Dealing with unknown words by simple decomposition: feasibility studies with Italian prefixes.

Bruno Cartoni
ISSCO/TIM ETI, University of Geneva
40 bd du Pont-d'Arve, 1211 Genève 4
bruno.cartoni@eti.unige.ch

Abstract

In this article, we present an experiment that aims to evaluate the feasibility of a superficial morphological analysis, to analyse unknown constructed neologisms. For any morphosyntactic analyser, lexical incompleteness is a real problem. This lack of information is partly due to lexical creativity, and more especially to the productivity of some morphological processes. We present here a set of word formation rules based on constructional morphology principles that can be used to improve the performance of an Italian morphosyntactic analyser. These rules use only simple computing techniques in order to ensure efficiency because any improvements in coverage must not slow down the entire system. In the second part of this paper, we describe a method for constraining the rules, and an evaluation of these constraints in terms of performance. Great improvements are achieved in reducing the number of incorrect analyses of unknown neologisms ("noise"), although this is at the cost of some increase in "silence" (correct analyses which are no longer produced). This classic trade-off between "noise" and "silence", however, can hardly be avoided and we believe that this experiment successfully demonstrates the feasibility of superficial analysis in improving performance and points the way to other avenues of research.

1. Introduction

Various NLP applications such as syntactic tagging rely on lexical resources, and lexical incompleteness can be problematic. Apart from proper nouns and spelling mistakes, most unknown words come from the lexical creativity of every natural language. This concept of creativity covers different linguistic phenomena, such as lexical borrowings, onomatopoeia. And the most interesting phenomenon is constructed neologisms - new words constructed using existing lexical items. In Italian, as in other European languages, the most frequent construction operation to create new words is suffixation, followed closely by prefixation (Iacobini, 2004). In this paper, we propose to use very simple Word Formation Rules (Aronoff, 1976) (hereafter WFR) to decompose and analyse unknown constructed words, in order to improve lexical coverage of a morphosyntactic analyser. We ran a set of WFRs on a large number of unknown words and manually evaluated the results. For the rules that did not provide satisfying results, we elaborated constraints, based on linguistic knowledge and technical considerations. We then evaluated the new constrained rules, both with and without applying these constraints, and we analysed the improvement in the performance of the analyser.

2. Reference Data And Corpus

The main aim of this experiment is to improve the lexical coverage of the morphosyntactic analyser Mmorph (Petitpierre & Russel, 1995) that is used by the Tatoo[1] tagger. We automatically compared the Italian lexicon of Mmorph with two large corpora [*ilSole24ore* [2], containing about 1.88 million occurrences, and the CoLFIScorpus[3],of about 136'884 types), and produced two lists of unknown lexical units (225'075 from the *ilSole24ore* corpus, and 58'926 from the CoLFIS one). We then excluded proper nouns from the lists (using a simple routine based on capitalisation) and we got two lists of 41'027 unknown words that can potentially be considered as neologisms (13'415 types from the *Ilsole24ore* corpus, and 27'612 types from the CoLFIS corpus).

3. WFRs to Analyse Unknown Words

We concentrated on prefixation operations for various reasons:

- after suffixation, prefixation is the most productive process for constructing new words in Italian;
- prefixes are a relatively closed class of affixes, and so can be easily listed;
- their allormorphic variations are well-known and so can be easily listed too;
- apart from a few exceptions, prefixes are intra-category, which mean that they don't change the grammatical function of the base, which is an asset for automatic analysis. If they do so, such behaviour is well-described in the linguistic literature and can be consequently easily formalised (see below).

