INFORMATICA, 2000, Vol. 11, No. 1, 19–40

Automatic Stressing of the Lithuanian Text on the Basis of a Dictionary

Pijus KASPARAITIS
Department of Informatics, Faculty of Mathematics, Vilnius University
Naugarduko 24, 2006 Vilnius, Lithuania
e-mail: pkasparaitis@yahoo.com

Received: January 2000

Abstract. The paper deals with one of the components of text-to-speech synthesis of the Lithuanian language, namely – automatic text stressing. The present work substantiates the necessity to divide words into fixed and variable parts used to build different grammatical forms, as well as to store only those parts rather than the whole worlds in the dictionary. According to the inflexion method, all words of the Lithuanian language are divided into three groups (noun-adjectives, verbs and non-inflectional words) and each group is analysed separately. The type of information, as well as the form in which it is to be stored, has been established for each group and the algorithm by means of which the grammatical form of a word can be recognised and stressed, has been presented.

Key words: text-to-speech synthesis, automatic text stressing.

1. Introduction

One of the most recent areas of applying computer technique is text-to-speech synthesis. Text-to-speech synthesis can be divided into the following stages: dividing the text into syllables, automatic stressing, transcription, generation of a sound. The present work is concerned with automatic stressing of the Lithuanian text. Stressing means defining the place of the stress in a word and the accent.

Rules of stressing are presented in many Lithuanian grammars, however, in most cases the latter are concerned with the definitions, which are absolutely inadequate for work with a computer. E.g., rules for stressing nouns are presented in "A Grammar of Modern Lithuanian" (Ambrazas *et al.*, 1996). A noun is defined as "an independent part of speech, to which words denoting names of things, phenomena, actions and features belong, and which has independent categories of gender, case and singular/plural". Hardly any of these concepts can be easily formalised so that a computer could verify whether the word is or is not a noun.

Perhaps the most acceptable stressing method for a computer to use for stressing words is to store a list of certain words that have already been stressed in its memory. Such a method applies to non-inflectional words. However, most words are inflectional. Consequently, this method can cause the following problems:

1) the dictionary will take too much space in the computer memory;

2) the search in a large dictionary will take too long;

3) compiling such a dictionary will be too time-consuming;

4) new words might appear which will be impossible to predict in advance.

Let us try to approximately assess the size of the dictionary. The Dictionary of Modern Lithuanian, 1993 contains about 50000 words and the Dictionary of International Words, 1985 contains 21159 words. The number of words totals about 70000. The longest word in Lithuanian is made of 31 letters. If words were non-inflectional, such a dictionary would require about 2 MB of memory and could be compiled within a reasonable time.

1065 endings and 252 prefixes are used to build different grammatical verb forms. The total number is 268380 combinations. Though some of these combinations are impossible to be formed and some of the endings are impossible to be added to a certain stem, the number of combinations is still large. It is obvious that there is no point in storing all grammatical forms in the dictionary and it takes too much time to compile such a dictionary. An automatic compilation of a dictionary of stressed words poses a problem of the same complexity as does stressing during synthesis.

Thus, it is a good idea to store parts of words (prefixes, stems and endings) in the dictionary together with the information necessary for building and stressing words. The present work discusses what information is needed and in what way it should be stored. However, before proceeding with the analyses, additional requirements for the stressing algorithm should be considered.

2. Additional Requirements for Stressing Algorithm Used in Speech Synthesis

In this work we restrict ourselves to stressing separate words, i.e., no contextual information is used. E.g., in the combinations of words “žmonių galvos” and “žmonės galvos” the inflexion of the word “žmonės” helps recognise that in the first case we have the future tense of the verb “galvoti” and in the second case – the nominative case plural of the noun “galva”. These words have stresses in different places. Unfortunately, so far no thorough investigations have been carried out to define how the structure of a sentence can be used for automatic stressing.

Requirement 1. The stressing algorithm must provide for the possibility to include additional information about stressing depending on the context, in case such information becomes available.

Requirement 2. If the word can be stressed in several ways with nearly the same probability and if the stress position changes the meaning (e.g., “kalvo∼s” and “kal∼vos”), it is preferable to leave that word unstressed rather than to make a mistake in stressing it. People listening to a synthesised speech quite often miss the unstressed words, however, they always catch the words that are stressed incorrectly. Words are worth stressing only in case one stressing variant is much more frequent statistically than the other, e.g., the locative case of “kieme` ” is used more often than its vocative case “kie∼me”.

Sometimes the words with different accents in the same syllables containing a vowel or the diphthongs “ie”, “uo” sound very similarly, e.g., “li′epų” and “Lie∼pų”. In such cases one stressing variant can be chosen as well.

Requirement 3. The stressing algorithm should comprise all the words used in Lithuanian (including names of places, surnames and international words) irrespective of their having some features atypical of the Lithuanian language.

Requirement 4. In case a word can be stressed in several ways and the stress does not change its meaning, e.g., "deguo∼nis" and "deguoni`s" (Ambrazas *et al.*, 1996), one stressing variant is chosen and that variant will always be used. There is no need to provide for possibilities to stress a word in several ways.

In order to meet requirements 1 and 2, stressing has been divided into two stages:

1. Attempts have been made to recognise what grammatical forms and what words the word being analysed coincides with. E.g., the word "galvos" can be: a) the genitive case singular of the noun "galva", b) the nominative case plural of the noun "galva", c) the future tense of the verb "galvoti". Each grammatical form is stressed. No final decision about stressing is made at this stage.

2. It is being verified whether all grammatical forms are stressed in the same way. E.g., the noun "pi′eva" has the same stress in the nominative, instrumental and vocative cases. Provided all the stressing variants are identical, the word is stressed; otherwise the contextual information can be used to find a single stressing variant. In case such information is unavailable, grammatical forms with low probabilities (e.g., vocative cases of nouns) can be rejected. If only those grammatical forms, which are stressed in the same or in a similar way remain, the word is stressed, otherwise the word is left unstressed.

Dividing the stressing process into two stages is convenient from that point of view that it enables us to use absolutely different algorithms for nouns, adjectives, verbs, etc. Besides, it is unnecessary to predict all possible identical grammatical forms.

3. Stressing in Other Languages

Most languages have a fixed stress (Girdenis, 1995), i.e., the stress position is regulated by strict rules. In most cases these are simple propositions indicating the distance of the stress from the beginning or the end of the word. Three models of a fixed stress have been established on the basis of this distance:

1) the stress in the first syllable. This stressing system is characteristic of the Latvian, Czechish, Slovakian, Icelandic, Estonian, Finnish, Hungarian languages;

2) the stress in the last syllable. This stressing model is typical of most Turkic and the Persian (Tadjik) languages. The French language has a similar stressing system, however, there certain groups of words rather than separate words are stressed;

3) the stress in the penultimate syllable. The Polish language has this stressing system.

More complicated stressing models, in which the stress position depends on the quantity and quality of syllables, are also possible. E.g., in the Mongolian language the stress falls on the first long syllable, however, the first syllable is stressed if all the syllables in a word are short.

