E. Giovannetti, A. Bellandi, D. Dattilo, A.M. Del Grosso, S. Marchi,
A. Pecchioli, S. Piccini

THE TERMINOLOGY OF THE BABYLONIAN TALMUD: EXTRACTION, REPRESENTATION AND USE IN THE CONTEXT OF COMPUTATIONAL LINGUISTICS

1. Introduction

Emiliano Giovannetti

Few sacred texts can boast the vastness, complexity, and richness of the content of the Talmud. The Talmud represents the fundamental text for Judaism, after the Tanak, and it constitutes a real mine of historical, cultural, social, legal, and scientific information. The translation of the Talmud into Italian is being carried out within the homonymous project.[1]

The text taken as reference for the translation is the Vilna edition (1835) of the Babylonian Talmud, accompanied by the traditional commentaries by Raši (Rabbi Šelomo Iṣḥaqi, France-Rhineland XI cent.) and *Tosafot*. In the translation process other commentaries and references have been taken into account, such as *'En Mišpaṭ - Ner Miṣwa*, the *Masoret ha-Šas* and, *Tora Or* by rav Yehošua' Bo'az leVet Baruk (Salvatore Boniforte de Benedetti, XVI cent.), the textual corrections of Rabbi Yoel Sirqis, the author of *Bait Ḥadaš*, the notes of *Ghilayon ha-Šas* by rabbi Aqiva Eiger (1761-1837) and the medieval comments by rav Nissim Gaon. The text adopted for the *Mišna* and *Gemara* is the fully diacritized version by rav Adin Steinsaltz, developed in the context of a private agreement with PTTB and Milta Management Ltd.

A variety of research and technological activities have been already carried out within the project, focussing on the Talmud languages.[2] In this paper, a computational linguistics approach to managing the talmudic terminology will be illustrated. One of the main objectives of the computational linguistics team involved in the Project is, indeed, to provide each word of the Talmud and its translation with linguistic data. Such data may comprise morphosyntactic information and relationships holding between the word itself and other words (such as synonymy, hyperonymy, etc.). The starting point of this complex and time-consuming objective was the terminology, to be intended as the subset of the Hebrew/Aramaic/Italian lexicon occurring in the Talmud and its translation and constituted by the most significant and meaningful words with respect to the talmudic domain. The examined talmudic terms, in virtue of the richness of (syntactic and semantic) relationships holding among them, provided an ideal basis to experiment with the adoption of the Meaning-Text Theory by encoding a first set of particularly meaningful lexemes in the form of an Explanatory and Combinatory Dictionary (ECD).

The structure of the paper reflects the different phases of the workflow we adopted. After a survey of the available resources related to the lexicon of the Talmud in section 2, section 3 describes the task for the automatic extraction of the terms from the Italian translation of the Talmud. The candidate corresponding Hebrew/Aramaic terms are then defined via the alignment technique described in section 4. The process of structuring the terms in the form of an ECD is detailed in section 5, followed by some examples

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

[2] E. Giovannetti, D. Albanesi, A. Bellandi, and G. Benotto, *Traduco: A collaborative web-based CAT environment for the interpretation and translation of texts*, in «Digital Scholarship in the Humanities» 32 (suppl_1) (2017), pp. 47-62; A. Bellandi, D. Albanesi, A. Bellusci, A. Bozzi, and E. Giovannetti, *The Talmud System: a Collaborative web Application for the Translation of the Babylonian Talmud Into Italian*, in «Proceedings of The First Italian Conference on Computational Linguistics», R. Basili *et al.* (eds.), Pisa University Press, Pisa, pp. 53-57.

and linguistic considerations in section 6. The web application used to browse and manage the terminology is shown in section 7, while section 8 concludes the paper presenting some final discussions and the next steps of the research.

2. *Related works*

Since this work deals with the construction of a talmudic terminology, an overview of existing resources, both in printed and digital format, will be provided.

2.1. *Available printed resources*
David Dattilo

Among the most widespread printed works describing the domain of the Talmud and including lexicographic information it is worth mentioning *Sefer Ha-ʿaruk* by Natan ben Yeḥiel ʿAnaw (1035-1160), a vast compendium of linguistic, bibliographic and encyclopedic data, still in use today. A modern and heavily revised edition of the *Ha-ʿaruk* was edited by Kohut and published in eight volumes (*ʿAruk Ha-šalem*) between 1878 and 1892. We also cite the *Lexicon Hebraicum et Chaldaicum cum brevi Lexico Rabbinico Philosophico* by Johannes Buxtorf, printed in Basel in 1607 and reprinted in Glasgow in 1824.

In the modern age we can find some dictionaries and talmudic encyclopedias describing the lexicon and the terminology, among which it is worth mentioning: *A Dictionary of the Targumim, the Talmud Babli and Yerushalmi, and the Midrashic Literature*, by Marcus Jastrow (London and New York, 1886-1903); *Milon talmud* by Baruch Karu (Jerushalaim - I vol. 1956/57; Tel Aviv, 1967 II vol.); *Madrik La-Talmud* by Rabbi Adin Steinsaltz, Jerushalaim, 1998; *Reference Guide to the Talmud*, by Adin Even-Israel Steinsaltz (Jerushalaim 2014).

2.2. *Available digital resources*
Alessandra Pecchioli

As already mentioned in our previous article,[3] there are also some digitized linguistic reference resources for the Talmud language.

The already cited Marcus Jastrow's dictionary, written specifically to describe late Hebrew and its contemporary Aramaic forms as found in rabbinic literature, was made available in a digitized edition[4] as well. Still regarding the same languages, we also cite, for Hebrew, the *Historical Dictionary* project of the Academy of Hebrew Language[5] and, for Aramaic, the *Comprehensive Aramaic Lexicon* (CAL).[6] In addition, some interesting academic digital projects are currently underway, which mainly address the study of rabbinic Hebrew, including *Digital Mishna*[7] and *Creating Annotated Corpora of Classical Hebrew Texts* (CACCHT, a joint project of the ETCBC and the Theological Seminary at Andrews University).[8] Although the main objective of these two projects is not to create a lexicon, both involve the linguistic annotation of text, which could be exploited to support the construction of lexical resources.