In total, we identified 46 prefixes from theoretical references such as Iacobini 2004 (*a, anti, arci, auto, co, con, contro, de, dis, ex, extra, in, infra, inter, intra, iper, macro, macro, maxi, mega, meta, micro, mini, multi, neo, non, oltre, para, pluri, poli, post, pre, pro, retro, ri, s, semi,*

[1] The ISSCO Tagger Tool : http://issco-www.unige.ch/staff/robert/tatoo/tatoo.html

[2] http ://www.ilsole24ore.com/, corpus MLCC 1997, published by ELRA

[3] Most frequent Type of Italian, available on http://alphalinguistica.sns.it/CoLFIS/

sopra, sotto, sovra, sub, super, sur, trans, ultra, vice) for which we elaborated WFRs to analyse unknown prefixed words. For every allomorphic variation including, in some cases, the possibility of a hyphen between the prefix and the base, a different WFR had to be written, since we deal essentially with simple pattern matching of character strings.
So, for the 46 prefixes described above, we elaborated 72 WFRs of the following form:

```
WFR (X):
    Z/CAT = X/PREF[Y/base]
    Y/CAT ∈ Lit
```

where X is a prefix and Y a lexical element existing in the reference lexicon, whose grammatical category (CAT) corresponds to the one(s) acceptable for the prefix. For example, for the prefix *iper* (EN: *hyper*) that can only be placed before a noun or an adjective, we created the following rule:

```
WFR(iper):
    Z/CAT = iper/PREF[Y/base]
    Y/CAT ∈ Lit
    Y/CAT = Noun | Adj
```

This rule produces the analysis shown below for the prefix *iper* (*hyper*), in the syntax used in the Mmorph tool:

```
"iperinflazione" = "inflazione"
Noun[gender=feminine number=singular]
```

These WFRs are designed to be in a simple additional module to the analyser. That means that they can only exploit information from the analyser (category, gender and number) and the character string itself. Indeed, one of the essential features of this little module is portability, because we do not want to weigh down the global system. Behind this practical consideration, our linguistic motivation is to consider new words as transparent and unambiguous. Indeed, we think that neologisms cannot be ambiguous, because they need to be as transparent as possible to be produced and understood. We also wanted to investigate whether superficial analysis based on character strings would be enough to deal with constructed neologisms.

4. Results of Applying WFRs

Applying the set of rules to the corpus of unknown words produced 2820 analyses of neologisms constructed using almost every prefix from the list, with varying frequencies. We manually evaluated these analyses in terms of compositionality. A derived word can be described as *compositional* if its meaning is predictable from the meanings of its constituents (Apothéloz 2002).
We found that 1980 words were correctly analysed by these rules (which represents 70% of the total number of words analysed). This performance is **not** considered to be good enough, especially because it means that the 30 % of incorrect analyses might cause more problems than the 70% of the words which were correctly analysed.

However, every rule does not exhibit the same performance. So we evaluated each rule in terms of its performance (i.e. the percentage of correct analyses which it produced). In terms of distribution of the performance, we found the following results:

- 43 WFRs (for 31 prefixes) produced correct analyses in 100% of cases;
- 7 WFRs (for 7 prefixes) produced correct analyses in 90% and 99% of cases;
- 15 WFRs (for 14 prefixes) produced correct analyses in between 50 % and 89 % of cases;
- 7 WFRs (for 7 prefixes) produced correct analyses in less than 50 % of cases;

The two first groups of rules seem to be efficient enough. We also found that some morphosyntactic categories tend to be less ambiguous than others (for example, the WFR for *de + verb* produces 37 % of good analyses compared to the WFR *de + noun* that produces 63 %). Thanks to these rules, 1585 unknown words were correctly analysed. Obviously, the "good" performance of these rules has to be weighed against the number of occurrences extracted (100% of good analyses out of 2 occurrences extracted can hardly be called a significant result). The number occurrences analysed by a particular rule might also be an indicator of the "productivity" of a prefix.
We also noted that every rule which included a hyphen produced perfect analyses. Amongst prefixed words without hyphens, less than a half were always analysed correctly. So, observing this distribution of good performance, we can conclude that our superficial approach based on character strings is limited to those prefixes that are the most compositional and the least ambiguous.
However, for the 29 WFRs that produced analyses that were not always correct, it is interesting to consider what causes the problem (i.e. what provoke the bad analyses). We found that the main causes of incorrect analyses were:

