Computer Science Journal of Moldova, vol.**16**, no.2(47), 2008

Specific features in automatic processing of the formations with prefixes*

Mircea Petic

Abstract

This article contains the information about the defining and analysis of some rules which will permit the automatic retrieving of the romanian prefixed formations and enriching the lexical resources by automatic derivation with the romanian prefixes *ne-* and *re-*. There are also described some peculiarities of the romanian prefixes.

Keywords: lexical resources, automatic prefixation, derivatives, algorithm for word extraction

Introduction

Linguistic electronic resources form the fundamental support for drawing up automatic tools for processing the linguistic information.

The need of the linguistic resources enriching is satisfied not only by word borrowing from other languages, but also by the use of some exclusively internal processes. Derivation means creation of a new word or a word with other meaning by adding some prefixes or suffixes to existent lexical bases. Prefixes are productive from the lexical point of view, when after removing the affix remains an attested word, and non-productive, when the remaining sequence after this operation is not attested as a word, though etymologically we recognize the prefix [1] in the word. The units from which the derivatives are formed, are called bases. In this article, as derivatives affixes, prefixes will be studied.

*The article is carried out as a part of the INTAS project, ref. nr. 05-104-7633

The scope of this article is the defining and analysis of some rules which will permit the automatic retrieving of the formations with prefixes and enriching the lexical resourses by automatic derivation with the romanian prefixes ***ne-*** *and* ***re-****.*

1 The classification of the prefixes

The 86 Romanian simple prefixes described in [2] can generate 5680 derivatives. These derivatives can be classified in four fundamental categories, such as:

- inherited derivatives from Latin, for example, (a) *închide* (in engl. (to) *close*), (a) *deschide* (in engl. (to) *open*), (a) *rămîne* (in engl. (to) *remain*);
- borrowing from other languages, for example, *deservire* (in engl. *service*), *nonsens* (in engl. *nonsense*), *prefabricat* (in engl. *prefab*);
- imitations by foreign models, for example, *concetăţean* (in engl. *fellow citizen*), (a) *întrevedea* (in engl. (to) *discern*);
- Romanian internal creations, for example, (a) *dezrobi* (in engl. (to) *emancipate*), *nefericit* (in engl. *unhappy*), (a) *înţărca* (in engl. (to) *wean*).

The compound prefixes obtained from two simple prefixes as a unique element of derivation are: *apar-*, *metem-* and *meten-*, *ram-* and *ran-*, *rim-* and *rin-*, *sco-*, *sper-*. It is worth mentioning that compound prefixes, analysed together with the wordbase and not relating to a formation prefixed before with one of the prefixes, we do not confuse with a double successive prefixation, for example, in *pre/străne*pot (in engl. *the son or daughter of the / great-grandson*), or (a) *re/înscrie* (in engl. (to) *re/write down*).

Developed prefixes, that result from combination of a simple prefix with a non-prefixal element, an insignificant fragment of a prefix or

a root, are: *destr-* (destrauri), *pres-* (presfira), *zăs-* (zăstimp, in engl. *period*). Sometimes it is difficult to distinguish a compound prefix from a developed one. Thus, a complex prefix *năs-*, based on a simple prefix *nă-*, may be considered compound with simple prefix *s-* (*z-*) in formations such as *năzbate* and developed in *năzduh* or *năzvăţa* (in engl. to *spoil*), where *z-* is recognized because of false analysis of the word *văzduh* (in engl. *air*) or (a) *dezvăţa* (in engl. (to) *unlearn*) [1].

As to the prefix's etymology, 12 prefixes are inherited from Latin (*des-*, *în-*, *stră-*), 13 – from Slavic (*ne-*, *răs-*), 18 came from ancient Greek, being taken by a Slavic or Latino-Romanic intermediate (*anti-*, *arhi-*, *hiper-*, *hipo-*) [3]. 29 prefixes with a multiple etymology are neologic inherited from Latin, French and/or Italian (*ante-*, *circum-*, *co(n)-*, *contra-*, *ex-*, *extra-*, *non-*, *post-*, *re-*, *ultra-*).

2 Some quantitative characteristics of the prefixes

Traditionally prefixes are described by a limited set of features: etymology, derivational model (with pointing the part of speech of the obtaining word and of its bases, for example, the verb (a) *dezdoi* (in engl. (to) *unbend*) is constructed from the numeral *doi* (in engl. *two*) and the prefix *dez-*) and the meaning of the obtaining word. Step by step derivational models began to be characterized by a set of qualitative attributives, such as: utility, regularity, productivity, frequency, etc. There aren't clear definitions of those notions. That is why we can only suppose the real meaning of those terms. Only formulating a set of quantitative definitions allow us to obtain some of the quantitative characteristics for the derivational affixes [4].

