Towards the Classification of the Finnish Internet Parsebank: Detecting Translations and Informality

Veronika Laippala[1,2] **Jenna Kanerva**[3] **Anna Missilä**[2]
Sampo Pyysalo[3] **Tapio Salakoski**[3] **Filip Ginter**[3]
[1] Turku Institute for Advanced Studies, University of Turku, Finland
[2] School of Languages and Translation Studies, University of Turku, Finland
[3] Department of Information Technology, University of Turku, Finland
first.last@utu.fi

Abstract

This paper presents the first results on detecting informality, machine and human translations in the Finnish Internet Parsebank, a project developing a large-scale, web-based corpus with full morphological and syntactic analyses. The paper aims at classifying the Parsebank according to these criteria, as well as studying the linguistic characteristics of the classes. The features used include both lexical and morpho-syntactic properties, such as syntactic n-grams. The results are practically applicable, with an AUC range of 85–85% for the human, $\sim$ 98% for the machine translated texts and 73% for the informal texts. While word-based classification performs well for the in-domain experiments, delexicalized methods with morpho-syntactic features prove to be more tolerant to variation caused by genre or source language. In addition, the results show that the features used in the classification provide interesting pointers for further, more detailed studies on the linguistic characteristics of these texts.

1 Introduction

With its growing size and coverage, the Internet has become an attractive source of material for linguistic resources, used both for linguistics and natural language processing (NLP) applications (Baroni et al., 2009; Kilgarriff and Grefenstette, 2003). However, automatically collected, very large corpora covering all the text that can be found are very heterogeneous, which may complicate their usage. In linguistics, the origin of the corpus texts is of primary importance (Biber et al., 1998; Sinclair, 1996), and also in many NLP applications, such as automatic syntactic analysis, linguistic variation across different domains affects the results significantly (Laippala et al., 2014).

This paper presents the first results on the linguistic variation in the Finnish-language Internet by analyzing informality, machine translations and human translations in the Finnish Internet Parsebank[1], an on-going project aiming at a large-scale, web-based corpus of Finnish with full morphological and syntactic analyses. The current version consists of 3.2 billion tokens and 241 million sentences.

This article has two main objectives. The first aim is to develop classification methods in order to detect informality, machine translations and human translations from the Parsebank. This would facilitate the use of the Parsebank, as searches or applications could be targeted only at certain parts of the corpus. In the classification, the features used include syntactic n-grams, little subtrees of dependency syntax analyses developed for Finnish by Kanerva et al. (2014), originally produced for English by Goldberg and Orwant (2013).

Secondly, the study points research directions for the analysis of the linguistic characteristics of the text classes. The automatic classification based on the data-driven combination of lexical, syntactic and morphological features offers a new approach to the linguistic study of these texts and their characteristics, as traditional linguistic studies often concentrate on the analysis of a limited number of preselected features.

The study consists of three sets of classification experiments and their analyses. In the first, texts are classified according to the level of formality to standard and informal. In the second, the classification is done to human translated texts and texts originally written in Finnish, and in the last, to machine translated texts and texts originally written in Finnish.

[1] http://bionlp.utu.fi/finnish-internet-parsebank.html

2 Related Work

Regarding human translation detection, *translationese* is a term originally coined by linguists to refer to features typical of translated texts. Baker (1993) was the first to define potential *translation universals*: features that all translated texts hypothetically share.

The existence of translationese has also been tested by studies applying machine learning. Baroni and Bernardini (2006) use monolingual corpora to experiment with for instance lemmas and POS tags, providing evidence that an algorithm can perform better than humans in recognizing human translated texts. Other similar studies include Ilisei et al. (2010) presenting a language-independent system based on average sentence length or lexical richness, Popescu (2011) using solely character 5-grams (ignoring sentence boundaries) to detect English translations, and Avner, Ordan and Wintner (2014) concentrating on morphological properties in Hebrew.

