10.12871/CLICIT2014111

The Talmud System: a Collaborative web Application for the Translation of the Babylonian Talmud Into Italian

Andrea Bellandi, Davide Albanesi,
Alessia Bellusci, Andrea Bozzi, Emiliano Giovannetti
Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, 56124, Pisa - Italy
{name.surname}@ilc.cnr.it

Abstract

English. In this paper we introduce the Talmud System, a collaborative web application for the translation of the Babylonian Talmud into Italian. The system we are developing in the context of the "Progetto Traduzione del Talmud Babilonese" has been designed to improve the experience of collaborative translation using Computer-Assisted Translation technologies and providing a rich environment for the creation of comments and the annotation of text on a linguistic and semantic basis.

Italiano. *In questo articolo presentiamo il Sistema Talmud, un'applicazione web collaborativa per la traduzione del Talmud babilonese in italiano. Il sistema, che stiamo sviluppando nel contesto del "Progetto Traduzione del Talmud Babilonese", stato progettato per migliorare l'esperienza di traduzione collaborativa utilizzando tecnologie di Computer-Assisted Translation e fornendo un ambiente ricco per la creazione di commenti e l'annotazione del testo su base linguistica e semantica.*

1 Introduction

Alongside the Bible, the Babylonian Talmud (BT) is the Jewish text that has mostly influenced Jewish life and thought over the last two millennia. The BT corresponds to the effort of late antique scholars (*Amoraim*) to provide an exegesis of the *Mishnah*, an earlier rabbinic legal compilation, divided in six "orders" (*sedarim*) corresponding to different categories of Jewish law, with a total of 63 tractates (*massekhtaot*). Although following the inner structure of the *Mishnah*, the BT discusses only 37 tractates, with a total of 2711 double sided folia in the printed edition (Vilna, XIX century). The BT is a comprehensive literary creation, which went through an intricate process of oral and written transmission, was expanded in every generations before its final redaction, and has been the object of explanatory commentaries and reflexions from the Medieval Era onwards. In its long history of formulation, interpretation, transmission and study, the BT reflects inner developments within the Jewish tradition as well as the interactions between Judaism and the cultures with which the Jews came into contact (Strack and Stemberger, 1996). In the past decades, online resources for studying Rabbinic literature have considerably increased and several digital collections of Talmudic texts and manuscripts are nowadays available (Lerner, 2010). Particularly, scholars as well as a larger public of users can benefit from several new computing technologies applied to the research and the study of the BT, such as (i.) HTML (Segal, 2006), (ii.) optical character recognition, (iii.) three-dimensional computer graphics (Small, 1999), (iv.) text encoding, text and data mining (v.) image recognition (Wolf et al., 2011(a); Wolf et al., 2011(b); Shweka et al., 2013), and (vi.) computer-supported learning environments (Klamma et al., 2005; Klamma et al., 2002). In the context of the "Progetto Traduzione del Talmud Babilonese", the Institute for Computational Linguistics of the Italian National Research Council (ILC-CNR) is in charge of developing a collaborative Java-EE web application for the translation of the BT into Italian by a team of translators. The Talmud System (TS) already includes Computer-Assisted Translation (CAT), Knowledge Engineering and Digital Philology tools, and, in future versions, will include Natural Language Processing tools for Hebrew/Aramaic, each of which will be outlined in

detail in the next Sections.

2 Description of the System

The general architecture of the TS is represented in Figure 1. Each system component implements specific functionalities targeted at different types of users. Translators and revisors are assisted in the translation process by CAT technologies, including indexers and a Translation Memory (TM); philologists and linguists are enabled to insert notes, comments, semantic annotations and bibliographical references; domain experts are allowed to structure relevant terms into glossaries, and, possibly, into domain ontologies; researchers and scholars can carry out complex searches both on a linguistic and semantic basis; editors are enabled to produce the printed edition of the translation of the BT in an easier manner, by arranging translations and notes in standard formats for desktop publishing software. In what follows, we briefly outline the TS main components and the progress state of their development.

