The Formal Patterns of the Lithuanian Verb Forms

Loïc Boizou
Centre of Computational Linguistics (http://donelaitis.vdu.lt)
Vytautas Magnus University
Kaunas, Lithuania
l.boizou@hmf.vdu.lt

Abstract

This paper describes the formal structure of the Lithuanian verbs, emphasizing the difference between two kinds of formal patterns called primary and secondary. This short outline attemps to highlight some salient aspects of different descriptive levels (traditionnal model, formalized model and implemented model).

1 Introduction

In Lithuanian, morphology plays a considerable role in both domains of inflection and derivation. This property is obvious for both nouns and verbs. Although Lithuanian verbal morphology in general is quite thoroughly described, automatic processing gives some opportunities to consider the question somehow differently.

The aim of this study is to describe the formal patterns of Lithuanian verbs from the perspective of the written language. The study is restricted to the conjugated forms, putting the main emphasis on the word forms instead of lexemes. It covers only the question of analysis and interpretation : the problem of lemmatization is left beyond the scope.

This paper gives a short theoretical account of the question, it raises some problematic issues concerning the formal interpretation of verbal word forms and outlines a possible implementation of verbal formal patterns using the analyzer ALeksas.

2 Primary vs secondary verb forms

According to Stankiewicz (Stankiewicz, 1999), verbs tend to be formally simpler than nouns. Lithuanian seems to conform to this remark, as, in spite of the complexity of the Lithuanian conjugation, verbs forms follow a limited set of patterns. There are two families of patterns, the primary and the secondary ones, which differ in many respects. It must be emphasized that these patterns describe verbal word forms, not verbal lexemes. Indeed, while verbal word forms are either primary or secondary, verbal lexemes may have

	simple	mixed	complex
present	*perk-a*	*mieg-a*	**auk-oj-a**
preterit	*pirk-o*	**mieg-oj-o**	**auk-oj-o**
infinitive	*pirk-ti*	**mieg-o-ti**	**auk-o-ti**

Table 1: Types of verbal lexemes

- only primary forms (simple verbs)
- only secondary forms (complex verbs) or
- a combination of primary and secondary forms (mixed).

These three categories, well known in the Lithuanian grammatical tradition (ex: DLKG, (Ambrazas (red.), 1996), are shown in the table 1 (primary verb forms in italics, secondary ones in bold). The given forms are the usual ones in the Lithuanian lexicographic tradition, present tense, preterit, infinitive, called by Hoskovec (Hoskovec, 2009) the lemmatic root triplet.

Furthermore, all the verb forms in the past iterative tense are secondary ones, since there are made with the suffix dav-, ex. *dainuodavome* 'we used to sing, we often sang', *šokdavo* 'used to dance, often danced'.

2.1 The primary verbs forms

The number of verbs with primary verb forms reaches few thousands item, but this list seems to be a closed set. From a lexematic perspective, these verbs often show complex models of inflexion, with vowel alteration, infixation, inflexional suffixes. In general, they offer a large panel of paradigmatic variety.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 50–57

2.1.1 Formal structure

Primary forms follow quite a simple pattern[1] (with optional elements in brackets):

(ModPfx+) (Pfx+) (Refl+) Root' (+Enlargement) +Ending[2]

Each element of the structure may appear only once.

The root The root is in fact a lexically actualized root (hence the notation *Root'*), where the vocalism is fully specified. In some cases, actualized roots contain an infix *-n-*, or *-m-* in prelabial context, (*randa*[3] vs *rado*, *tampa* vs *tapo*).

All the primary verb roots are monosyllabic, with the exception of GĄLANDA 'to sharpen' (at least in a synchronic perspective). They match the following pattern (Ambrazas (red.), 1996), where all consonant parts are optional :

Spir + Occl + Son + V + Son + Occl + Spir

The root may be extended by optional elements, enlargements and prefixes.