Previous studies on classifying machine translated texts mostly rely on different combinations of lexical and grammatical features as well. Aharoni, Koppel, and Goldberg (2014) use a set of function words, POS tags and a mix of the two to classify texts, whereas Arase and Zhou (2013) concentrate on indicators based on sentence-internal coherence, also called the *phrase salad* phenomenon (Lopez, 2008).

Despite their relative infrequency, some previous work also concentrate on classifying informality. Unlike those concerning translation classification, these concentrate on lexical rather than morphological features. Lahiri, Mitra, and Lu (2011) explore the *Formality Score*, a frequency list based on the differences of word classes in a corpus. Mosquera and Moreda (2011) define the most relevant features of informality to be the frequencies of spelling mistakes, interjections, and emoticons. These same individual features have also been studied as signs of informality in many linguistic studies (Lehti and Laippala, 2014).

3 Data

3.1 Finnish Internet Parsebank and Syntactic N-grams

The current version of the Finnish Internet Parsebank consists of 3.2 billion tokens and 241 million sentences. It is produced by crawling the Finnish web with the Spiderling web crawler[2]. Being designed for collecting text corpora, it can be targeted to crawl only pages in a specific language. In addition, it can automatically remove boilerplate text, such as lists and menus from the output. The output is deduplicated at the web page level and fully morphologically and syntactically analyzed using the parsing pipeline by Haverinen et al˙ (2013).

Syntactic n-grams are little subtrees of dependency syntax analyses, originally produced for English by Goldberg and Orwant (2013) and recently for Finnish by Kanerva et al. (2014). Instead of the linear context used with flat n-grams, the context for syntactic n-grams is defined by the syntactic representation. Possible configurations include combinations from one to four arcs. In addition to the syntactic and lexical information, complete syntactic n-grams include the part-of-speech (POS) categories of the words together with their morphological features (see Figure 1). Some of these analyses and/or the words can be also be deleted in order to obtain the desired level of description granularity.

3.2 Translations

The source data for machine translation comes for most part from WaCky (Baroni et al., 2009). The corpora used were ukWaC for English, frWaC for French, and deWaC for German. These languages were chosen based on both their common usage and availability. A random sample was taken from each of the corpora and machine translated using Google Translate (2015). The resulting translation was then parsed using the parsing pipeline by Haverinen et al. (2013). The part of data marked "randomPB" in Table 1 is a random selection from the Parsebank, manually identified as machine translated.

The Finnish human translations are taken from the *Corpus of Translated Finnish* (CTF) (Mauranen, 2000). The 10-million-word CTF is categorised into different genres based on the classifications by publishers and reviewers. It can be divided into three different sub-corpora: firstly, a corpus of translated Finnish where English is the source language, secondly, a corpus of original Finnish, and thirdly, a substantially smaller corpus of translated Finnish with multiple source languages (for example Russian, French, and Ger-

[2]http://nlp.fi.muni.cz/trac/spiderling

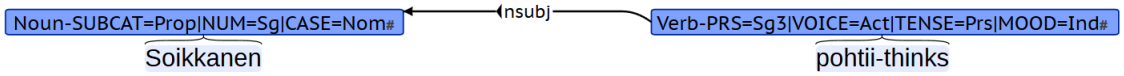

Figure 1: Syntactic 2-gram with word-level, POS-level and morphological analyses.

man). The question of data availability governed the text selection, and therefore only translations with English as the source language were chosen as training set.

In order to expand the available test data and ensure the performance of the algorithm, the biography section with English as the source language was kept as as an out-of-domain test corpus not used in the training.

3.3 Standard and Informal Data

While standard language is relatively simple to define as a variant following the recommendations and guidelines of a language, informality is less so.

In Finnish, the language variant defined as more free and less premeditated (Institute of Languages of Finland, 2014) is generally referred to as "everyday language / arkikieli". Despite its somewhat misleading name, the term is used for language variants that could also be called informal, regardless of the topic of the discussion or text (Grönros, 2006). Typical instances of informality are for instance playful and subjective expressions (Mäkinen, 1989).

