Morph-it!

A free corpus-based morphological resource for the Italian language

Eros Zanchetta and Marco Baroni
SITLEC / SSLMIT
University of Bologna / University of Bologna
eros@sslmit.unibo.it / baroni@sslmit.unibo.it
http://sslmit.unibo.it/morphit

1. Introduction

Lexicons and morphological analysers are at the core of many NLP applications, such as lemmatisation, POS tagging and morphology generation. Unfortunately, since the creation of a lexicon tends to be a long and labour intensive task (especially for highly inflectional languages), to date, there are no freely available lexicons for the Italian language. This is the reason why we embarked on the task of creating our own lexicon and then decided to make it freely available. In this paper we describe our method for the rapid creation of a lexicon using a mixture of corpus based techniques and manual checking.

Our main source of linguistic data was the "Repubblica" corpus (Baroni et al. 2004), we extracted lemmas and inferred morphological information not present in the original corpus (i.e. gender) using distributional as well as morphological cues. With that information we then generated inflected forms for all extracted lemmas.

The project is not yet complete and a first evaluation of the quality of the resource suggests that more words from everyday language should be added. Also proper nouns, loan words, diminutive adjectives and a large number of forms of verbs with clitics attached are still missing.

So far the project has been carried out by two people working part time on it, for a total of about 600 person hours.

In this paper we illustrate the process of creating a lexicon of the Italian language, the same methodology can however be easily adapted and replicated in other knowledge-poor morphological extraction projects in different languages.

2. The raw material

Our main source of linguistic data was the "Repubblica" corpus, a large corpus (approximately 380 million tokens) of newspaper Italian containing all the articles published by "La Repubblica" (one of Italy's most read newspapers) between 1985 and 2000. The corpus is annotated with lemmas and POS tags (ADJ, NOUN, VER etc.), for more information on the "Repubblica" corpus see Baroni et al. 2004.

In the early stages of development, we also used a 25-million-tokens corpus created using the BootCat Tools. The BootCat Tools are a set of Perl scripts used to automatically construct "disposable" web based corpora using a few seed words. For the purpose of creating our corpus, seed words were extracted from "Repubblica" (see

Baroni and Bernardini 2004). The downloaded texts were processed using Van Noord's TextCat language guesser (http://www.let.rug.nl/~vannoord/) to eliminate non Italian pages and then tagged using TreeTagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html).

3. Methodology

Early in the planning stage, we assumed that a simple extraction of words from the corpus would not yield a sufficient number of verbal and adjectival forms, mainly for two (strictly related) reasons: 1) Italian is an inflectional language and rare or unusual inflected forms might not appear at all in the corpora; 2) the corpora we used had been automatically annotated, therefore, since we could not rule out annotation errors, when extracting lemmas we had to leave out all low frequency types to avoid introducing too many errors in the lexicon.

This initial assumption was proved to be right later on by the comparison between our lexicon and *Colfis*, a frequency lexicon of the Italian language: 83% of the forms *Colfis* was missing with respect to those found in *Morph-it!* were verbs and 13.6% were adjectives (see below for more details on the evaluation process).

The most promising method for obtaining a complete lexicon appeared to be identifying lemmas and then generate all inflected forms by rule, a non-trivial task given the number of irregularities found in Italian morphology. We used the tagger's lemmatisation to identify potential lemmas and then inferred more information using various techniques. We employed different strategies to extract and process lemmas belonging to the 4 categories into which we divided the lexicon: verbs, adjectives, nouns, adverbs and function words

3.1 Verbs

We started by collecting all lemmas tagged as verbs by TreeTagger (TT), and discarding all those not ending in -are/-ere/-ire/-rsi (the suffixes of the three canonical conjugations in Italian, plus reflexive verbs formed by infinitive + the reflexive particle si), since lemmas with different suffixes would be a clear indication of tagging errors. Lemmas ending in -rsi which had no equivalent verb ending in -re were inserted in a special list of reflexive only verbs.

We manually created a separate list of highly irregular inflected forms (i.e. “essere”, “andare” etc.) and we created a stop list containing the lemmas in this file to prevent the script from generating regularised versions of irregular verbs.

