AnIta: a powerful morphological analyser for Italian

Fabio Tamburini[◇], **Matias Melandri**[*]

[◇]Dept. of Linguistics and Oriental Studies, University of Bologna, Italy
fabio.tamburini@unibo.it

[*]Dept. of Computer Science, University of Bologna, Italy
mats.mela@gmail.com

Abstract

In this paper we present AnIta, a powerful morphological analyser for Italian implemented within the framework of finite-state-automata models. It is provided by a large lexicon containing more than 110,000 lemmas that enable it to cover relevant portions of Italian texts. We describe our design choices for the management of inflectional phenomena as well as some interesting new features to explicitly handle derivational and compositional processes in Italian, namely the wordform segmentation structure and Derivation Graph. Two different evaluation experiments, for testing coverage (Recall) and Precision, are described in detail, comparing the AnIta performances with some other freely available tools to handle Italian morphology. The experiments results show that the AnIta Morphological Analyser obtains the best performances among the tested systems, with Recall = 97.21% and Precision = 98.71%. This tool was a fundamental building block for designing a performant PoS-tagger and Lemmatiser for the Italian language that participated to two EVALITA evaluation campaigns ranking, in both cases, together with the best performing systems.

Keywords: Morphological Analyser, Derivation Graph, Italian Language, Evaluation

1. Introduction

Stemming and lemmatisation are fundamental tasks at low-level Natural Language Processing (NLP) in particular for morphologically complex languages involving rich inflectional and derivational phenomena. These tasks are usually based on powerful morphological analysers able to handle the complex information and processes involved in successful wordform analysis.

After the seminal work of Koskenniemi (1983) (see also the recent books (Beesley and Karttunen, 2003; Roark and Sproat, 2006) for general overviews) introducing the two-level approach to computational morphology, a lot of successful implementations of morphological analysers for Western European languages has been produced (Beesley and Karttunen, 2003; Cöltekin, 2010; Pianta et al., 2008; Schmid et al., 2004; Tzoukermann and Libermann, 1990). Although this model has been heavily challenged by some languages (especially semitic languages (Gridach and Chenfour, 2010; Kiraz, 2004)), it is still the reference model for building such kind of computational resources at least for Western European languages.

These models usually implement two different operations: a) analysis, which extracts all the information connected with a wordform associating it to a standardised notation "lemma+features" – for example the form *libri* ('books') becomes "*libro*+Noun+Masc+Plur" and the form *amo* (which is ambiguous in Italian, and may correspond to 'I love' or to 'hook') is associated to two different lemmas, "*amare*+Verb+Ind+Pres+1p+Sing" and "*amo*+Noun+Masc+Sing" – and b) generation, the opposite operation, which associates to a structure "lemma+features" the corresponding wordform – for example the structure "*dormire*+Verb+Ind+Pres+1p+Sing" is associated to the wordform *dormo* ('I sleep').

In the late nineties some corpus-based/machine-learning methods were introduced to automatically induce the information for building a morphological analyser from corpus data (see the review papers (Creutz and Lagus, 2007; Hammarström and Borin, 2011)). These methods seem to be able to induce the lexicon from data, avoiding the complex work of manually writing it, despite some reduction in performance.

1.1. Italian Morphology

Italian is one of the ten most widely spoken languages in the world. It is a highly-inflected Romance language and simple words can be modified by, essentially, three morphological processes: inflection, derivation and compounding. This section will give a short overview of Italian morphological phenomena.

Inflection

Words belonging to inflected classes (adjectives, nouns, determiners and verbs) exhibit a rich set of inflection phenomena. There are essential two basic types of inflection: noun inflection and verb inflection.

Noun inflection, also shared with adjectives and determiners, has different suffixes (or inflection endings) expressing at the same time both gender and number, while verb inflection presents a rich variety of inflection endings for tense, mood, person and number. Both nouns and verbs can present a large set of regular inflections and a relevant range of irregular behaviours changing the wordform base. All inflection phenomena are realised by using different suffixes, and some morphophonological rules have to be applied to adjust the orthographic form in the juncture between the base and the inflectional ending.

