Enriching a Multilingual Terminology Exploiting Parallel Texts: An Experiment on the Italian Translation of the Babylonian Talmud

Angelo Mario Del Grosso, Emiliano Giovannetti, Simone Marchi
Istituto di Linguistica Computazionale "A. Zampolli"
{name.surname}@ilc.cnr.it

Abstract

English. Parallel texts can represent an extremely useful source of information in a number of text and linguistic processing tasks. In this work we show an experiment conducted on the Italian translation of the Babylonian Talmud, a text we have analyzed and processed to support in the construction of a multilingual Hebrew/Aramaic/Italian terminological resource. The approach we adopted comprised: i) the TEI encoding of the text, ii) the automatic extraction of the Italian terms, iii) the addition of Hebrew/Aramaic terms via word-by-word alignment, iv) the revision of the obtained results.

Italiano. I testi paralleli possono costituire una fonte estremamente utile di informazioni per numerosi task di elaborazione del testo e della lingua. In questo lavoro illustriamo un esperimento condotto sulla traduzione italiana del Talmud babilonese, un testo che abbiamo analizzato ed elaborato per supportare la costruzione di una risorsa terminologica multilingue in Ebraico, Aramaico e Italiano. L'approccio adottato comprende: i) la codifica TEI del testo, ii) l'estrazione automatica dei termini italiani, iii) l'aggiunta dei termini ebraici e aramaici tramite tecniche di allineamento parola per parola, iv) la revisione dei risultati ottenuti.

1 Introduction

Translation is the only way of making a text accessible to people that do not understand the language the original text is written in. Translation, in other words, allows to build bridges between peoples and cultures. It is no coincidence that it has been through a translation, contained in the well-known Rosetta Stone, that Egyptian hieroglyphs could be deciphered. The work we here describe is based upon a similar principle: how to exploit the translation in a "known" language of a text written in an "unknown" language to derive some linguistic information from the latter. In our case, the "known" language is a language for which tools and resources are available to automatically extract information from a text written in that language. Viceversa, the "unknown" language is the one that poses analytical problems, as it typically happens in projects involving ancient texts and languages. In particular, as detailed in the following section, we wanted to experiment a way of supporting the construction of a multilingual terminology by exploiting an existing translation.

The use of parallel texts in support to lexicon construction is a field known as bilingual lexicon extraction, and it has a wide scientific literature (see for example (Fung, 1998), (Tufiş et al., 2004), (Gutierrez-Vasques, 2015)). From a more applicative point of view, tools and software libraries have been implemented to assist developers in implementing the word-by-word text alignment necessary to process parallel texts. Giza++[1] and the Berkeley aligner[2], for example, have been largely adopted for these tasks. More in general, and in the context of Digital Humanities, the idea of exploiting parallel texts has been adopted in a number of initiatives, among which we point out the Perseus project, where the project team, together with the Von Humboldt professorship G. Crane within the Global Philology

[1] http://www.statmt.org/moses/giza/GIZA++.html
[2] https://code.google.com/archive/p/berkeleyaligner/

project, analyzed and implemented a collection of technologies and tools to envisage the "complexities of working with a historical record that contains far more languages than any individual could study, much less master" (Crane, Gregory et al., 2019).

2 Objectives and motivation

The experiment we here illustrate, still in progress, was conducted on the Babylonian Talmud Italian translation, in the context of the homonymous project[3]. The project, in addition to the development of the software Traduco used to support in the translation of the Talmud (Giovannetti et al., 2016), envisages the construction of a multilingual (Hebrew-Aramaic-Italian) terminological resource to support a number of activities, such as boosting the Translation Memory System with terminological information and creating an ontology of the talmudic domain. As described in Section 3, the Italian portion of the resource was built with the aid of a terminology extractor exploiting linguistic analysis tools for Italian. However, no tool or linguistic resource was available to automatically process the three main ancient languages appearing in the Talmud, namely, mishnaic Hebrew, biblical Hebrew and babylonian Aramaic. To obviate to this issue, and to the difficulty of automatically detecting the source terms through standard extraction processes, we chose to exploit the data produced in the last seven years of project activities, i.e. the available translated tractates of the Talmud. The results of the experiment suggested more ways of exploiting the obtained list of term-pairs in addition to the enrichment of the terminological resource, for example, as it will be discussed in the final version of the paper, to help in the lemmatization of semitic languages.

