БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

КИБЕРНЕТИКА И ИНФОРМАЦИОННИ ТЕХНОЛОГИИ • Том **2**, № 2
CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **2**, No 2

София . 2002 . Sofia

The Bulgarian Dictionary in Multilingual Lexical Data Bases*

Ludmila Dimitrova, Radoslav Pavlov**, Kiril Simov****

** Institute of Mathematics and Informatics, 1113 Sofia*
*** Institute of Information Technologies, 1113 Sofia*
**** Central Laboratory for Parallel Processing, 1113 Sofia*

Abstract: *This paper describes the process of preparing Bulgarian lexical databases for the CONCEDE EC project whose aim is to harmonise the methodology, tools and resources for building Lexical Data Bases (LDBs) in a general-purpose document-interchange format, for six Central European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. The selection of the words on the basis of their frequency in naturally occurring texts - Orwell's 1984 – ensures that the project produce the lexical databases useful for real applications.*

Keywords: *dictionary encoding, lexical databases, document type definition*

1. