Semi-automatic Construction of Cross-period Thesaurus

Chaya Liebeskind, Ido Dagan, Jonathan Schler
Computer Science Department
Bar-Ilan University
Ramat-Gan, Israel
liebchaya@gmail.com, dagan@cs.biu.ac.il, schler@gmail.com

Abstract

Cross-period (diachronic) thesaurus construction aims to enable potential users to search for modern terms and obtain semantically related terms from earlier periods in history. This is a complex task not previously addressed computationally. In this paper we introduce a semi-automatic iterative Query Expansion (QE) scheme for supporting cross-period thesaurus construction. We demonstrate the empirical benefit of our scheme for a Jewish cross-period thesaurus and evaluate its impact on recall and on the effectiveness of lexicographer manual effort.

1 Introduction and Background

In the last decade, there is a growing interest in applying Natural Language Processing (NLP) methods to historical texts due to the increased availability of these texts in digital form (Sporleder, 2010; Sánchez-Marco et al., 2011; Piotrowski, 2012). The specific linguistic properties of historical texts, such as nonstandard orthography, grammar and abbreviations, pose special challenges for NLP. One of this challenges, which has not been addressed so far, is the problem of bridging the lexical gap between modern and ancient language.

In this paper, we address the interesting task of cross-period thesaurus (a.k.a. diachronic thesaurus) construction. A thesaurus usually contains thousands of entries, denoted here as *target terms*. Each entry includes a list of *related terms*, covering various semantic relations. A cross-period thesaurus aims to enable the potential user to search for a modern term and get related terms from earlier periods. Thus, in a cross-period thesaurus the target terms are modern while their related terms are ancient. In many cases, while the actual modern term (or its synynom) does not appear in earlier historical periods, different aspects of that term were mentioned. For example, in our Jewish historical corpora, the modern term *birth control*, has no equivalent ancient term, However, different contraceptive methods were described in our historical texts that are semantically similar to *birth control*. Thus, a related term is considered similar to the target term when it refers to the same concept.

The goal of our research is to support constructing a high-quality publishable thesaurus, as a cultural resource on its own, alongside being a useful tool for supporting searches in the domain. Since the precision of fully automatically-constructed thesauri is typically low (e.g. (Mihalcea et al., 2006)), we present a semi-automatic setting for supporting thesaurus construction by a domain expert lexicographer. Our recall-oriented setting assumes that manual effort is worthwhile for increasing recall as long as it is being utilized effectively.

Corpus-based thesaurus construction is an active research area (Curran and Moens, 2002; Kilgarriff, 2003; Rychlý and Kilgarriff, 2007; Liebeskind et al., 2012; Zohar et al., 2013). Typically, two statistical approaches for identifying semantic relatedness between words were investigated: first-order (co-occurrence-based) similarity and second-order (distributional) similarity (Lin, 1998; Gasperin et al., 2001; Weeds and Weir, 2003; Kotlerman et al., 2010). In this research, we focus on statistical measures of first-order similarity (see Section 2). These methods were found to be effective for thesaurus construction as standalone methods and as complementary to second-order methods (Peirsman et al., 2008). First-order measures assume that words that frequently occur together are topically related (Schütze and Pedersen, 1997). Thus, co-occurrence provides an appropriate approach to identify highly related terms for the thesaurus entries.

Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 29–35,
Sofia, Bulgaria, August 8 2013.

In general, there are two types of historically-relevant corpora: ancient corpora of ancient language, and modern corpora with references and mentions to ancient language (termed here *mixed corpora*). Since in our setting the thesaurus' target terms are modern terms, which do not appear in ancient corpora, co-occurrence methods would be directly applicable only over a mixed corpus. In a preliminary experiment, we applied the Liebeskind et al. (2012) algorithmic scheme, which applies first-order similarity and morphological aspects of corpus-based thesaurus construction, on a mixed corpus of our historical domain. We observed that the target terms had low frequency in this corpus. Since statistical co-occurrence measures have poor performance over low statistics, the experiment's results were not satisfactory. We therefore looked for ways to increase the number of documents in the statistical extraction process, and decided that applying query expansion (QE) techniques might be a viable solution.

We recognized two potential types of sources of lexical expansions for the target terms. The first is lexical resources available over the internet for extracting different types of semantic relations (Shnarch et al., 2009; Bollegala et al., 2011; Hashimoto et al., 2011). The second is lists of related terms extracted from a mixed corpus by a first-order co-occurrence measure. These lists contain both ancient and modern terms. Although only ancient terms will be included in the final thesaurus, modern terms can be utilized for QE to increase thesaurus coverage. Furthermore, expanding the target term with ancient related terms enables the use of ancient-only corpora for co-occurrence extraction.