- Homographic character strings between a "potential constructed word" and another lexical unit (for example, the foreign words *interest* was analysed as *inter + est*).
- Conflict with other constructional operations (*infrastrutturale* is not *infra+strutturale* but *infrastrutture + ale* – an adjective derived from *infrastructure*). It is also interesting to note that some of these wrong analyses (in terms of compositionality) do not, in the end, cause incorrect morphological analyses, but such cases are pure coincidences which we should not rely on.
- Incorrect analysis of badly spelled words, when they happen to be homographic with a potentially constructed word.
- Shortness of the prefix: We noticed that major problems appeared with short prefixes (the WFRs for mono-character prefixes such as *a* or *s*

produced very few good analyses – 22 % and 14 % respectively)

As a preliminary conclusion at this point, we might have decided that we should use only rules that has been validated on large corpora and that always give good results. Then we could simply not use WFRs that generate more problems than they solve. But some WFRs produced so few wrong analyses that we decided to find a new method to constrain WFRs, keeping in mind that they have to be portable and light, and not resource-consuming. We present this method below.

5. Morphological Structure of the Base: an Indicator of Constructivity

Iacobini (2004) states the semantic value of the base constrains the use of one or another prefix. For example, only verbs of action or process can be reiterate, so only those verbs can be prefixed by the reiteration prefix *ri*.
Moreover, this semantic value is sometimes shown in the suffix of the base.
This constraint can be easily implemented because suffixes can be simply expressed in terms of character strings, and then included in the WFRs. By examining linguistic surveys on prefixes, we isolated a number of constraints for the use of prefixes:

- The prefix *in* occurs very frequently with adjectives constructed with *–bile,* and consequently with nouns constructed with *–bilità,* like in *inguidabile* (EN: *undrivable*) or *ingovernabilità* (EN: ungovernability).
- The prefix *co-* appears frequently with process nouns (ending in*–ione, mento,* like *cofidenziamento* EN: *cofinancing*) or agentive nouns (ending in *–tore,* like *coproduttore* EN: *coproductor*)
- The deverbal prefixes *ri-* and *de-* create new verbs that can be frequently nominalised. The typical suffixes of nominalization of verbs in Italian are *–mento, -zione,* and *-aggio*. We can consequently assume that constructed words in *de-* or *ri-* are frequently nouns ending with such suffixes (like *riaffermazione* EN: *re-affirmating*, or *deconcentrazione* EN: *deconcentration*).

These phenomenon led us to believe that the « constructivity » of the base allows certain forms of prefixation. So we added some constraints to the WFRs specifying these particular endings to the base, as shown in the example below. Since our analysis is based only on character strings, we refer to this as an « indicator of constructivity ». Indeed, some endings are too well-established in the history of the formation of particular word to be considered as modern suffixes, but they still seem to be indicative of a specific semantic value, that allows only certain prefixes. Here below, we present a WFR for the prefix *ri* that can construct a noun that ends with the indicators of constructivity that are typical for process nouns (*-zione, -mento, -aggio*). Again, since we used only character strings, both plural and singular forms have to be declared:

```
WFR (ri):
     Z/Noun => ri/PREF [Y/Noun]
     Y/Noun = [a-z]*zione/i |
              [a-z]*mento/i |
              [a-z]*aggio/i
     Y/Noun ∈ Lit
```

We developed constraints for 5 WFRs dealing with the following prefixes and base types: *ri+noun, de+noun, in+noun, in+adjective,* and *co+noun* and then ran tests to evaluate their performance. This evaluation is described in the next section.

6. Evaluation

To evaluate the efficacy of the constraints we wanted to compare the performance of the rules with and without constraints. It was our assumption that to improve the performance of the analyser it is most important to reduce "noise" (wrong analyses). Actually, "silence" (the absence of an analysis of a word) would only reflect the *status quo* in the performance of the analyser.
For the 5 WFRs we found constraints for, we applied the rules with and without the constraints defined above to the data and then manually checked the compositionality of the analysed words.
In tables 1 and 2 below, we present the results achieved with and without the application of **the constraints of indicators of constructivity**. Each prefix listed in the tables stands for all its allomorphs, and in some cases, for occurrences with or without a hyphen.