By number of the registered derivatives in [2], the most productive prefixes are *în-* (their number is 571), *ne-* (449), *des-* (395), *re-* (338).

In [5] it had been established some quantitative characteristics for these romanian prefixes, namely: number of words which begin with these prefixes, number of the derivatives with these prefixes, the repartition of letters which follow after the mentioned prefixes, the distri-

bution of the parts of speech (noun, adjective, verb) for every prefix mentioned above.

The romanian prefix *în-* can be attested in four different phonological forms, namely *in-*, *im-*, *în-* and *îm-*. One of them is a phonetic variant *îm-* before *b*, *p* consonant, for example, (a) *îmbătrîni* (in engl. (to) *grow old*), (a) *împărţi* (in engl. (to) *divide*), etc. Sometimes it is possible that the romanian prefixes *în-* and *îm-* are attested as the prefixes *in-* and *im-*. The romanian prefix *ne-* has no other literary phonetic variants. The prefix *des-* has the following literary phonetic variants: *dez-*, before the roots that begin with a (semi)-vowel or a voiced consonant and *de-* obvious before the roots that begin with *s*, *ş*. The prefix *re-* is a neologic prefix and sometimes substitues the ancient variant *ră-*. As a source for establishing these quantitative characteristics served [5], that contains about 30 thousands words divided in flection groups depending on their forming way. These words are grouped by part of speech, but the noun is classified also into gender. It permits to establish the announced quantitative characteristics mentioned above. These characteristics has been obtained by some programs developed in C programming language.

Taking into account all phonological forms of prefix *în-* for every form (*in-*, *im-*, *în-* and *îm-*), some specific quantitative characteristics have been found. Thus, it has been calculated the number of words that began with the particles *in-*, *im-*, *în-* and *îm-*. The total number of these words is 1145 [5].

The same situation is also with the prefix *re-* which has another phonological form *ră-*, but it is an old one. For the prefix *des-* it has been calculated separately the quantitative characteristics for phonologic forms *des-* and *dez-*. As the romanian prefix *ne-* has no phonological variants, in the calculations only this particle has been considered. The details of quantitative characteristics for these prefixes are presented in Table 1.

Taking into consideration the information obtained and arranged in Table 1 we can say that there are four letters, namely *ă*, *k*, *q* and *y*, before which the particles mentioned above have not been found. In addition, the words in which the presence of the prefix is recognized

Table 1. The quantitative characteristics for the phonetic variants of the studied prefixes [5]

pre-fixes	pho-netic vari-ants	num-ber of words	num-ber of deriva-tives	%	part of speech	total let-ters	conso-nants	vowels
in	în	583	107	18	V,N,A	21	17	4
	in	360	47	13	N,V,A	14	9	5
	im	126	15	12	N,V,A	7	2	5
	îm	79	19	24	V,N,A	2	2	0
ne		216	36	17	N,A,V	19	14	5
des	des	191	39	20	V,N,A	13	8	5
	dez	101	48	47	V,N,A	12	7	5
re		454	81	18	N,V,A	20	15	5
average				21				

where V denotes the verb, N – the noun and A – the adjective.

by relating it to a wordbase existent in Romanian language represent about a fifth part of the words that began with these particles. In general, more often we find consonant after the studied particles. The words that begin with the mentioned particles more often are nouns as a part of speech, then come the words that are verbs and finally adjectives. Actually, some prefixes contain more phonological forms and it makes more difficult to establish some quantitative characteristics.

3 Retrieving analysable formations with prefixes from a lexicon

For all categories of formations only the words in which the prefix presence was distinguished by referring to a wordbase existent in the Romanian language had been taken into considerations [2]. They can be grouped in derivatives of the type: *analyzable*, *semianalyzable*, and

non-analyzable. In the analyzable formations both are distinguished either the prefix or the wordbase. For example, in the word *poimîine* (in engl. *the day after tomorrow*), *poi-* is the prefix, and *mîine* (in engl. *tomorrow*) is an adverb (wordbase).

As a source for analyzable formations with simple prefix extracting a lexicon[1] had served which contains not only graphic representations of the words, but also its part of speech. This lexicon contains reusable resources of Romanian language (it has approximately 100000 of wordbases). It needs to be mentioned that a word can have several entrances for different parts of speech [7].