Enlargements Enlargements (the term used by Hoskovec (2009) is taken from Benveniste) make up a small group of consonant suffixes appearing only after the root (this principle excluding the possibility of iteration). The list includes *-st-* (*tirpsta* 'melt'), *-d-* (*pildo* 'fill'), *-s-* (*linksi* 'to nod' from LINKSI).

Prefixes Primary verb forms may include one prefix. Proper prefixes, which are mainly of prepositional origin, belong to a narrow set : *ap(i)-*, *at(i)-*, *į-*, *iš-*, *nu-*, *pa-*, *par-*, *per-*, *pra-*, *pri-*, *su-* , *už-*. As a rule, prefixation does not modify the inflection of the base verb.

Beside the proper prefixes, some modal prefixes (in fact prefixed particles) can be added to prefixed or unprefixed verb forms : *ne-* (negation), *be-* (duration), *te-* (restriction) and the combinations *tebe-* (continuation), *nebe-* (interruption). These modal prefixes appear at the absolute beginning, before proper prefixes.

The status of the particle *ne-* is specific, since it can also be used like a proper prefix, for example NERIMSTA 'to worry'. Besides, the related verb SUNERIMSTA deviates from the normal pattern.

Paradigms and desinences Lithuanian is characterized by the coexistence of several verbal desinential systems (for the description of the formal structure of the desinences, see (Chicouène and Skūpas, 1998) and (Hoskovec, 2009)) presented in the table 2. The form of desinences may change before the reflexive clitic.

From a lexematic point of view, the present tense (*-a*, *-ia*, *-o* or *-i*) and the preterit (*-o* or *-ė*) show a concurrence between desinential system, while other tenses and moods have a unique paradigm (future *-i'*, imperative *-i"*). All concurrent systems appear with some primary verb forms, for example:

- -a : *tinka*, *verda* (present)
- -ia : *keikia*, *taria* (present)
- -i : *nori*, *žiūri* (present)
- -o : *šaudo*, *sako* (present), *tirpo* (preterit)
- -ė : *liepė*, *valdė* (preterit)

Possible combinations of present and preterit desinential systems draw a set of eight theorically conjugation paradigms, seven being eventually used by the system (numbers indicate conjugation class and subclass in the Grammar of the contemporary Lithuanian(Ambrazas (red.), 1996)), as shown in the table 3.

The quantitative weigth of each model greatly varies from one isolated verb (I-IV) to some several hundreds (I-I, taking in account primary verbs only). Furthermore, a full description of the paradigms needs to integrate enlargements and root alterations, but given the present approach focused on the words forms and not on the lexemes, this question will not be further discussed.

The split description of the reflexive clitic In Lithuanian, the reflexive clitic *(*-si) may appear in two different positions: if the verb is prefixed (even by a modal prefix), the reflexive clitic is between the prefix(es) and the root, ex. *ne-at-si-kelė* 'did not wake up'; else, the clitic is at the end of the word (possibly with a formal alteration of the desinence), ex. *džiaugia-si* 'is/are delighted'.

Taking in account the efficiency of the implementation, the traditionnal unified description of the of the clitic was abandoned. Thus, the model includes

[1] The present outline is quite brief, for a more complete description, see (Hoskovec, 2009), which largely inspired this presentation.

[2] ModPfx: modal prefix, Pfx: (non modal) prefix, Refl: reflexive marker.

[3] The paper follows the notational convention of Matthews (1991): small capitals for lexemes, italic for word forms.

	a	ia	o	ė	i	i’	i”	ė’
1sg	renku	šaukiu	sakau	liepiau	žiūriu	kviesiu	×	imčiau
2sg	renki	šauki	sakai	liepei	žiūri	kviesi	lauk	imtum(ei)
3	renka	šaukia	sako	liepė	žiūri	kvies	×	imtų
1pl	renkame	šaukiame	sakome	liepėme	žiūrime	kviesime	laukime	imtumėme
2pl	renkate	šaukiate	sakote	liepėte	žiūrite	kviesite	laukite	imtumėt(e)

Remark the paradigm of the conditionnal (ė’) is made of two components, a suffixal segment *-tum-* and a desinence of type ė. Nonetheless, given the lack of stability of both components, the whole structure is considered in the present model as a specific desinential system.

