Translationese Traits in Romanian Newspapers: A Machine Learning Approach

Iustina Ilisei[1] and Diana Inkpen[2]

[1] Research Institute in Information and Language Processing,
University of Wolverhampton
Wulfruna Street, Wolverhampton WV1 1LY, United Kingdom
iustina.ilisei@gmail.com
[2] School of Information Technology and Engineering,
University of Ottawa,
800, King Edward Street, Ottawa, ON, K1N 6N5, Canada
diana@site.uOttawa.ca

Abstract. This paper presents a machine learning approach to the investigation of the translationese effect on Romanian newspapers texts. The aim is to train a learning system to distinguish between translated and non-translated texts. The classifiers achieve an accuracy well above the chance level, the results confirming the existence of translationese manifestation. Also, the experiments investigate whether there are any traits of the simplification universal within translated text. The learning system is enhanced with features previously proposed to stand for this universal, and their impact on the learning model is assessed.

Key words: Translationese, Machine Learning, Translation Studies

1 Introduction

Beginning in the eighties, certain studies noted certain characteristics exhibited by translated language compared to non-translated texts. These unnatural 'fingerprints' suspected to be characteristic of translated texts were first described by Gellerstam and the effect was called translationese [1]. In the nineties, the topic was studied intensely and the translation universals theory was proposed by [2, 3] describing four hypotheses: simplification, explicitation, transfer, and convergence. Translated language is believed to manifest certain universal features, as a consequence of the translation process, translated texts presenting their own specific lexico-grammatical and syntactic characteristics [4–6]. Toury enriched this theory by adopting a different view and by proposing two laws of translation: the law of standardisation and the law of interference [7]. Laviosa continued this line of research by proposing features for the simplification universal in a corpus-based study [8].

However, the existence of translation universals continues to be a polemical and highly debated issue in translation studies. While some scholars, like [9, 10], claim the universality aspect of the proposed hypotheses, others [11] emphasise that the value of these assumptions stands in their explanatory power. Despite some evidence of the existence of such tendencies on translated texts, there is still a remarkable challenge in defining the specific features which characterise each translation universal presented.

On one hand, the main reason to investigate these hypotheses is to raise awareness among translators about their conscious or unconscious effects on translated texts, and the relationship between language and culture [8]. Bringing unconscious tendencies to light will emphasise translators' decisions, and hence should pave the way for future development of more accurate and natural translations, with more "desired effects and fewer unwanted ones"[12]. On the other hand, the automatic identification of these hypotheses' effects on texts can be a module in various natural language processing frameworks: it can improve web-based parallel corpus extractors' by finding the candidate parallel texts [13], or it can be integrated into statistical machine translation systems to learn the direction of the translation [14].

The objective of the current study is to model a language-independent learning system able to distinguish between translated and non-translated texts. The advantages of such a methodology are obvious: the system has a wide applicability for other languages, and thus, the "universal"character of this hypotheses is easier to investigate. An additional goal is to investigate whether the simplification features previously proposed [8, 15], according to the simplification universal, are influencing the learning model. In the same line of research similar supervised learning techniques were employed on different languages, for medical and technical texts in Spanish [16], and also for Italian [17]. It must be noted that most of the studies employed on translationese use a corpus-based approach [3, 8], whereas others adopt the advantages offered by machine learning techniques.

2 Related Work

Translationese has been previously studied by different scholars, most of them employing a corpus-based approach and comparing different patterns between translated and non-translated texts. Interesting patterns have been outlined, with important differences being discovered [18]. However, the inception of the pattern under investigation is almost always based on scholars' intuition, and clearly, the approach adopted is hand-engineered. Machine learning frameworks brought interesting results for translationese [17, 16] and new pathways of research were recently created. These methodologies are obviously complementary and new pathways to further investigations are outlined.