3 Methodology

Basically, the proposed approach makes use of a word-by-word alignment technique applied to a text in translation. The overall extraction process, leading to the enrichment of the terminological resource, followed a four step approach: i) encoding of the parallel text in TEI, ii) extraction of the Italian terms using a customized term extractor, iii) application of a word-by-word alignment technique to the parallel textual segments of the Talmud, iv) manual revision of the obtained alignment for the detection of the Hebrew/Aramaic terms corresponding to the Italian ones.

3.1 TEI encoding of the parallel text

We have first modeled and encoded the available parallel text (i.e. the Talmud and its Italian translation) by means of the best practices dictated by the Text Encoding Initiative (TEI), whose schema is currently the *de facto* standard to encode text-bearing objects (TBO) within the most authoritative scholarly projects involving literary inquiries. Actually, the choice to adhere to TEI environment provides benefits both to scholars, offering a standard model for the digital representation of critical texts, and to technicians, concerning modularity, data management, and, in particular, independence related to specific development choices. We have adopted the hierarchical text-group technique in order to encode the basic textual segments in three different modalities: 1) the original talmudic text; 2) the Italian translated text; 3) the literal Italian translated text. Moreover, the linkage among the different textual fragments has been conducted by means of the linkgroup technique [4]. Section 3.3 will illustrate the word-by-word alignment task that has been developed.

3.2 Extraction of the target terms

As mentioned before, given the lack of NLP tools and resources for Ancient Hebrew and Aramaic we could carry out the automatic extraction of the terminology only on the italian translation of the Talmud. For this purpose we used T2K^2 (Dell'Orletta et al., 2014), a platform for linguistic analysis available at the Institute of Computational Linguistics (ILC) of the Italian National Research Council (CNR). T2K^2 includes a stochastic module for terminology extraction which appeared adequate for our experimental purposes. We applied the extractor to four of the already translated and revised tractates of the Talmud,

[3] https://www.talmud.it

[4] Module number 16 of the TEI guidelines - Groups of Links. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SAPTLGen/html/SA.html#SAPTLG

namely: Berakhòt, Rosh haShanà, Ta'anìt and Qiddushìn. The corpus made of textual (plain-text UTF-8) documents was analyzed with T2K^2 and the obtained output was furtherly processed in order to remove erroneous terms deriving from Part-Of-Speech tagging errors, and to sort the extracted terms by means of the TF-IDF (Term Frequency-Inverse Document Frequency) statistical measure. An important outcome of the TF-IDF is to permit to measure the relevance of each term for each tractate in which it appears: a high value of TF-IDF represents a high degree of relevance in the context of a specific tractate. The Table 1 shows some examples of term relevance by tractate.

Berakhòt			Qiddushin		
Terms	tfidf	freq	Terms	tfidf	freq
Birkàt haMazòn	0.0239	120	documento	0.0159	123
emissione di seme	0.0209	105	perutà	0.0155	120
Shemà	0.0205	206	qiddushìn	0.0142	55
sogno	0.0166	167	schiava	0.0119	46
gabinetto	0.0076	38	terra di Israele	0.0107	83
frutto della terra	0.0072	36	divorzio	0.0104	40
benedizione sul vino	0.0062	31	rapporto sessuale	0.0101	78
tipi di cibi	0.0060	30	padrone	0.0100	186
pane dalla terra	0.0056	28	schiava ebrea	0.0088	34
bisogni	0.0047	47	trovatello	0.0080	31

Table 1: The first ten Italian terms extracted from two of the four analyzed tractates and ordered by tf-idf.

3.3 Extraction of the source terms via alignment

Word-by-word text alignment is a very useful technique to help understanding cross-lingual properties of parallel texts while processing only one half of the whole resource (Tiedemann, 2011). In order to add the Hebrew and Aramaic terms to the terminological resource we are building up from the Talmud, we set up the alignment process at token granularity. Specifically, we used an open source library realized by the Berkeley University (Liang et al., 2006) to develop a tool for the linking of Hebrew/Aramaic textual segments with the corresponding Italian translations.