Automatic stressing is quite a simple matter in the languages with a fixed stress. The Lithuanian language, like the Russian, Bulgarian, Serbian–Croatian, Italian, Spanish and

English languages, has a free stress. In some languages, which have a free stress, many words with the same ending have the same stress (e.g., this is the case in the Italian language (Nebbia, 1990)), therefore statistical methods can be used in stressing. In the languages with a complicated transcription (e.g., English or German (Paulus, 1998)) stressing and transcribing are done simultaneously with the help of a dictionary of stressed and transcribed words or parts of words.

Since transcription is a relatively simple matter in the Lithuanian language (Kasparaitis, 1999), it is more convenient to separate stressing and transcribing. This paper presents one of the possible models of automatic stressing of the Lithuanian text with the help of a dictionary, as well as stressing and inflexion rules.

4. Stressing Parts of Speech on the Basis of a Dictionary

As mentioned above, it is impossible to compile a dictionary containing all grammatical forms of all Lithuanian words in a computer memory. This chapter presents one of the methods of compiling a dictionary of main parts of words and of recognising different grammatical forms of the word, as well as stressing them on the basis of that information.

It is appropriate to divide all words into three groups according to the inflexion method: 1) declinable (nouns, adjectives, some pronouns and numerals), 2) conjugated (verbs and non-conjugated forms of verb, e.g., participles) 3) non-inflectional. Each of these groups are analysed below.

4.1. *Stressing Nouns and Adjectives Based on Dictionary*

4.1.1. *Division into Parts of Word*

All nouns and adjectives can be divided into two components: the stem and the ending. Inflexion of the noun or adjective means building of their different grammatical forms by adding an ending to the stem. Building of new words by means of suffixes and prefixes is not discussed in this chapter. Prefixes and suffixes are treated as a part of the stem. Hence, adjectives with prefixes, as well as the diminutives are regarded as words with different stems.

Besides, division of words into the stem and the ending enables the stressing process to be carried out in two stages:

1) defining whether the stress is in the stem or in the ending (defining the stress position in a word),

2) defining the stress position in the stem or in the ending respectively.

4.1.2. *Inflexion*

Nouns and adjectives have singular and plural. Presently only the adjectives of the feminine and masculine gender shall be discussed (the neuter gender will be considered later) because only these adjectives are inflectional. Though not all the nouns have singular and plural, in this chapter they will be treated as such because:

1) most of the nouns and all adjectives have singular and plural,

2) this renders the model simpler,
3) sometimes it is unnecessary to know that the word does not possess some grammatical form. E.g., the word "žirklės" has only plural (The Dictionary of Modern Lithuanian, 1993). Providing this word is assumed to have singular too, there will be no problems since no singular form will appear in a real text.

There are 6 cases in the Lithuanian language. Traditionally, the seventh, the vocative case, will be attributed to them. The plural form of the vocative case is identical to that of the nominative case. Thus, the number of cases in singular and plural totals 13: the singular nominative case (using the Lithuanian abbreviation it will be denoted vv), the singular genitive case (vk), the singular dative case (vn), the singular accusative case (vg), the singular instrumental case (vį), the singular locative case (vt), the singular vocative case (vš), the plural nominative case (dv), the plural genitive case (dk), the plural dative case (dn), the plural accusative case (dg), the plural instrumental case (dį), the plural locative case (dt).

In addition to numbers and cases, the adjectives have two genders, four degrees (excluding one type of adjectives), pronominal and non-pronominal forms. All together – 16 combinations. Hence, it is convenient to assume that 16 sets of endings can be added to the stem of the adjective. Some nouns have the feminine and masculine gender too, e.g., "šern-as – šern-ė", "ligon-is – ligon-ė", "mokytoj-as – mokytoj-a", "inžinier-ius – inžinier-ė". As there are many combinations of masculine and feminine endings (one ending of the masculine adjective corresponds to only one ending of the feminine adjective) and the majority of nouns have only one gender, it is more convenient to assume that there are two different stems.

4.1.3. *Types of Stems*

Endings belonging to a certain set can be added to the stem of the noun or the adjective. The set usually contains one ending per case. Sometimes there are several endings corresponding to one case (e.g., vt "vėj-uje" and "vėj-yje"), however, there might not be a single ending either (e.g., most adjectives have no vocative case). Thus, each set of endings contains 13 groups of endings. The set of endings is defined by the declension of the stem (there are 5 declensions of nouns and 4 declensions of adjectives) or by the declension paradigm of the stem (there are 12 declension paradigms of nouns, 5 declension paradigms of adjectives of the masculine gender and 4 declension paradigms of adjectives of the feminine gender). Another grouping of stems is used in this work, which are called types of stems. The new grouping is necessary because some groups of identical endings belong to different declension paradigms (e.g., the words "rank-a" and "sauj-a"), and some sets with different endings – to the same paradigm (e.g., "peil-is" and "arklys"). The principal criterion used in forming types of stems was as follows: words, which have only different endings in a certain case, cannot belong to a single type. 19 types of stems of nouns and 48 types of stems of adjectives have been distinguished. Suffixes used to form degrees of comparison of adjectives are treated as a part of the ending. To simplify the model the fact that some groups of endings of adjectives are identical was not taken into consideration. The formation of types on the basis of the endings of certain cases is presented in Table 1.

