THE LONG JOURNEY FROM THE CORE TO THE REAL SIZE OF LARGE LDB

Elena PASKALEVA, Kiril SIMOV, Mariana DAMOVA, Milena SLAVCHEVA

Linguistic Modeling Laboratory, Center for Informatics
and Computer Technology, Bulgarian Academy of Sciences
Acad. G. Bonchev St. 25a, 1113 Sofia, Bulgaria
fax:+359-2-707273, e-mail:HELLEN@BGEARN.bitnet

1. Introduction: The Meanings Of "Large"

Large Lexical Data Bases are one of the earliest applications of NLP. The initial stage of their rise, with the admiration for the automation of lexicographic work itself, came to an end long ago. In the following stages Lexical Data Bases (LDB) began to extend considerably the range of their application and the scope of CL problems put forward by them [see Calzolari 1991, Calzolari and Zampolli 1988 and Boguraev et al.1988]. It is worth discussing a new version of LDB (for a concrete new language) only in the present-day context of these problems. This does not, however, relieve the creators of LDB for a new language of the solution of the trivial problems standing at the lower foot of the ladder used to "storm" the lexical wealth of language. After overcoming these obstacles, there is prototype version available or a core of LDB, which cannot be called large especially when its volume is concerned. Speaking of volume, quite naturally, the following question arises: in what direction should the linguistic knowledge be extended, so that the system could be defined as *large*? Shall we say "large" in the literal sense, having in mind the number of entries in DB, or does "large" mean "deep", i.e. the richer linguistic information in the lexical entry means a larger scope of linguistic phenomena included in DB?

It is obvious that the researchers who have climbed higher up the ladder mentioned above (in the works quoted above) are interested in the second sense of the attribute "large", as the first type of expansion of the basis has long been a fact for them.

This paper is an attempt to share the experience of researchers who have climbed up the first few steps of the ladder, and who are clearly conscious of the height they still have to reach (on account of the fact that they began to build an LDB in the early 90s). This consciousness makes them speed up the process of climbing the first few steps (i.e. , to make the base large in physical volume), in order to continue at a higher speed the expansion of the base with regard to the scope of linguistic knowledge (i.e. to build a "deeper" large DB).

The intellectualization, hence the speeding up of the first type of expansion of the base through the creation of special programming tools for representing, correcting and enriching the linguistic knowledge in a separate entry, is a task we have already confronted with at the Linguistic Modeling Laboratory when working on an LDB for Russian and Bulgarian.

This paper is about the programming tools accomplishing the interface with the linguist who has at his disposal a nuclear prototype DB and whose task is to turn it into a really large DB.

2. Designing The Core Of The System: The Volume Of Linguistic Knowledge

In the Linguistic Modelling Laboratory the idea of creating a large DB was in fact a natural continuation of the research work on an exhaustive formalized description of Bulgarian inflexional morphology in the form of procedures for morphological analysis and synthesis including the entire scope of phenomena in Bulgarian word inflexion. This goal is achieved through the system MORPHO-ASSISTANT [see MORPHO-ASSISTANT 1990]. The exhaustiveness of the morphological description is guaranteed by: a) a full list of inflexional types of Bulgarian inflected words; b) a full list of all types of graphemic changes -

the so-called alternation types; c) inclusion of all possible doublets; d) inclusion of morphosyntactic phenomena (i.e. going beyond the framework of the separate wordform) of first order. It concerns the description of complex verb tenses, as well as the exhaustive classification of verbs with regard to their voice behaviour. This classification makes possible the analysis and synthesis of complex structures consisting of a verb and declined reflexive and pronominal clitics (in this configuration some vocables of verbs are given in Bulgarian dictionaries). Semantic characteristics are included in the description only if they are relevant to the word inflexion.

The first project variant of MORPHO-ASSISTANT is served by a minimal dictionary of lexemes and morphemes reflecting all the phenomena described above. The next most natural question concerns the representativity of the lexical base of the system. Thus the ambition of the creators of a computer system with an exhaustive morphological knowledge quite naturally turned into an ambition for creating a large lexical data base for Bulgarian and Russian (the former containing 60 000, and the latter - 100 000 units). The simplicity of the transition: lexical base of a morphological component ---> lexical data base was ensured by the programming language chosen for the two products, namely PROLOG.

