the number of words and the number of X in the sentence.

4.2 Description and interpretation of resulting factors

Although PCA produces a number of factors equal to the rank of the matrix (21), only the first 5 account for more than 5% of the total variance.

Each factor is described by coefficients (or *loadings*) for each initial variable. Thus, features with high absolute loadings for a given factor can be used to interpret it, by studying its position in the initial data space. Table 2 shows the loadings for each of the first 5 factors for each feature (absolute values greater than 0.3 are shown in boldface).

Feature	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
WORD_LENGTH	0.06	-0.12	-0.19	**0.68**	-0.25
NUMERALS	0.12	0.09	-0.17	**-0.55**	0.07
SENT_LENGTH	**0.35**	0.17	-0.27	-0.01	-0.05
FUNCTION_WORDS	0.17	0.26	**0.46**	-0.02	0
NODE_RATIO	**0.32**	-0.26	0.19	0	0.09
COORDINATIONS	0.23	0.19	-0.18	0	0
PREP_A-DE	0.12	**-0.30**	-0.22	**-0.33**	-0.02
PRONOUN_INCIDENCE	0.08	**0.35**	**0.42**	-0.06	0.02
SYNT_SPAN	**0.30**	-0.12	0.19	-0.01	0.01
OBJECT_LENGTH	0.24	-0.22	0.07	-0.03	0.23
SUBORDINATIONS	0.21	**0.32**	-0.05	0.03	-0.08
COMPLEX_TENSE	0.21	0.27	-0.12	0.05	-0.10
VERB_SATURATION	0.26	-0.06	**0.34**	-0.02	-0.22
SYNT_DEPTH	**0.34**	-0.19	-0.13	-0.05	0.06
PREP_OTHER	0.09	-0.13	0.07	0.28	**0.45**
SUBJECT_LENGTH	0.16	-0.22	0.09	-0.06	**-0.60**
COMMAS	0.01	0.12	-0.18	-0.11	-0.24
NOMINALISATION	0.28	0.11	-0.25	0.04	-0.09
NP_INCIDENCE	0.23	**-0.32**	-0.05	-0.06	0.01
INFINITIVE	0.21	0.17	-0.20	0.09	**0.41**
RELATIVE_PRONOUN	0.20	0.26	-0.12	0.05	-0.06

TABLE 2: Factor Loadings

Each factor is described more precisely in the following paragraphs. The associated percentage corresponds to the part of the total variance explained by the factor.

4.2.1 Factor 1: Sentence length (26%)

The first factor has high positive scores with features related to sentence length. This simply means that the number of words (and thus of syntactic links) in a sentence is the main source of variation across the sentences in our study.

Sample sentence with high positive score: The longest sentence in our corpus contains 102 words and is not reproduced here for practical reasons)

Sample sentence with high negative score: *Saler, poivrer.*[7]

[7] *Salt, pepper* (both infinitive verbs).
In this translation and the following, we use rough literal transposition in order to preserve the global syntactic structure of the sentences.

This was naturally expected: long sentence are intuitively complex, and several features take sentence length into account, such as syntactic depth and span. However, isolating this simple dimension is a necessary step in the study of the others.

4.2.2 Factor 2: Nominal vs. verbal (10.5%)

The second factor is much more interesting and enlightening than the first. A sentence with a high positive score on this factor is likely to contain subordinate or relative clauses, and many function words and verbal constructs. On the contrary, a sentence with a high negative score will have large noun phrases and prepositions. This can be interpreted as the duality between nominal and verbal status of sentences. This can be more easily seen in the following examples.

Sample sentence with high positive score: *Le roi et ses chasseurs voyant le bel animal se mirent à sa poursuite, mais ne purent l'encercler, et lorsqu'ils pensèrent y être parvenus, il bondit et disparut dans les taillis et disparut.*[8]

Sample sentence with high negative score: *Ces facteurs contradictoires conduisent à un freinage de l'augmentation de la puissance et des coefficients de sécurité, par la limitation de la masse du moteur à une valeur acceptable.*[9]

Although being of roughly the same length, the first sentence mainly consists of an enumeration of short verbal clauses, while the second has extremely long noun phrases and only one verb. Both can thus be said to be complex (as they both have high scores for this factor), but for completely different reasons.

4.2.3 Factor 3: Syntactic connexity (8.4%)

The third factor is more difficult to interpret, but can be seen as an indication of the connexity of the constituents the sentence. This factor has high positive loadings on function words, pronouns, and verb saturation, which means that a sentence with a high score on this factor has many syntactic links with fewer words. High negative loadings are observed for long sentences with nominalisations, infinitives and commas. A better insight can be reached by looking at the following examples:

Sample sentence with high positive score: *Vous marcherez immédiatement après nous...*[10]

Sample sentence with high negative score: *Cette manifestation répondait à l'appel lancé, une semaine plus tôt, par 550 personnalités d'Europe, Russie comprise, intitulé « tapis rouge, silence et crime », à l'initiative notamment d'André Glucksmann.*[11]

[8] *The king and his hunters, seeing the beautiful animal, started to pursue it but could not surround it, and when they thought they had succeeded, it leaped away and fled into the thickets and disappeared.*

[9] *These contradictory factors lead to a slowing down in the increase in power and security coefficients, through the limitation of the engine weight to an acceptable value.*

[10] *You will walk closely behind us...*

[11] *This demonstration was an answer to the call, a week earlier, by 550 European personalities, including Russia, named "red carpet, silence and crime", initiated namely by André Glucksmann.*

The first sentence is what can be seen as a canonical and highly connected sentence, while the second contains a quote, appositions, and adverbial clauses, all leading to disconnected syntactical parts. Although interesting, this factor has to be carefully interpreted, as it is heavily dependent on the parser's results.