3. *The automatic extraction of the Italian terms*
Simone Marchi

In computational linguistics, term extraction is considered as the first step in the task of ontology learning from texts.[9] For our

[3] A. Pecchioli, D. Albanesi, A. Bellandi, E. Giovannetti, S. Marchi, *Annotazione Linguistica Automatica dell'Ebraico Mishnaico: Esperimenti sul Talmud Babilonese*, in «Materia Giudaica» 23 (2018), pp. 281-291, especially paragraph 3.1 pp. 286-287.

[4] M. Jastrow, *A Dictionary of the Targumim, the Talmud Babli and Yerushalmi, and the Midrashic Literature*, London and New York, 1886-1903, available at http://www.tyndalearchive.com/TABS/Jastrow/.

[5] http://maagarim.hebrew-academy.org.il.

[6] http://cal.huc.edu/.

[7] https://www.digitalmishnah.org/, cfr A. Pecchioli *et al.*, *Annotazione Linguistica*, cit.

[8] https://github.com/ETCBC/CACCHT.

[9] P. Buitelaar, P. Cimiano, B. Magnini, *Ontology learning from text: An overview*, in P. Buitelaar, P. Cimiano, B. Magnini (eds.), *Ontology Learning from Text: Methods, Evaluation and Applications*, IOS Press, Amsterdam 2005.

purposes, a (candidate) term was defined as a simple (single-word) or complex (multi-words) nominal structure with modifiers. More precisely, candidate terms have to match the following regular expression:[10]

/S(A|E|EA|S)*[AS]?/

where S stands for *noun*, A stands for *adjective*, E/EA stand for *preposition*. In other words, a candidate term must start with a noun represented in the regular expression by "S" - optionally followed by an arbitrary number of adjectives or prepositions or nouns - represented by "(A|E|EA|S)*", where the asterisk stands for "0 or more times" – and, optionally, ending with an adjective or a name – represented by [AS]?, where the question mark stands for "zero or one" between adjective and name.

Given the lack of linguistic tools for Mishnaic Hebrew and Babylonian Aramaic, the extraction of candidate terms from the Italian translation was performed by using the Italian term extractor called T2K (Term to Knowledge). T2K is a linguistic platform developed at ILC-CNR[11] and composed of a set of linguistic tools designed to extract semantic and lexical information from collections of texts. Text resources given as input to T2K are processed in a pipeline of linguistic modules. The first tool is the tokenizer, which identifies each single token (i.e. common words, dates, numbers, punctuation, etc.); tokens are sent to the morphological analyser (the second step of the linguistic chain) which provides an ambiguous morphological analysis of each token: e.g. the italian word "*porto*" can be analysed as the present tense of "*portare*" (to bring) or the noun "*porto*" (harbour); once each token is associated with its morphological analysis, the part-of-speech tagger choose the best analysis using context information such as part-of-speech of previous words, information about the next words, etc.; the last step is the term extraction: based on syntactic rules and stochastics algorithms the term extractor identifies all the potential terms in the corpus and creates a list of candidate terms in order to permit a verification step performed by a scholar.

In the following fragment of text there are three terms highlighted in bold identified by T2K: "*documento di divorzio*" (divorce document), "*procuratore*" (prosecutor) and "*posto*" (place). Each word is followed by its part-of-speech (in brackets):

la (RD) donna (S) che (PR) può (VM) prendere (V) il (RD) *documento* (S) *di* (E) *divorzio* (S) direttamente (B) , (FF) può (VM) anche (B) nominare (V) un (RI) *procuratore* (S) che (PR) lo (PC) prenda (V) al (EA) *posto* (S) suo (AP)

All three terms match the regular expression illustrated above: the parts-of-speech of the multiword expression "*documento di divorzio*" are S E S (noun preposition noun), while the part-of-speech of "*procuratore*" and "*posto*" is simply S (noun).

A total of 4166 terms were automatically extracted from the current corpus composed by four of the already translated tractates (i.e. *Berakhot*, *Roš Ha-Šana*, *Ta'anit*, and *Qiddušin*). The obtained terms were then filtered using a statistical measure called *tf-idf*. The tf-idf measures the relevance of a term to a document in a corpus and consists of two parts: the term frequency (*tf*) and the inverse document frequency (*idf*). In simple words, a high value of tf-idf means that the term appears frequently in

[10] A regular expression is a string composed by a simple characters mixed with special characters that can be used to locate (match) text. While simple characters are mostly letters and numbers, special characters are useful to describe special meaning: the caret ^ is an anchor at the beginning of the string, the dollar sign $ is an anchor at the end of the string, the question mark ? means 0 or one time, the asterisk or star * means 0 or more times, the plus sign + means 1 or more times, the opening parenthesis (and the closing parenthesis) are useful to create groups, etc.; for details, see A.V. Aho and J.D. Ullman. 1992. *Foundations of computer science (Chapter 10: Patterns, Automata, and Regular Expressions)*. Computer Science Press, Inc., USA.

[11] F. Dell'Orletta, G. Venturi, A. Cimino, and S. Montemagni, *T2K²: a System for Automatically Extracting and Organizing Knowledge from Texts*, in N. Calzolari *et al.* (edd.), in «*Proceedings of 9th Edition of International Conference on Language Resources and Evaluation*», Reykjavik 2014.

a few number of documents (thus being specific for those documents). Viceversa, a low value of tf-idf means that the term is distributed in many different documents (for example, the terms "*rabbi*", "*master*", "*baraita*" which occur widely in all the Talmud tractates).[12] Table 1 shows a list of the most relevant terms (i.e. with a higher value of tf-idf) extracted from the four tractates mentioned above.

