The Role of a Computational Lexicon for Query Expansion in Full-Text Search

Emiliano Giovannetti, Davide Albanesi, Andrea Bellandi, Simone Marchi, Mafalda Papini, Flavia Sciolette
Istituto di Linguistica Computazionale, Via G. Moruzzi 1, 56124, Pisa
name.surname@ilc.cnr.it

Abstract

English. This work describes the first experiments conducted with a computational lexicon of Italian in a context of query expansion for full-text search. An application, composed of a graphical user interface and backend services to access the lexicon and the database containing the corpus to be queried, was developed. The text was morphologically analysed to improve the precision of the search process. Some examples of queries are given to show the potential of a text search approach supported by a complex and stratified lexical resource.

Italiano. *Il presente lavoro illustra i primi esperimenti condotti con un lessico computazionale dell'italiano in un contesto di query expansion per la ricerca full-text. È stata sviluppata una applicazione composta da una interfaccia grafica utente e un backend di servizi che permette l'accesso sia al lessico che al database contenente il corpus da interrogare. Il testo è stato analizzato morfologicamente al fine di migliorare la precisione del processo di ricerca. Alcuni esempi di query sono forniti al fine di mostrare le potenzialità di un approccio di ricerca sul testo supportato da una risorsa lessicale complessa e stratificata.*

1 Introduction

The need of techniques going beyond the mere "search by keyword" in the querying of textual resources dates back to the dawn of computational linguistics. Seminal works in the 60s on the development of the very first question answering (QA) systems already included linguistic resources as support datasets. To bring some "old school" examples, the "General Inquirer" QA system (Stone et al., 1962) used a thesaurus for "coding words as to concept membership" while Simmon's "Protosynthex" was equipped with a synonym dictionary (Simmons et al, 1963) to "expand the meaning of the question's words to any desired level". One of the first works specifically focussed on the use of a lexical resource for NLP tasks was about COMPLEX (for "COMPutational LEXicon"), a resource developed at IBM (Klavans, 1988).

The support of linguistic resources has proved its potential in the field of information retrieval (IR) too, as highlighted in many of Bill Woods' works, culminating in the introduction of his conceptual indexing technique and the conceptual taxonomy resource (Woods, 1997) and later refined in an article entitled "Linguistic Knowledge can Improve Information Retrieval" (Woods, et al, 2000). More recently, other researchers have stressed the importance of the availability of a "Lexical Knowledge Base" (another way to refer to a computational lexicon) in tasks such as Word Sense Disambiguation, since their use, in some contexts, can outperform supervised systems (Agirre et al., 2009).

The use of linguistic resources in QA of the earliest period of computational linguistics can be considered as the precursor of "query expansion" (QE), the technique that Manning and Raghavanat describe as the most used "local method" in IR to tackle those situations in which "the same concept may be referred to using different words" (Manning et al., 2008).

Though QE may be obtained in different ways (among which query reformulations based on query log mining) we are here interested in those applications that make use of lexical resources.

Most of the works, published from the 90s to nowadays (proving that QE is still being investigated), exploit WordNet (Fellbaum, 1998), the *de facto* and most widespread ontological (or lexical, depending from the point of view) multilingual resource. Ellen Vorhees was one of the first and used WordNet's IS_A relations to improve text retrieval (Vorhees, 1993). Moving on directly to the most recent works, WordNet has been used with all its ontological features to expand queries in a semantic text search context in (Ngo et al., 2018) while in (Azad and Deepak, 2019) the authors combined WordNet and Wikipedia for QE, exploiting the first to expand individual terms and the second to expand phrase terms.

The research work here illustrated places itself in the context of full-text search carried out using a lexical resource-driven QE technique. However, the focus of this research, differently from that of the cited works, is not on the specific QE technique and the relative evaluation, but on the resource we chose to exploit, introduced in the next section, in place of WordNet and on the frontend and backend technologies implemented to query the text, as described in details in Section 3. The advantages derived from the adoption of a rich and highly structured computational lexicon will also be remarked through some query examples shown in Section 4. The developed application can be freely accessed and used to query the corpus[1].

2 The Context and the Resource

This work stems from the activities conducted by the Institute of Computational Linguistics of CNR (ILC-CNR) in the context of the Talmud Translation Project[2]. The need of providing a way to query the Italian translation of the Talmud[3] on a linguistic basis was the initial spark that led to the idea of experimenting the use of a computational lexicon for Italian. As a matter of fact, this resource (described below) represents a "linguistic mine" which has never been exploited for tasks of full-text search or information retrieval.