In this paper, we adopt the term informality to refer to language that does not follow the general language guidelines and/or includes other structural or lexical instances untypical for standard language. As noted by Grönros (2006), informality is subjective with frequent borderline cases. In order to operationalize the concept, we rely both on human annotation and on data representing informal texts and apply two sources of data: the weak data set, a collection of large unannotated corpora that are expected to be biased towards standard or informal sentences based on their venue of publication, and second, the gold data set, a smaller corpus of sentences drawn at random from the Parsebank and manually annotated to identify sentence formality.

The annotation process is described in Section 3.4 and the unannotated data used in Table 2. The different parts of the standard language section of the unannotated data are derived from two sources: the news and the Europarl sections from

Standard language corpora	Words
News text	27 121
Europarl	18 946
Academic research papers	1 400 281
Biographies	337 642
Popular science	632 102
Total	2 416 092
Informal language corpora	Words
Popular blogs	21 791
Online discussion forums:	
- the Finnish yellow press	54 091
- the main Finnish newspaper	93 425
- a big Finnish online community	65 966
Total	236 083

Table 2: Informal and standard language corpora.

Turku Dependency Treebank (TDT) (Haverinen et al., 2013), and the academic research articles, biographies and popular science books from CTF (Mauranen, 2000). On the informal side, the blogs are from TDT, and the forum discussions from the main Finnish newspaper website from a private corpus collected by a research group from the School of Languages and Translation Studies from the University of Turku. The rest of the forum discussions are collected for the purposes of this study.

3.4 Formality Annotation

For the formality classification task, we annotated a random sample of sentences from the Finnish Parsebank. The manual annotation involved assigning each sentence into one of three categories: standard Finnish, informal Finnish, or not Finnish. Only the former two categories were considered in the experiments, with sentences identified as not being Finnish discarded after initial annotation.

The manual annotation effort started with simple initial guidelines (approximately one page) that were applied by two annotators working independently on the same sample of sentences. The two sets of annotations were then compared, differences resolved to generate merged consensus annotation, and the guidelines refined to identify the desired annotation in cases where disagree-

	Genre+Lang	Train	Devel	Test	Total
MT	DE	12 166	12 165	7228	31 559
	EN	17 664	17 663	14 511	49 838
	FR	23 662	23 662	19 117	66 441
	RandomPB			4468	4468
	TotalMT	53 492	53 490	40 856	147 842
MT OOD	DE	12 166	12 165		24 331
	EN	17 664	17 663		35 327
	FR			66 441	66 441
MT OOD2	DE			31 559	31 559
	EN	17 664	17 663		35 327
	FR	23 662	23 662		47 324
HT	AcadDE			66 774	66 774
	AcadEN	428 365	428 364	158 215	1 014 944
	AcadFR			57 373	57 373
	BioEN			151 517	151 517
	ChildEN	306 856	306 856	43 437	657 149
	DetEN	72 243	72 243	200 746	345 232
	EntEN			270 332	270 332
	FicEN	537 904	537 903	188 935	1 264 742
	PopEN	133 355	133 355	188 587	455 297
	TotalHT	1 478 723	1 478 721	1 407 628	4 365 072
HT OOD	BioEN			151 517	151 517
Orig	AcadFI	525 326	525 326	267 418	1 318 070
	BioFI			309 941	309 941
	ChildFI	256 768	256 767	94 590	608 125
	DetFI	31 374	31 374	176 251	238 999
	EntFI			235 885	235 885
	FicFI	551 318	551 318	105 244	1 207 880
	PopFI	193 554	193 554	203 143	590 251
	TotalFI	1 558 340	1 558 339	1 392 472	4 509 151
Orig OOD	BioFI			309 941	309 941
Orig-PB	WebFI	53 493	53 492	40 640	147 625
Total		**3 144 048**	**3 144 042**	**2 804 352**	**9 092 442**

