A Quantitative Insight into the Impact of Translation on Readability

Alina Maria Ciobanu, Liviu P. Dinu
Center for Computational Linguistics, University of Bucharest
Faculty of Mathematics and Computer Science, University of Bucharest
alina.ciobanu@my.fmi.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

In this paper we investigate the impact of translation on readability. We propose a quantitative analysis of several shallow, lexical and morpho-syntactic features that have been traditionally used for assessing readability and have proven relevant for this task. We conduct our experiments on a parallel corpus of transcribed parliamentary sessions and we investigate readability metrics for the original segments of text, written in the language of the speaker, and their translations.

1 Introduction

Systems for automatic readability assessment have been studied since the 1920s and have received an increasing attention during the last decade. Early research on readability assessment focused only on shallow language properties, but nowadays natural language processing technologies allow the investigation of a wide range of factors which influence the ease which a text is read and understood with. These factors correspond to different levels of linguistic analysis, such as the lexical, morphological, semantic, syntactic or discourse levels. However, readability depends not only on text properties, but also on characteristics of the target readers. Aspects such as background knowledge, age, level of literacy and motivation of the expected audience should be considered when developing a readability assessment system. Although most readability metrics were initially developed for English, current research has shown a growing interest in other languages, such as German, French, Italian or Portuguese.

Readability assessment systems are relevant for a wide variety of applications, both human- and machine-oriented (Dell'Orletta et al., 2011). Second language learners and people with disabilities or low literacy skills benefit from such systems, which provide assistance in selecting reading material with an appropriate level of complexity from a large collection of documents – for example, the documents available on the web (Collins-Thompson, 2011). Within the medical domain, the investigation of the readability level of medical texts helps developing well-suited materials to increase the level of information for preventing diseases (Richwald et al., 1989) and to automatically adapt technical documents to various levels of medical expertise (Elhadad and Sutaria, 2007). For natural language processing tasks such as machine translation (Stymne et al., 2013), text simplification (Aluisio et al., 2010), speech recognition (Jones et al., 2005) or document summarization (Radev and Fan, 2000), readability approaches are employed to assist the process and to evaluate and quantify its performance and effectiveness.

1.1 Related Work

Most of the traditional readability approaches investigate shallow text properties to determine the complexity of a text. These readability metrics are based on assumptions which correlate surface features with the linguistic factors which influence readability. For example, the average number of characters or syllables per word, the average number of words per sentence and the percentage of words not occurring among the most frequent *n* words in a language are correlated with the lexical, syntactic and, respectively, the semantic complexity of the text. The Flesch-Kincaid measure (Kincaid et al., 1975) employs the average number of syllables per word and the average number of words per sentence to assess readability, while the Automated Readability Index (Smith and Senter, 1967) and the Coleman-Liau metric (Coleman and Liau, 1975) measure word length based on character count rather than syllable count; they are func-

Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014, pages 104–113,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

tions of both the average number of characters per word and the average number of words per sentence. Gunning Fog (Gunning, 1952) and SMOG (McLaughlin, 1969) account also for the percentage of polysyllabic words and the Dale-Chall formula (Dale and Chall, 1995) relies on word frequency lists to assess readability. The traditional readability approaches are not computationally expensive, but they are only a coarse approximation of the linguistic factors which influence readability (Pitler and Nenkova, 2008). According to Si and Callan (2001), the shallow features employed by standard readability indices are based on assumptions about writing style that may not apply in all situations.

Along with the development of natural languages processing tools and machine learning techniques, factors of increasing complexity , corresponding to various levels of linguistic analysis, have been taken into account in the study of readability assessment. Si and Callan (2001) and Collins-Thompson and Callan (2004) use statistical language modeling and Petersen and Ostendorf (2009) combine features from statistical language models, syntactic parse trees and traditional metrics to estimate reading difficulty. Feng (2009) explores discourse level attributes, along with lexical and syntactic features, and emphasizes the value of the global semantic properties of the text for predicting text readability. Pitler and Nenkova (2008) propose and analyze two perspectives for the task of readability assessment: prediction and ranking. Using various features, they reach the conclusion that only discourse level features exhibit robustness across the two tasks. Vajjala and Meurers (2012) show that combining lexical and syntactic features with features derived from second language acquisition research leads to performance improvements.

