DEVELOPMENT OF TOOLS TO INFLECT COMPOUND WORDS IN ROMANIAN

Elena Boian[1], S. Cojocaru[1], A. Colesnicov[1], L.Malahova[1*], Tatiana Baltaga[2]

Institute of Mathematics and Computer Science, Academy of Sciences of Moldova[1]
State University of Moldova, Department of Mathematics and Information Technologies[2]
e-mail: baltaga@usm.md, lena@math.md sveta@math.md kae@math.md mal@math.md

ABSTRACT

One of basic functions at management of electronic dictionaries is its population with new words. The population of words can be manual when word by word is entered with morphologic attributes, or automated when a special tool is used to inflect simple and compound words. An approach to compound words inflexion in Romanian is discussed.

Keywords: morphologic dictionary, compound words, automated inflexion

INTRODUCTION

Morphologic dictionary contains words of a natural language with the attributes that are specific for this language morphology. Specific attributes make the essence of a morphologic dictionary. To generate the actual paradigm we use a program for automatic inflexion. The dictionary serves a primary open base to set new attributes for different domains of a natural language [1-2]. The dictionary structure for a highly inflective language is a difficult problem

One of basic functions at management of electronic dictionaries is its population with new words [3]. The population of words can be manual when word by word is entered with morphologic attributes (we can call it also the declarative mode), or automated when a special tool is used to inflect simple and compound words (procedural mode) [4–5]. The problem of the declarative mode is the necessity to have all word-forms ready for each new word. The procedural method can meet a problem if the inflexion rules for a new word were not programmed. The manual population is very difficult, because we need to enter word-forms and the corresponding morphological information. It is therefore naturally to try to automate the process.

The word inflexion program facilitates dictionary population. Information concerning part of speech (noun, verb, etc.) and maybe something else (e.g., gender for nouns) can be obtained

interactively. Then deep linguistic analysis with possible additional questions (usually about the alternation or suffixes) permits to predict the structure of the forms.

Special attention is given to nouns and adjectives declination, and verbs conjugation because they generate a lot of flexions. Pronouns, articles and numerals are not so numerous and may be entered in the dictionary manually. The problem of the complete formalization of the inflection process is very difficult, and we solved it partially. We ask the user for additional information in some special cases. This information concerns morphological categories, and consonant and vowel alternations.

A part that was not implemented till now is the automated inflexion of compound Romanian words.

Methods of word composition in Romanian are **prefixing, suffixing** and **colliding [6–8]**. Prefixing and suffixing form simple words. These words are inflected by the existing word inflexion program [4–5].

Another mode of word formation is **colliding.** The collided word units in itself constructive elements that were combined during language evolution, or two or more words, which produced a new word with a new contents [6-8].

The present word supposes as its purpose further development of existing tools of word inflexion for Romanian [4–5] and their generalization for compound words, and the use of the developed tools for population of a Romanian lexicon with word-forms and morphological information for compound words [1-2].

Word colliding

We mentioned above that word **colliding** is made in two modes. Table 1 shows the first mode, when a word is varied by additional auxiliary elements. The second mode is a colliding of two different words with their sense making a new word with a new content (Table 2).

Table 1. *The first mode of compound word formation*

Part of speech	Constructive elements	Examples
Common nouns	words that has in their structure elements, which can not be used independently: *aero-, balneo-, micro-, neuro-, -cid, -naut, etc.*	*aerodrom, balneolog microfon, neurology, genocid, astronaut,*
	composed of two or more words where only a final component is inflected	*binecuvântare*
	composed of two or more words	*metaloplasiie*
Adjective	composed of existing words or radixes: adverb+ adjective *ne+ mai+* adjective	*binefăcător* *nemaiauzit, nemaivăzut*
	composed of an auxiliary element and an existing word: *mono-, uni-, bi-, multi-,poli-, semi-*	*monocelular, monofazic, univalent, univoc, bianual, bisilabic multimi-lenar, multinaţional, polisemantic, politehnic, semiovscur, semioficial*
Verb	composed of *bine+* verb	*binecuvînta*

	composed of *auto*+ verb	*autoadministra*
	composed of *tele*+ verb	*telecomanda*
	composed of adverb *nemai*+ participle or gerund	*Nemaicunoscutt, nemaipomenit, nemaiauzind, etc.*
Numeral	formed from two or more words	*noiăsprezece, treizeci*
Pronoun	words that has in their structure elements, which can not be used independently:	*dumneata, dumnealui, dumneavoastră* (polite personal pronouns); *sieşi, sineşi* (reflexive pronouns); *însumi, însuţi, însuşi, ş.a.* (identity pronouns); *acelaşi, aceeaşi; ş.a.* (demonstrative pronouns)*; careva, vreun, vreunul, vreuna, vreo* (indefinite pronouns)
Adverb	preposition + adverb:	*adeseori, apururea, deasupra, înapoi*
	preposition + cardinal numeral :	*acasă, alături, alene, alocuri, deopotrivă, departe, deseară, devale, devreme, pesemne, etc.*
	prepositions *în* + *de* + another part of speech	*Intruna, delaolaltă, laolaltă; îndeaproape, îndeajuns, îndelung, înaemînă, îndeobşte, îndeosebi, îndeseară, etc.*
	adjective + noun	*deseori, rareori*
	pronoun + adverb	*astfel, altfel, alaltăieri, alaltăseară, altădată, altcîndva,altcum, altcumva*; *altcîndva, altcum, altcumva*
	adverb + adverb:	*bunăoară, etc.*
	adverb + auxiliary element	*cîndva, cumva,încotrova, undeva, oarecînd, oarecum, oarecît, oareunde, orictnd, oricum, oriîncotro, oriunde, oricît, etc.*
	adverb + *şi*	*cîtuşi, iarăşi, totuşi*
	several heterogeneous elements	*cîteodată, nicicacum, nicidecum, numaidecît, orişicînd, orişicum, orişiunde, pasămite, încămite, etc.*
Preposition	composed from constructive elements combined in a single word	*despre, dinspre, dintre, deasupra, dindărătul*
Conjunction	composed from constructive elements combined in a single word	*aşadar, deoarece, fiindcă, încît*