We finally generated the inflected forms using a set of Perl script centred around *MyConjugate*, a modified version of Aldo Capini's *Lingua::IT::Conjugate* (http://dada.perl.it/). We ran the script on our list of lemmas, merged the list with the one containing the special verbs and then looked for errors. To quickly spot possible sources of problems, we compared our lexicon with the corpus, extracting forms that had been identified as non-unknown verbs by the tagger but were missing from our generated list. By manually examining the list we isolated 5 types of errors:

1. many verbs (2648 lemmas) with an attached clitic (e.g. “abituarsi”) had not been

generated; using a simple regular expression we created a separate list of these verbs and dealt with them separately;

2. a large number (361 lemmas) of truncated verbs (e.g. “portar”) was also missing. For this category too we generated a special lists and processed it separately;

3. another rather large group (266 lemmas) was formed by those that we dubbed “iscere” verbs, that is verbs belonging to the third conjugation terminating in “-isco” in the 1st person of the present indicative (e.g. “addolcire”, “usufruire”). A few experiments revealed that these verbs were inflected correctly only if the script was instructed to behave as if their lemma terminated in “-iscere” instead of “-ire”, hence the name. All the “iscere” verbs were added to the list of exceptions in the *MyConjugate* module.

4. the fourth group was formed by clusters of prefixed forms and their roots (e.g. “togliere/distogliere”, “giungere/ingiungere/raggiungere”). We isolated the 78 roots (57 verbs, 21 bound stems), added all the prefixes each root could occur with and again manually updated *MyConjugate* to take into account all possible exceptions.

5. the final group was formed by all those errors that did not fit into any of the preceding categories. Members of this group were manually checked one by one and corrections were made to the scripts or to the lists to take them into account.

Once we had completed the analysis of the errors and made all the necessary adjustments, we regenerated the list of verbs and started the error correction procedure again. The whole process was repeated several times until we were satisfied with the result. We can summarise this iterative refinement process as follows:

- comparison – we compared the generated list with the corpora to isolate missing forms

- analysis – mismatching forms were usually the result of three types of error: a) high frequency lemma assigned to the wrong category by TT b) irregular verbs/adjectives generated as if they were regular, c) bugs in our scripts;

- adjustment – corrections of the errors in stem lists, exception lists and scripts;

- regeneration of inflected verbs

We stopped when we got to the point where no more errors could be found using this methods. At the time of writing, version 0.31 of *Morph-it!* contains 6,159 verbal lemmas and 396,120 inflected verbal forms.

3.2. Adjectives

The generation of adjectives was somewhat less problematic than the generation of verbs, still it was very time consuming since there were no tools available for the inflection of adjectives and we had to create our own scripts from scratch.

The main script uses simple regular expression matching to determine the morphological features of lemmas and relies on manually compiled lists for the inflection of irregular and invariable forms.

The script generates all forms (masculine/feminine/singular/plural) of the positive and superlative grades. Diminutive adjectives were purposefully excluded from generation because, unlike superlatives, they are rarely productive in modern Italian. We opted to generate regularised forms (as well as the "correct" irregular ones) of the superlative of irregular adjectives. For example, the correct superlative of "aspro" ("sour" but also "harsh") is "asperrimo", but we also generated the regularised form "asprissimo". We did this because, deprecated as they may be, these forms are widely attested, especially on the Web, and since one of the main functions of this lexicon is to provide a useful tool for NLP applications, we preferred to be descriptive rather than prescriptive in our representation of the Italian language.

Here again, we began our work by extracting from our corpora all forms tagged as adjectives. We discarded low frequency types and divided the remaining ones into two groups:

1. potential adjectives with inconspicuous morphological features were placed in the main list for generation. Adjectives whose lemma TT was unable to determine with certainty were also added to this list;

2. candidates ending in a consonant were placed in a special list of possible loan/invariable adjectives;

Both lists were manually checked: non-adjectives were discarded and invariable adjectives were added to a separate stop list. We also manually compiled a special list containing a dozen irregular adjectives.

Using the same method employed for verbs, we merged all the lists, generated the lexicon and then compared it with the corpus, extracting forms that had been identified as adjectival forms by TT but were missing from our generated list. By analysing the errors we were able to make the necessary adjustments to the scripts or the lists and then regenerated the lexicon.

Here again the process was repeated several times, until we were unable to find any more errors. At the end of the process we had 9,442 adjectival lemmas and 72,683 inflected adjectives.

3.3. Nouns

3.3.1. Stage one

The TT tags nouns as such, but does not specify gender and number. Thus, our task was to find out the gender and number of words tagged as nouns. Given the highly irregular nature of nominal morphology, and the presence of nouns that only occur as singular or plural, in most cases we looked for corpus evidence for both the singular and plural form of a noun, rather than generating forms that were not attested in the corpus.