Table 1
Types of stems

Type of stem	Cases, endings, examples
	Nouns
1	vv “–as” after a hard consonant (“namas”), vt “–e” (“name”);
2	vv “–as” after “j” (“vėjas”), vt “–uje” or “–yje” (“vėjuje” or “vėjyje”);
3	vv “–ias” (“kelias”), vt “–yje” (“kelyje”);
4	vv “–is” after a hard consonant (“brolis”), vk “–io” (“brolio”);
5	vv “–is” after “j” (“kūjis”), vk “–o” (“kūjo”);
6	vv “–ys” after a hard consonant (“arklys”), vk “–io” (“arklio”);
7	vv “–ys” after “j” (“žvejys”), vk “–o” (“žvejo”);
8	vv “–a” (“ranka”, “sauja”), vk “–os” (“rankos”, “saujos”);
9	vv “–ia” (“vyšnia”), vk “–ios” (“vyšnios”);
10	vv “–i” (“marti”, “pati”), vk “–ios” (“marčios”, “pačios”);
11	vv “–ė” after a hard consonant (“bitė”), vk “–ės” (“bitės”), dk “–ių” (“bičių”);
12	vv “–ė” after “j” (“skerssijė”), vk “–ės” (“skerssijės”), dk “–ų” (“skerssijų”);
13	vv “–is” (“krosnis”), vk “–ies” (“krosnies”), vn “–iai” (“krosniai”);
14	vv “–is” (“žvėris”), vk “–ies” (“žvėries”), vn “–iui” (“žvėriui”);
15	vv “–us” after a hard consonant (“sūnus”), dn “–ums” (“sūnums”);
16	vv “–us” after “j” (“pavojus”), dn “–ams” (“pavojams”);
17	vv “–ius” (“sodžius”), dn “–iams” (“sodžiams”);
18	vv “–uo” (“akmuo”), vk “–ens” (“akmens”);
19	vv “–uo”, “–ė” (“sesuo”, “duktė”), vk “–ers” (“sesers”, “dukters”);
	Adjectives
20–23	vv “–as”, “–a” after a hard consonant (“geras”, “gera”, “gerasis”, “geroji”);
24–27	vv “–as”, “–a” after “j” (“abuojas”, “abuoja”, “abuojasis”, “abuojoji”);
28–31	vv “–ias”, “–ia” (“žalias”, “žalia”, “žaliasis”, “žalioji”);
32–35	vv “–is”, “–ė” (“didelis”, “didelė”, “didysis”, “didžioji”), dv “–i” (“dideli”);
36–39	vv “–ys”, “–ė” (“kairys”, “kairė”, “kairysis”, “kairioji”);
40–41	vv “–is”, “–ė” after a hard consonant (“medinis”, “medinė”), dv “–iai” (“mediniai”);
42–43	vv “–is”, “–ė” after “j” (“ilgakojis”, “ilgakojė”), dv “–ai” (“ilgakojai”);
44–47	vv “–us”, “–i” after a hard consonant (“gražus”, “graži”, “gražusis”, “gražioji”);
48–51	vv “–us”, “–i” after “j” (“gajus”, “gaji”, “gajusis”, “grajoji”);
52–55	vv “–esnis”, “–esnė” (“geresnis”, “geresnė”, “geresnysis”, “geresnioji”);
56–59	vv “–ėlesnis”, “–ėlesnė” (“gerėlesnis”, “gerėlesnė”, “gerėlesnysis”, “gerėlesnioji”);
60–63	vv “–iausias”, “–iausia” (“geriausias”, “geriausia”, “geriausiasis”, “geriausioji”);
64–67	vv “–ausias”, “–ausia” (“gajausias”, “gajausia”, “gajausiasis”, “gajausioji”).

4.1.4. *Stress Paradigms*

Nouns and adjectives are divided into four stress paradigms in the Lithuanian language. The stress paradigm is defined by the stress position in a word in the dative case plural and the accusative case plural. Such a classification is insufficient for automatic stressing of the text because the words belonging to the same stress paradigm can have the stress

Table 2

Sets of stresses

Stress paradigm	No.	vv	vk	vn	vg	vį	vt	vš	dv	dk	dn	dg	dį	dt	Types of stem
I	1	1	1	1	1	1	1	1	1	1	1	1	1	1	all
II	2	1	1	1	1	0	0	1	1	1	1	0	1	1	1
	3	1	1	1	1	0	1	1	1	1	1	0	1	1	2,4,5,11,12,40–43
	4	0	1	1	1	0	1	1	1	1	1	0	1	1	8,9
	5	1	1	1	1	1	1	1	1	1	1	0	1	1	15,16,17
III	6	1	1	1	1	1	0	1	0	0	0	1	0	0	1,3
	7	0	1	1	1	1	0	0	0	0	0	1	0	0	6,7
	8	0	0	1	1	1	0	1	1	0	0	1	0	0	8,11,12,21,25,29, 37,45,49
	9	0	0	1	1	0	0	0	1	0	0	1	0	0	13,14,15
	10	0	0	1	1	1	0	0	1	0	0	1	0	0	18,19
	11	1	1	0	1	1	0	1	0	0	0	1	0	0	20,24,28,32
	12	1	0	1	1	1	0	1	1	0	0	1	0	0	33
	13	0	1	0	1	1	0	–	0	0	0	1	0	0	36
	14	0	0	0	1	1	0	–	1	0	0	1	0	0	44,48
IV	15	1	1	1	1	0	0	1	0	0	0	0	0	0	1
	16	1	1	1	1	0	0	0	0	0	0	0	0	0	3
	17	0	1	1	1	0	0	0	0	0	0	0	0	0	6,7
	18	0	0	1	1	0	0	1	1	0	0	0	0	0	8–12,21,23,25,27, 29,31,37,39,45,47, 49,51
	19	0	0	1	1	0	0	0	1	0	0	0	0	0	13,14,15,18
	20	1	1	0	1	0	0	–	0	0	0	0	0	0	20,24,28,32
	21	1	0	1	1	0	1	–	1	0	0	0	0	0	33
	22	0	1	0	1	0	0	–	0	0	0	0	0	0	22,26,30,34,36,38, 46,50
	23	0	0	0	1	0	0	–	1	0	0	0	0	0	44,48
	24	0	0	0	0	0	0	–	0	0	0	0	0	0	52–67

in different places in other cases. E.g., the words “pirštas” and “ranka” belong to the second stress paradigm, however, in vv they are stressed “pir∼št-as” – “rank-a` ” and in vt - “piršt-e` ” – “ran∼k-oje”.

A new concept – a set of stresses – has been introduced. Each stress paradigm will be divided into sets of stresses. The set of stresses defines the stress position (either in the stem or in the ending) in each case. The set of stresses is defined by the declension and stress paradigm. The simplest model would be to have one set of stresses for each combination of the declension and the stress paradigm. The total number of combinations would amount to 268 (67 types of the stem * 4 stress paradigm). Many different stems

can be stressed by means of the same set of stresses. The total number of different sets of stresses used is 24 (see Table 2).

In this Table "0" means that the stress is in the ending, "1" – the stress is in the stem and "–" – the word does not have that case. Words belonging to some type of stems cannot be stressed by means of certain stress paradigm. Thus, the right column does not represent all types of stems.

It is convenient to have one more Table, which can assist in finding the set of stresses making use of the type of the stem and the stress paradigm.

4.1.5. *Information about Stems*

As mentioned above, it is convenient to divide stressing of nouns and adjectives into two stages: first to determine if the stress is in the stem or in the ending and then to find the stress position either in the stem or in the ending, respectively. Consequently, two databases are needed:

1) the database of stems in which information about the stress position in the stem, as well as information about the stress position in a word is stored here. The stress position in a word is defined by the type of the stem and the stress paradigm. Both these attributes are features of the stem;

2) the database of endings containing information about the stress position in the ending.

The stress position in the stem (in case the stress is in the stem) and the accent do not depend on the ending.

Summing up the above information, the entity relationship diagram (ERD) (Barker, 1994) of the database of stems could be presented as follows (Fig. 1).

The attribute *Name* of the entity **Stem** is a textual representation of the stem. Certain endings can cause assimilation in the ending of the stem. Therefore it is worth discussing the most convenient form of storing the textual representation of stems in the database and how to search for them.

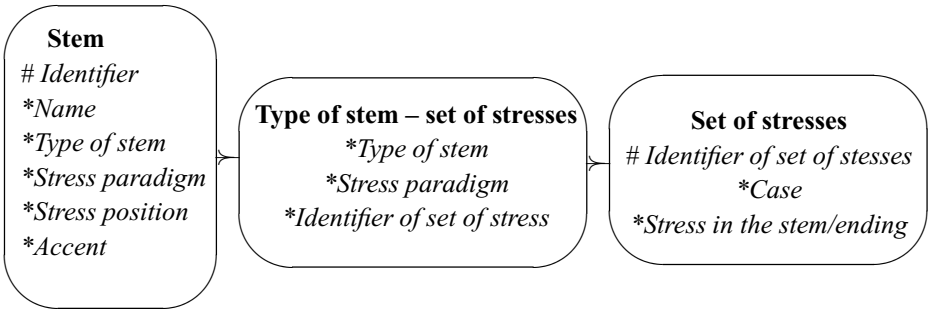

Fig. 1. Entity relationship diagram of the data base of stems.

