Automatic Derivational Morphology Contribution to Romanian Lexical Acquisition

Petic Mircea

Institute of Mathematics and Computer Science of the ASM,
5, Academies str., Chisinau, MD-2028, Republic of Moldova
mirsha@math.md

Abstract. The derivation with affixes is a method of vocabulary enrichment. The consciousness of the stages in the process of the lexicon enrichment by means of derivational morphology mechanisms will lead to the construction of the new derivatives automatic generator. Therefore the digital variant of the derivatives dictionary helps to overcome difficult situations in the process of the new word validation and the handling of the uncertain character of the affixes. In addition, the derivatives groups, the concrete consonant and vowel alternations and the lexical families can be established using the dictionary of derivatives.

1 Introduction

The linguistic resources represent the fundamental support for automatic tools development in the processing of linguistic information. The lexical acquisition constitutes one of the most important methods for lexical resources enrichment. The examination of the problems referring to automatization is one of the main aspects in the process of the linguistic resources creation.

The need of the lexical resources enrichment is satisfied not only by borrowings of words from other languages, but also by the use of some exclusively internal processes [1]. Inflection, derivation and compounding are the most useful ways of word formation in Romanian language. *Inflection* is the generation of word forms without changing their initial meaning. *Derivation* means the creation of a new word by adding some affixe(s) to existent lexical base. *Compounding* is made up of the words which exist and are independent or with the elements of thematic type. In this article the derivation is investigated.

The aim of this article is to study the derivational morphology mechanisms which permit the automatic lexical resources acquisition for Romanian language.

In this context the paper is structured in the following way. Firstly, the derivational particularities of Romanian language are described, and then the most important stages in the automatic derivation are the subject to review. A special description is dedicated to the methods of new word validation. The uncertain particularities of Romanian affixes are examined. The solution for the uncertainty was found in the digital variant of the Romanian derivatives dictionary. The new electronic resource can be used in the process of detection of the derivatives groups. Moreover, it permits the vowel and consonant alternations examination and the construction of lexical families in the derivation process.

Special issue: Natural Language Processing and its Applications.
Research in Computing Science 46, 2010, pp. 67-78

Received 22/11/09
Accepted 16/01/10
Final version 09/03/10

2 Derivational Particularities of Romanian Language

The majority of Romanian language derivational mechanisms are also common for other European languages, especially for those of Latin origin.

The *affix* represents any morpheme which remains beside the root, when a word is segmented. It includes *prefixes* (added before the root) and *suffixes* (added after the root) [2]. Romanian language has 86 prefixes [3] and more than 600 suffixes. There are two types of suffixes: lexical and grammatical. *Lexical suffix* is a group of letters (or a single letter) which is added after the root and forms a new word. *Grammatical suffix* is a group of letters (or a single letter) which is added after the stem and, as a result, a form of the same word is obtained [4]. Roots, prefixes and suffixes represent the *morphemes* [5].

The word formed by adding a prefix or a suffix is called *derivative*. For example, the derivative *frumuseţe*, consists of the root *frumos* and the suffix *-eţe*. Often suffixes and prefixes are not added directly to roots but to a lexical stem, which represents a root and, at least, a prefix or a suffix, for example, *străbătător* is formed from the prefix *stră-* and the stem *bătător*, which consists of the root *bate* and the suffix *-ător* [2].

The derivatives can be classified in three groups: analyzable, semi-analyzable, and non-analyzable. In the *analyzable* derivatives both the affix and the root are distinguished. *Semi-analyzable* derivatives represent the words in which only the affix is distinguished. In the *non-analyzable* derivatives we can not distinguish neither the affix nor the root [3].

3 Stages in the Automatic Derivation

In the process of the derivational morphology automatization it is important to examine the following steps: the analysis of the affixes, the elaboration of affixes detection algorithm, the formalization of the derivational rules and the validation of the new words generated by specific algorithms.