Berakhot	Roš Ha-Šanah	Ta'anit	Qiddušin
seminal emissions	year	rain	*qiddušin*
blessing	month	day	woman
bathrooms	day	fast	father
fruit of the earth	*šofar*	water	money
blessing on wine	obligation	blessing	slave
Šema'	New Year's Eve	person	part
bread from the earth	*nisan*	funeral service	owner
meal	tithe	fasts	daughter
midnight	tribunal	might of the rains	*get*
fruit of the tree	sound	water libation	bill of divorce

Table 1: An excerpt of the most relevant terms extracted from each tractate (the Italian terms have been translated in to English).

4. The addition of the Hebrew/Aramaic terms
Angelo Mario Del Grosso

Parallel texts, also called bitexts,[13] can be exploited for Natural Language Processing (NLP) tasks and for the construction of linguistic resources.[14] This is typically done by automatically detecting cross-linguistic relationships between the source and the target languages.

In this work, the talmudic translatants of the Italian terms – previously extracted (see section 3) – have been provided via a word-by-word alignment technique. The process was carried out in two steps.

Firstly, the corpus was segmented to map the parallel textual units according to different levels of granularity describing the main logical hierarchy of the resource. In such a way, a formal structure of the bitexts has been encoded exploiting the following hierarchical levels: 1) tractate, 2) chapter, 3) block, 4) logical unit, and 5) segment. Each textual passage was then identified by a unique string obtained as a concatenation of exact references for these levels. For instance, the string *19.1.2.2.4* references the parallel segment "l'uomo dà i qiddušìn?" - "*ha'iš m^{e}qaddeš?*".

At the end of this first phase of the process, the aligned corpus was encoded by adopting the *TEI guidelines*[15] to ensure long-term preservation and further computational analyses.

Afterwards, a word-by-word alignment tool was developed to obtain a finer-grained mapping where each Italian term is connected with its original Hebrew/Aramaic translatants. This outcome not only allows scholars to have a support

[12] https://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html.

[13] J. Tiedemann, *Bitext Alignment*, in G. Hirst (ed.), *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool, 2011.

[14] A.M. Del Grosso, E. Giovannetti, S. Marchi, *Enriching a Multilingual Terminology Exploiting Parallel Texts: An Experiment on the Italian Translation of the Babylonian Talmud*, in «Atti del IX Convegno Annuale Associazione per l'Informatica Umanistica e la Cultura Digitale», *Quaderni di Umanistica Digitale*, Università Cattolica del Sacro Cuore, Milano 2020, pp. 119-124.

[15] *Guidelines for Electronic Text Encoding and Interchange*: https://www.tei-c.org.

in the construction of multilingual lexicons but also to enhance the performance of the translation tool.

The implemented aligner analysed the content of each parallel segment and identified the possible links between words belonging to the different languages, on the basis of their position in the text. To bring an example, the output provided by the tool concerning the parallel segment introduced above is the sequence *0-0 4-6 2-5 3-4 2-3 1-2 0-1*. In this format, each pair of numbers defines the link between the source term and the target term taking into account the position of the word within each segment. So, the pair *1-2* corresponds to the word pair *ha-ʾiš*-l'uomo.

The tool recognises words as "alignable" by means of statistical models leveraging translation evidence. Consequently, the appropriate statistical parameters are learnt from the extant observable data through a dedicated training phase.[16] The aligner, indeed, is grounded on statistical machine translation techniques[17] able to generate a sequence of words in a source language which is considered the translation of another sequence of words in a target language. Within this framework, the alignment tool proved to fit well in the automatic analysis of the Talmud along with its translation, taking into account literary translations, deletions, insertions, and transpositions.[18] More technically, the aligner is based on stochastic generative models (i.e., IBM and HMM models)[19] and on a collection of algorithms[20] modelling the probability of linking two words in a bilingual term pair within a definite parallel segment.

For the time being, 219.000 words have been used to train the probabilistic models which are distributed on 42.000 segments extracted from the available tractates. The obtained statistical model provides, for each Italian term, a list of Hebrew/Aramaic words which are the most likely candidates as translatants of the selected term and vice-versa. Table 2 points out the case of the term *geṭ*.[21] It illustrates the Hebrew/Aramaic words corresponding to the term *geṭ* along with other related Italian words. In this latter example it is worth noting that the list of related words shows *giṭṭekh* and *bᵉgeṭ* linked respectively to the Italian words "con" and "tuo". These kinds of outcomes could be exploited for further morphological, syntactic and semantic analyses (see section 6).

Hebrew term (entropy)	*Italian term*	*other Italian related words*
גֵּט (0.172)	ghet	ora; documento; meùn; essenziale; valido; liberazione; appartiene; emancipazione; divorzio; procuratore;
בְּגֵט (0.058)	ghet	esce; con; scalzamento; confronto; acquisisce; utilizzato;
גִּיטָּה (0.290)	ghet	divorzio, dato, specificare, parlando; ha; suo; il; suoi;
גִּיטָּךְ (0.403)	ghet	tuo; dia; tu;
גִּיטֵּי (0.203)	ghet	prendi; ricevi; mio; posto; valido; arrivi
גִּיטַּיךְ (0.454)	ghet (Aramaic)	tuo; beva; tu;

Table 2: Example of Hebrew/Aramaic terms together with the corresponding Italian words (the transliteration of גֵּט follows the editorial rules of the project).

[16] P. Liang, B. Taskar, D. Klein, *Alignment by agreement*, in «Proceedings of the North American Chapter of the Association for Computational Linguistics», ACL, New York City 2006, pp. 104-111.

[17] P. Koehn, *Statistical Machine Translation*, Cambridge University Press, Cambridge, New York 2010.

[18] The aligner uses a software library developed by the NLP team at the Berkeley University: https://code.google.com/archive/p/berkeleyaligner.

[19] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, R.L. Mercer, *The mathematics of statistical machine translation: Parameter estimation*, in «Computational Linguistics» 19:2 (1993), pp. 263-311.

