Creating a Web-based Lexical Corpus and Information-extraction Tools for the Semitic Language Maltese*

La creación de un corpus léxico basado en Internet y herramientas para la extracción de información léxica para la lengua semítica maltés

Jerid Francom
Wake Forest University
323 Greene Hall, PO Box 7566
Winston-Salem, NC 27109 (USA)
francojc@wfu.edu

Dainon Woudstra, Adam Ussishkin
University of Arizona
Douglass Building, Room 200E
Tucson, AZ 85721 (USA)
{dainon, ussishki}@u.arizona.edu

Abstract: In this paper, we document the creation of a web-based lexical corpus and a set of lexical tools for the less-resourced Semitic language Maltese. The benefits and shortcomings of using the web as a source of textual information are discussed in addition to the practical steps taken to develop and evaluate the resulting resources. We believe this preliminary work sets the groundwork for further development of Semitic language resources, as well as for less-resourced languages in general, and contributes to a growing interest to apply corpus data in theoretical analysis.
Keywords: Semitic languages, Maltese, Web-based corpus, Lexical corpus, minority languages, theoretical linguistics

1 Introduction

For many languages other than English, the quality and quantity of available resources is quite limited. To address this gap, many researchers have recently focused their efforts on documenting and creating database resources for these less-studied languages (McEnery, Xiao, and Tono, 2006). In addition to the clear role that electronic text resources play in computational applications, corpora are increasingly playing a larger role in the testing and development of linguistic theories, and a wide range of languages is crucial to the success and applicability of these theories.

In what follows, we document the creation of a Maltese lexical corpus developed in order to further psycholinguistic research on the mental organization of Semitic lexicons. The primary goal of this project is to create a sizable lexical corpus and a set of lexical calculation tools capable of producing a filtered set of lexical items for psycholinguistic experimentation. The sources for this effort include text extracted from the web and collaborative efforts from other scholars working in the area. We discuss both theoretical and practical issues in creating corpora in general and this corpus more specifically, elaborate the steps taken to bring this project to fruition and report the statistical results of our efforts. The research reported in this paper contributes to corpus linguistics as well to other areas of linguistics, including formal approaches to language investigation and psycholinguistics.

2 Web as corpus

A corpus can be thought of as a collection of texts. Traditionally, collections of texts have been amassed from prose found in print (Francis, 1975; Johansson, 1982; Sinclair, 1987), though the web has become an increasingly popular source for corpus creation due to several factors. The web contains a large amount of text already in electronic form, obviating tedious and time-consuming processes for converting print to electronic media. The web also provides access to a wide range of languages, language varieties, and genres that may be difficult to acquire in print.

The web as a data source for linguistic corpus creation is not without its theoretical and practical pitfalls. One consequence of sampling is that the data may not be as representative of the body from which it is

* We gratefully acknowledge funding from the United States National Science Foundation (BCS-0715500) to Adam Ussishkin.

extracted as assumed both in terms of balance and sparseness. Thus, the corpus linguist must make pragmatic decisions based on the ultimate purpose of the corpus, and evaluate the degree to which the data sufficiently fills this function.

Another concern is that not all languages are equally represented on the web. Evidence from various web experiments show that a handful of languages dominate 90% of the web, most notably English with over 70% (Xu, 2000; Kilgarriff and Grefenstette, 2003). This exacerbates the sparseness problem for languages with less content on the web.

The web also presents a level of variability not typical for print sources. Languages where non-standard web characters are lexically contrastive raises the possibility of conflation and misrepresentation in token counts. Finally, text found on the web is inherently more variable than text in print. Rigorous publishing standards for print are not always adhered to on the web, resulting in a larger number of typographical errors (Ringlstetter, Schulz, and Mihov, 2006).

In sum, there are a number of practical advantages that motivate data extraction from the web along with concerns that must be addressed in web-based corpora construction in order to ensure its theoretical integrity.

3 Development of the corpus

In what follows, we describe the development of a web-based lexical corpus for Maltese. This corpus represents two stages: 1) bottom-up construction of Maltese data including seed selection, web extraction, data filtering and data indexing and 2) a collaborative effort with a Maltese computational linguist providing pre-filtered text prepared for tokenizing and data indexing.