Derivation

Nouns, adjectives and verbs form the base for deriving new words through complex combinations of prefixes and suffixes added to a base form and through conversion processes. A large number of affixes can be combined in various ways in order to derive new words: for example the word '*riallineabilità*' – 'realignability' – is formed by adding two prefixes, '*ri-*' and '*a-*', and two suffixes, '*-bile*' and '*-ità*', to the base '*linea*'. Deciding the actual order of the derivational processes is not obvious and, in most cases, cannot be established in a clear way. We will discuss this problem in detail in one of the following sections.

Compounding
Also compounded forms are quite frequent in Italian. Compounding regards the combination of two base forms to produce a new form: in Italian various combinations of word classes are acceptable, for example Noun+Noun, '*pesce+cane*' – 'shark', Adj.+Adj., '*dolce+amaro*' – 'bittersweet', Verb+Verb, '*gira+volta*' – 'wirl, somersault', Verb+Noun, '*canta+storie*' – 'storyteller', and so on. Not all combinations of word classes, even if attested in some cases, are really productive.
Some compounds can be written also by connecting the two forms with an hyphen or even by keeping the two words separate, but this kind of orthographic spelling is not usually handled by a morphological analyser.

1.2. Computational tools to handle Italian Morphology

From a computational point of view there are some resources able to manage the complex morphological information of the Italian language. On the one hand we have open source or freely available resources, such as:

- *Morph-it* (Zanchetta and Baroni, 2005) an open source lexicon that can be compiled using various packages implementing Finite State Automata (FSA) for two-level morphology (SFST-Stuttgart Finite State Transducer Tools and Jan Daciuk's FSA utilities). It globally contains 505,074 wordforms and 35,056 lemmas. The lexicon is quite small and, in order to be used to successfully annotate real texts, it requires to be extended. Moreover, the lexicon is presented as an annotated wordform list and extending it is a very complex task. Although it uses FSA packages it does not exploit the possibilities provided by these models of combining bases with inflection endings, thus the addition of new lemmas and wordforms requires listing all possible cases.

- *TextPro/MorphoPro* (Pianta et al., 2008) a freely available package (only for research purposes) implementing various low-level and middle-level tasks useful for NLP. The lexicon used by MorphoPro is composed of about 89,000 lemmas, but, being inserted into a closed system, it cannot be extended in any way. The underlying model is based on FSA.

On the other side we have some tools not freely distributed that implement powerful morphological analysers for Italian:

- *MAGIC* (Battista and Pirrelli, 2000) is a complex platform to analyse and generate Italian wordforms based on a lexicon composed of about 100,000 lemmas. The lexicon is quite large, but it is not available to the research community; ALEP is the underlying formalism used by this resource.

- *Getarun* (Delmonte, 2009) is a complete package for text analysis. It contains a wide variety of specific tools to perform various NLP tasks (PoS-tagging, parsing, lemmatisation, anaphora resolution, semantic interpretation, discourse modeling...). Specifically, the morphological analyser is based on 80,000 roots and large lists of about 100,000 wordforms. Again the lexicon is quite large, but, being a close application not available to the community, it does not allow to profitably use such resource to develop new NLP tools for the Italian language.

2. The *AnIta* Morphological Analyser

In this paper we present *AnIta*, a morphological analyser for Italian based on a large hand-written lexicon and two-level rule-based finite-state technologies. The motivations for such choice can be traced back, on the one hand, to the availability of a large electronic lexicon ready to be converted for such models and, on the other hand, on the the aim of obtaining an extremely precise and performant tool able to cover a large part of the wordforms found into real Italian texts (this second requirement drove us to choose a rule-based manually-written system instead of unsupervised machine-learning methods for designing the lexicon).
It is quite common, in computational analysis of morphology, to implement models covering most of the inflectional phenomena involved in the studied language. Implementing the management of derivational and compositional phenomena in the same computational environment is less common and morphological analysers covering such operations are quite rare (e.g. (Schmid et al., 2004; Tzoukermann and Libermann, 1990)).
The implementation of derivational phenomena in Italian considering the framework of two-level morphology has been extensively studied by (Carota, 2006); she concludes that "...the continuation classes representing the mutual ordering of the affixes in the word structure are not powerful enough to provide a motivated account of the co-selectional restriction constraining affixal combination. In fact, affix co-selection is sensitive to semantic properties." Considering this results we decided to implement only the inflectional phenomena of Italian by using the considered framework and manage the other morphological operations by means of a different annotation scheme.
The development of the AnIta morphological analyser is based on the Helsinki Finite-State Transducer package (Lindén et al., 2009).
Considering the morphotactics combinations allowed for Italian, we have currently defined about 110,000 lemmas, 21,000 of which without inflection, 51 continuation classes to handle regular and irregular verb conjugations (following the proposal of (Pirrelli and Battista, 1996) for the latter) and 54 continuation classes for noun and adjective declensions. In Italian clitic pronouns can be attached to the end of some verbal forms and can be combined together

to build complex clitic clusters. All these phenomena have been managed by the analyser through specific continuation classes.

Nine morphographemic rules handle the transformations between abstract lexical strings and surface strings, mainly for managing the presence of velar and glide sounds in the edge between the base and the inflectional ending.

The Appendix shows a lexicon fragment for three simple lemmas.

The management of inflectional phenomena for Italian is fairly standard and do not require special devices or complex solutions in the implementation.

The most interesting feature introduced into AnIta concerns the complex morphological annotation devised to mark the derivational and compounding processes. AnIta is able to produce wordforms where the various morphemes (base, prefixes and suffixes) are clearly marked and segmented.

2.1. A first extension: wordform segmentation

In order to describe derivational phenomena, we devised a first level of annotation able to mark the internal segmentation of word forms. Each form will be associated with a linear structure that can be described by the following regular expression schema:

/ (PREF>)*BASE(<SUFF)*(∼CLITCL)?(-INFLEND)? /

where PREF, BASE, SUFF, CLITCL and INFLEND are strings that represent a prefix, a base, a suffix a clitic cluster and an inflectional ending, respectively.

The insertion of this annotation inside a corpus allows for a large number of sophisticated queries by using regular expressions, for example:

/dis>.+/	wordforms prefixed with dis-
/.+<on-[eia]/	wordforms suffixed with -one (-oni, -ona)
/in>.+<ità/	wordforms simultaneously prefixed with in- and suffixed with -ità

We followed two simple rules to segment the lemmas for marking derivational phenomena: (a) segment a lexicon entry only if its base is an Italian independent word clearly recognisable (we have thus excluded all the bases taken from Greek, Latin or other languages); (b) we decided to keep the affix unchanged, maintaining all possible variations (geminations, clipping, phonetic readjustments, ...) onto the base.

While this first level of morphological annotation allows for a large number of complex queries, it is still unsuitable to represent some fundamental information. First of all, it does not contain any indication about the lexical class of the bases and of the derived forms and, secondly, the representation of Italian complex words it provides is not enough detailed and powerful. A more complete annotation schema, able to complete this first level segmentation, has to be devised in order to capture the complex details of Italian morphological processes.

2.2. Derivation Graphs for representing morphological processes

Two problems are pressing while annotating real texts. First of all, the derivational processes underlying some wordforms cannot be easily described as single derivational trees; instead, a single derived word can involve different possible interpretations giving rise to different trees; consequently, a one dimension model is unsuitable to account for such complex words. Moreover, in order to be able to retrieve all possible morphological combinations, we need to incorporate into the corpus annotation information about the lexical classes both of the bases and of the complex words derived by affixation and to make it available for the users.

We will present the proposed solution to these problems by discussing an example. Let us consider the complex word *s>componi<bile* 'decomposable'. This form can be described as the result of two possible derivational paths, and, consequently, it can be represented by two different trees (we represent the tree using the parenthesised notation indicating the class of the derived form as a subscript):

$$[[s > componi_V]_V < bile]_A$$
$$[s > [componi_V < bile]_A]_A$$

Choosing one of these options, and, consequently, discarding the other, is a strong theoretical choice, since it is impossible to determine, on empirical grounds, whether the adjective *scomponibile* is derived from the adjective *componibile* by adding the prefix *s-* or from the verb *scomporre* by adding the suffix *-bile*. See also (Mahlow and Piotrowski, 2009) and (Celata and Bertinetto, 2010) for similar discussions about this problem.

The formal structure that naturally extends a tree is the "graph". If we consider each element intervening in a derivational process (the base and the affix(es)) as the nodes of a graph (keeping the information on the nature of the affix, as in the segmentation annotation) and the "derivation relation" as the formal device for defining the edges of the graph, we can build the "Derivation Graph" (DG) for the form scomponibile as in Figure 1. The edges have arrows which mark the direction in which a derivation can take place and the class of the derived word.

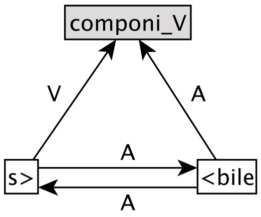

Figure 1: The Derivation Graph for the wordform *scomponibile*.

In order to navigate a graph, two rules must be obeyed:

- the starting point is always the base, that is the upper element (highlighted with grey in Figure 1);
- each edge must be always travelled in the opposite direction of the arrow.

Therefore it is possible to reconstruct all the possible interpretations of a derivational process by navigating the DG following a simple rule: every path in the graph starting from the base and built reversing the derivation relation (i.e. traveling the edges in the opposite direction of the arrows) that includes all the nodes at once leads to a possible interpretation of the derivational history of a complex word, and produces a tree describing this process.

From a theoretical/computational point of view there are various ways of representing a graph structure, depending on the intended final use of such information. One of these methods consists in listing all the graph edges. Using this representation we can describe an entire graph as a single string considering the concatenation of graph edges. For example, the DG in Figure 1 can be expressed through the following list of its edges:

s_V>componi_V#componi_V<bile_A#
s_A>bile_A#s_A<bile_A

where the character '#' acts as a separator between the edges.

Once each wordform in the corpus has been annotated with the string expressing the DG associated with it, the construction of simple but extremely powerful queries is possible with any corpus management program permitting the use of regular expressions in corpus queries, such as the IMS/Corpus Workbench. Some examples of these queries are given below:

/.+_V<bile_A/	all the instances/concordances in which the suffix '-bile' forms an adjective from a verb;
/s_A>.+_A/	all the instances/concordances in which the prefix 's-' forms an adjective from another adjective;
/.+_V<.+_A/	all the instances/concordances in which a suffix forms an adjective from a verb.

In Italian, it is common that a single affix derives words belonging to different lexical classes. It is the case, for example, of the word $[oper_N < aio]_{N/A}$. In order to take these cases into account, we propose to encode all the possible combinations of the four major lexical classes (N, A, V, D(=adverb)) by using the simple encoding schema depicted in Table 1. So, a problematic word like operaio can be associated with the structure $[oper_N < aio]_C$. In this way, operaio will be included in the results from queries aimed at extracting derived nouns and derived adjectives from the corpus.

Code	Combination	Code	Combination
A	**A (Adj)**	I	A+D+N+V
B	A+D	J	D+N
C	A+N	K	D+V
D	**D (Adv)**	L	D+N+V
E	A+V	**N**	**N (Noun)**
F	A+D+N	O	N+V
G	A+D+V	**V**	**V (Verb)**
H	A+N+V		

Table 1: Encoding schema for expressing all the possible combinations of multiple word classes.

The class encoding schema we propose covers all the combinations that are logically possible, although most of them are not attested in Italian. Again, a system based on regular expression searches will help find all the relevant combinations while querying the corpus.

Table 2 show some examples of complete analysis performed by AnIta. The second column – *Morphological analysis* – shows the inflectional analysis of the wordform, while the third column – *Segmentation* – depicts the internal wordform segmentation. The fourth analysis – *Derivational Process* – contains the DG of each wordform, when a derivation process is present.

Please refer to (Grandi et al., 2011) for a complete description of DG, the annotation schema and for the theoretical grounds and consequences involved by this model.

Wordform	**Morphological analysis**
adulti	l_adulto+NN+MASC+PLUR
	l_adulto+ADJ+MASC+PLUR
ricercai	l_ricercare+V_FIN+IND+PAST+1+SING
mangiarglielo	l_mangiare+V_NOFIN+INF+PRES+C_GLI+ C_LO
impareggiabile	l_impareggiabile+ADJ+MASC+SING
	l_impareggiabile+ADJ+FEMM+SING
scomponibile	l_scomponibile+ADJ+MASC+SING
	l_scomponibile+ADJ+FEMM+SING
capostazione	l_capostazione+NN+MASC+SING

Wordform	**Segmentation**	**Derivational Process**
adulti	adult-i	-
	adult-i	-
ricercai	ri>cerc-ai	ri_V>cercai_V
mangiarglielo	mangi-ar∼glielo	-
impareggiabile	im>pareggia<bil-e	pareggia_V<bile_A# im_A>bile_A
scomponibile	s>componi<bile	s_V>componi_V# componi_V<bile_A# s_A>bile_A# s_A<bile_A
capostazione	capo+stazione	-

Table 2: Some examples of AnIta analyses. Various special characters mark univocally the different components in the wordform segmentation ('-' for inflectional suffixes, '<' for derivational suffixes, '>' for prefixes, '∼' for clitic clusters and '+' for compounds).