[20] Specifically, machine learning algorithms.

[21] The number in brackets is the entropy of the word, which denotes the confidence about the translatant: the lower the entropy, the higher the likelihood the translatant is correct.

The aligned segments can present inaccuracies as they are automatically generated. In order to overcome this issue, a revision phase of the alignment – conducted by experts – is necessary. To this end a web-based application was implemented to browse, identify, select and correct the outcomes of the aligner, where appropriate[22] (see section 7).

5. The Explanatory Combinatorial Dictionary
Silvia Piccini

The Italian terminology repertoire – extracted according to the methods described in section 3 – and the Hebrew/Aramaic equivalent terms – identified in the alignment task – have been structured and formalised according to the principles of the Explanatory and Combinatorial Lexicology (ECL).[23]

This theoretical framework, elaborated by Mel'čuk and his colleagues in the context of the linguistic Meaning-Text Theory (MTT),[24] consists in a highly formalized lexicographic model, where each term receives a rigorous, explicit and systematic description of its semantics and combinatorics.

The use of ECL principles in the compilation of specialized dictionaries is rather complex and time-consuming. Nevertheless, such a model was chosen and, as a consequence, a trilingual Hebrew-Aramaic-Italian Explanatory and Combinatory Dictionary (ECD) for the Talmud terminology has been built, for three main reasons.

First, the ECD model has been successfully adopted in the terminological field, especially in North America, where many ECD for specific domains of knowledge, such as DiCoEnviro[25] (environmental terminology), DiCoInfo[26] (computer science terminology), JuriDico[27] (legal terminology), etc., were created and made available online by the *Observatoire de linguistique Sens-Texte* (University of Montreal).

Secondly, the model underlying the ECD was considered particularly appropriate for the domain of Talmud, as the Talmud project, in its essence, is a translation project and the Explanatory and Combinatorial Lexicology is a fundamental component of the MTT, developed by Mel'čuk and the Moscow Semantic Circle precisely in the context of machine translation research. The ECD gives, indeed, a complete and exhaustive documentation of the semantic, syntactic and collocational properties of a large number of lexical units, constituting in this respect a powerful support for automatic translation tasks.

Thirdly, the theoretical model underlying an ECD has proven to be particularly effective in advanced NLP applications.[28] In this regard, the terminological data formalized in the ECD of the Talmud will be exploited to improve the results obtained in the word alignment task, so to determine more precise correspondences between words and sentences in the two parallel

[22] Any interested scholar can learn more about how the web application for the alignment verification works by reading A.M. Del Grosso, E. Giovannetti, & S. Marchi. *Enriching a Multilingual Terminology Exploiting Parallel Texts: An Experiment on the Italian Translation of the Babylonian Talmud*, in «Atti del IX Convegno Annuale AIUCD 2020», pp. 119-124.

[23] I.A. Mel'cuk, A. Clas, A. Polguère, *Introduction à la lexicologie explicative et combinatoire*, Duculot, Bruxelles 1995, p. 256.

[24] A.K. Žolkovskij - I.A. Mel'čuk, О семантическом синтезе, «Проблемы кибернетики» 19 (1967), pp. 177-238; I.A. Mel'čuk, *Опыт теории лингвистических моделей «Смысл-Текст»*, Nauka, Moscow 1974, p. 314. For a synthesis of the MTT theory see: J. Milićević, *A short guide to the Meaning-Text linguistic theory*, in «Journal of Koralex» 8 (2006), pp. 187-233.

[25] http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi.

[26] http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi.

[27] http://olst.ling.umontreal.ca/cgi-bin/juridico/search.cgi.

[28] Cf. L. Wanner, *Lexical functions in lexicography and natural language processing*, Vol. 31 in «Studies in language companion series», J. Benjamins Publishing Company, Amsterdam; Philadelphia 1996, pp. xx + 355; J. Apresjan *et al.*, *Lexical Functions in Actual NLP Applications*, in L. Wanner (ed.), *Selected Lexical and Grammatical Issues in the Meaning-Text Theory*. In honour of Igor Mel'čuk, Amsterdam, Benjamins Academic Publishers 2007, pp. 199-230.

texts, especially in the case of idiomatic syntactico-semantic patterns. At the same time, as in a virtuous circle, the sentence-aligned corpus will contribute to render the dictionary building process as automatic as possible.

In a nutshell, the description of the structure of an ECD entry will follow.

An ECD entry *L'* – called *lexie* – can be either a lexeme or a phraseme (more precisely, an idiom). *L'* is a linguistic sign, namely an ordered triplet composed of a signified (*signifié*), a signifier (*signifiant*) and some syntactic properties determining the behaviour of *L'*, when co-occurring with other signs:

L' = <'Signified' ; /Signifier/ ; Σ(syntactics)>

The triadic nature of the linguistic sign is reflected in the tripartite structure of the dictionary entry, made of a semantic zone, a syntactic zone and a lexical co-occurrence zone.

The *semantic zone* provides the lexicographic definition of the *L'* meanings, written according to a pre-established and well-developed lexicographic metalanguage. An ECD definition is composed of two separate parts, as follows:

definiendum ≡ *definiens*

The *definiendum* is expressed in a propositional form, which is composed of the lexeme *L'* and the Semantic Actants[29] (SemAs) – if any – required by the *L'* semantic nature and indicated by means of variables (X, Y, Z, W). The *definiens* is represented by the decomposition of the *L'* meaning into semantically simpler units (*L'* = *'L1'* ⊕ *'L2'* ⊕ … ⊕ *'Ln'*), to avoid the vicious circles generated by the use of synonyms in the definition.

The *syntactic zone* specifies for each SemA introduced in the definition (X, Y, Z, W) the corresponding surface morphosyntactic expressions. The government pattern of *L'* is thus described by means of a rectangular matrix having as many columns as the SemAs and as many rows as the potential morphosyntactic realisations of each SemA.

