Three Aspects of Language Modelling

G. Chikoidze, E. Dokvadze, N. Javashvili, L. Lortkipanidze
Dpt. of Language and speech systems, Institute of Control Systems, Georgia
gogi@gw.acnet.ge ninojav@gw.acnet.ge, liana@gw.acnet.ge

From some point of view Language Modeling (LM) can be considered as a some axis of the linguistics. The thing is that just in the frames of it the different basic components of language should be unified and as a result brought into accord and conformity. In some sense just in this their interaction in the course of LM activity shows itself their true structure, import and their natural place in a system as a whole. The typical instances of a such interaction and reciprocal influence are: dictionary and grammar, different levels of language (from phonetic-phonological-up to semantic-pragmatical), analytical and synthetical direction of speech/text processing, etc.

Just one more, though somewhat more global and general, dimension of such relations is here under consideration: that is, we shall here touch the question of a triple relation between aspects of language knowledge, its use and its acquisition. This direction of investigations newly began and as far only some sketches of the language knowledge/use relation are ready for demonstration, though even they are not sufficiently tested and don't guarantee complete correctness of their functioning.

The most general scheme of the knowledge/use relation (on the morphologic level) can be represented by the Fig.1.

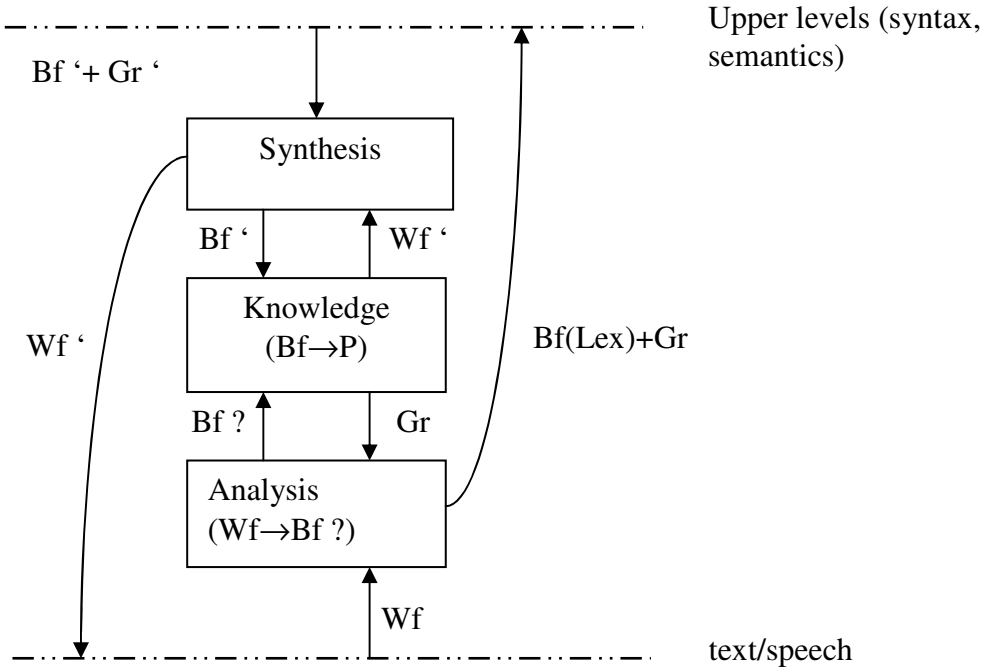

Fig.1. The symbols of the scheme mean: Bf, Bf '-Basic form, Wf, Wf '- arbitrary Word form, the question sign underlines that denotatum of the symbol is only hypothetic, Gr – Grammatic features (of Wf), Lex – Lexical information, corresponding to Bf.

According to the wide accepted opinion that language knowledge can be represented by a generative grammar the productive components (synthesis and analysis) are based on the morphologic generator Bf→P (described in [1]), which transforms each basic (dictionary) form into all members of the paradigm (P) corresponding to the input Bf. Just this component of the scheme is the most accomplished: its object is Russian morphology and it is based on [2].