Table 2: Verb desinential systems

	preterit	
pres	o	ė
a	I-I	I-II
ia	I-IV	I-III
o	III-II	III-I
i	II	×

Table 3: Verbal paradigms

Before consonant	Before present (paradigm -a)	Before preterit (paradigm -o)
-in-, -en-		
-o-, -ė-, -y-	-oj-, -ėj-, -ij-	
-au-, -uo-	-auj-, -uoj-	-av-

Table 4: Compatibility of the verbal suffixes

1. a reflexive prefix (unable to take the initial position) ;
2. a set of reflexive endings;
3. a rule of incompatibility between them.

It allows to give a simpler formalization of the morphological expression of reflexivity in Lithuanian, since the split avoids to handle a single morphological unit with two different positions in the morphematic structure.

2.2 The secondary verb forms

The set of secondary verbs is an open collection, insofar as it includes verbal derivatives and borrowed verbs.

2.2.1 Formal structure

The pattern of secondary verbs is made of an arbitrarily complex morphological structure containing an actualized root, followed by a verb suffixe and a desinence. That is, suffixes may follow an already suffixed base (*mok-y-toj-auja* ’work(s) as a teacher’), a prefixed one (*nebe-žiūrės* ’won’t look any more’, *ne-už-ant-spauduoja* ’do(es) not stamp’) or even a compound one (*šun-uodegavo* ’toadied’, *su-daikta-vardėjo* ’became a noun’). It results from the derivational role of suffixes.

The set of verbal suffixes is quite restricted : *-o-* , *-ė-*, *-au-*, *-uo-*, *-av-*, *-i-*, *-y-*, *-in-*, *-en-*. There are some cases of combined suffixes : *-st-y-*, *-d-in-* (both with enlargement), *-in-ė-*, *-tel-ė-*. Besides, Lithuanian shows several examples of formal variants (although most of them are not productive any more) with different initial segments : ex. *-dė-*, *-sė-*, *-ė-*; *-dy-*, *-sy-*, *-y-*. It seems to confirm the remark of Kuryłowicz (Kuryłowicz, 1936) about the trend of initial segments in complex suffixes to lose their individual function, the whole combination becoming a free variant of the second suffix used separately.

Contrary to primary verb forms, which show a great variety of paradigms, secondary forms are compatible only with the -a paradigm for the present and with the -o paradigm for the preterit (as type I-I). Given the compatibility of suffixes with endings, different groups of suffixes may be recognized (see table number 4).

The prefixes *-en-* and *-in-* may appear before all types of ending (-a, -o and consonant[4]). The suffixes *-oj-*, *-ėj-* and *-ij-* appear before vowels, *-o-*, *-ė-* and *-y-* before consonant. The suffixes *-auj-* and *-uoj-* appear only before the present and *-au-* /*-uo-* before consonants, while *-av-* appears only before past endings.

The remarks concerning primary verbs about modal prefixes and reflexive clitic are shared by secondary verbs.

[4] All tenses other than the present and the preterit, that is, future, conditional, past iterative, imperative, are made with a consonantic onset.

noun	verbal derivative
-as (auksas 'gold')	-uo-j-a (auksuoja 'to gild')
-a (auka 'sacrifice')	-o-j-a (aukoja 'to sacrifice')
-ė (dėmė 'stain')	-ė-j-a (dėmėja 'to stain')
-is (dalis 'part')	-i-j-a (dalija 'to share')

Table 5: Reactulisation of the noun desinential base

2.2.2 General features of the suffixation

The weak specification of verbal suffixes It must be emphasized that most of the Lithuanian proper suffixes and enlargements are not specific to either derivation or inflexion. That is, they are general morphological devices. In fact, many morphological markers (prefixes, suffixes, endings) are obviously multi-functional in Lithuanian.