The *lexical co-occurence zone* presents the most characteristic feature of the ECD, namely the lexical functions, a powerful mechanism introduced in the MTT to represent paradigmatic relations as well as syntagmatic restricted lexical co-occurrence.[30]

The lexical function has to be intended as a multi-value mathematical function, consisting of a triple of elements {f, L, Li}:

f(L) = {Li}

where *f* represents a general semantic relationship holding between an argument *L* (named *keyword*) and another lexeme *Li*. *Li* constitutes the *value* of *f* and usually represents a set of quasi-synonymous lexical units expressing a specific meaning when applied to *L* by means of *f*.

Mel'čuk identified approximately 60 standard lexical functions divided into paradigmatic and syntagmatic functions. Lexical functions convey a general and abstract meaning. Let us consider, for example, the syntagmatic lexical function *Magn*, expressing intensification:

Magn (criticize) = bitterly, harshly, seriously, severely, strongly

In compiling an ECD the lexicographer proceeds by encoding, first, all the entries belonging to the same semantic field in order to guarantee uniformity, exhaustiveness and coherence of the description.

Our analysis focused on the *Qiddušin* tractate and, more specifically, on the terms denoting the legal documents stating the contractual relationships between the woman and her (future) husband.

Before getting into the details of an ECD entry, it is worth underlining that our trilingual ECD is conceived as a linguistic research tool and, therefore, is not aimed at the general public, but mainly at linguists, translators, or more generally at scholars interested in the Talmud language. Needless to say, it can also be exploited in NPL applications.

[29] The semantic actant is the participant involved in the *L'* being described.

[30] For details, see *inter al.* I. Mel'čuk, *Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon*, in Wanner, *Lexical functions* cit., pp. 37-102.

6. Examples and linguistics issues
Alessandra Pecchioli

As underlined in the previous section the construction of an ECD constitutes a complex task. This task becomes even more complex in the case of Talmud, as we are dealing with a text characterised by a great linguistic heterogeneity. Two languages, Hebrew and Aramaic, alternate in the various sections and often even intertwined within the same sentence. The language used in the Talmud can basically be defined as a literary language characterized by a profound stratification.

In addition, it is traditionally recognized that the style of the Talmud authors is bare, essential and so elliptical as to discourage even the average reader trying to understand the right meaning. Besides, as already mentioned in the Introduction, the edition taken as reference within the Project, namely the printed edition of Vilna, contributes to make the task even more difficult because, as already stated by E. Yechezkel Kutscher, the classic printed editions, especially for Aramaic sections, are not entirely reliable, not being designed by today's scientific standards.[31]

An ECD devoted to formalizing the Hebrew and Aramaic terminology of the Talmud constitute an *unicum* as, for the first time, the focus is on dead languages. One of the main characteristics of many ancient languages is that they are found in very small *corpora* compared to modern spoken languages, with no possibility of verification by native speakers.[32] As a consequence, an ECD devoted to dead languages needs adaptations. In addition, in the light of all the above, it is important to emphasize that extracting sufficiently sharp semantic data and lexical functions from such a text is not a trivial task at all. This entails that scholars studying such languages have to rely only on their linguistic and conceptual knowledge of the domain.

In the previous section it has already been mentioned that the construction of the ECD moves by exploring semantic fields. The *Qiddušin* tractate was therefore chosen, devoted to describing the act of purchasing an asset. Within this vast subject, several more specific semantic fields can be identified. Our analysis focussed on the concept related to the dissolution of the bond that the purchased object establishes with the owner, starting from the moment of its acquisition. In *Qiddušin* such an object can be a woman, a slave, an animal, movable or immovable property, the sanctuary, etc. More specifically, a characteristic marked terminology is used when such a particular bond between the buyer – generally a man – and other human beings – typically a woman or a slave – is cancelled (for example *geṭ*, *gerušin*, *qanah* 'et 'aṣmah, etc). Among all the lexical units, the term *geṭ* was chosen.

Table 3 illustrates the morphological, semantic and syntactic description of the Hebrew entry *geṭ* based on data taken from *Qiddušin*. Nevertheless, it has been helpful to consult other texts to support the interpretation of the meaning of the keyword.[33]

Two meanings were identified, expressed respectively by roman numerals I and II. The first is non-specific and denotes the generic meaning of legal document, the second is more

[31] E.Y. Kutscher, *Talmudic Text Samples*, in F. Rosenthal (ed.), *An Aramaic Handbook: with contributions by Z. Ben-Hayyim [et al.]*, Harrassowitz, Wiesbaden 1967; *Idem*, *sub* "Aramaic" in *Encyclopaedia Judaica*, vol. 2, Keter Publishing House, Detroit 2007², p. 354.

[32] For example the Bible, one of the largest texts, which composes the ancient Hebrew *corpus*, has an average of 432,691 tokens, *vs* 110,691,482 tokens of the British National Corpus (C. Christodouloupoulos - M. Steedman, *A massively parallel corpus: the Bible in 100 languages*. Lang Resources & Evaluation 49, 375-395, 2015, tab. 3, https://link.springer.com/article/10.1007/s10579-014-9287-y#citeas).

[33] For example, other Talmud tractates, lexicons and encyclopedias to substitute, at least in part, the native speaker's function. In our case, as just noted, the *corpus* of this language is extremely small as far as number and topics are concerned. Consequently, if the analysis focuses only on one tractate, the semantic features of lexemes will be particularly biased. However, all these considerations should not be of discouragement, because this is the condition of most of the ancient texts and because our main objective is to demonstrate that building an ECD could help clarify many relations and so remedy some of the problems.

marked and is the most relevant in our tractate.[34] In the table only this latter meaning will be illustrated in detail. As the Government Pattern section shows, *geṭ*$_{II}$ can introduce only one argument, which can be realised as a noun or by means of the third-person pronoun suffix feminine singular. The expression "*If C1 = Λ, then X = našim*" underlines that *geṭ*$_{II}$ can only be accompanied by the term *našim*. If *geṭ*$_{II}$ occurs in the absolute state, then the default interpretation is "*geṭ našim*". Unlike *geṭ*$_{II}$, which is regularly used in the absolute state to denote the "document of divorce", *geṭ*$_{I}$ is always accompanied by several nouns (for example šiḥur and *ḥaliṣa* in *Qiddušin*).