Besides the lexicon it was used a list of 86 simple prefixes that are registered in [2]. Also phonological forms of those prefixes had been added. Being important the peculiarity of those prefixes and derivatives described above the algorithm for automatic extraction of the analysable formations with simple prefixes had been developed. Lately it was implemented in a program written in C programming language.

Taking into account the peculiarity of the Romanian simple prefixes and of the available lexicon a simple but distinctively used for automatic extraction of the analysable formations with simple prefixes algorithm had been elaborated. The essence of the algorithm is the following: it reads a prefix and when listing the words from a lexicon, chooses those that can be integrated in the established constrains, namely if after removing the prefix it remains an attested in the dictionary word and its part of speech coincides with the initial one. Corresponding to this algorithm, it could be added also the instructions for counting both the analysable formations and of the words which begin with the particles that coincide with the respective prefixes [8].

Using the program, the analyzable formations with simple prefixes had been found and extracted for all prefixes mentioned in [2]. In Table 2 the concrete numbers of the analyzable formations for the ten most numerous prefixes obtained with described program comparatively with those in the source [2] are given. The results presented in this table are expectable for some prefixes, but for the other are really amazing.

[1]The lexicon is available on the site http://imi201.math.md/elrr/

Table 2. The number of analysable formations (NAF)

prefixes	NAF program	NAF source [2]
ne-	1500	449
re-	970	338
de-	702	250
in-	632	217
a-	610	276
s-	474	73
pre-	441	186
co-	338	19
con-	282	106
anti-	257	281

Verifying the analyzable formations with the prefixes *ne-* and *re-* we can say that the precision of a good word selection is very high, the cases of the word of the kind *rece* (in engl. *chill*) are singular. The same can be said also about the prefixes *pre-*, *con-* and *anti-*.

For the other prefixes many words had been obtained that were not analyzable formations with these prefixes, for example for the prefix *s-*: *scară* (in engl. *ladder*), *seră* (in engl. *hothouse*), and so on. But the number of analyzable formations with the prefix *anti-* is not greater that those registered in [2]. So it can be presumed that some of the words are not present in any of lexicographic sources.

In order to retrieve analysable formations with compound prefixes it was used the same algorithm described in [2] and a program in C programming language has been developed on its base.

The concrete data obtained by the mentioned above program for 10 compound prefixes, comparatively with those presented in [2], are given in Table 3. In this table we can bring into evidence the big number of analysable formations obtained by program comparatively to the number recorded in [2] for the compound prefixes *ra-* and *sco-*. It can be explained by confusing the prefixation with double successive

Table 3. The number of analysable formations (NAF)

the prefix	NAF program	NAF source [2]
ra-	64	2
sco-	22	2
ram-	4	1
apar-	4	1
ran-	3	1
rin-	3	1
sper-	2	1
metem-	1	1
meten-	1	1
rim-	0	1

prefixation with simple prefixes *a-* and respectively *co-* after that with *re-* (*r-*) and, respectively, *s-*. In other cases the expectable results were obtained, except the prefix *rim-* for which we didn't find the analysable formation *rimbomba*, which isn't present in the lexicon.

The obtained results suggest the idea that the algorithm for automatic extraction of the analysable formations with simple prefixes is not universal. It can be used in automatic extraction of the analyzable formations with long simple prefixes. For short simple prefixes the information about only graphic representation and part of speech proved to be insufficient for several simple prefixes. They need, probably, the semantic information about analyzable formations with such prefixes.

Although, we certainly can say that for long simple prefixes, the presented algorithm satisfies with a high precision the automatic extraction of the analyzable formations with simple prefixes. The results proved that the most numerous simple prefixes in analyzable formations registered in [2] coincide with the results of the program [9].

Nevertheless, the algorithm mentioned above will not help us to find the derivatives like: *deschis* (in engl. *opened*), *închis* (in engl. *closed*), (a) *combina* (in engl. (to) *combine*) and (a) *îmbina* (in engl. (to)

join), *interbelic* (in engl. *interwar*) and *antebelic* (in engl. *prewar*). These words aren't analysable formations. They are semi-analysable formations. The prefix derivatives of this type are the words in which we can distinguish only the prefix as opposed to the other prefixed formations or compound words with a common root, inexistent as an independent word. So the method by which we recognize the analysable formations is not valid for semianalysable formations.

4 Automatic prefixation

According to quantitative characteristics discussed in section 3, the more frequent ones are *în-*, *ne-*, *des-*, *re-*. As the prefixes *în-* and *des-* have more phonetic variants and need more information in order to formulate prefixation rules, afterwards we will consider only the affixation by the romanian prefixes *ne-* and *re-*.