2.1 The Parole-Simple-Clips Lexicon

"PAROLE-SIMPLE-CLIPS" (PSC) is a computational lexicon of Italian, developed from 1996 to 2003 by ILC-CNR (Ruimy et al., 2002). Currently, the resource is stored as a MySQL database available on CLARIN[4], and represents a *unicum* among the available linguistic resources for Italian, thanks to its richness and articulated structure of data. Based on the Generative Lexicon theory (Pustejovsky, 1995), the schema on which the linguistic information is encoded is composed of four distinct, but strictly interconnected layers of analysis: phonology, morphology, syntax, and semantics.

In these features lies the motivation of this work, since the available linguistic information may be combined in ways that go well beyond what resources such as WordNet allow to do in the context of text search support. Even considering semantics alone, the information in PSC is detailed with fine-grained features that are not described in WordNet's network of synsets: PSC encodes the meaning of each lexical sense as an array of information, including "templates" (see below), semantic traits, semantic roles, and argumental structures.

In this work, we document the first steps in the use of PSC for QE. At this stage we used: i) the Morphological Units, classified according to their POS, which represent the lemmas of the computational lexicon; ii) the Phonological Units that represent the inflected forms of the lemmas; iii) the Semantic Units (SemUs), that describe the senses expressed by the words. Furthermore, we considered the following morphological and semantic information: i) morphological traits (e.g. gender, number); ii) relations between SemUs (at the moment limited to synonymy and hyponymy); iii) the association between SemUs and "templates", representing sets of senses, labeled according to one of the types represented in the Simple Ontology (Lenci et. al., 2001). The other parts

[1] https://klab.ilc.cnr.it/talmudSearch/

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

[3] The corpus here queried is limited to eight tractates of the babylonian Talmud: Rosh Hashanah, Berakhot, Ta'anit, Kiddushin, Chagigah, Beitza, Sukkah, and Megillah

[4] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-88.

of linguistic information will be the subject of future works, according to an incremental approach.

3 The Process and the Application

The whole search process involves a series of steps that can be summarized as follows (see Fig. 1 for a schematic functional architecture of the application):

i) the user inserts a first set of data to formulate the desired query in the Graphical User Interface;
ii) the interface requests, via Web API, the lexicon backend services which return the linguistic data matching the initial query;
iii) the user completes the query taking into account the linguistic data and starts the search;
iv) the interface executes the query expansion and requests, via Web API, the text backend services which collect, tag, and return the matching textual portions of the Talmud;
v) the interface shows the results to the user.

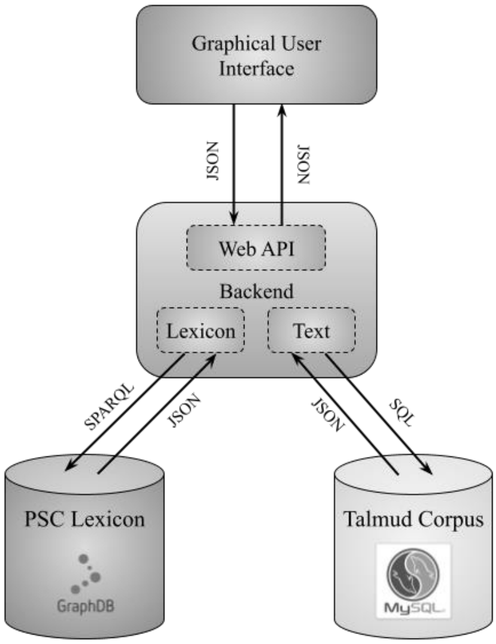


Figure 1. Functional architecture of the application.

First of all, to make the lexicon efficiently queryable, it needed to be transformed from relational data into linked data (Section 3.1). At the same time, a list of services to query both PSC and the database storing the Italian translation of the Talmud needed to be developed in order to answer to the interface requests (Section 3.2). The interface itself was designed on the basis of the available linguistic information exposed from PSC and developed accordingly (Section 3.3). Finally, to improve the precision of the search process, the queried corpus was also POS-tagged (Section 3.4).

3.1 A First Conversion of PSC

The first phase of our work was to consider the relational database of PSC as the data source for the generation of a first Linked Data (LD) conversion. Two main reasons led to the need for a conversion of PSC: i) to ease the reuse of the lexicon itself, in virtue of the intrinsic nature of LD, ii) the possibility of performing automated reasoning on data if appropriately modeled taking into account ontological principles, for example to compute inferred closures, infer new knowledge on the basis of class taxonomies, property hierarchies, and so on. Accordingly to the LD principles, we first had to look for existing vocabularies for the modeling of lexicons.