For example, the enlargement *-st-* can be used as an inflexion marker indicating the present[5] (*dingsta* 'disappear(s)' vs *dingo* 'disappeared') or as a derivation marker indicating the repetition of the process (PJAUNA 'cut (once)' vs. PJAUSTO 'cut (repeating the process several times)').

It is possible to give a similar example with the suffix *-ė(j)-*, in (rare) derivational use, cf ČIULPĖJA 'to touch several times' vs. ČIULPA 'to touch', and in inflexional use, *kalbėjo* 'talked' vs. *kalba* 'talk(s), is/are talking'.

Desinential traces The semantic typology of the denominal verbal suffixes given in the Grammar of the Contemporary Lithuanian (Ambrazas (red.), 1996) proved to be in many respects inaccurate. Indeed, this description offers a widespread synonymy and homonymy rising doubt on the relevancy of the classification.

In fact, despite their formal identity, some morphological elements are not suffixes, but a remaining part of the desinential vocalic base of the base lexeme. Selected allomorphs depend on the paradigms of declension of the base noun, as shown in the table 5.

Desinential traces are only formal elements bereft of semantic motivation, and thus, invalidating the semantic categorization. The formal concordance between traces and suffixes must be emphasized : it accounts to a large extent for the shortcomings in the presentation of verbal suffixes in the Lithuanian tradition. The description of the desinential traces explains some essential features of the morphological system. From a historical perspective, it must be noticed that, although such cases are not exceptional, it concerns old derivatives, remaining as a legacy. As a general feature in morphology (Kerleroux, 2005), not all the system is semantically motivated

Main semantic values The core semantic system is made of the suffixes *-in-*, *-ė-*, *-uo-*, *-au-*, *-av-* (counterpart to *-uo-* and *-au-* in the preterit), *-telė-*.

By comparing some verbs like GELTONUOJA 'to appear green', GELTONĖJA 'to become green' and GELTONINA 'to make sth green', it seems possible to state that *-uo-* is intransitive and static, *-ė-* intransitive dynamic and *-in-* transitive dynamic. That's why *-in-* (and some other related suffixes *-din-*, *-y-*, *-dy-*) is used for causative verbs, since they also involve a transitive dynamic process. Thus, *-in-* may derive verbs from nouns (RUSINA 'to russify s.o.' from RUSAS 'Russian') as well as from verbs (TALPINA 'to make sth fit into' from TELPA 'to fit into')

The suffix *-au-* expresses activity (MOKYTOJAUJA 'to work as/be a teacher (MOKYTOJAS)', UOGAUJA 'to pick berries (UOGOS)').

The suffix *-telė-* (deverbal verbal suffix) expresses a very short process (ŽVELGIA 'to look', ŽVILGTELĖJA 'to glance').

Borrowed verbs Some borrowed verbs follow the semantic system, ex. BROKERIAUJA (unconventional) 'to deal', SPORTUOJA 'to practice sport', EUROPINA 'to europeanize s.o./sth', EUROPĖJA 'to europeanize, to acquire european features', but for most of them the question of the present productivity is open.

Contemporary loanwords seem to use the suffixes *-uo-* (and its allomorphs *-uoj-* / *-au-*) or *-in-* as general integrators (Corbin, 1986) into the verb class, ex. SKENUOJA 'to scan', DEVALVUOJA 'to devaluate', SINCHRONIZUOJA 'to synchronize', GUGLINTI (unconventional) 'to use Google', TVITINTI (unconventional) 'to use Twitter'. Although in such cases the suffix seems to be unspecified regarding semantic value and transitivity, the couple FORMUOJA 'to form sth' / FORMUOJASI 'to form (intr)' is an (older) example where opposition is renewed by resort to the category of reflexivity.