The information included in the so designed Bulgarian LDB brings it near to the so called grammatical dictionaries (such as the known to all Slavonic scholars Grammatical dictionary of A.A.Zaliznyak [see Zaliznyak 1977].

In this way, the core of LDB includes the following portions of linguistic knowledge: a) a list of all grammatical formatives participating in the inflexion; b) a list of the grammatical categories characterizing the inflexion; c) a list of the full paradigms of the different parts of speech; d) a list of the inflexional types of the inflected words - each inflexional type is a set of correspondences between a member of the paradigm and the formative representing it; e) a list of the types of alternation describing the letter changes in the stem as a result of alternation and the conditions for these changes (determined by the grammatical categories of the member of the paradigm for which they are valid); f) procedures for morphological analysis and synthesis; g) a dictionary of lexemes for which the principle of minimal representativeness is applied - the set of lexemes should make possible the representation of each linguistic fact from a) to e) applying the procedures in f); h) an exhaustive description of the completeness of the paradigm for all lexemic units (marking the possible defects in the inflexion).

3. How To Make The Data Base Really "Large" In A Short Time?

The problem of expansion of the prototype core system to the volume of a real LDB is solved depending on the sources of this expansion. A standard source for collecting lexical elements in the required volume are obviously the existing dictionaries. But as the action takes place in the early 90s, it is natural to rely at least on machine-readable dictionaries (MRD) in their great variety of volume and type of data. When such a civilized solution is found, the problem of completing the base to its real volume is reduced to the creation of programming tools for recoding the information in MRD and its eventual completion in interface mode [see Boguraev and Briscoe 1987]. A similar approach was used for the construction of the Russian LDB which we are developing together with the Department for Machine Fund of Russian in the Institute of Russian language (Moscow). As the Dictionary of Zaliznyak - the base of the Russian LDB - is machine readable, the work on its representation in MORPHO-ASSISTANT format is reduced to the construction of a recoding program accomplishing the translation from the specific notation of the grammatical information in its entries into the corresponding classes of inflexion and alternation. The information for possible defects in the paradigm is the only one, introduced manually. As for the Bulgarian LDB the problem is a little different. Bulgarian lexicography has not its own grammatical dictionary even in a man-readable form. Here the problem is, how from the great number of one-language or spelling dictionaries, normative grammars and handbooks in morphology we can determine the units of LDB and give the necessary information for each one of them according to the principles established in the construction of the core. The first problem was solved by choosing the vocabulary of the latest Bulgarian spelling dictionary (60,000 words). Represented as a text file, this lexicon served us as a MRD consisting only of the vocables of the dictionary entries (the information about the word inflexion in the spelling dictionary usually point out the exceptions and difficulties). Thus the first task in the process of expanding the LDB core came out: the determination of the dictionary information for a given entry.

The rich morphological system of Bulgarian (145 classes of word inflexion, 72 classes of alternation, great variety of not regularly determined cases of incompleteness in the rich paradigm of Bulgarian words) makes the specification of this information even if carried out by a highly qualified linguist a difficult and not safe from mistakes task. We did our best to make improvements by creating a special software, i.e. "linguist friendly" programs speeding up the process of filling the entries. This "linguist friendly" software consists of two basic programming packages: programs for filling LDB entries and programs for revising LDB entries.

3.1. Filling A Lexical Entry In A Friendly Way

The above mentioned information in LDB core determines also the content of the LDB entry which, most generally speaking, consists of the following portions of linguistic information: part of speech, characteristics of the lexeme (depending on the part of speech: gender, animateness and person of nouns, aspect and transitivity of verbs, etc.), inflexional type, type of alternation, defects in the paradigm of the lexeme. The considered software models partially the filling of each LDB entry, performed by a human, without relieving entirely the user of the duty to use his linguistic competence in the bottle-neck points.

The system has the following functions: calculating the calculable, correcting the erroneous and simplifying the difficult. In the first two functions we use the knowledge about the links between various linguistic categories and their values. The specific part of linguistic knowledge, accessible to man only and necessary for fulfilling the third function of the programming environment, is reduced to elementary routine work on building the concrete paradigm. The three functions mentioned above are performed by a system of menus reflecting the relations in the linguistic knowledge.

