Stylometry in Computer-Assisted Translation: Experiments on the Babylonian Talmud

Emiliano Giovannetti[1], **Davide Albanesi**[1], **Andrea Bellandi**[1],
David Dattilo[2], **Felice Dell'Orletta**[1]
[1] Istituto di Linguistica Computazionale, Via G. Moruzzi 1, 56124, Pisa
name.surname@ilc.cnr.it
[2] Progetto Traduzione Talmud Babilonese S.c.a r.l., Lungotevere Sanzio 9, 00153 Roma
david.dattilo@talmud.it

Abstract

English. The purpose of this research is to experiment the application of stylometric techniques in the area of Computer-Assisted Translation to reduce the revision effort in the context of a collaborative, large scale translation project. The obtained results show a correlation between the editing extent and the compliance to some specific linguistic features, suggesting that supporting translators in writing translations following a desired style may actually reduce the number of following necessary interventions (and, consequently, save time) by revisors, editors and curators

Italiano. *Lo scopo di questa ricerca è la sperimentazione dell'applicazione di tecniche stilometriche nell'area della Traduzione Assistita dal Calcolatore per ridurre il lavoro di revisione nel contesto di un progetto di traduzione collaborativo di ampia scala. I risultati ottenuti mostrano una correlazione tra l'entità delle modifiche effettuate e la conformità ad alcune specifiche caratteristiche linguistiche, suggerendo che supportare i traduttori nel processo traduttivo seguendo uno stile desiderato possa effettivamente ridurre il numero di interventi necessari (e, quindi, risparmiare tempo) da parte di revisori, redattori e curatori.*

1 Introduction

The Progetto Traduzione Talmud Babilonese[1] (PTTB) is a research and education project carrying out the digitized Italian translation of the Babylonian Talmud (BT), a fundamental book of the Jewish tradition, covering every aspect of human knowledge: law, science, philosophy, religion and even aspects of everyday life. The translation of the Talmud has been assigned to more than 50 scholars comprising expert translators, trainee translators, instructors, editors and curators.

The translated text is accompanied by the explanations and comments on specific words and subjects, and also by illustrative sheets for the various scientific, historical and linguistic topics addressed inside the Talmudic discussions. However, the Project objectives include more than the translation of the Talmud: the whole work has been set up to be completely digital. Everything, from the very first activities of assigning users to the translation of specific chapters to supporting in the definition of the final printing layout, revolves around Traduco, a collaborative web-based Computer-Assisted Translation (CAT) tool developed within the Project.

Today, many CAT tools, both commercial and freely distributed, are already available, but they have been designed for the translation of technical manuals or domain-specific texts (legislative, medical) with the main purpose of speeding up the translation process.

The BT is a very complex text in many ways: its content, the different, ancient, languages it is composed of (though mainly Babylonian Aramaic and Mishnaic Hebrew), and the history of its composition over the centuries. For these reasons, the approach we adopted for the development of Traduco had to take into account the needs of translators working on a text with very particular interpretative issues. Traduco allows a user to distinguish the literal part of the translation (in bold, see Fig.1) from explicative additions, included by translators to make the most difficult passages clearer to readers. Indeed, a full understanding of this kind of texts requires a translation

[1] www.talmud.it (last access: 25/07/2017)

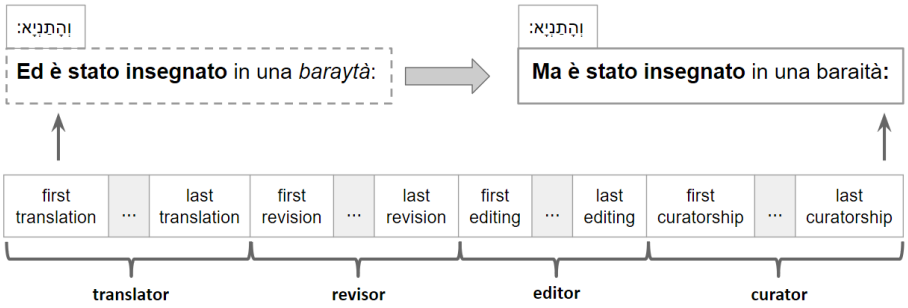

Figure 1: The life cycle of a translated string.

to be enriched with comments, notes, and glossary entries. Furthermore, due to the complexity of the inner structure of the BT, Traduco allows users to split autonomously their translations into "strings" (representing, typically, a sentence, see Fig.1), gathered into "logical units" . Finally, Traduco provides a collaborative and training environment allowing a translator to instantly consult translations done by others, when portions of text (and sometimes even a single word) are difficult to interpret and translate. For a comprehensive description of how Traduco works refer to (Giovannetti et al., 2017). The size and complexity of the text and the need to produce a printed version of the BT translation, required a team of users composed of translators, revisors, editors, curators and supervisors.