[5] In fact, it might be better described as a co-marker, whose value results from the combination of the enlargement and the desinence.

After this brief general presentation, it must be dealt with the more practical aspects of recognition and interpretation of the verbal word forms.

3 Interpretation of the word forms

In a perspective of NLP, we consider the description of the verb structures as a way to extract or analyse some word forms in a symbolic framework.

3.1 Recognition of the verb forms

The recognition of verb forms without lexicon is not an easy task. The main problem is that primary verb forms are highly ambiguous (it is general tendency in Lithuanian, (Rimkutė and Grigonytė, 2006)). Different factors explain this situation:

- formal simplicity : nouns can be formally more complex than verbs, but they can also be as simple as primary verbs ;
- ambiguity of endings : desinences tend to be short segments often a vowel or a vowel and a consonant (although some Lithuanian desinence may be dissyllabic), leading to a widespread homonymy, increased by the high number of inflexion categories and the concurrence of multiple paradigms ;
- prosodic deficiency : the written language does not mark prosody, which conveys very useful grammatical information in the spoken language;
- extended conversion : the same root can be easily actualized according to different parts of speech : ex. *kalbos* 'languages', *kalbu* 'I speak', *kalbus* 'loquacious'.

As a consequence, analysis without dictionary frequently generates multiple interpretations. For example, formally *rauda*, *teka* and *gera* are all possible verbs (candidate verbs), but while *teka* is really a verb ('flows'), *gera* is an adjective ('good'), and *rauda* might be either a verb ('cries') or a noun ('lamentation').

It is yet possible to recognize few well marked verb forms. Thus, some prefixes are specific to the verb category : *pri-*, *nu-*, *su-*, *api-*, *ati-*. With few exceptions, words with monosyllabic roots combined with these prefixes are all verb forms. There are two systematic exceptions concerning some deverbal noun forms :

- words with the element *-t-*, which seems to be a general mark of deverbalization in Lithuanian. It appears in the infinitive forms (ex. TARTI 'to pronounce, to utter'), in the so called past passive participle (ex. LAUKTAS 'waited') and in some other deverbal nouns (TARTIS 'pronunciation', to be compared to infinitive, NAŠTA 'burden'[6]). Such derivatives may all present the given prefixes, therefore bases ended with *-t-* are ambiguous in almost all cases;
- verbal adjective derived by conversion, ex. NUMANUS, PRIVALUS.

In such cases, the recognition must rely on some unambiguous desinences.

Furthermore, the frequency of the conversion strongly limits the possibility to use marked forms to tag less marked ones. For example given a form *tebekalba*, which is obviously a verb to the 3rd person present, it does not imply that *kalba* (without the modal particle) is always a verb (*kalba* 'speaks'), since it can also be a noun (*kalba* 'language'). Nevertheless, it may be possible to recognize some verbo-nominal roots. This idea given by Patrice Pognan (oral communication) comes from the Semitic tradition and would lead to a distinction between verbo-nominal roots (SKUBA 'to hurry', SKUBA 'haste', SKUBUS 'urgent') and purely nominal ones (MEDIS 'tree', ŠUO 'dog').

It is usually easier to deal with secondary forms, since longer forms are usually more marked. Nonetheless, there are some systematic interferences which stem from noun suffixes such as *-ija* (collective suffix), *-ėjas* and *-tojas* (agent suffixes), ex. genitive masc. sg. *kėpėjo* 'backer's' vs. *girdėjo* 'heard'.

3.2 Grammatical interpretation

Once a word form is recognized as verbal the next step is to interpret its grammatical features.