Although most readability approaches developed so far deal with English, the development of adequate corpora for experiments and the study of readability features tailored for other languages have received increasing attention. For Italian, Franchina and Vacca (1986) propose the Flesch-Vacca formula, which is an adaptation of the Flesch index (Flesch, 1946). Another metric developed for Italian is Gulpease (Lucisano and Piemontese, 1988), which uses characters instead of syllables to measure word length and thus requires less resources. Dell'Orletta et al. (2011) combine traditional, morpho-syntactic, lexical and syntactic features for building a readability model for Italian, while Tonelli et al. (2012) propose a system for readability assessment for Italian inspired by the principles of Coh-Metrix (Graesser et al., 2004). For French, Kandel and Moles (1958) propose an adaptation of the Flesch formula and François and Miltsakaki (2012) investigate a wide range of classic and non-classic features to predict readability level using a dataset for French as a foreign language. Readability assessment was also studied for Spanish (Huerta, 1959) and Portuguese (Aluisio et al., 2010) using features derived from previous research on English.

1.2 Readability of Translation

According to Sun (2012), the reception of a translated text is related to cross-cultural readability. Translators need to understand the particularities of both the source and the target language in order to transfer the meaning of the text from one language to another. This process can be challenging, especially for languages with significant structure differences, such as English and Chinese. The three-step system of translation (analysis, transfer and restructuring) presented by Nida and Taber (1969) summarizes the process and emphasizes the importance of a proper understanding of the source and the target languages. While rendering the source language text into the target language, it is also important to maintain the style of the document. Various genres of text might be translated for different purposes, which influence the choice of the translation strategy. For example, for political speeches the purpose is to report exactly what is communicated in a given text (Trosborg, 1997).

Parallel corpora are very useful in studying the properties of translation and the relationships between source language and target language. Therefore, the corpus-based research has become more and more popular in translation research. Using the *Europarl* (Koehn, 2005) parallel corpus, van Halteren (2008) investigates the automatic identification of the source language of European Parliament speeches, based on frequency counts of word n-grams. Islam and Mehler (2012) draw attention to the absence of adequate corpora for studies on translation and propose a resource suited for this purpose.

2 Our Approach and Methodology

The problem that we address in this paper is whether human translation has an impact on readability. Given a text T_1 in a source language L_1 and its translations in various target languages $L_2, ..., L_n$, how does readability vary? Is the original text in L_1 easier to read and understand than its translation in a target language L_i? Which language is closest to the source language, in terms of readability? We investigate several shallow, lexical and morpho-syntactic features that have been widely used and have proven relevant for assessing readability. We are interested in observing the differences between the feature values obtained for the original texts and those obtained for their translations. Although some of the metrics (such as average word length) might be language-specific, most of them are language-independent and a comparison between them across languages is justified. The 10 readability metrics that we account for are described in Section 3.2.

We run our experiments on *Europarl* (Koehn, 2005), a multilingual parallel corpus which is described in detail in Section 3.1. We investigate 5 Romance languages (Romanian, French, Italian, Spanish and Portuguese) and, in order to excerpt an adequate dataset of parallel texts, we adopt a strategy similar to that of van Halteren (2008): given n languages $L_1, ..., L_n$, we apply the following steps:

1. we select L_1 as the source language
2. we excerpt the collection of segments of text T_1 for which L_1 is the source language
3. we identify the translations $T_2, ..., T_n$ of T_1 in the target languages $L_2, ..., L_n$
4. we compute the readability metrics for $T_1, ..., T_n$
5. we repeat steps $1 - 4$ using each language $L_2, ..., L_n$ as the source language, one at a time

We propose two approaches to quantify and evaluate the variation in the readability feature values from the original texts to their translations: a distance-based method and a multi-criteria technique based on rank aggregation.

3 Experimental Setup

3.1 Data

Europarl (Koehn, 2005) is a multilingual parallel corpus extracted from the proceedings of the European Parliament. Its main intended use is as aid for statistical machine translation research (Tiedemann, 2012). The corpus is tokenized and aligned in 21 languages. The files contain annotations for marking the document (<*chapter*>), the speaker (<*speaker*>) and the paragraph (<*p*>). Some documents have the attribute *language* for the *speaker* tag, which indicates the language used by the original speaker. Another way of annotating the original language is by having the language abbreviation written between parentheses at the beginning of each segment of text. However, there are segments where the language is not marked in either of the two ways. We account only for sentences for which the original language could be determined and we exclude all segments showing inconsistent values.