3.1 Seed selection and web extraction

The first step in creating a web-based corpus is to select the sources that will serve as the seeds for web extraction. A preselection list of URLs was obtained through a Google search. The most prevalent sources were newspapers and blogs. As discussed in the previous section, there are a number of criteria that need to be negotiated in developing a corpus from the web. From the standpoint of this project and general considerations of size, non-target language contamination, proofreading standards and corpus balance, a decision was made to pursue the online newspaper sources as our primary extraction seeds. [1]

Another aspect critical to the validity of a Maltese lexical corpus is character encoding. Of the potential sources, not all of them fully encoded all Maltese characters, including Maltese-specific characters (ċ, ġ, ħ, ż). Given these considerations, this project produced three seed candidates for extraction: Illum (`http://www.illum.com.mt/`), L-Orizzont (`http://www.l-orizzont.com/`) and Malta Right Now (`http://www.maltarightnow.com/`).

Extraction from the web often employs one of three main approaches. 1) Web-based searches through popular search engines, 2) more advanced search-engine based extraction via API[2] interfaces and 3) independent web crawling.

Both web-based searches and API searches through search engines have inherent drawbacks. First, search-engine based extraction is limited in terms of the number of queries one can run and the type of queries that can be performed. This highlights the 'brittle' aspect of depending on commercial parties to provide academically relevant services. Maybe more importantly, search engines do not freely disclose the sources of their databases and indexing lists. Therefore there is no way to determine the relevance of the sample to the population of interest.

Given these shortcomings, the current project selected to perform a web-crawl on the selected seed URLs. This approach avoids the natural restrictions placed on search-engine based extraction as all aspects of the process are independently maintained including extraction, documentation and interactive searching. After investigating a number of web-crawling tools including Heritrix (Mohr et al., 2004), WIRE (Castillo and Baeza-Yates, 2005) and Nutch (Khare et al., 2004), we opted to employ the open-source UNIX utility Wget. This decision reflected a desire to develop feature-rich strategies that can be easily obtained, configured

[1] See (Ghani, Jones, and Mladenic, 2005) for methodologies for extracting minority languages more generally from the web.

[2] Application Program Interface: Google provides a set of tools for building software applications, which interface with search engine queries.

and deployed by other scholars without having to compromise adherence to web protocol standards and best practices such as respecting robots.txt[3] and easing remote host server load.

Our particular extraction efforts with Wget included a number of syntax flags illustrated in figure 1.

Figure 1: WGET Syntax

```
wget -r -w3 -U LabBot=host.address.edu/
-A htm, html, asp, php, cfm, shtml
http://www.sitename.com.mt/ -o log
--output-document=outputfile.htm
```

Through this implementation we recursively crawled the sites identified previously (-r). The server was hit on an interval of once every three seconds (-w3) and each time the crawler reported the name of our bot and our web address. Our efforts and intentions were documented here with expressing our full compliance to stop and destroy data on request, an effort to respect best practices. The crawler was specified to only extract pages that conformed to a pre-screening of page extensions found to contain viable prose in text readable form (-A). Finally, all activities were recorded in a log file and the output appended to a standard web-content file (output-document=outputfile.htm) stored offline for processing.

The results of this crawl produced two key file types, output files and log files. Output files contained raw source code from those pages found inside the root directory of each of the target seeds. The log files contained connection information, URL name, file size downloaded and connection status. In the case of two of the three site seeds used in this corpus construction, the URL also contained the publishing date of the article or piece downloaded. This conveniently provided date range information for archived articles.

3.2 Data filtering, indexing and results

An attempt to remove non-target textual information including links, titles to articles and other redundant site text was conducted. This 'boilerplate' information contributes to 'noisy' or 'dirty' data that skews and obscures relevant corpus frequency data. A 'wrapper' approach to collecting relevant data can be used to extract text which occurs between certain open and close HTML tags (i.e., $<$ *body* $>, <$ *div* $>, <$ *span* $>$, etc.). Some of these tags may include specific CLASS or ID values, which may be useful for isolating sections of the HTML code for extraction. JavaScript and HTML comments, which occur sporadically between open and close tags, required additional filtering.