4.1.6. *"D" and "t" Assimilation at the End of Stems*

DEFINITION. Endings that begin with "i" before a vowel, excluding "e", are called soft endings. All other endings are called hard endings.

PROPOSITION 1. If the stem ends in "d" or "t" (in this case the ending is always hard) and while declined it acquires the soft ending, "d" changes to "dž" and "t" – to "č" at the end of the stem.

PROPOSITION 2. The opposite proposition is incorrect. E.g., the words "Sočio" and "Mačio" have soft endings in the vk case but in the vv case they have hard endings ("Sočis", "Mačys"), however, they still retain "č". If a stem ends in either "č" or "dž" before the hard ending, the same stem is retained with all the endings.

As a rule, a stem in the vv case is the stem of a word. A word can have hard endings in all cases, e.g., "banda", "bandža", "pučas", "puta"; soft endings in all cases, e.g., "valdžia", "risčia"; a hard ending in the vv case, however, a soft ending in another case, e.g., "medis – medžio", "kirtis – kirčio"; a soft ending in the vv case and a hard ending in another case, e.g., "kurčias – kurtiems", "bergždžias – bergždiems". Further conclusions that follow from Propositions 1 and 2 are not presented here. The final algorithm for storing stems and searching for them is as follows:

1) If a word has a hard ending before "d" or "t" and a soft ending before "dž" or "č" stems ending in "d" or "t" are stored. Stems that remain after the ending has been removed are stored for all other words.

2) If the word being analysed has the stem ending in "dž" or "č" and the ending is soft, two stems are used for the search: one stem that ends in "dž" or "č" and the other stem that ends in "d" or "t". One stem that remains after the ending has been removed is used for the search in all other cases.

Thus, some words should be stored with the stems other than those we are accustomed to, e.g., the word "bergždžias" is stored with the stem "bergžd" (because there exists the form "bergždieji"), and the word "kurčias" – with the stem "kurt" ("kurtieji").

4.1.7. *Information about Endings*

The stress position in the ending and its accent is the feature of the ending and does not depend on the stem it is added to. This is quite an interesting and unusual feature of the ending because the same ending can be added to the stems of a different type and it can denote different cases, e.g., vv "sūn-u` s" and dg "nam-u` s". This happens quite often.

Taking into account the said feature, it is convenient to store separately the ending with its attributes (name, stress position, accent, feature of softness) and the information about in what case and what type of stem this ending can be added to. Fig. 2 represents this in terms of the ERD.

The above-mentioned feature is characteristic merely of full endings. In dialects some endings have short variants which can:

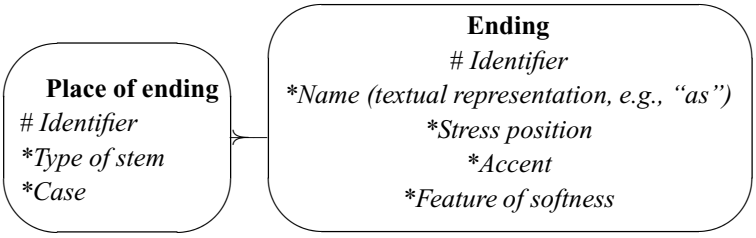

Fig. 2. Entity relationship diagram of the database of endings.

1) be identical to the full ending of the same stem in other cases but they can have different stress positions or accents, e.g., "ger-a´ jam" (vn) and the short ending "ger-a∼-jam" (vt).

2) two short endings can be identical, however, they can have different stress positions or accents, e.g., "ger-o´ siom" (dn) and "ger-o∼siom" (dį).

Short endings mentioned in Point 1 were not included in the list of endings. Only one ending mentioned in Point 2 was included in the list. All other short endings were included in the list of endings. The list contains (including short endings) 355 endings of nouns and adjectives (begining with "-as", "-o", "-ui", . . . and ending in . . ., "-ausiomis", "-ausiose").

4.1.8. *Stressing Other Grammatical Forms and Parts of Speech*

Adjectives of the neuter gender have the same stress position and accent as those of the masculine gender vv case ("ge∼ras – ge∼ra", "gražu` s – gražu` "), thus, it is convenient to treat endings of the neuter gender as one more ending of the masculine gender vv case. Similarly endings of adjectives of the neuter gender of the comparative ("geriau∼", "gerėliau∼") and superlative ("geria´usia") degrees are also treated as endings of the singular masculine gender vv case of corresponding degrees.

Stressing of adverbs formed by means of the stem of the adjective ("gerai∼", "tam∼siai") is more complicated. Endings of adverbs are treated as endings of one more case. A special feature is stored for this case in the set of stresses indicating that an additional algorithm is to be used to define the stress position. This algorithm makes use of the following features: the number of syllables, the type of stem, the stress paradigm and the length of a stem.

Ordinal numerals and the numeral "vienas" are treated as adjectives. All the cases of the numerals "du", "dvi", "trys", "keturi", "keturios", . . ., "devyni", "devynios" are stored in the database of the non-inflectional words. The numerals from "vienuolika" to "devyniolika" and "šimtas", "tūkstantis", "milijonas", "milijardas" are treated as nouns. The numerals "dešimtis", . . ., "devyniasdešimtis" are treated as nouns, however their short forms in vv case "dešimt", . . ., "devyniasdešimt" are stored in the base of the non-inflectional words.

The pronouns "kitas", "visas", "kiekvienas", "tūlas", "manas", "tavas", "savas", "šitas" are treated as adjectives. The following forms "šituo∼", "šitie∼" and "šituo∼s" are

stored in the base of non-inflectional words only. The following pronouns “toks”, “šioks”, “šitoks”, “anoks”, “koks”, “joks”, “visoks”, “vienoks”, “kitoks”, “kažkoks” are assumed to be declined and stressed in the same way as the adjective “žalias”, with their vv, vg and dv cases (“to´ks”, “to´kį”, “tokie∼”) being stored in the base of the non-inflectional words. All grammatical forms of all other pronouns are also stored in that base.

4.1.9. *Common Algorithm of Stressing Adjectives and Nouns*

1) All the endings are taken for each word being stressed and it is verified whether the attribute *Name* coincides with the end of the word being stressed. In case they coincide, with the last letter of the stem being a vowel, after the ending has been removed, such an ending is considered to be improper. E.g., the endings “-ų” and “-ių” are proper for the end of the word “kačių”, however, after the ending “-ų” has been removed, the stem ends in the vowel “i”. The following list is compiled: Remaining stem, Ending (the identifier of an ending). It is quite often that this list contains more than one entry. E.g., the following hypotheses of separating the stem and the ending in the word “žaliuosiuose” are formulated:

A. “žal-”, “-iuosiuose”,
B. “žali-”, “-uosiuose”,
C. “žaliuos-”, “-iuose”,
D. “žaliuosi-”, “-uose”,
E. “žaliuosiu-”, “-ose”,
F. “žaliuosiuos-”, “-e”.