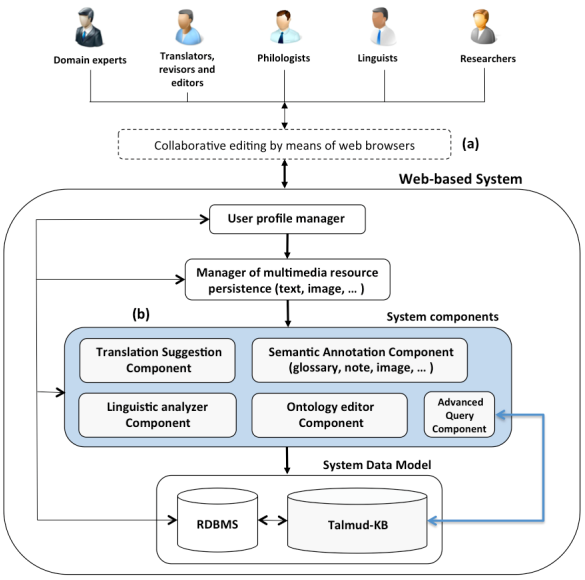

Figure 1: The Talmud System's architecture. (a) Collaborative editing - (b) Component based structure.

2.1 Translation Suggestion Component

We chose to adopt a Translation Memory (TM) based approach due to the literary style of the BT. Composed in a dialogical form and characterized by formulaic language, the BT presents several standard expressions. Furthermore, as an exegetical text, the BT contains innumerable quotation from the Bible, the *Mishnah*, other tannaitic sources and even from amoraitic statements discussed in other passages of the BT itself.

To the best of our knowledge, our implementation mainly contemplates aspects related to the specific needs of the translators community working on the BT in a collaborative environment, that the main non commercial CAT tools (OpenTM, OmegaT, Olanto, Transolution) and commercial ones (Dèjà Vu, Trados, Wordfast, Multitrans, Star Transit) do not take suitably into account (see (Bellandi et al., 2014(b)) for details). These specific requisites can be generalized to other complex ancient texts, where the emphasis of the translation work shall concern the quality instead of the translation pace. Exhibiting exceptionally concise sentences, which remain often unclear even to expert Talmudists, the BT cannot be treated and translated as a modern text. It is worth considering the Matecat Project[1], where the authors combine CAT and machine translation (MT) technologies, providing both suggestions by MT which are consistent with respect to the whole text, and methods for the automatic self-correction of MT making use of the implicit feedback of the user. The lack of linguistically annotated resources, and large collections of parallel texts regarding the languages present in the BT, prevented us to consider any statistical MT toolkit. We implemented a TM enabling translators to re-elaborate the plain and literal translation of the text and integrate it with explicative additions. The TM is organized at the segment level. A segment is a portion of original text having an arbitrary length. We formally defined the translation memory $M_{BT} = \{(s_i, T_i, A_i, c_i)\}$ with i ranging 1 to n, as a set of n tuples, where each tuple is defined by:

- s_i, the source segment;
- $T_i = \{t_i^1, \ldots, t_i^k\}$, the set of translations of s_i with $k \geq 1$, where each t_i^j has its literal part $\overline{t}_i^j$, and its contextual information $\widetilde{t}_i^j$, with $1 \leq j \leq k$;
- $A_i = \{a_i^1, \ldots, a_i^k\}$, the set of translators id of each translation of s_i in T_i with $k \geq 1$;
- c_i, the context of s_i referring to the tractate which belongs to;

Each segment's translation is obtained by differentiating the "literal" translation (using the bold