Table 1: Translation detection statistics: number of words. (MT - machine translation, HT - human translation, Orig - original, Orig-PB - original from Parsebank, OOD - out of domain; DE - German, EN - English, FI - Finnish, FR - French; Acad - academic, Bio - biography, Child - children, Det - detective, Ent - fictional entertainment, Fic - fiction, Pop - popular non-fiction.)

ment was found. This double annotation protocol was repeated until an 89% intern-annotator agreement was reached on a batch of 100 sentences. After this initial development and refinement of annotation guidelines and annotator training, annotation proceeded independently to categorize a total of 3300 sentences.

After the primary annotation, the final training, development and test sets were prepared as follows. First, 218 sentences identified as *not Finnish* and 27 sentences that were duplicates of other sentences in the data were discarded. The remaining sentences were then down-sampled to 3000, which were split into a training set of 1500, a development set of 500, and a test set of 1000 sentences. The random split was stratified to roughly preserve the distribution of standard and informal sentences in each subset. Table 3 shows the final statistics of the annotated corpus.

	Train	Devel	Test	Total
Standard	957	321	639	1917
Informal	543	179	361	1083
Total	1500	500	1000	3000

Table 3: Formality annotation statistics: number of sentences

4 Experiments & Results

4.1 Experimental setup

We evaluate performance using two standard sets of metrics. First, we report classification precision, recall and their balanced harmonic mean, the F_1-score (*F-score* for short). As these metrics are sensitive to the distribution of positive and negative instances in the test data, we also report the area under the receiver operating characteristic curve (AUC), which corresponds to the probability that a randomly chosen negative and a randomly chosen positive example will be correctly ranked by the classifier. As AUC is invariant with respect to the positive-negative distribution, it is more readily applicable for comparing performance across datasets that have a different balance of examples of these classes.

In the sentence formality classification task, we consider the informal class positive and the standard class negative for the purposes of calculating precision, recall and F-score.[3] In the translation recognition task, we similarly consider the translation classes positive and the original Finnish class negative. In the informality detection, the classification is done at the sentence-level, where as the translations are classified in segments of five sentences.

In all three classification tasks, a bag-of-words (BOW) approach is used as a simple baseline method. Leaning purely on lexical items can however lead to a topic-wise classification which would decrease the performance when classifying texts with wide range of topics, such as Internet texts from the Finnish Parsebank. Therefore, we also run two delexicalized approaches giving linguistically interesting results on the morpho-syntactic characteristics of the corpora as well. First, we derive features from the morphology of the tokens, using combined POS and morphological feature uni-, bi- and trigrams. During the preliminary experiences, we noticed that the fine-grained morphology carries some features that signal very reliably the text class but do not carry any linguistically interesting information, such as morphological features indicating proper nouns, capitalization and whether the Finnish morphological analyzer recognized the token. As our aim is not only to classify the texts but also to analyze the linguistic characteristics of the resulting classes, these analyses are discarded from the morphology. In the second delexicalized approach, the feature set is expanded to include syntactic information giving the opportunity to recognize more complex sentence structures (see Figure 1).

The machine learning is carried out using a linear classifier trained with the stochastic gradient descent method.[4] We optimized the learning rate in preliminary experiments and set it constant throughout the rest of the study, since a per-experiment grid search is unlikely to result in any substantial gains. This allows us a very fast turnaround in the various runs, and the classifier is performance-wise roughly equal to linear SVMs on our data — which we verified in preliminary experiments as well.

4.2 Standard / Informality Classification

Table 4 presents the results of the informality classification task trained with different feature combinations on different data sets. All the methods are tested on the manually annotated gold standard.