We use the following strategy: because for the Romance languages there are very few segments of text for which the *language* attribute is consistent across all versions, we take into account an attribute *L* if all other Romance languages mention it. For example, given a paragraph *P* in the Romanian subcorpus, we assume that the source language for this paragraph is Romanian if all other four subcorpora (Italian, French, Spanish and Portuguese) mark this paragraph *P* with the tag *RO* for language. Thus, we obtain a collection of segments of text for each subcorpus. We identify 4,988 paragraphs for which Romanian is the source language, 13,093 for French, 7,485 for Italian, 5,959 for Spanish and 8,049 for Portuguese. Because we need sets of approximately equal size for comparison, we choose, for each language, a subset equal with the size of the smallest subset, i.e., we keep 4,988 paragraphs for each language.

Note that in this corpus paragraphs are aligned across languages, but the number of sentences may be different. For example, the sentence *"UE trebuie să fie ambiţioasă în combaterea schimbărilor climatice, iar rolul energiei nucleare şi energiilor regenerabile nu poate fi neglijat."*[1], for which Romanian is the source language,

[1] Translation into English: *"The EU must be ambitious in the battle against climate change, which means that the role of nuclear power and renewable energy sources cannot be discounted."*

is translated into French in two sentences: *"L'UE doit se montrer ambitieuse dans sa lutte contre les changements climatiques."* and *"L'énergie nucléaire et les sources d'énergie renouvelables ne peuvent donc pas être écartées."*. Therefore, we match paragraphs, rather than sentences, across languages.

As a preprocessing step, we discard the transcribers' descriptions of the parliamentary sessions (such as *"Applause"*, *"The President interrupted the speaker"* or *"The session was suspended at 19.30 and resumed at 21.00"*).

According to van Halteren (2008), translations in the European Parliament are generally made by native speakers of the target language. Translation is an inherent part of the political activity (Schäffner and Bassnett, 2010) and has a high influence on the way the political speeches are perceived. The question posed by Schäffner and Bassnett (2010) *"What exactly happens in the complex processes of recontextualisation across linguistic, cultural and ideological boundaries?"* summarizes the complexity of the process of translating political documents. Political texts might contain complex technical terms and elaborated sentences. Therefore, the results of our experiments are probably domain-specific and cannot be generalized to other types of texts. Although parliamentary documents probably have a low readability level, our investigation is not negatively influenced by the choice of corpus because we are consistent across all experiments in terms of text gender and we report results obtained solely by comparison between source and target languages.

3.2 Features

We investigate several shallow, lexical and morpho-syntactic features that were traditionally used for assessing readability and have proven high discriminative power within readability metrics.

3.2.1 Shallow Features

Average number of words per sentence. The average sentence length is one of the most widely used metrics for determining readability level and was employed in numerous readability formulas, proving to be most meaningful in combined evidence with average word frequency. Feng et al. (2010) find the average sentence length to have higher predictive power than all the other lexical and syllable-based features they used.

Average number of characters per word. It is generally considered that frequently occurring words are usually short, so the average number of characters per word was broadly used for measuring readability in a robust manner. Many readability formulas measure word length in syllables rather than letters, but this requires additional resources for syllabication.

3.2.2 Lexical Features

Percentage of words from the basic lexicon. Based on the assumption that more common words are easier to understand, the percentage of words not occurring among the most frequent *n* in the language is a commonly used metric to approximate readability. To determine the percentage of words from the basic lexicon, we employ the representative vocabularies for Romance languages proposed by Sala (1988).

Type/Token Ratio. The proportion between the number of lexical types and the number of tokens indicates the range of use of vocabulary. The higher the value of this feature, the higher the variability of the vocabulary used in the text.

3.2.3 Morpho-Syntactic Features

Relative frequency of POS unigrams. The ratio for 5 parts of speech (verbs, nouns, pronouns, adjectives and adverbs), computed individually on a per-token basis. This feature assumes that the probability of a token is context-independent. For lemmatization and part of speech tagging we use the *DexOnline*[2] machine-readable dictionary for Romanian and the *FreeLing*[3] (Padró and Stanilovsky, 2012; Padró, 2011; Padró et al., 2010; Atserias et al., 2006; Carreras et al., 2004) language analysis tool suite for French, Italian, Spanish and Portuguese.