Several strategies to combat this issue were considered and tested. Machine learning techniques, such as CFR++,[4] used to recognize and tag webpage content, may be used to filter data. The most effective for this project, given reduced number of sites to filter, was to create the above mentioned 'wrapper' effectively identifying unique tags surrounding relevant text. The source code under scrutiny, for both news sources, showed consistent CSS tagging for ID and CLASS descriptors. This strategy provided robust exclusion of irrelevant data, but unfortunately did not exclude all irrelevant non-target text. Regular expressions were used to minimize these types of text as a final processing adjustment.

The strategy of shingling (Gibson, Wellner, and Lubar, 2008), which computes the resemblance between two documents. Shingling would effectively reduce the chances of processing near duplicate or identical texts. The extracted text does contain words from non-target languages which was not filtered out of the final data. The application of text categorization tool (TextCat (Cavnar and Trenkle, 1994), for example) would potentially reduce the amount of foreign text. Our original implementation does not include these technique/tools, but it may provide useful in future improvements of this data.

The tokenization process included the data obtained from the web-crawling procedures described in addition to a sizable set of data from Dr. Albert Gatt.[5] To tokenize both the webpage content and Gatt corpora, words were split according to morphologi-

[3] For more information about web ethics and web crawling see (Eichmann, 1995)

[4] `http://crfpp.sourceforget.net/`

[5] This supplementary data included text from Kulħadd, Leħen is-Sewwa, Il-Mument and In-Nazzjon online newspapers. Of note, Kulħadd, Leħen is-Sewwa and In-Nazzjon data were extracted from non-overlapping date ranges to the web-crawl described here.

cal boundary characters such as the apostrophe and any non-Maltese alphabet characters such as hyphens, slashes, among a few others. This converted any complex words into single token strings. All special characters and numbers were removed from the corpora. Multiple white space characters were minimized to a single space and punctuation was removed. The resulting long string data was split into an array at each remaining white space character. The resulting array was reprocessed into a hash table with corresponding counts for each token.

The database structure was designed according to a simple set of criteria. First, a token column is needed to textually represent each unique token, second a total count column and, finally one column of the token counts for each corpus processed. The database was designed to keep data collected from different corpora separate. For instance, if the user wanted to query only frequencies which occur in In-Nazzjon, its respective count information must be retained separately.

The two sources of data results in 3,323,325 total tokens, of which 53,396 are unique. Web-crawling produced 58.9%[6] of the total database and 40.2% from the corpora provided by Gatt. The largest percent of unique words was found in In-Nazzjon and the smallest percent from Leħen is-Sewwa (In-Nazzjon = 79.1%, Malta Right Now = 31.3%, Kulħadd = 15.3%, L-Orizzont = 9.3%, Leħen is-Sewwa = 8.6%). The following corpus counts are illustrated by (total count—unique count): Malta Right Now (1,927,598—8,165), In-Nazzjon (1,240,923—42,240), Kulħadd (69,908—8,165), L-Orizzont (60,982—4,944), Leħen is-Sewwa (24,914—4,577).

4 Corpus interface

In this section we describe the creation of a web interface to provide access to the lexical corpus.[7]. One goal was to provide international, cross-platform and requirement-free access to this collection. Thus, the interface was implemented using PHP,[8] an open-source technology with robust compatibility with data-driven websites that does not require any special software on the user end.[9] Another goal was to construct an intuitive graphic interface that provided encoding neutral interactive queries and facilitated seamless comparisons between various lexical calculations. In what follows we describe in detail these aspects of the user interface.

4.1 Basic tools

At the most basic level the interface includes detailed documentation concerning the sources, citations, counts and date ranges of the collection. This information can be found outside the registered-users area of the site in order to provide a good-faith effort to document our efforts.

Once inside the registered-users area the user is presented with a language selector and corresponding virtual keyboard (Figure 2). The design of this interface and the database underlying the web portal is designed to be extensible. In this way, any future plans to incorporate other languages can proceed without major modification to the interface.[10] On selecting a language, the relevant database token counts are presented. These counts are dynamically updated as the database is refreshed with new content.

The interface is also composed of a general interactive search field and a set of lexical calculation tools. The search field supports full use of POSIX regular expressions and is supported by a graphical web keyboard layout. The keyboard layout serves as an encoding-neutral input source that will allow users to query the Maltese and other language corpora with appropriate graphemes that may not happen to be installed on the local machine. In addition, the keyboard includes a number of characters employed in regular expression syntax.