As to synthesis it obviously does not create any serious problems in the context of the Bf→P system if we suppose that its input is just Bf' which is at the same time the generator's input also; and the choice of the required form is immediately defined by the grammatical part of the

input (Gr'). Only desirable result in this case is the acceleration of the dictionary unit search which contains the information necessary for the generative process. In principle the Bf→P system as a such includes some means for that end however it's desirable somewhat increase their efficiency.

Essential heavier version of this problem characterizes the analytic component of system. In this case a direct mode of comparison in course of dictionary search is changed by the attempts to find some alikeness between the input Wf and Bf's which label the dictionary units. Such more specific process exacts some more effective means for reduction of the search area in dictionary and in the generated paradigm (P) both. The corresponding operational expenses can be very roughly evaluated by the expression:

$$\tfrac{1}{2} l \times (G + p \times C),$$

where l is the number of dictionary units, G-average operational expenses of generation of a single, paradigm(P), p-average number of paradigm members and C- expenses needed for a single comparison of input Wf with one of the P members. This expression implies the most simple and for that the most expensive way of analysis: Bf→P generates in turn the paradigms of each dictionary unit till one of their members will not be identified with the input Wf. Naturally, our aim should be a reduction of values of l and g parameters.

The scheme in Fig.2 proposes the partially solution of this task.

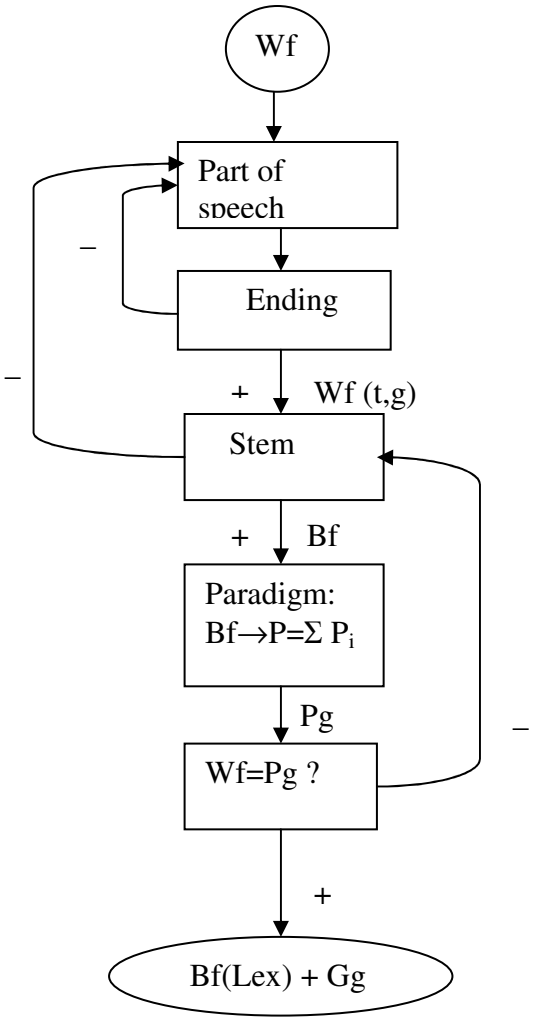

Fig.2. General scheme of morphologic analysis, including some means for reduction of l, p parameters of expression (1) (Comments in the text).

In the beginning the process ascribes to Wf some part of speech (PS) characteristics (block 0) and then tries on the basis of this supposition to analyze the ending of Wf (block 1); if this step of processing fails it changes its previous PS hypothesis and addresses to the block 1 again; when

the latter is satisfied it produces, firstly, the set of types (t) to which should belong Wf (if it really is characterized by the lastly supposed PS) and, secondly, the bunches of grammatical features g which may it express (if the PS hypothesis is correct again). After this the process addresses to dictionary and tries to find there the units (of t-type of given PS) which satisfy the conditions of likeness between Wf – stem and Bf which represents this unit as its label (block 2); after this block 3 generates the paradigm (P) of Bf and block 4 tries to identify Wf with members of P which satisfy g-condition. The cycle of the scheme continues till the requirements of block 4 are fulfilled.