	Analysed words	Correct analysis	%
ri- (+noun)	188	119	63 %
de- (+noun)	35	22	63 %
in- (+noun)	66	19	24 %
in- (+adj)	53	27	51 %
co- (+noun)	32	21	67 %

Table 1 : Performance of the rules, without indicator of constructivity

	Analysed words	Correct analysis	%
ri- (+noun)	64	62	97 %
de- (+noun)	18	18	100 %
in- (+noun)	11	11	100 %
in- (+adj)	9	9	100 %
co- (+noun)	7	7	100 %

Table 2: Performance of the rules, with indicator of constructivity

As we can see, we achieved great improvement in reducing noise thanks to these indicators of constructivity.

Indeed, the 4 of the 5 WFRs get 100% correct analyses with these constraints (table 2), although some of them hardly reached the 50 % without them (table 1). However silence - the number of words which received no analysis because of the constraints - has also increased.
This improvement (and the consequent reduction of the number of analysed words) was replicated when we applied the same constraints to a set of established neologisms found in Adamo and Della Valle dictionary (Adamo, G. and V. Della Valle, 2003), as shown in table 3 below:

	Prefixed Neologisms	Neologisms that match the constraints	%
ri- (+noun)	15	12	80 %
de- (+noun)	11	11	100 %
co- (+adj)	9	6	67 %

Table 3 : Performance of the constraints on established neologisms

In the context of improving the performance of a morphosyntactic analyser, it is most important to provide correct information, without adding new errors. Consequently, in evaluating this kind of method, we have concentrated on the issue of reducing noise, The good performance achieved by the application of the constrained rules (table 2) shows that the noise is clearly reduced. However, comparing the number of correctly analysed words of the two tables, we can see that constraining the rules has resulted in the loss of some occurrences (silence).

7. Conclusion

In this paper, we show that for a large number of prefixes, simple identification based on character strings is sufficient to improve the performance of a morphosyntactic analyser. We also show that the use of theoretical and practical information led us to elaborate a basic simple module that improves the performance of the application in terms of reducing noise.

The silence caused by the constraints has to be weighed against the number of words that would have been unknown otherwise. Moreover, prefixation is a very productive process, and listing all the possible prefixed new words is a time-consuming and almost impossible task. Within this perspective, finding simple rules that correctly analyse new formations is an interesting and promising field of research to deal with the incompleteness of lexical resources

In the future, we may want to address other constructional processes, but we are already aware that not every process is compatible with such simple treatment based on character strings. For example, suffixation provokes too many morphographemic variations (on the base and on the suffix itself) to be treated by simple decomposition. But other processes such as combining forms (Fradin 2000), especially those that are placed before the base, might be transparent enough, at least in terms of morphosyntactic analysis, to be decomposed in such manner.

8. References

Adamo, G. and V. Della Valle (2003), Neologismi quotidiani: un dizionario a cavallo del millennio 1998-2003. Lessico intellettuale europeo, Florence, Italie: Leo S. Olschki.

Apothéloz, D. (2003) La construction du lexique français, Paris: Orphys

Aronoff, M. (1976) Word Formation in Generative Grammar. Linguistic Inquiry Monographs. Cambridge: The MIT press.

Fradin, B. (2000). Combining Forms, Blends and related phenomena. In A. Thornton and U. Dolsescha (Eds.), Extragrammatical and Marginal Morphology (p. 11-59). Munich: Lincorn Europa.

Iacobini, C. (2004). I prefissi. In M. Grossmann and F. Rainer (Eds), La formazione delle parole in italiano (p. 99-163). Tübingen: Niemeyer

Petitpierre, D. and G. Russel (1995) Mmorph, The Multext Morphology. Issco (Technical Report): Genève.