The search fields maintain previous query strings, and thus provide access to quick comparisons between the three calculators: a lexical frequency calculator, a lexical uniqueness point calculator, and a neighborhood density calculator. These tools provide the user with

[6]The Illum data is not included in the database to date.

[7]This site can be found at `http://dingo.sbs.arizona.edu/~psycol/resources/` Registration is required.

[8]PHP is server-side HTML embedded scripting language for Hypertext Preprocessing.

[9]One exception is that users must have cookies enabled on the client machine in order to interact fully with the database.

[10]Currently access to a Hebrew corpus and Khalkha Mongolian corpus is available. Hebrew: 60,052,261 tokens; Mongolian: 259,264 tokens.

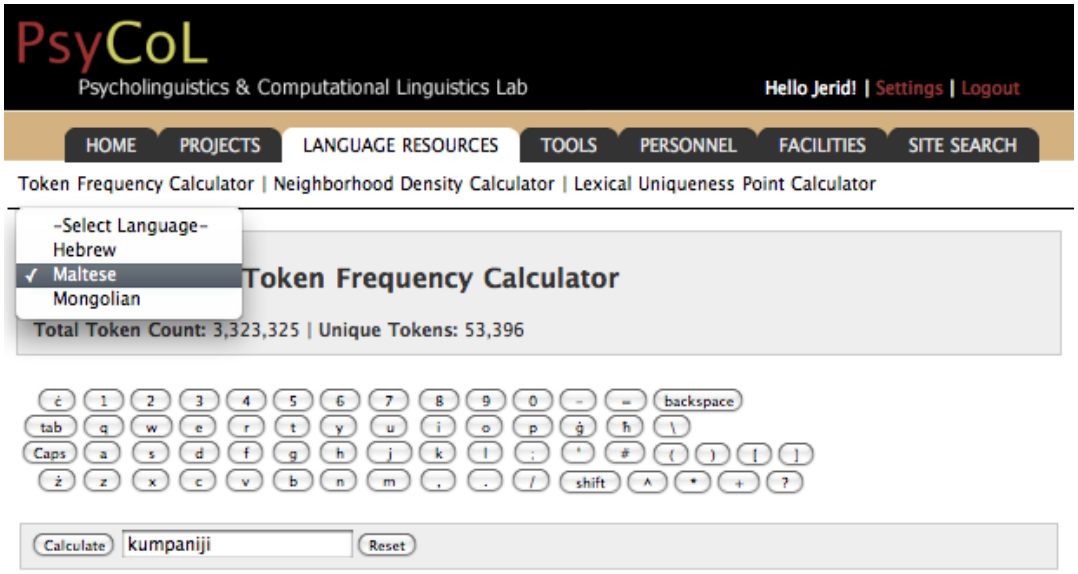

Figure 2: Basic web interface with language selector, query window and virtual keyboard

the opportunity to retrieve information useful for lexical statistics, item selection for experiments, and other uses.

4.2 Lexical frequency

Using POSIX regular expressions the lexical frequency calculator returns the total token count and number of unique tokens. The number of queried token counts divided by the total number of token counts in the corpus $\frac{c}{N}$ is known as lexical frequency. The returned query displays all tokens that match the regular expression with their respective total count and individual corpus counts. The return values for each token includes its counts per million and natural log frequency (Figure 3).

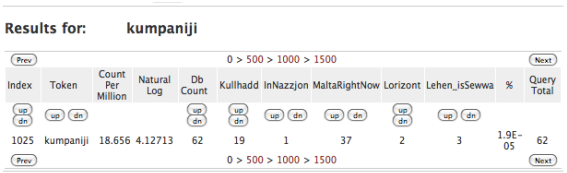

Figure 3: Token frequency sample output.

4.3 Lexical uniqueness point

The point at which a set of graphemes is no longer a subset of some other set of graphemes is known as a lexical uniqueness point. The uniqueness point can be indexed from left to right, or right to left. Marslen-Wilson's Cohort model (1978) suggests lexical processing makes optimal use of the lexicon during real-time use. Subsequent studies (Wurm, 2007) have shown reaction time effects corresponding to the Cohort model. Both calculations are included in the current web interface. This lexical tool queries the database for the desired string, which may be a word of the language but need not be. This string is then compared to each and every entry that contains it. If the string is not unique, the number and list of overlapped words is returned. If the string is unique, the point at which the input no longer overlaps is highlighted and two indices are provided (Figure 4). Not all unique strings are words, and this fact is also reported.

