REFLECTING ACCENTUATION IN THE RUSSIAN MORPHOLOGICAL DICTIONARY OF THE MULTIFUNCTIONAL LINGUISTIC PROCESSOR ETAP-3[1]

V. G. Sizov (sizov@iitp.ru)

O. Iu. Podlesskaia (olga@iitp.ru)

Laboratory of Computational Linguistics, Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russian Federation

Our work is aimed at the introduction of accentual information into the morphological dictionary of the multifunctional linguistic processor ETAP-3. A special formal description language has been created, and special rules for most of the basic accentual schemes have been designed. Special algorithms have been written for morphological analysis and synthesis.

Key words: ETAP, ETAP-3, accent, accentuation, morphological dictionary.

1. Introduction. Problem statement

Accentual information (information about the location of accent, or stress, in word forms) in morphological dictionaries of text processing systems allows solving important and useful tasks. First, this information is necessary for high-quality speech synthesis, especially combined with means of homonymy disambiguation (disambiguation of word forms that have the same form but different readings, e, g., *vse* ‘everybody’ vs *vsjo* ‘everything’). Accentual information may also be used for automatic disambiguation in texts tagged with stress diacritics and in accentuated corpora.

Most text processing systems that operate with accentual information for Russian words are based on the grammatical dictionary of Russian language (GD) by acad. A. A. Zaliznyak [1]. Among these systems are morphological processors Dialing [2], Starling (Starostin˚S. A., [3]), text-to-speech synthesis system «Multifon» (Lobanov B. M, [4]), and some others. In modern corpus Slavistics the stress-tagging problem was stated in the Russian national corpus (RNC) (see [5]). RNC has accentual tagging: disambiguated texts have been automatically tagged with stresses[2].

[1] This work has been supported by Russian foundation for Basic Research (grant № 10-07-90001 Bel_a).

[2] The accentual tagging was done using programs Mystem (Yandex, [6]) and Dialing.

While working on the international project “Intellectual speech synthesis model based on deep linguistic text analysis”, it was decided to enrich the morphological dictionary of multifunctional linguistic processor ETAP-3 with accentual information [7]. It was found that in order to reduce the number of errors during speech synthesis, linguistic processor ETAP-3 should be able to disambiguate homonymy, and place accents during morphological synthesis. So, among the goals of the projects was a specific task — reflecting accentuation in the morphological dictionary of ETAP-3.

2. Main features of the formal morphological model of ETAP

The morphological part of text processing systems is used for morphological analysis of the input text and for morphological synthesis of the output text. The aim of the morphological analysis is to find all possible morphological sets (lexeme names and corresponding morphological characteristics) for all word forms in the input text. The aim of the morphological synthesis is to generate the word forms from lexeme names and their morphological characteristics. To improve the performance of analysis and synthesis components, the dictionary compiler generates all possible word forms with all possible characteristics for the forms that have entries in the dictionary[3], after that the pairs «lexeme + word form with morphological features» that form the paradigm are stored in the final-state transducer [10].

Russian is well-known for its rich morphology, and manual introduction of all word forms with stresses is hardly possible. Therefore, inflection description in the morphological dictionary assumes that for each word form there should be specified (a) non-changeable **base** — a part of word that is common for the word forms in the paradigm and (b) inflectional parts of the word: **prefixes**, **themes**[4], **suffixes**, **endings** and **particles** (*-sja/-sj*). Regular sequences of inflectional parts are described using standard objects (STO). A simple STO may be a part of one or more complex STO. In most cases, to create a paradigm of a word one needs to specify the base and provide a reference to one or more standard objects. At present, Russian morphological dictionary of ETAP-3 contains about 130,000 dictionary entries and nearly 1000 standard objects.

The introduction of accentual information in the dictionary assumes highlighting of the stressed vowels with the sign of reversed accent (`), if the stress is strong, or with the tilde (~), if the stress is weak, and also the interchange between «e» and «jo». This task is far from trivial because there are three different types of stress alternations in Russian:

[3] In the ETAP-3 system each dictionary entry describes one lexeme.