In the context of the Semantic Web, the *de facto* standard for representing lexical information is the lemon model (Cimiano et al., 2016). Its core module, called OntoLex, allows to represent grammatical, basic morphological and semantic information by means of three main classes: Lexical Entry, Form (lemma and inflected forms), and Lexical Sense. Lemon relies on external vocabularies to define semantic relations between senses: in this conversion we modelled PSC's synonymy and hyponymy with LexInfo ontology[5]. Currently, the converted resource includes 72006 lexical entries (48735 nouns, 6522 verbs, and 11830 adjectives), 469726 inflected forms, and 57130 senses. Explicit lexico-semantic relations include 1803 meronyms, 4060 synonyms, and 44487 hyponyms. This initial conversion of PSC as Linked Data was purely functional to the linguistic querying of the Italian translation of the Babylonian Talmud[6]. Therefore, it was decided to convert a selected number of linguistic data to be exploited for the process of query expansion. At the time of writing this

[5] https://lexinfo.net/

[6] We remark that the conversion of PSC Simple is not the focus of this work, but it was necessary for performing linguistic searches experiments on the Italian translation of the Talmud.

proposal, a complete conversion of PSC as LOD (Linked Open Data) is in progress. This complete conversion will also take full advantage of the already available works on the resource as documented in (Khan et al., 2018) and (Del Gratta et al., 2015).

3.2 Setting up the Backend

Once the computational lexicon was converted, the implementation of the querying system continued with the creation of the backend services needed to access both the lexicon and the database storing the text to be queried. Regarding the lexicon, a GraphDB[7] repository, containing all the converted data, was set up. The access to the repository was implemented with a set of REST services that can be invoked from any web client[8]. The services have been based on the already available backend of LexO, a collaborative web tool for the creation and editing of lemon lexical resources (Bellandi, 2021). At the same time, a list of analogous services was made available to retrieve the textual portions of the corpus matching the expanded queries coming from the frontend of the system. The Italian translation of the babylonian Talmud is currently stored as a MySQL database, where each segment of text appears both in its original and POS-tagged version (see 3.4).

3.3 The Graphical User Interface

The GUI (Fig. 2) set up to query the corpus was developed using Angular[9], one of the most widespread frameworks for frontend Web development, which provides high levels of portability and scalability. In this first version of the search system, the interface was conceived as a sort of "hub" of the whole architecture: from the one side to interact with the user and from the other side to invoke the services exposed by GraphDB and the Talmud database. The interface is divided into two sections. In the left-hand column, the available tractates of the Talmud that can be queried are represented as a tree allowing the user to specify the search context at different levels of granularity. The right-hand section contains the search parameters, where the user can choose between three types of search using the available tabs: Keyword, Form/Lemma, or Semantic Traits.

The first one is the classic keyword-based search. The second type, via the Form/Lemma tab, allows to search for a specific word form or the set of inflected forms of a given lemma by specifying some morphological constraints. By entering a word in the text field, the GUI invokes the lexicon backend services to retrieve the lemmas corresponding to the indicated parameters and displays them with their different senses. Users can then proceed with the search or they can select one or more lemmas and apply to them morphological constraints by clicking on the three dots icon on their right. The selection of at least one of the senses enables the semantic extension search feature: a drop-down menu allows users to look for all the other senses in the lexicon appearing as hypernyms, hyponyms, or synonyms at a specified distance. The forms obtained with this extension are subject to the propagation of the morphological constraints applied to the lexical entry to which they are linked, whether explicit (entered from the interface) or implicit (in the case of a search by form). Finally, the "semantic traits" tab provides two template trees on which multiple selections are possible: the first click selects a template with all its descendants, the second deselects the descendants, and the third deselects the node itself. When the selection changes, the lexicon is queried to obtain the list of senses linked to the chosen templates. Users can then select the desired senses which will be used to retrieve the forms of the relative lemmas to be used in the QE.

All the entered data are used to compose the expanded query, which will be constituted by all the inflected forms provided by the lexicon and matching the indicated morphological constraints, semantic extension, or templates.

The results coming from the backend services accessing the Talmud database are then displayed in a table on the right-hand side, upon which a panel lists the forms retrieved from the lexicon and used for the QE.