The most essential facility for the linguist (in the third function of the programming environment) is the determining of the inflexional and derivational classes. In accordance with the chosen grammatical characteristics of a given lexeme, a so called "diagnostic paradigm" is automatically formed . The number of its members is greatly reduced (as it is possible to calculate some functional dependencies). When processing this diagnostic part, the user fixes the correct wordforms of the lexeme, i.e. he determines the inflexion and eventually edits the stem in case of alternation. After creating the diagnostic part of the paradigm, the inflexional type and the type of alternation come out automatically. If the input values do not correspond to the information from the core, the system answer is either wrong combination of formatives (so it has to be corrected), or necessity of introducing a new classificational type. As the richest paradigm in Bulgarian inflexion - a verbal one - consists of 52 forms, we are satisfied with the achieved maximum speed of filling the entries - 80 entries per hour (on an XT computer).

The error control (in the third function of the software) is exercised only over dependencies between the combinations of the grammatical categories and the formatives expressing them (separately or as a whole), but cannot check the authenticity of the specific lexical information which is filled in (for example stem features, paradigm defects, etc.). That is why a considerable part of the responsibility for the correct filling of the lexical entries is shifted on another software product ensuring their revision and updating.

3.2. Friendly Tool-Kit For Updating The Lexical Entries

The LDB organization of the Bulgarian grammatical computer dictionary (in ARITY PROLOG) saves us the boring subsequent updating of lexical entries. The input lexical entries are grouped in a natural way depending on the values of the grammatical characteristics. The grouping specifies the entities to be processed simultaneously. The minimal group for viewing/updating is the group of lexemes with equal values in all fields of the entry. This grouping, however, can be optimized from a linguistic point of view as well, according to the actual hierarchy of the linguistic knowledge in question.

The linguistic knowledge hierarchy, correlated with the objects grouped in such a way, can be seen in the screen of the system, given below in the Fig.1.