The wordform segmentation and the DG are not implemented directly into the main morphological analyser, but they are implemented as secondary two-level transducers that take the output of the first analyser as input and produce the proper wordform segmentation and DG.

The information about the internal segmentation has been inserted directly into the lemma string and multiple affixation is handled by inserting multiple symbols (e.g. *ri>ab>bassa<ment-o*). With regard the DG, we inserted directly in the annotation the string representing the edges list that results from the analysis process.

AnIta has been successfully used to annotate CORIS, a large reference corpus of contemporary Italian (Rossini Favretti et al., 2002): such annotations allow for complex corpus queries, as in the examples described before, enabling the CORIS user to retrieve corpus data in a very powerful way.

The core section of AnIta, composed by the 8,000 most frequent lemmas corresponding to the De Mauro (2000) definition of ”base disctionary” (divided into three further classes ‘Fundamental’, ‘High Use’ and ‘High availability’), will be made available freely to the research community[1].

3. Evaluation

In the literature, various possibilities have been proposed for evaluating morphological analysers:

(a) compare the results produced by the morphological analyser with a manually checked set of data (Gold Standard) as in (Faaß, 2011; Mahlow and Piotrowski, 2009; Sawalha and Atwell, 2008). This approach requires, on the one hand, the availability, or production, of an expensive gold standard that, for this reason, is usually quite small. On the other hand we can evaluate the obtained results on a fine-grained basis checking coverage and also classification accuracy among the different morphological possibilities;

(b) compare the analyser coverage against well attested lexicons/dictionary as in (Zanchetta and Baroni, 2005). This approach requires the availability of a large electronic lexical resource and tests the analyser on a standard well attested lexicon, leaving aside most of the terminology that you can find in real texts. Moreover, we have to test the analyser only against the base form of each lemma and cannot verify the correct recognition behaviour for each wordform;

(c) the third possibility involves the computation of the analyser coverage over a large corpus as in (Cöltekin, 2010; Keselj and Sipka, 2008; Schmid et al., 2004; Yablonsky, 1999). Testing the morphological analyser on authentic texts gives a good measure of its coverage performances when working on real data.

3.1. First evaluation step

As a first step, we chose to evaluate AnIta using the third method. We followed the same procedure suggested in (Schmid et al., 2004) for evaluating SMOR, a morphological analyser for German. We extracted the wordform frequency list from CORIS and compute the total amount of wordforms identified by the analyser multiplied for its respective frequency inside the corpus (multiplying each analyser output for the wordform frequency is simply a trick to speed up the process, but it does not change the final results). For testing, we considered only wordforms satisfying the regular expression `/[a-zA-Z]+'?/`, as the purpose of this evaluation is to test the analyser on real words excluding all non-words (numbers, codes, acronyms, ...), quite frequent in real texts.

The metric used for the evaluation of the AnIta coverage is the Word Error Rate (WER), as suggested in (De Pauw and de Schryver, 2008), consisting in the ratio between the total number of tokens recognised by the analyser divided by the total number of tokens analysed (those satisfying the regular expression described before). It is worth noting that in this experiment WER, as defined before, is equivalent to make a measure of the Recall obtained by the system defined as the number of true positives (wordforms that were to be analysed and that were analysed by the system) divided by the sum of true positives and false negatives (wordforms that were to be analysed but that were not analysed by the system) (Faaß, 2011), or, in other words, WER = 1 – Recall.

Number of CORIS tokens	110303560
Number of analysed tokens	86297311
Number of analyses types	592500

Table 3: Statistics for the evaluation data extracted from CORIS.

Thanks to the availability of Morph-It and TextPro we were able to compare AnIta performances against these other two, commonly used, tools for Italian language. See Table 3 for a complete overview of the experiment figures.

Table 4 shows results for all the three tested morphological analysers. The AnIta results are presented considering the two options with (AnIta-PN) and without (AnIta) the insertion of a Proper-Noun list (3,461 entries comprising person names, cities, countries, etc.) into the analyser lexicon. In both cases AnIta outperforms the other systems obtaining smaller WER, as absolute value, which is significantly lower than the others. The insertion of a proper-noun list into the lexicon proved to increase the analyser performance quite significantly.