[4] The theme is the first verbal affix after the root in morphological dictionary of ETAP-3 system, for example: *ris-**u**-ju ‘drawing’, ris-**ova**-nn-yj ‘painted’, tolk-**nu**-t’ ‘push’.*

- Alternations in STO, for example, in the endings:(*mope`d-**ami** ‘by motorbikes’ / stol-**a`mi** ‘by tables’*); in suffixes (*umn- **e`jsh**-ij ‘the most clever’ / razu`mn-**ejsh**-ij ‘the wisest’*);
- Alternations within the paradigm of the lexeme: ***kra`sn**-yi ‘red’ / krasn-**éjsh**-ij ‘the reddest’ / krasn-**a`** short feminine form for ‘red’.*
- Alternations within a morpheme: *risk-**ova`**-tj ‘risk’ / risk-**o`va**-nn-yj ‘risky’.*

Among typical Russian alternations is also «e» / «jo» alternation in morphemes (*kon-jo`m ‘by horse’ / lo`s-em ‘by elk’*) and in paradigms (*jozh ‘hedgehog’ / ezha` ‘(without) hedgehog’*). It is almost always connected with the stress location: accentuated vowel is pronounced as *jo* and vowel without accent as *e*.

Due to diversity of such alternations (and *e/jo* alternations), introduction of the stress (and also the letter *jo*) to the bases and inflectional parts of the words will require that the existing STO should be divided and, consequently, necessitate the changes in every dictionary entry that may need years to be completed. To avoid this, a description of stress alternations that requires minimal changes in existing STO and MD entries should be used.

In GD, where declination / conjugation and accentual schemes are described separately, regularities of morphological paradigms and regularities of stress alternations within the paradigm are, generally speaking, uncorrelated. This fact allowed us to compile the morphological entry in two steps: at the first step the paradigm is constructed with no regard to the stress while at the second step the accents are assigned. The analysis of the accentual patterns has shown that in order to place the stress one needs:

- The accentual pattern of the word;
- Morphological characteristics of the word form;
- Morphemic structure of the word form (e. g., in the word form *doroga* ‘road’ the first five letters make up the base, and the sixth letter is the ending);
- Stress location for morphemes consisting of more than one syllable (*doro`g-a*);
- (in some cases) Letters that build up a morpheme (*vetr-y* ‘winds’ vs. *vetr-a* ‘winds’);
- (in some cases) The type of declination / conjugation according to GD.

This implies that the current set of STO can practically remain unchanged, only accents in polysyllabic morphemes should be set, and e should be replaced by jo, where needed. Then the final accent setting and the choice of e/jo will be made in the already built paradigm, so that all information required for this will be obtained from this paradigm.

3. Accentuation rules

The conditions that have to be met in order to assign the correct accentual scheme are checked by the special rules. These rules are language-independent, which allows using them for accentuation in other languages. Formal language of linguistic rules ETAP-3 FORET was taken as a base for the formal language needed for the accentual rules.

As rules of linguistic processor ETAP-3, the accentual rules contain logical expressions that check truth of the conditions in them and the instructions that are fulfilled if the check has shown the truth of the conditions. The logical expressions contain predicates united in conjunctions, disjunctions and parentheses expressions. Rules may be united in blocks of rules. A block of rules consists of head word and then (optionally) name of block and the list of rules. Blocks that have names are called named blocks, other blocks are called unnamed blocks.

We will look at rule specification and the working algorithm in more detail.

3.1. Instructions

Accentual rules' **instructions** are of two types: stress location and stress location shift. The instructions of the first type set preliminary accents in word forms on the morpheme specified by instruction. Their names (*pref:*, *osn:*, *tm:*, *sf:*, *ok:* and *chs:*) correspond to the type of the morpheme that bears the stress. Stress location may be specified in the instruction by pointing out the number of the stress-bearing syllable. If the stress location within the morpheme is not pointed out explicitly, then the instruction preserves the original stress location in the morpheme. If the stress location is pointed out explicitly, then the instruction deletes the stress and sets the new stress in the position mentioned. The stress occurring in all other morphemes of the word form is deleted. If the morpheme mentioned in the instruction lacks vowels or is absent in this word form altogether, then stress is located on the last syllable of the preceding morpheme. The instruction may perform additional actions that change the appearance of the word forms, such as weak stress instead of strong stress or change of accentuated vowel (for example, replacement of the stressed *e* by *jo*[5]. The stress location change instructions make corrections to the previously located stress — they shift it a number of syllables to the left or right. If the number is positive, there is a right shift, if it is negative, then the shift is to the left.

