Memoirs of the Scientific Sections of the Romanian Academy
Tome XXXIV, 2011

GENERATIVE MECHANISMS OF ROMANIAN DERIVATIONAL MORPHOLOGY

MIRCEA PETIC

Institute of Mathematics and Computer Science, Academy of Sciences of Moldova,
Chisinau, Republic of Moldova
E-mail: mirsha@math.md

The article studies the problem of a Romanian morphologic derivative generator development. The starting point is a lexicon that provides storage of derivatives in a lexicon, which contains not only the graphical representation of derivatives but also their constituent morphemes. This allows studying and formulating rules that would generate derivatives taking into account certain restrictions. The article deals with different cases of derivation, such as: semi-analyzable derivation, regular derivation, the projection process of the derivation with prefixation on the following suffixation, and derivatives with prefixes im-/in.

Key words: automatic generation of derivatives, lexicon, prefixes, suffixes.

1. INTRODUCTION

Modern dictionaries face some shortcomings, becoming objects of research for lexicographers. Since dictionaries are constantly filled with new entries thanks to language development, the complete vocabulary development task remains practically impossible. Therefore, the automatic and/or semi-automatic completion of the linguistic resources with words, automatically generated relying on existing ones by exclusively internal processes (in particular, by derivation with prefixes and suffixes), is a significant source of vocabulary enrichment. One of the well known applications of automatic derivation system for Romanian language was FAVR in the Mac environment ELU, which aimed to complete coverage of the inflectional morphology. Then, prefixes and suffixes were described by means of lexical or grammatical paradigm. FAVR approach intended to use semantic information in order to describe and generate derivatives. In this article we will focus only on graphic representation of the word and not on its semantic information.

To automate the process of derivation, in this case, it is necessary:

- to establish rules that can be applied to stems in order to obtain new derivatives;
- to establish conditions in which these rules can be applied;

- if the above restrictions do not guarantee the correctness of the generated words - to develop and implement a validation mechanism.

To solve these problems we need to perform a preliminary study of the derivational process. For this purpose, several sources have been used as lexicographical support, namely, the electronic version of the dictionary of derivatives [1], www.dexonline.ro and RRTLN[1] (*Resurse Reutilizabile ale Tehnologiei Limbajului Natural*).

The aim of this article is to study specific particularities of the derivatives, to establish and simulate some generative mechanisms in the derivational morphology.

First of all, the electronic version of the dictionary is described, revealing the characteristics of the dictionary in terms of statistical data of the derivatives and their constituents. Next, a particular case of semi-analyzable derivation is given. It causes more questions about the way of its generation. Later, the situation of regular derivations was studied, which implies the generation of the derivatives by changing the word gender, generation of augmentatives and diminutives. A special section is dedicated to the projection process of the derivation with prefixation on the following suffixation. Finally, the peculiarities of the set features are given on the automatic generation of the derivatives with prefixes *im-*/*in-*.

Taking it into account, the novelty of this study consists in formulating several derivational rules, implementation of them into some modules, establishing the list of wrong generated derivatives and, in order to solve this discrepancy, the formulation of some restrictions or/and exceptions.

2. THE LEXICON OF DERIVATIVES

The lexicon represents the electronic variant of the dictionary derivatives, elaborated by S. Constantinescu [1]. The lexicon contains only the graphical representation of derivatives and their constituent morphemes, without any information about their part of speech and their stems. For easier processing of the lexicon entries, a regular expression was developed, which represents the following derivative structure:

```
derivat = (+morfem)*.morfem(-morfem)*
```

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

where `+morfem` represents a prefix, `.morfem` is a stem, and `−morfem` is a suffix. An example of an entry in the lexicon is:

```
antistatal=+anti.stat-al
reprogramabil=+re.programa-bil
```

In order to find out the statistical characteristics of the lexicon (Table 1), there were elaborated algorithms and developed corresponding programs [4].

Table 1

The statistical characteristics of the lexicon

Characteristics	**Number**
Derivatives	15300
Roots	6800
Prefixes	42
Suffixes	433

For lexicon processing by extracting derivative groups with the same affixes, there were found that a small group of prefixes and suffixes constitutes the majority of derivatives [4].

Thus, 12 of the 42 prefixes form 88.2% of all derivatives with prefixes, registered in the lexicon. The derivatives formed by the following prefixes are the most numerous (given in descending order): *ne-*, *re-*, *în-*, *des-*, *pre-*, *anti-*, *auto-*, *sub-*, *dez-*, *supra-*, *de-* and *îm-*.