For the system trained on the manually annotated gold data set, the best AUC, 73%, is produced by the simple BOW method, indicating that the vocabulary is the most distinguishing characteristic of informality, as already proposed by previous studies discussed in Sections 2 and 3.3. For the system trained on the more heterogeneous weak data set, the delexicalized methods are slightly better. This could suggest that the delexicalized methods are more robust and better adapted to variation in the test setting, and also proves that informal text does have morpho-syntactic characteristics as well. In addition, even if the performance of the systems trained on the weak data set is not excellent, its performance on the gold data set proves that the training of such a system without manual annotation is possible.

Although outperformed by the BOW method, the results of the classification based on morpho-

[3] The assignment of positive and negative label does not affect AUC.

[4] Implemented in the well-known *Vowpal Wabbit* package (Agarwal et al., 2011).

	Train	AUC	Pre	Rec	F-score
Bag-of-words	Weak	64.23	47.50	47.37	47.43
POS+feat	Weak	63.25	48.01	46.81	47.41
POS+feat+syntax	Weak	66.22	49.86	49.31	49.58
Bag-of-words	Gold	73.34	66.67	39.89	49.91
POS+feat	Gold	69.71	59.84	41.27	48.85
POS+feat+syntax	Gold	70.03	60.00	40.72	48.51

Table 4: Results for the detection of informal sentences. *POS+feat* refers to unigram, bigram and trigram sequences including the POS and other morphological tags of the tokens. *Syntax* refers to sequences of syntactic relations generated from the delexicalized syntactic n-grams.

logical and syntactic features are also reasonable and provide interesting information on the structural characteristics of the classes. Table 5 shows some of the most significant morphological + syntactic features of the informal text class with similar features grouped together. The tendency is clear with interjections and pronouns forming the majority of the ten most important features. In fact, this supports the findings by Mosquera et Moreda (2011).

Rank	Feature
1	Interj / Punct
2	Interj
5	Interj / ROOT
3	Pron+NUM_Sg_CASE_Nom
6	Pron+NUM_Pl+CASE_Nom
10	Pron+NUM_Sg+CASE_Nom / N+NUM_Sg+CASE_Gen

Table 5: Most significant features for the informal class grouped together. The rank means the significance rank of the feature in the classification.

4.3 Human Translated / Original Text Classification

The results of the human translation detection, shown in Table 6, support the existence of translationese: especially in the general training setting where the test set includes both domain and out-of-domain data, the best detection AUC is 87.19%. For the general test setting, the best results are obtained with the simple BOW method. However, when tested against an out-of-domain data set consisting of biographies, a genre not included in the training data set, the other methods perform clearly better, showing that a delexicalized method is more easily generalizable than the ones based on lexicon, and that the classification is also possible based on morphological and syntactic structures. Furthermore, these structural features are more interesting for the linguistic study of the characteristics of the classes than simple words.

Table 7 and Table 8 show some of the most significant features of human translated texts and texts originally written in Finnish, with similar features grouped together. Some of them reflect translation universals found in previous studies. In particular, the noun+verb combinations (ranks 2 and 14) in the translations, and the pronoun+verb combinations (rank 1) in the original Finnish support the previous results (Nevalainen, 2003; Laviosa-Braithwaite, 1995) on the frequency of pronouns on original texts on one hand, as well as on the lexical repetition on the other.

However, some of the results also contest previous studies. In the original Finnish data, many n-grams (ranks 5,7,8) seem to describe simple verb+argument fragments, which could easily reflect simplification, a typical feature of translationese studied by Blum-Kulka and Levenston (1983) and Laviosa (2002). Also nonfinite structures appear in both classes, even though according to a previous study by Puurtinen (2003), these would be typical of translated texts.

4.4 Machine Translated / Original Text Classification

The results of the machine translation detection are shown in Table 9. They reflect the effortlessness of the task: the results attain an $\sim 98\%$ AUC and a $\sim 91\%$ F-score for all data sets. For the out-of-domain tasks where the test sets are composed of translations from a language not included in the training set, the results are equally good, indicating that the source language does not have a significant effect. This has practical advantages, as machine translations can be detected without collecting training data from all possible languages.