Each instruction has a **priority —** an integer-valued characteristic that is assigned explicitly or implicitly. If during compilation a word form fulfills the conditions of several rules, then only those rules are valid that have the instructions with higher priorities.

The accentual system allows that several rules operate on the same word form, and the stress is located in various positions. In this case several copies of one word form are created, and each copy is treated by a separate rule. Such mechanism describes alternative accentuation (for example, *tvoro`g / tvo`rog 'curds', kazaki` / kaza`ki 'Cossacks'*). Each word form of the lexeme *tvorog* is treated by two rules: one sets stress

[5] In most cases *e / jo* alternations obey the following rule — «*jo* is stressed, *e* is without stress». The option "change symbol" allows describing non-standard alternation in such word forms as *izrjok* — *izrekshij,* where *e* is also stressed although it is not converted into *jo*.

on the ending, and the other on the base. Word form *tvorog* in nominative singular has a zero ending, that is why an accent ascribed by the second rule is moved to the last syllable of the base.

3.2. Predicates

Predicates that are part of conditions in accentual rules are divided into predicates that check morphological characteristics and predicates that check the string of letters in the morphemes. Predicates of the first type coincide with the name of the checked characteristic and are true if this characteristic is in the list of characteristics of this word form. Predicates of the second type check if there is a certain morpheme in the word form, and (if there is one) — they check a symbol string that is in the regular expression in the predicate.

3.3. The scope of the rules

The scope of the rules may vary from a unique dictionary entry to the whole dictionary. To simplify the description of the scope, the rules were divided into **general**, **template**[6] and **dictionary rules**.

General rules are applied to word forms of all dictionary entries and stored in a special file along with other language specificities.

Template rules are applied to subsets of entries, from twenty up to several thousand elements (as a rule these are entries with the same accentual schemes). They are stored in the file of standard objects under the standard objects class named *acct*. These subsets mostly correspond to the main accentual schemes from the Grammatical dictionary by acad. A. A. Zaliznyak. The entries to which the template rules are applied contain references to these templates.

Dictionary rules are applied to the specific dictionary entries and stored directly in the entries.

3.4. The algorithm of processing the accentual rules

General, template and dictionary rules are applied to the paradigm of the lexeme, consequently to each word form: at first the stress location rules, then stress location shift rules. The priority of the rules is also taken into account: at first instructions with the highest priority are applied.

[6] This type of rules has been named in the same way as the syntactic rules of a similar type in the ETAP-3 processor [3],[4].

3.5. Accentual rules. Examples

We will illustrate how the rule functions by the exampled of the lexeme *TRAKTOR* ‘tractor’. This lexeme has a stress on the first vowel of the base in singular and in plural for the form *traktory*, while for the alternative plural form *traktora* the stress is on the ending, and the stress location varies in other plural forms.

The dictionary entry of the accentuated morphological dictionary for this lexeme looks as follows:

(1) ENTRY:TRAKTOR acct:c_a
base:tra`ktor f:1,end:'a'nom,pl,'a'acc,pl t:6

At the first compilation stage, the lexeme paradigm will look like:

(2) *tra`ktor — S,SG,MASC,NOM,INAN*
tra`ktor|a — S,SG,MASC,GEN,INAN
tra`ktor|u — S,SG,MASC,DAT,INAN
tra`ktor — S,SG,MASC,ACC,INAN
tra`ktor|om — S,SG,MASC,INS,INAN
tra`ktor|e — S,SG,MASC,LOC,INAN
tra`ktor|y — S,PL,MASC,NOM,INAN
tra`ktor|a — S,PL,MASC,NOM,INAN
tra`ktor|ov — S,PL,MASC,GEN,INAN
tra`ktor|am — S,PL,MASC,DAT,INAN
tra`ktor|y — S,PL,MASC,ACC,INAN
tra`ktor|a — S,PL,MASC,ACC,INAN
tra`ktor|a`mi — S,PL,MASC,INS,INAN
tra`ktor|ah — S,PL,MASC,LOC,INAN
tra`ktor|o — S,MASC,INAN,COMP

(the last line corresponds to the word form with the tag COMP, which is used in composite words such as *traktorostroenie* ‘manufacturing tractors’).