One of the hypotheses of translation theory is the simplification universal. The universal is described as the tendency of translators to produce simpler and easier-to-follow texts [2], and some empirical results sustaining the universal were provided [8]. Laviosa investigates lexical patterns for English and the obtained results show a relatively low proportion of lexical words over function words in translated texts, and a high proportion of high-frequency words compared to the low-frequency words. Moreover, great repetition of the most frequent words and less variety in the most frequently-used words has been emphasised [19].

Furthermore, a corpus-based approach which tests the statistical significance of the features proposed for the investigation of the simplification universal has been presented for Spanish [15, 20]. The experiments were on both the medical and the technical domains, and the translated texts were produced by both professional and semi-professional translators. In [15], the simplification universal is confirmed only for the lexical richness attribute. The results for the following features appear to be against this universal: complex sentences, sentence length, depth of syntactical trees, information load, number of senses per word. The experiments in [20] revealed that translated texts exhibit lower lexical density and richness, seem to be more readable, have a smaller proportion of simple sentences and appear to be significantly shorter, and that discourse markers were used significantly less often. Simplification fingerprints were found on the technical translations and seemed to show that texts written by non-professional translators do not have such simplification traits.

For Italian, a different perspective over translationese is given by the supervised learning approach employed by Baroni and Bernardini [17]. They investigate whether a computer system can distinguish between translated texts from non-translated ones in the Italian language. A special corpus for this purpose was compiled and the results of an SVM classifier depend heavily on lexical cues, on the distribution of n-grams of function words and on morpho-syntactic categories. In particular, they notice that elements such as personal pronouns and adverbs also influence the framework. Therefore, it is proved that shallow data representations can be sufficient to automatically distinguish professional translations from non-translated texts with a high accuracy, and it was hypothesised that this representation captures the distinguishing features of translationese. Moreover, the difficulty for humans to differentiate translated and non-translated texts is emphasised and explicit evidence of the superiority of automatic knowledge-poor system on the same task is shown. In contrast to their study, current experiments avoid the exploitation of n-gram indicators or any type of language-dependent attributes, being able to reuse the same learning model on various languages. The bag-of-words model (unigrams) was avoided to prevent learning to classify according to texts' topic. Additionally, the Romanian language has not been previously studied from this point of view, and since this effect on translated texts is claimed to be a 'universal', applying the learning model to a new language is a novel contribution in the field.

3 Methodology

The chosen methodology consists in supervised machine learning techniques, with the aim to model a learning system to distinguish between translated and non-translated texts. Therefore, a training and a test dataset were created, comprising instances from both classes at a ratio of 2:1 of non-translated:translated texts. By using the Weka tool[3] [21, 22], classifiers are trained by including and excluding the attributes proposed for the simplification universal within the feature vectors. As a result, the success rate would indicate whether the model is influenced to some extent by the simplification features.

Probably the best resource for the investigation of translationese is a comparable corpus (containing translated and non-translated texts in the same language) [23], and hence, this approach would avoid any foreign interference [24]. As no study of the Romanian language has been employed for translationese, a dedicated type of resource did not exist. For this reason a comparable corpus has been specially compiled for this task, consisting of newspaper articles published between 2005-2009. The translated subcorpus is collected from the Southeast European Times[4] comprising 223 articles written after the year 2005. The non-translated subcorpus comprises 416 documents from the same time-span, in the same domain, from a reputable newspaper from Romania called 'Ziua'[5]. The corpus has in total 341320 tokens (200211 for the translated subcorpus and 141109 tokens for the non-translated subcorpus). The selected articles are written by various translators, so the possibility of a specific style playing a role in the classification task is avoided. Also, the texts are translated from various languages into Romanian, an advantage that assures a high likelihood that all the discovered patterns are not due to one particular source language.

The collected dataset was randomly divided into a training set of 639 texts and a test set of 148 texts, while keeping the same ratio of translated and non-translated class instances in the training and test set. In order to extract the feature vector for the learning process, all the texts of the corpora were first tagged using the part of speech tagger provided as a web service by the Research Institute for Artificial Intelligence[6], the Romanian Academy [25, 26].

The learning system exploits thirty-eight language-independent features extracted from the tagger's output, including both the 'translationese features' and the 'simplification features'. As the translationese effect is considered to happen at the morphological level of the texts [8, 7], the first set of attributes captures general language features, in the current study being referenced as 'translationese features':

- the proportion in texts of grammatical words (the parts of speech considered to belong to this class: determiners, articles, prepositions, auxiliary verbs, pronouns, conjunctions, and interjections);
- the proportion of nouns in texts;
- the proportion of verbs in texts;
- the proportion of adjectives in texts;
- the proportion of adverbs in texts;
- the proportion of numerals in texts;
- the proportion of pronouns in texts;

[3] http://www.cs.waikato.ac.nz/ml/weka

[4] http://www.setimes.com

[5] http://www.ziuaveche.ro

[6] http://www.racai.ro/webservices/

- the proportion of prepositions in texts;
- the proportion of determiners in texts;
- the proportion of articles in texts;
- the proportion of conjunctions in texts;
- the proportion in texts of grammatical words per lexical words (the lexical words class is represented by nouns, verbs, adjectives, adverbs, and numerals);
- the proportion of interjections in texts;
- the proportion of proper nouns in texts;
- the proportion of common nouns in texts;
- the proportion in texts of verbs in the first person plural;
- the proportion in texts of verbs in the first person singular;
- the proportion in texts of verbs in the second person plural;
- the proportion in texts of verbs in the second person singular;
- the proportion in texts of verbs in the third person plural;
- the proportion in texts of verbs in the third person singular;
- the proportion of auxiliary verbs in texts;
- the proportion of copulative verbs in texts;
- the proportion of modal verbs in texts;
- the proportion in texts of verbs in the indicative mood;
- the proportion in texts of verbs in the subjunctive mood;
- the proportion in texts of verbs in the imperative mood;
- the proportion in texts of verbs in the infinitive mood;
- the proportion in texts of verbs in the gerund mood;
- the proportion in texts of verbs in the participle mood;
- the proportion of comparative adjectives in texts;
- the proportion of positive adjectives in texts;
- the proportion of superlative adjectives in texts.

The data representation for the learning system comprises all the above parameters and also includes the following previously proposed simplification features [19, 15, 16]:

- the lexical richness as the proportion of type lemmas per tokens;
- the sentence length as the proportion of number of words per sentence;
- the word length in terms of number of characters normalised by the tokens number;
- the number of simple sentences[7] normalised by the number of sentences;
- the information load as the proportion of lexical words to tokens.

Morpho-syntactic categories have been previously used as features in a similar classification task, [17], showing that non-clitic personal pronouns and adverbs are distinguishing features of translationese in a study on Italian texts. Also [16,

[7] Given that the tagger does not provide any syntactic information, the following algorithm has been employed to compute this feature: sentences with one or zero personal verbs are considered to be 'simple sentences'.

27] use similar features, such as proportion of numerals, adjectives, finite verbs, pronouns and nouns are among the most useful attributes in a classification task on Spanish medical and technical texts. The current experiments have the advantage of considering even more in depth each feature, in order to investigate if the sub-categories of these morphological features have a particular influence on the current learning model.

The classifiers applied for the categorisation task are the following: Jrip, Decision Tree, Naive Bayes, and SVM [22]. The evaluation results are outlined in the next section. These particular classification algorithms were chosen because the rules produced by Jrip and decision trees classifiers provide an output with what has been learnt, Naive Bayes because it is known to work well with text, and SVM because it is known to achieve high performance.

4 Evaluation

The results obtained with the data representation including the simplification features are compared to the accuracy obtained by the system without these. The accuracies for the 10-fold cross-validation evaluation on the training data and the accuracy for the test dataset evaluation are reported in Table 1. The training dataset comprises 639 instances (223 for the translation class and 416 for non-translation class instances), and the test dataset selected comprises 148 instances (49 for the translation class and 99 for non-translation class).

Table 1. Classification Results: Accuracies for several classifiers

	Excluding Simplification Attributes		**Including Simplification Attributes**	
Classifier	*10-fold cross-validation*	*Test set*	*10-fold cross-validation*	*Test set*
Baseline	65.10%	66.89%	65.10%	66.89%
Naive Bayes	91.71%	91.89%	95.46%	94.59%
SVM	97.18%	95.95%	98.90%	98.65%
Jrip	92.80%	93.24%	92.80%	97.30%
Decision Trees	92.64%	91.89%	94.52%	95.27%

The baseline classifier considers the majority class from the dataset, therefore the baseline is 64.5% since the dominant class is the non-translated one. The results shown are definitely an argument towards the existence of translationese, an effect that was hypothesised only twenty years ago. The best accuracy is obtained by the SVM classifier with a 98.90% value for the 10-fold cross validation and with a 98.65% value for the randomly selected test dataset. Moreover, the SVM classifier performed very well for the Spanish [16] and the Italian [17] language on the same categorisation task, even though different domains were involved in those experiments. These success rates are impressive and prove, without doubt, that an automatic system is able to distinguish between translated and non-translated texts. However, the removal of the 'simplification features' leads to a slightly decreased accuracy for all the classifiers, the Naive Bayes technique registering the biggest difference of approximately 4%. Nevertheless, as an overall perspective, all the classifiers' performances are outstanding, with accuracies ranging between 91.71% and 98.90%

In order to observe the rules considered by the classifiers, the pruned tree output from the JRip classifier and the Decision Tree output are outlined in figures 1 and 2. These two classifiers provide an intuitive output for more detailed data analysis [28].

```
Rule 1: (LexicalRichness <= 0.492095) and (Prepositions >= 0.106925)
=> class=translated (175.0/9.0)
Rule 2: (Nouns <= 0.302041) and (Prepositions >= 0.089489)
and (InformationLoad <= 0.001211)
=> class=translated (37.0/0.0)
Rule 3: (Prepositions >= 0.118367) and (InformationLoad <= 0.001507)
and (SentenceLenght <= 27.333334)
=> class=translated (9.0/0.0)
Rule 4: (CommonNouns <= 0.24838) and (InformationLoad <= 0.001329)
and (SimpleSentences >= 0.73913)
=> class=translated (8.0/1.0)
Rule 5: (LexicalRichness <= 0.485342) and (Adjectives >= 0.098787)
and (CommonNouns <= 0.259965)
=> class=translated (4.0/0.0)
Rule 6: => class=non-translated (406.0/0.0)
```

Fig. 1. JRip classifier rules output.

As can be seen in the JRip classifier's output, the first rule, quite frequently used, considers the lexical richness and the proportion of prepositions features, whilst more information for this classifier is provided by information load, sentence length and the proportion of nouns (the second and third rule). The fourth and fifth rule also use the proportion in texts of common nouns, simple sentences and adjectives. On the other hand, the decision trees classifier seems to give a much higher priority to the proportion of nouns, positioning this feature on the first level of the classification tree (figure 2), while lexical richness does not appear at all among the attributes considered by this classifier. Similar to the JRip output, information load and prepositions have an important role in the categorisation task. Moreover, the decision tree algorithm also considers the common nouns, the word length, the third person singular and first person verbs, the determiners and the adverbs in their last levels of the pruned tree.

Furthermore, the feature selection evaluators' output is exploited in order to see the ranking of the attributes, regardless of any classifier. The Information

```
Nouns <= 0.318261
|   InformationLoad <= 0.001387
|   |   Prepositions <= 0.093886
|   |   |   CommonNouns <= 0.232852
|   |   |   |   VerbsPersOneSingular <= 0.000472: non-translated (4.0)
|   |   |   |   VerbsPersOneSingular > 0.000472: translated (3.0)
|   |   |   CommonNouns > 0.232852: non-translated (29.0)
|   |   Prepositions > 0.093886
|   |   |   VerbsPersThreeSingular <= 0.033898
|   |   |   |   Pronouns <= 0.083582
|   |   |   |   |   Nouns <= 0.311526: translated (171.0)
|   |   |   |   |   Nouns > 0.311526
|   |   |   |   |   |   Adverbs <= 0.045584: translated (14.0)
|   |   |   |   |   |   Adverbs > 0.045584: non-translated (4.0/1.0)
|   |   |   |   Pronouns > 0.083582
|   |   |   |   |   VerbsPersOnePlural <= 0.004237: translated (12.0/1.0)
|   |   |   |   |   VerbsPersOnePlural > 0.004237: non-translated (3.0)
|   |   |   VerbsPersThreeSingular > 0.033898
|   |   |   |   Determiners <= 0.030864: translated (5.0)
|   |   |   |   Determiners > 0.030864: non-translated (6.0/1.0)
|   InformationLoad > 0.001387: non-translated (58.0)
Nouns > 0.318261
|   Prepositions <= 0.12724
|   |   Nouns <= 0.327212
|   |   |   WordLength <= 5.537314: non-translated (36.0/1.0)
|   |   |   WordLength > 5.537314: translated (3.0)
|   |   Nouns > 0.327212: non-translated (269.0/1.0)
|   Prepositions > 0.12724
|   |   InformationLoad <= 0.001662: translated (13.0/1.0)
|   |   InformationLoad > 0.001662: non-translated (9.0)
```

Fig. 2. Pruned tree output from the Decision Tree classifier.

Gain and Chi-squared algorithms provide the information from Table 2. The first twenty-six attributes are shown in the figure, as the rest of them are given a null value, and, consequently, they have been omitted from the table. The ranking provided by these two algorithms gives approximately the same type of information. This tendency is similar to the study on the Spanish language [16].

The first four features which most influence the classification are: information load, proportion of nouns, proportion of prepositions, and lexical richness, two of which are considered to stand for the simplification universal. They are shortly followed by another set of five features: proportion of common nouns, proportion in texts of grammatical words per lexical words, third singular verbs, numerals, grammatical words, and simple sentences. The scores provided for the two ranking filters drop for the rest of the items.

Regarding the simplification features investigated in these experiments, the ranking algorithms place three of them among the top most influencing features: information load - actually being the most useful feature of all, lexical richness as has been previously hypothesised [3], and proportion of simple sentences in texts (item ranked among the first also in the study on Spanish texts [16]).

4.1 The Simplification Learning Model

In order to bring light on the simplification hypothesis, the learning model has been trained using only the simplification features (information load, lexical

Table 2. Attributes Ranking Filters.

Chi-squared		**Information Gain**	
321.4558	InformationLoad	0.4367	InformationLoad
320.3795	Nouns	0.4207	Nouns
311.8009	Prepositions	0.4082	Prepositions
287.4316	LexicalRichness	0.3922	LexicalRichness
271.1993	CommonNouns	0.3391	GrammaticalWordsPerLexicalWords
258.2133	GrammaticalWordsPerLexicalWords	0.3387	CommonNouns
186.1927	VerbsPersThreeSingular	0.2319	VerbsPersThreeSingular
174.5388	Numerals	0.2304	Numerals
167.4469	GrammaticalWords	0.2081	GrammaticalWords
130.8715	SimpleSentences	0.17	SimpleSentences
59.031	VerbsIndicative	0.0743	Determiners
50.5388	Determiners	0.0738	VerbsIndicative
49.3664	Conjunctions	0.0639	Conjunctions
48.0164	Adverbs	0.0565	Adverbs
45.7126	ProperNouns	0.0537	ProperNouns
42.9458	VerbsParticiple	0.0507	VerbsParticiple
38.1205	VerbsGerund	0.0487	SentenceLenght
34.758	SentenceLenght	0.0441	VerbsGerund
29.3952	VerbsPersOnePlural	0.0327	VerbsPersOnePlural
28.0491	WordLength	0.0312	WordLength
24.395	Pronouns	0.0308	Pronouns
24.1608	Verbs	0.0294	Verbs
22.8642	VerbsPersTwoSingular	0.029	VerbsSubjonctive
21.9677	VerbsAux	0.0287	VerbsPersTwoSingular
18.515	VerbsSubjonctive	0.0249	VerbsAux
7.1135	AdjectivesSuperlative	0.0128	AdjectivesSuperlative
.....		.....	