Italian terms	most likely Hebrew term	other candidates Hebrew terms
benedizione (2.1)	בְּרָכָה (0.41)	מְבָרְכִין (0.29), מְבָרֵךְ (0.09)
Shemà (1.1)	קְרִיאַת (0.53)	שְׁמַע (0.44)
preghiera (2.2)	תְּפִלָּה (0.30)	תְּפִלַּת (0.15), תְּפִלָּה (0.13), תְּפִלָּתוֹ (0.16)
pane (1.9)	לֶחֶם (0.27)	פַּת (0.16), רִיפְתָּא (0.16), הַפַּת (0.22)
anno (2.14)	שָׁנָה (0.32)	שָׁנָה (0.15), הַשָּׁנָה (0.10), הַשָּׁנָה (0.19)
mese (1.93)	חֹדֶשׁ (0.25)	הַחֹדֶשׁ (0.21), לַחֹדֶשׁ (0.24)
giorno (1.90)	יוֹם (0.36)	הַיּוֹם (0.14), בַּיּוֹם (0.09), יוֹמָא (0.19)
shofàr (0.87)	שׁוֹפָר (0.77)	בְּשׁוֹפָרוֹת (0.08)
obbligo (2.1)	יָצָא (0.34)	חוֹבָה (0.09), חוֹבָתוֹ (0.22)
schiavo (0.82)	עֶבֶד (0.80)	וְעֶבֶד (0.09)

Table 2: Some examples of Italian-Hebrew/Aramaic aligned terms. Italian terms with high entropy (such as "preghiera") have been aligned with multiple Hebrew/Aramaic terms: the confidence that the term "תְּפִלָּה" (the one with the highest likelihood) is the actual translation of preghiera is low.

To carry out the word-by-word alignment, the tool implements generative models that have been studied during the last decades by the IBM researchers and by the Machine Translation community

(Brown et al., 1993). In particular, it adopts the IBM Model-1 with the extension of the Hidden Markov Model paradigm (Östling and Tiedemann, 2016). The alignment task employed non-supervised machine learning algorithms adopting probabilistic models to calculate the likelihood estimation of aligning a term in a known language to a term in a foreign language. The expressiveness of these kinds of alignment models is particularly suitable in the literary domain, where translations tend to be more interpretative and less literal. Eventually, for each Italian term the computed probabilistic alignment model provided a list of Hebrew/Aramaic candidate words. In Table 2, the numbers reported next to the Italian terms represents the entropy measure, which indicates the confidence of the translated word. The numbers next to the Hebrew/Aramaic words indicate the likelihood that word is the translation of the corresponding Italian word.

3.4 Manual revision

The aligner developed so far is based on statistical approaches which are, inherently, prone to errors. For this reason, the alignment environment requires a tool to validate and manually annotate the obtained outputs. We are thus developing a Web application able to manage and process the aligned text segments. As shown in 1, we have provided the proofreader with the possibility to annotate each word with a number of language and textual traits, namely lemma, Part-of-Speech, type of text, and language.

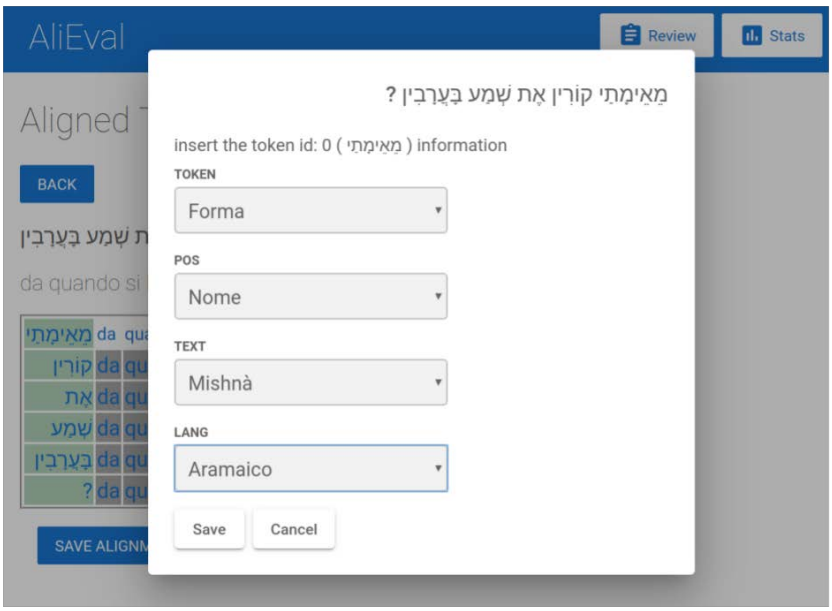

Figure 1: The annotation component of the proofreader.