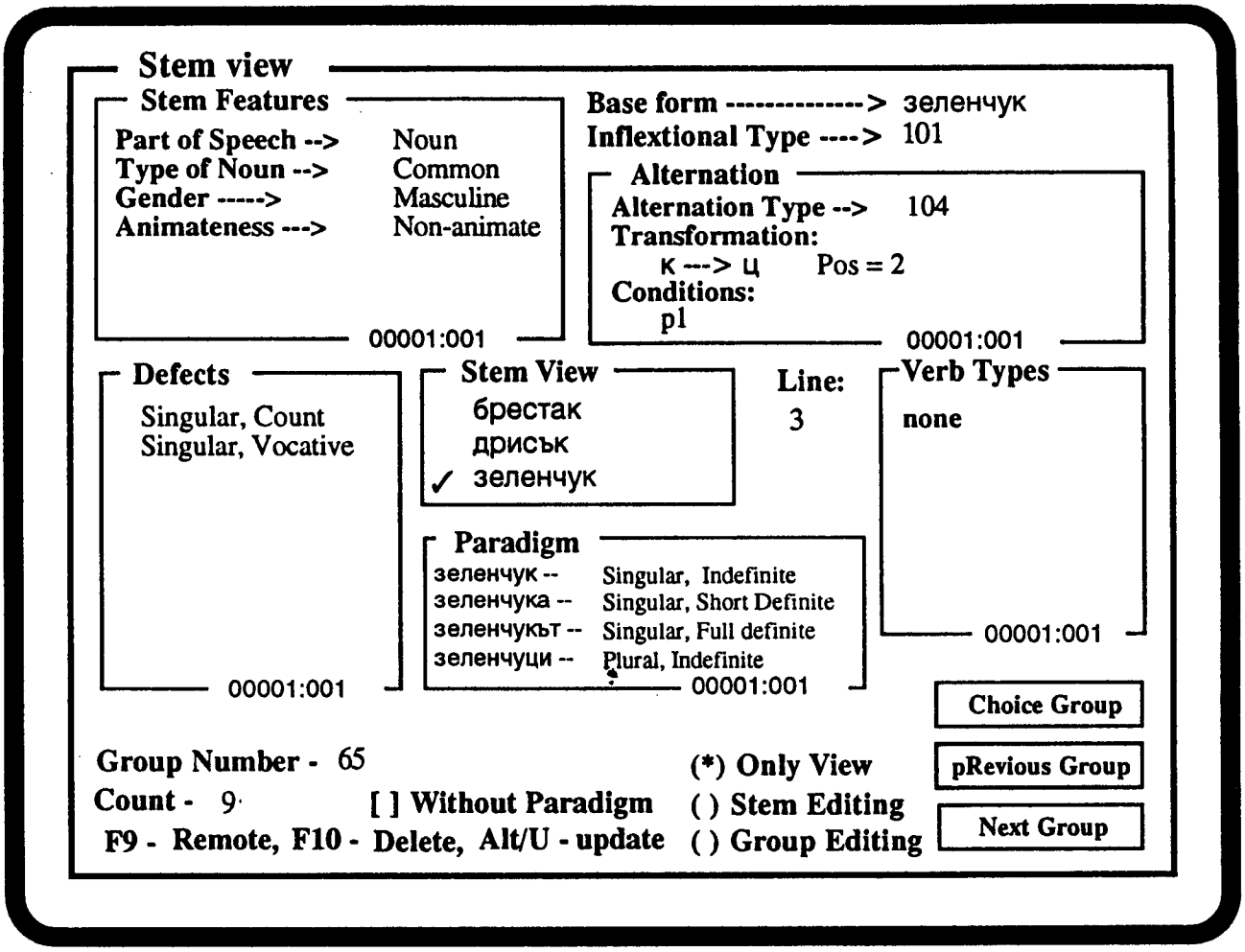

Figure 1: View/Update Screen

The screen information has not only illustrating but editing functions as well. The characteristics are represented on the screen by windows and string fields with a dynamic reflection of the links between the attributes and the values of the features of each choice. Besides the static characteristics from LDB, the screen reflects the results from the generating procedure in the special window "Paradigm", containing the members of the full paradigm of the chosen lexeme. The editing in "Paradigm" window may invoke changes in the characteristics of the entry - the procedural testing is the best control (for example the deletion of a member of the paradigm causes a change in the window "Defects"; the correction of an inflexion in "Paradigm" leads to a change in the information about the inflexional type, etc.).

4. How To Use The Large Bulgarian LDB?

The flexible programming tools described in I and II aim at speeding up the process of creating the Bulgarian LDB through making easier the task of the constructors. These tools, however, are only an intermediate device for achieving the final goal which deserves to be discussed in detail in this last paragraph. What are the benefits of the final users? The screens below illustrate its potential capacities. We are not going to discuss the standard capacities of LDB which can be seen in Fig. 2 and 3.