Figure 4: Lexical uniqueness sample output.

4.4 Neighborhood density

Research has shown that lexical access is sensitive to the number of lexical neighbors a given target has (Goldinger, Luce, and Pisoni, 1989; Cluff and Luce, 1990; Luce and Pisoni, 1998). The neighborhood density tool was designed to provide users with the ability to retrieve the number of neighbors of a given query string. The corpus density calculations are computed per query. The initial method of preprocessing this information was

rejected due to repopulation of data. As more data is added to the site, the neighborhood density would require reprocessing. To make this calculation on-the-fly, a language specific alphabet generates all combinations of token neighbors based on individual characters: insert a letter, remove a letter and replace a letter (i.e., a Hamming distance of 1). Each resulting possible neighbor is added to the query string. The resulting output contains the database counts for each neighboring token and the corpus density (weight) of these counts (Figure 5).

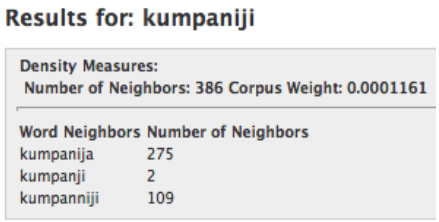

Figure 5: Neighborhood density sample output.

5 *Corpus evaluation*

The Maltese corpus described here shows both strengths and limitations as a language resource. First, the size of this collection represents quite a large sampling of the language, though "large" is a relative term. The size of electronic corpus data for English is quite humbling (i.e., Web 1T 5-gram Version 1, 1 trillion word tokens; North American News Text Corpus, 350 million word tokens; English Corpus Concordance, 18 million). To our knowledge, our Maltese corpus constitutes the largest lexical corpus in existence for Maltese.[11]

Another strength found in these resources is their general accessibility. Users can perform a number of calculations and queries without concern for special client-side software, and there is no need for local data storage. Finally, the corpora and lexical tools developed here were constructed with extensibility in mind. The underlying database structure is abstract enough that any language conforming to the encoding and database design can be accessed by the lexical tools with minor modification.

[11] Work reported by Kevin Scannell `http://borel.slu.edu/crubadan/` constitutes another large Maltese data set at 518,275 words.

There are potential limitations to our work that highlight crucial issues in corpus linguistics. First, all corpus creation must deal with the inadvertent inclusion of 'dirty' or 'noisy' data. These data can arise from various sources including author error in the case of misspelled or mistyped words from the original source and faulty filtering strategies attributable to the corpus designers in the form of web programming artifacts HTML, JavaScript, CSS, etc. We are actively engaged in improving our filtering process in order to address this issue.

Another limitation inherent in corpus development concerns the representativeness of word frequencies. The text extracted from the web is a pseudo-random sample of the target language, which provides a certain level of assurance that the corpus will contain a 'natural' balance of the language and the frequency of the words therein. It is important to bear in mind, however, that this limitation extends beyond the work presented here and must be assumed more generally as a part of all corpus development and analysis.

There are several limitations of concern that hold more specifically for our current enterprise and which constitute continuing areas of work. First, our efforts to design a maximally extensible resource are limited when data is collected via the web due to the idiosyncratic nature of web programming and coding practices. The wrapper approach adopted to filtering the raw web data retrieved cannot be performed automatically. Given the limited number of seeds (2) in our first web-crawl a machine-learning approach was not a requirement. As our project grows, however, we hope to explore more automatic strategies for teasing apart target data from extraneous web noise including wrapper (Prasad and Paepcke, 2008) and machine-learning techniques (Baroni and Ueyama, 2006; Spousta, Marek, and Pecina, 2008).

A second shortcoming is more specific to our primary goal of retrieving a filtered set of lexical items for psycholinguistic experimentation. Currently queries cannot be processed in batch. The ability to obtain lexical calculations for a set of items in one process will facilitate operations with the system. Plans to enable more robust queries and batch output is earmarked as a next step in the development of this set of language re-

source tools.