Following these observations, we present an iterative interactive QE scheme for bootstrapping thesaurus construction. This approach is used to bridge the lexical gap between modern and ancient terminology by means of statistical co-occurrence approaches. We demonstrate the empirical advantage of our scheme over a cross-period Jewish domain and evaluate its impact on recall and on the effectiveness of the lexicographer manual effort.

The remainder of this paper is organized as follows: we start with a description of the statistical thesaurus construction method that we utilize in our scheme. Our main contribution of the iterative scheme is described in Section 3, followed by a case-study in Section 4 and evaluation and summary in Sections 5 and 6.

2 Automatic Thesaurus Construction

Automatic thesaurus construction focuses on the process of extracting a ranked list of candidate related terms (termed *candidate terms*) for each given target term. We assume that the top ranked candidates will be further examined (manually) by a lexicographer, who will select the eventual related terms for the thesaurus entry.

Statistical measures of first-order similarity (word co-occurrence), such as *Dice coefficient* (Smadja et al., 1996) and *Pointwise Mutual Information* (PMI) (Church and Hanks, 1990), were commonly used to extract ranked lists of candidate related terms. These measures consider the number of times in which each candidate term co-occurs with the target term, in the same document, relative to their total frequencies in the corpus.

In our setting, we construct a thesaurus for a morphologically rich language (Hebrew). Therefore, we followed the Liebeskind et al. (2012) algorithmic scheme designed for these cases, summarized below. First, our target term is represented in its lemma form. For each target term we retrieve all the corpus documents containing this given target term. Then, we define a set of candidate terms, which are represented in their surface form, that consists of all the terms in all these documents. Next, the Dice co-occurrence score between the target term and each of the candidates is calculated, based on their document-level statistics in the corpus. After sorting the terms based on their scores, the highest rated candidate terms are clustered into lemma-based clusters. Finally, we rank the clusters by summing the co-occurrence scores of their members and the highest rated clusters constitute the candidate terms for the given target term, to be presented to a domain expert.

3 Iterative Semi-automatic Scheme for Cross-period Thesaurus Construction

As explained in Section 1, our research focuses on a semi-automatic setting for supporting cross-period thesaurus construction by a lexicographer. In this work, we assume that a list of modern target terms is given as input. Then, we automatically extract a ranked list of candidate related terms for each target term using statistical measures, as detailed in Section 2. Notice that at this first step related terms can be extracted only from the mixed

corpora, in which the given (modern) target term may occur. Next, a lexicographer manually selects, from the top ranked candidates, *ancient* related terms for the thesaurus entry as well as terms for QE. The QE terms may be either ancient or modern terms from the candidate list, or terms from a lexical resource. Our iterative QE scheme iterates over the QE terms. In each iteration, a QE term replaces the target term's role in the statistics extraction process. Candidate related terms are extracted for the QE term and the lexicographer judges their relevancy with respect to the original target term. Notice that if the QE term is modern, only the mixed corpora can be utilized. However, if the QE term is ancient, the ancient corpora are also utilized and may contribute additional related terms.

The algorithmic scheme we developed for thesaurus construction is illustrated in Figure 1. Our input is a modern target term. First, we automatically extract candidates by statistical co-occurrence measures, as described in Section 2. Then, a domain-expert annotates the candidates.

The manual selection process includes two decisions on each candidate (either modern or ancient): (i) whether the candidate is related to the target term and should be included in its thesaurus entry, and (ii) whether this candidate can be used as a QE term for the original target term. The second decision provides input to the QE process, which triggers the subsequent iterations. Following the first decision we filter the modern terms and include only ancient ones in the actual thesaurus.

The classification of a candidate term as ancient or modern is done automatically by a simple classification rule: If a term appears in an ancient corpus, then it is necessarily an ancient term; otherwise, it is a modern term (notice that the converse is not true, since an ancient term might appear in modern documents).

In parallel to extracting candidate related terms from the corpus, we extract candidate terms also from our lexical resources, and the domain expert judges their fitness as well. Our iterative process is applied over the expansions list. In each iteration, we take out an expansion term and automatically extract related candidates for it. Then, the annotator selects both ancient related terms for the thesaurus and suitable terms, either modern or ancient, for the expansion list for further iterations.