richness, sentence length, word length, and simple sentences). For this reason this model is furthermore called 'the simplification learning model'. The training and the test datasets are the same as in the previous experiments, and also the same classifiers have been used. The results on the simplification learning model, as shown in Table 3, are lower, but nevertheless remarkable. This impact on the classifiers is expected, as this learning model uses only five features and, as can be seen from the previous analysis presented, there are other attributes which contribute in the original learning model.

Table 3. Classification Accuracies for the Simplification Learning Model

	10-fold	*Test*
Classifier	*cross-validation*	*set*
Baseline	65.10%	66.89%
NaiveBayes	88.58%	85.81%
SVM	87.64%	87.84%
Jrip	88.42%	93.24%
J48	88.89%	96.62%

The new learning model is able to classify the texts with accuracies between 87.64% to 88.89% on 10 fold cross-validation, and reaching up to 93.24% for the test dataset. These values may be considered an argument in favour of the simplification universal.

5 Conclusions and Further Research

This paper presents a new study on the investigation of universals of translations, for the Romanian language. A supervised learning approach is employed to identify the most informative features that characterise translations compared to non-translated texts. Additionally, an analysis of the impact on classification of the features previously proposed for the ’simplification universal’ is conducted. The accuracies of the learning model in the categorisation task have outstanding values, and reach up to 98.65% value on a randomly generated test dataset. All the classifiers register a decreased success rate when the simplification features are removed. However, the lowest result is still well above the chance level. The performance analysis on the classifiers’ output reveals that the learning model relies highly on the following attributes: information load, lexical richness, proportion in texts of nouns, prepositions, grammatical words to lexical words, third person singular verbs, numerals and simple sentences. For future work, the inclusion of other features considered to stand for different translation universals in the learning model may bring different arguments towards their validity.

Acknowledgements

We would like to express our gratitude to members of staff at Research Group of Computational Linguistics from University of Wolverhampton, and in particular to Georgiana Puşcaşu, Dr. Constantin Orăsan, and Prof. Ruslan Mitkov for their invaluable suggestions, guidance and support. Also, we are thankful to Claudiu Mihăilă from University of Manchester for his prompt and constructive comments.

References

1. Gellerstam, M.: Translationese in Swedish novels translated from English. Translation Studies in Scandinavia. Lund: CWK Gleerup (1986)
2. Baker, M.: Corpus Linguistics and Translation Studies Implications and Applications. In: Text and Technology: In Honour of John Sinclair. Amsterdam & Philadelphia: John Benjamins (1993) 233–250
3. Baker, M.: Corpus-based Translation Studies: The Challenges that Lie Ahead. In: Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager. Amsterdam & Philadelphia: John Benjamins (1996) 175–186
4. Borin, L., Prütz, K.: Thorough a dark glass: part of speech distribution in original and translated text. In: Computational Linguistics in the Netherlands. Amsterdam: Rodopi (2001) 3044
5. Hansen, S.: The Nature of Translated Text. Saarbrücken: Saarland University (2003)

