View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by CiteSeerX

Building a computational lexicon by using SQL

Alessandro Mazzei
Dipartimento di Informatica
Universit degli Studi di Torino
Corso Svizzera 185, 10149 Torino
mazzei@di.unito.it

Abstract

English. This paper presents some issues about a computational lexicon employed in a generation system for Italian (Mazzei et al., 2016). The paper has three goals: (i) to describe the SQL resources produced during the construction of the lexicon; (ii) to describe the algorithm for building the lexicon; (iii) to present an ongoing work for enhancing the lexicon by using the syntactic information extracted from a treebank.

Italiano. *Questo lavoro descrive la costruzione di un lessico computazionale per la generazione automatica dell'italiano (Mazzei et al., 2016). Il lavoro ha tre obiettivi: (i) descrivere alcune risorse SQL prodotte funzionalmente alla costruzione del lessico; (ii) descrivere l'algoritmo per la costruzione del lessico; (iii) presentare un lavoro in divenire per migliorare il lessico che usa l'informazione sintattica estratta da un treebank.*

1 Introduction

A number of free large multilingual resources covering Italian have been released, e.g. MultiWordnet, UniversalWordnet, BabelNet (Pianta et al., 2002; de Melo and Weikum, 2009; Navigli and Ponzetto, 2012). Moreover, several lexical corpora have been built specifically for Italian, as the detailed map of the Italian NLP resources produced within the PARLI project shows[1]. Unfortunately most resources are designed to represent lexical semantics rather than morpho-syntactic relations among the words. As a consequence, these resources cannot be employed in statistical or rule-based natural language morho-syntactic analyzer or generator.

A notable exception is the PAROLE-SIMPLE-CLIPS lexicon, that is a four-level (i.e. phonological, morphological, syntactical, semantic) general purpose lexicon composed by $53,044$ lemmata (Ruimy et al., 1998). Unfortunately, a strong limitation for the usage of PAROLE-SIMPLE-CLIPS is the licence, since it is not freely available for research or commercial use.

Rule-based natural language realization engines, that are systems performing linearisation and morphological inflections of a proto-syntactic input tree (Gatt and Reiter, 2009), need wide coverage morpho-syntactic information as knowledge-base. In other words, to perform realization, that is the last step of natural language generation (Reiter and Dale, 2000), one needs two main kinds of linguistic knowledge: (i) the grammatical/syntactical knowledge that specifies the syntactic rules of the language and which is usually encoded into formal rules; (ii) the morphological and lexical knowledge, which is usually encoded into a computational lexicon. In the porting of the SimpleNLG system to Italian (henceforth *SimpleNLG-IT*) (Mazzei et al., 2016), we have used the grammar (Patota, 2006) as the linguistic reference for the syntax: we have encoded the Italian syntactic inflections and word ordering by using IF-THEN-ELSE rules in Java. However, since Italian has a high number of irregularities for verb and adjective inflections, we needed for a specifically designed computational lexicon too. We needed for a lexicon that has both a good coverage and a detailed account of the morphological irregularities.

In order to build this specific lexicon, that we have called *SimpleLEX-IT*, we have decided to merge three free resources for Italian, namely *Morph-it!* (Zanchetta and Baroni, 2005), the *Vo-*

[1] http://parli.di.unito.it/resources_en.html

cabolario di base della lingua italiana (De Mauro, 1985) and, for some specific issues, Wikipedia. The differences between the three resources can be referred to both the reasons for which the authors developed them and the adopted methodology and approach they applied in their development: the first is a hand-made list of *basic* words; the second one is an extensional corpus based morphological lexicon; the third one is a collection of encyclopedic entries about irregular verbs in Italian.

This paper is organized as follows: in Section 2 we describe the conversion of the three lexical resources used into a relational database; in Section 3 we provide some details about the algorithm used to build SimpleLEX-IT; in Section 4 we describe a work in progress to enrich the lexicon by using the syntactic information extracted from a treebank; finally, Section 5 closes the paper with conclusions.

2 Using relational database for representing linguistic data

In order to merge different lexical resources we needed to convert them in a common computational representation. We used a relational database[2] (SQL henceforth) since all the three resources are originally provided as text files, organized as tables or simple list.

