A Corpus of Native, Non-native and Translated Texts

Sergiu Nisioi◇, Ella Rabinovich*, Liviu P. Dinu◇, Shuly Wintner*
◇Center for Computational Linguistics,
University of Bucharest
*Department of Computer Science,
University of Haifa
sergiu.nisioi@gmail.com, ellarabi@csweb.haifa.ac.il,
ldinu@fmi.unibuc.ro, shuly@cs.haifa.ac.il

Abstract

We describe a monolingual English corpus of original and (human) translated texts, with an accurate annotation of speaker properties, including the original language of the utterances and the speaker's country of origin. We thus obtain three sub-corpora of texts reflecting native English, non-native English, and English translated from a variety of European languages. This dataset will facilitate the investigation of similarities and differences between these kinds of sub-languages. Moreover, it will facilitate a unified comparative study of translations and language produced by (highly fluent) non-native speakers, two closely-related phenomena that have only been studied in isolation so far.

Keywords: Corpus linguistics, Translation, Bilingualism, Second language acquisition

1. Introduction

Interlanguage is the entire linguistic system that emerges when second language learners express meaning in the target language (Selinker, 1972). One of the main characteristics of this system are the so-called *cross-linguistic influences* - a cover term proposed by Kellerman and Sharwood-Smith (1986) to denote various phenomena that stem from language contact situations such as transfer, interference, avoidance, borrowing, etc. Language transfer, in particular, is the influence resulting from similarities and differences between the target language (L2) and any other language that has been previously acquired (L1) (Odlin, 1989).

The centerpiece of interlanguage resides in the concept of *fossilization*, i.e., "the permanent cessation of target language learning before the learner has attained the L2 norms at all levels of linguistic structure" (Selinker, 1972; Han, 2013). In other words, the language learning process *rarely* (Lardiere, 2006) leads to a full, or native-like acquisition of the target language. Under this assumption, non-native speakers are distinguishable from native speakers, no matter how proficient they are in the target language. However, researchers (Birdsong, 1992; White and Genesee, 1996; Long, 2003) indicate that fossilization does not occur on all levels of linguistic structure, rather stabilization might occur on certain levels, while others continue to develop.

Toury (1979) claimed that interlanguage is not only present in the context of non-native speakers, but also, to a certain degree, in translated texts, which presumably reflect both the artifacts of the translation process and traces of the original language. Research in translation studies (Frawley, 1984; Baker, 1996) indicates that translated texts have unique characteristics. Gellerstam (1986) suggested that the differences between original and translated texts do not indicate poor translation but rather a *statistical phenomenon*, caused by a systematic transfer of the source language into the target one.

Specific phenomena that characterize the contact between two languages in non-native utterances or translated texts have so far been studied in isolation, both in the linguistic literature and in terms of computational investigations. In this paper we describe a corpus constructed from original English utterances (where we differentiate between native and non-native speakers based on their country of origin) and English translations from a variety of European languages. This corpus will be instrumental for research that aims to uniformly address both second language acquisition (more specifically, the language of highly fluent non-native speakers) and human translation.

The corpus we describe is based on Europarl (Koehn, 2005), and is the first corpus that allows a uniform comparative study of both phenomena (translation and language acquisition). The texts contained in the corpus are uniform in terms of style, respecting the European Parliament's formal standards. In addition, the original English utterances are accurately annotated with speakers' data, including knowledge about the speakers' native language[1].

Corpora of original and translated language are essential for empirical investigation of theoretically-motivated hypotheses from the field of translation studies (Baker, 1996). In particular, these corpora have been extensively used for investigation of transfer of the source language into the target one (van Halteren, 2008; Popescu, 2011; Koppel and Ordan, 2011). Learner corpora are a different type of resource, constructed from texts written by non-native language learners, most often students acquiring a foreign language. Such corpora can reveal useful insights about the developmental process of language acquisition.

Furthermore, such corpora also have practical applications, e.g., for automatic error correction (Dale and Kilgarriff, 2011). Another prominent computational application is the task of native language identification of English learners (Koppel et al., 2005; Tetreault et al., 2013; Nisioi, 2015a). Similarly, corpora of translated texts have been instrumental in automatic identification of translation, where much research demonstrates that machine learning techniques can discriminate between original and translated

[1]The corpus is freely available at `http://nlp.unibuc.ro/resources/ENNTT.tar.gz`

texts with very high accuracy in both supervised (Baroni and Bernardini, 2006; Ilisei and Inkpen, 2011; Volansky et al., 2015) and unsupervised scenarios (Rabinovich and Wintner, 2015; Nisioi, 2015b). Such studies can be of much use for training better translation and language models for machine translation (Lembersky et al., 2012; Lembersky et al., 2013). The corpus that we describe will facilitate computational research into the similarities and differences between the two types of language contact (second language learning and translation), hopefully leading to better solutions for the related computational tasks.