The analysis of the affixes consists in the establishing of its quantitative features. The the quantitative features for some prefixes and suffixes were set up according to the list of Romanian prefixes and suffixes. On the base of the elaborated algorithms [6], programs were developed that allowed to find out:

- the number of words which begin (end) with some prefixes (suffixes);
- the number of words derived from these affixes;
- the repartition of letters followed by the mentioned affixes;
- the part of speech distribution for some Romanian affixes.

Taking into account the obtained results, it was noticed that some prefixes and suffixes have more phonological forms. That is why it is difficult to set up the quantitative characteristics of the affixes [6].

As not all the words end (begin) with the same suffixes (prefixes), some algorithms were elaborated for enabling the automatic extraction of the derivatives from the lexicon . The elaborated algorithms took into account the fact that being $x, y \in \Sigma^+$, where Σ^+ is the set of all possible roots, and if $y = xv$ then v

is the suffix of y and if $y = ux$ then u is the prefix of y. In this context both y and x must be valid words in Romanian language, and u and v are strings that can be affixes attested for Romanian language [3]. The problem of consonant and vocalic alternations was neglected in the case of the algorithm derivatives extraction. This fact does not permit the exact detecting of all derivatives.

For a certain number of affixes was found out some derivational rules which allow the generation of new derived words unregistered in the dictionaries [7]. In the process of the completion of linguistic resources by automatic derivation appears a natural tendency of using the most frequent affixes. But, in fact, the use of most productive affixes become to be problematic because of their irregular behavior [7]. That is why those affixes which permit to formulate simpler rules of behavior without having many exceptions were taken for the research. As a consequence, these rules operate with prefixes *ne-* and *re-*, and also with suffixes *-tor* and *-bil* (the last being frequent in the derivational process with the prefix *ne-*). The lexical suffix *-iza* was also included in the research, having neological origin and being actually very productive in the strong relation with the lexical suffixes *-ism* and *-ist*.

After derivatives generation process, not all obtained words can be considered valid. The set of words needs to pass the step of validation, which represents an important level in the correct detection of the generated derivatives, thanks to the programs based on derivation rules. Also, it is possible to validate the words with the help of linguists, but it requires more time and there is the possibility of making mistakes.

On the other hand, validation can be done by checking the derived words presence in the electronic documents. In this situation it is possible to use verified electronic documents, such as different Romanian corpora. Unfortunately, the corpora have insufficient number of words, which can be used in new derivatives validation. Therefore, the simplest way is to work with the existent documents on Internet, which are considered to be unverified sources. In this case the difficulties appear in setting up the conditions which assure a valid word. So, it was attempted to determine the indices of frequency [7] for some Romanian affixes. Thus, a program which extracts the number of appearances of words in Google searching engine for Romanian language was elaborated.

4 Digital Variant of the Derivatives Dictionary

The derivatives dictionary [8] has only the graphic representation of the derivative and the constituent morphemes without any information about the part of speech of the derivatives and of its stems. The digital variant of the dictionary [8] is obtained after the process of scanning, OCR-izing and the correction of the original entries. This electronic variant of the dictionary [8] becomes important as it is difficult to establish the criterion for validation of new derived words. Nevertheless, it permits the detecting of the derivatives morphemes with its type (prefix, root and suffix) and constitutes an important electronic linguistic resource.

Practically, the inputs in this dictionary are constructed being based on an unsure scheme. It is not clear where the affixes and the root are. In order to exclude the uncertainty in the input of the digital variant of the dictionary, it was made a regular expression that represents the structure of the derivatives:

$$derivative = (+morpheme)^*.morpheme(-morpheme)^*$$

where $+morpheme$ is the prefix, $.morpheme$ is the root and $-morpheme$ is the suffix.

Table 1. The most frequent prefixes

Prefix	Number of distinct derivatives
ne-	571
în-	293
re-	281
des-	109
pre-	109

To find out the statistic characteristics of the dictionary algorithms were elaborated and then programs were developed. It was calculated that dictionary consists of 15.300 derivatives with 42 prefixes (Table 1), 433 suffixes (Table 2) and over 6800 roots.