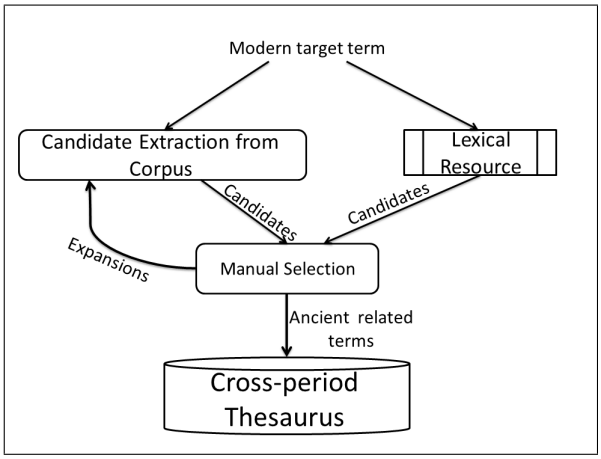

Figure 1: Semi-automatic Algorithmic Scheme

For efficiency, only new candidates that were not judged in pervious iterations are given for judgement. The stopping criterion is when there are no additional expansions in the expansions list.

Since the scheme is recall-oriented, the aim of the annotation process is to maximize the thesaurus coverage. In each iteration, the domain expert annotates the extracted ranked list of candidate terms until k sequential candidates were judged as irrelevant. This stopping criterion for each iteration controls the efforts to increase recall while maintaining a low, but reasonable precision.

In our setting, we extract ancient related terms for modern terms. Therefore, in order to utilize co-occurrence statistics extraction, our scheme requires both ancient and mixed corpora, where the first iteration utilizes only the mixed corpora. Then, our iterative scheme enables subsequent iterations to utilize the ancient corpora as well.

4 Case Study: Cross-period Jewish Thesaurus

Our research targets the construction of a cross-period thesaurus for the Responsa project[1]. The corpus includes questions on various daily issues posed to rabbis and their detailed rabbinic answers, collected over fourteen centuries, and was used for previous IR and NLP research (Choueka et al., 1971; Choueka et al., 1987; HaCohen-Kerner et al., 2008; Liebeskind et al., 2012; Zohar et al., 2013).

The Responsa corpus' documents are divided to four periods: the 11^{th} century until the end of the 15^{th} century, the 16^{th} century, the 17^{th} through the 19^{th} centuries, and the 20^{th} century until to-

[1]Corpus kindly provided: http://biu.ac.il/jh/Responsa/

day. We considered the first three periods as our ancient corpora along with the RaMBaM (Hebrew acronym for Rabbi Mosheh Ben Maimon) writings from the 12^{th} century. For the mixed corpus we used the corpus' documents from the last period, but due to relatively low volume of modern documents we enriched it with additional modern collections (Tchumin collection [2], ASSIA (a Journal of Jewish Ethics and Halacha), the Medical-Halachic Encyclopedia[3], a collection of questions and answers written by Rabbi Shaul Israeli[4], and the Talmudic Encyclopedia (a Hebrew language encyclopedia that summarizes halachic topics of the Talmud in alphabetical order). Hebrew Wiktionary was used as a lexical resource for synonyms.

For statistics extraction, we applied (Liebeskind et al., 2012) algorithmic scheme using Dice coefficient as our co-occurrence measure (see Section 2). Statistics were calculated over bigrams from corpora consisting of 81993 documents.

5 Evaluation

5.1 Evaluation Setting

We assessed our iterative algorithmic scheme by evaluating its ability to increase the thesaurus coverage, compared to a similar non-iterative co-occurrence-based thesaurus construction method. In our experiments, we assumed that it is worth spending the lexicographer's time as long as it is productive, thus, all the manual annotations were based on the lexicographer efforts to increase recall until reaching the stopping criterion.

We used Liebeskind et al. (2012) algorithmic scheme as our non-iterative baseline (*Baseline*). For comparison, we ran our iterative scheme, calculated the average number of judgments per target term (88) and set the baseline stopping criterion to be the same number of judgements per target. Thus, we ensured that the number of judgements for our iterative algorithm and for the baseline is equal, and thus coverage increase is due to a better use of lexicographer's effort. For completeness, we present the results of the non-iterative algorithm with the stopping criterion of the iterative algorithm, when reaching *k* (*k=10* was empirically selected in our case) sequential irrelevant candidates (*First-iteration*).