Moreover, from 433 suffixes recorded in the lexicon, 52 represent 87.7% of all derivatives with suffixes. The most numerous derivatives have proved to be, in frequency descending order, the following suffixes: *-re*, *-tor*, *-toare*, *-eală*, *-ie*, *-ătoare*, *-iza*, *-oasă*, *-ar*, *-ător*, *-ească*, *-os*, *-aş*, *-esc*, *-tură*, *-iţă*, *-ist*, *-uţă*, *-el*, *-i*, *-ui*, *-ătură*, *-eşte*, *-ism*, *-a*, *-ărie*, *-ică*, *-ime*, *-itate*, *-ioară*, *-işor*, *-işoară*, *-ic*, *-uleţ*, *-că*, *-ean*, *-iş*, *-easă*, *-bil*, *-uţ*, *-at*, *-oaică*, *-uşor*, *-an*, *-oi*, *-uliţ*, *-iu*, *-enie*, *-istă*, *-al* and *-ea*.

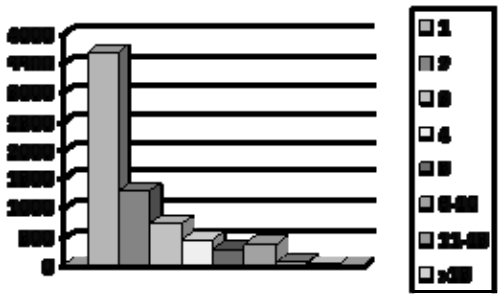

Fig. 1 - Statistical data referring the number of derivatives in the lexical families

It should be noted that not all derivated roots form a proportional number of derivatives [4]. Thus, there are words for which the maximum number of derivatives is recorded, namely: **bun** (32 derivatives), **alb** (25 derivatives), **şarpe** (22 derivatives), **roată** (22 derivatives), **om** (20 derivatives). In the lexicon, a large number of words (3657) appear with a single derivative (Fig. 1). It can be acknowledged that only a small number of roots can be really productive in the process of derivation.

So, there is a limited number of productive affixes, as there is a limited number of productive roots. That is why it will be useful to concentrate the attention on these roots and affixes.

3. PARTICULARITIES OF THE SEMI-ANALYZABLE DERIVATIVES

According to Iorgu Iordan in [3], derivatives in Romanian can be grouped into: analyzable, semi-analyzable and non-analyzable.

In *analyzable* derivatives, both prefix and stem are recognized. In the *semi-analyzable* derivatives, only the affix is recognized, in opposition to other derivatives or compounds (for example, *deschis – închis*). Some derivatives became *non-analyzable* because of the phonetic, morphological or semantic evolution, or because of the wordbase dissappearance.

The semianalyzable derivative detection seems to be a more complicated matter than of the *analyzable* ones, because the stems that have formed the derivatives are not already known. Thus, to study this process it is necessary to have the derivatives of this

type with the coresponding marking. Therefore, the lexicon [1], which contains derivatives with their morphemic structure is useful. Analyzing the content of the lexicon, there was found that it includes such derivatives, and the major part, the semianalysable derivatives, refers to *des-/în-* prefixes. Automatic extraction of these words was done, identifying 57 of such derivatives. So, some derivatives are important in order to clear up the main point of the derivational mechanism.

The particular characteristic for the semi-analyzable derivatives is the utilization of regressive derivation, namely the removal of an affix and not its addition. A good example is the word crucişătură, formed by the verb încrucişa and the suffixation with the –ătură. The correctness of this word formation process is confirmed by the lexicon entry:

crucişătură=.(în)crucişa-ătură

So, for this case was observed that the entry describes the formation process of the derivative. The verb *încrucişa* derived regressively by removing the prefix *în-* and adding of the suffix *–ătură*. The entry was marked as regressive derivation by the prefix taken between parentheses. On the www.dexonline.ro resource, the same information is given as follows:

[în]crucişa + suf. -ătură

where the markers are similar to the above ones (formation of derivative *crucişătură*). Beyond that word, there are other derivatives of the same type, for example, *cingătoare*, *chiondoreală*, *fundătură*, *greunatic*, *lăuntric*, *notătoare*, etc.

Another example of semi-analyzable derivatives are the derivatives with prefixes *des-*(*dez-*). For further examples, the following words prefixed in a semi-analyzable way are considered:

despăduri=+des.(îm)păduri
dezvălui=+dez.(în)vălui

Thus, one can say that there is an exchange of prefixes, mentioning that there are not such stems as *păduri*, and *vălui*. It is interesting that in the lexicon only verbs with semi-analyzable derivatives have been found, even besides the verb *despăduri* there is the noun *despădurire*, which is not semi-analyzable.