6 Conclusion

We have documented the creation of corpora and corresponding lexical tools for the Semitic language Maltese. These efforts coincide with a growing interest in developing corpus resources for less-resourced languages. In this project we highlighted the benefits and shortcomings of using the web as a source of textual information and pointed to emerging methods that may facilitate data extraction in future work. We believe this preliminary work sets the groundwork for further development and contributes to a growing interest to apply corpus data in theoretical analysis.

References

[Baroni and Ueyama2006] Baroni, M. and M. Ueyama. 2006. Building general-and special-purpose corpora by Web crawling. In *Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application*, pages 31–40.

[Castillo and Baeza-Yates2005] Castillo, Carlos and Ricardo Baeza-Yates. 2005. Wire: an open-source web information retrieval environment. In *Workshop on Open Source Web Information Retrieval (OSWIR)*.

[Cavnar and Trenkle1994] Cavnar, W.B. and J.M. Trenkle. 1994. N-gram-based text categorization. *Ann Arbor MI*, 48113:4001.

[Cluff and Luce1990] Cluff, M.S. and P.A. Luce. 1990. Similarity neighborhoods of spoken two syllable words: Retroactive effects on multiple activation. *The Journal of the Acoustical Society of America*, 87:S125.

[Eichmann1995] Eichmann, D. 1995. Ethical web agents. *Computer Networks and ISDN Systems*, 28(1-2):127–136.

[Francis1975] Francis, WN. 1975. Problems of Assembling, Describing, and Computerizing Corpora. *Research Techniques and Prospects. Papers in Southwest English*, (1).

[Ghani, Jones, and Mladenic2005] Ghani, R., R. Jones, and D. Mladenic. 2005. Building Minority Language Corpora by Learning to Generate Web Search Queries. *Knowledge and Information Systems*, 7(1):56–83.

[Gibson, Wellner, and Lubar2008] Gibson, J., B. Wellner, and S. Lubar. 2008. Identification of Duplicate News Stories in Web Pages. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

[Goldinger, Luce, and Pisoni1989] Goldinger, S.D., P.A. Luce, and D.B. Pisoni. 1989. Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28(5):501–518.

[Johansson1982] Johansson, S., editor. 1982. *Computer corpora in English language research*. Bergen: NAVF.

[Khare et al.2004] Khare, R., D. Cutting, K. Sitaker, and A. Rifkin. 2004. Nutch: A flexible and scalable open-source web search engine. *Oregon State University*.

[Kilgarriff and Grefenstette2003] Kilgarriff, A. and G. Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3).

[Luce and Pisoni1998] Luce, P.A. and D.B. Pisoni. 1998. Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, 19(1):1.

[Marslen-Wilson and Welsh1978] Marslen-Wilson, W.D. and A. Welsh. 1978. Processing Interactions and Lexical Access during Word Recognition in Continuous Speech. *Cognitive Psychology*, 10(1):29–63.

[McEnery, Xiao, and Tono2006] McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based language studies: an advanced resource book*. Routledge.

[Mohr et al.2004] Mohr, G., M. Kimpton, M. Stack, and I. Ranitovic. 2004. Introduction to heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04), Bath, UK*.

[Prasad and Paepcke2008] Prasad, J. and A. Paepcke. 2008. Coreex: content extraction from online news articles. In

Proceeding of the 17th ACM conference on Information and knowledge management. ACM New York, NY, USA.

[Ringlstetter, Schulz, and Mihov2006] Ringlstetter, C., K.U. Schulz, and S. Mihov. 2006. Orthographic errors in web pages: Toward cleaner web corpora. *Computational Linguistics*, 32(3):295–340.

[Sinclair1987] Sinclair, J.M. 1987. *Looking up: an account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. Collins ELT.

[Spousta, Marek, and Pecina2008] Spousta, M., M. Marek, and P. Pecina. 2008. Victor: the Web-Page Cleaning Tool. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 12–17.

[Wurm2007] Wurm, L.H. 2007. Semantic processing in auditory lexical decision: Ear-of-presentation and sex differences. *Cognition & Emotion*, 99999(1):1–26.

[Xu2000] Xu, Jack. 2000. Multilingual search on the World Wide Web. In *Proceedings of the Hawaii International Conference on System Sciences HICSS*, volume 33.