As for recognition, secondary forms are easier to interpret. Given the strict limitation of paradigms, their grammatical interpretation is unambiguous. The only exception concerns the inference between the preterit forms of some derivatives in *-avo* and the past iterative tense (in *-davo*). It arises when the suffixal variant *-av-* is preceded

[6]The same element appears in many deverbal compound suffixes (ĮKURTUVĖS, LENKTYNĖS, SPAUSTUVĖ, KASTUVAS, JUNGTUKAS, DEGTINĖ, TEIKTINAS), but in such cases word forms are not monosyllabic any more.

by *-d*, for example pret. *maldavo* 'begged' (cf. pres. *maldauja*, past iter. *maldaudavo*) vs past iter. *maldavo* 'used to grind' (cf. pres. *mala* 'grinds').

For primary forms, problems mainly arise from the tenses using enlargement, which can be :

- future (enlargement *-s*, allomorph *-š*) : *kep-s-iu* fut. 'I will cook' (pres. *kep-u*) vs *juos-iu* pres. 'I wear a belt' (fut. *juos-ė-s-iu*)
- imperative (enlargement *-k*) : *tar-k-ime* imper. 'let's say' (pres. *tar-iame*) vs. *tik-ime* pres. 'we believe' (imper. *tik-ė-k-ime*);
- the 1^{st} pers. of the conditional : *dirb-č-iau* cond. (pret. *dirb-au*) vs *kviečiau* pret. (cond. *kvies-čiau*)

The previous examples involved the category of tense and mood, but the *-i* paradigm imply a systematic ambiguity between the 2^{nd} pers. sg. and the 3^{rd} pers. (ex. *nori*)

From the point of view of the implementation, all these formal interferences require either a lexicon and/or the handling of multiple interpretations (possibly solvable by a following syntactic or semantic disambiguation).

4 Formal approach

The present section gives a short account of a possible formalization of the verbal lexical structure. This is a complementary approach to lexicon-based analysis, since it allows to provide interpretations for verbal word forms absent from a given lexical database, be it neologisms, rare verbs or occasionalisms.

4.1 Patterns

The verbal patterns may be expressed by regular expressions.

Conventions Conjunction is indicated by direct concatenation (ex. $\alpha\beta$), disjunction by a vertical stroke ($\alpha|\beta$). Generic symbols are indicated by upper case letters (for morphological classes) or by Greek letters (for phonological classes, more precisely from a phonographic point of view).

Generic symbols:

X	base of arbitrary morphological complexity
A	desinences of type -a
C	desinences of the conditional
Ė	desinences of type -ė
I	desinences of type -i
I'	desinences of type -i'
I"	desinences of type -i"
'A	desinences of type -ia
O	desinences of type -o
T	A | 'A | O | Ė | I
T'	sI' | kI" | C
V	T | T'
Σ	syllable
Σ^{α}	syllable ending in α
$\Sigma^{-\alpha}$	syllable ending by coda other than α
σ'	s | š | ž
σ	s | š
δ'	t | d | s
δ	t | d
γ	k | g

Primary verbal patterns

ΣdO|Ė	(ex. *įvykdė*)
ΣstA|'A|O|Ė	(ex. *tirpsta*)
$\Sigma^{\sigma'}$t|A|O|Ė	(ex. *klysta*, *laužta*, *pjaustė* + *niežti*)
Σ^{σ}I'	(ex. *kvies*)
Σ^{k}I"	(ex *bėk*)
ΣV	

Remark: all the preceding patterns may be preceded by (M)(P)(si), where M is a modal prefix and P a proper prefix.

In fact, some patterns are slightly more strict than ΣV. For example, ΣsI' (implied by ΣV) should be defined as $\Sigma^{-\delta'}$sI' (and similarly $\Sigma^{-}\delta'$C and $\Sigma^{-}\gamma$kI"). It is a consequence of some morphophonological alternations (*t*+*s* $\rightarrow$ *s*, *d*+*s* $\rightarrow$ *s*, *s*+*s* $\rightarrow$ *s*). But since such configurations are impossible in Lithuanian, the approximation ΣsI' is sufficient.