System	WER (Recall)
AnIta-PN	2.79% (97.21%)
AnIta	3.55% (96.45%)
TextPro	5.01% (94.99%)
Morph-It	6.52% (93.48%)

Table 4: Evaluation WER/Recall results for the CORIS coverage experiment.

3.2. A second evaluation step

The second evaluation step is aimed at measuring the precision of the AnIta Morphological Analyser. For this second experiment we chose to implement an evaluation scheme of type (a): the wordforms contained in the Gold Standard corpus used in the EVALITA 2011 Lemmatisation Task (Tamburini, 2012a), annotated manually both with PoS-tags and correct disambiguated lemmas, were provided to the systems and analysed, verifying if the correct lemma extracted from the Gold Standard is one of the option provided by the tested system. The actual Precision score is then computed as the ratio between the number of wordforms for which the correct lemma is part of the analyser proposed solutions divided by the total number of recognised wordform (for which we have at least one possible solution). To

[1] Please contact the first author for downloading the core version of the AnIta Morphological Analyser.

provide a complete picture, we introduced also a measure of ambiguity, computed as the number of wordforms having more than one lemma as possible solution divided by the number of wordforms recognised by the system. Table 5 depicts the results of this second evaluation step. Unfortunately, for this kind of evaluation, we can compare AnIta precision only with Morph-It, because TextPro implements also the disambiguation step, thus the results are, in this case, not comparable. The Precision exhibited by AnIta is quite high, both as absolute value and compared with the other system performance.

System	Precision	Ambiguity
AnIta-PN	98.71%	53.82%
Morph-It	90.21%	47.99%

Table 5: Evaluation of systems' precision.

Using the AnIta morphological analyser as fundamental resource, we built a new Part-of-Speech tagger, derived from the one presented in (Tamburini, 2007), and a new lemmatiser program able to solve the ambiguity described before and choose the correct lemma among the possibilities provided by the AnIta Morphological Analyser for an ambiguous wordform (Tamburini, 2012b). This Lemmatiser system participated to the Lemmatisation Task of the EVALITA 2011 evaluation campaign (http://www.evalita.it/) obtaining very good accuracy scores. These results are mainly due to the AnIta's large lexicon that allows for a high coverage of Italian texts.

4. References

Marco Battista and Vito Pirrelli. 2000. Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane. Technical report, ILC-CNR.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.

Francesca Carota. 2006. Derivational morphology of italian: Principles for formalisation. *Literary and Linguistic Computing*, 21:41–53.

Chiara Celata and Pier Marco Bertinetto. 2010. Per un?analisi morfologica del lessico italiano. *Quaderni del Laboratorio di Linguistica*, 9:1–13.

Cagri Cöltekin. 2010. A freely available morphological analyzer for turkish. In *Proc. of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process*, 4:3:1–3:34, February.

Tullio De Mauro. 2000. *Il dizionario della lingua italiana*. Paravia.

Guy De Pauw and Gilles-Maurice de Schryver. 2008. Improving the computational morphological analysis of a swahili corpus for lexicographic purposes. *Lexikos*, 18:303–318.

Rodolfo Delmonte. 2009. *Computational Linguistic Text Processing - Lexicon, Grammar, Parsing and Anaphora Resolution*. Nova Science Publisher, New York.

Gertrud Faaß. 2011. A user-oriented approach to evaluation and documentation of a morphological analyzer. In Cerstin Mahlow and Michael Piotrowski, editors, *Systems and Frameworks for Computational Morphology*, volume 100 of *Communications in Computer and Information Science*, pages 46–66. Springer Berlin Heidelberg.

Nicola Grandi, Fabio Montermini, and Fabio Tamburini. 2011. Annotating large corpora for studying italian derivational morphology. *Lingue e Linguaggio*, X:227–244.

Mourad Gridach and Noureddine Chenfour. 2010. XMODEL: An XML-based Morphological Analyzer for Arabic Language. *International Journal of Computational Linguistics*, 1:12–26.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Comput. Linguist.*, 37:309–350.

Vlado Keselj and Danko Sipka. 2008. A suffix subsumption-based approach to building stemmers and lemmatizers for highly inflectional languages with sparse resources. *INFOTHECA, Journal of Informatics and Librarianship*, IX:23a–33a.