The whole translation workflow can be described by following the "life-cycle" of each string (Fig.1). It all starts as soon as the coordinator of the translation assigns a chapter to a specific translator: the first phase of the work, the **translation**, begins. The translation is carried out by scholars having two distinct profiles: expert translators, working autonomously, and trainee translators, these latter being constantly supported by instructors monitoring online their work and providing face-to-face lectures. Once the translation of a specific chapter is concluded, the **revision** phase starts. Revisors are chosen among the most expert scholars involved in the Project and their main task is to verify if translators have understood correctly the meaning of each string. They also have to check if the domain terms (if present) have been appropriately annotated and explained in the relative glossary entry. After the content has been revised, the **editing** starts. In this phase, a formal and linguistic control of the translation is carried out, where the editors ensure that the translated strings are syntactically and orthographically correct. Contextually, each string can be enriched, if needed to help in the understanding of the text, with pictures and tables. The last phase is the **curatorship**, during which one more general control of the translation is done before proceeding with the final exporting and printing of the volume. As we showed in a previous work (Bellandi et al., 2016), the introduction of Natural Language Processing techniques in CAT tools can bring concrete advantages to the translation work and pave the way to innovative research in the area of NLP for Digital Humanities.

One way to ease the translation of a text as the BT is to assist translators in writing, in the first place, good translations requiring as few corrections as possible by revisors, editors and curators. In other words, we want to find a way of alerting a user about to submit a new translation by highlighting specific characteristics of the sentence that may further require a revision and, thus, slow down the overall translation process.

To do that, we chose to experiment the application of stylometric measures to Italian translations. The assumption we would like to prove is that translations being more compliant to the style of revisors will actually require less revisions. If that will be demonstrated, we may develop a strategy to alert translators of potential "unfit" translations and suggest a way to improve them in order to minimize the following editing for revision, editing, and curatorship.

2 Background

Over the last ten years, Natural Language Processing (NLP) techniques combined with machine learning algorithms started being used to investigate the "form" of a text rather than its content. The range of tasks sharing this approach to the analysis of texts is wide, ranging e.g. from na-

tive language identification (see among the others (Koppel et al., 2005) and (Wong and Dras, 2009)), author recognition and verification (see e.g. (van Halteren, 2004), authorship attribution (see (Juola, 2008) for a survey), genre identification (Mehler et al., 2011) to readability assessment (see (Dell'Orletta et al., 2014) for an updated survey) or tracking the evolution of written language competence (Richter et al., 2015). Besides obvious differences at the level of the considered task, they share a common approach: they succeed in determining the language variety, the author, the text genre or the level of readability of a text by exploiting the distribution of features automatically extracted from texts. To put it in van Halteren words (van Halteren, 2004), they carry out "linguistic profiling" of texts, i.e. "the occurrences of a large number of linguistic features in a text, either individual items or combinations of items, are counted" in order to determine "how much [...] they differ from the mean observed in a profile reference corpus".

To the best of our knowledge, however, no research has been documented in literature about the application of stylometric or readability techniques to Computer-Assisted Translation. For this reason, a comparison with existing approaches and results was not possible.

On the other hand, the use of stylometry and readability in translation studies is described in several works, especially in the analysis of literary texts (Heydel and Rybicki, 2012), (Kolahi and Shirvani, 2012), (Acar and İŞİSAĞ, 2017), (Huang, 2015) and some of them provide useful indications on how the personal writing style (being it, in our case, that of a translator or a revisor) can influence the final translation (Baker, 2000) and (Rybicki, 2012).

3 Methodology

To construct the dataset we exploited the versioning features of Traduco. As a matter of fact, every version of most of textual resources (currently: strings, notes, and glossary entries) is stored in the database. It is thus possible to compare earlier versions of translations (i.e. those inserted by translators) with the latest ones (i.e. those that have been completely revised) in order to analyse the differences between them. For the experiment, we built two datasets using textual segments of different granularity: blocks for the DS_{bl} dataset and logical units for DS_{lu}.

In more details, each dataset has been built as a set of textual segment pairs extracted from the translations of the tractates Berakhot and Ta'anit, respectively composed, in their revised versions, of 216138 and 81696 tokens. Given a pair (s_1, s_2), the first component s_1 represents the last translation of a block or logical unit inserted by the translator[2] and the second component s_2 its very last version (i.e. that following the revision, editing and curatorship phases). Concerning the size, DS_{bl} was composed of 554 blocks and DS_{lu} of 4303 logical units. Each logical unit is composed, in average, by 5.62 strings, while each string is composed, in average, by 12.5 tokens.