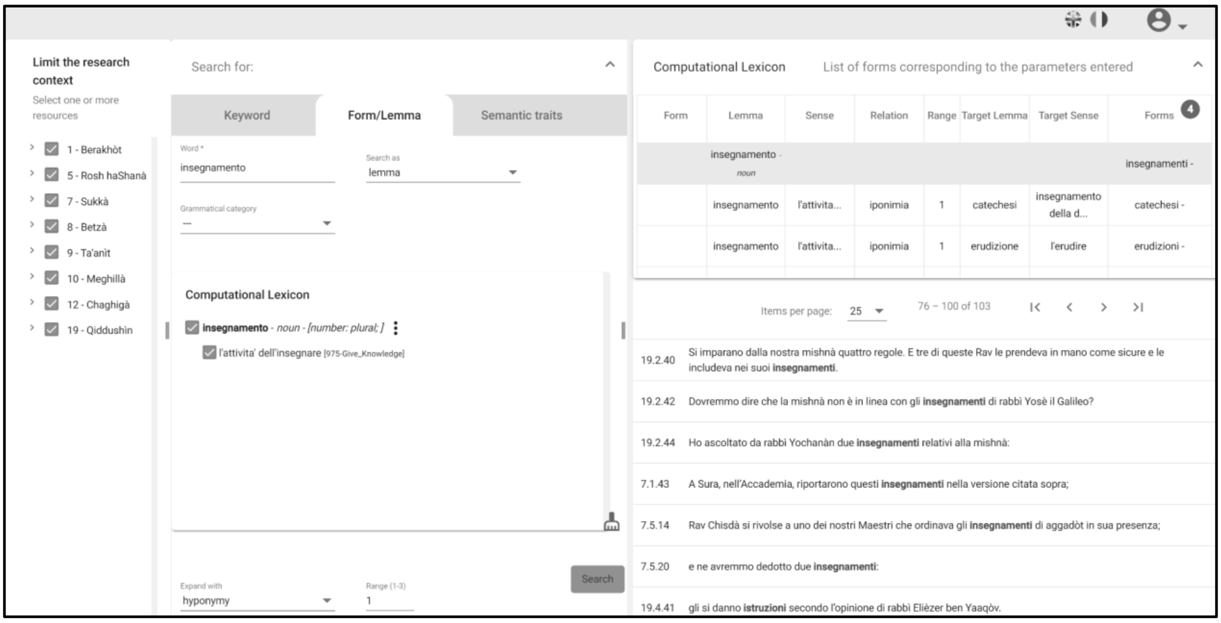
[7] Ontotext GraphDB is a highly efficient and robust graph database with RDF/OWL and SPARQL support (https://graphdb.ontotext.com/documentation/free/free/graphdb-free.html)

[8] The source code of the REST services is available at https://github.com/andreabellandi/LexO-backend

[9] https://angular.io/

3.4 POS-Tagging of the Text

For the purpose of reducing the lexical ambiguity in cases where a searched word could match with homographs, the corpus was automatically analyzed and annotated with morphological information.



Limit the research context

Select one or more resources

- ☑ 1 - Berakhòt
- ☑ 5 - Rosh haShanà
- ☑ 7 - Sukkà
- ☑ 8 - Betzà
- ☑ 9 - Ta'anìt
- ☑ 10 - Meghillà
- ☑ 12 - Chaghigà
- ☑ 19 - Qiddushìn

Word *: insegnamento

Search as: lemma

Grammatical category: ---

Computational Lexicon

☑ **insegnamento** - *noun - [number: plural;]*

☑ l'attivita' dell'insegnare [975-Give_Knowledge]

Expand with: hyponymy

Range (1-3): 1

Computational Lexicon

Form	Lemma	Sense	Relation	Range	Target Lemma	Target Sense	Forms 4
	insegnamento - *noun*						insegnamenti -
	insegnamento	l'attivita...	iponimia	1	catechesi	insegnamento della d...	catechesi -
	insegnamento	l'attivita...	iponimia	1	erudizione	l'erudire	erudizioni -

Items per page: 25 76 – 100 of 103

19.2.40 Si imparano dalla nostra mishnà quattro regole. E tre di queste Rav le prendeva in mano come sicure e le includeva nei suoi **insegnamenti**.

