Constructing an Annotated Resource for Part-Of-Speech Tagging of Mishnaic Hebrew

Emiliano Giovannetti[1], **Davide Albanesi**[1], **Andrea Bellandi**[1], **Simone Marchi**[1], **Alessandra Pecchioli**[2]

[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
alepec3@gmail.com

Abstract

English. This paper introduces the research in Part-Of-Speech tagging of mishnaic Hebrew carried out within the Babylonian Talmud Translation Project. Since no tagged resource was available to train a stochastic POS tagger, a portion of the Mishna of the Babylonian Talmud has been morphologically annotated using an ad hoc developed tool connected with the DB containing the talmudic text being translated. The final aim of this research is to add a linguistic support to the Translation Memory System of Traduco, the Computer-Assisted Translation tool developed and used within the Project.

Italiano. *In questo articolo è introdotta la ricerca nel Part-Of-Speech tagging dell'Ebraico mishnaico condotta nell'ambito del Progetto Traduzione Talmud Babilonese. Data l'indisponibilità di risorse annotate necessarie per l'addestramento di un POS tagger stocastico, una porzione di Mishnà del Talmud Babilonese è stata annotata morfologicamente utilizzando uno strumento sviluppato ad hoc collegato al DB dove risiede il testo talmudico in traduzione. L'obiettivo finale di questa ricerca è lo sviluppo di un supporto linguistico al sistema di Memoria di Traduzione di Traduco, lo strumento di traduzione assistita utilizzato nell'ambito del Progetto.*

1 Introduction

The present work has been conducted within the Babylonian Talmud Translation Project (in Italian, Progetto Traduzione Talmud Babilonese - PTTB) which aims at the translation of the Babylonian Talmud (BT) into Italian.

The translation is being carried out with the aid of tools for text and language processing integrated into an application, called *Traduco* (Bellandi et al., 2016), developed by the Institute of Computational Linguistics "Antonio Zampolli" of the CNR in collaboration with the PTTB team. Traduco is a collaborative computer-assisted translation (CAT) tool conceived to ease the translation, revision and editing of the BT.

The research described here fits exactly in this context: we want to provide the system with additional informative elements as a further aid in the translation of the Talmud. In particular, we intend to linguistically analyze the Talmudic text starting from the automatic attribution of the Part-Of-Speech to words by adopting a stochastic POS tagging approach.

The first difficulty that has emerged regards the text and the languages it contains. In this regard we can say, simplifying, that the Babylonian Talmud is essentially composed of two languages which, in turn, correspond to two distinct texts: the Mishna and the Gemara. The first is the oldest one written in mishnaic Hebrew, one of the most homogeneous and coherent languages appearing in the Talmud that, for this reason, has been chosen to start from in the POS tagging experiment.

The main purpose of linguistic analysis in the context of our translation project is to improve the suggestions provided by the system through the so-called Translation Memory (TM).

Moreover, on a linguistically annotated text it is possible to carry out linguistic-based searches, useful both for the scholar (in this

case a talmudist), and, during the translation work, for the revisor and the curator, who have the possibility, for example, to make bulk editing of polysemous words by discarding out words with undesired POS.

The rest of the paper is organized as follows: Section 2 summarizes the state of the art in NLP of Hebrew. The construction of the linguistically annotated corpus is described in Section 3. The training process and evaluation of the POS taggers used in the experiments is detailed in Section 4. Lastly, Section 5 outlines the next steps of the research.

2 State of the art

The aforementioned linguistic richness and the intrinsic complexity of the Babylonian Talmud make automatic linguistic analysis of the BT particularly hard (Bellandi et al., 2015).

However, some linguistic resources of ancient Hebrew and Aramaic have been (and are being) developed, among which we cite: i) the Hebrew Text Database (Van Peursen and Sikkel, 2014) (ETCBC) accessible by SHEBANQ[1] an online environment for the study of Biblical Hebrew (with emphasis on syntax), developed by the Eep Talstra Centre for Bible and Computer of the Vrije Universiteit in Amsterdam; ii) the Historical Dictionary[2] project of the Academy of the Hebrew Language of Israel; iii) the Comprehensive Aramaic Lexicon (CAL)[3] developed by the Hebrew Union College of Cincinnati; iv) the Digital Mishna[4] project, concerning the creation of a digital scholarly edition of the Mishna conducted by the Maryland Institute of Technology in the Humanities.