Table 2. The most frequent suffixes

Suffix	Number of distinct derivatives
-re	2793
-tor	605
-toare	522
-ală	514
-ie	400

5 Uncertain Characters of the Affixes

One of the main problems concerning Romanian derivation is the uncertainty of morpheme boundaries. In many cases different people, or even the same person put in different situations, divides the same word into segments in different ways. Besides the segmentation in morphemes (for example: anti-rachet) or in the allomorph variants, a word form can be segmented in different ways [8]

such as: *syllables* (for example: an-ti-ra-che-tă), *sounds* (for example: a-n-t-i-r-a-ch-e-t-ă) and *letters* (for example: a-n-t-i-r-a-c-h-e-t-ă). All these types of segmentation have nothing to do with segmentation in morphemes, because they do not indicate the morpheme boundaries.

The segmentation in morphemes implies the detection of the morphemes with its types (root, prefix, and suffix). Unfortunately, not all morphemes are different. It was observed that the derivatives can contain roots which can be also suffixes or prefixes (for example: *-re*, *-tor*, *-os*, *-uşor*, *-uliţă*). Also, there are morphemes where the root, prefix and suffix coincide, for example, the morpheme an is a prefix in the word *an*istoric, in the word *an*işor is a root and in the word americ*an* is a suffix.

Taking into account this uncertain character of the morphemes boundaries within the derivatives, it appears a natural tendency of applying to the probabilistic method in measuring this uncertainty.

Let X be a discrete variable with the values $\{x_1, x_2, x_3, x_4, x_5, x_6\}$, where
x_1 represents the case when a string is a prefix, for example, *an*istoric;
x_2 – a part of a prefix, for example, *an*tevorbitor;
x_3 – a root, for example, *an*işor;
x_4 – a part of a root, for example, um*an*;
x_5 – a suffix, for example, americ*an*;
x_6 – a part of a suffix, for example, comedi*an*t.

Table 3. Prefixes with the lowest entropy

String	Number of prefixes	Number of parts of prefixes	Number of roots	Number of parts of roots	Number of suffixes	Number of parts of suffixes	Entropy
super	6	0	0	0	0	0	0.0000
ultra	12	0	0	0	0	0	0.0000
arhe	1	0	0	0	0	0	0.0000
com	1	0	0	63	0	0	0.1161
auto	87	0	0	6	0	0	0.3451

Let $p(x_k)$ be the probability of the x_k event. Of course, not all the affixes can correspond to all the values of the discrete variable X. Therefore, let consider the set:

$$Y = \{y_k | p(x_k) \neq 0, 1 \leqslant k \leqslant 6\}.$$

Let $H(Y)$ be the entropy of the Y:

$$H(Y) = -\sum_{i=1}^{6} p(y_i) \times log_2 p(y_i)$$

For example, the prefix *răs-* corresponds to three values of the discrete variable X. So, the set is $Y = \{y_1, y_2, y_4\}$. The respective probabilities are $p(y_1) = 0,34313, p(y_2) = 0,01960, p(y_4) = 0,63725$.

$$H(Y) = -(p(y_1) \times log_2 p(y_1) + p(y_2) \times log_2 p(y_2) + p(y_4) \times log_2 p(y_4)) = 1,05495$$

It is worth being mentioned that if the uncertainty is higher, it means that the number of cases will increase, so the entropy will increase too, otherwise it will decrease. For example for the prefix *ultra-*, as a result, it is used only as a prefix, that is why the entropy is 0 (Table 3). In the case of the prefix *an-* the entropy is 1,2322 (Table 4), because it can correspond to all six values of the discrete variable X.

Table 4. Prefixes with the highest entropy

String	Number of prefixes	Number of parts of prefixes	Number of roots	Number of parts of roots	Number of suffixes	Number of parts of suffixes	Entropy
re	281	118	0	4708	3218	975	1.6005
intra	2	0	2	1	0	0	1.5219
an	1	105	1	1058	74	200	1.2322
de	71	209	0	547	0	0	1.2000
auto	571	0	0	426	0	39	1.1790

6 The Analysis of the Consonant and Vowel Alternations

The problem of derivation consists not only in the detection of the derivational rules for separate affixes, but also in the examination of the concrete consonant and vowel alternations for the affixes. It is important that not all affixes need vowel and consonant alternations in the process of derivation. On the purpose of precising which affixes have alternations in the process of derivation the digital variant of the derivatives dictionary was studied (Table 5).