19.2.42 Dovremmo dire che la mishnà non è in linea con gli **insegnamenti** di rabbì Yosè il Galileo?

19.2.44 Ho ascoltato da rabbì Yochanàn due **insegnamenti** relativi alla mishnà:

7.1.43 A Sura, nell'Accademia, riportarono questi **insegnamenti** nella versione citata sopra;

7.5.14 Rav Chisdà si rivolse a uno dei nostri Maestri che ordinava gli **insegnamenti** di aggadòt in sua presenza;

7.5.20 e ne avremmo dedotto due **insegnamenti**:

19.4.41 gli si danno **istruzioni** secondo l'opinione di rabbì Elièzer ben Yaaqòv.

Figure 2. The graphical user interface showing the example of lemma “insegnamento”.

In particular, we parsed all the sentences of the eight tractates of the babylonian Talmud with Stanford's Stanza tools (Qi et al., 2020) using the pre-trained model based on the UD Italian ISDT treebank[10]. The tool was configured to use the processors for tokenization, multi-word token expansion, and Part-of-Speech tagging, which also includes the attribution of morphological traits. Each morphologically annotated textual segment was then stored in the MySQL database to return just the forms matching with the morphological constraints coming from the GUI.

4 Examples of Queries

In this last section, we show a concrete application of the approach by introducing some query examples. Each query can also be tested by the reader by accessing the available application.

The first two examples show the search for words with specific morphological traits and the application of semantic extension. In these cases, the “Form/Lemma” type of search is selected. In the first example, the word “insegnamento” (teaching) is inserted as a lemma. The system finds it in the lexicon and shows it as a noun with one single sense. The user then adds a morphological constraint by setting the “number” trait to “plural”. Finally, the user extends the search to direct hyponyms (distance = 1) and submits the query.

This is a simple case of propagation of the morphological traits through semantics. The lexicon contains the two following key information: i) the fact that the sense of “insegnamento” has three hyponyms: erudizione” (erudition), “istruzione” (instruction), and “catechesi” (catechesis); ii) all the inflected forms and the relative morphological traits of the searched word and its three hyponyms. On the basis of these data, the system composes the final query, which allows to search for all the plural forms of the four lemmas as nouns. As a result, 103 textual segments are retrieved, containing the words “insegnamenti” (97 matches) and “istruzioni” (6 matches) (Fig. 2).

The second example involves the verb “permettere” (to permit/allow), searched as a lemma, with morphological constraints on the finite mood (“indicative”, “subjunctive”, “imperative”, “conditional”). In addition, the user selects just one of the two available senses of the verb (the one with the definition “dare a

[10] https://universaldependencies.org/treebanks/it_isdt/index.html

qlcu la possibilità' di fare qlco" - to give sb the chance to do smth -) and then extends the search to its synonyms. In this case, the lexicon proposes two synonyms of the selected sense: the (single) senses of words "concedere" and "consentire". The resulting expanded query retrieves from the database a total of 405 matches, containing 334 strings of "permettere" (for 131 available forms of the lexicon), 44 strings of "concedere" (for 45 available forms) and 27 strings of "concedere" (for 41 forms).

The last type of search is structured as a more explorative querying of the corpus. In the semantic traits tab, the user can choose one or more between noun/verb or adjectival templates (group of senses), to look for all words relative to a specific semantic field, such as objects, weather verbs, metalanguage, etc.

In this example, the user selects the template "Air animal", which appears as a "leaf" of the sub-tree under the parent-node "Entity". Once the template is chosen, the system retrieves from the lexicon all the relative senses and shows them in a window. It is then possible to select all the available 165 senses or just some of them. Finally, the user can run the search: the system composes the expanded query and retrieves 226 textual segments of the Talmud containing words (both as lemmas and inflected forms) with senses referring to the semantic field of "Air animal": "uccello" (bird), "mosca" (fly), "cavallette" (grasshoppers), and so on.

Among future developments, a feature for a "grouped" selection of multiple templates will be added, that will allow to search for textual segments containing co-occurrences of words referring to the specified templates. To bring an example, the grouped selection of templates "Color" and "Earth animal" will retrieve segments containing multiword expressions such as "vacca rossa" (red cow), "gatta nera" (black she-cat), "oche bianche" (white gooses), etc.

5 Conclusion

As shown in this paper, the availability of a rich and structured linguistic resource (as the computational lexicon we have taken into account) seems to provide an edge over the standard query expansion techniques for full-text search based on WordNet. Now that a very first portion of the resource has been made available (though with a preliminary conversion) and the web application has been implemented, the road is cleared for the next steps.