Hypotheses B, D and E are rejected because the stem ends in a vowel. Besides, it is verified whether the stem remains after the ending has been removed, as some endings are identical to the whole Lithuanian words, e.g., “o”, “į”, “imi”.

Endings and stems can consist of a different number of syllables and letters. Besides, there are no features marking the boundary between the stem and the ending. Finding the longest ending does not give desired results either as, e.g., the endings “-ajai” and “-ai” match the end of the word “samurajai”, however, it is only with the help of the second one that a correct separation is achieved. Consequently, formulating all possible hypotheses is the only possible way. The same method is applied to verbs.

2) For each entry on the list we verify whether the ending is soft or hard (the attribute the *Feature of softness* is being verified). If the ending is soft and the stem ends in “č” or “dž”, the list is supplemented with one more entry in which “č” is changed to “t” and “dž” is changed to “d” at the end of the stem (see Chapter “D” and “T” assimilation at the end of stems”).

3) Now the search for matching stems is made in the database of stems and a new list is compiled: Identifier of the stem, Type of the stem, Identifier of the ending. Without doubt, not a single matching stem has been found for some of the entries of the earlier list, whereas many matching stems have been found for others, e.g., the stem “žal-“ is the stem of the words “žalias”, “žalas” and “žala”. For each entry of the new list the *Type of stem* of which belongs to adjectives of the masculine gender (the stems of adjectives are stored by indicating this type of the stem), the list is supplemented with 15 entries,

in which the type of the stem corresponds to all possible combinations of the gender, degree and pronominal/non-pronominal forms. For the adjectives, which possess neither pronominal forms nor degrees, the list is supplemented with one entry corresponding to the feminine gender.

Why do we first remove the ending and then look for the stem? By removing the ending first it is possible to determine whether the ending is soft and whether the assimilation rules are to be applied.

4) Assuming the stems and endings are stored in the relational database generated according to the ERD presented in Fig. 1 and 2, the further search is possible to be made by means of the following query:

SELECT *Stress in the stem/ending*
FROM **Stem, Type of stem – set of stresses, Set of stresses, Place of ending, Ending**
WHERE
Set of stresses. *Identifier of set of stresses* = **Type of stem – Set of stresses**. *Identifier of set of stresses*
AND **Type of stem – set of stresses**. *Type of stem* = Type of stem
AND **Type of stem – set of stresses**. *Stress paradigm* = **Stem**. *Stress paradigm*
AND **Stem**. *Identifier* = Identifier of stem
AND **Set of stresses**. *Case* = **Place of ending**. *Case*
AND **Place of ending**. *Type of stem* = Type of stem
AND **Place of ending**. *Identifier* = Identifier of ending

During this search no matching entries can be found for an individual entry on the list or as many as several matching entries can be found.

5) Now we go through all the entries found. If the attribute *Stress in the stem/ending* shows that the stress is in the stem, **Stem.** *Identifier* is used to find the *Stress position* and *Accent*, and if the stress falls on the ending – **Ending.** *Identifier* is used.

6) In case the stress position and accent is the same for all the entries, the word is stressed. If it is not, either a certain algorithm is used to find one stressing variant or the word is left unstressed.

4.2. *Stressing of Verbs on the Basis of Dictionary*

The present Chapter is concerned with both conjugated and non-conjugated forms of verbs, such as participles, half-participles, verbal adverbs, infinitives, however, they all will be called by a common name – verbs.

Let us assume any verb to consist of the stem and the ending. Besides, it may have an optional group of prefixes. The Chapter will deal merely with the formation of different grammatical forms of the same verb by means of adding endings and prefixes to the stem.

The formation of new words by means of suffixes (e.g., “neš-ti” and “neš-io-ti”) will not be discussed. Such suffixes are treated as a part of the stem.

Some grammatical forms are built by adding not only the ending but also a suffix to the stem, e.g., “neš-tin-as”. Such suffixes are treated as a part of the ending.

Dividing verbs into the stem, the ending and the prefix (as in the case of nouns and adjectives) enables us:

1) to store stems, endings and prefixes in separate databases and thus to decrease the amount of information;

2) to divide the stressing process into two stages: determining which part of the word is stressed (the stem, the ending or the prefix) and finding the stress position in a corresponding part of the word.

4.2.1. *Classes of Conjugation*

All grammatical forms of verbs are formed by means of three main stems: the stem of the present tense, the stem of the past tense and the stem of the infinitive. These stems can either be identical, e.g., "neš-a", "neš-ė", "neš-ti", or different, e.g., "kert-a", "kirt-o", "kirs-ti". To simplify the model, all these stems will be assumed to be different.

All endings of the verb can be divided into three groups according to the stem they can be added to. Some endings can belong to many groups, e.g., the ending "-o" can be found in the verb of the present tense "mat-o" and in the verb of the past tense "kirp-o". In order to simplify the model, these groups will be treated as separate.

Class of conjugation defines the group of endings that can be added to a certain stem by building grammatical forms in the present and past tenses. Endings of the third person define these groups. According to the Lithuanian grammar, three classes of conjugation are distinguished in the present tense (1. "a" and "ia", 2. "i", 3. "o") and two classes – in the past tense (1. "o", 2. "ė"). Since the endings "a" and "ia" have a somewhat different form, they are more convenient to be treated as the endings of different classes.

Some endings can be added to the stem taking into account the endings in the present and past tenses. E.g., "ius" ("rašius") can be added only to the stems which acquire the ending "ė" ("rašė") in the past tense and the ending "o" ("rašo") in the present tense. Thus it is more convenient to analyse classes of conjugation of the present and past tenses together and to form classes of conjugation by means of the endings of both tenses.

Consequently, the following classes of conjugation (the ending of the present tense – the ending of the past tense) have been established: 1) "a – o", 2) "a – ė", 3) "ia – o", 4) "ia – ė", 5) "i – o", 6) "o – o", 7) "o – ė".

4.2.2. *Rules of Stressing*

The ending of the verb determines its grammatical form. The prefix "te-" is also used for this purpose. Irrespective of the fact that some grammatical forms have the same ending, all grammatical forms are being discussed. A verb can have the stress in the prefix, the stem and the ending. The stress position is defined by certain rules. All grammatical forms can be classified into groups according to the rules of stressing.

The following rules have been formulated (examples of grammatical forms that are stressed in compliance with the rule are given in brackets).

Rule 1. If a verb has a prefix and the stress is retracted to the prefix – the prefix is to be stressed; if the last syllable of the stem is stressed and its accent is not falling – the ending is to be stressed, otherwise the stress falls on the stem ("kerpu", "kirpau").

Rule 2. If a verb has a prefix and the stress is retracted to the prefix – the prefix is to be stressed, otherwise the stem is to be stressed (“kerpa”, “kerpant”, “kerpantis”, “kerpamas”, “kerpančiai”, “kirpo”).