4.2.4 Factor 4: Lexical complexity (5.9%)

The fourth factor can simply be interpreted as a measure of the *lexical* complexity of a sentence. The very high positive loading with the word length, and high negative loadings with short words (numerals and prepositions) are self-explanatory. This factor is notably uncorrelated with all other syntactical features, as all other loadings are equal or close to zero.

Sample sentence with high positive score: *En cas d'insuffisance tricuspidienne fonctionnelle, réversibilité temporaire sous traitement médical.*[12]

Sample sentence with high negative score: *La tête de la vis sans fin sert de logement à deux pistons de soupape 9 et 10 perpendiculaires à l'axe de la vis.*[13]

Although both phrases belong to a technical genre, the second one uses notably shorter words (*vis*, *fin*, *axe*).

4.2.5 Factor 5: Subject complexity (5.8%)

The fifth factor, although accounting for only 5.8% of total variance, is interesting, as it focuses on a specific part of a sentence: the subject. High negative loadings for this factor are related to subject length, while positive loadings are observed for infinitive and *other* prepositions. A better insight can be obtained by looking at sample sentences with high scores on this factor.

Sample sentence with high positive score: *Préparer les autres éléments, écaler et couper les oeufs durs, peler les tomates et les couper en rondelles, peler les échalotes et les hacher, épépiner les poivrons et les émincer finement, couper les olives et les coeurs de palmier en rondelles, peler l'orange à vif et la couper en rondelles.*[14]

Sample sentence with high negative score: *Un ensemble : filtre-régulateur de pression lubrificateur brouillard d'huile, monté sur la tuyauterie de raccordement du groupe d'unités et à proximité de celui-ci, assurera ces fonctions.*[15]

As can be seen in these examples, high negative scores are obtained for sentences with a very complex subject (in the second example, it covers 90% of the sentence), while sentences with no subjects at all (i.e. infinitive or imperative main verbs) get high positive scores.

[12] *In case of functional tricuspidian insuffisance, temporal reversibility under medical treatment.*

[13] *The head of the worm is used as a slot for two valve pistons 9 and 10 perpendicular to the axle of the screw.*

[14] *Prepare the other ingredients, peel and cut the hard-boiled eggs, peel the tomatoes and slice them, peel the shallots and mince them, remove the seeds from the peppers and mince them, slice the olives and the palm hearts, peel the orange and slice it.*

[15] *A set consisting of a filter regulating the lubricating oil pressure, mounted on the pipe linking the group of units, and close to it, will ensure these functions.*

4.3 Other results

The five dimensions described above shed a broader light on what the wide notion of complexity covers. Most readability metrics, for example, took only the first and third factors into account, and controlled languages mostly address issues corresponding to factors 2 and 4. In addition to the interpretation of factor loadings, we have also been able to observe that different text genres (i.e. the subparts of our corpus) are positioned in distinctly different areas of the resulting 5-dimension space, thus meeting some of Biber's conclusions from the sentence level.

5 Application to Information Retrieval

Although the five dimensions described above are very interesting as a new way to describe sentence complexity in a linguistics point of view, we believe that they can also be used for specific NLP tasks. We will describe in this section how it can be useful for an information retrieval (IR) task to study the relation between the text of a query and the results of an IR system. The objective can be either to predict a difficult query, to identify which language phenomena are problematic to a system or, more ambitiously, to apply different IR techniques depending on the linguistic characteristics of the query.

5.1 Previous studies

Although most of the studies in this area focus on IR-specific features (mostly based on query terms frequency and distribution), or semantic aspects (mostly polysemy), some (Mothe and Tanguy, 2005; Mandl and Womser-Hacker, 2002; Moreau *et al.*, 2007) have used linguistic features.

These works have two objectives: the first is to predict the query difficulty, a task which has known increased interest in the last few years. The second is to build adaptive IR systems, in which different techniques are used to process different queries. In both cases, linguistic features are used as descriptive variables for queries, and the nature and quality of these features are of course essential. All experiments rely on the data and results of previous IR evaluation campaigns (TREC and CLEF).

Mothe and Tanguy (2005) and Mandl and Womser-Hacker (2002) have measured correlation between linguistic features and the average precision and/or recall obtained by different IR systems over sets of queries. The following conclusions have been reached:

- longer queries (i.e. which higher number of words) and queries with proper nouns lead to better results;
- syntactic distance and syntactic depth are negatively correlated to both recall and precision;
- for some campaigns, prepositions, conjunctions and complex (suffixed) words have a negative correlation with either recall or precision.