However, we did use the presence of both a singular and a plural form that could be connected by a plausible singular<->plural correspondence rule as evidence that an analysis was correct (e.g., the fact that the corpus contained both gatto in plausible masculine singular contexts and gatti in plausible masculine plural contexts, and that o<->i is a possible correspondence rule, was considered as evidence in favour of both the masculine singular analysis of gatto and the feminine plural analysis of gatti).

Our general strategy was to look for the occurrence of forms in contexts in which they were immediately preceded by an article/determiner, or by a sequence of article/determiner and adjective, where the article/determiner was unambigously associated to a specific set of morphosyntactic features (e.g., “il” is unambigously masculine and singular; “l'” was not considered since it could be both masculine and feminine).

For each noun, we collected token and type statistics in unambiguous contexts. For example, let's suppose that the form “paperelle” (“duckies”) appeared in the following contexts:

- le piccole paperelle (art + adj + noun, where the article is unambigously feminine plural)

- alcune paperelle (det + noun, where the article is unambigously feminine plural)

- lo portano paperelle (pronoun + verb + noun sequence wrongly tagged by TT as art + adj + noun, where lo, as article, is unambigously masculine singular)

Suppose further that in the corpus “le piccole paperelle” occurs 3 times, “alcune paperelle” occurs 7 times and “lo portano paperelle” 1 time. Then, paperelle would have a type frequency of 2 and a token frequency of 10 in feminine plural contexts and both a type and a token frequency of 1 in masculine singular contexts.

The correspondence rules were built by looking at standard Italian grammar references, and are reported in table 1:

Gender	Singular	Plural	Example
m	o	i	gatto / gatti
m	e	i	cane / cani
m	co	chi	cocco / cocchi
m	go	ghi	lago / laghi
m	a	i	poeta / poeti
m	io	i	armadio / armadi
m	αvowel	αvowel	cinema / cinema
m	αconsonant	αconsonant	sport / sport
f	a	e	aorta / aorte

Gender	Singular	Plural	Example
f	e	i	miriade / miriadi
f	ca	che	albicocca / albicocche
f	ga	ghe	aringa / aringhe
f	cia	ce	focaccia / focacce
f	gia	ge	bolgia / bolge
f	αvowel	αvowel	radio / radio
f	αconsonant	αconsonant	star / star

Table 1: singular/plural formation rules

Finally we added a special rule for “gender-benders”, i.e. nouns that are masculine in the singular form and feminine in the plural form (e.g. “braccio/braccia”, “uovo/uova”).

We then extracted lists of couples satisfying any of the above rules and where both the masculine and the plural occurred in an ambiguous context, with a minimum frequency of 2 (type count) and 10 (token count). The test was applied to all possible gender/number combination (we did not apply the “pigeonhole principle” at this stage, i.e. both “braccio/braccia” and “braccio/bracci” were considered). We then checked the lists manually, discarding errors: this allowed us to compute the reliability for each prediction rule, where by reliability we mean the ratio between the number of pairs we kept after manual checking and the total number of pairs generated using the rule.

Singular	Plural	Kept/Total	Reliability
go	ghi	50/50	1.0000
o	i	3107/3204	0.9697
e	i	1577/1630	0.9674
co	chi	129/134	0.9626
io	i	647/681	0.9500
consonant (invariable)	consonant (invariable)	557/588	0.9472
a	i	410/484	0.8471
vowel (invariable)	vowel (invariable)	232/591	0.3925

Table 2: reliability of prediction for singular/plural rules (masculine nouns)

Singular	Plural	Kept/Total	Reliability
gia	ge	10/10	1.0000
cia	ce	54/54	1.0000
ga	ghe	38/38	1.0000
e	i	1199/1240	0.9669

Singular	Plural	Kept/Total	Reliability
ca	che	165/171	0.9649
a	e	2573/2667	0.9647
consonant (invariable)	consonant (invariable)	325/349	0.9312
vowel (invariable)	vowel (invariable)	134/691	0.1939

Table 3: reliability of prediction for singular/plural rules (feminine nouns)

Nouns not satisfying any of the above rules (i.e. no plural was found applying any rule to the singular and vice versa) but with high frequency in an unambiguous context were placed in a special list; their missing form was generated using simple heuristics (a => e, o => i etc.) and the results were checked manually.