Once the datasets were ready, we had to attribute to each pair a "revision measure" to quantify the difference between s_1 and s_2 in terms of both words and characters. For this purpose we chose to adopt the Levenshtein distance. Since Traduco is equipped with a spell checker, we assumed that the presence of typos should not impact on the revision measure significantly.

As the next step we investigated the presence of linguistic features extracted from those texts belonging to the s_1 component of the pairs correlating with the revision measures. For this purpose, the considered texts were automatically POS tagged by the Part-Of-Speech tagger described in (Cimino and Dell'Orletta, 2016) and dependency parsed by the DeSR parser (Attardi et al., 2009) using multilayer perceptron as learning algorithm. For the specific concerns of this study, we focused on a wide set of features ranging across different linguistic description levels which are typically used in studies focusing on the "form" of a text, e.g. on issues of genre, style, authorship or readability. This represents a peculiarity of our approach: we resort to general features qualifying the lexical and grammatical characteristics of a text, rather than ad hoc features, specifically selected for a given text type or task. The set of selected features is organised into four main categories defined on the basis of the different levels of linguistic analysis automatically carried out (tokenization, lemmatization, morphosyntactic tagging and dependency parsing): i.e. raw text features, lexical features as well as morpho-syntactic and syntactic features.

[2] sometimes translators insert a draft version of a translation, to be completed later: for this reason we chose to take the last translation available.

	DS_{lu}		DS_{bl}	
features	**char**	**token**	**char**	**token**
Number of tokens	**0.65**	**0.68**	**0.84**	**0.85**
Arity of verbs	**0.62**	**0.64**	**0.83**	**0.83**
Number of main verbs	**0.62**	**0.64**	**0.83**	**0.83**
Number of prepositional 'chains'	**0.57**	**0.60**	**0.81**	**0.82**
Number of sentences	**0.49**	**0.53**	**0.80**	**0.80**
Number of verb roots	**0.49**	**0.53**	**0.79**	**0.79**
Number of subord clauses	**0.37**	**0.38**	**0.68**	**0.68**
% of verbs with 5 syntactic dependent	-	-	**0.37**	**0.36**
% of first person singular of verbs	-	-	**0.31**	0.32
% of subjunctive auxiliary-verbs	-	-	0.31	0.30
% of locative modifier	-	-	0.31	0.31
% of second person plural	-	-	0.31	0.31
% of verb in infinitive mood	-	-	0.30	**0.32**
% of demonstrative determiner	-	-	0.30	-
% of "balanced" punctuation	-	**0.33**	-	-
Average of length of dependency links	**0.35**	**0.37**	-	-
Longest dependency links	**0.34**	**0.34**	-	-
Average of main verbs for sentence	**0.33**	**0.32**	-	-
Average length of subord clauses	**0.31**	**0.31**	-	-

Table 1: Spearman's rank correlation coefficients (in bold with $p < 0.001$, otherwise with $p < 0.05$) calculated on both datasets and the two revision measures (distance per character and per token); values below 0.3 have been discarded.

To conclude our experiment we applied the Spearman's rank correlation coefficient to assess the presence of a statistical dependence between our revision measures and the calculated linguistic features.

4 Evaluation

The results (filtered by keeping just the features providing coefficients greater or equal than 0.3) are summarized in Table 1. Apart from the expected correlations between the size of the texts (represented by raw text features such as "Number of tokens" and "Number of sentences") and the revision measures, we found some significative correlations, in relation to morphosyntactic and syntactic features. Most of the morphosyntactic features involve verbs: the presence of main verbs, the mood, the tense, etc.

Some of the syntactic features showing a correlation, such as the length of dependency links, the length of subordinate clauses and the number of prepositional chains, are particularly interesting. As a matter of fact, these linguistic features are typically used as indicators of linguistic complexity: indeed, portions of translated text constituted of long and articulated syntactic structures appear to be more subjected to revisions. As expected, the correlation of some of these syntactic features, such as the number of prepositional chains, appears to be proportional to the size of the analysed text (as in the blocks wrt the logical units in the datasets), since the presence of deeper syntactic structures increases and the text, at least in principle, gets more linguistically complex.

