Reports of Enlarged Session of the
Seminar of I.Vekua Institute
of Applied Mathematics
Vol. 19, N1, 2004

RECOGNITION OF GEORGIAN WORDFORMS AND THEIR MORPHOLOGICAL CATEGORIES BY COMPUTER

Antidze J., Mishelashvili D.

I.Vekua Institute of Applied Mathematics
Tbilisi State University

Received in 15.05.04

The work describes how to devide Georgian wordforms on morphemes and identify morphological categories for them by help of computer. The aim of our work is to create such computer program, which would be optimal with regard of problem resolution speed and the volume of used computer memory (the priority was given to the speed of problem resolution). To satisfy the aim we used morphological analyzer of the natural language [1].

The algorithm for identification of morphological categories of verbs is based on the classification of verbs, which is described in the work [2]. The creation and fast realization of the algorithm becomes also simpler by separation of the problem of verbform division on morphemes from the verification of constraints, which should be satisfied by the morphemes. The algorithm is universal, because it gives all possible variants of morphemes for particular wordform and by adding constraints on appropriate place; we can reject false combination of morphemes in time. All these make the process of search much faster. The algorithm is realized by using of special instrumental tool [1].

The algorithm provides only division of a wordform on morphemes and finds corresponding morphological categories. The morphemes, which might meet in a wordform, are divided on classes according to the wordform speech part. For example: in case of noun we have the following classes: the part (parts) of a noun, which are not changed by declination and number; morphemes of declination; signs of number; prepositions and etc. In case of verbs, together with other morphemes, as a lexical entry, the root of verbs is used, not the stem. The stem is find by help of diathesis, classes, vowel prefixes and signs of themes, so as it is described in [2].

Every morpheme in the corresponding file is accompanied by the features and is presented by names and values. For example, one of the features for the stem of noun is the particularity of declination and number for stems. According to this feature, stems are divided on different types. In this case value of the feature will be the number designating the type. If a root of a verb belongs to more than one diathesis or classes, such root is repeated in the root's file. In other cases all roots are unique in the root's file.

The constraints, which are putted on wordform morphemes, are logical expressions in which the features and their values are participating. Often constraints differ only from each other by features and their values. By help of computer could be created the text of such constraints. To achieve this we use well known method of macrogeneration, which makes the speed of the constraints writing process much faster.

Let’s look in more details on recognition of Georgian verbforms. In case of verbforms we have the following possible sequence of morpheme classes: preverb, person sign prefix, vowel prefix, root, d-passive, contact, sign of theme, rank, person sign suffix and number. In a verbform must be one representative of each class in the above order. Of couse some classes can also not be represented by their elements in a concrete verbform. In this case we mean zero (empty) morpheme of this class. In contact class we included the repetition of sign of theme, for example, ebin morpheme. We do the same for some verbforms - already existing suffixes are added to person sign suffix class. The aim of this is to reduce the number of morpheme classes. This is possible in cases, where this addition does not create any problem for realization of our aim. Increasing of number of morpheme classes will lead to the increasing of the time for finding morphemes in a verbform significantly.

For creation of constraints, we should give to every morpheme the feature, that is typical for it and different values should be given to this feature according to the different situation, where the morpheme is used. For example, if we take the conjugation of a verb in person and number, we must have a feature for corresponding verbroot, which indicates a particularity of it in the conjugation. Having such features and their values for morphemes gives us the opportunity to create constraints for concrete use of this particular morpheme. The way, how effectively we will be able to create the features and their values will determine the correct creation of constraints and this, on the other hand, will influence the speed of correct determination of morphemes in wordforms. The creation of constraints for different paradigms is routine work, because the same constraints can be repeated in different paradigms or these constraints can differ from each other only with values of features. In such cases it is better to create generalized constraint or the united constraint, where we will use as parameters such features, the values of which are different in different constraints. Such parameters are called macrovariables and unified constraints are called macros, where instead of parameters macrovariables are used. Macros has own name and we can call this macros to concrete place using this name and macroarguments, where every macrovariable has corresponding macroargument. Macroarguments are the corresponding values of feature of macrovariable. On the place of a macrocall the text of macros is copied by help of special program - macrogenerator, where macrovariables are replaced by corresponding macroarguments. This procedure is well known in programming and is called macrogeneration. By using macrogeneration linguist can compose macros and use it for similar constraints by changing macroarguments. Macrogeneration is foreseen in morphological analyzer. For example, for the similar paradigms of verbs we are composing one macros and by using different macrocalls we receive the texts of constraints for all similar paradigms. It is possible to use inside of macros a macrocall to other macros. Macroarguments can be any texts, in particular a list. In case of list, the elements of the list are connected by logical “or”. It means that list is short form of writing of subexpressions connected by “or””. So macros is short form of writing of similar constraints and use of macros will decrease significantly the time used for writing of constraints. For example, the determination of verb’s person and number is necessary for every rank of a paradigm and using macros will decrease the time for writing of constraints. The constraints of ranks are different by signs of ranks, signs of theme and by using of prepositions. So in the macros of ranks, macrovariables will be above listed signs. We can create macros for different types and classes analogically to the described ones. Using a macros we decrease the time, which we need for writing of constraints.

Constraints for recognizing third series of verbforms are separated from first and second series of verbforms. This made it possible to widen the area of macros use, because the constraints, which we use for third series, are significantly different from the constraints used for first and second series of verbform. By help of morphological analyzer we can derive words, also this is not our aim. Finding out the wordform and recognizing its morphological categories gives us the opportunity to unify it with syntactical parsing of a sentence. The work, which was conducted, ensured us, that the morphological analyzer is very smart instrument for computer morphological analysis for such natural languages, which have well-developed flexible system for creation of wordforms.

Used instrument gives to the linguist the possibility to reduce the time for creation of program and make the experiments for checking his morphological system. The correction of the program is very simplified as well [1].

REFERENCE

1. Antidze J., Mishelashvili D. Instrumental tool for computer morphological analysis of some natural languages, in the volume.
2. Melikishvili D. Conjugation system of Georgian Verbs, Tbilisi, 2001.