Lexical density. The proportion of content words (verbs, nouns, adjectives and adverbs), computed on a per-token basis. Grammatical features were shown to be useful in readability prediction (Heilman et al., 2007).

4 Results Analysis

Our main purpose is to investigate the variability of the feature values from the original texts to their translations. In Table 1 we report the values

[2] http://dexonline.ro

[3] http://nlp.lsi.upc.edu/freeling

obtained for 10 readability metrics computed for the *Europarl* subcorpora for Romanian, French, Italian, Spanish and Portuguese. The readability metrics we computed lead to several immediate remarks. We notice that, generally, when representing the values for a feature *F* on the real axis, the values corresponding to the translations are not placed on the same side of the value corresponding to the original text. For example, considering feature *F3* (the percentage of words from the basic lexicon), and taking Romanian as the source language, we observe that the value for the original text is between Italian (on the left side) and the other languages (on the right side).

In the absence of a widely-accepted readability metric, such as the Flesch-Kincaid formula or the Automated Readability Index, for all 5 Romance languages, we choose two other ways to evaluate the results obtained after applying the 10 readability features: a distance-based evaluation and a multi-criteria approach.

In order to compute distance measures reliably, we normalize feature values using the following formula:

$$f_i' = \frac{f_i - f_{min}}{f_{max} - f_{min}},$$

where f_{min} is the minimum value for feature *F* and f_{max} is the maximum value for feature *F*. For example, if *F* = *F1* and the source language is Romanian, then f_{min} = 26.2 and f_{max} = 29.0.

4.1 Preliminaries

In this subsection we shortly describe the two techniques used. The experimented reader can skip this subsection.

4.1.1 Rank Aggregation

Rank distance (Dinu and Dinu, 2005) is a metric used for measuring the similarity between two ranked lists. A ranking of a set of n objects can be represented as a permutation of the integers $1, 2, ..., n$. S is a set of ranking results, $\sigma \in S$. $\sigma(i)$ represents the rank of object i in the ranking result σ. The rank distance is computed as:

$$\Delta(\sigma, \tau) = \sum_{i=1}^{n} |\sigma(i) - \tau(i)|$$

The ranks of the elements are given from bottom up, i.e., from n to 1, in a Borda order. The elements which do not occur in any of the rankings receive the rank 0.

In a selection process, rankings are issued for a common decision problem, therefore a ranking that "combines" all the original (base) rankings is required. One common-sense solution is finding a ranking that is as close as possible to all the particular rankings.

Formally, given m partial rankings $\mathcal{T} = \tau_1, \tau_2, ..., \tau_m$, over a universe $\mathcal{U}$, the rank aggregation problem requires a partial ranking that is as close as possible to all these rankings to be determined. In other words, it requires a means of combining the rankings. There are many ways to solve this problem, one of which is by trying to find a ranking such that the sum of rank distances between it and the given rankings is minimal. In other words, find σ such that:

$$\Delta(\sigma, \mathcal{T}) = \sum_{\tau \in \mathcal{T}} \Delta(\sigma, \tau)$$

is minimal. The set of all rankings that minimize $\Delta(\sigma, \mathcal{T})$ is called the aggregations set and is denoted by $agr(\mathcal{T})$.

Apart from many paradoxes of different aggregation methods, this problem is NP-hard for most non-trivial distances (e.g., for edit distance, see (de la Higuera and Casacuberta, 2000)). Dinu and Manea (2006) show that the rank aggregation problem using rank distance, which minimizes the sum $\Delta(\sigma, \mathcal{T})$ of the rank distances between the aggregation and each given ranking, can be reduced to solving $|\mathcal{U}|$ assignment problems, where $\mathcal{U}$ is the universe of objects. Let $n = \#\mathcal{U}$. The time complexity to obtain one such aggregation (there may be more than one) is $\mathcal{O}(n^4)$.

We then transform the aggregation problem in a categorization problem as follows (Dinu and Popescu, 2008): for a multiset *L* of rankings, we determine all the aggregations of *L* and then we apply voting on the set of *agr(L)*.