The situation is that there are more derivatives without alternations, especially in the case of the prefix derivation. The lack of vowel and consonant alternations in the process of derivation is observed with the following most frequent prefixes: *ne-*, *re-*, *pre-*, *anti-*, *auto-*, *supra-*, and *de-*. The prefixes *în-*, *des-*, *sub-*, *dez-*, and *îm-* use the vowel and consonant alternations in the process of derivation (Table 6).

There are several types of vowel and consonant alternations in the process of derivation with the prefixes:

Table 5. Statistics about the derivatives

Affix	Number of derivatives witout alternations	Number of derivatives with alternations
prefix	1134	224
suffix	6809	6381
prefix and suffix	632	191
total	8575	6796

– the addition of a letter to the end of the root, for example, *şurub* → *înşuruba, bold* → *îmboldi, plin* → *împlini*;

– the changing of the final letter in the root, for example, *lînă* → *dezlîna, purpură* → *împurpura, puşcă* → *împuşca*;

– the changing of the final letter in the root and the addition of the letter to the end of the root, for example, *avut* → *înavuţi, compus* → *descompune, păros* → *împăroşa, blînd* → *îmblînzi*;

– the changing of the vowels in the root, for example, *cataramă* → *încătărăma, primăvară* → *desprimăvăra, rădăcină* → *dezrădăcina, platoşă* → *împlătoşi*;

– the changing in the prefix, for example, *şoca* → *deşoca, pat* → *supat*;

– the avoiding of the double consonant, for example, *spinteca* → *despinteca, braţ* → *subraţ*.

Table 6. Prefixes with vowel and consonant alternations

Prefix	Number of derivatives witout alternations	Number of derivatives with alternations	Total number of derivatives
în-	33	115	148
des-	57	6	63
sub-	81	2	83
dez-	32	2	34
îm-	3	38	41
total	206	163	369

In the most of cases in the process of derivation with prefixes *în-* and *îm-* the alternations are used because of the part of speech changing, especially from adjectives and nouns to verbs.

The process of derivation with suffixes does not attest cases without consonant and vowel alternations. It means that there are situations when the derivation is made up with minimum number of alternations (Table 7) and with maximum cases of changes in the root (Table 8). The possible vowel and consonant alternations are so varied that it is difficult to describe them all in a chapter, but it is possible, at least, to classify them:

Table 7. Suffixes with fewer derivatives formed without alternations

Suffix	Number of derivatives witout alternations	Number of derivatives with alternations	Total number of derivatives
-re	2782	11	2793
-tor	561	43	605
-toare	478	35	522
-iza	221	23	244
-tură	205	27	232

– the changing of the final letter in the root, for example, *alinia* → ***aliniere***, *aşchia* → ***aşchietor***, *cumpăra* → ***cumpărător***, *curăţi* → ***curăţător***, *delăsa* → ***delăsător***, *depune* → ***depunător***, *faianţă* → ***faianţator***, *fărîma* → ***fărîmător***, *împinge* → ***împingător***, *transcrie* → ***transcriitor***;

– the removing of the last vowel in the root, for example, *răşchia* → ***răşchitor***, *acri* → ***acreală***, *aduna* → ***adunătoare***;

– the removing of the final vowel in the root and the changing of the letter before the last one, for example, *zeflemea* → ***zeflemitor***, *ascunde* → ***ascunzătoare***;

Table 8. Suffixes with fewer derivatives formed without alternation

Suffix	Number of derivatives witout alternations	Number of derivatives with alternations	Total number of derivatives
-eală	19	495	514
-ătoare	5	281	286
-ător	5	279	284
-ar	110	249	359
-ie	166	234	400

– the changing of two final letters in the root, for example, *bea* → ***băutor***, *bea* → ***băutoare***, *încăpea* → ***încăpătoare***;