After that for each word form of this paradigm the conditions of those rules are checked that correspond to the nouns, in particular:

(3)

- base:(0,"*~")=COMP+^V; [The stress in COMP is always on the base and weak]
acct:c_a [tra`ktor, traktora`, tra`ktory]
- base:=S;
end:=PL+^(NOM|ACC+INAN);
end:{2}=PL+(NOM|ACC+INAN)+search("a",end:)

The word form tra`ktor|o — S,MASC,INAN,COMP fulfills the conditions of the template rule base:=S; and the basic rule base:(0,"*~")=COMP+^V; While the priority of the basic rules is higher, the word form is constructed with the instruction _base:(0,"*~"),_ that changes strong stress in the base to the weak one:

tra~ktoro — S,MASC,INAN,COMP

The word forms *tra`ktor|a — S,PL,MASC,NOM,INAN, tra`ktor|a — S, PL, MASC, ACC, INAN* fulfill the conditions of the template rule *base:=S;* and the dictionary rule *end:{2}=PL+(NOM|ACC)+search("a",end:).* The priority of the dictionary rule is higher, therefore the stress in these forms will be on the endings:

traktora` — S,PL,MASC,NOM,INAN
traktora` — S,PL,MASC,ACC,INAN

The remaining word forms with characteristic PL satisfy the conditions of the template rule *base:=S;* and dictionary rule *end:=PL+^(NOM|ACC);* while the word forms with characteristic SG only satisfy the rule *base:=S.* Therefore word forms with characteristic PL will be duplicated, the stress in one copy will be placed on the first vowel of the base, and the stress in the second copy will be placed on the ending. Word forms marked with SG fulfill only the rule “base:=S”, and the stress will be placed on the first vowel of the base.

4. Automatical introduction of accentual information to MD of ETAP-3

Introduction of accentual information to MD will be done mostly automatically. Rules for accent setting in accordance with accentual schemes used in GD were written and tested. The correspondences between MD and GD entries were established. The last step of introduction consists in supplying MD entries with accentual rules that correspond to the accentual schemes in GD entries.

4.1. Working on main accentual rules

While writing the rules describing accentual schemes used in GD, some problems had to be solved. Most of them are concerned with systematic discrepancies between MD and GD dictionaries.

First, the different aspect verb forms in MD are usually merged into one entry, while in GD they are always separated. To place the accents correctly two rules were used. Each rule was applied only to the word forms of one aspect (it was done so by mentioning the aspect in the rule conditions).

The second systematic difference between the dictionaries is the lack of the comparative and superlative degrees in the paradigm of the adjectives and adverbs in GD.

Accentual schemes for adjectives in GD were supplied with additional accentual rules for the comparative and superlative degrees and adverbs were also fully accentuated.

When we studied the group of these word forms, we could establish a new regularity: adjectives with the stress located on the base (this is one of the accentual schemes) retain it in the forms of superlative. The only exception seems to be *boga`t-yj — bogat-e`jsh-ij*. Special complementary rules were added to the adjective templates for correct accentuation in the comparative and superlative degree forms in MD.

More problems arise in case of several GD entries being merged in one MD entry because of synonymy (GD entries *chitat' 'read'/ prochitat' 'have read'/ prochest' 'have read'* correspond to MD entry *CHITAT'*, GD entries *povorachivat' 'turn'/ povertyvat' 'turn'/ povernut' 'have turned'/ povorotit' 'have turned'* correspond to MD entry *POVORACHIVAT')*. In such entries the word forms with the same sets of morphological characteristics often correspond to different accentual schemes. This may complicate the conditions of the rules in accentual schemes. To solve this problem we have inserted a special separator that divides morphological dictionary entry into parts that correspond to GD. These parts include corresponding STO and stress location rules, therefore a separate compilation and stress location is possible, and after that separate word forms are merged into one paradigm.

