Dicionário-Aberto: A Source of Resources for the Portuguese Language Processing

Alberto Simões[1], Álvaro Iriarte Sanromán[1], and José João Almeida[2]

[1] Center for Humanistic Studies, Minho's University
{ambs,alvaro}@ilch.uminho.pt
[2] Computer Science and Technology Center, Minho's University
jj@di.uminho.pt

Abstract. In this paper we describe how Dicionário-Aberto, an online dictionary for the Portuguese language, is being used as the base to construct diverse resources that are relevant in the processing of the Portuguese language.

We will briefly present its history, explaining how we got here. Then, we will describe the resources already available to download and use, followed by the discussion on the resources that are being currently developed.

Keywords: dictionary, lexicography, open resource, thesauri.

1 A Brief History

Dicionário-Aberto[1] started in June 2005, when a few people felt that the Portuguese language was missing an open dictionary for use (any use). The process of creating a dictionary from scratch is difficult and expensive. When most of the interested persons are engineers and computer scientists, this task gets more difficult. As one member of this group was responsible for the transcription of Portuguese books for the Project Gutenberg [1], the idea of transcribing a full dictionary appeared. An old dictionary with expired copyright was searched and chosen[2], digitalized and the transcription process started using the Project Gutenberg Distributed Proofreaders web interface. A detailed description of this process is described in [8].

The transcription was performed by volunteers, using a simple textual syntax, very similar to a subset of common wiki syntaxes. In March, 2010, the full transcription was concluded (different validation rounds were performed for each page). This textual document was converted to a more formal syntax, based on XML.

[1] Available at http://www.dicionario-aberto.net/

[2] The chosen dictionary was "Novo Diccionário da Língua Portuguesa, Cândido de Figueiredo, 1913." It was not chosen by its lexicographic quality, but only because of a set of circumstantial facts.

H. Caseli et al. (Eds.): PROPOR 2012, LNAI 7243, pp. 121–127, 2012.

A subset of the TEI [3] (Text Encoding Initiative) format for dictionaries was chosen. Simões and Almeida [9] describe this conversion process.

The chosen dictionary used an old orthographic form (prior to 1943/45 agreements). To be part of Project Gutenberg the books must be transcribed in original form, so the transcribed documents needed orthographic modernization to be useful. This task was automated and at the present moment, a set of volunteers are approving the modernized entries[3] Dicionário-Aberto has now 128 521 entries, and about 8% of the entries were verified for modernization errors.

The new orthographic agreement (1990) will require a new modernization process. Fortunately, this will be easier to automate, as there are a couple of good conversion tools available [2].

In the future, Dicionário-Aberto will be open as a dictionary Wiki, where the community can edit corrections or add new words. To guarantee quality (and a somewhat controlled language) a two tier process will be implemented: a change or addition (or even deletion) will be available right away, but in a "non official" status, until a moderator approves the change.

2 Currently Available Resources

With the current status of Dicionário-Aberto as described in the previous section, there are some resources that can be downloaded and used to learn about the Portuguese language, its history and to process automatically using natural language processing techniques.

Original TXT and TEI Transcriptions

The more basic resources are the plain text files with the original transcriptions, both in wiki or TEI formats. These resources are available in 28 separate files, one for each letter, plus a geographic and an onomastic appendixes. These documents (specially the TEI version as it is annotated and is easier to process automatically) are mostly useful to study the Portuguese language before the 1943/45 orthographic agreement.

Current Database Snapshot

A view of the Dicionário-Aberto web-site database is available to download and use. It is an SQL document that can be imported in any MySQL database server (and probably in other tools with minor changes). Figure 1 shows the structure of this view (further tables might be available in the future, accordingly with new resources being developed).

This is the better way to use the current database in offline mode, as it enables the user to access the current versions for all dictionary entries, as well as the previous versions (before orthographic modernization). Therefore this format can be used not just to perform text-mining in the dictionary but also

[3] Unfortunately the modernization process had some false positive substitutions.

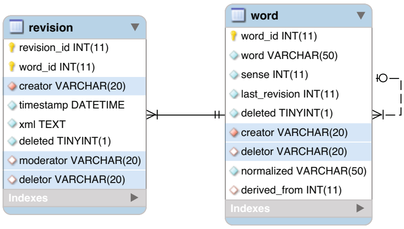

Fig. 1. Dicionário-Aberto database view

for contrastive studies. The database is regularly checked for quality control, checking the database for consistency and validating all XML snippets. Another test, that validates the dictionary completeness (that all cross-references have a valid target) is currently disabled given the modernization process.