– the changing of the first letter of the suffix, for example, *bîntui* → ***bîntuială***, *murui* → ***muruială***;

– the removing of the final letter in the root with the vowel changing, for example, *cană* → ***căneală***, *atrage* → ***atrăgătoare***, *bate* → ***bătătoare***;

– the removing of the two letters in the suffix, for example, *căpia* → ***căpială***, *încleia* → ***încleială***;

– the removing of the final letter and that of the vowel inside the root, for example, *coace* → ***cocătoare***;

– the removing of the final vowel and the changing of the final consonant, for example, *descreşte* → *descrescătoare, închide* → *închizătoare, încrede* → *încrezătoare, promite* → *primiţătoare*;

– the changing in the root, for example, *rîde* → *rîzătoare, recunoaşte* → *recunoscătoare, roade* → *rozătoare, sta* → *stătătoare, şedea* → *şezătoare, vedea* → *văzătoare, şti* → *ştiutor.*

7 Lexical Families

The set of derivatives with a common root and meaning represents a lexical family. So, the second part of the definition is very important because there is a tendency of grouping the words in lexical families only by a common root. Therefore the following words: *alb*, *albastru*, and *albanez* can be considered lexical family, so ignoring the fact that lexical family consists of the words that have the appropriate meaning. In addition, a lexical family consists of the words which are different in what concern the grammatical categories, beginning with the same root. The word base can have the same form of the root for all words from the family, for example: *actor actoraş, actoricesc, actorie.*

When in a derivative the affixes are trimmed, it means that only the root remains. The root can suffer little alternations: *ţară* (*ţar-/ ţăr-*); but it is never changed. During the derivation, it was observed that not all the lexical units derive directly from the root, some of them derive from previous derivatives, for example,

$$[\text{cruce}]_{noun} \rightarrow [\text{cruciş}]_{adj,adv} \rightarrow [\text{încrucişa}]_{verb} \rightarrow [\text{încrucişator}]_{adj,noun}.$$

Lexical family consists also of compounds, which contain the word base from the respective families. So, the compound word *cal-de-mare* is in the lexical family of the word *cal* and also in that of the word *mare.* But, it will not be in the lexical family of the word *călare*, which is the word base of the other family: *călari, călarie, călarime, călaresc, călaraş.*

In the electronic variant of the derivatives dictionary the most numerous lexical families are of the root **bun** (32 derivatives with the prefixes: *stră-, îm-, ne- and în-*; and the suffixes: *-el, -eţe, -ătate, -ic, -uţă, -icea, -icel, -icică, -işoară, -işor, -iţă, -uţ, -re, -i, -ariţă, -atic, -eală, -ească, -esc, -eşte, -toare, -tor*, and *-ie*); **alb** (25 derivatives with the prefix: *în-*; and with the suffixes: *-eaţă, -ei, -eţ, -eală, -icioasă, -icios,-iliţă, -re, -i, -ime, -ineaţă, -ineţ, -ior, -işor, -itoare, -itor, -ie, -itură, -iţă, -ui, -uie, -uleţ, -uş*, and *-uţ*); **şarpe** (22 derivatives without any prefixes and with the suffixes: *-ar, -aş, -ărie, -ească, -esc, -eşte,-işor, -oaică, -oaie, -oi, -ui, -eală, -re, -toare, -tor, -tură, -urel*, and *-uşor*); **roată** (22 derivatives without any prefixes and with the suffixes: *-ar, -easă, -ie, -it, -iţă, -aş,-at, -ată, -i, -cică, -re, -tor, -toare, -tură, -ilă, -at, -iş, -ocoală*, and *-ocol*); **om** (20 derivatives with the prefixes: *ne-* and *supra-*; and with the suffixes: *-ime, -oasă, -os, -oi, -uleţ, -uşor, -ească, -esc, -eşte*, and *-ie*). In the same context there are over 3000 roots with a single derivative.

There were found out that 7 prefixes and namely *a-*, *arhe-*, *para-*, *dis-*, *i-*, *im* and *întru-*, are not attached directly to roots, but only to stems. Also there are several suffixes, that are not attached to root.