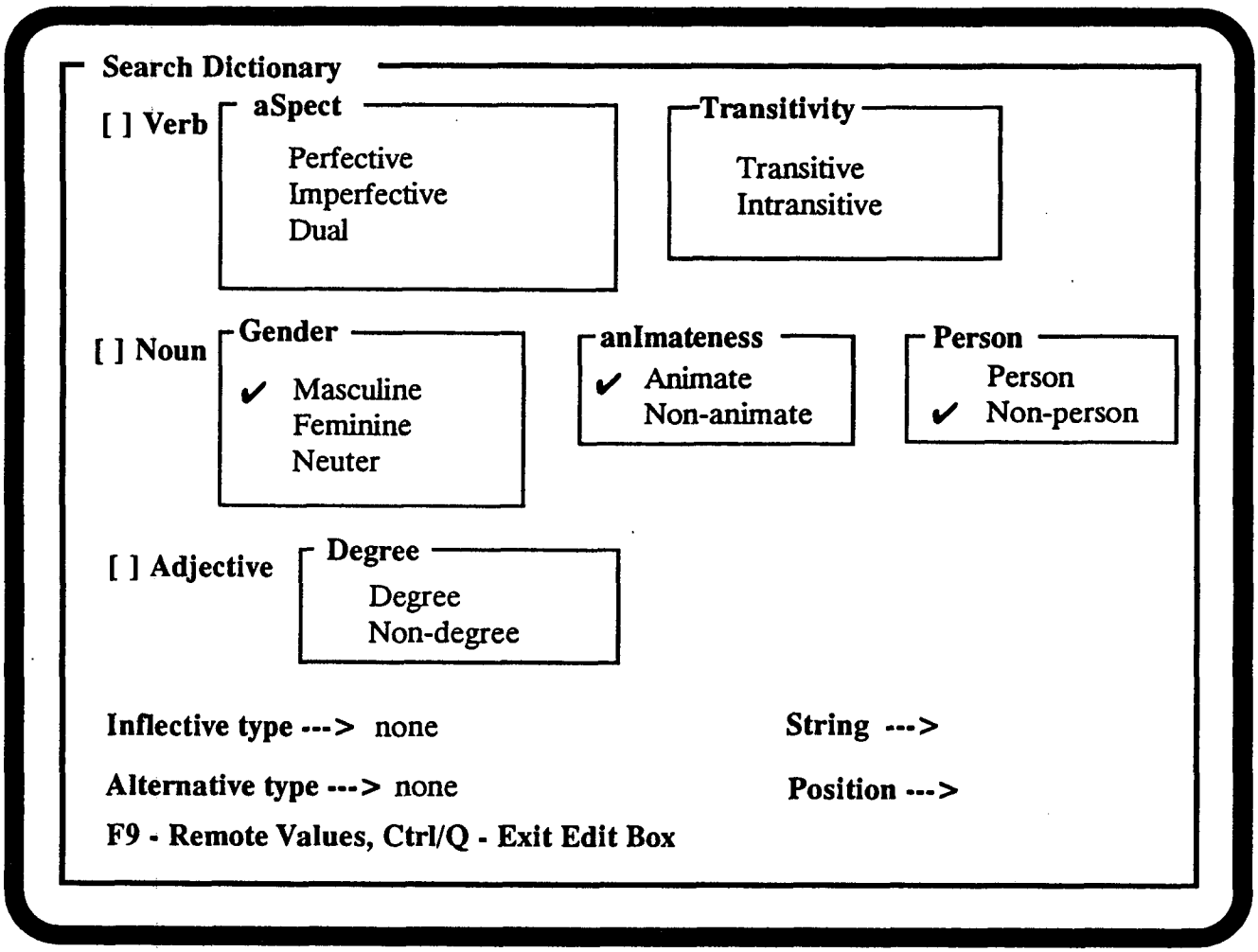

Figure 2: Query screen

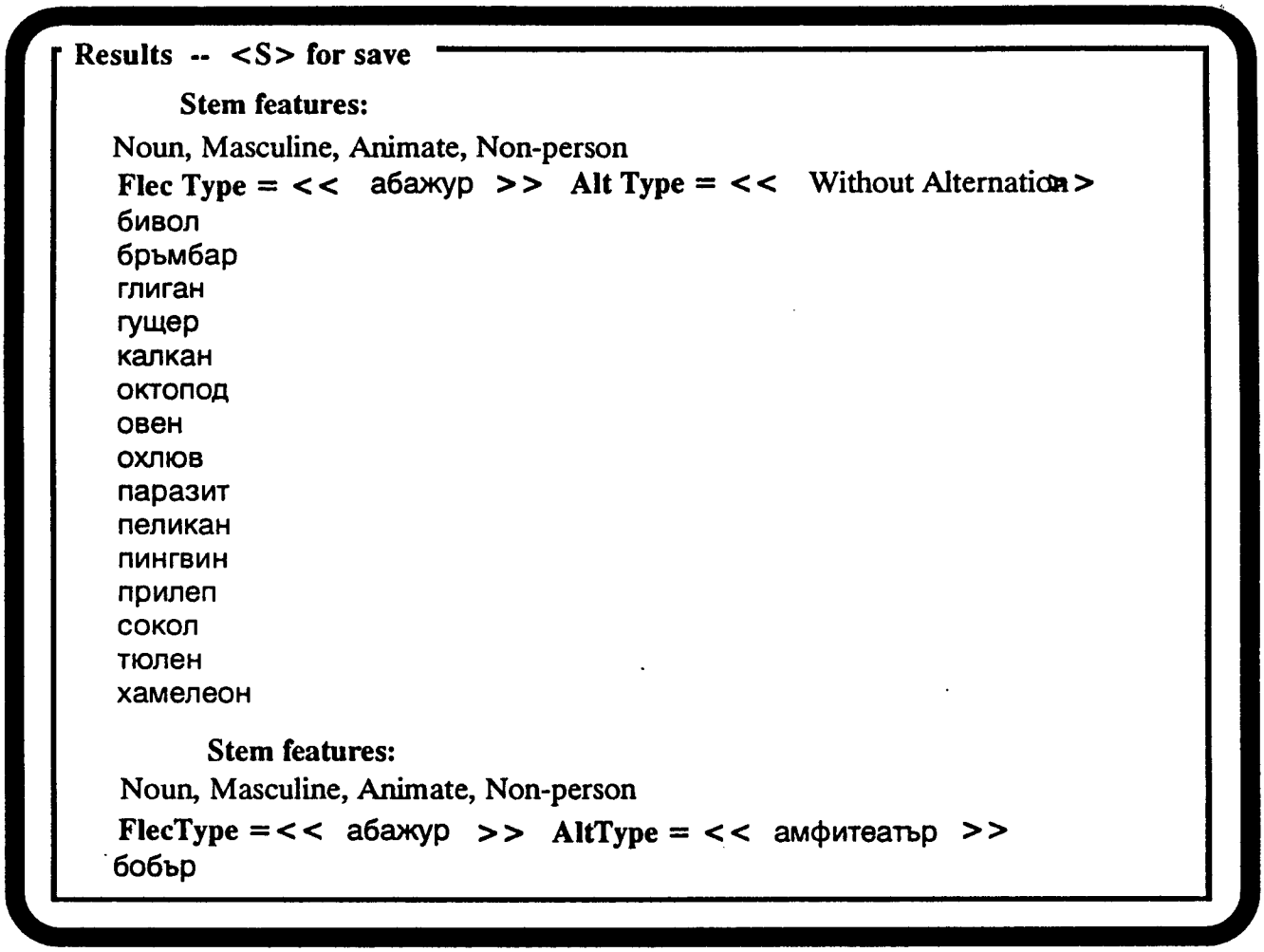

Figure 3: Search results

We shall consider some expansions of the standard LDB operations:

a) As we can see in figure 3., the output lists of lexemes, extracted by given features, are in addition automatically grouped according to the characteristics which do not participate in the searching. The lexemes of such a group are alphabetically ordered.

b) There is a special searching-by-string-and-position procedure. What is more, it processes a level deeper than the graphemic one - namely, the morphemic level.

The result represents groups of lexemes with the same letter combinations in the given position. Using a special option, the searching procedure ensures preliminary elimination of the prefix elements and search of the given string at the beginning of the rest of the lexeme (see fig. 4 and 5). In this way the output includes families of words of first approximation.