[1] http://www.matecat.com/matecat/the-project/

style) from explicative additions, i.e. "contextual information". Segments exhibiting the same literal part may convey different contextual information. By the term "context", we refer to the tractate to which the source segment belongs. The translation environment we created allows to acquire the segment to be translated, to query the TM, and to suggest the Italian translations related to the most similar strings. Since the BT does not exhibit a linguistic continuity, thus preventing an automatic splitting into sentences, we opted for a manual segmentation. Each translator selects a certain source segment to translate from a specific window of the system's GUI, which contains the specific tractate of the BT. This process may have a positive outcome: translators, being forced to manually detect the segments, could acquire a deeper awareness of the text they are about to translate. Clearly, the manual segmentation implies the engagement of the translators in a deep cognitive process aimed to establish the exact borders of a segment. The thorough reflection of the segmentation affects deeply also the final translation, by orienting the content and nature of the TM. So far, we could not include neither grammatical nor syntactic information in the similarity search algorithm (see, Section 2.4). Thus, we adopted similarity measures based on edit distance, by considering that two source segments are more similar when exhibiting the same terms in the same order. The novelty of this approach consists in the way we rank suggestions with the same value, based on external information, stored as metadata inside the TM, i.e. (i.) authors of translations and (ii.) the context (the tractate of reference). These informations are highly valuable, enabling (i.) translators to evaluate the reliability of the suggested translations according to the scientific authority of their authors, and (ii.) revisors to pervain to a more coherent, homogeneous and fluent translation. Since each suggested translation can be shown with or without its contextual information, each translator is enabled to approve and choose the literal translation, editing only the contextual information. Thus, our system relieves human translators from routine work, but always enabling them to control and orientate the translation process. Such a system is particularly useful for a complex ancient text such as the BT, which demands the linguistic and scholarly input of human users. Finally, Figure 2 shows the TM performance in terms of redundancy rate, roughly estimated by conducting a jackknife experiment (Wu, 1986). Redundancy curves are drawn by considering the ranking of the similarity function. The percentage of source segments found both verbatim and fuzzy in the memory appears to grow logarithmically with time (and consequently with the size of the memory).

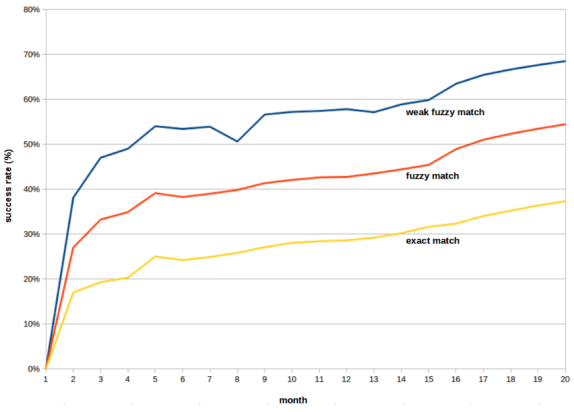

Figure 2: Redundancy of the translation memory in function of time.

2.2 Knowledge Engineering

Dealing with ethics, jurisprudence, liturgy, ritual, philosophy, trade, medicine, astronomy, magic and so much more, the BT represents the most important legal source for Orthodox Judaism and the foundation for all the successive developments of *Halakhah* (legal knowledge) and *Aggadah* (narrative knowledge). By means of an annotation module, translators can then semantically annotate arbitrary portions of text on the basis of the above fields. To date, the annotation process exploits an initial set of five predefined semantic classes: people's names, animals, plants, idiomatic expressions (e.g., *the Master said*), concepts (e.g., *Terumah*). This functionality allows the creation of specialized glossaries and, when fully implemented, the automatization of the annotation process. Furthermore, it enables experts specialized in the various Talmudic subjects to annotate, in a collaborative environment, relevant and technical terms and, eventually, structure them in a Talmudic Knowledge base (Talmud-KB, in Figure 1), using a formal knowledge representation language. To face the plurality of opinions, which generally originates in a collegial environment when assigning semantic labels, especially in the context of translation, the TS is fitted to enable domain

experts to represent uncertain knowledge through "weighted" relations, according to their scientific confidence (Bellandi and Turini, 2012; Danev et al., 2006; Ding et al., 2005).