8 The Process of Derivative Groups Establishing

8.1 Derivative Groups with Prefixes

The derivatives were extracted separately for every affix and were compared with the flection groups. Thanks to flection groups from [9, 10] and derivatives from morphological dictionary [8] it was attempted to detect the derivation groups. The morphological dictionary [8] consists already of many derivatives.

So, in order to acquire information referring to affixes, which can be attached to roots from the dictionary [9] special programs were developed. The derivatives those which have the following prefixes proved to be the most numerous, in a descending order: *ne-*, *re-*, *în-*, *des-*, *pre-*, *anti-*, *auto-*, *sub-*, *dez-*, *supra-*, *de-*, and *îm-*. The rest of the prefixes have an unsignificant number of derivatives. These 12 prefixes from 42 form 88.2 per cent from all derivatives with prefixes, registered in this electronic dictionary.

Firstly, it was set up the flection groups of the roots, which correspond to derivatives with prefixes without any suffixes. Nevertheless, there is a big number of flection groups for a single prefix, for example,

$$\begin{aligned}
&\texttt{nescris=+ne.scris}_{N29}\\
&\texttt{nescris=+ne.scris}_{N1}\\
&\texttt{nescris=+ne.scris}_{N24}\\
&\texttt{nescris=+ne.scris}_{A4}\\
&\texttt{nescris=+ne.scris}_{M6}.
\end{aligned}$$

For every prefix was set up the most frequent flection group of the derivatives roots. So, the roots with verbal flection group V201 is attached mostly to the following prefixes: *re-*, *des-*, *auto-*, *dez-*, *de-*; masculine noun flection group M1: *anti-*, *sub-*, *îm-*; feminine noun flection group F1: *în-*, *supra-*; verbal flection group V401: *pre-*; adjectival flection group A2: *ne-*.

Secondly, it was extracted the flection group of the roots that correspond to derivatives with prefixes that was first derivated with suffixes. In this case, there is the probability that there are roots with the same flection groups that form derivatives with different suffixes, for example,

$$\begin{aligned}
&\texttt{subordonator=+sub.ordona}_{V201}\texttt{-tor}\\
&\texttt{subordonatoare=+sub.ordona}_{V201}\texttt{-toare}.
\end{aligned}$$

In the same procedure as that described above the derivatives are used with the most frequent flection groups of the derivatives roots: verbal flection group V201 uses mostly the following prefixes: *pre-*, *dez-*, *auto-*, *sub-*, *de-*, *supra-*; verbal flection group V401: *ne-*, *des-*; masculine noun flection group M1: *îm-*; adjectival flection group A1: *anti-*; neutral noun flection group N24: *în-*.

During the investigation there were found out that the following prefixes are attached especially to the nouns: *ne-, în-, anti-, sub-, supra* and *în-*. Another group of prefixes is in the most of cases attached to the verbs: *de-, dez-, auto-, pre-, des-* and *re-*.

8.2 Derivative Groups with Suffixes

In the same way as for the prefixes, in order to decide to which roots can be attached the concrete suffix from the morphological dictionary, special programs were developed, which extracted derivatives separately for every suffix, and after that, it is compared with the flection group.

The most numerous derivatives proved to be, in a 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*. The rest of the suffixes have an insignificant number of derivatives. From 430 suffixes registered in this electronic dictionary 52 of them form 87.7 per cent from suffix derivatives.

It was retrieved the flection group of the roots that correspond to derivatives with suffixes. The words in the dictionary [9] have several entries for different flection groups, for example, the verb *învîrti* takes part from verbal flection groups `V401` and `V305` (the same part of speech), and the word *croi* belongs to different parts of speech: noun flection group `N67` and verbal flection group `V408`.