4. THE PROJECTION OF PREFIXATION ON THE FOLLOWING SUFFIXATION

The projection of derivatives represents a method of word formation of the prefixed words from the suffixed words of the same root. According to Spanish researchers [5], the Spanish verb *amortizar* can be derivated with the prefix *des-* obtaining *desamortizar*. Also, *amortizar* can be derivated with suffixes *–cion* and *–able*. So, the derivative with prefix *des-* can derivate with the suffixes *–cion* and *–able*. The hypothesis is that derivatives can inherit/project the derivatives with suffixes of the stem whose the prefixation was realized. So, the mechanism of derivative generation implies the following enounce: if R is a root of a word, S_i are the possible suffixes of the root R, namely $R \rightarrow RS_i$, and P is the corresponding prefix to the root R, namely $R \rightarrow PR$, then there exist suffixes S_i for which $R \rightarrow PRS_i$.

In the case of the Romanian language, for the stem *capitula*, we can find in www.dexonline.ro the following derivatives with suffixes: *capitulant*, *capitulantă*, *capitulard*, *capitulare*, *capitulaţie*. However, for re-prefixed word *recapitula* we can find the followind suffixed words: *recapitulare*, *recapitulaţie*. Thus, there are two derivatives carrying out the projection of the derivation by suffixes. Trying the projection of the derivative *capitulant* → *recapitulant*, we obtain an inexistent word.

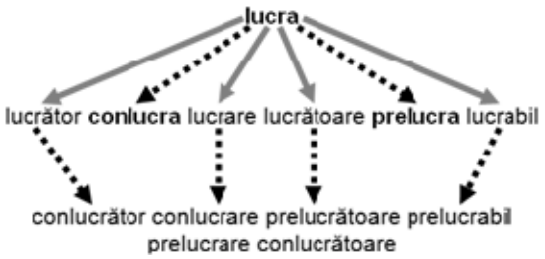

Fig. 2 - The derivative projection in the case of the word *lucra* for Romanian language

Starting from the existing lexicon with the morpheme structure [1], using some programs, there were extracted certain groups derived by projection means, registering 363 stems from which it is possible the projection of derivatives. Most of them consist only of a suffixed derivative and a prefixed one. However, we can argue that the method is useful in derivative generating for Romanian language. An example of such

derivation (Fig. 2) can be applied not only in the case o a single prefix, but there can be situations of several prefixes.

5. STUDYING DIMINUTIVES, AUGMENTATIVES, AND DERIVATION BY GENDER CHANGING

According to Serbian researchers, new Serbian words with predictable meanings were created using derivation. The predictable meanings represent the derivatives obtained by the amplification of the meaning, namely generation of the diminutives (*profesorčić*), augmentatives (*profesorčina*), and changing the word gender (male *professor* → female *professorka*) [2].

DIMINUTIVE PROCESSING

Diminutive suffixes in Romanian language are the following: *-aş* (copilaş), *-uc* (sătuc), *-el* (bătrânel), *-iţă* (fetiţă), *-uţă* (caruţă), *-ică* (floricică), *-uleţ* (ursuleţ), *-iş* (podiş), *-uţ* (căluţ). According to [1], there were established the number of derivatives with each diminutive suffix (Table 2). Beside this, all these suffixes need the vocalic and/or consonantal alterations in the process of derivation, for example, *s**a**t - s**ă**tuc, c**a**r - c**ă**ruţă, cal – c**ă**luţ,* the alternation being *a->ă; f**a**tă – f**e**tiţă: a ->e; fl**oa**re – fl**o**ricică: oa ->o; fr**at**e- fr**ăţ**ior: at -> ăţ* etc.

Table 2

The number of diminutive derivatives for concrete suffixes

Suffix	**Number of derivatives**
-aş	327
-iţă	249
-el	221
-uţă	208
-ică	139
-uleţ	104
-iş	101
-uţ	88
-uc	7

In the process of derivation of these suffixes, the most numerous class is the group of nouns, mentioning the following suffixes *–uc*, *-el*, *-aş*, *-iş*, *-uţ*, *-uţă*, *-iţă*, *-uleţ* and *-ică*. The less numerous are adjectives, in the case of the following suffixes: *-uc*, *-el*, *-iş*,

-uţ and *–ică.* The rarest case was observed with the suffix *-iş*, made up from the verb *zbura,* the adverb *zburiş*.

AUGMENTATIVE PROCESSING

Augmentative suffixes in Romanian language are the following: -andru (copilandru); -an (băietan); -oi/oaie (căsoi, căsoaie).