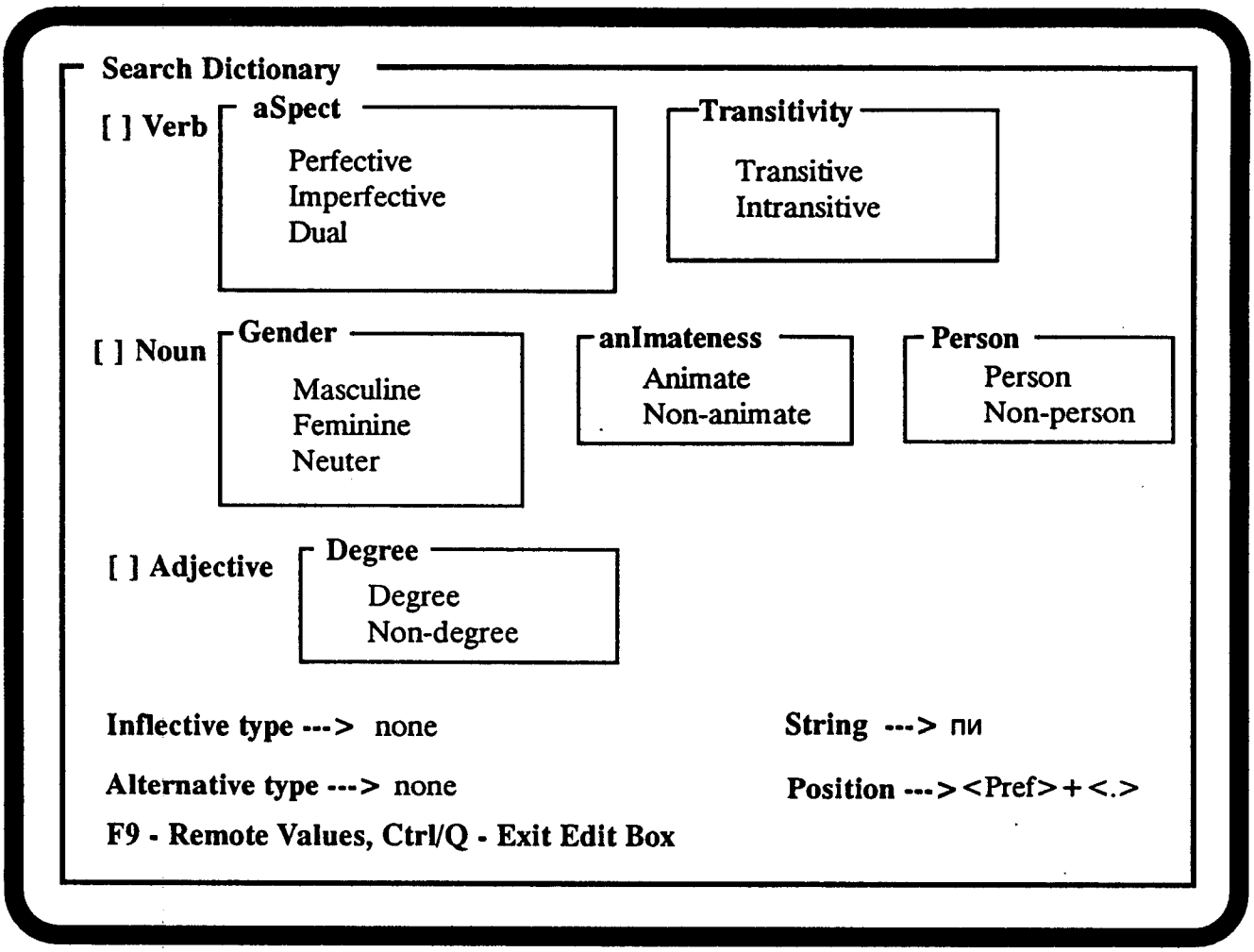

Figure 4: Query screen with prefix-eliminating search

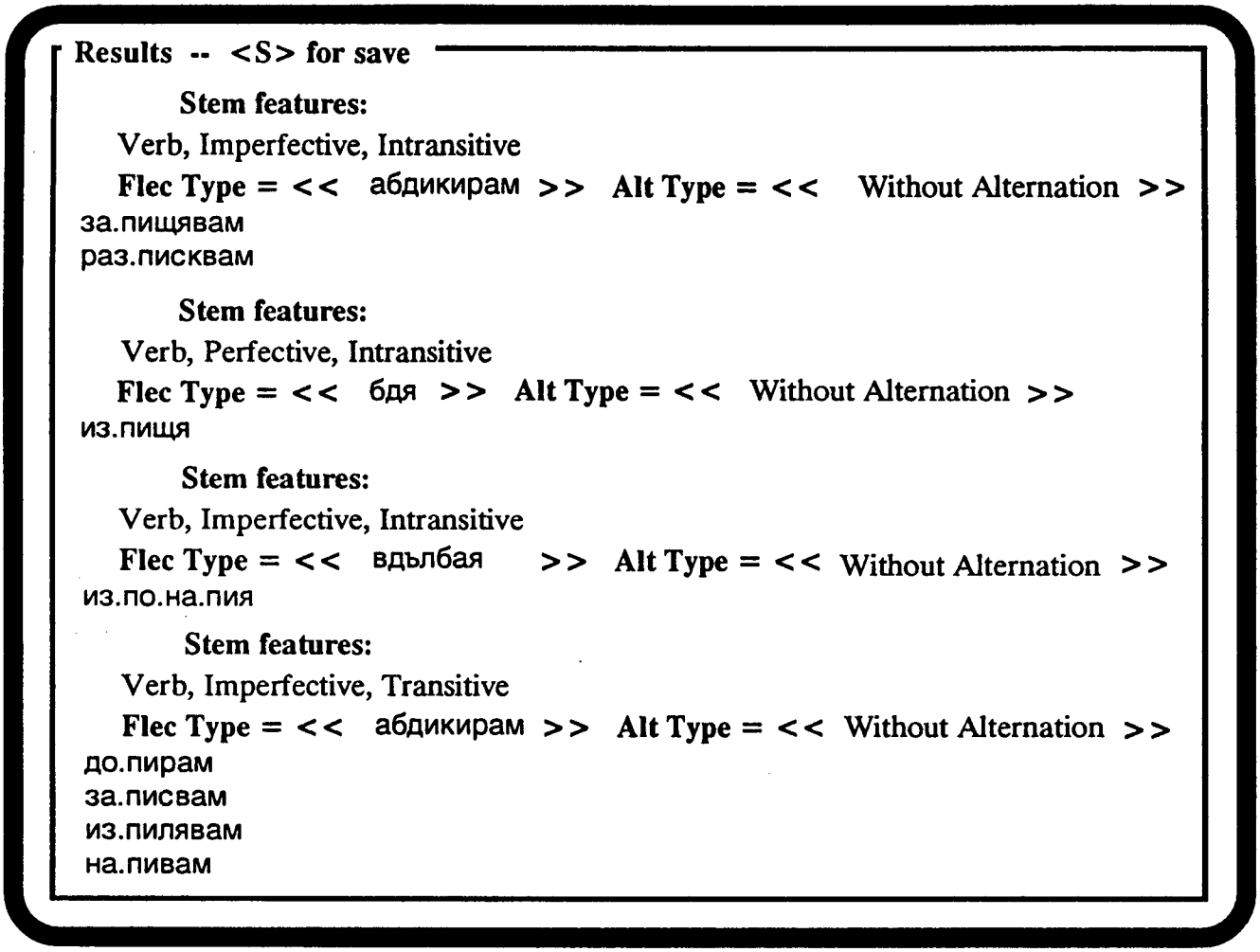

Figure 5: Results from prefix-eliminating search - word family of first approximation

5. Future Development: The Journey To The Lkb

Being aware of the long way to the creation of a real LKB (Lexical Knowledge Base), we would like to write about the first steps we have made in this direction which coincides with the main goal of the CL group in the Linguistic Modeling Laboratory. It is the creation of the base of linguistic knowledge for Bulgarian.

1. The inclusion of the procedures of analysis and synthesis (realized in the system MORPHO-ASSISTANT) in LDB makes possible not only the expansion of the searching procedures but the accomplishment of the following transitions as well:

a) from a text corpus to LDB (using the analysis of MORPHO-ASSISTANT);

b) from any LDB entry to arbitrary parts of its paradigm (using the synthesis of MORPHO-ASSISTANT);

2. The linguistic results from the string searching can be considerably deepened by the creation of software tools for editing the family words of first approximation in dialog mode. In such a way, the real

family of derivationally related words can be constructed. Their accumulation and connection with the main LDB will make possible the automated creation of a Bulgarian morphemic computer dictionary (which does not exist even in a traditional form) and a knowledge base for the derivational morphology.

3. The acquired experience in creating a flexible software environment, facilitating the filling of the lexical entry, makes it possible to create, in the same style, a procedure for completing the LDB with information about the accentual characteristics of words. A description of Bulgarian word inflexion neglecting the accentual information cannot be regarded as a complete one, because the movable stress in Bulgarian is an essential part of the inflexional mechanism.