	Train	Test	AUC	Pre	Rec	F-score
Bag-of-words	HT + Orig	HT + Orig	87.19	75.71	79.90	77.75
POS+feat	HT + Orig	HT+ Orig	84.98	75.28	74.17	74.72
POS+feat+syntax	HT + Orig	HT + Orig	86.26	75.17	77.76	76.57
Bag-of-words	HT + Orig	HT OOD + Orig OOD	81.80	61.20	58.82	59.99
POS+feat	HT + Orig	HT OOD + Orig OOD	86.05	60.99	72.84	66.39
POS+feat+syntax	HT + Orig	HT OOD + Orig OOD	85.00	59.16	73.15	65.42

Table 6: Results for the detection of human translations. *POS+feat* refers to unigram, bigram and trigram sequences including the POS and other morphological tags of the tokens. *Syntax* refers to sequences of syntactic relations generated from the delexicalized syntactic n-grams. The data sets are presented in Table 1.

Rank	N-gram
1	**C+SUBCAT_CC / Pron+SUBCAT_Dem+NUM_Pl+CASE_Ill** *and / to-those*
15	V+NUM_Sg+CASE_Nom+VOICE_Act+PCP_PrfPrc+CMP_Pos /… V+PRS_Sg3+VOICE_Act+TENSE_Prt+MOOD_Ind / **C+SUBCAT_CC** *broken / took / and*
2	**N+ / V+PRS_Sg3+VOICE_Act+TENSE_Prt+MOOD_Ind** */A / took*
14	Punct **N+ / V+PRS_Sg3+VOICE_Act+TENSE_Prt+MOOD_Ind** */. / A / took*
4	**Pron+SUBCAT_Rel+NUM_Pl+CASE_Nom /…** Pron+SUBCAT_Pers+NUM_Pl+CASE_Gen / Adv+POSS_Px3 *who / ours / together-with*
6	N+NUM_Pl+CASE_Ins / N+NUM_Sg+CASE_Ela / **Pron+SUBCAT_Rel+NUM_Sg+CASE_Nom** *with-fingers / from-town / which*
5	Punct / **V+NUM_Sg+CASE_Ine+POSS_Px3+VOICE_Act+INF_Inf2 /…** **V+NUM_Sg+CASE_Ine+VOICE_Act+INF_Inf3** *. / while-he-was-going / taking*
7	N+NUM_Sg+CASE_Ade / **V+NUM_Sg+CASE_Abe+VOICE_Act+INF_Inf3 / N+NUM_Pl+CASE_Par** *at-the-table / without-understanding / dogs*

Table 7: Most significant features in the human translation class, followed by example lexicalizations. The features are POS n-grams with morphological features. The first column refers to the feature ranks in the classification, 1 being the most significant feature.

The best results for the general test setting are obtained with the syntactic n-grams, while the weakest ones are obtained with the BOW method. Although the BOW's AUC is comparable to other methods, the recall for the general setting is 81.58%, and 66.34% and 58.54% for the out-of-domains. This implies that the most significant features of the machine translations are not lexical and that the structural information included in the POS, morphological and syntactic analyses is needed, most importantly when generalizing to domains not included in the training data.

5 Conclusion

This paper proves that a reliable detection of informality, human and machine translations is realistic. As shown already by Aharoni et al. (2014), machine translations can be detected at an extremely high level of accuracy. In addition, our results indicate that the source language does not affect the results significantly. For human translations, the detection task is obviously more difficult. However, our results achieve a very applicable AUC of $\sim$ 86%, both for the general setting and the out-of-domain one, showing that genre variation has some but not a dramatic effect on the results. For the informality detection, the results are applicable, although they can still be improved. For this class in particular, more studies on genre variation is needed in order to improve the classification features and thereby results.