Method	RT	R	Pro	J
First-iteration	50	0.31	0.038	1307
Baseline	63	0.39	0.024	2640
Iterative	151	0.94	0.057	2640

Table 1: Results Comparison

To evaluate our scheme's performance, we used several measures: total number of ancient related terms extracted (RT), relative recall (R) and productivity (Pro). Since we do not have any predefined thesaurus, our micro-averaged relative-recall considered the number of ancient related terms from the output of both methods (baseline and iterative) as the full set of related terms. Productivity was measured by dividing the total number of ancient related terms extracted (RT) by the total number of the judgments performed for the method (J).

5.2 Results

Table 1 compares the performance of our semi-automatic iterative scheme with that of the baseline over a test set of 30 modern target terms. Our iterative scheme increases the average number of extracted related terms from 2.1 to 5, i.e., increasing recall by 240%. The relative recall of the first-iteration (0.31) is included in the relative recall of both the baseline and our iterative method. Iterating over the first iteration increases recall by 300% (from 50 to 151 terms), while adding more judgements to the non-iterative method increases recall only by 26% (to 63 terms). The productivity of the iterative process is higher even than the productivity of the first iteration, showing that the iterative process optimizes the lexicographer's manual effort.

Table 2 shows examples of thesaurus target terms and their ancient related terms, which were added by our iterative scheme[5]. Since the related terms are ancient Halachic terms, we explain them rather than translate them to English.

We further analyze our scheme by comparing the use of ancient versus modern terms in the iterative process. Although modern related terms were not included in our cross-period thesaurus, in the judgement process the lexicographer judged their

[2]http://www.zomet.org.il/?CategoryID=170

[3]http://medethics.org.il/website/index.php/en/research-2/encyclopedia-of-jewish-medical-ethics

[4]http://www.eretzhemdah.org/data/uploadedfiles/ebooks/14-sfile.pdf

[5]To facilitate readability we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexico-graphic order, are abgdhwzxTiklmns`pcqršt.

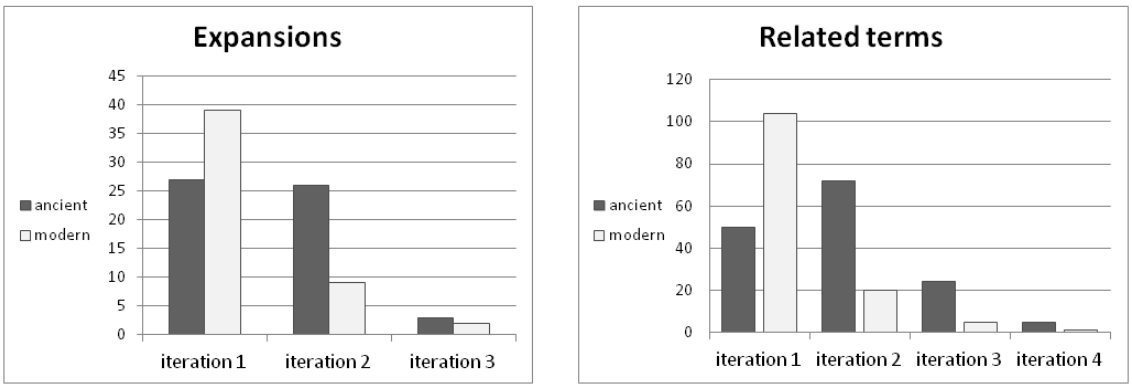

Figure 2: The extraction of ancient terms versus modern terms in the iterative process

Target term	**Related term**
zkwiwt iwcrim (copyright)	*hsgt gbwl* (trespassing)
	iwrd lamnwt xbrw ([competitively] enter his friend's profession)
	`ni hmhpk bxrrh (a poor man is deciding whether to buy a cake and another person comes and takes it)
hmtt xsd (euthanasia)	*rwb gwssin lmith* (most dying people die)
	xii š`h (living for the moment)
hpsqt hriwn (abortion)	*xwtkin h`wbr* (killing the fetus)
	hwrg npš (killing a person)
	rwdp (pursuer, a fetus endangering its mother's life)
tiknwn hmšpxh (birth control)	*šlwš nšim mšmšwt bmwk* (three types of women allowed to use cotton diaphragm)
	dš mbpnim wzwrh mbxwc (withdrawal method)
hprt xwzh (breach of contract)	*biTwl mqx* (cancelling a purchase)
	dina dgrmi (indirect damage)
	mqx t`wt (erroneous bargain)
srwb pqwdh (insubordination)	*mwrd bmlkwt* (rebel against the sovereign [government])
	imrh at pik (to disobey)
	Avner and *khni Nob* (a biblical story: king Saul ordered to slay Ahimilech together with 85 priests. Avner, the captain of Saul's guard, disobeyed the order.)