4. Following the tradition in creating "linguist friendly" software, we are planning the filling of the syntactic part of lexical entries (and some other information). Unfortunately, we should say that Bulgarian lexicography is not so friendly to computational linguists and has not supplied them (and not only them) with suitable syntactic dictionaries including information about the subcategorization of lexical units. In spite of the delay in creating LDB (due to historical and technological reasons) and the lack of traditional lexicographic sources on which to rely, the CL group hopes to rank in the forefront of CL investigations using advanced computer technologies.

References:

Boguraev et al. 1988: Boguraev, B., E. Briscoe, N. Calzolari, A. Cater, W. Meijs, A. Zampolli. *Acquisition of Lexical Knowledge for Natural Language Processing Systems.* Proposal for ESPRIT Basic Research Actions, August 1988.

Calzolari and Zampolli 1988: Calzolari, N., A. Zampolli. *From Monolingual to Bilingual Automated Lexicons: Is There a Continuum?* In: Lexicographica, 4/1988.

Calzolari 1991: Calzolari, N. *Structure and access in an automated lexicon and related issues.* In: Automating the Lexicon, Walker D., A.Zampolli, N.Calzolari (editors). Cambridge University Press, 1991?

MORPHO-ASSITANT 1990: Simov, K., G. Angelova and E. Paskaleva. *MORPHO-ASSISTANT:The proper treatment of morphological knowledge.* Proc. COLING'90, vol.3, 453 - 457.

Zaliznyak 1977, *Grammaticheskii slovar russkogo yazika: Slovoizmenenie.* Moskva 1977.