The first resource that we exploited for populating SimpleLEX-IT is Morph-it! (Zanchetta and Baroni, 2005). The dataset released in the Morph-it! project consists of a lexicon organized according to the inflected word forms, with associated lemmas and morphological features. The lexicon is provided by the authors as a text file where the values of the information about each lexical entry are separated by a tab key. It is an alphabetically ordered list of triples form-lemma-features. An example of the annotation for the form *corsi* (*ran*) is:

```
corsi correre-VER:ind past+1+s
```

where the features are the part of speech (PoS, `VER`b), the mood of the verb (`ind`icative), the tense (`past`), the person (`1`), and the number (`s`ingular). The last released version of Morph-it! (v.48, 2009-02-23) contains $505,074$ different forms corresponding to $35,056$ lemmas. It has been realized starting from a large newspaper corpus, nevertheless it is not balanced and a small number of also very common Italian words are not included in the lexicon, e.g. *sposa* (bride), *ovest* (west) or *aceto* (vinegar). Morph-it! represents extensionally the Italian language by listing all the morphological inflections, i.e. adjective, verbs, nouns inflections are represented as a list rather than by using morphological rules. We converted Morph-it! in SQL by exploiting its original feature structure: we used one single attribute to represent one single feature[3]. We used one table to collect all the lemmata and seven tables, with a different number of attributes, to collect the various inflected forms:

- the table *lemmata* is formed by 3 attributes: a lemma, its PoS and its ID (integer). This table contains $34,725$ records. A number of lemmata belonging to the original version of Morph-it! have been excluded in our conversion: proper nouns, emoticons and cardinals beginning with a digit (e.g. *15mila*).
- the tables *det_demo_table, pro_demo_table, pronou_table* are used to collect inflected form of demonstrative determiners (116 records, 4 attributes: *ID_word, form, ID_lemma, number, gender*), demonstrative pronouns (95 records, 5 attributes: *ID_word, form, ID_lemma, number, gender*), personal pronouns (63 records, 7 attributes: *ID_word, form, ID_lemma, person, number, gender, clitics*).
- the tables *adv_table, adj_table, nou_table, ver_table* are used to collect inflected form of adverbs ($1,594$ records, 3 attributes: *ID_word, form, ID_lemma*), adjectives ($72,367$ records, 6 attributes: *ID_word, form, ID_lemma, kind, number, gender*), nouns ($35,618$ records, 5 attributes: *ID_word, form, ID_lemma, number, gender*) and verbs ($392,139$, 8 attributes: records: *ID_word, form, ID_lemma, mode, time, person, number, gender*) respectively.

The second resource we exploited for populating SimpleLEX-IT is the "Vocabolario di base della lingua italiana" (*VdB-IT* henceforth), a collection of around $7,000$ words created by the linguist Tullio De Mauro and his team[4] (De Mauro, 1985). The development of this vocabulary has been mainly driven by the distinction between the

[2]We used the PostgreSQL database .

[3]Morph-IT! is provided with a script that allows for a naive conversion into SQL that use one single table and one single attribute for all the features.

[4]The second edition of the vocabulary has been announced (Chiari and De Mauro, 2014) and it is going to be released (p.c.).

most *frequent* words (around 5.000) and the most *familiar* words (around 2.000). VdB-IT is therefore organized in the following three sections:

- the *vocabolario fondamentale* (fundamental vocabulary), which contains $2,000$ words featured by the highest frequency into a balanced corpus of Italian texts (composed of novels, movie and theater scripts, newspapers, basic scholastic books); *amore* (love), *lavoro* (work), *pane* (bread) are in this section.
- the *vocabolario di alto uso* (vocabulary of high usage), which includes other $2,937$ words with high frequency, but lesser than the *vocabolario fondamentale*; *ala* (wing), *seta* (silk), *toro* (bull) are in this section
- the *vocabolario di alta disponibilità* (vocabulary of high availability), is composed of $1,753$ words not often used in written language, but featured by a high frequency in spoken language, which are indeed perceived as especially familiar by native speakers; *aglio* (garlic), *cascata* (waterfall), *passeggero* (passenger) are in this section.

The list of lemmata of VdB has been converted in SQL by using one single table, called *lemmadema* (6540 records), which have two attributes, i.e. an ID (integer) and the lemma.