Table 2: Examples for the iterative scheme's contribution

relevancy too. In Figure 2, we report the number of modern related terms in comparison to the number of ancient related terms for each iteration. In parallel, we illustrate the number of ancient expansions in proportion to the number of modern expansions. The x-axis' values denote the iterations, while the y-axis' values denote the number of expansions and related terms respectively. For each iteration, the expansions chart presents the expansions that were extracted while the related terms chart presents the extracted related terms, of which the ancient ones were included in the thesaurus. Since the input for our scheme is a modern target terms, the first iteration extracted more modern related terms than ancient terms and utilized more modern expansions than ancient. However, this proportion changed in the second iteration, probably thanks to the ancient expansions retrieved in the first iteration.

Although there are often mixed results on the effectiveness of QE for information retrieval (Voorhees, 1994; Xu and Croft, 1996), our results show that QE for thesaurus construction in an iterative interactive setting is beneficial for increasing thesaurus' coverage substantially.

6 Conclusions and Future Work

We introduced an iterative interactive scheme for cross-period thesaurus construction, utilizing QE techniques. Our semi-automatic algorithm significantly increased thesaurus coverage, while optimizing the lexicographer manual effort. The scheme was investigated for Hebrew, but can be generically applied for other languages.

We plan to further explore the suggested scheme by utilizing additional lexical resources and QE algorithms. We also plan to adopt second-order distributional similarity methods for cross-period thesaurus construction.

References

D. Bollegala, Y. Matsuo, and M. Ishizuka. 2011. A web search engine-based approach to measure semantic similarity between words. *Knowledge and Data Engineering, IEEE Transactions on*, 23(7):977–990.

Yaacov Choueka, M. Cohen, J. Dueck, Aviezri S. Fraenkel, and M. Slae. 1971. Full text document retrieval: Hebrew legal texts. In *SIGIR*, pages 61–79.

Yaacov Choueka, Aviezri S. Fraenkel, Shmuel T. Klein, and E. Segal. 1987. Improved techniques for processing queries in full-text systems. In *SIGIR*, pages 306–315.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition - Volume 9*, ULA '02, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Caroline Gasperin, Pablo Gamallo, Alexandre Agustini, Gabriel Lopes, Vera De Lima, et al. 2001. Using syntactic contexts for measuring word similarity. In *Workshop on Knowledge Acquisition and Categorization, ESSLLI*, Helsinki, Finland.

Yaakov HaCohen-Kerner, Ariel Kass, and Ariel Peretz. 2008. Combined one sense disambiguation of abbreviations. *Proceedings of ACL08: HLT, Short Papers*, pages 61–64.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Junichi Kazama, and Sadao Kurohashi. 2011. Extracting paraphrases from definition sentences on the web. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1087–1097.

Adam Kilgarriff. 2003. Thesauruses for natural language processing. In *Proceedings of the Joint Conference on Natural Language Processing and Knowledge Engineering*, pages 5–13, Beijing, China.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-geffet. 2010. Directional distributional similarity for lexical inference. *Nat. Lang. Eng.*, 16(4):359–389, October.

Chaya Liebeskind, Ido Dagan, and Jonathan Schler. 2012. Statistical thesaurus construction for a morphologically rich language. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 59–64, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st national conference on Artificial intelligence - Volume 1*, AAAI'06, pages 775–780. AAAI Press.

Yves Peirsman, Kris Heylen, and Dirk Speelman. 2008. Putting things in order. First and second order context models for the calculation of semantic similarity. In *9es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008)*. Lyon, France.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157.

Pavel Rychlý and Adam Kilgarriff. 2007. An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 41–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cristina Sánchez-Marco, Gemma Boleda, and Lluís Padró. 2011. Extending the tool, or how to annotate historical language varieties. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hinrich Schütze and Jan O. Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Inf. Process. Manage.*, 33(3):307–318, May.

Eyal Shnarch, Libby Barak, and Ido Dagan. 2009. Extracting lexical reference rules from wikipedia. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 450–458, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Comput. Linguist.*, 22(1):1–38, March.

Caroline Sporleder. 2010. Natural language processing for cultural heritage domains. *Language and Linguistics Compass*, 4(9):750–768.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 81–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11, New York, NY, USA. ACM.

Hadas Zohar, Chaya Liebeskind, Jonathan Schler, and Ido Dagan. 2013. Automatic thesaurus construction for cross generation corpus. *Journal on Computing and Cultural Heritage (JOCCH)*, 6(1):4:1–4:19, April.