The output of the aligner is formatted as a sequence of strings like `0-0 4-6 2-5 3-4 1-2 1-1` representing the word pairs that have been aligned. The order of the pairs is not significant, while the number within the pair represents the position of the word within the source-target strings; for example, in the two strings "עַד סוֹף הָאַשְׁמוּרָה הָרִאשׁוֹנָה" and "fino alla fine della prima veglia" the pair 0-0 would indicate the word pair "עַד-fino" (Hebrew is read from right to left). More details about the proofreader will be provided in the final version of the paper. Eventually, the revision process will allow to build a ground truth and/or a gold training set and consequently put in place a complete validation process of the alignment results.

4 Preliminary results, discussion and next steps

As it was shown, a parallel text can be exploited fruitfully via text alignment techniques to help in the construction of a multilingual terminology. Our reference scenario was the Italian translation of the

Babylonian Talmud, carried out in the context of the homonymous project. At the current stage of the work, 219.000 tokens have been analyzed, distributed on 42.000 textual segments extracted from the four aforementioned tractates which have been translated so far.

In addition to their use in populating the terminological resource, the obtained term-pairs may be also exploited in other ways. The first two applications we are going to investigate are: the boosting of text search, as recently experimented also in (Andonovski et al., 2019), and the support in the automatic processing of the source language.

Concerning the next steps of this research, once a significant number of segments (and, thus, of the terms appearing in the segments) will have been revised by the expert of the Talmud, a formal evaluation of the accuracy of the approach will be carried out. Fig. 2 shows an example of revision of the alignment.

Figure 2: An example of use of the proofreader: the output of the automatic alignment (at the top) and the relative revision (at the bottom).

Besides, we intend to improve the performance of the approach by taking into account the variety of texts and languages that coexist inside the Talmud before the application of the aligner. As a matter of fact, the Babylonian Talmud is constituted by two (macro) texts, i.e. the Mishna and the Gemara, which, in turn, incorporate portions of other texts, such as, for example, quotes from the Tanakh (the Hebrew Bible). In the particular case of the Talmud, each text is written in a specific language: the Mishna in Mishnaic Hebrew, the Gemara in Babylonian Aramaic and the Tanakh in Biblical Hebrew. The idea is to automatically classify each segment of the Talmud on the basis of the text it belongs to and, after that,

to apply the aligner on each textual class composed of linguistically homogeneous segments. By doing this, we expect a better accuracy from the aligner and, ideally, no need from the revisor to indicate the language of each segment.

Acknowledgement

This work was conducted in the context of the TALMUD project and the scientific cooperation between S.ca r.l. PTTB and ILC-CNR.

References

Jelena Andonovski, Branislava Šandrih, and Olivera Kitanović. 2019. Bilingual lexical extraction based on word alignment for improving corpus search. *The Electronic Library* .

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2):263–311. https://www.aclweb.org/anthology/J93-2003

Crane, Gregory, Jovanovic, Neven, Sklaviadis, Sophia, De Luca, Margherita, Šoštarić, Petra, Foradi, Maryam, Cottrell, Kate, Tauber, James, Shamsian, Farnoosh, and Palladino, Chiara. 2019. Confronting Complexity of Babel in a Global and Digital Age. In *Complexities*. ADO, Utrecht, Netherlands. https://dev.clariah.nl/files/dh2019/boa/0611.html

Felice Dell'Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. T2K^2: a system for automatically extracting and organizing knowledge from texts. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Machine Translation and the Information Soup*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 1–17.

Emiliano Giovannetti, Davide Albanesi, Andrea Bellandi, and Giulia Benotto. 2016. Traduco: A collaborative web-based cat environment for the interpretation and translation of texts. *Digital Scholarship in the Humanities* 32(suppl_1):i47–i62.

Ximena Gutierrez-Vasques. 2015. Bilingual lexicon extraction for a distant language pair using a small parallel corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. pages 154–160.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT-NAACL '06, pages 104–111. https://doi.org/10.3115/1220835.1220849

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics* 106:125–146. https://doi.org/10.1515/pralin-2016-0013

Jöho Tiedemann. 2011. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Dan Tufiş, Ana Maria Barbu, and Radu Ion. 2004. Extracting Multilingual Lexicons from Parallel Corpora. Computers and the Humanities 38(2):163–189. https://doi.org/10.1023/B:CHUM.0000031172.03949.48