George Anton Kiraz. 2004. *Computational Nonlinear Morphology: with emphasis on Semitic Languages*. Cambridge University Press.

Kimmo Koskenniemi. 1983. *Two-level morphology: A general computational model for word-form recognition and generation*. Ph.D. thesis, University of Helsinki.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. HFST Tools for Morphology—An Efficient Open-Source Package for Construction of Morphological Analyzers. In *Systems and Frameworks for Computational Morphology*, pages 28–47, Zurich.

Cerstin Mahlow and Michael Piotrowski. 2009. SMM: Detailed, Structured Morphological Analysis for Spanish. *Polibits: Computer Science and Computer Engineering with Applications*, 39:41–48.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The textpro tool suite. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Vito Pirrelli and Marco Battista. 1996. Monotonic paradigmatic schemata in italian verb inflection. In *Proc. COLING'96*, pages 77–82.

Brian Roark and Richard Sproat. 2006. *Computational Approaches to Morphology and Syntax*. Oxford University Press.

Rema Rossini Favretti, Fabio Tamburini, and Cristiana De Santis. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In Andrew Wilson, Paul Rayson, and Tony McEnery, editors, *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, pages 27–38. Lincom-Europa, Munich.

Majdi Sawalha and Eric Atwell. 2008. Comparative evaluation of arabic language morphological analysers and stemmers. In *Proc. of COLING 2008 (Poster Volume)*, pages 107–110, Manchester.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1263–1266, Lisbon.

Fabio Tamburini. 2007. CORISTagger: a high-performance PoS tagger for Italian. *Intelligenza Artificiale*, IV:14–15.

Fabio Tamburini. 2012a. The evalita 2011 lemmatisation task. In *Working Notes of EVALITA 2011*. CELCT, Trento, Italy.

Fabio Tamburini. 2012b. The anita-lemmatiser. In *Working Notes of EVALITA 2011*. CELCT, Trento, Italy.

Evelyne Tzoukermann and Mark Libermann. 1990. A finite-state morphological processor for spanish. In *Proc. of COLING'90*, pages 277–281.

Serge A. Yablonsky. 1999. Russian morphological analysis. In *Proc. of VEXTAL '99*, pages 83–90, Venice.

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

Appendix

```
LEXICON root
    l_adulto+NN:adult+NN SufNomO;
    l_verde:verd SufAggE;
    l_mangiare:mang+V_ARE SufVerbAre;

LEXICON SufNomO
    +MASC+SING:+MASC+SING⋆o #;
    +FEMM+SING:+FEMM+SING⋆a #;
    +MASC+PLUR:+MASC+PLUR⋆i #;
    +FEMM+PLUR:+FEMM+PLUR⋆e #;
    +MASC:+MASC SufModMasc;
    +FEMM:+FEMM SufModFemm;

LEXICON SufAggE
    +ADJ+MASC+SING:+ADJ+MASC+SING⋆e #;
    +ADJ+FEMM+SING:+ADJ+FEMM+SING⋆e #;
    +ADJ+MASC+PLUR:+ADJ+MASC+PLUR⋆i #;
    +ADJ+FEMM+PLUR:+ADJ+FEMM+PLUR⋆i #;
    +ADJ+MASC:+ADJ+MASC SufModMasc;
    +ADJ+FEMM:+ADJ+FEMM SufModFemm;
    +ADJ:+ADJ SufIssimo;
    +ADV+SUP:+ADV+SUP⋆issimamente #;

LEXICON SufVerbAre
    +V_NOFIN+INF+PRES:+INF+PRES⋆are #;
    +V_NOFIN+INF+PRES:+INF+PRES⋆ar SufClit;
    +V_FIN+IND+PRES+1+SING:+IND+PRES+1+SING⋆o #;
    +V_FIN+IND+PRES+2+SING:+IND+PRES+2+SING⋆i #;
    +V_FIN+IND+PRES+3+SING:+IND+PRES+3+SING⋆a #;
    +V_FIN+IND+PRES+1+PLUR:+IND+PRES+1+PLUR⋆iamo #;
    +V_FIN+IND+PRES+2+PLUR:+IND+PRES+2+PLUR⋆ate #;
    +V_FIN+IND+PRES+3+PLUR:+IND+PRES+3+PLUR⋆ano #;
    ...
```