Rule 3. If the stem in the present tense is neither polysyllabic nor belongs to the class of conjugation „o-*“ – the ending is to be stressed, otherwise the stem is to be stressed (“tekerpie”, “kerpamam”, “kirpdama”, “kirptam”, “kirpsimam”, “kirptinam”, “kirpte”, “kirptinai”, “kirptai”).

Rule 4. The stem is to be stressed (“temokai”, “kirpus”, “kirpęs”, “kirpusiai”, “kirpti”, “kirpdavau”, “kirpsiu”, “kirpčiau”, “kirpk”, “kirpdamas”, “kirpdavus”, “kirpsiant”, “kirpdavęs”, “kirpsiąs”, “kirpsimas”, “kirptinas”).

Rule 5. If the stress is retracted to the prefix – the ending is to be stressed; otherwise the stem is to be stressed (“kerpąs”).

Rule 6. If the stem in the present tense is neither polysyllabic nor belongs to the class of conjugation „o-*“, the accent is not falling and there are only the vowels “a” or “e” in the stem – the ending is to be stressed, otherwise the stem is stressed (“kerpamai”).

Rule 7. If the stem is neither polysyllabic nor the accent is falling – the ending is to be stressed, otherwise the stem is stressed (“kirptų”).

Rule 8. If the stem is not polysyllabic, the accent is not falling and the verb has a prefix – the prefix, is to be stressed, otherwise the stem is to be stressed. If there are short stressed vowels “a” or “e” in the stem – the accent changes to the rising one (“kirptas”).

Rule 9. If the vowel assimilation (see Chapter “Assimilation rules of the stem endings”) was applied to the stem – the accent is short, if the last syllable in the stem is stressed and the accent is falling – the accent changes to the rising one. The stem is to be stressed (“li˜ s”, “lanky∼s”).

The above rules imply that the following information is necessary for the stress position and the accent to be defined:

1) whether a verb has a prefix;
2) whether the stress is retracted to the prefix;
3) the stressed syllable in the stem;
4) the accent of the stress in the stem;
5) whether the stem is polysyllabic;
6) class of conjugation;
7) whether there are vowels „a“ or „e“ in the stem.

The condition „whether the stress is retracted to the prefix“ should be divided into two following conditions: „whether the stress is retracted to the prefix in the present tense“ and „whether the stress is retracted to the prefix in the past tense“. Seven cases when the stress is not retracted to the prefix in the present tense and four cases when it is retracted have been established (Vaitkevičiūtė, 1997). Therefore it is more appropriate to store alongside the stem, as an attribute, the feature indicating whether the stress is or is not retracted to the prefix in the present tense.

Opposite is the case with the past tense. The stress is retracted to the prefix only in case the verb belongs to the class of conjugation "a-ė" or "ia-ė" and if it does not have the falling accent (Vaitkevičiūtė, 1997). Hence, it is easy to determine whether the stress is retracted to the prefix and it is not necessary to store this feature as an attribute.

Consequently, Rules 1 and 2 can be rewritten separately for the present and past tenses.

The condition "whether the stem is polysyllabic" can also be divided into two cases: "whether the stem in the present tense is polysyllabic" and "whether the stem of the infinitive is polysyllabic". These two features could be stored as attributes in the database, however, it is easy to count syllables algorithmically. Besides, the class of conjugation defines the number of syllables in many cases. Stems of classes of conjugation "a-ė", "ia-o", "ia-ė", "i-o", "o-o" and "o-ė" are not polysyllabic in the present tense. Stems of classes of conjugation "a-ė" and "ia-ė" are not polysyllabic in the infinitive. Stems of classes of conjugation "ia-o", "i-o", "o-o" and "o-ė" have many syllables in the infinitive. The number of syllables should be counted algorithmically only for the stems belonging to the class of conjugation "a-o".

4.2.3. *Assimilation Rules of Stem Endings*

In case the firs letter in the ending is "s" and the stem ends in "s", "z", "š", "ž", the endings "s" disappear and "z", "ž" become "s", "š", respectively, e.g., "kirp" + "siu" = "kirpsiu", "mes" + "siu" = "mesiu", "megz" + "siu" = "megsiu", "neš" + "siu" = "nešiu", "vež" + "siu" = "vešiu" (Ambrazas *et al.*, 1996). Since first of all we look for matching endings, all possible variants of endings are needed. Therefore, the list of endings beginning with "s" is supplemented with the same endings that begin with "š", e.g., "-siu" and "-šiu".

In case the ending that begins with "s", "š" or "k" matches the end of the word, several stems are used for the search. The examples below show how these stems are formed:

"meg-siu" – "meg", "meg+s" and "meg+z";

"ve-šiu" – "ve+š" and "ve+ž";

"au-kime" – "au", "au+k" and "au+g".

By adding a soft ending to the stem that ends in "d" or "t" those letters change to "dž " or "č", respectively. However, unlike nouns stems of verbs cannot end in "dž" or "č" and have a hard ending. Therefore, upon finding "dž" or "č" at the end of stem, the latter are changed into "d" and "t", respectively. Only the forms of stems that end in "d" and "t" are stored in the database of stems.

The ending "-s" (the third person of the future tense) deserves special mention, as it is this ending alone that can cause the change of vowels in the stem. The grammar rule runs as follows: if a verb has either "y" or "ū" in monosyllabic stems of the present tense and the infinitive, and the stem of the past tense has "i" or "u", respectively, the third person of the future also has either "i" or "u". Since the aim of solving the problem of automatic stressing is to recognise the form of the future tense rather than to build it, the following algorithm is necessary: if, after removing the ending "-s", the stem ends either in "i" or "u", to supplement the list of stems used for the search by one more stem and to remember

that the assimilation rule has been applied to that stem. The letter “i” is changed to “y” and “u” is changed to “ū” in that stem. Later, when looking for the stem in the database, provided the assimilation rule has been applied to the stem, not only the textual form of the stems must coincide but also the stem of the infinitive must be monosyllabic, the stem of the present tense must end in either “y” or “ū”, and the stem of the past tense must end in either “i” or “u”, respectively.

Is it not possible to verify immediately after the ending “s” has been removed whether the stem is monosyllabic? No, it is not, because the stem may have a group of prefixes, e.g., “neprilyti – neprilis”.

4.2.4. *Prefixes*

In the Lithuanian language there are 14 prefixes intended for the formation of verbs: “ap”, “api”, “at”, “ati”, “į”, “iš”, “nu”, “pa”, “par”, “pra”, “pri”, “su”, “už”, “per”. Besides, alongside prefixes, the particle “ne”, “nebe”, “tebe”, “be”, the reflexive particle “si” and the particle of the imperative mood “te” can be used. Prefixes and particles can form groups of prefixes. They can appear in the following sequence: the particles “te”, “tebe” or “be”, the particles “ne” or “nebe”, a prefix and finally a reflexive partition, e.g., “te-neapsi”. The last element of the group of prefixes is always stressed, with the exception of the groups containing the prefix “per” which is always stressed.