Table 3

The number of augmentatives for concrete suffixes

Suffix	**Number of derivatives**
-andru	4
-an	74
-oi	74
-oaie	26

According to the lexicon [1] mentioned above, there was established the number of derivatives for each augmentative suffix (Tabel 3). All these suffixes can be attached to nouns, to obtain new nouns. The suffixes *–an*, *-oi*, *-oaie* can be attached to adjectives in order to obtain augmentative words. Note that in the process of derivation with these suffixes were attested vocalic and/or consonantal alternations, for example, *c**a**să - c**ă**soi/c**ă**soaie,* the alternation is *a->ă, băie**t** – băie**ţ**andru t->ţ,* etc.

THE GENDER CHANGING OF THE WORDS

In the case of the Romanian language, the gender changing word can be achieved by switching to other corresponding suffixes, for example, *-tor* ↔ *-toare*, *-esc* ↔ *-ească*, etc. Thus, it was observed that the gender changing is made with the help of suffixation, not with prefixation.

The lexicon mentioned above consists of suffixed derivatives with *–tor*, *–toare,* and together with *–tor* and *–toare*. So, following the information from Figure 3, there are 148 words (nouns and/or adjectives) of the form ω′=ωtor, which could derive into the words of the form ω″=ωtoare. Similarly, there are 42 words (nouns and/or adjectives) of the form β′=βtoare which could derive into the words of the form β″=βtoare. Nevertheless, these 190 words generated in an automatic way should be validated. First of all, words were checked on their presence in RRTLN. 122 from all generated words

were present there. The remaining words were checked in electronic documents on Internet, and 49 of 68 derivatives have been validated. Thus 95% of the generated words were valid.

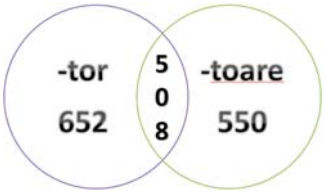

Fig. 3 - The number of the derivatives with the suffixes *–tor* and *-toare*

The same situation is with the pair of the suffixes *–esc* and *–eas*că. According to the same lexicon [1], it consists of 274 of derivatives with suffixes *–esc*, and 249 with the suffix *–eas*că. Note, that 229 of the derivatives are suffixes both with *-esc* and *–eas*că. It is natural to assume that the words (nouns and/or adjectives) of the form ω′=ω*esc*, could derivate into the words of the form ω″=ω*eas*că. Similarly, the words (nouns and/or adjectives) of the form β′=β*eas*că could derivate into the words of the form β″=β*toare*. Generating in an automatic way those derivatives which lack in the case of gender and checking them in an automatic way in the electronic documents, it was established that with the help of RRTLN there were validated 43 words of all 65 generated words. Another 12 of the 22 remaining derivatives were validated using a web application based on Google search engine opportunities. So, 84% of the obtained words were validated.

6. ASPECTS IN GENERATING THE DERIVATIVES WITH PREFIXES IN-/IM-

There are several classes of derivatives with the prefixes *in-*/*im-*. We will describe the case when the generated derivatives have the negative meaning. Romanian language has several prefixes that give the negative meaning to the derivatives, namely: *a-*, *i-*, *ne-* and *im-/in-*.

The derivatives with the prefixes *im-/in-*, as a rule, are adjectives, rarely nouns and verbs. The most numerous derivatives with prefix *in-/im-* are adjectives formed with the

suffix *–bil*, for example, *incurabil*, *inestimabil*, etc. So, being the adjectives of the form $\omega'=\omega bil$, they form derivatives of the form $\omega''=\beta\omega bil$, where $\beta \in \{in\text{-}, im\text{-}\}$.

Another well established group is that of adjectives derivated with the suffixes *–ent* and *–ant*: *inaderent*, *incoerent*, *independent*, etc [3]. Similar, by being the adjectives $\omega'=\omega\gamma$, they form derivatives $\omega''=\beta\omega\gamma$, where $\beta \in \{in\text{-}, im\text{-}\}$ and $\gamma \in \{\text{-}ent, \text{-}ant\}$. In both cases, the choice of the β depends on the first letter of the adjective ω, namely in the case when the letter is *b* or *p,* then $\beta=im\text{-}$, in other cases it is *in-*.

Other classes of derivatives with prefixes *in-/im-* are insignificant. Note, that there exist derivatives with *im-/in-* from the stems already derived with prefixes.