Apart from the aforementioned resources, to date there are no available NLP tools suitable for the processing of ancient north-western Semitic languages, such as the different Aramaic idioms and the historical variants of Hebrew attested in the BT. The only existing projects and tools for the processing of Jewish languages (Kamir et al., 2002) (Cohen and Smith, 2007) have been developed for modern Hebrew, a language that has been artificially revitalized from the end of the XIX century and that does not correspond to the idioms recurring in the BT. Among them we cite HebTokenizer[5] for tokenization, MILA (Barhaim et al., 2008), HebMorph[6], MorphTagger [7] and NLPH[8] for morphological analysis and lemmatization, yap[9], hebdepparser[10], UD_Hebrew[11] for syntactic analysis. We conducted some preliminary tests by starting with MILA's (ambiguous) morphological analyzer applied to the three main languages of the Talmud:

1. *Aramaic*: Hebrew and Aramaic are different languages. There are even some cases in which the very same root has different semantics in the two languages. In addition, MILA did not recognize many aramaic roots, tagging the relative words, derived from them, as proper nouns.

2. *Biblical Hebrew*: MILA recognized most of the words, since Modern Hebrew preserved almost the entire biblical lexicon. However, syntax of Modern Hebrew is quite different from that of Biblical Hebrew, leading MILA to output wrong analyses.

3. *Mishnaic Hebrew*: this is the language where MILA performed better. Modern Hebrew inherits some of the morpho-syntactic features of mishnaic Hebrew, however, the two idioms differ substantially on the lexicon, since in modern Hebrew many archaic words have been lost (Skolnik and Berenbaum, 2007).

In the light of the above, we decided to create a novel linguistically annotated resource to start developing our own tools for the processing of ancient Jewish languages. In the next section, we will describe how the resource was built.

3 Building the resource

The linguistic annotation of Semitic languages poses several problems. Although we here discuss the analysis of Hebrew, many of the critical points that must be taken into account are

[1] shebanq.ancient-data.org
[2] maagarim.hebrew-academy.org.il
[3] cal.huc.edu
[4] www.digitalmishnah.org
[5] www.cs.bgu.ac.il/~yoavg/software/hebtokenizer
[6] code972.com/hebmorph
[7] www.cs.technion.ac.il/~barhaim/MorphTagger
[8] github.com/NLPH/NLPH
[9] github.com/habeanf/yap
[10] tinyurl.com/hebdepparser
[11] github.com/UniversalDependencies/UD_Hebrew

common to other languages belonging to the same family. As already mentioned in the previous section, the first problem concerns the access to existing linguistic resources and analytical tools which, in the case of Hebrew, are available exclusively for the modern language.

One of the major challenges posed by the morphological analysis of Semitic languages is the orthographic disambiguation of words. Since writing is almost exclusively consonantal, every word can have multiple readings. The problem of orthographic ambiguity, crucial in all studies on large corpora (typically in Hebrew and modern Arabic), does not prove to be so difficult when the text under examination is vocalized.

The edition of the Talmud used in the project is actually vocalized and the text, consequently, is orthographically unambiguous. An additional critical aspect is represented by the definition of the tagset. Most of the computational studies on language analysis have been conducted on Indo-european languages (especially on English).

As a result, it may be difficult to reuse tagsets created for these languages. Not surprisingly, there are still many discussions about how it is better to catalog some POS and each language has its own part under discussion. Each tagset must ultimately be created in the light of a specific purpose. For example, the tagging of the (Modern) Hebrew Treebank developed at the Technion (Sima'an et al., 2001) was syntax-oriented, while the work on participles of Hebrew described in (Adler et al., 2008) was more lexicon-oriented. We considered the idea of adopting the tagset used in the already cited Universal Dependency Corpus for Hebrew. However, its 16 tags appeared to be too "coarse grained" for our purposes.[12] In particular, the UD tagset lacks of all the prefix tags that we needed. For this reason we decided to define our own tagset.

Once the tagset has been defined, it remains to decide which is the most suitable grammatical category to associate with each token. You can collect essentially two types of information, the problem is how and if you can keep both, in particular: i) the definition of the token from a syntagmatic perspective (i.e. what the token represents in context) and ii) the lexical information that the token gives by itself (without context). To give a couple of examples:

- Verb/noun: אִשְׁתּוֹ אֶת הַמַּדִּיר → is הַמַּדִּיר "the one who makes a vow" or "the vowing"? (the one who consecrates his wife): should it be assigned to verb or noun category?