The first critical issue that will need to be faced involves the limitedness of the resource, covering most - but not all - the lemmas, forms, and senses of standard contemporary Italian and that lacks many domain-related terms or senses. To fill this gap the resource will have to be updated and enriched with more entries.

At the same time, as anticipated, a more in-depth and rigorous conversion of PSC will have to be carried out, a process that will probably take a lot of time and research effort and that for the sake of this first experiment would have been premature and unnecessary. As soon as the whole conversion will be ready, the rest of the information encoded in the lexicon will be made available and integrated in the search process.

Though the benefits of the availability of a computational lexicon wrt WordNet (or a similar resource) may seem obvious in a context of QE for full-text search, an empirical evaluation would be desirable. However, the set up of a benchmark conceived for this purpose appears anything but easy, mainly due to the lack of comparable works or evaluation campaigns focussing on the role of linguistic resources as support.

In conclusion, we believe these first experiments carried out by querying the talmudic text appear promising, especially considering that only a small part of the lexicon has been used. In addition, the support in the disambiguation provided by the POS tagging of the text suggests that an hybridization of a resource-driven QE technique with a deeper stochastic annotation of the corpus to be queried may constitute an interesting experimental field to be investigated.

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

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2009. Knowledge-Based WSD on Specific Domains: Performing better than Generic Supervised WSD. In *IJCAI'09: Proceedings of the 21st international joint conference on Artificial intelligence*. 1501-1506.

Andrea Bellandi. 2021. LexO: An Open-source System for Managing OntoLex-Lemon Resources. *Language Resources & Evaluation.* https://doi.org/10.1007/s10579-021-09546-4

Ontology-Lexicon Community Group (W3C). Phillip Cimiano, John P. McCrae, and Paul Buitelaar (eds). 2016. *Lexicon Model for Ontologies: Community Report*. https://www.w3.org/2016/05/ontolex/#overview

Riccardo Del Gratta, Francesca Frontini, Fahad Khan, and Monica Monachini. 2015. Converting the PAROLE SIMPLE CLIPS Lexicon into RDF with lemon. *Semantic web* 6: 387-392.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database.* MA: MIT Press.

Azad Hiteshwar Kumar, and Akshay Deepak. 2019. A new approach for query expansion using Wikipedia and WordNet. *Information sciences* 492: 147-163.

Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography* 13(4): 249–263.

Fahad Khan, Andrea Bellandi, Francesca Frontini, and Monica Monachini. 2018. One Language to rule them all: Modelling Morphological Patterns in a Large Scale Italian Lexicon with SWRL. In *Proceedings of the 11th International Conference on Language Resources and Evaluation - LREC2018, 2018, Miyazaki, Japan*. hal-01832652

Judith Klavans. 1988. COMPLEX: a computational lexicon for natural language systems. In *COLING '88: Proceedings of the 12th conference on Computational Linguistics*. https://doi.org/10.3115/991719.991802

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*, Cambridge University Press.

Vuong M. Ngo, Tru H. Cao, and Tuan M. V. Le. 2018. WordNet-Based Information Retrieval Using Common Hypernyms and Combined Features. preprint arXiv:1807.05574.

James Pustejovsky. 1995. *The Generative Lexicon.* MA: MIT Press.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*.

Nilda Ruimy, Monica Monachini, Raffaella Distante, Elisabetta Guazzini, Stefano Molino, Marisa Ulivieri., Nicoletta Calzolari, and Antonio Zampolli. 2002. Clips, a multi-level italian computational lexicon: A glimpse to data. In *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC02).

Robert F. Simmons, Sheldon Klein, and Keren McConlogue. 1963. Indexing and dependency logic for answering English questions. *American Documentation* 15(3): 196-204.

Philip J. Stone, Robert F. Bales, J. Zvi Namenwirth, and Daniel Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science* 7(4): 484–498.

Ellen M. Voorhees. 1993. Using WordNet to disambiguate word senses for text retrieval. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.* https://doi.org/10.1145/160688.160715

William A. Woods. 1997. *Conceptual indexing: A better way to organize knowledge*. Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA, April. www.sun.com/research/techrep/1997/abstract61.html.

William A. Woods, Lawrence A. Bookman, Ann Houston, Robert J. Kuhns, Paul Martin, and Stephen Green. 2000. Linguistic knowledge can improve information retrieval. In *ANLC '00: Proceedings of the sixth conference on Applied natural language processing.* https://doi.org/10.3115/974147.974183