According to the linguistic resource www.dexonline.ro, there are about 4946 words that begin with the combination of letters *in,* and 1249 – with the combination of letters *im*. The number of those words ending with *ant* are corresponding to 38 and 13, with *ent* 61 and 12, and respectively *bil* 220 and 43, totally being 387 words. As some words can be several parts of speech, as it is the case of the words ending with *bil*, it was possible to filter and to obtain a number of 293 only adjectives.

Moreover, the adjectives with the suffix *–bil* form also derivatives with the prefix *ne-*, which offer the same negative meaning. In this way, it is useful to verify the derivatives with the help of the searching engine, for example www.google.com.

The case of derivation with the prefixes *in-/im-* from the words ending with *–ant* or *–ent* is more problematic, by the fact is that there are many words that would not form the derivatives with *in-/im-* because they are also nouns, not only adjectives.

The derivational lexicon [1] has only one derivative with prefixes *im-/in-,* namely *impermeabilizare*. So, it is possible to generate derivatives with the corresponding prefixes. In automatic way, it was established that the lexicon [1] consists of 62 derivatives with the suffix *–bil*, 1 derivative with the suffix *–ent,* and 37 with *–ant*.

Examining the words from the lexicon and concatenating the prefix respectively to those which correspond to the categories established by the rules above, without any vocalic or/and consonantal alternations, an algorithm of derivation with prefixes *im-/in-* was developed as a corresponding capable module to generate new words. As a result, 100 derivatives with the prefix *in-/im-* were obtained.

Initially, the generated words were checked whether they are present in the RRTLN. There was established that 7 of them are present. Other 93 have been verified in electronic documents from Internet with the help of the searching engine www.google.com. As a result, only 14 derivatives of the generated words have been found through the electronic documents. Four words, *inofertant*, *imprelucrabil*, *inrezolvabil*, *intrasabil*, have been found only one time. The data about other derivatives are presented in Table 4.

Table 4

The number of diminutive derivatives for concrete suffixes

Derivative	**Number of appearances**
Invindecabil	7
Indefrişabil	8
Indifuzabil	8
Infiltrabil	8
Indeşirabil	56
Insubstituibil	77
Imprecizabil	94
Injustificabil	181
Injucabil	353
Incadrabil	4710

In the case of the derivatives generated with prefixes *in-/im-* and with a small number of appearances though the electronic documents, it is clear that the attached prefix with the negative meaning was not applied correctly. This is the ambigous situation of the derivatives *inrezolvabil* (1) – *nerezolvabil* (1280), *insubstituibil* (77) – *nesubstituibil* (222), *injucabil* (353) - *nejucabil* (3050). However, there are derivatives with a large number of appearances that have another variant of the prefix with a small number of appearances: *inabordabil* (2810) - *neabordabil* (699), *inacceptabil* (67900) - *neacceptabil* (7140), *incalculabil* (24000) - *necalculabil* (469).

7. CONCLUSIONS

The article studied specific particularities of the derivatives, established and simulated some generative mechanisms in the derivational morphology.

Generation of derivatives is not a trivial problem, because the process does not have a regular mechanism. Solution to store all derivatives of a dictionary is a reasonable one

because these derivatives still will not cover the full diversity of the language, this being in continuous evolution. This article treated different cases of derivation, such as: semi-analyzable derivation, regular derivation, the projection process of the derivation with prefixation on the following suffixation and derivatives with prefixes *im-*/*in-*.

Nevertheless, the approach to generate constraint derivatives according to constraint rules for derived groups is a mechanism of over-generation, when the validation phase excludes many wrong formed words. That is why, it was established a list of wrong generated derivatives and in order to solve this discrepancy we formulated some restrictions or/and exceptions. Well defined rules will increase the level of the correct word generation.

REFERENCES

1. CONSTANTINESCU S., *Dicţionar de cuvinte derivate*, Editura Herra, Bucureşti, 2008.
2. DUŠKO V., KRSTEV C., *Derivational Morphology in a E-Dictionary of Serbian*, Proceedings of the 2nd Language & Technology Conference (Zygmunt Vetulani, ed.), Poznan, Poland, 2005, 139-143.
3. IORDAN I., *Limba română contemporană*, Bucureşti, 1970, 66-99.
4. PETIC M., *Automatic derivational morphology contribution to Romanian lexical acquisition*, Special issue: *Natural Language Processing and its Application. Research in Computing Science,* Mexico, 2010, vol. 46, ISSN: 1870-4069, 67-78.
5. SANTANA O., PEREZ J., CARRERAS F., RODRIGES G., *Suffixal and Prefixal Morpholexical Relationships of Spanish*, Lecture Notes in Artificial Intelligence, Springer-Verlag, 2004, 407-418.

Received August 9, 2010
Revised January 14, 2011