At the end of stage one we had 13,370 lemmas and 27,126 inflected forms.

3.3.2. Stage two

By the time we had completed stage one, adjectives had become available. We extracted all non ambiguous adjectival forms (i.e. discarding adjectives such as “rosso”, that could also be a name, and “facile” that could be either masculine or feminine) and repeated the process we had done at the beginning of stage one, looking for nouns appearing in the same contexts indicated above plus:

4. adjective + noun

5. noun + adjective

We considered these last two as the same context for purposes of counting (“amico simpatico” and “simpatico amico” were considered identical tokens). This time we applied the rules in order of reliability and followed the pigeonhole principle: whenever a form satisfied a rule, lower ranking rules were not applied to it. Also, we used a different frequency thresholds for each rule, with higher ranking rules needing a lower frequency to be applied (i.e. masculine “go/ghi” rule was applied to forms having a type/token frequency of at least 2/2, while masculine “a/i” rule was applied only when type/token frequencies of 4/4 or higher were attested). Here again output was manually checked to minimise errors.

Finally we looked for words with just plural or singular form, but with robust distributional cues indicating their number/gender, and we collected those as well, again checking manually the final result.

At the end of stage two, we had 17,331 noun lemmas and 35,282 nominal word forms.

3.4. Function words

Under the label “function words” we group many different grammatical categories:

- modal, auxiliary, causative and aspectual verbs;
- adverbs;
- articles;
- clitics;
- conjunctions;
- determiners;
- pronouns;
- interjections;
- punctuation and sentence markers;
- numerals;
- prepositions;

All these forms were extracted from the reference corpora and manually checked. The extraction procedure was quite straightforward, except for adverbs ending in “-mente”, a highly productive (655 out of 837 adverbs in our lexicon have this termination) yet ambiguous suffix. We narrowed down the list of potential “-mente” adverbs by applying a simple heuristic method: we extracted and manually checked all short (that is one syllable + “mente”) forms ending in “-mente”, since this typically cued an adjective (i.e. “demente”, “clemente” etc.) Longer forms in “-mente” (i.e. that contained more than one vowel before the suffix) were kept.

At the end of the extraction process we had 1,437 lemmas and 2,742 word forms belonging to the function words category.

3.5. The final product

As of the time of writing, the lexicon is formed by 31,955 lemmas and 506,827 word forms and is available in two formats:

1. a finite state transducer for use with the SFST (http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html) tools and Jan Daciuk's FSA utilities (http://juggernaut.eti.pg.gda.pl/~jandac/fsa.html)

2. a human readable list of word forms with their lemma and morphological features. We distinguish between derivational features, that pertain to the lemma, and inflectional features, that pertain to the wordform. Derivational and

inflectional features are separated by a colon. The derivational features are in upper case and they are dash-delimited. The inflectional features are in lower case and they are plus-sign-delimited.

Form	Lemma	Features
rimpinzeremmo	rimpinzare	VER:cond+pre+1+p
abominevoli	abominevole	ADJ:pos+m+p
dabbenaggine	dabbenaggine	NOUN-F:s
ostensibilmente	ostensibilmente	ADV

Table 4: human readable version of *Morph-it!*

The resource is freely available (under a "Creative Commons" licence) and can be downloaded from the project's home page: http://sslmit.unibo.it/morphit).

4. Evaluation

We carried out a first evaluation of our lexicon by comparing it to another lexicon, *Colfis, Corpus e Lessico di Frequenza dell'Italiano Scritto* (Laudanna et al. 1995), a frequency lexicon of written Italian built using a manually constructed, balanced 3 million tokens corpus. Although (for obvious reasons) related, the two lexicons are nonetheless clearly different: *Colfis* is directed specifically towards psycholinguistic research and aims to be a faithful representation of what Italians actually read; *Morph-it!*'s goal, on the other hand, is providing a tool for the practical needs of NLP applications and therefore aims for maximum coverage.

Despite these differences, we feel that a comparison between the two will give us some insight into what can be improved in our lexicon.

4.1. Test 1

First we wanted to see how many high frequency words were missing from *Morph-it!*. We considered the top 10,000 ranks of *Colfis*'s frequency list and performed a few cleaning operations on them. We discarded all but the forms containing alphanumeric characters, hyphens and apostrophes. Then we eliminated all forms beginning with an hyphen or formed by more than one word. This reduced the list to 9139 items. Now we ran the comparison between the two lexicons, which yielded 518 mismatches.