4.2. Working on dictionary entry correspondence tables

At present, the table of corresponding accentual schemes and rules and the table of corresponding entries have been made. For the latter table a program was created that finds the correspondence between MD and GD entries, between such morphological characteristics as part of speech, gender, animacy, and so on. There are nominal lexemes in GD that have both masculine and feminine and animated and non-animated forms in the (*zanuda 'bore'* ma//fa, *mikrob* '*microbe'* m//ma), while in MD it is not so. Conversely, in MD many verb entries merge verbs of two aspects. The found pairs were combined into larger groups.

Morphologically homonymous MD entries (for example, DERZHATEL' 'a man / a device', or USTANOVKA — 'installation' — 'action' vs. 'object') were combined in a table of «many-to-many» correspondence. The created tables need post-editing for making the correspondences «one-to-one».

Accentual information from GD was transferred to those entries in MD that have been put in the «one-to-one» correspondence table. During this transfer in the dictionary entry of MD: 1) an accentual rule was assigned that corresponded to the accentual scheme from GD; 2) a base was accentuated, and the stress location was defined according to the entry in GD. As a result, 65 000 entries were given accentual information (there are about 129 000 entries in MD altogether). The transfer of accentual information from other tables is also in process now.

Words that have not been put in the tables were accentuated manually. Among these words are 10 000 adverbs, the most part of which is absent in GD.

4.3. Creating the set of accentual rules for adjectives

The main principles of accentual rules may be shown on the example of accentual rules for adjectives. In GD there are 12 most frequent accentual schemes: *a, a’, a/b, a/b’, a/c, a/c’, a/c”; b, b’, b/c, b/c’, b/c”*. The first letter shows accentual scheme for full forms, the second letter — for short forms. If the second letter is the same as the first, it must be omitted.

Accentual scheme *a* means that word forms have stress on the base and accentual scheme *b* means that stress is always on the ending, excluding comparative and superlative that are not part of the paradigm in GD. Scheme *c* (only for short forms) means that accent falls on the ending in feminine singular forms. According to these schemes accentual rules may be built.

At first accents for full and short forms as in GD are created:

Rule adj_a (corresponding to scheme *a*): the accent falls on the base

(4) *acct:adj_a [suro`v-yj ‘severe’, udo`bn-yj ‘comfortable’]*
base:=A;

This rule is applied to adjectives in MD that have the scheme a in GD.

Rule adj_a1 (corresponding to scheme a’): the accent falls on the base while for BREV+FEM accent also falls on the ending:

(5) *acct:adj_a1 [vla`stn-yj ’powerful’ (vla`sten, vla`stn-a/vlastn-a`, vla`stn-o, vla`stn-y) end:=BREV+FEM;*

base:=A;

Rule adj_ab (scheme a/b) prescribes that the accent is placed on the ending in short forms and on the base in all other forms:

(6) *acct:adj_ab [zdoro`v-yj (zdoro`v, zdorov-a`, zdorov-o`, zdorov-y`)]*
end:=BREV;
base:=^BREV;

Other rules are listed without detailed description:

(7) *acct:adj_ab1 [sve`zh-ij ‘fresh’ (sve`zh, svezh-a`,svezh-o`, sve`zh-i/svezh-i`);*
scheme ***a/b’]***
end:=BREV;
base:=^(SG+BREV);

acct:adj_ac [tse`l-yj ‘whole’ (tse’l, tsel-a`, tse`l-o, tse`l-y); scheme ***a/c]***
end:=BREV+FEM;
base:=^(BREV+FEM);

acct:adj_ac1 [ми́л-ый ‘dear’ (mi`l, mil-a`, mi`l-o, mi`l-y/mil-y’); scheme ***a/c’]***

end:=BREV+(FEM|PL);
base:=^(BREV+FEM);

acct:adj_ac2 [be`l-yj 'white' (bel, bel-a`, be`l-o/bel-o`, be`l-y/bel-y`) scheme ***a/c"]***
end:=BREV+^MASC;
base:=^(BREV+FEM);
acct:adj_b [smeshn-o`j 'funny' (smesho`n, smeshn-a`, smeshn-o`, smeshn-y`); scheme ***b]***
end:=A;