4.1.2 Cosine Distance

Cosine distance is a metric which computes the angular cosine distance between two vectors of an inner product space. Given two vectors of features, *A* and *B*, the cosine distance is represented as follows:

$$\Delta(A, B) = 1 - \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

When used in positive space, the cosine distance ranges from 0 to 1.

Source	Target	Features									
Language	Language	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
RO	RO	26.2	5.61	0.67	0.06	0.66	0.15	0.29	0.16	0.05	0.11
	FR	29.0	5.06	0.79	0.03	0.59	0.13	0.35	0.06	0.04	0.06
	IT	27.4	5.57	0.63	0.04	0.61	0.16	0.30	0.10	0.04	0.06
	ES	28.3	5.18	0.81	0.04	0.53	0.15	0.24	0.09	0.03	0.03
	PT	26.8	5.31	0.78	0.04	0.58	0.14	0.30	0.08	0.04	0.02
FR	RO	24.6	5.35	0.70	0.06	0.64	0.17	0.26	0.14	0.06	0.13
	FR	27.4	4.86	0.81	0.04	0.58	0.14	0.32	0.05	0.06	0.09
	IT	25.7	5.46	0.65	0.05	0.61	0.17	0.28	0.09	0.05	0.07
	ES	26.3	5.11	0.82	0.05	0.53	0.16	0.23	0.08	0.04	0.04
	PT	25.1	5.21	0.80	0.05	0.58	0.16	0.29	0.07	0.05	0.02
IT	RO	29.7	5.46	0.69	0.06	0.62	0.16	0.27	0.15	0.05	0.12
	FR	32.4	5.00	0.80	0.04	0.58	0.14	0.33	0.06	0.05	0.08
	IT	30.9	5.48	0.64	0.05	0.61	0.16	0.28	0.10	0.05	0.07
	ES	31.8	5.15	0.82	0.04	0.53	0.16	0.23	0.09	0.04	0.03
	PT	30.5	5.28	0.79	0.04	0.58	0.15	0.29	0.07	0.05	0.02
ES	RO	27.6	5.33	0.70	0.06	0.64	0.17	0.26	0.14	0.06	0.13
	FR	29.9	4.91	0.81	0.04	0.58	0.14	0.32	0.05	0.05	0.09
	IT	27.9	5.45	0.66	0.05	0.60	0.17	0.28	0.09	0.05	0.08
	ES	31.1	5.02	0.83	0.05	0.52	0.16	0.22	0.08	0.05	0.04
	PT	28.2	5.17	0.81	0.05	0.57	0.16	0.28	0.07	0.05	0.02
PT	RO	29.3	5.58	0.67	0.05	0.65	0.15	0.28	0.16	0.05	0.12
	FR	32.8	5.04	0.80	0.03	0.58	0.13	0.34	0.06	0.04	0.07
	IT	30.9	5.56	0.62	0.04	0.60	0.15	0.29	0.10	0.04	0.06
	ES	32.5	5.15	0.81	0.03	0.53	0.15	0.24	0.09	0.03	0.03
	PT	30.9	5.28	0.79	0.04	0.57	0.14	0.30	0.08	0.04	0.02

Table 1: Values for readability metrics applied on *Europarl*. The first column represents the source language (the language of the speaker). The second column represents the target language (the language in which the text is written / translated). The features F1 - F10 are as follows:

- F1 - average number of words per sentence
- F2 - average number of characters per word
- F3 - percentage of words from the basic lexicon
- F4 - type / token ratio
- F5 - lexical density
- F6 - relative frequency of POS unigrams: verbs
- F7 - relative frequency of POS unigrams: nouns
- P8 - relative frequency of POS unigrams: adjectives
- F9 - relative frequency of POS unigrams: adverbs
- F10 - relative frequency of POS unigrams: pronouns

	RO	FR	IT	ES	PT
RO	–	0.571	0.138	0.582	0.292
FR	0.513	–	0.505	0.491	0.328
IT	0.075	0.416	–	0.502	0.212
ES	0.531	0.423	0.545	–	0.256
PT	0.300	0.227	0.252	0.275	–

Table 2: Cosine distance between feature vectors. The first column represents the source language and the first line represents the target language.

4.2 Experiment Analysis: Original vs. Translation

Our main goal is to determine a robust way to evaluate the variation in readability from the original texts to their translations, after applying the 10 readability features described in Section 3.2.