2. Corpus pre-processing and annotation of the translation direction

The Europarl corpus is extracted from the collection of the proceedings of the European Parliament (Koehn, 2005), dating back to 1996. The transcriptions are produced as follows: (1) the utterances of the speakers are transcribed; (2) the transcriptions are sent to the speaker who may suggest minimal editing without changing the content; (3) the edited version is then translated (i.e., the texts are not a product of simultaneous interpreting). The EU Parliament requires the translations to be produced by *native* speakers of the target language (Pym et al., 2013). In each subcorpus, each paragraph is annotated with meta-information; in particular, the original language in which the paragraph was uttered. Unfortunately, the meta-information pertaining to the original language of the utterances is frequently missing and in some (rare) cases this information is inconsistent: the source language tag is not identical across translations of the same paragraph. Additionally, the Europarl corpus contains several bilingual (sentence-aligned) sub-corpora with no annotations.

To minimize the risk of erroneous information, we process the corpus as follows. First, we propagate the meta-information from the monolingual texts to the bilingual sub-corpora, such that each sentence pair is annotated with the original language in which it was uttered. We iterate this process five times, extracting the meta-information from the original monolingual corpora in five languages: English, French, German, Italian, and Spanish; note that not all monolingual corpora are identical: some are much larger than others. For the same reason, not all English sentences are represented in the five monolingual corpora, and therefore some sentence pairs have less than five annotations for original language. We restrict the final corpus only to those sentences that have five non-contradicting annotations. The speaker information (ID, original language) in the filtered corpus is therefore highly accurate.

A detailed description of the preprocessing and annotation procedure can be found in Rabinovich et al. (2015).

3. Indication of non-native English speakers

3.1. Manual speaker disambiguation

The native language of a member of the European Parliament (MEP) is considered to be the one corresponding to the country he or she represents. We acknowledge that the country information is not strictly identical with the native language of a MEP, since EU states can be multilingual and one can be part of a minority group within that country. The native state of each MEP is extracted from the Europarl website[2], along with the standardized version of their names, places of birth, pictures and IDs.

A major obstacle is aligning the standardized names with all the different variants used to refer to the MEPs. For example, the name *Nuala Ahern* (id: 2230) from Ireland can be found in the following forms in the corpus: "Ahern", "Ahern (Verts/ALE).", "Ahern, Nuala (Verts/ALE)" and more. We were faced with building a many to one relation, from all the different variants used in Europarl proceedings to an actual entry that contains the standardized name, id, country of origin and other details. In total we manually linked over two thousand variants of names with the corresponding id and standard name.

3.2. Crawling the sessions

An additional ambiguity found in the proceedings is related to the usage of the same name for two different members of the parliament: "Ryan" can be encountered as a reference to either Eoin Ryan (ID: 28113) or Richie Ryan (ID: 220), both from Ireland.

We solved these ambiguities by crawling all the sessions again and using the picture ID of each speaker as an indicator if two names refer to the same person or not. We assume that the editors do not use the same name for two MEPs in the same session, so we tag all the ambiguous references of a name per session. The heuristics were applied on the sessions after 1999, since before this date the website had a different structure that is not compatible with our approach. In total, we disambiguated 78 different MEPs that present this type of ambiguity, while the small number of remaining ambiguities were completely eliminated from the extraction process.

3.3. Corpus properties

Table 1 reports some statistical data on the corpus (after tokenization). We augmented each sentence in the obtained dataset with the following information:

NAME	speaker's name as it appears written in the session
LANGUAGE	original language in which the sentence was uttered
SESSION_ID	the name of the corresponding protocol source file
SEQ_SPEAKER_ID	sequential number of the speaker within a session

Sentences uttered in English are annotated with additional information:

STATE	the EU state represented by the MEP
MEPID	the ID used by the Europarl website to display the MEPs online images

[2]http://www.europarl.europa.eu/meps/en/directory.html

sub-corpus	# sentences	# tokens	# types
native English	116,341	3,051,082	36,323
non-native English	29,734	783,742	18,419
translations into English	738,597	22,309,296	80,254

Table 1: The number of sentences, tokens, and types corresponding to each sub-corpus

4. Learner corpus vs. non-native corpus

This resource is novel in comparison to other corpora which contain non-native productions, such as ICLE (Sylviane Granger, Estelle Dagneaux, Fanny Meunier and Magali Paquot, 2003), EFCAMDAT (Jeroen Geertzen, Theodora Alexopoulou and Anna Korhonen, 2014) or TOEFL-11 (Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, Martin Chodorow, 2014). Among other factors influencing learner corpora, there are at least three major differences compared to Europarl: first, the texts reflected in learner corpora are restricted to the requirements of the exercises of the tests; second, student learners have various proficiency levels (beginner, intermediate, advanced) and errors can still be frequently encountered; and, finally, according to Horwitz et al. (1986), students taking a test can find themselves under the influence of foreign language anxiety. This type of anxiety frequently appears in testing situations in which students having acquired certain grammatical rules can still produce errors because of the pressure of being exposed to evaluation.