Although correlation values were generally low, these results indicated a link between some linguistic complexity of queries and the lower performance of an IR system, although the complexity is not related to simple length.

5.2 Sentence complexity in monolingual CLEF campaigns

Following the same approach, we compared individual features and complexity factors as predictors of a query difficulty.

We used past CLEF monolingual French results (years 2001 to 2004, 50 queries each year), and computed the correlation between average precision and each score corresponding to our individual 21 features. It appears that:

- some of the previous results were confirmed: query length is positively correlated to precision, and syntactic depth and span are negatively correlated;
- however, all correlation values are quite low and are found in the -0.2 – +0.2 interval.

We then computed the corresponding scores for each query on the 5 factors, and again calculated the correlation between these 5 scores and the average precision. We found that:

- Factor 2 (nominal vs. verbal) is positively correlated to precision, meaning that nominal queries are more difficult that verbal ones;
- Factor 5 (subject complexity) is positively correlated to precision, meaning that queries with complex subjects are more difficult.

On the whole, significant correlation scores are higher for factors 2 and 5 than for individual features, including the ones with high loadings. Compared to previous results, these factors are more precise to identify queries where the formulation makes use of nominalisations and complex NPs. These queries are apparently more difficult to process by an IR system as the information is more often expressed in documents through verbal structures. These first results are encouraging, and we will investigate the concrete use of our complexity factors in adaptive IR systems.

6 Conclusion and further work

The work presented here is a first step in the investigation of sentence complexity in NLP. Starting from several different approaches to this wide and fuzzy linguistic phenomenon, we propose a small set of automatically measurable features that can be used to characterise sentences and texts.

Furthermore, we managed to identify some of the dimensions of sentence complexity, which give a better insight on this phenomenon. Of course, many other linguistic aspects of complexity have been ignored, being too specific for a small corpus, or needing more sophisticated techniques to be taken into account.

The practical results can now be more thoroughly evaluated. This evaluation can be performed with a psycholinguistics perspective, as it has been done for some of the initial complexity features, such as readability metrics. But we also propose a practical evaluation, by using the complexity factors in NLP tasks. This allows us to get back to the initial research areas who addressed complexity in the first place. Our first experiment in information retrieval for measuring query difficulty is promising: this aspect will be more thoroughly investigated and applied to other languages.

References

R.H. Baayen (2008), *Analyzing Linguistic Data*, Cambridge University Press.

D. Biber, S. Conrad, and R. Reppen (1998), *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge University Press.

D. Bourigault (2007), Un analyseur syntaxique opérationnel: Syntex, *Research Report: CNRS & Université de Toulouse-Le Mirail.*

K. Brown, editor (2006), *Encyclopedia of Language and Linguistics second Edition*, Elsevier.

S. Conrad and D. Biber (2001), *Variation in English: Multi-Dimensional Studies*, Longman Pub Group.

W.H. DuBay (2004), The Principles of Readability, *Costa Mesa, CA: Impact Information.*

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 and Computers*, 36(2):193–202.

J. Karlgren (1996), Stylistic variation in an information retrieval experiment, in *Proceedings of the NeMLaP-2 Conference.*

L. L. Lee (1974), *Developmental Sentence Analysis*, Northwestern Univ. Press.

T. Mandl and C. Womser-Hacker (2002), Linguistic and Statistical Analysis of the CLEF Topics, in *proceedings of the CLEF 2002 Workshop.*

T. Mitamura, K. Baker, E. Nyberg, and D. Svoboda (2003), Diagnostics for interactive controlled language checking, in *Proceedings of 4th Controlled Language Applications Workshop.*

M. Y. Monod and V. Prince (2005), Automatic Summarization Based on Sentence Morpho-Syntactic Structure: Narrative Sentences Compression, in *proceedings of NLUCS*, pp. 161–167.

F. Moreau, V. Claveau, and P. Sébillot (2007), Combining linguistic indexes to improve the performances of information retrieval systems: a machine learning based solution, in *proceeding of the eighth RIAO conference*, Pittsburgh.

J. Mothe and L. Tanguy (2005), Linguistic Features to Predict Query Difficulty - a Case Study on Previous TREC campaigns, in *Proceeding of the SIGIR workshop on predicting query complexity*, p. 7–10.

S. Rosenberg and L. Abbeduto (1987), Indicators of Linguistic Competence in the Peer Group Conversational Behavior of Mildly Retarded Adults., *Applied Psycholinguistics*, 8:19–32.

K. Sagae, A. Lavie, and B. MacWhinney (2005), Automatic Measurement of Syntactic Development in Child Language, *Ann Arbor*, 100.

H. S. Scarborough (1990), Index of Productive Syntax., *Applied Psycholinguistics*, 11:1–22.

A. Siddharthan (2006), Syntactic Simplification and Text Cohesion, *Research on Language & Computation*, 4(1):77–109.

STE (2007), Simplified Technical English, Aerospace and Defence Industries Association of Europe. Specification report.

M. J. Voss (2005), Determining Syntactic Complexity Using Very Shallow Parsing, *Artificial Intelligence Center The University of Georgia. CASPR Research Report.*