5 Conclusions

The experiment described in this paper proves that the application of NLP to CAT contexts can open new research perspectives and, more importantly, may be of concrete help in real usage translation scenarios. The proposed methodology can be applied, in principle, to any translation project in which a revision phase is a part of the whole translation workflow and where an history of the edits is maintained. The same analysis could be performed on different languages depending solely on the availability of the suitable NLP tools. Some of the NLP techniques adopted for the stylometric analysis of Italian may also be adapted to the processing of Mishnaic Hebrew and Aramaic (the

main source languages). The automatic linguistic analysis of Mishnaic Hebrew, for example, is being experimented (Pecchioli, 2017). However, an analysis of the style (or complexity) of the source text, though interesting in a historical text analysis perspective, would be pointless in the specific context of revision support in computer-assisted translation.

The correlation we found between the revision measures and some linguistic features (some of which are actually used as indicators of linguistic complexity) is the first step towards the design of a technique aimed at providing users a way of writing translations less prone to revisions. In this way, the whole translation workflow would benefit from a reduced time in the revision, editing and curatorship phases. Once the approach will be defined, the relative software will be implemented as a new component of Traduco. Moreover, the possibility of suggesting a way of writing "better" translations (at least wrt revisor's style) will be exploited in the education of trainee translators.

6 Acknowledgment

This work was partially supported by the project TALMUD and carried out in the context of the scientific partnership between S.c.a r.l. "Progetto Traduzione del Talmud Babilonese" and ILC-CNR and on the basis of the regulations stated in the "Protocollo d'Intesa" (memorandum of understanding) between the Italian Presidency of the Council of Ministers, the Italian Ministry of Education, Universities and Research, the Union of Italian Jewish Communities, the Italian Rabbinical College, and the Italian National Research Council (21 January 2011).

References

Alpaslan Acar and Korkut Uluç İŞİSAĞ. 2017. Readability and comprehensibility in translation using reading ease and grade indices. *International Journal of Comparative Literature and Translation Studies*, 5(2):47–53.

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, and Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of the 2nd Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, (EVALITA 2009)*.

Mona Baker. 2000. Towards a methodology for investigating the style of a literary translator. *Target. International Journal of Translation Studies*, 12(2):241–266.

Andrea Bellandi, Giulia Benotto, Gianfranco Di Segni, and Emiliano Giovannetti. 2016. Investigating the application and evaluation of distributional semantics in the translation of humanistic texts: a case study. In *Proceedings of the* 2^{nd} *Workshop on Natural Language Processing for Translation Memories*, pages 6–11.

Andrea Cimino and Felice Dell'Orletta. 2016. Building the state-of-the-art in POS tagging of italian tweets. In *Proceedings of Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. In John Benjamins Publishing Company, editor, *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics, 165:2*, pages 163–193.

Emiliano Giovannetti, Davide Albanesi, Andrea Bellandi, and Giulia Benotto. 2017. Traduco: A collaborative web-based cat environment for the interpretation and translation of texts. *Digital Scholarship in the Humanities*, 32(suppl_1):i47–i62.

Magda Heydel and Jan Rybicki. 2012. The stylometry of collaborative translation. woolf's night and day in polish. In *Digital Humanities 2012 Conference Abstracts*, pages 212–217.

Libo Huang. 2015. Readability as an indicator of self-translating style: A case study of eileen chang. In *Style in Translation: A Corpus-Based Perspective*, pages 95–111. Springer.

Patrick Juola. 2008. Authorship attribution. In *Now Publishers Inc*.

Sholeh Kolahi and Elaheh Shirvani. 2012. A comparative study of the readability of english textbooks of translation and their persian translations. *International Journal of Linguistics*, 4(4):344.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics, vol. 3495, LNCS, Springer–Verlag*, pages 209–217.

Alexander Mehler, Serge Sharoff, and Marina (Eds.) Santini. 2011. Genres on the web. computational models and empirical studies. In *Springer Series: Text, Speech and Language Technology*.

Alessandra Pecchioli. 2017. Elaborazione del linguaggio naturale (nlp) in ebraico: il caso dell'analisi linguistica automatica applicata all'ebraico mishnaico

del talmud. Oral communication, sep. XXXI Convegno AISG 2017 - Nuovi studi sullEbraismo, 4-6 settembre 2017, Ravenna, Italy.

Stefan Richter, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2015. Tracking the evolution of written language competence: an nlpbased approach. In Cristina Bosco, Sara Tonelli, and Massimo Zanzotto, editors, *Proceedings of the Second Italian Conference on Computational Linguistics - CLiC-it 2015*, pages 236–240.

Jan Rybicki. 2012. The great mystery of the (almost) invisible translator. *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*, 231:231–248.

Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In John Benjamins Publishing Company, editor, *Proceedings of the Association for Computational Linguistics (ACL04)*, pages 200–207.

Sze-Meng Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*.