A natural approach is to use an evaluation methodology based on a distance metric between feature vectors to observe how close translations are in various languages, with respect to readability. The closer the distance is to 0, the more easily can one language be translated into the other, in terms of readability. Briefly, our first approach is as follows: for each source language *L* in column 1 of Table 1, we consider the feature vector corresponding to this language from column 2 and we compute the cosine distance between this vector and all the other 4 vectors remaining in column 2, one for each target language. The obtained values are reported in Table 2, on the line corresponding to language *L*.

Table 2 provides not only information regarding the closest language, but also the hierarchy of languages in terms of readability. For example, the closest language to Romanian is Italian, followed by Portuguese, French and Spanish. Overall, the lowest distance between an original text and its translation occurs when Italian is the source language and Romanian the target language. The highest distance is reported for translations from Romanian into Spanish.

The second approach we use for investigating the readability of translation is multi-criteria aggregation: since the 10 monitored features can be seen as individual classifiers for readability (and in various papers they were used either individually or combined as representative features for predicting readability), we experiment with a multi-criteria aggregation of these metrics in order to predict which language is closest to the source language in terms of readability.

For segments of text having the source language *L*, we consider each feature F_i, one at a time, and we compute the absolute value of the difference between the F_i value for the original text and the F_i values for its translations. Then, we sort the values in ascending order, thus obtaining for each language *L* and feature F_i a ranking with 4 elements (one for each translation) determined as follows: the language having the lowest computed absolute value is placed on the first position, the language having the second to lowest computed absolute value is placed on the second position, and so on. Finally, we have, for each language *L*, 10 rankings (one for each feature) with 4 elements (one for each translation), each ranking indicating on the first position the target language which is closest to the source language with regard to readability measured by feature F_i. In case of equal values for the computed absolute distance, we consider all possible rankings.

Given these rankings, the task we propose is to determine which target language is closest to the source language in terms of readability. To solve this requirement, we apply multi-criteria aggregation based on rank distance. For each language, we aggregate the 10 corresponding rankings and determine the closest language with respect to readability across translation. The results we obtain for Romance languages after the rank aggregation are as follows: the closest translation language for Romanian is Italian (followed by Portuguese, Spanish and French). Conversely, for Italian the closest language is Romanian (followed by Portuguese, French and Spanish). For French, Portuguese occupies the first position in the ranking (followed by Spanish, Italian and Romanian). For Spanish, Portuguese ranks first (followed by Italian, French and Romanian), while for Portuguese, Italian is the closest language (followed by French, Spanish and Romanian).

The obtained results are very similar to those computed by the cosine distance and reported in Table 2. The only difference regarding the closest language in terms of readability is that rank aggregation reports Italian as being closest to Portuguese, while the cosine distance reports French instead. However, the differences between the first two ranked languages for Portuguese, namely French and Italian, are insignificant.