2.3 Digital Philology

The system also responds to the specific needs of philological work and specialized analyses of the text, allowing to insert annotations at various levels of granularity. The parts of the Italian translation that appear in bold, for example, correspond to literal translations, while those in plain are explicative additions, i.e phrases added to make concepts expressed in Hebrew/Aramaic understandable to an Italian reader. Other annotations of greater granularity include: i) the addition of (explanatory) notes by translators and revision notes by revisors, ii) semantic annotations based on predefined types (see 2.2) designed to offer greater philological precision to the analysis and indexing of the text and for the construction of glossaries. A further element designed to perform more in-depth analysis of the translated text is provided by a dedicated component to introduce, in a standardized way, partially precompiled bibliograpghic references (e.g. for biblical citations to be completed with chapter and verse numbers) and names of Rabbis.

2.4 Language Analysis

Within the BT, we distinguish: (i.) quotations of portions from the *Mishnah*, (ii.) long amoraic discussions of mishnaic passages aimed at clarifying the positions and lexicon adopted by the *Tannaim*, and (iii.) external tannaitic material not incorporated in the canonical *Mishnah*. The content and philological depth of the BT implies an elevated degree of linguistic richness. In its extant form, the BT attests to (i.) different linguistic stages of Hebrew (Biblical Hebrew, Mishnaic Hebrew, Amoraic Hebrew), (ii.) different variants of Jewish Aramaic (Babylonian Aramaic and Palestinian Aramaic), and (iii.) several loanwords from Akkadian, ancient Greek, Latin, Pahlavi, Syriac and Arabic. To date, there are no available Natural Language Processing (NLP) tools suitable for processing ancient North-western Semitic languages, such as the different Aramaic idioms attested to in the BT, and for detecting the historical variants of Hebrew language as used in the Talmudic text. Several computational studies have been recently carried out on Modern Semitic Languages, including Modern Hebrew, and two high quality NLP tools are implemented for this language (Itai, 2006; HebMorph, 2010). Nevertheless, Modern Hebrew has been through a process of artificial revitalization from the end of the XIX century and does not correspond to the idioms recurring in the BT, even not to Biblical Hebrew or Mishnaic Hebrew. For this dissimilarity between the new and the ancient Hebrew languages, the existing NLP tools for Hebrew are highly unfit for processing the BT. In its multifaceted form, the "language" of the BT is unique and attested to only in few other writings. In addition, only few scholars have a full knowledge of the linguistic peculiarities of the BT and even fewer experts in Talmudic Studies are interested in collaborating to the creation of computational technologies for this textual corpus. These two main reasons have prevented, so far, the development of NLP tools for the BT, which would require a huge and very difficult effort probably not entirely justified by the subsequent use of the new technologies developed. The only attempts in these direction have been conducted within the Responsa Project on rabbinic texts, including the BT, and the Search And Mining Tools with Linguistic Analysis (SAMTLA[2]) on the corpus of Aramaic Magic Texts from Late Antiquity (AMTLA), some of which are written in Jewish Babylonian Aramaic, the dialect characterizing the BT. In the future phases of our project, we aim to develop some language resources for processing the linguistic and dialectic variants attested to in the BT.

3 Conclusion

We here introduced the Talmud System, a collaborative web application for the translation of the Babylonian Talmud into Italian integrating technologies belonging to the areas of (i.) Computer-Assisted Translation, (ii.) Digital Philology, (iii.) Knowledge Engineering and (iv.) Natural Language Processing. Through the enhancement of the already integrated components (i., ii., iii.) and the inclusion of new ones (iv.) the TS will allow, in addition to the improvement of the quality and pace of the translation, to provide a multi-layered navigation (linguistic, philological and semantic) of the translated text (Bellandi et al., 2014(c)).

[2] http://samtla.dcs.bbk.ac.uk/

4 Acknowledgements

This work has been conducted in the context of the research project TALMUD and 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 (21/01/2011).