Table 2. The second mode of compound word formation

Part of speech	Constructive elements	Examples
Common noun	common noun formed from two or more independent common nouns, where the first element is inflected	*tonă-kilometru* (*tone-kitometru*)
	common noun formed from two or more independent common nouns, where the second element is inflected	*liber-cugetător (liberului-cugetător)*
	common noun formed from two or more independent common nouns, where all elements are inflected	*maşină-unealtă (maşinii-unelte)*
	common noun formed from two or more independent common nouns, where no elements are inflected	*apă-albă*
Proper noun	proper noun formed from two or more nouns in nominative-accusative case	*Grinâuţi-Moldova*
	proper noun followed by another noun, being personal or geographic name	*Ana-Maria*
	proper noun followed or forwarded by a qualificating noun or adjective	*Alb-împărat*
	proper noun followed or forwarded by a common noun	*Vadul-Leca*

	other structure	*Lacul Sărat, Polul Nord*
Adjective	adjective + adjective	*alb-galbui,*
Numeral	formed of two or more words linked by prepositions, from which at least one of numerals is a simple one	*al doilea (a doua), patruzeci şi opt cinci zecimi; de două ori, de trei ori, etc.*
Pronoun	with different elements of composition	*nici unul, nici un, nici una, nici o, etc.* (negative pronoun); *Alteţa sa, Sanctitatea sa, etc.* (polite pronoun)
Adverb	noun + noun, noun + adverb, adverb + adverb, verb + verb, interjection + interjection	*calea-valea, an-vară, cîine-cîineşte, vrând-nevrând, treacă-meargă, încet-încet, hodoronc-tronc, ţac-pac, etc.*
	adjective *ast*ă + noun	*astă-iarnâ, astă-toamnă, astâ-vară, etc*
	adverb + noun	*azi-dimineaţă ieri-noapte, ieri-seară, etc.*
	preposition *după* + noun	*după-amiază, după-masă, după-prînz, etc.*
	preposition întru, dintru + *adverb*	*într-aceea, dintr-acolo, într-adevăr, într-adins, într-aiurea, etc.*
Preposition	composed from several simple prepositions	*de la, de către, de pe, de pe la, de după, de sub pe lângă, până pe la, până pe lângă, etc.*
	prepositional locations of different structure	*afară de, în afară, pe aproape de. de jur împrejurul*
Conjunction	compound conjunctions are many different structures	*fără ca, pentru ca, pentru că, pentru ce, de ce, ca şi cum, ca şi cînd, pe cîtă vreme, cît timp, ori de cîte ori, etc.*
Interjection	formed from one or more simple interjections	*cioc-cioc, hop-hop, hi-hi-hi, ha-ha-ha, tronca-tronca, etc.*
	repeating simple interjections	*miau,miau; mac,mac; boca,boca; etc.*

THE INFLEXION PROCESS

The inflexion process is as follows. The word to be inflected is entered and indicated as simple or compound. If the word is simple, the inflexion program works as described in [4–5]. If the compound word is formed in accordance with Table 1, the inflexion program works in the usual mode described in the [4–5]. If the compound word is formed as described in Table 2, additional information is needed to inflect this word. After inflection the manual correction of the resulting forms is possible. During manual correction the user has a possibility to edit the whole visible screen, however not inserting and deleting lines and without block operations. After the manual correction the words can be introduced into the dictionary.

CONCLUSIONS

The presented data and described methods are the first step to further development of automated tools for Romanian lexicon population, including the inflexion of compound words.

REFERENCES

1. Studii de Materiale privitoare la Formarea Cuvintelor in Limba Română Academia Republicii Populare Române. Institutul de Lingvistică din Bucureşti. Volumul I. Editura Academiei Republicii Populare Române, Bucureşti, 1959.
2. E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, V. Demidova, L. Malahova: "Reusable lingvistic resources for romanian." - Second Conference of the Mathematical Society of the Republic of Moldova, August, 17-19, 2004, Chişinău, Republic of Moldova. – p. 54–57.
3. E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, V. Demidova, L. Malahova: "Lexical resources for Romanian". - în "Memoriile ştiinţifice ale Academiei Române", Bucureşti, România. 2004.
4. S. Cojocaru: Romanian Lexicon: Tools, Implementation, Usage. – In: Dan Tufis & Poul Andersen (eds.), Recent Advances in Romanian Language Technology. – ISBN 973–27–0626–0, Editura Academiei, 1997, I, pp. 107–114.
5. E. Boian, S. Cojocaru. The inflexion regularities for the Romanian language. – Computer Science Journal of Moldova, Vol. 4, Nr. 1 (10), 1995, pp.40–58.
6. E. Boian, S. Cojocaru, L. Malahova. Instruments pour applications linguistiques. La terminologie en Roumanie et en République de Moldova, Hors série N4, 2000.
7. Studii de Materiale privitoare la Formarea Cuvintelor in Limba Română Academia Republicii Populare Române. Institutul de Lingvistică din Bucureşti. Volumul I. Editura Academiei Republicii Populare Române, Bucureşti, 1959.
8. Gramatica limbii române.– Bucureşti, Ed. Acad. Române, 1977.
9. Iorgu Iordan, Vladimir Robu. Limba Română Contemporană. Editura Didactică şi Pedagogică, Bucureşti, 1978.