INTEX 4.0 for Bulgarian

(Error checking as an INTEX application)

Svetla KOEVA, Stoyan MIHOV

1. Introduction

INTEX under NextStep has been available since 1992. A module for Bulgarian under NextStep was implemented between 1995-1997 at LML (Linguistic Modelling Laboratory, http://www.lml.bas.bg).

For several years INTEX has ported to Windows95-NT. Here we present the Bulgarian lexical and grammatical knowledge integrated in INTEX for Windows. The investigation is being developed as a part of the Joint research project "Computer Representation of grammatical Knowledge of Bulgarian" between LADL (Laboratory for Information Retrieval Systems and Linguistics at the University of Paris-7, http://www.ladl.jussieu.fr) and CMD (Department for Computer Modeling of Bulgarian at the Institute for Bulgarian language, Bulgarian Academy of Sciences, http://www.ibl.bas.bg).

2. Preprocessing the text

The preprocessing in INTEX includes the segmentation of the text in sentences, the identification of unambiguous compound words and the distinguishing of special tokens (contractions, elisions etc.). We supply INTEX with Bulgarian lexical and grammatical knowledge as fol-

✉ Svetla KOEVA, Department for Computer Modeling of Bulgarian
Institute for Bulgarian language, Bulgarian Academy of Sciences
e-mail: svetla@ibl.bas.bg
Stoyan MIHOV, Sector for Linguistic Modeling, Center for Parallel Processing
Bulgarian Academy of Sciences
e-mail: stoyan@lml.bas.bg

lows:

– FST for sentence recognition in Bulgarian;
– Dictionary for Bulgarian unambiguous compounds;
– FSTs for identifying some specific Bulgarian tokens.

2.1. Identifying sentences

The FST named Sentence.fst is applied to the text in MERGE mode. The output of the FST — {S} is inserted in the text at the sentence borders. Here bellow is presented Sentence.fst for Bulgarian (Fig. 1), which describes some specific features of Bulgarian punctuation (for example lexical abbreviations, Bulgarian compound words that end with a capital letter, graphical abbreviations, etc.) together with the trivial cases.

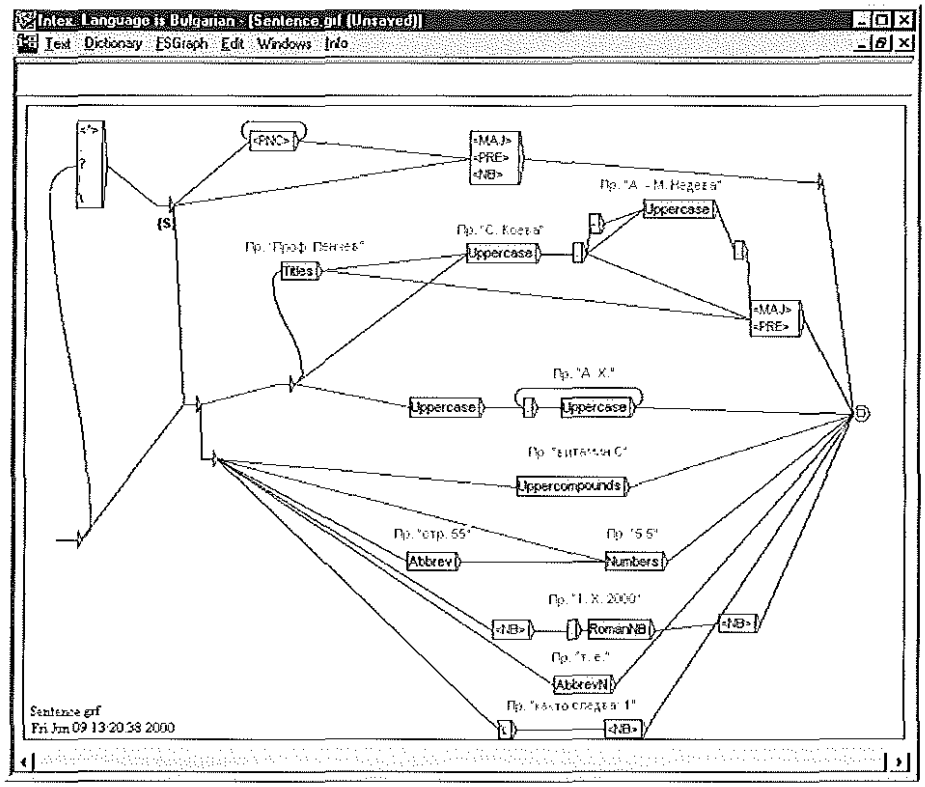

Figure 1

2.2. Identifying unambiguous compounds

The preprocessing includes the identifying and tagging of the unambiguous compound words in the text. The dictionary Norm.dic in which Bulgarian unambiguous compound words are listed is used in this operation.

Most of the compound words are unambiguous and it is good to recognize them during the preprocessing in order to avoid wrong readings, i.e.:

Cherni vryh (Black peak)

Cherni {*cherni*, cher.A:p} *vryh* (*vryh*, vryh.N+M:s)
{*cherni*, cheren.A:p}
{*chrni*, chernja.V+IMP+T:E2s:E3s:I2s}

The correct reading is: *Cherni vryh*
(*Cherni vryh*, Cherni vryh.N+M:s)

In general we classified Bulgarian unambiguous compound words as personal names (Baba Jaga), geographical names (*Tihi okean*—The Pacific), names of institutions (University of Sofia), terms (stone age), complex prepositions (*blagodarenie na*—thanks to) and complex conjunctions (*za da* — in order to).