acct:adj_bc [zhiv-o`j 'alive' (zhiv, zhiv-a`, zhi`v-o, zhi`v-y) ; scheme ***bc]***
base:=BREV+(NEUTR|PL);
end:=^(BREV+(NEUTR|PL));

acct:adj_bc1 [skup-o`j 'stingy' (skup, skup-a`, sku`p-o, sku`p-y/skup-y`);
scheme ***bc']***
base:=BREV+(NEUTR|PL);
end:=^(BREV+NEUTR);

acct:adj_bc2 [dryann-o`j 'bad' (drya`nen, dryann-a`, dryann-o`/ drya`nn-o, dryann-y`/drya`nn-y) ; scheme ***bc"]***
base:=BREV+^FEM;
end:=^(BREV+MASC);

For the comparative and superlative word forms, general rules are written (because they are applied to all schemes excluding ***a***):

- a rule that places stress on the base in the comparative forms (*bo`l'-she 'bigger', glu`b-zhe 'deeper'*)

(8) *base:=COMPAR+search("^e$",sf:);*

- a rule that places stress on the suffix in the comparative forms (*smel-e`e 'more boldly', vesel-e`j 'funnier'*)

(9) *(9)* sf:=COMPAR+search("^é?[ей]$",sf:);[7]

- a rule that places stress on the suffix in the superlative forms (*velich-aj`sh-ij 'the greatest', umn-e`jsh-ij 'the cleverest'*)

(10) *sf:=SUPER*.

[7] The regular expression "^é?[ей]$" means that the first character of the line has to be *е* followed by an optional stress symbol, while the last character of the line is *е* or *й*.

These rules are applied to almost all adjective accentual schemes (11 in GD), except scheme *a*, where the stress is always on the base. For this reason, a template rule for *a*-adjectives should have a higher priority (*acct:adj_a base:{2}=A*).

By default, the general rules have a higher priority than the template rules. That is why the instructions *adj_a — adj_bc2* are not applied to the comparative and superlative word forms, even if these word forms meet the conditions of these rules. This is true for adjectives with accentual schemes ***a' — b/c"***, and not true for adjectives with accentual scheme ***a,*** where the accent falls on the base also in the comparative and the superlative. To cancel the action of the general accentual rules mentioned above, the template rule for scheme ***a*** is assigned a higher priority than that of the general rules:

(11) *acct:adj_a [суро́в-ый, удо́бн-ый]*
*осн**:{2}**=A;*

However, among the adjectives of accentual scheme a, there is an exception: the word *bogaty 'rich'* that has a superlative *bogat-e`jsh-ij* with the stress on the suffix, rather than on the base. For a correct description of the word forms in the superlative, the dictionary entry:

(12) *ENTRY:BOGATY acct:adj_a* ***acct:sf:{3}=SUPER;***
base:boga`t no:compar t:211 har:A,compar,osn:boga`che
har:A,compar,att,base:poboga`che

This rule has a higher priority yet, 3, whereby the accent set for scheme a of a word *bogatyj* by the rule *acct:adj_a* is suppressed.

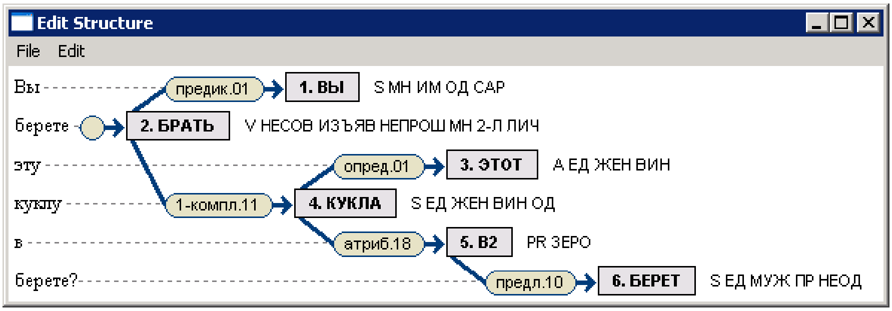
5. *Morphological analysis and synthesis using accentual information*