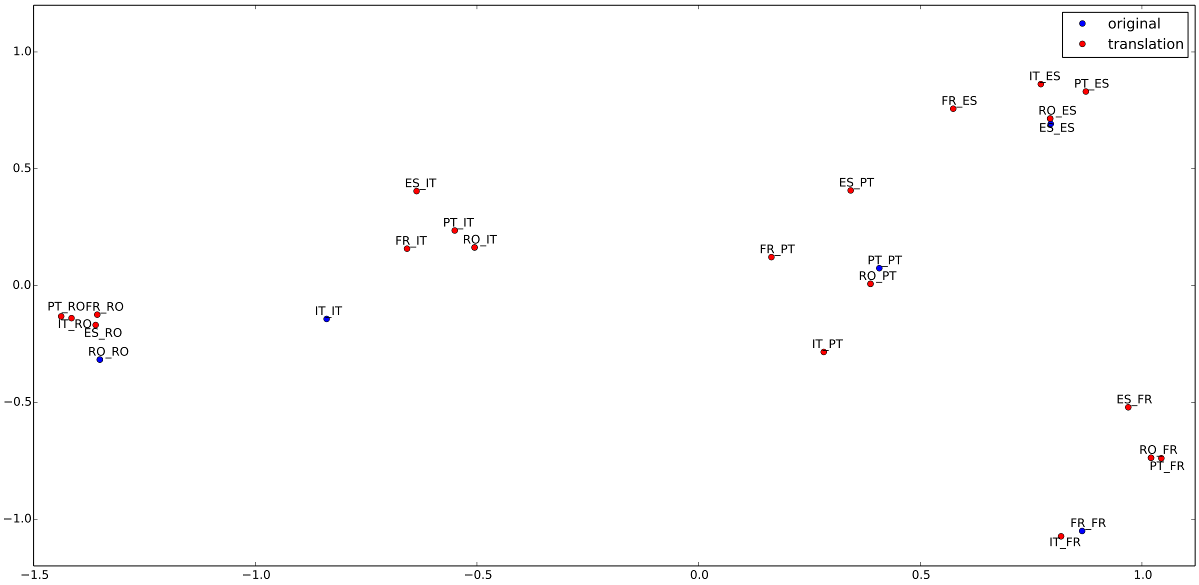

Figure 1: PCA. Languages are annotated in the figure as follows: $L_1\_L_2$, where L_1 is the source language and L_2 is the target language.

4.3 PCA: Original vs. Translation

In Figure 1 we employ Principal Component Analysis (PCA) to perform linear data reduction in order to obtain a better representation of the readability feature vectors without losing much information. We use the Modular toolkit for Data Processing (MDP), a Python data processing framework (Zito et al., 2008). We observe that clusters tend to be formed based on the target language. rather than based on the source language. While for Romanian and Italian the original texts are to some extent isolated from their translations, for French, Spanish and Portuguese the original texts are more integrated within the groups of translations. The most compact cluster corresponds to Romanian as a target language.

5 Conclusions

In this paper we investigate the behaviour of various readability metrics across parallel translations of texts from a source language to target languages. We focus on Romance languages and we propose two methods for the analysis of the closest translation, in terms of readability. Given a text in a source language, we determine which of its translations in various target languages is closest to the original text with regard to readability. In our future works, we plan to extend our analysis to more languages, in order to cover a wider variety of linguistic families. We are mainly interested in the 21 languages covered by *Europarl*. Moreover, we intend to enrich the variety of the texts, beginning with an analysis of translations of literary works. As far as resources are available, we plan to investigate other readability metrics as well and to combine our findings with the views of human experts. We believe our method can provide valuable information regarding the difficulty of translation from one language into another in terms of readability.

Acknowledgements

The authors thank the anonymous reviewers for their helpful and constructive comments. The contribution of the authors to this paper is equal. Research supported by a grant of the Romanian National Authority for Scientific Research, CNCS UEFISCDI, project number PN-II-ID-PCE-2011-3-0959.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability Assessment for Text Simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, IUNLPBEA 2010*, pages 1–9.

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In *Proceedings of* *the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 2281–2286.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, pages 239–242.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Kevyn Collins-Thompson and James P. Callan. 2004. A Language Modeling Approach to Predicting Reading Difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL 2004*, pages 193–200.

Kevyn Collins-Thompson. 2011. Enriching Information Retrieval with Reading Level Prediction. In *SIGIR 2011 Workshop on Enriching Information Retrieval*.

Edgar Dale and Jeanne Chall. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

C. de la Higuera and F. Casacuberta. 2000. Topology of Strings: Median String is NP-complete. *Theoretical Computer Science*, 230(1-2):39–48.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies, SLPAT 2011*, pages 73–83.

Anca Dinu and Liviu P. Dinu. 2005. On the Syllabic Similarities of Romance Languages. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2005*, pages 785–788.

Liviu P. Dinu and Florin Manea. 2006. An Efficient Approach for the Rank Aggregation Problem. *Theoretical Computer Science*, 359(1):455–461.

Liviu P. Dinu and Marius Popescu. 2008. A Multi-Criteria Decision Method Based on Rank Distance. *Fundamenta Informaticae*, 86(1-2):79–91.

Noemie Elhadad and Komal Sutaria. 2007. Mining a Lexicon of Technical Terms and Lay Equivalents. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, BioNLP 2007*, pages 49–56.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A Comparison of Features for Automatic Readability Assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING 2010*, pages 276–284.

Lijun Feng. 2009. Automatic Readability Assessment for People with Intellectual Disabilities. *SIGACCESS Access. Comput.*, (93):84–91.

Rudolf Flesch. 1946. *The Art of plain talk*. T. Harper.