4.1 The prefix *ne-*

According to [2] in The Dictionary of the Romanian language (DRL) there are registered 449 analysable formations, considering only one formation from a group (though its number is greater), having a common base: for example, *nemişcare* (in engl. *immobility*), but not *nemişcat* (in engl. *motionless*), *nemişcător* (in engl. *motionless*). Although, DRL has only a part of all existent formations with the prefix *ne-*, the number of derivatives from the DRL is small and it does not illustrate the real productivity of this romanian prefix (as a rule, the dictionary does not include obvious formations). Since the derivation with *ne-* forms an open system, it does not contain the formations from other sources.

It is important to note the preference of this prefix to the adjectives of participle origin (in our case that ends in: *-at*, *-it*, *-ut*, *-ît*, because the others face more difficulties in participle recognition), and also those derivatives with the suffixes *-tor*, *-bil*, *-os* in relation with verbal bases. In DRL, the formations from the adjectives derivated by other suffixes are, as a rule, fortuitous. The creations with *ne-* from the primary

adjective are, in general, not numerous and unusual.

The formations with *ne-* from the verbs are not numerous. The derivations from verbs, being not specific for the prefix *ne-*, can be from adjectives of participle origin or some derivatives with suffixes. The prefix *ne-* is grammaticalized in relation with gerund, participle and supine of the verbs, replacing completely the negation *nu* (engl. *not*).

The derivation with romanian prefix *ne-* constitues an open system. Wordbases to which this prefix can be attached are of various origin, where the prefix *ne-* can be attached, valid in literary language in general, it does not function in popular and familiar ones, where in spontaneous constructions, the prefix *ne-* can deny any kind of word. The prefix *ne-* is productive in all literary styles of the language.

Thus, from those mentioned above we can infer the following rules for the prefix *ne-*:

- from the adjectives derivated by suffixes *-tor*, *-bil*, *-os* we form adjectives derivated by the prefix *ne-*;
- from the participle ended with *-at*, *-it*, *-ut*, *-ît* we form adjectives derivated by prefix *ne-*;
- from the gerund we form adjectives derivated by prefix *ne-*.

4.2 The prefix *re-*

According to [2] in DRL the number of analysable derivatives with *re-* is 283. To these words we can add 55 derivatives from other sources (for example: Dictionary of neologisms, Morphological Dictionary, Actual Romanian Language, etc), the most numerous derivatives are those verbal obtained from verbs (analysable formations). Many of semi-analysable formations with the romanian prefix *re-* is related with those semianalysable derivatives with romanian prefixes *în-* (*in-*), *con-*.

The prefix *re-* is not attached, as a rule, to the words which begin with r+vowel [2]. In fact, there are some derivatives of this case: *reromanizare* (in engl. *reromanization*), *rerafinare* (in engl. *repurifying*), *reruralizare* (in engl. *reruralization*).

The wordbases at which the prefix *re-* is attached are of different origin, both ancient (latine, slave, hungarian, turkish, greek), and modern (neologic).

Actually, the romanian prefix *re-* can be attached to any verbal wordbase. Another observation is that the nouns formed from the present infinitive of the verb by adding the suffix *-re* form derivative by prefix *re-*. The derivatives with this prefix can be found in all literary styles of the Romanian language.

So, we can define the following rules for the romanian prefix *re-* from the infinitive of the verbs we form:

- verbs derivatives with prefix *re-*;
- nouns derivatives both with prefix *re-* and the suffix *-re*.

4.3 The methodology of automatic prefixation

The rules formulated above need knowledge only about graphic representation of a word and its part of speech. The source [6] contains 28932 words which are divided into flection groups depending on its way of forming. Dividing into a set of other flection groups, as it was observed from the rules above, does not influence the process of forming the derivatives with the romanian prefixes *ne-* and *re-*. It is important to mention that the obtained derivatives based on formulated rules above will inherit the flection group of the wordbase of the derivative, except only the case of noun forming from the verb.

Examining the words from the lexicon and concatenating the prefix respectively to those ones which correspond to the categories established by the rules above, an algorithm of analysable derivation with romanian prefixes *ne-* and *re-* is implemented in a unit capable to generate new words.