In contrast to learner corpora, Europarl contains high-proficiency English of advanced speakers who master the language well enough to express themselves fluently. Members of the European Parliament have the right to use any of the EU's 24 official languages when speaking in Parliament, and the fact that some of them prefer to use English further suggests that foreign language anxiety is less present, if not completely absent, and that the speakers have a considerable degree of confidence in their language skills. In addition, the texts in the European Parliament share common aspects of formal style that make the statistical analysis unbiased with respect to genre (Brooke and Hirst, 2011). Even given these circumstances, however, we hypothesize that it is possible to distinguish native from non-native utterances based on the idea that fossilized linguistic structures are present in the language of non-native speakers.

5. Evaluation and results

We conduct two sets of experiments: (1) a three-way classification of native, non-native and translated productions, and (2) an investigation of lexical diversity and vocabulary richness, as reflected by the type-to-token ratio (TTR) of all production types.

We pre-process the (tokenized) datasets by splitting them into chunks of approximately 2000 tokens, respecting sentence boundaries and preserving punctuation. We use function words (conjunctions, preposition, adverbs, etc.) as features to classify the texts. These are non-topical words that reflect grammatical and structural properties of the text, and have been shown to be effective in numerous text classification studies on both translated and non-native productions (see Section 1.). The features are weighted using the log-entropy scheme (Dumais, 1991), an approach suggested by previous studies (Jarvis et al., 2012) on native language identification. We use a support vector classifier (libsvm) with a linear kernel (Chang and Lin, 2011) and parameter tuning based on grid search. We ensured that each class is represented by an equal number of training examples to have a uniform baseline. The ten-fold cross-validation results of the three-way classification (discriminating between native, non-native and translated texts) are reported as a confusion matrix in Table 2; the overall classification accuracy is 90.78%.

actual / predicted	native	non-native	translated
native	92.86	4.75	2.38
non-native	7.44	88.13	4.43
translated	2.68	5.95	91.37

Table 2: Confusion matrix listing the percentage of classified and misclassified documents

Two main observations emerge from the classification results: (1) native texts are easily distinguishable from the other two classes, and (2) translated texts are often misclassified as non-native, but not vice-versa. This could be an indicator that translations and non-native utterances reflect different types of interlanguage, despite the undoubted similarities between the two.

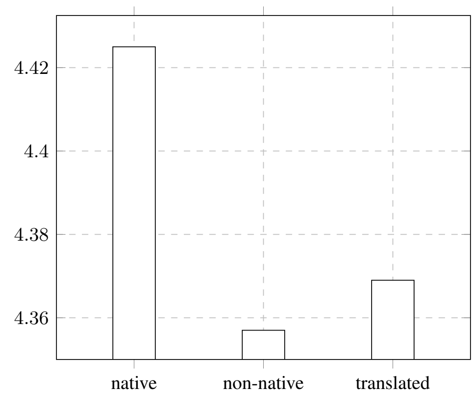

Figure 1: Logarithmic TTR comparison – native, non-native and translated productions

Furthermore, translated texts tend to exhibit less lexical diversity and vocabulary richness (Al-Shabab, 1996), reflected by their lower type-to-token ratio compared to that of native productions. We calculate the logarithmic TTR of non-native texts and compare it to that of native texts and translations. As shown in Figure 1, the log-TTR of native texts is higher than that of translated texts. Moreover, the TTR of non-native productions is lower than that of native texts, and, more pertinently, it is also lower than that of translations, mirroring the fact that the lexical diversity of (highly competent) non-native speakers is poorer than that of translators, who translate into their mother tongue.

6. Conclusions and future work

We developed a high-quality English Europarl dataset comprising native, non-native and translated texts. The corpus is uniformly processed and accurately annotated; it will be instrumental in research of second language acquisition, translation studies and, more prominently, in unified investigations of transfer-related, as well as source-language independent phenomena across both domains. Our future plans include conducting this cross-disciplinary comparative study, as well as extending the corpus to additional domains and languages.

Acknowledgments

This research was supported by the Israeli Ministry of Science and Technology and partly by UEFISCDI, PNII-IDPCE-2011-3-0959. We are grateful to Noam Ordan for his advices and helpful suggestions.