Secondary verbal patterns

XinA|O|T'	
XenA|O|T'	
XėjA|O	
XojA|O	
XijA|O	
XėT'	
XoT'	
XyT'	
XuoT'	
XuojA	
XaujA	
XavO	
XdavO	(past iterative tense)

4.2 The analyzer formalism

The model is implemented with ALeksas (Boizou, 2009), a morphological analyzer of the Lithuanian language, based on a structural description of the lexicon by formal patterns expressed by finite state

automata. The data are given in quite a rough format, with a numerical input state, a numerical output state and a transition symbol.

```
Ex.  (1, 2, "pri").
```

However, the formalism is extended by some features which give to the description a more natural linguistic expression:

- complex symbols
- generic symbols and inheritance
- grammatical values recording

Besides, ALeksas allows recursive structures: a transition by an automaton is possible. Thus, the automaton representing the root (made of characters) is nested in the automaton of the lexical structure (made of morphological elements). The aim is to avoid the mixing of different levels of description (morphemic vs graphematic).

Complex symbols In the first example, the symbol was a simple string (`"pri"`), but ALeksas allows complex symbols made of a string, a set of grammatical values and a set of operations on the grammatical context (see *Value recording*).

```
Ex.  (250, 30, { "is"; TM(DSN),
 CAS(NOM), GNR(M), NB(SG) ; }).
```

The three mentioned parts of symbol are separated by semi-colons (in the previous example, the last part is empty). Features encoding grammatical information are made of a name of feature and a corresponding value in brackets. The number of features bound to a symbol is free.

Generic symbols and inheritance With the aim of minimizing redundancy in automata, ALeksas allows the use of generic symbols. For example, instead of listing all the prefixes as transitions, it is possible to declare a transition by a generic symbol (written without quotes) :

```
Ex.  (1, 2, Pfx).
```

All generic symbols must be defined in the header of the file :

```
Ex.  PFX = "pri" | "su" | "nu" |
                ...  .
```

Generic symbols may be recursive, that is, a generic symbol (ex. `ExtSfxXV`) may derived another one (ex. `SfxV`).

```
Ex.  SfxV = ExtSfxXV | { "d" ;
    SEMT(fact) } | { "st" ;
         SEMT(fact) }.
```

Generic symbols can also be associated to grammatical features shared by derived symbols. This property is very close to the concept of inheritance in object-oriented programming. For example, all the symbols derived from the generic symbol `Pfx` may inherit the value *prefix* (`PFX`) for the feature *morphological type* (`MT`) :

```
Ex.  (5, 5, { Pfx ; MT(PFX) ;
        [>PFX] }).
```

The last component of the complex symbol, operations on the context, is described in the next paragraph.

Value recording ALeksas is also extended by a grammatical recording device, which allows to carry relevant grammatical informations while progressing in the automaton. Such data are recorded in a register expressing the grammatical context. The register of ALeksas presents many similarities both in design and in purpose with the registers of Cohen-Sygal and Wintner (2006).

ALeksas defines four operations which may be carried on the register, two mutators and two accessors :

>X : adds the symbol X

<X : suppresses the symbol X

+X : asserts the presence of the symbol X

-X : asserts the absence of the symbol X

All these operations, which can be combined by &, appear in the third part of the complex transition symbol.

```
Ex.  (5, 5,  "si" ; REFL ;
   [+PFX&-REFL&>REFL] ).
```

The purposes of the register are :

1. to insure of more natural expression of some relations between morphological units;
2. to transfer information between the different levels (root automaton ↔ lexical automaton);
3. to minimize the automaton, especially in case of distant grammatical dependencies.