4.4 The results

The developed program based on the algorithm mentioned above enriched the lexicon with 417 derivatives with the romanian prefix *ne-*

and 9014 derivatives with the prefix *re-*. In Table 4 there are the results obtained for the analysable derivatives rules with the romanian prefixes *ne-* and *re-* for adjectives and verbs. Taking into consideration that the lexicon had already 216 words that began with *ne-*, the growth is about 1,9 times. In the same way the number of the words that begin with *re-* is 454, so the growth is 19 times. As a result, the growth of derivatives is 9430 words with the romanian prefixes *ne-* and *re-*, that represents about 1/3 of the initial lexicon. In the process of flection of these words with the help of the programs described in [10] we will obtain 250340 of flected forms.

Table 4. The results obtained for analysable derivation rules with romanian prefixes *ne-* and *re-* for adjectives and verbs

prefixes	part of speech	suffixes	nr. of derivatives
ne-	Adjective	-tor	72
		-os	46
		-bil	20
	Participle	-at, -it, -ut, -ît	261
	Gerund	-ind	17
re-	Verb		4507
	Nouns from verb	-re	4507
Total number of wordbases with prefixes *ne-* and *re-*			9430

Thouth the statistics presented in Table 4 is important, there are some moments that can be brought into the evidence. The number of the derivatives obtained by the romanian prefixes *ne-* and *re-* includes also those that are already contained in [5] and in DEX.

So, there is one derivative prefixed by *ne-* that is already included in [5] and 20 that are in DEX. In this case the number of true new words automatically generated is 397, that represents 95,43% of the initially generated words.

Analogous, the number of already existent words with romanian

prefix *re-* in [5] among those automatically generated is 27, and in DEX 298. Thus, the number of really new words is 8698, that constitutes 96,49% from those automatically generated by the romanian prefix *re-*.

If we will exclude the words that had already been prefixed with the affixes *ne-* and *re-* the number of the remained derivates will constitute 391 and, respectivelly, 8504. Although among the remained words there are also the ambigous ones. There are 362 perfectly valid words with prefix *ne-* and 7834 with prefix *re-*, that represent 91.18% and respectivelly 92.12% of new automatically generated words.

5 Conclusions

As only two romanian prefixes *ne-* and *re-* have been studied in new words generation, it proved their good productive properties in automatic affix derivation. In addition, those derivational rules, except the fact that the information about the part of speech and graphic representation is needed, require also the semantic and etymologic information of the wordbases to which they are attached. It was ascertained that in the existent lexicon it is difficult to distinguish the derivatives and the words that begin with the same sequence of letters of the correspondent prefix. In general, the derivatives generated by the correspondent rules would inherit the flection group from the derivatives wordbase, except the case when the verb is transformed in the noun.

References

[1] Iorgu Iordan. *Limba română contemporană*. Bucureşti, Editura Academiei, 1970.

[2] Formarea cuvintelor în limba română. (vol. Colectiv), Bucureşti, Editura Academiei, 1970.

[3] Al. Graur, Mioara Avram (redactori responsabili). *Formarea cuvintelor în limba română*, vol. II, Bucureşti, Editura Academiei, 1978.

[4] B.I. Bartkov, T.B. Bartkov, E.A. Golovatskaia, L.K. Karashchuk. *Qualitative derivatology of contemporary english prefixes of negation.* http://old.festu.ru/ru/structure/library/library/vologdin/v2000-I/173.htm (in russian)

[5] M. Petic. *Unele caracteristici cantitative ale celor mai productive prefixe în limba română.* In Proceedings of the Vth International Conference on Microelectronics and Computer Science, Chişinău, september 2007. pp. 161–164.

[6] A. Lombard, C. Gâdei. *Dictionnaire morphologique de la langue roumain.* Bucureşti, Editura Academiei, 1981, 232 p.

[7] Ciubotaru C, Cojocaru S.,.Boian E., Colesnicov A., Malahova L., Demidova V., Burlaca O. *Resurse Lingvistice Reutilizabile.* În Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii Române. Iaşi, noiembrie 2006, pp. 75–79.

[8] M. Petic. *Automatic extraction of the analysable formations with simple prefixes.* In Proceedings of the Second International Conference of Young Scientists Computer Science and Engineering-2007, Lvov, october 2007, pp. 215–217.

[9] M. Petic. *Prefixe compuse în formaţiile analizabile.* In International Conference of Young Researchers. Scientific Abstracts, Chişinău, november 2007, pp. 215.

[10] E. Boian, S. Cojocaru. *The inflexion regularities for the Romanian Language.* Computer Science Journal of Moldova, 4, 1(7), 1995, pp. 40–58.

M.Petic, Received June 12, 2008

M.Petic
Institute of Mathematics and Computer Science,
Academy of Sciences of Moldova
5, Academiei str., Chisinau,
MD 2028, Moldova
E–mail: *mirsha@math.md*