The third part, entirely dedicated to the lexical functions, deserves more attention. Identifying lexical functions in the Talmud lexicon is a non-trivial task, as underlined several times, and it is in this context that the most relevant adaptations have been implemented. In them we find the complexity of the relationships that make up a sentence.

Here is an example. As is well known, the Talmud is rich of Bible quotations because of its way of proceeding. In particular, in *Qiddušin* 3b, Dt 24 is quoted to explain how the divorce had to take place. In this context we find the biblical equivalent of *geṭ*$_{II}$. To describe this relationship, a new label "*bibl*" was introduced to specify that the lexical function value is represented by a term occurring only in the Bible context. This label underlines that the reader comes across a diachronic variant, which belongs to the older language stage of the Bible. As proof of this is the fact that the term is translated into a version understandable for readers of the time, i.e. *sefer kortah*. Needless to say, other interpretations are possible. However, this example illustrates how the formalism adopted in the Explanatory Combinatorial Lexicology asks the lexicographer to reflect on certain linguistic dynamics and to make them explicit. The result of this burdensome activity is a dictionary that – more than traditional dictionaries – can shed light on the true functioning of the language in context.

The Talmud ECD, thus, could help to know more deeply the languages of this text and the relationships existing between them and previous *corpora*.

GEṬ, noun, masc.

I. Official written document attesting the juridical relationship between X and Y
II. Official written document stating that X is no longer tied to Y by the marriage bond

II. *geṭ* of X = Official written document stating that X is no longer tied to Y by the marriage bond

GOVERNMENT PATTERN

1 = X
1. N_{pl} 2. Suffix of personal pronoun

C1.1 : ***našim***
C1.2 : ***-ah***
If C1 = Λ, then X = ***našim***
C1 : ***giṭṭey našim; giṭṭah***

[34] Unfortunately in an ECD the semantic features are taken for granted. Nonetheless, knowing the process that leads to their identification would be particularly interesting to understand the complexity of the task.

MORPHOLOGICAL DESCRIPTION

	Absolute st.	Construct st.
Sing.	*geṭ*	*geṭ/giṭ-*
Pl.	*giṭṭin*	*giṭṭey*

LEXICAL FUNCTIONS[35]

Gener	:	*giṭṭey našim*	Relation expressing the closest generic concept
Anti	:	šetar	Relation expressing the semantic operation of negation
$Syn_{\subset}$	:	get_{1}	Less specific synonym
$Syn_{\subset}$	:	**bibl** *sefer keritut*	Equivalent term occurring in a Bible quotation (Dt 24,1)
$Syn_{\supset}$	:	*me'un*; *mequšar*	More specific synonym
$Syn_{\cap}$	:	$gerušin_{pl}$	Intersecting meanings
S_1	:	*ba'al*; 'ab	Name of the first actant
S_2	:	*'iša*	Name of the second actant
S_3	:	"*harey 'at m^{e}šullaḥat*"; "*harey 'at m^{e}gorešet*"; "*harey 'at muteret l^{e}khol 'adam*"	Content of *geṭ*
Ver	:	*ze~*; *harey~*	Relation describing intended requirements
Instr	:	*b^{e}*	Preposition with the meaning 'by means of'
$Caus_1Func_0$	:	*katab~*	Cause of the creation of *geṭ*
$Oper_1$	:	*natan~*	Object of a verb
$Oper_2$	:	*qibbel~*	Object of a verb
Degrad	:	*en~*	Contrary of Ver
$Func_0$	:	*ṣarikh*	Complement of modal verb

[35] For symbols and abbreviations of Lexical Functions see Mel'čuk, 1996 *op. cit*. The first nine Lexical F identified are paradigmatic; the following relations are all syntagmatic.

CONTEXTS[36]

כָּתַב גֵּט וּנְתָנוֹ בְּיַד עַבְדָּהּ יָשֵׁן וּמְשַׁמַּרְתּוֹ – הֲרֵי זֶה גֵּט. נֵיעוֹר (TB, *Qiddušin*, 44b)

If one wrote a bill of divorce and placed it in the hand of his wife's slave, who is sleeping, but she is guarding him, then it is a valid bill of divorce.

קְטַנָּה שֶׁאָמְרָה "הִתְקַבֵּל לִי גִּיטִּי" אֵינוֹ גֵּט, עַד שֶׁיַּגִּיעַ גֵּט לְיָדָהּ (TB, *Qiddušin*, 44b)

In the case of a minor girl who said to an agent: Receive my bill of divorce for me, it is not a valid bill of divorce until the bill of divorce reaches her possession.

הָא נַעֲרָה – הֲרֵי זֶה גֵּט! (TB, *Qiddušin*, 44b)

But in the case of a young woman, the bill of divorce is valid.

אִם גֵּט – לָמָּה מֵיאוֹן? אִם מֵיאוֹן – לָמָּה גֵּט (TB, *Qiddušin*, 44b)

If she needs a bill of divorce, why does she require refusal? Conversely, if she requires refusal, why does she require a bill of divorce?

Table 3: ECD entry example of *geṭ* based on data from *Qiddušin*.

7. *The interface to the resource*
Andrea Bellandi

In the field of lexicography, the application of computational techniques has profoundly changed the way the linguistic resources are built. At the same time, a significant contribution is given by the increasing use of the Semantic Web technologies (RDF, OWL, SPARQL, etc.) and of the Linked Data paradigm, to create FAIR[37] (i.e. findable, accessible, interoperable, and reusable) lexical and terminological resources. In light of this trend, a tool – called LexO[38] – has been developed at the Institute of Computational Linguistics.