New requirements to the morphological component of ETAP-3 must be met while adding accentual information:

- morphological analyzer must recognize word forms from the input text both with accents and *jo* and without;
- morphological analyzer must disambiguate strictly and operate with texts that have consequent *jo* distinction. This mode should not confuse *e* and *jo*: (*on osel 'he has collapsed' ≠ on osjol 'he is an ass'*);
- morphological synthesizer must generate output text both with *jo* and accentuation and without it;
- when word forms with alternative accentuations are produced (for example, *profe`ssoram / professora`m '(to) professors'*), it should be possible to have the most useful word form (implicitly) or the word form you need.

Let us look at the new possibilities of ETAP-3 with accentual information in more detail. The syntactic analyzer is used for disambiguation also in a special pre-processing mode of speech-synthesis. Phrases are assigned with syntactic structures and disambiguated in this mode. They are also accentuated and sent to speech synthesizer.

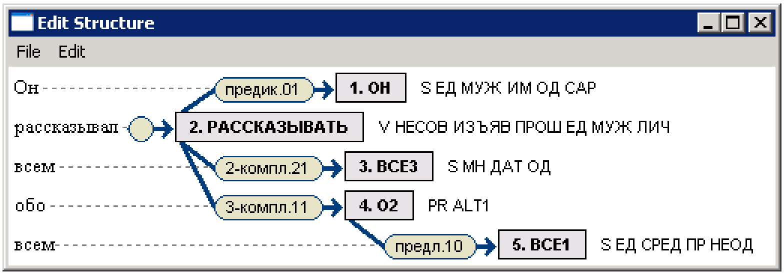
For the phrase «*Vy berete etu kuklu v berete*?» 'Are you taking this doll in a beret', the text-to-speech synthesizer Multifon will mistake the second occurrence of the word form *berete* 'beret' for the verb *berjote* 'take' But with the ETAP-3 pre-processing mode activated we will have the correct syntactic structure of this sentence:



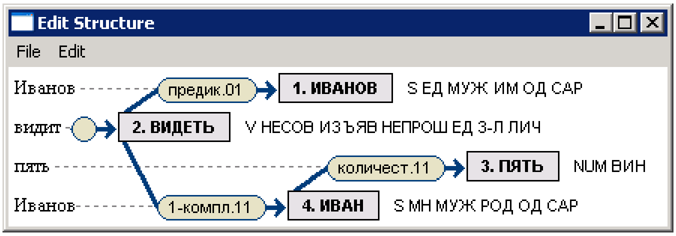
Pic. 1. Phrase "Vy berete etu kuklu v berete?"

Then the morphological synthesizer produces accentuated and disambiguated sentence from this structure: *"Vy` berjote e`tu ku`klu v bere`te?"*.

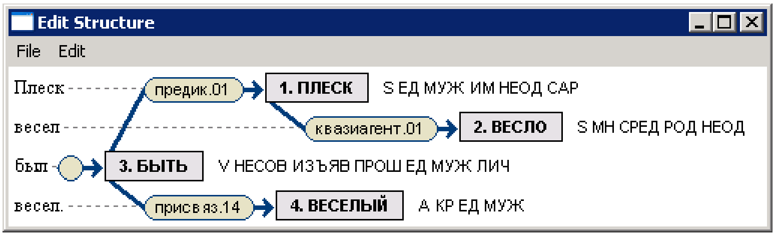
Below are several additional examples of homonymy in phrases: 1) *On rasskazyval vsem obo vsem* 'He told everybody about everything'; 2) *Ivanov vidit pjatj Ivanov* 'Ivanov sees five Ivans'; 3) *Plesk vesel byl vesel* 'The splash of the oars was merry'. ETAP-3 produces correct syntactic structures:



Pic. 2. Phrase "On rasskazyval vsem obo vsem"



Pic. 3. Phrase "Ivanov vidit pjatj Ivanov"



Pic. 4. Phrase "Plesk vesel byl vesel"

The morphological synthesizer produces accentuated disambiguated texts for these structures: *"O`n rasska`zyvajet vse`m o`bo vsjo`m" / "Ivano`v vi`dit dvu`h Iva`nov" / "Ple`sk vjo`sel by`l ve`sel"*. This text is then sent to the speech synthesizer.