6. Teich, E.: Cross-linguistic Variation in System and Text. Berlin:Mouton de Gruyter (2003)
7. Toury, G.: Descriptive Translation Studies and Beyond. Amsterdam: John Benjamins (1995)
8. Laviosa, S.: Corpus-based Translation Studies. Theory, Findings, Applications. Amsterdam & New York: Rodopi (2002)
9. Tymoczko, M.: Computerized corpora and the future of translation studies. Meta **43:4** (1998) 652–659
10. Bernardini, S., Zanettin, F.: When is a Universal not a Universal? In: Translation Universals. Do they exist? Amsterdam: Benjamins (2004) 5162
11. Toury, G.: Probabilistic explanations in translation studies. Welcome as they are, would they qualify as universals? In: Translation Universals: Do they exist? Amsterdam: John Benjamins (2004) 15–32
12. Chesterman, A.: A Causal Model for Translation Studies. In: Intercultural Faultlines. Research Models in Translation Studies I. Textual and Cognitive Aspects. St. Jerome (2000) 15–27
13. Resnik, P., Smith, N.: The web as a parallel corpus. Computational Linguistics **29(3)** (2003) 349380
14. Goutte, C., Kurokawa, D., Isabelle, P.: Improving smt by learning translation direction. In: Statistical Multilingual Analysis for Retrieval and Translation, Barcelona, Spain (2009)
15. Corpas, G.: Investigar con corpus en traduccin: los retos de un nuevo paradigma. Frankfurt am Main, Berlin & New York: Peter Lang (2008)
16. Ilisei, I., Inkpen, D., Pastor, G., Mitkov, R.: Identification of Translationese: A Supervised Learning Approach. In: CICLing 2010, Lecture Notes in Computer Science 6008. Springer, Heidelberg (2010) 503–511
17. Baroni, M., Bernardini, S.: A new approach to the study of translationese: Machine-learning the difference between original and translated text. Literary and Linguistic Computing **21, 3** (2006) 259–274
18. Olohan, M., Baker, M.: Reported 'that' in translated english: Evidence for subconcious processes of explicitation? Across Languages and Culture **1:2** (2000) 141–158
19. Laviosa, S.: Core patterns of lexical use in a comparable corpus of English narrative prose. In: The Corpus-Based Approach. Volume Special Issue of Meta. Montral: Les Presses de L'Universit de Montral (1998) 557–570
20. Corpas, G., Mitkov, R., Afzal, N., Pekar, V.: Translation universals: Do they exist? a corpus-based nlp study of convergence and simplification. In: Proceedings of the AMTA, Waikiki, Hawaii (2008)
21. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: An update. SIGKDD Explorations **11(1)** (2009) 10–18
22. Witten, I.H., Frank, E.: Data Mining : Practical Machine Learning Tools and Techniques. Second edition edn. Morgan Kaufman (2005)
23. Olohan, M.: Introducing Corpora in Translation Studies. Routledge (2004)
24. Pym, A.: On Toury's laws of how translators translate. In: Beyond Descriptive Translation Studies. Benjamins (2008) 311–328
25. Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D.: Racai's linguistic web services. In: Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco. Number ISBN 2-9517408-4-0, ELRA - European Language Ressources Association (2008)

26. Tufiş, D., Ştefănescu, D., Ion, R., Ceauşu, A.: RACAI's Question Answering System at QA@CLEF 2007. In: Advances in Multilingual and Multimodal Information Retrieval (CLEF 2007), Lecture Notes in Computer Science. Volume 5152. Springer-Verlag (2008) 3284–3291
27. Ilisei, I., Inkpen, D., Corpas, G., Mitkov, R.: Towards simplification: A supervised learning approach. In: Proceedings of Machine Translation 25 Years On, London, United Kingdom. (2009)
28. Quinlan, J.R.: Induction of decision trees. Machine Learning **1** (1986) 81–106