The third resource that we used for the lexicon creation is Wikipedia. Our reference grammar (Patota, 2006) reports a partial list of the principal Italian irregular verbs, but we decided to use the larger list of verbs reported in Wikipedia[5] (*VerIrr* henceforth). Another linguistic distinction for Italian verbs reported in Wikipedia[6] (*VerInc* henceforth) has been exploited in the lexicon: the *incoativi* verbs are a subclass of the third conjugation that have a special behavior in the present time (e.g. *capire*). So, in order to produce the correct conjugation of these verbs in SimpleNLG-IT, they needed to be marked in the lexicon. Both these lists of verbs have been converted in SQL by using two distinct tables which have two attributes, i.e. an ID (integer) and the verb in the infinitive form. The two tables are *verbiirregolari* (858 records) and *verbiincoativi* (726 records).

A notable advantage of the SQL representation for linguistic resources is the possibility to extract intrinsic information with simple queries. Indeed, we found that Morph-it! and VdB share $4,086$ nouns and $1,448$ verbs, but there are 245 lemmas belonging to VdD and not belonging to Morph-it!: most of these words are nouns, for instance *lavapiatti, chimica, inclinta*, but we found too a systematic difference for verbs. Indeed, VdB consider as *proper* reflexive a number of verbs, for instance *avvantaggiarsi, sdraiarsi*. In contrast, these verbs are are treated as *improper* reflexive in Morph-it!, which annotates *avvantaggiare* and *sdraiare* as their lemmata.

3 Building SimpleLEX-IT 1.0

In this section we describe the algorithm used to build the computational lexicon SimpleLEX-IT, which is based on the three resources described in the Section 2, and that has been used in SimpleNLG-IT.

A computational lexicon can be split in two major classes: open and closed classes. The closed class, that are usually composed by function words (i.e. prepositions, determiners, conjunctions, pronouns, etc.) is one to which new words are very rarely added. In contrast, the open classes, that is usually composed by lexical words (i.e. nouns, verbs, adjectives, adverbs), accept the addition of new words. We adopted the same strategy of (Vaudry and Lapalme, 2013): we built by hand the closed part of the Italian lexicon and we built automatically the open part by using the available resources.

In order to build the open class for SimpleLEX-IT we needed both a large coverage and a detailed account of morphological irregularities, also considering their high frequency in Italian. Moreover, in order to have good time execution performance in the realiser (cf. (De Oliveira and Sripada, 2014)), a trade-off between the size of the lexicon and its usability for our task must be achieved, which consists in assuming a form of word classification where fundamental Italian words are distinguished from the less-fundamental ones. In order to balance completeness and efficiency in SimpleLEX-IT, we put in the lexicon the open classes words belonging to the intersection of VdB-IT and Morph-it!.

We reported in Algorithm 1 the process used for the insertion and the annotation of the words belonging to the open classes in SimpleLEX-IT. Note that in order to recognize proper reflexive verbs, we check if the infinitive form of the verb

[5] https://it.wikipedia.org/wiki/Verbi_irregolari_italiani

[6] https://it.wikipedia.org/wiki/Verbi_incoativi

```
foreach adverb ∈ Morph-it! ∩ VdB-IT do
    Add the adverb in normal form into SimpleLEX-IT
end
foreach adjective ∈ Morph-it! ∩ VdB-IT do
    Add the adjective in normal form (masculine-singular) and in
    feminine-singular, masculine-plural, feminine-plural forms, into
    SimpleLEX-IT
end
foreach noun ∈ Morph-it! ∩ VdB-IT do
    Add the noun in normal form (singular), the plural form, and the
    gender into SimpleLEX-IT
end
foreach verb ∈ Morph-it! ∩ VdB-IT do
    if verb ∈ VerIrr then
        Add all the inflections for the indicativo presente,
        congiuntivo presente, futuro semplice, condizionale,
        imperfetto, participio passato, passato remoto into
        SimpleLEX-IT
    else
        if verb ∈ VerInc then
            Set active the incoativo feature in the entry
        end
        if the verb is properly reflexive (i.e. "...rsi") then
            Set active the reflexive feature in the entry
        end
        Add the verb in normal form into SimpleLEX-IT
    end
end
```

Algorithm 1: The algorithm for building the adverbs, adjectives, nouns and verbs in SimpleLEX-IT

has the postfix "*rsi*", since MorphIT! contains this inflection as its normal form. In Table 1 we reported some statistics about SimpleLEX-IT composition. Most of the lexicon is composed by nouns (58%), followed by verbs (21%), adjectives (19%), and adverbs (2%).