For every suffix it was established the most frequent flection groups for the derivatives roots. So, from the roots with the masculine noun flection group `M1` flection group can be generated the derivatives with the following suffixes: *-ie, -ească, -aş, -esc, -iţă, -el, -i, -eşte, -ism, -ime, -ic, -iş, -oaică, -an, -oi, -iu, -uţ*; neutral noun flection group `N24`: *-oasă, -os, -arie, -işor, -uleţ, -ean, -uţ, -uşor, -al*; feminine noun flection group `F1`: *-ar, -uţă, -ică, -işoară, -uliţă*; verbal flection group `V401`: *-tor, -toare, -eală, -tora, -enie*; verbal flection group `V201`: *-re, -ătoare, -ător, -ătură, -bil*, `N1`: *-ist, -at, -al* (also neutral noun flection group `N24`), adjectival flection group `A1`: *-iza, -itate*, feminine noun flection group `F135`: *-ioară*, masculine noun flection group `M20`: *-că*, neutral noun flection group `N11`: *-istă*, feminine noun flection group `F43`: *-ea*.

After the examinations of the most frequent suffixes it was found that the following suffixes are attached mostly to the nouns: *-ie, -iza, -oasă, -ar, -ească, -os, -aş, -esc, -tură, -iţă, -ist, -uţă, -el, -i, -ui, -eşte, -ism, -a, -ărie, -ică, -ioară, -işor, -işoară, -ic, -uleţ, -că, -ean, -iş, -easă, -uş, -at, -oaică, -uşor, -an, -oi, -uliţă, -iu, -enie, -istă, -al, -ea*, and *uţ*. Another group of suffixes, and namely: *-re, -tor, -toare, -eală, -ătoare, -ător, -tură, -ătură*, and *-bil*, is attached mostly to verbs. Only two suffixes are attached in the most of cases to adjectives: *-ime* and *-itate*.

9 Conclusions

The process of the derivatives generator construction needs the detailed studying of the affixes and the derivatives features. The new derivatives validation is one of the steps in automatic derivation that raises many questions.

In the case it is difficult to set up the criterion for words validation by means of Internet, it is important to use the digital variant of the derivatives dictionary, which will permit the establishing of the morphemes of the derivatives with its type (prefix, root and suffix). It was proved to be useful in the detection of entire variety of consonant and vowel alternations in the process of derivation with prefixes and suffixes.

Another notion connected with the lexical derivation is the lexical family, that offers the possibility of acquiring sets of affixes possible to attach to the concrete roots.

The grouping of the derivatives by flection classes can give the possibility of finding the morphological characteristics of the words. The following step in the research should answer whether the most numerous flection groups for affixes can serve for the process of automatic generation of new valid derivatives.

References

1. Tufiş, D., Barbu, A.M., Revealing Translator's Knowledge: Statistical Methods in Constructing Practical Translation Lexicons for Language and Speech Processing, *International Journal of Speech Technology* 5, 2002, pp. 199-209.
2. Hristea, T., *Sinteze de limba română*, Bucureşti, 1984, pp. 66-99.
3. Graur, Al., Avram, M., *Formarea cuvintelor în limba română*, vol. II Editura Academiei, Bucureşti, 1978.
4. *Gramatica limbii române. Cuvîntul*, Editura Academiei Române, Bucureşti, 2005.
5. Hausser, R., *Foundations of Computational Linguistics. Human-Computer Communication in Natural Language*, 2nd Edition, Revised and Extended, Springer, 2001, 580 p.
6. Petic, M., Specific features in automatic processing of formations with prefixes, *Computer Science Journal of Moldova*, 4 1(7), 2008, pp. 209-222.
7. Cojocaru, S., Boian, E., Petic, M., Stages in automatic derivational morphology processing, *International Conference on Knowledge Engineering, Principles and Techniques, KEPT2009, Selected Papers*, Cluj-Napoca, July 24, 2009, pp. 97-104.
8. Constantinescu, S., *Dicţionar de cuvinte derivate. Editura Herra*, Bucureşti, 2008.
9. Lombard, A., Gâdei, C., *Dictionnaire morphologique de la langue roumaine*, Editura Academiei, Bucureşti, 1981, 232 p.
10. S.Cojocaru, M.Evstunin, V.Ufnarovski, Detecting and correcting spelling errors for the Roumanian language. *Computer Science Journal of Moldova*, vol.1, no.1(1), 1993, pp 3-21.