LexO is based on the lemon model (Lexicon Model for Ontologies),[39] which can be considered a *de facto* standard for the representation of computational lexicons in the Semantic Web. The linguistic information is organized by means of 3 modules: i) the lexicographic module (OntoLex) concerning the structure of each entry, e.g., lemma, other forms, senses, ii) the variation and translation module, aiming to represent the variation of relations across entries in the same or different languages (e.g., dialectal, register, and translation relations, or morphological and orthographic ones), iii) the syntactic and semantics module, devoted to describing the syntactic behaviour of an entry, its valence (the syntactic arguments involved by the situation

[36] The English translation of the Talmud is taken from Rabbi Dr Tzvi Hersh Weinreb (ed.), *Koren Talmud Bavli, Kiddushin*, commentary by Rabbi Adin Even-Israel Steinsaltz, Koren Publishers, Jerusalem 2015.

[37] M. Wilkinson, M. Dumontier, I. Aalbersberg, *et al.*, *The FAIR Guiding Principles for scientific data management and stewardship*, in «Scientific Data», vol. 3, 160018 (2016) doi:10.1038/sdata.2016.18.

[38] A. Bellandi, E. Giovannetti, S. Piccini, A. Weingart, *Developing LexO: A Collaborative Editor of Multilingual Lexica and Termino-ontological Resources in the Humanities*, in F. Frontini *et al.* (eds.), *Proceedings of the Language, Ontology, Terminology and Knowledge Structures*, ACL, Montpellier, France 2017, pp. 85-94.

[39] J.P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar and P. Cimiano, *The OntoLexLemon Model: Development and Applications*, in «Proceedings of Electronic Lexicography in the 21st Century», Lexical Computing CZ s.r.o., Brno, Czech Republic 2017, pp. 19-21.

the word refers to) and the link to the ontological structures representing the entry meaning.

Figure 2 shows the structure of the entry *geṭ*. The canonical form, described in the box on the left, typically corresponds to the lemma of the word and is determined by the lexicographic conventions adopted in that language. For each lexical sense, illustrated in the boxes on the right, a definition in natural language is provided. By clicking the plus button inside the lemma box, users can also add information concerning other forms and new lexical senses. By selecting the "variation and translation" or "syntactic and semantics" tab, a user can enrich the lexical entry with the suitable information.

The translation variation module consists of two kinds of relations: 1) the semantic relations holding between senses and including terminological relations (dialectal, register, chronological, discursive, and dimensional variation) and the translation relation ; 2) the relations linking lexical entries and/or forms, which describe, for example, the morphological and orthographic variations of the word.

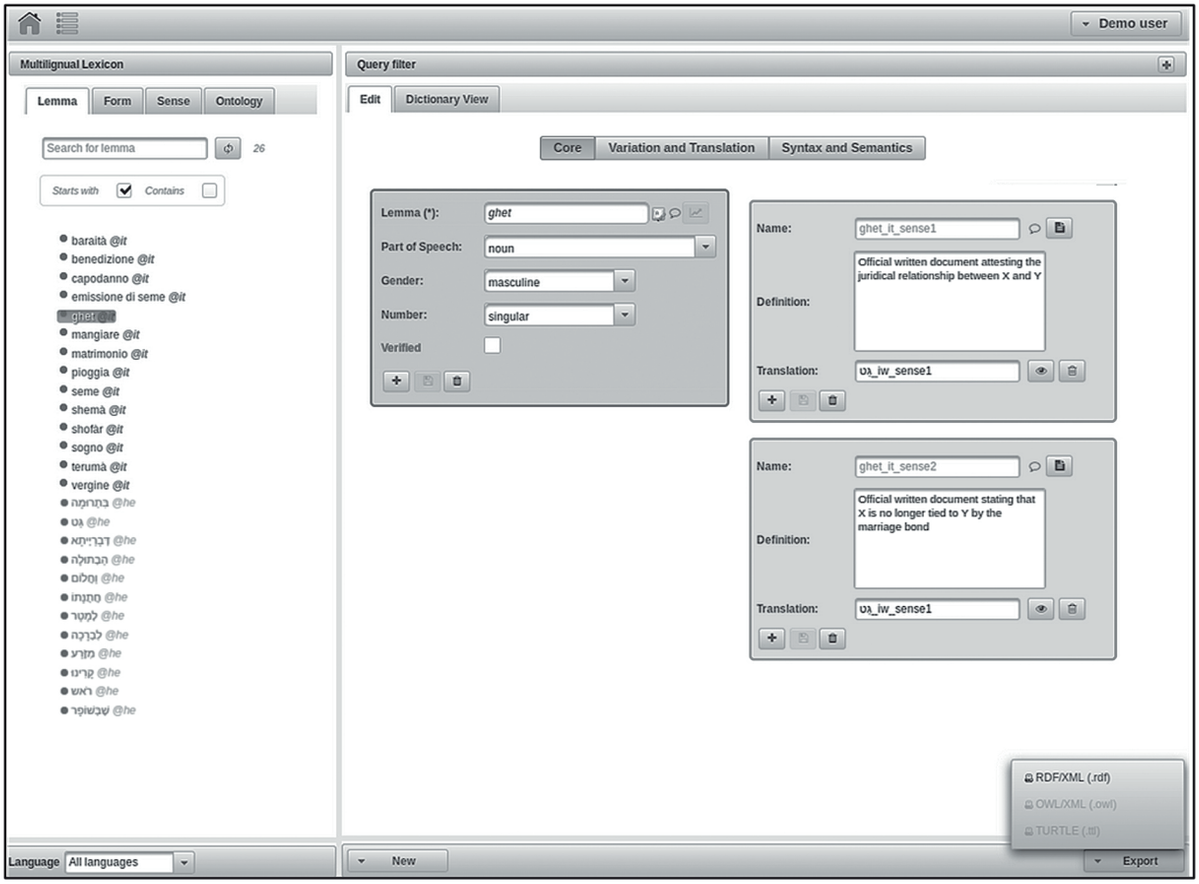

Fig. 1 - LexO's main interface.