PoS	Number	%
Adverb	146	2
Verb (irr.)	283	4
Verb (reg.)	1168	17
Adjective	1333	19
Noun	4092	58
Total	7022	100

Table 1: Number of adverbs, adjectives, nouns and verbs in SimpleLEX-IT.

4 Work in Progress: adding information from a treebank

The Universal Dependency Treebank (UDT) is a recent project that releases freely available treebanks for 33 languages (in this work, version 1.2) (Nivre et al., 2016). Each UDT is split in three sections, *train*, *dev* and *test*, which can be exploited in the evaluation of NLP/NLG systems.

We are working on the idea of adding more information in SimpleLEX-IT by using UDT-IT, i.e. the Italian section of UDT. A specific case that we are currently considering regards auxiliary verbs. The current version of SimpleNLG-IT does not manage auxiliary verbs: in order to produce some complex verb tense, e.g. *passato prossimo*, the user needs to give in input to the realiser the correct auxiliary, i.e. *essere* (to be, e.g. *Io sono nato a Napoli*) or *avere* (to have, e.g. *Io ho amato la scuola.*). Our reference grammar reports complex rules based on lexical semantics in order to choose the correct auxiliary verb and, unfortunately, these rules have many exceptions. So, we can use UDT-IT to empirically decide the correct auxiliary in SimpleNLG-IT. By following this idea, we converted UDT-IT in SQL by exploiting its original feature structuree. We used one table to collect information about the sentences, and one table to collect information about the words:

- the table *sentence_ud* is formed by 4 attributes: an ID (integer), the original treebank (i.e. TUT, ISST, etc.), the original ID, the section (i.e. DEV, TRAIN, TEST).
- the table *words_ud* is used to collect all the words of the UDT-IT. It uses 21 attributes: one attribute *id_sentence*, contains the id of the sentence in the table *sentence_ud*, and 20 attributes correspond to the featured used in the UD annotation.

In order to find the correct auxiliary for a specific verb in UDT-IT, we need to exclude passive, reflexive and modal verb constructions in the query. We found 512 verbs of SimpleLEX-IT that are used in UDT-IT with an auxiliary. It is interesting to note that 60 verbs are used both with the auxiliary *essere* and with the auxiliary *avere*: this is grammatical for some verbs (e.g. *vivere*), but more often we found an annotation error in the UDT-IT.

Finally, note that another possible use of UDT-IT regards the evaluation of the lexicon. In future work we plan to quantify the coverage of SimpleLEX-IT by using the TEST section of the UDT-IT.

5 Conclusions

In this paper we have presented some issues about the computational lexicon SimpleLEX-IT. We described the algorithm used to build the lexicon, three SQL resources produced as side effects of the lexicon building and a work in progress about the extraction of syntactic information from UD-IT.

All the resources described in this paper can be downloaded at `https://github.com/alexmazzei/SimpleLEX-IT`.

References

Isabella Chiari and Tullio De Mauro. 2014. The New Basic Vocabulary of Italian as a linguistic resource. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *1th Italian Conference on Computational Linguistics (CLiC-it)*, volume 1, pages 93–97. Pisa University Press, December.

Tullio De Mauro. 1985. *Guida all'uso delle parole*. Libri di base. Editori Riuniti.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

Rodrigo De Oliveira and Somayajulu Sripada. 2014. Adapting simplenlg for brazilian portuguese realisation. In *Proc. of INLG 2014*.

Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *Proc. of ENLG 2009*, ENLG '09.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. SimpleNLG-IT: adapting SimpleNLG to Italian. In *Proc. of INLG 2016*. TO APPEAR.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1:A Multilingual Treebank Collection. In *Proc. of LREC'16*, may. TO APPEAR.

Giuseppe Patota. 2006. *Grammatica di riferimento dell'italiano contemporaneo*. Guide linguistiche. Garzanti Linguistica.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Nilda Ruimy, Ornella Corazzari, Elisabetta Gola, Antonietta Spanu, Nicoletta Calzolari, and Antonio Zampolli. 1998. The European LE-PAROLE project: the Italian Syntactic Lexicon. In *Proceedings of the First International Conference on Language resources and Evaluation* , pages 241–248.

Pierre-Luc Vaudry and Guy Lapalme. 2013. Adapting simplenlg for bilingual english-french realisation. In *Proc. of ENLG 2013*.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the italian language. *Corpus Linguistics 2005*, 1(1).