A group of prefixes can contain only one prefix or a particle, with the exception of the reflexive particle.

As in the case with endings it is necessary to compile a list of all possible hypotheses of removing the prefix and the stem, e.g., 1) A. “neper-“, “-skaito”, B. “ne-“, “-perskaito”; 2) A. “neper-“, “-inti”, B. “ne- “, “-perinti”; 3) A. “prisi-“, “-rinko“, B. “pri-“, “-sirinko”; 4) A. “prisi-“, “-rpo”, B. “pri-“, “-sirpo”.

The question is as follows: what is to be removed first – the prefix or the ending? It should be noted that a verb cannot have both a reflexive ending and a reflexive particle in the group of prefixes at the same time, i.e., if a verb has a prefix, a reflexive particle can only be in the group of prefixes. Thus, if the prefix is removed first, it can be established whether it contains a reflexive particle, and if so, the ending should be searched among the endings that are not reflexive. If the ending is removed first, it can be established whether it is reflexive, and if so, the word cannot have any prefix at all.

How is it possible to determine whether the ending is reflexive? Sometimes it is impossible to find a reflexive particle in the ending. E.g., the ending „-antis“ can be the ending of a non-reflexive particle and the ending of a reflexive verbal adverb. Therefore, it is advisable to store together with the ending one more attribute indicating whether the ending is reflexive.

Prefixes and particles formed 252 groups of prefixes. Since this is not a large number, groups of prefixes that have already been formed can be stored together with the information about the stress position and the accent in the computer memory. This method makes stressing of prefixes quite easy because the position of the stressed letter can be stored. In other cases the method of building a group of prefixes presented above is to be used.

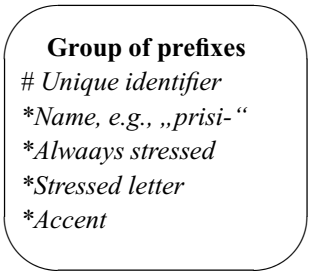

Fig. 3. ERD of the database of prefixes.

As in the case of nouns and adjectives, prefixes, endings and stems are convenient to be stored in separate databases. The ERD of the database of prefixes is presented in Fig. 3.

Group of prefixes has the feature *Always stressed* if it contains the prefix “per-”.

4.2.5. *Endings*

Taking into account the above said, information necessary to stress endings is presented in Fig. 4.

The attribute *Feature of prefix “te-”* enables us to find those grammatical forms, which can have this prefix.

Attributes *Stressed letter* and *Accent* are optional because some endings are never stressed.

The database of 1065 verb endings has been created. The short endings were added to the database following the same criteria as in the case of nouns and adjectives. The list of endings was supplemented with the endings the first letter in which was “š” (see Chapter “Assimilation rules of the stem endings”). Removing identical endings whose attributes coincide or whose attributes can be combined has decreased the number of endings.

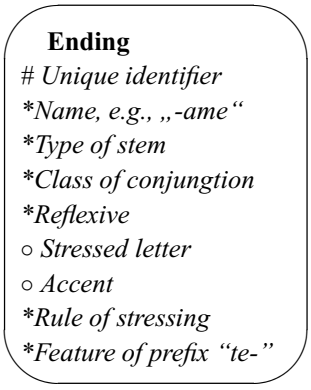

Fig. 4. The ERD of the database of endings of verbs.

4.2.6. *Stems*

The ERD of the database of stems is presented in Fig. 5.

The attribute *Type of stem* defines whether the stem is that of the present tense, the past tense or the infinitive.

The **Ending** attribute *Class of conjugation* is a binary template whose unities are placed in those positions to the stems of classes of conjugation of which this ending can be added to. The **Verb** attribute *Class of conjugation* can have the unity in one position only.

Attributes *Polysyllabic stem in the present tense* and *Polysyllabic stem of the infinitive* can be calculated, therefore it is unnecessary to store them.

To define the stress position in the prefix and the ending a stressed letter is indicated. In the case of stems the stressed syllable will be indicated.

4.2.7. *Common Algorithm of Stressing Verbs*

1. In the database of endings the search for the endings that match the end of the stem are to be made and they are to be removed from the stem (together with the prefix). If the stem ends in a vowel and the ending begins with a vowel, such an ending does not fit. Provided no stem remains after the ending has been removed, such an ending does not fit either. E.g., the word “antis” is identical to the ending “-antis”. The following list is compiled: Stem (with a prefix), Identifier of the ending.

2. The list is supplemented with the entries received after the assimilation rules have been applied: (“dž ” and “č” before the soft ending; “s”, “š”, “z”, “ž” before the endings that begin with the letter “s”; “k” and “g” before the endings that begin with the letter “k”; “ū” and “y” before the ending “-s”. For more details see Chapter “Assimilation rules of the stem endings”).

3. In case the ending is not reflexive, all the groups of prefixes that match the beginning of the word are to be found and removed from the stem. Provided the group of prefixes contains the particle “te” (rather than “tebe”) and the ending does not have any corresponding attribute, such a group does not fit. If no stem remains after the prefix

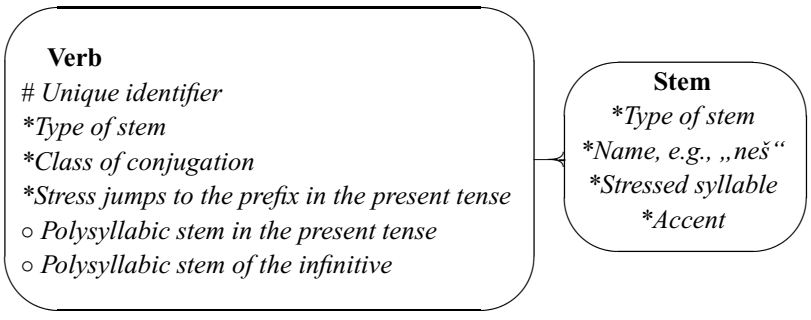

Fig. 5. The ERD of the database of verb stems.

has been removed, such a group is unfit either, e.g., the word “peri”. A new list is to be compiled: Identifier of the group of prefixes, Stem, Identifier of the ending. The list must contain at least one entry without a prefix for each stem.

4. All coinciding stems are found for every entry made in Point 3 of the compiled list of stems. The stem type defined by the ending is to be taken into account. Moreover, the class of conjugation of the stem must coincide with the template of the class of conjugation of the ending. A new list is to be compiled: Identifier of the group of prefixes, Identifier of the stem, Identifier of the ending.

5. Steps 6–8 are to be repeated for all the entries made in Point 4 and in this way to compile a list of stressed words.

6. In case there is a group of prefixes which is always stressed the attribute *Place of stress-prefix* is formed. Move to step 8.

7. The place of the stress is found according to the attribute of **Ending**. *Rule of Stressing*. Some rules form the attribute *Accent*.

8. In case *Place of stress* is a prefix, the word is stressed according to the attributes of the group of prefixes *Stressed letter* and *Accent*. Provided *Place of stress* is a stem, the word is stressed according to the attribute of the stem *Stressed syllable* and the attribute *Accent* which was formed in step 7. In case this attribute has not been formed, the word is stressed according to the attribute of the stem *Accent.* If *Place of stress* is the ending, the word is stressed according to the attributes of the ending *Stressed letter* and *Accent*.