Modernization Rules

With the modernization process we obtained two versions of the dictionary in two different versions of Portuguese. Although they are similar in great extend, they can be considered two different languages, and therefore all typical approaches to align and extract bilingual dictionaries automatically can be applied.

Tables 1 present two types of orthographic modernization rules that can be extracted (these lists were constructed using `lexdiff` [2]). The first one maps full words and the second maps letter sequences. The latter is easier to apply in documents with words that are not in the dictionary (but that share a subsequence of letters with some other word), but the first is better for accuracy. The third column is the rule confidence about its transformation.

Table 1. Word and pattern rules for orthographic modernization

Full-word Rules		
gênero	género	100%
aquelle	aquele	100%
effeito	efeito	100%
fórma	forma	100%
pessôa	pessoa	100%
camillo	camilo	100%
póde	pode	99%
sôbre	sobre	98%
ás	às	90%

Pattern Rules		
sôa	soa	99%
ffe	fe	95%
ph	f	95%
lle	le	90%
llo	lo	88%
aes	ais	87%
gên	gén	87%
bôa	boa	80%
cci	ci	40%

Note, however, that these dictionaries were calculated with the current version of Dicionário-Aberto, where some hundred words were not modernized correctly by the automatic process[4].

[4] Being a dictionary, it includes very odd words that are not easily modernized by general rules.

Morphologic Dictionary

All the dictionary entries include partial morphological information (at least its main category, and in some situations the genre or type of verb), making it possible to extract a list of words with associated morphologic information. Table 2 resumes the size of this dictionary and distribution by main morphologic categories.

Table 2. Morphologic distribution of Dicionário-Aberto entries

Nouns		**Adjectives**	**Adverbs**	**Verbs**		**Locutions**
masc.	*fem.*			*transitive*	*intransitive*	
45 657	38 488	30 469	2 962	10 147	4 016	579

REST API

To enable the use of the dictionary in the cloud a simple web service API based on REST principles is available. It supports queries both in XML and JSON, and lets the user query for definitions, given a specific word, or search for near misses, prefixes and suffixes. This enables the development of mobile applications. An example of such application is the iPhone interface to Dicionário-Aberto[5]

3 Resources under Development

Further work is being done to make the Dicionário-Aberto experience more interesting, enabling new services in the web site, but also to develop new resources that can be used by natural language processing researchers.

Reverse-Order Dictionary

Reverse-Order Dictionaries are not very common (do not confuse with Reverse Dictionaries, discussed below). They let the user browse the dictionary searching by the end of the word, instead of its beginning (looking up for suffixes instead of prefixes).

One of their applications is the construction of a rhyme dictionary (notice however that this will be a partial rhyme dictionary, as some words with different orthography have the same sound, like *doce* and *fosse*).

Another use of these kind of dictionaries is the study of the morphology of a language [6], like the study of suffix productivity (productivity for some scientific terminology suffixes — *-ato, -eto, -ito* — the productivity of effect/result words — *-data, -ção, -são, -ança, -ância*, etc.).

[5] Developed by log.oscon, available from the Apple AppStore. Further details can be found at `http://log.pt/dicionarioaberto/`

Reverse Search or Reverse Dictionary

In a standard (paper) dictionary, the query can only be performed browsing the list of alphabetically sorted lemmas. In machine-readable dictionaries, the search should not be limited to the lemmas. The user should be able to search the full entries, including its definition, examples, etymology or any other section.

This reverse search capabilities transforms electronic dictionaries in ideological or conceptual databases, also known as analogical or onomasiological dictionaries. This feature is available for some languages like Spanish[6] and English[7].

There is a long tradition of onomasiological dictionaries for the European languages. Some ideological dictionaries were prepared during XIX and XX centuries[8], that allowed the reader to search an idea or concept in a descriptors structure similar to a thesaurus [11], or a structured list of concepts sorted by subjects (summary tables) together with a list of hypernyms or broader terms (categories, general ideas) that lead the reader to the searched word.

The reverse search functionality can help surpassing some of the limitations present in paper versions of these dictionaries. This functionality can not be only seen as a simple query tool. Imagine the potentiality of reverse search if the dictionary authors use a controlled language to write their definitions, just like a descriptors thesaurus.

Its main usage is to search a word that we know adequate for a specific situation, but that we can not remember at that moment, or to search for a more specific word, or even to check if there is some word to express some concept [4].

Dicionário-Aberto will not only offer this functionality for end-users through the web interface, and for cloud applications through the REST API, but also make available the reverse index, for offline processing.