For the machine translation experiment in the general setting, the features composed of POS, morphological and syntactic information performed the best, while for the human translation and informality detection, the BOW reached better results. However, in out-of-domain settings, the BOW is clearly outperformed by the other

Rank	N-gram
1	**Pron+NUM_Sg+CASE_Nom / V+PRS_Sg1+VOICE_Act+TENSE_Prt+MOOD_Ind** *I / ran*
4	**V+CASE_Ine+VOICE_Pass+INF_Inf2 / /...** **V+PRS_Pe4+VOICE_Pass+TENSE_Prs+MOOD_Ind / A+NUM_Pl+CASE_Par+CMP_Pos** *if-it-is-needed / we-put / more-funny*
9	N+NUM_Pl+CASE_Nom / N+NUM_Sg+CASE_All / **V+NUM_Sg+CASE_Ill+VOICE_Act+INF_Inf3** *children / to-the-school / to-read*
12	A+NUM_Pl+CASE_Tra+CMP_Pos / **V+PRS_Sg1+VOICE_Act+MOOD_Pot /...** **V+NUM_Sg+CASE_Ins+VOICE_Act+INF_Inf2** *to-wise / might / resulting-from*
5	N+NUM_Pl+CASE_Nom / **V+PRS_Sg2+VOICE_Act+TENSE_Prt+MOOD_Ind** *children / you-said*
7	Pron+NUM_Sg+CASE_All / **V+PRS_Sg3+VOICE_Act+TENSE_Prt+MOOD_Ind / Punct** *for-him / he-said / .*
8	N+NUM_Pl+CASE_Par / **V+PRS_Pl3+VOICE_Act+TENSE_Prt+MOOD_Ind / N+NUM_Sg+CASE_Ine** *dogs / they-said / in-house*

Table 8: Most significant features in the original Finnish class, followed by example lexicalizations. The features are POS n-grams with morphological features.

	Train	**Test**	**AUC**	**Pre**	**Rec**	**F-score**
Bag-of-words	MT + Orig-PB	MT + Orig-PB	98.03	99.10	80.58	88.88
POS+feat			98.06	96.22	86.41	91.05
POS+feat+syntax			98.35	98.89	86.17	92.09
Bag-of-words	MT OOD + Orig-PB	MT OOD + Orig-PB	95.37	99.51	66.34	79.61
POS+feat			97.56	98.64	82.84	90.05
POS+feat+syntax			98.17	97.56	85.37	91.06
Bag-of-words	MT OOD2 + Orig-PB	MT OOD2 + Orig-PB	97.31	97.96	58.54	73.28
POS+feat			97.56	98.64	82.84	90.05
POS+feat+syntax			98.03	99.40	82.51	91.57

Table 9: Results for classifying machine translated text and text originally written in Finnish. *POS+feat* refers to unigram, bigram and trigram sequences including the POS and other morphological tags of the tokens. *Syntax* refers to sequences of syntactic relations generated from the delexicalized syntactic n-grams. The data sets used are described in Table 1.

approaches. This demonstrates that while word-based methods can be useful for well defined contexts, different levels of delexicalizations are more tolerant for linguistic variation caused by for instance differences in genre or the source language, making them further applicable for the Parsebank classification.

In addition, it is important to notice that even if they were not ranked first for all the tasks, the delexicalized methods reached good results. indicating that morpho-syntactic differences between the texts classes can be captured by automatic classification. From a linguistic perspective studying the characteristics of the text classes, this is very promising. Also our findings on the distinguishing features of the studied classes reflect this: by supporting some previous findings and contesting others, the delexicalized classification method provides material for linguistic studies. Even if a detailed analysis of all of the features is not possible in the scope of this article, the utility of the approach is demonstrated.

The article offers multiple possibilities for future studies. In particular, the most significant text class features pointed out by the classification offer several research directions. In addition, the method can be extended to the study of other lexical and morpho-syntactic characteristics of other genres. Naturally, an obvious next step would also be the classification of the entire Internet Parsebank.