- Adjective/verb: עַד וְלִגְמוֹר לְהַתְחִיל יְכוֹלִין אִם יַתְחִילוּ - לַשּׁוּרָה נִגְּעוּ שֶׁלֹּא → is יְכוֹלִין adjective or verb (given that most of the mishnaic language dictionaries provide both options)?

We could discuss about which category would be the best for each and why, but, for now, we decided to keep both by introducing two parallel annotations, by "category" (without context) and by "function" (in context). The tagset we used for this work are the following: *agg.*, *avv.*, *cong.*, *interiez.*, *nome pr.*, *num. card.*, *num. ord.*, *pref. art.*, *pref. cong.*, *pref. prep.*, *pref. pron. rel.*, *prep.*, *pron. dim.*, *pron. indef.*, *pron. interr.*, *pron. pers.*, *pron. suff.*, *punt.*, *sost.*, *vb.*

One could also envisage the refining of the tagset by adding: interrogative, modal, negation, and quantifier (Adler, 2007) (Netzer and Elhadad, 1998) (Netzer et al., 2007).

As anticipated, in order to build the morphologically annotated resource, all of the Mishna sentences were extracted from the Talmud and annotated using an ad hoc developed Web application (Fig. 1).

All the annotations have been made with the aim of training a stochastic POS tagger in charge of the automatic analysis of the entire Mishna: to obtain a good accuracy it was thus necessary to manually annotate as many sentences as possible. To date, 10442 tokens have been annotated.

The software created for the annotation shows, in a tabular form, the information of the analysis carried out on a sentence by sentence basis.

The system, once a sentence is selected for annotation, checks whether the tokens composing it have already been analyzed and, in

[12] github.com/UniversalDependencies/UD_Hebrew-HTB/blob/master/stats.xml

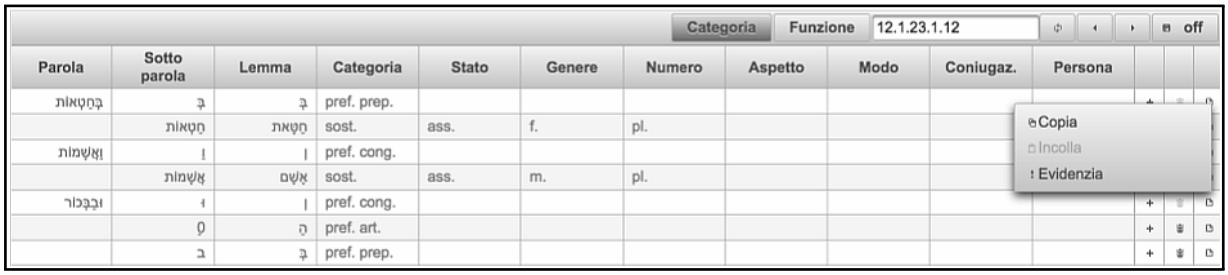

Figure 1: The interface for the linguistic annotation of the corpus to be used to train the POS tagger

case, calculates a possible subdivision into subtokens (i.e. the stems, prefixes and suffixes constituting each word) by exploiting previous annotations. If the system finds that a word is associated with multiple different annotations, it proposes the most frequent one.

Regarding the linguistic annotation, the grammar of Pérez Fernández (Fernández and Elwolde, 1999) was adopted and, for lemmatization, the dictionary of M. Jastrow (Jastrow, 1971).

The software allows to gather as much information as possible for each word by providing a double annotation: by "category" to represent the POS from a grammatical point of view, and by "function" to describe the function the word assumes in its context. For the POS tagging experiments, described below, we used the annotation made by "function".

4 Training and testing of POS taggers

Once the mishnaic corpus has been linguistically annotated three of the most used algorithms for POS tagging have been used and evaluated: HunPos (Halácsy et al., 2007), the Stanford Log-linear Part-Of-Speech Tagger (Toutanova et al., 2003), and TreeTagger (Schmid, 1994). The three algorithms implement supervised stochastic models and, consequently, they need to be trained with a manually annotated corpus.

To evaluate the accuracy of the algorithms we adopted the strategy of *k-fold cross validation* (Brink et al., 2016), with k set to 10, and thus dividing the corpus in 10 partitions.