Ontology View

Hugo Oliveira [7] has been working in the creation of Onto.PT, a lexical ontology for the Portuguese Language. Oliveira performed some experiments with different resources, and Dicionário-Aberto was also covered. A similar approach was also performed by Simões et al [10]. With these experiments in mind, and the interesting results obtained, a new view for Dicionário-Aberto is being developed: a (lexical) ontology view.

This view will enable the user to query the dictionary, using the standard search or one of the two new methods described earlier, and together with the definition, the examples, and the etymology, consult a thesaurus-like structure. This structure includes a set of relations (like synonymy, antonymy, hyperonymy,

[6] Reverse dictionary search in the Dictionary of the Real Spanish Academy, by Gabriel Alberich: `http://dirae.es/`

[7] OneLook Reverse Dictionary, by Doug Beeferman, that searches more than one thousand indexed dictionaries: `http://www.onelook.com/reverse-dictionary.shtml`

[8] Some examples are listed by Martínez de Sousa [5].

instances/species/genres, actor/action, etc) to other dictionary entries (in case of multi-word expressions each component word will have its own link).

The extraction uses a set of patterns, just like the methods described by the mentioned authors. We decided not to reuse the extracted data because in the future, as stated in the introduction, Dicionário-Aberto will be a Wiki, making it crucial to have an automatic method to recalculate the ontology. The extraction method will run everyday in an unsupervised way.

The ontology completeness will be guaranteed by a set of completion rules. For example, the synonymy relation is symmetric, the hyperonymy relation is inverse of the hyponymy relation, the hyperonymy relation can be seen as a special case of a transitive relation, antisymmetric relations, anti-reflexive, etc. These properties can be described in a mathematical notation and used to ensure that entries that do not have a reference to related words can still get the relation information.

This kind of feature will make the dictionary much more interesting for the end user but also for the natural language processing researcher. For the first, it will make entries browsable by concept. For the second, will be a complement of a Portuguese word-net, as concepts will also include definitions.

4 Conclusions

In this document we described the current status of Dicionário-Aberto, an open project to the development of a knowledge and feature rich dictionary for the Portuguese language, both to be used as a standard dictionary but also as a resource for natural language processing tasks. We defend that resources construction should be automatic, especially in a case like Dicionário-Aberto that will be open for the community to cooperate. This guarantees that the extracted resources can grow in size and quality at the same time as its main resource.

References

1. Project Gutenberg. Project Gutenberg Literary Archive Foundation (November 2011), http://www.gutenberg.org/
2. Almeida, J.J., Santos, A., Simões, A.: Bigorna – a toolkit for orthography migration challenges. In: Calzolari, N., et al. (eds.) Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, pp. 227–232. European Language Resources Association (May 2010)
3. TEI Consortium, editor. TEI P5: Guidelines for Electronic Text Encoding and Interchange, chapter 9. Dictionaries. TEI Consortium (January 2012), http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html, Version 2.0.1 edition, December 22, 2011
4. Porto da Pena, J.A.: Dicionário de uso del Español (2003), http://cvc.cervantes.es/actcult/mmoliner/diccionario/ (retrieved November 3, 2011)
5. Martínez de Sousa, J.: Diccionario de lexicografía práctica. Biblograf, Barcelona (1995)

6. Millán, J.A.: Zigzag, gong, ping-pong, iceberg. donde se descubre que hay diccionarios inversos, y su utilidad manifiesta para el progreso de la humanidad (1999), `http://jamillan.com/inverso.htm` (retrieved November 3, 2011)
7. Oliveira, H.G., Gomes, P.: Onto.PT: automatic construction of a lexical ontology for Portuguese. In: 5th European Starting AI Researcher Symposium (STAIRS 2010) (August 2010)
8. Simões, A., Farinha, R.: Dicionário Aberto: Um novo recurso para PLN. Vice-versa (16), 159–171 (2011)
9. Simões, A., Almeida, J.J.: Processing XML: a rewriting system approach. In: Simões, A., da Cruz, D., Ramalho, J.C. (eds.) XATA 2010 — 8^a Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas, Vila do Conde, Maio, pp. 27–38 (2010)
10. Simões, A., Almeida, J.J., Farinha, R.: Processing and extracting data from Dicionãrio Aberto. In: Calzolari, N., et al. (eds.) Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, pp. 2600–2605. European Language Resources Association (May 2010)
11. van Slype, G.: Les langages indexation: conception, construction et utilisation dans les systmes documentaires. Les Editions d'Organisation, Paris (1987)