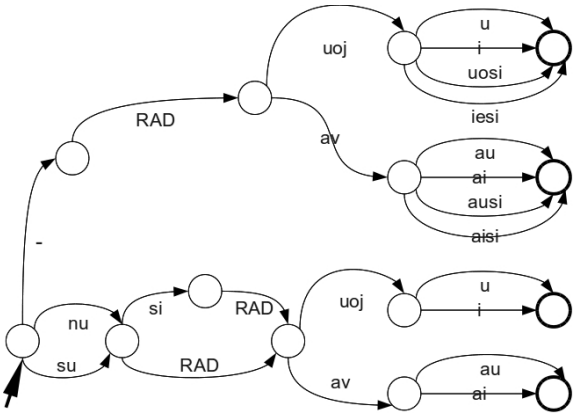

Figure 1: Sketch of an automaton

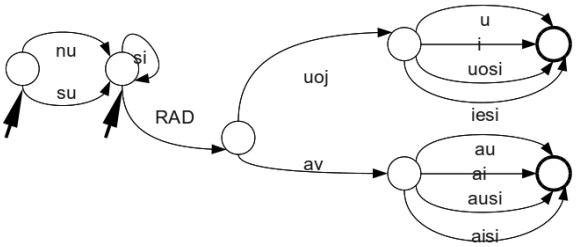

Figure 2: Corresponding compressed automaton

A characteristic example of the third point is given by the reflexive clitic. With a usual automaton, the structure (with a small illustrative subset) would be as in figure 1. The upper part of the figure describes unprefixed word forms, in which desinences can be either reflexive or not, while the lower part describes the prefixed verbal forms, reflexive or not, in which desinences cannot be reflexive. Obviously, the structure is partially double.

The use of the register and the mechanism of extended symbols, which are able to encode a part of the grammatical information, allows a significant reduction of the automaton, as shown in figure 2 (the rule of incompatibility between two elements with a REFL feature avoiding the conflict between a reflexive prefix and a reflexive ending).

However, such an enhancement from the point of view of expressivity involves a higher degree of complexity in processing and an increase in the time of execution.

5 Conclusion

Despite strong restrictions on the verbal patterns, which belong to two very different subsets, recognition and analysis of verb forms raise many problems. The challenge mostly arise from the fact that the simplest formal patterns are, to a great extent, shared by verbs and nouns and from multi-functional nature of many morphological elements.

The model has to be tested, in order to determine more precisely possible gaps in the description and to evaluate the efficiency of the proposed approach, especially by comparison with lexicon-based and statistical methods.

In further works, the question of lemmatization or, alternatively, the recognition of paradigmatically related verb forms, has to be addressed, so as to set the connections between word forms, which are essentially syntactic units, and lexemes, that is, lexical units.

References

Vytautas Ambrazas (red.). 1996. *Dabartinės lietuvių kalbos gramatika*. Mokslo ir enciklopedijų leidykla, Vilnius.

Loïc Boizou. 2009. Analyse lexicale automatique du lituanien. Master's thesis, Institut national des langues et civilisations orientales, Paris.

Michel Chicouène and Laurynas-Algimantas Skūpas. 1998. *Parlons lituanien*. L'Harmattan, Paris.

Yael Cohen-Sygal and Shuly Wintner. 2006. Finite state registered automata and their uses in natural languages. *Lecture Notes in Computer Science*, pages 43–54.

Danièle Corbin. 1986. *Morphologie dérivationnelle et structuration du lexique*. Max Niemeyer, Tübingen.

Tomáš Hoskovec. 2009. *Formální morfologie litevštiny ve funkčním popisu jazyka*. Slovanská ústav, Praha.

Françoise Kerleroux. 2005. Morpho-logie : la forme et l'intelligible. *Langage*, 152:12–32.

Jerzy Kuryłowicz. 1936. Dérivation lexicale et dérivation syntaxique. *Bulletin de la société linguistique de Paris*, 37:79–92.

Peter Matthews. 1991. *Morphology*. Cambridge university Press, Cambridge.

Erika Rimkutė and Gintarė Grigonytė. 2006. Automatizuotas lietuvių kalbos morfologinio daugiareikšmingumo ribojimas. *Kalbų Studijos*, 9:: 30–37.

Edward Stankiewicz. 1999. Grammatical categories and their formal patterns. *Travaux du Cercle linguistique de Prague*, (3):71–90.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955