6. Conclusion

As a result, in place of the morphological dictionary of ETAP-3 system we will have a fully accentuated large-size Russian morphological dictionary supplied, among other things, with completely formalized rules for creating accentual paradigms of new lexemes. Another important feature of this dictionary is a sufficiently full accentuation coverage of degrees of comparison for adverbs and adjectives. The accentual data on these word forms in other dictionaries is rather scarce. The accentual information from the dictionary will allow one to use the syntactic analyzer for the disambiguation during speech synthesis. The information on the accents may also be used for accentual tagging of the Syntactic corpus of Russian language (SynTagRus).

The main factor that facilitated the introduction of accentual information into the large dictionary was the creation of a formal language for describing this information. This language is entirely independent from the formalism used for the description of paradigms. The mechanism of automatic transfer of the data from GD into MD was essential, too. Such formal language may be effectively used for the introduction of accentual information into morphological systems for Russian that use the formalism different from that of GD. Due to its maximum linguistic independence, this formal language may be used to create accentual rules for other natural languages besides Russian.[8]

References

1. *Apresian Iu. D, Boguslavskii I. M., Iomdin L. L., Lazurskii A. V., Mitiushin L. G., Sannikov V. Z., Tsinman L. L.* 1992. Linguistic Processor for Complex Informative Systems [Lingvisticheskii Processor dlia Slozhnykh Informatsionnykh Sistem].
2. *Apresian Iu. D., Boguslavskii I. M., Iomdin L. L., Lazurskii A. V., Pertsov N. V., Sannikov V. Z., Tsinman L. L.* 1989. Linguistic Supply for ETAP-2 [Lingvisticheskoe Obespechenie Sistemy ETAP-2].
3. *Grishina E. A.* 2009. The “History of Russian Accent” Corpus [Korpus «Istoria Russkogo Udarenia»]. Natsionalyi Korpus Russkogo Iazyka. Novye Rezultaty i Perspektivy.
4. *Iomdin L. L., Lobanov B. M.* 2009. Syntaxic Correlates of Prosodic Marked Sentence Elements and its Role in the Synthesis of Text-to-speech [Sintaksicheskie Korreliaty Prosodicheski Markirovannykh Elementov Predlozheniia i ikh Rol’ v Zadachakh Sinteza Rechi po Tekstu], available at: http://www.dialog-21.ru/dialog2009/materials/html/23.htm
5. *Kazennikov A. O.* 2008. The Use of Final Automatons dor Morphological Analysis and Synthesis basing on ETAP Dictionaries [Ispol’zovanie Konechnykh Avtomatov dlia Morfologicheskogo Analiza i Sinteza na osnove Slovarei Sistemy ETAP]. Sbornik Trudov 31 Konferentsii Molodykh Uchenykh i Specialistov IPPI RAN "Informatsionnye Tekhnologii i Sistemy " (Proc. of the 31 Conference "Information Technologies and Systems"): 201–205.
6. *Lobanov B. M.* 2007. «Muiltifon» — a Personalized Text-to-speech Synthesis System for Slavic Languages. Linguistic Polyphony : 849–866.
7. *Segalovich I.* A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. MLMTA-2003, available at: http://download.yandex.ru/company/iseg-las-vegas.pdf
8. *Sokirko A.* 2001. A Short Description of Dialing Project, available at: http://www.aot.ru/docs/sokirko/sokirko-candid-eng.html
9. *Zalizniak A. A.* 2007. Russian Grammar Dictionary [Grammaticheskii Slovar’ Russkogo Iazyka].
10. *www.starling.rinet.ru*

[8] It is evident that the introduction of accentual information into the morphological dictionary is on the one hand useful for the language orthographies in which the letter structure is close to the phonemic structure. In this case, the information on the word’s letter structure and the stress location is sufficient for speech synthesis On the other hand, identifying the stress location in this language should not be too trivial, like in Spanish, where the stress location is determined by means of simple rules and is explicitly marked in exceptions. Among such languages with non-trivial stress location are besides Russian, Byelorussian, Ukrainian and most probably German.