References

Andrea Bellandi, Alessia Bellusci, Emiliano Giovannetti, Enrico Carniani. 2014(a). Content Elicitation: Towards a New Paradigm for the Analysis and Interpretation of Text. Mohamed H. Hamza, ed., In *Proceedings of the IASTED International Conference on Informatics*, pp. 507-532.

Andrea Bellandi, Franco Turini. 2012. Mining Bayesian Networks Out of Ontologies. *Journal of Intelligent Information Systems*, 38(2):507-532.

Andrea Bellandi, Alessia Bellusci, Emiliano Giovannetti. 2014(b). Computer Assisted Translation of Ancient Texts: the Babylonian Talmud Case Study. In *Proceedings of the 11th International Natural Language Processing and Cognitive Systems*.

Andrea Bellandi, Alessia Bellusci, Amedeo Cappelli, and Emiliano Giovannetti. 2014(c). Graphic Visualization in Literary Text Interpretation. In *Proceedings of the 18th International Conference on Information Visualisation*. Paris, France. July 15-18.

Boris Danev, Ann Devitt, Katarina Matusikovai. 2006. *Constructing Bayesian Networks Automatically using Ontologies*. Second Workshop on Formal Ontologies Meets Industry.

Zhongli Ding, Yun Peng, Rong Pan. 2005. *BayesOWL: Uncertainty Modeling in Semantic Web Ontologies*. Soft Computing in Ontologies and Semantic Web Springer-Verlag.

HebMorph - Morphological Analyser and Disambiguator for Hebrew Language. 2010. http://code972.com/hebmorph.

Alon Itai. 2006. Towards a Research Infrastructure for Language Resources. In *Proceedings of the Language Resources and Evaluation Conference*.

Ralf Klamma, Marc Spaniol, Matthias Jarke. 2005. MECCA: Hypermedia Capturing of Collaborative Scientific Discourses about Movies. *Informing Science Journal*, 8:3-38.

Ralf Klamma, Elisabeth Hollender, Matthias Jarke, Petra Moog, Volker Wulf. 2002. Vigils in a Wilderness of Knowledge: Metadata in Learning Environments. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, pp. 519-524.

Heidi Lerner. 2010. Online Resources for Talmud Research, Study and Teaching. *Association for Jewish Studies*, pp. 46-47.

MateCat Project. A CAT Tool for Your Business. Simple. Web-Based, 2012. http://www.matecat.com/matecat/the-project/.

Eliezer Segal. A Page from the Babylonian Talmud. http://www.ucalgary.ca/ elsegal/TalmudPage.html.

Eliezer Segal. 2006. Digital Discipleship: Using the Internet for the Teaching of Jewish Thought. H. Kreisel, ed., *Study and Knowledge in Jewish Thought*, pp. 359-373.

Roni Shweka, Yaacov Choueka, Lior Wolf, Nachum Dershowitz. 2013. Automatic Extraction of Catalog Data from Digital Images of Historical Manuscripts. *Literary and Linguistic Computing*, pp. 315-330.

David L. Small. 1999. *Rethinking the Book, unpubl. PhD Dissertation*. Massachusetts Institute of Technology, http://www.davidsmall.com/portfolio/talmud-project/.

H. L. Strack, G. Stemberger. 1996. *Introduction to Talmud and Midrash* . tr. and ed. by M. Bockmuehl, pp. 190-225.

Lior Wolf, Liza Potikha, Nachum Dershowitz, Roni Shweka, Yaacov Choueka. 2011(a). Computerized Palaeography: Tools for Historical Manuscripts. In *Proceedings 18th IEEE International Conference on Image Processing*, pp. 3545-3548.

Lior Wolf, Lior Litwak, Nachum Dershowitz, Roni Shweka, Yaacov Choueka. 2011(b). Active Clustering of Document Fragments using Information Derived from Both Images and Catalogs. In *Proceedings IEEE International Conference on Computer Vision*, pp. 1661-1667.

Chien-Fu Jeff Wu. 1986. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14(4):1261-1295.