Manual inspection of the mismatches revealed that 472 of them were proper nouns (a category still not covered by Morph-it!), while the remaining 46 belonged to different categories:

Mismatches	Category
12	adverbs
11	nouns
6	verbs
4	interjections
3	adjectives
3	adverbs + clitic

Mismatches	Category
2	pronouns + clitic
2	WH forms
1	prefixes
1	prepositional articles
1	numeral determiners

Total: 46 missing forms

Table 5: results of test 1

These results seem to indicate that the lexicon (with a few exceptions) is fairly complete. The fact that most of the missing forms were orthographic variants (typically truncated forms such as “mezz'”, “bell'” etc.) of forms present in the lexicon suggests that it would probably be better to couple *Morph-it!* with a morphological guesser than to invest too much on expanding it.

4.2. Test 2

In the second test we compared the whole content of the two lexicons and obtained a list of mismatches. We then took a random sample of 300 forms missing from *Colfis* and 300 forms missing from *Morph-it!* and examined them.

Forms missing from Colfis		Forms missing from Morph-it!	
249	verbs	**136**	proper nouns
41	adjectives	**37**	nouns
5	nouns	**35**	unidentified forms
4	errors in *Morph-it!*	**28**	forms beginning/ending in -
1	numeral determiners	**20**	verbs + clitic
		15	loans
		11	adjectives
		6	abbreviations
		5	adverbs
		3	verbs
		3	errors in *Colfis*
		1	interjections

Table 6: results of test 2

Most of the forms missing from Colfis were verbs (83%) and adjectives (13.6%). This confirms our initial assumption that by generating inflected forms, the lexicon would be more exhaustive than if we had limited ourselves to the extraction of forms from the corpora. Also, the low incidence of errors in the sample (1.3%) seems to suggest that the lexicon is relatively correct.

The inspection of the forms missing from *Morph-it!* offers some interesting insights into what can be improved. Setting aside proper nouns (45.3% of missing forms belong to this category, which has not been included yet) and loan words, it's interesting to note that 12.3% of nouns are still missing. The reason for this probably lies in the source of our data, a newspaper corpus and a web corpus created using seeds extracted from the same newspaper corpus. In fact, many of the missing forms come from everyday language (e.g. “bibitone”) of from highly specialised domains (e.g. “desquamazione”), that is words that we do not expect to appear in a typical daily newspaper (or at least not to appear frequently).

Verbs with an attached clitic represent 6.6% of misses. This does not come as a surprise since these forms were not generated but extracted. We opted against generation because the use of clitics tends to be highly idiosyncratic and by indiscriminately generating all verbal forms with an attached clitics we would have introduced too many implausible forms.

Interestingly enough, although a few adjectives were missing (3.6%), none of them were diminutive. This seems to indicate that by not generating diminutive adjectives we did not leave out an exceedingly significant part of Italian morphology.

Conclusions

The goal of this project was to create a linguistic resource for the needs of the NLP community using corpus-based methods, and to create it using very limited resources. The method described in this paper allowed us to create a lexicon comprising 506,827 word forms and 31,955 lemmas in a relatively short time (600 person hours).

A preliminary evaluation of the resource indicates that the lexicon still lacks words from everyday vocabulary, as well as proper nouns, loan words and verb + clitic forms. Inspection of many of the errors emerged during evaluation also suggests that many of them could be avoided by complementing the lexicon with a guesser capable of morphological analysis.

Further directions for future work on the lexicon include introducing a distinction between coordinative and subordinative conjunctions and more generally improving the current features of function words.

References

Baroni, M. / Bernardini, S. / Comastri, F. / Piccioni, L. / Volpi, A. / Aston, G / Mazzoleni, M. (2004) Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian, in *Proceedings of the Fourth Language Resources and Evaluation Conference*, (Lisbon: ELDA), 1771-1774.

Baroni, M. and Bernardini, S. (2004) BootCaT: Boot strapping corpora and terms from the Web, in *Proceedings of the Fourth Language Resources and Evaluation Conference*, (Lisbon: ELDA), 1313-1316.

Laudanna, A / Thornton, A. M. / Brown, G. / Burani, C. / Marconi, L. (1995) Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente, in S. Bolasco, L. Lebart e A. Salem (ed.) *III Giornate internazionali di Analisi Statistica dei Dati Testuali. Volume I*, (Roma: Cisu), 103-109.