Thomas François and Eleni Miltsakaki. 2012. Do NLP and Machine Learning Improve Traditional Readability Formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR 2012*, pages 49–57.

Valerio Franchina and Roberto Vacca. 1986. Adaptation of Flesch readability index on a bilingual text written by the same author both in Italian and English languages. *Linguaggi*, 3:47–49.

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

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill; Fouth Printing edition.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL 2007*, pages 460–467.

F. Huerta. 1959. Medida sencillas de lecturabilidad. *Consigna*, 214:29–32.

Zahurul Islam and Alexander Mehler. 2012. Customization of the Europarl Corpus for Translation Studies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 2505–2510.

Douglas Jones, Edward Gibson, Wade Shen, Neil Granoien, Martha Herzog, Douglas Reynolds, and Clifford Weinstein. 2005. Measuring Human Readability of Machine Generated Text: Three Case Studies in Speech Recognition and Machine Translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2005*, pages 1009–1012.

L. Kandel and A. Moles. 1958. Application de l'indice de Flesch a la langue française. *Cahiers Etudes de Radio-Television*, 19:253–274.

J. Peter Kincaid, Lieutenant Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading*

Ease formula) for Navy enlisted personnel. Research Branch Report, Millington, TN: Chief of Naval Training.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

Pietro Lucisano and Maria Emanuela Piemontese. 1988. Gulpease. una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e Città*, 39:110–124.

G. Harry McLaughlin. 1969. Smog grading: A new readability formula. *Journal of Reading*, 12(8):639–646.

Eugene A. Nida and Charles R. Taber. 1969. *The Theory and Practice of Translation*. Leiden: E.J. Brill.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 2473–2479.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. FreeLing 2.1: Five Years of Open-source Language Processing Tools. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, pages 931–936.

Lluís Padró. 2011. Analizadores Multilingües en FreeLing. *Linguamatica*, 3(2):13–20.

Sarah E. Petersen and Mari Ostendorf. 2009. A Machine Learning Approach to Reading Level Assessment. *Computer Speech and Language*, 23(1):89–106.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2008*, pages 186–195.

Dragomir R. Radev and Weiguo Fan. 2000. Automatic Summarization of Search Engine Hit Lists. In *Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, RANLPIR 2000*, pages 99–109.

Gary A. Richwald, Margarita Schneider-Mufnoz, and R. Burciaga Valdez. 1989. Are Condom Instructions in Spanish Readable? Implications for AIDS Prevention Activities for Hispanics. *Hispanic Journal of Behavioral Sciences*, 11(1):70–82.

Marius Sala. 1988. *Vocabularul Reprezentativ al Limbilor Romanice*. Editura Academiei, Bucureşti.

Christina Schäffner and Susan Bassnett. 2010. Politics, Media and Translation - Exploring Synergies. In *Political Discourse, Media and Translation*, pages 1–29. Newcastle upon Tyne: Cambridge Scholars Publishing.

Luo Si and Jamie Callan. 2001. A Statistical Model for Scientific Readability. In *Proceedings of the 10th International Conference on Information and Knowledge Management, CIKM 2001*, pages 574–576.

E.A. Smith and R.J. Senter. 1967. Automated readability index. *Wright-Patterson Air Force Base. AMRL-TR-6620*.

Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical Machine Translation with Readability Constraints. In *Proceedings of the 19th Nordic Conference on Computational Linguistics, NODALIDA 2013*, pages 375–386.

Yifeng Sun. 2012. Translation and strategies for cross-cultural communication. *Chinese Translators Journal*, 33(1):16–23.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 2214–2218.

Sara Tonelli, Ke Tran Manh, and Emanuele Pianta. 2012. Making Readability Indices Readable. In *Proceedings of the 1st Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR 2012*, pages 40–48.

Anna Trosborg, editor. 1997. *Text Typology and Translation*. Benjamins Translation Library.

Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification Using Insights from Second Language Acquisition. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP*, pages 163–173.

Hans van Halteren. 2008. Source Language Markers in EUROPARL Translations. In *Proceedings of the 22nd International Conference on Computational Linguistics, COLING 2008*, pages 937–944.

Tiziano Zito, Niko Wilbert, Laurenz Wiskott, and Pietro Berkes. 2008. Modular toolkit for Data Processing (MDP): a Python data processing frame work. *Front. Neuroinform.*, 2(8).