Table 1 summarizes the results of the experiment by showing the tagging accuracy of the three tested algorithms. With a number of tokens slightly higher than ten thousands the

Tagging Accuracy		
Stanford	*Hunpos*	*Treetagger*
87,90%	86,34%	86,74%

Table 1: Accuracy of the three POS taggers.

Stanford POS tagger provided the best results over HunPos and Treetagger, with an accuracy of 87,9%.

5 Next steps

In this work, the tagging experiments have been limited to the attribution of the Part-Of-Speech: the next, natural step, will be the addition of the lemma. Furthermore, we will try to modify the parameters affecting the behaviour of the three adopted POS taggers (left at their default values for the experiments) and see how they influence the results.

Once the Mishna will be lemmatized, Traduco, the software used to translate the Talmud in Italian, will be able to exploit this additional information mainly to provide translators with translation suggestions based on lemmas, but also to allow users to query the mishnaic text by POS and lemma.

As a further step we will also take into account the linguistic annotation of portions of the Babylonian Talmud written in other languages, starting from the Babylonian Aramaic, the language of the Gemara, which constitutes the earlier portion of the Talmud.

Acknowledgments

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

References

Meni Adler, Yael Netzer, Yoav Goldberg, David Gabay, and Michael Elhadad. 2008. Tagging a hebrew corpus: the case of participles. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Menahem Meni Adler. 2007. *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*. PhD Thesis, Ben-Gurion University of the Negev.

Roy Bar-haim, Khalil Sima'an, and Yoad Winter. 2008. Part-of-speech Tagging of Modern Hebrew Text. *Nat. Lang. Eng.*, 14(2):223–251, April.

Andrea Bellandi, Alessia Bellusci, and Emiliano Giovannetti. 2015. Computer Assisted Translation of Ancient Texts: the Babylonian Talmud Case Study. In *Natural Language Processing and Cognitive Science, Proceedings 2014*, Berlin/Munich. De Gruyter Saur.

Andrea Bellandi, Davide Albanesi, Giulia Benotto, and Emiliano Giovannetti. 2016. *Il Sistema Traduco nel Progetto Traduzione del Talmud Babilonese*. IJCoL Vol. 2, n. 2, December 2016. Special Issue on "NLP and Digital Humanities". Accademia University Press.

Henrik Brink, Joseph Richards, and Mark Fetherolf. 2016. *Real-World Machine Learning*. Manning Publications Co., Greenwich, CT, USA, 1st edition.

Shay B. Cohen and Noah A. Smith. 2007. Joint Morphological and Syntactic Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Miguel Pérez Fernández and John F. Elwolde. 1999. *An Introductory Grammar of Rabbinic Hebrew*. Interactive Factory, Leiden, The Netherlands.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos: An Open Source Trigram Tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 209–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marcus Jastrow. 1971. *A dictionary of the Targumim, the Talmud Babli and Yerushalmi, and the Midrashic literature*. Judaica Press.

Dror Kamir, Naama Soreq, and Yoni Neeman. 2002. A Comprehensive NLP System for Modern Standard Arabic and Modern Hebrew. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, SEMITIC '02, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yael Dahan Netzer and Michael Elhadad. 1998. Generating Determiners and Quantifiers in Hebrew. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages*, Semitic '98, pages 89–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yael Netzer, Meni Adler, David Gabay, and Michael Elhadad. 2007. Can You Tag the Modal? You Should. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 57–64, Prague, Czech Republic. Association for Computational Linguistics.

Helmut Schmid. 1994. Part-of-speech tagging with neural networks. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING '94, pages 172–176, Stroudsburg, PA, USA. Association for Computational Linguistics.

Khalil Sima'an, Alon Itai, Yoad Winter, Alon Altman, and Noa Nativ. 2001. Building a treebank of modern hebrew text. *TAL. Traitement automatique des langues*, 42(2):347–380.

Fred Skolnik and Michael Berenbaum, editors. 2007. *Encyclopaedia Judaica vol. 8*. Encyclopaedia Judaica. Macmillan Reference USA, 2 edition. Brovender Chaim and Blau Joshua and Kutscher Eduard Y. and Breuer Yochanan and Eytan Eli sub v. "Hebrew Language".

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wido Van Peursen and Constantijn Sikkel. 2014. Hebrew Text Database ETCBC4. type: dataset.