Obviously activity of block 1 will essentially reduce the values of l and p both. Moreover reduction of the former value (l) will restrict the area of dictionary unit search for the synthesis too. Lastly, functions of this block may be quite useful for the aspect of language knowledge acquirement. For example, if some unit is lacking in the dictionary, we can give its paradigm to the system, then address the block 1, which will define to which type (t) should belong this paradigm and correspondingly its Bf. After this some additional component of the teaching system may define some other features also: for example, such obvious characteristic as LF (Lacking Forms) which marks the paradigms without some members (e.g. "писать" – ' to write ' has not verbal adverb and passive participle of present tense). Finally, system will try all branches of the t-type paradigm generation, and stop at one, which generates a paradigm fully identical with the input. As to the case, in which such a branch doesn't exist, it requires, of course, some more complicated means about which we have as far some conjectures only.

Thus we may conclude from above considerations that analysis and especially its "Ending" component (block 1) serve as a link connecting and relating analysis and synthesis, dictionary and grammar, the language knowledge, its use and its acquisition.

In conclusion we give a little fragment of block 1 (Fig. 3) without comments.

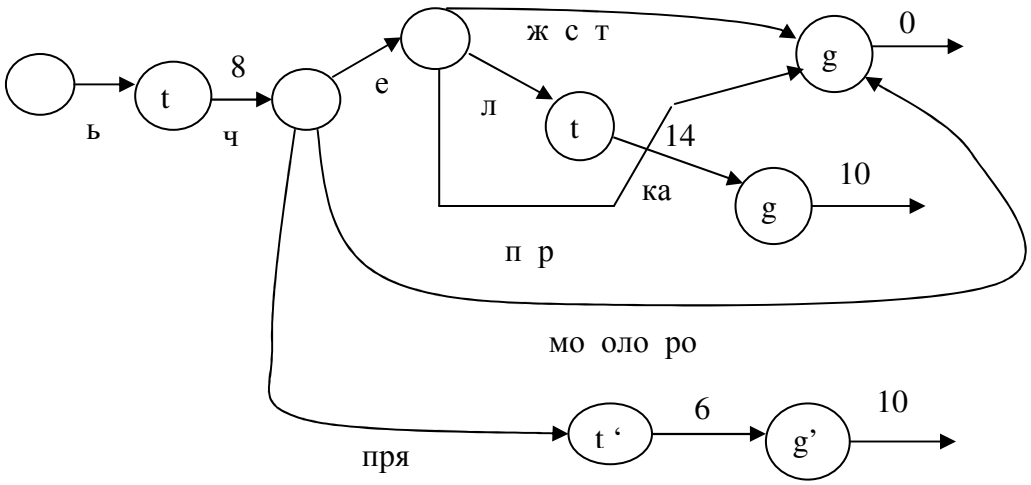

Fig.3. Fragment of block 1 (Fig.2) which analyses infinitive (g=0) endings of the type 8 (t=8); in some cases these forms are homonymous with imperative singular (g=10) of types 4 or 6 (t=4 or g=6); in a single case of "калечь"verb ('mutilate') we have the imperative form only.
In [3] is given a description of the system analyzing all endings of reflexive verbs.

Literature:

1. Г. Чикоидзе. Сетевое представление морфологических процессоров. ИСУ АН Грузии,2004
2. А. Зализняк, Грамматический словарь русского языка. Москва, изд.»Русский язык», 1977
3. Э. Доквадзе, Л.Лорткипанидзе, Г. Чикоидзе, Бессловарный сетевой морфологический тегер. Труды Международного семинара Диалог 2001 по компьютерной лингвистике и ее приложениям. Москва, 2001