7. Bibliographical References

Al-Shabab, O. S. (1996). *Interpretation and the language of translation: creativity and conventions in translation*. Janus, Edinburgh.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. *BENJAMINS TRANSLATION LIBRARY*, 18:175–186.

Baroni, M. and Bernardini, S. (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.

Birdsong, D. (1992). Ultimate attainment in second language acquisition. *Language*, 68(4):706–755.

Brooke, J. and Hirst, G. (2011). Native language detection with 'cheap' learner corpora. In *Conference of Learner Corpus Research (LCR2011)*, Louvain-la-Neuve, Belgium. Presses universitaires de Louvain.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Dale, R. and Kilgarriff, A. (2011). Helping our own: The hoo 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249. Association for Computational Linguistics.

Dumais, S. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236.

Frawley, W. (1984). Prolegomenon to a theory of translation. *Translation: Literary, linguistic and philosophical perspectives*, 159:175.

Gellerstam, M. (1986). Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, pages 88–95.

Han, Z. (2013). Forty years later: Updating the fossilization hypothesis. *Language teaching*, 46(02):133–171.

Horwitz, E. K., Horwitz, M. B., and Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, 70(2):pp. 125–132.

Ilisei, I. and Inkpen, D. (2011). Translationese traits in Romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications*, 2(1-2).

Jarvis, S., Castañeda-Jiménez, G., and Nielsen, R. (2012). Detecting L2 Writers' L1s on the Basis of Their Lexical Styles. In Scott Jarvis et al., editors, *Approaching Language Transfer through Text Classification*, pages 34–70. Multilingual Matters.

Kellerman, E. and Sharwood-Smith, M. (1986). *Crosslinguistic Influence in Second Language Acquisition*. Language Teaching Methodology Series. Pearson College Division.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA, June. Association for Computational Linguistics.

Koppel, M., Schler, J., and Zigdon, K. (2005). Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628, Chicago, IL. ACM.

Lardiere, D. (2006). *Ultimate Attainment in Second Language Acquisition: A Case Study*. L. Erlbaum.

Lembersky, G., Ordan, N., and Wintner, S. (2012). Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825, December.

Lembersky, G., Ordan, N., and Wintner, S. (2013). Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4):999–1023, December.

Long, M. H. (2003). Stabilization and fossilization in interlanguage development. *Language Research*, 12:335–73.

Nisioi, S. (2015a). Feature analysis for native language identification. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, Proceedings*, Lecture Notes in Computer Science. Springer.

Nisioi, S. (2015b). Unsupervised classification of translated texts. In Chris Biemann, et al., editors, *Natural Language Processing and Information Systems - 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015 Passau, Germany, June 17-19, 2015 Proceedings*, volume 9103 of *Lecture Notes in Computer Science*, pages 323–334. Springer.

Odlin, T. (1989). *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge Applied Linguistics. Cambridge University Press.

Popescu, M. (2011). Studying translationese at the character level. In Galia Angelova, et al., editors, *Proceedings of RANLP-2011*, pages 634–639.

Pym, A., Grin, F., Sfreddo, C., and Chan, A. L. (2013). *The status of the translation profession in the European Union*. Anthem Press.

Rabinovich, E. and Wintner, S. (2015). Unsupervised identification of translationese. *TACL*, 3:419–432.

Rabinovich, E., Wintner, S., and Lewinsohn, O. L. (2015). The haifa corpus of translationese. *CoRR*, abs/1509.03611.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1–4):209–232.

Tetreault, J., Blanchard, D., and Cahill, A. (2013). A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, GA, USA, June. Association for Computational Linguistics.

Toury, G. (1979). Interlanguage and its manifestations in translation. *Meta: Journal des traducteursMeta:/Translators' Journal*, 24(2):223–231.

van Halteren, H. (2008). Source language markers in EUROPARL translations. In Donia Scott et al., editors, *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 937–944, Morristown, NJ, USA. Association for Computational Linguistics.

Volansky, V., Ordan, N., and Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118, April.

White, L. and Genesee, F. (1996). How native is near-native? the issue of ultimate attainment in adult second language acquisition. *Second language research*, 12(3):233–265.

8. Language Resource References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, Martin Chodorow. (2014). *ETS Corpus of Non-Native Written English*. Web Download. Philadelphia: Linguistic Data Consortium, LDC2014T06, 1.0, ISLRN 640-546-772-297-1.

Jeroen Geertzen, Theodora Alexopoulou and Anna Korhonen. (2014). *EF-Cambridge Open Language Database*. University of Cambridge, Department of Theoretical and Applied Linguistics at the University of Cambridge in partnership with Education First, 1.0.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier and Magali Paquot. (2003). *International Corpus of Learner English ICLE v2*. Centre for English Corpus Linguistics - CECL, 2.0.