9. The common list of stressed nouns, adjectives and verbs should be verified to make sure that all the words are stressed in the same way. In the event they are not, rules to select a single stressing variant should be applied.

4.3. *Stressing of Non-inflectional Words*

Non-inflectional words and the exceptions of inflectional words should be stored in the database of non-inflectional words.

Non-inflectional words are as follows:

1) nouns, e.g., “foje` ”, “taksi` ”;

2) numerals, e.g., “dešimt”, “dvi` dešimt”, “pusan∼tro”;

3) adverbs, e.g., “dau∼g”, “namo∼”, “ryto´ j”, “ty´ čia”;

4) all particles, e.g., “ti` k”, “beve´ ik”, “da´ r”, “jau∼”, “neben∼t”, “nejau∼gi”;

5) all prepositions, e.g., “dėka` ”, “dė∼lei”, “lin∼k”, “vie∼toj”;

6) all conjunctions, e.g., “tačiau∼”, “arba` ”, “je´ igu”;

7) all interjections, e.g., “a∼čiū”, “dė∼kui”, “sudie∼”, “laba∼nakt”;

8) all onomatopoeic interjections, e.g., “žvi` lgt”, “tri` nkt”.

Exceptions of inflectional words are as follows:

1) some cases of some nouns, e.g., vk and vį cases of the word “petys” (“peties”, “petimi”), vv case of the words “viešpats” and “mėnuo”;

2) forms of the verb “bū´ ti”(“esu` ”, “esi` ”, “e∼sa”, “yra` ”, “e∼same”, “e∼sate”, “e` sti”);

3) the verbs “turi” and “gali” with the prefixes “ne”, “nebe”, “be”.

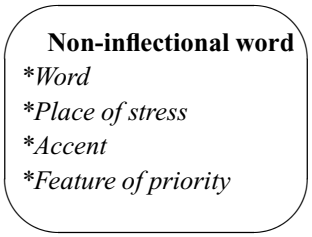

Fig. 6. The ERD of non-inflectional words.

As in the case of stressed nouns, adjectives and verbs, non-inflectional words are entered in the same list of stressed words and only then is one stressing variant selected from the list or the word is left unstressed. Words in the database of the non-inflectional words can be identical to the grammatical form of another word, e.g., the pronoun "me´ s" is identical to the future tense of the verb "me` s". No additional problems are created in this case. Simply some entries appear on the list of stressed words. Problems arise when the word is an exception to a general rule with respect to its place of stress. E.g., according to the rules of stressing verb, the verb "negali" is incorrectly stressed "ne` gali", therefore, a correctly stressed form "nega∼li", as an exception, is entered in the database of non-inflectional words. However, how is it possible, with these two words present on the list of stressed words, to establish which of the stressing variants is incorrect? The simplest way to do that is to store one more attribute – the feature of priority – together with such exceptions.

The structure of the database of non-inflectional words in terms of the ERD is presented in Fig. 6.

5. Results of Experiments and Directions in Future Work

All algorithms mentioned above were realised in the form of computer programs. The following databases were created: 8765 stems of verbs, 53277 stems of nouns and adjectives, and 2306 non-inflectional words. I thank E. Mitašiūnaitė and V. Zinkevičius for their help by compiling these databases. Most of the words were taken from the Dictionary of Modern Lithuanian, 1993 and the Dictionary of International Words, 1985. To establish the reliability of the algorithms presented in this paper, experiments were carried out with the texts covering about two pages of social and political journalism and fiction. In case the word could be stressed in more than one way, no algorithm to select one variant was applied and the word was left unstressed. Results are presented in Table 3.

Names, surnames, names of places, diminutives, adjectives with prefixes were not found in the database.

Directions in the future work:

1) to analyse the formation of words (first and foremost, that of nouns and adjectives) by means of prefixes and suffixes. This would enable the volume of the database to be

Table 3

Results of experiments

	Stressed correctly	Stressed incorrectly	Unstressed. Not found in the database	Unstressed. Many stressing variants	The total words
Social and political journalism	82.57 %	0 %	3.67 %	13.80 %	413
Fiction	81.53 %	0.20 %	1.20 %	17.07 %	498

reduced, there would be no necessity to predict all possible words to be formed by means of a certain suffix or prefix;

2) to analyse algorithms that make it possible to select one stressing variant;

3) to supplement the database of stems with names, surnames, and names of places.

References

Ambrazas, V., K. Garšva, A. Girdenis *et al.* (1996). *A Grammar of Modern Lithuanian*. 2nd ed. Mokslo ir enciklopedijų leidykla, Vilnius (in Lithuanian).

Barker, R. (1994). *CASE Method: Entity Relationship Modeling.* Oracle, Wokingham.

Dictionary of International Words (1985). Vyriausioji enciklopedijų redakcija, Vilnius (in Lithuanian).

Dictionary of Modern Lithuanian (1993). Mokslo ir enciklopedijų leidykla, Vilnius (in Lithuanian).

Girdenis, A. (1995). *Theoretical Fundamentals of Phonology*. Vilniaus Universitetas, Vilnius (in Lithuanian).

Kasparaitis, P. (1999). Transcribing of the Lithuanian text using formal rules. *Informatica*, **4**(10), 367–376.

Nebbia, L. (1990). Text-to-speech synthesis system for Italian: an overview. *CSELT Technical Reports,* Vol. XVIII, No. 2.

Paulus, E. (1998). *Speech Signal Processing: Analysis, Recognition, Synthesis*. Spektrum Akademischer Verlag, Heidelberg (in German).

Vaitkevičiūtė, V. (1997). *Stressing of Common Lithuanian.* Šviesa, Kaunas (in Lithuanian).

P. Kasparaitis was born in 1967. In 1991 he graduated from Vilnius University (Faculty of Mathematics). In 1996 he has been admitted as a PhD student in Vilnius University. Current research interests include text-to-speech synthesis and image processing.

Lietuvių kalbos teksto automatinis kirčiavimas remiantis žodynu

Pijus KASPARAITIS

Šiame darbe nagrinėjama viena iš lietuvių kalbos sintezės pagal tekstą sudedamųjų dalių – automatinis teksto kirčiavimas. Darbe pagrįsta būtinybė skaidyti žodžius į pastovias ir keičiamas žodžio dalis, iš kurių sudaromos įvairios gramatinės formos, ir skirtingose bazėse saugoti tik šias dalis, o ne visus žodžius. Pagal kaitymo būdą visi lietuvių kalbos žodžiai suskirstyti į tris grupes (daiktavardžius-būdvardžius, veiksmažodžius ir nekaitomus) ir atskirai išnagrinėta kiekviena iš jų. Kiekvienai grupei nustatyta, kokia informacija ir kokiu pavidalu turi būti saugoma bei pateiktas algoritmas, kaip pagal šią informaciją atpažinti žodžio gramatinę formą bei ją sukirčiuoti.