The syntactic and semantic module describes the syntactic behaviour of a word and its government pattern, namely the actants introduced by the word, their syntactic function and their morpho-syntactic realisation. These syntactic frames need also to be bound to the ontological structures representing their meaning. As a consequence, LexO makes it possible to map the argument of a predicate defined in an ontology and the syntactic argument introduced in a given syntactic frame. Finally, by selecting the "dictionary view" tab, a dictionary-like rendering of all the information related to the selected entry is shown in the central panel of LexO.

A customization of LexO has recently started with the objective of managing the most innovative aspect of the Explanatory and Combinatorial Lexicology, i.e. the lexical functions presented in section 5. In this regard, LexO aims to reuse *Lexfom*,[40] a lexical functions ontology model, which uses the vocabulary of the lemon model, in particular the classes "LexicalEntry" and "LexicalSense". Such an ontology will be integrated in LexO in order to make it possible for scholars to build an ECD resource. The dictionary view part will be extended, according to the layout presented in Table 3.

8. Discussion and next steps
Emiliano Giovannetti

We are well aware that the construction of a terminological resource structured in the form of an ECD requires much time and resources, also due to the complexity of the talmudic domain and the involved languages. For this reason we chose to adopt a development strategy structured in steps, which provides for the incremental addition of linguistic information to the talmudic terms, expressed both in the languages of the Talmud and in Italian. In this way, the multilingual resource will always be available for the tasks in which it will be exploited, but, at the same time, it will be possible to enrich it with more data at any time, depending on the availability of time and people with the appropriate skills. To ease the work of construction of the resource we intend to take advantage of linguistic analysis procedures to automatically suggest to the lexicographer candidate lexical functions related to a given lexeme. We will also exploit the research we have already conducted on the linguistic tagging of Mishnaic Hebrew,[41] in particular to integrate linguistic information (e.g. lemma, part-of-speech, etc.) in the process of automatic alignment described in section 4.

The terminological resource will be initially used in three contexts. First of all, within the project, to provide an additional support in the whole translation process, for use by translators, revisors and editors. In particular, translators will be able to count on a dedicated component of Traduco (the environment used for the translation of the Talmud) that, given a textual segment to be translated, will provide information about the domain terms appearing in the segment, including all the possible translations and all the available linguistic data. All these information will be exploited by users for a more precise interpretation of the text. Furthermore, content revisors and editors will be able to carry out much more accurate checks on the text and run advanced linguistic-based searches, for example to visualize all the textual passages including inflected forms of a certain verb, such as "to divorce".

Secondly, the terms collected in the resource will be interpreted as topics of discussion, and, as such, used for a more content-based analysis of the text. In particular, we intend to use the terminological resource as a component of a Talmudic Knowledge Base (TKB). This much broader and extensive resource, in addition to the terms and the text itself (and its translation), will include the concepts denoted by the terms, which will be appropriately structured in the form of an ontology. All the concepts will be formally associated with the Masters who deal with them in the talmudic discussions. Once the Masters and the topics denoted by the correspondent terms will be formalized in the TKB, it will be possible to query the text on a semantic basis by asking, for example, to visualize "all the passages of the Talmud where second generation amoraim discuss about bitter water".

Finally, thanks to the terminological resource it will be possible to offer Talmud scholars an additional tool for an advanced access to the text, language and contents of the text.

The terminological resource, as already suitably formatted according to the criteria of the Linguistic Linked Open Data (by virtue of the structure based on the lemon model), will be

[40] A. Fonseca, S. Fatiha, F. Lareau, *Lexfom: a Lexical Functions Ontology Model*, «Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon», the COLING 2016 Organizing Committee, Osaka 2016, pp. 145-155.

[41] A. Pecchioli, D. Albanesi, A. Bellandi, E. Giovannetti, S. Marchi, *Annotazione Linguistica Automatica dell'Ebraico Mishnaico: Esperimenti sul Talmud Babilonese*, in «Materia Giudaica» 23 (2018), pp. 281-291.

released, shared and connected to other pre-existing resources, thus making a significant contribution to the scientific community related to Jewish studies.

9. Acknowledgment

This work was supported by the TALMUD project and carried out within the scientific collaboration between S.c.a r.l. PTTB and ILC-CNR (09/11/2017).

Emiliano Giovannetti
e-mail: emiliano.giovannetti@ilc.cnr.it

Andrea Bellandi
e-mail: andrea.bellandi@ilc.cnr.it

David Dattilo
e-mail: david.dattilo@talmud.it

Angelo Mario Del Grosso
e.mail: angelo.delgrosso@ilc.cnr.it

Simone Marchi
e-mail: simone.marchi@ilc.cnr.it

Alessandra Pecchioli
e-mail: alepec3@gmail.com

Silvia Piccini
e-mail: silvia.piccini@ilc.cnr.it

SUMMARY

A formal digital structuring of the terminology of the Talmud is being carried out in the context of the Project for the Translation of the Babylonian Talmud into Italian. The terminological resource was encoded in the form of a multi-language Explanatory Combinatorial Dictionary (Hebrew-Aramaic-Italian) according to the principles of the Meaning-Text Theory. The construction of such a resource was supported by text processing and computational linguistics techniques aimed at automatically extracting terms from the Italian translation of the Talmud and aligning them with the corresponding Hebrew/Aramaic source terms. The paper describes the process that was set up for the construction of the terminological resource with the ultimate goal of illustrating the advantages of the adoption of a formal linguistic model. The terminological resource aims, indeed, to be a useful tool to deepen the characteristics of the languages of the Talmud, to help translators in their work and more generally, scholars in their study of the Talmud itself.

KEYWORDS: Babylonian Talmud; Computational Linguistics; Explanatory and Combinatorial Lexicology.