2.3. Identifying special tokens

The last stage of the preprocessing consists of identifying and tagging of the special tokens, such as abbreviated and contracted words. Applying the FST Norm.fst in REPLACE mode performs the identification of special tokens. The FST Norm.fst consists of several embedded FSTs. For example the FST Comparison.fst interprets Bulgarian comparative and superlative particles that are separated by the corresponding adjectives and adverbs by a hyphen. The FST Elision.fst describes the abbreviations in different positions in the sentence, the abbreviations after numbers, the units of measure, and the elisions with apostrophe and hyphen.

3. Applying dictionaries and lexical FSTs

3.1. Bulgarian INTEX dictionaries

Dictionaries applied by INTEX must be in a specific format, DELAF dictionaries for simple words and DELAFC dictionaries for compound words. Each lexical entry in the DELAS dictionary is associated with one FST that represents all corresponding inflected forms. DELAF dictionaries are generated from DELAS dictionaries by derivation of all paths of inflectional FSTs. For more detailed explanation of DELAS, DELAF and DELACF dictionaries in INTEX see GROSS AND PERRIN (1989), SILBERZTEIN (1986, 1987 and 1999, p. 16-24).

After the preprocessing procedure the simple words and the compound words in the text can be identified by the respective (DELAF and DELAFC) dictionaries and FSTs. The Bulgarian DELAF dictionary is generated from Bulgarian dictionary Gramatic2000 (KOEVA, 1999). The dictionary consists of about 80,000 lemmas and over 1,000,000 corresponding forms, basically all inflected simple words.

3.2 Lexical FSTs

The lexical FSTs are used to describe infinite sets of lexical entries (e.g. analytical numerals). The following FST Dnum.fst (Fig. 2) identifies and tags Bulgarian numerals from 2 to 9,999.

4. Bulgarian grammar rules

4.1. Disambiguation rules

Most of the interesting problems in computational linguistics, as well as the important applications in Natural language processing require a tagger—an automatic system that correctly associates the words with the grammatical categories and their values. The morphological analyzer produces all legitimate tags for the words that appear in the text. For instance the Bulgarian word kosi will receive in Intex the following tags:

Noun, Feminine, Plural, Indefinite "hair"
Verb, Present, 3 Person, Singular "he mows"
Verb, Perfect, 2 Person, Singular "you mew"
Verb, Perfect, 3 Person, Singular "he mew"
Verb, Imperative, 2 Person, Singular "mow!(now)"

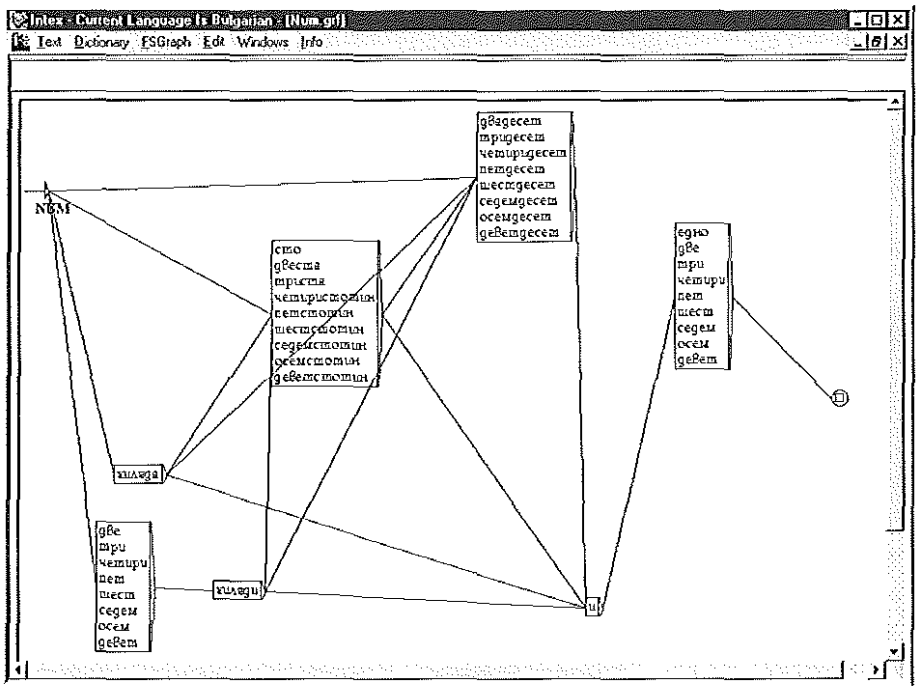

Figure 2

It is well known that there are two different approaches to the disambiguation—statistical (data-driven) and linguistic (constraint-based). The INTEX offers good language independent tool for the linguistically oriented disambiguation. We can formulate a few types of linguistic rules as FSTs. The contextual restrictions for the most frequently used ambiguous words can be defined to resolve ambiguity. For example a rule can say that:

- Forms like *hubavo* are adverbs, if there is not a noun (neuter, singular) in the sentence.
- Forms like *hubavo* are adjectives, if there is a noun (neuter, singular, indefinite) that immediately follows them.

The rules described above are certainly not sufficient to provide full disambiguation. An interesting observation (which is also true for Bulgarian) is that the most frequently used ambiguous words are usually words that are corpus independent (prepositions, pronouns, con-