Acknowledgments

This work has been supported by the Kone foundation and Emil Aaltonen foundation. We also thank CSC - IT Center for Science.

References

Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. 2011. A reliable effective terascale linear learning system. *CoRR*, abs/1110.4198.

Roee Aharoni, Moshe. Koppel, and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 289–295.

Yuki Arase and Ming Zhou. 2013. Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Long Papers*, pages 1597–1607.

Alexander Ehud Avner, Noam Ordan, and Shuly Wintner. 2014. Identifying translationese at the word and sub-word level. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Mona Baker. 1993. Corpus linguistics and translation studies: implications and applications. In Gill Francis and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 233–252. John Benjamins.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Douglas Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, Cambridge.

Soshana. Blum-Kulka and Edwards Levenston. 1983. Universals of lexical simplification. *Language Learning*, 28:399–415.

Yoav Goldberg and John Orwant. 2013. A dataset of syntactic-ngrams over time froma a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task; Semantic Textual Similarity*, pages 241–247. Association for Computational Linguistics.

Google. 2015. Google Translate.

Eija-Riitta Grönros. 2006. Arkikielesta yleiskieleen (From everyday language to standard language). *Kielikello*, 4.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2013. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, pages 1–39.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511. Springer.

Institute of Languages of Finland. 2014. *Kielitoimiston sanakirja / Dictionary of the Institute for Languages of Finland*. Number 35 in Kotimaisten kielten keskuksen verkkojulkaisuja. Kotimaisten kielten keskus / Institute for Languages of Finland.

Jenna Kanerva, Juhani Luotolahti, Veronika Laippala, and Filip Ginter. 2014. Syntactic n-gram collection from a large-scale corpus of internet finnish. In *Proceedings of the Sixth International Conference Baltic HLT*.

Adam Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on web as corpus. *Computational Linguistics*, 29:333–347.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 446–457. Springer Berlin Heidelberg.

Veronika Laippala, Timo Viljanen, Antti Airola, Jenna Kanerva, Sanna Salanterä, Tapio Salakoski, and Filip Ginter. 2014. Statistical parsing of varieties of clinical Finnish. *Artificial Intelligence in Medicine*, 61(3):131–136.

Sara Laviosa-Braithwaite. 1995. Comparable corpora: towards a corpus linguistic methodology for the empirical study of translation. In M. Thelen and B. Lewandowska-Tomaszczyzk, editors, *Translation and Meaning Part 3. Proceedings of the Maastricht Session of the 2nd International Maastricht-Lodz Duo Colloquium on "Translation and Meaning"*., pages 153–163. Hoogeschool Maastricht, Maastricht.

Sara Laviosa. 2002. *Corpus-based Translation Studies: Theory, Findings, Applications*. Rodopi, Amsterdam, New York.

Lotta Lehti and Veronika Laippala. 2014. Style in french politicians' blogs: Degree of formality. *Language at Internet*, 11.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49.

Kirsti Mäkinen. 1989. Sanojen tyyliväri. In *Nykysuomen sanavarat*, pages 200–212. WSOY.

Anna Mauranen. 2000. Strange strings in translated language: A study on corpora. In *Intercultural Faultlines. Research Models in Translation Studies 1*, pages 119–141. St. Jerome Publishing, Manchester.

Alejandro Mosquera and Paloma Moreda. 2011. The use of metrics for measuring informality levels in web 2.0 texts. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Sampo Nevalainen. 2003. Käännöskirjallisuuden puhekielisyyksistä - kaksinkertaista illuusiota? (on the informality of translated literature - a double illusion?). *Virittäjä*, (1):2–26.

Marius Popescu. 2011. Studying translationese on the character level. In *Proceedings of Recent Advances in Natural Language Processing*, pages 634–639.

Tiina Puurtinen. 2003. Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and Linguistic Computing*, 18(4):389–406.

John M. Sinclair. 1996. *Preliminary recommendations on Corpus Typology*. http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html.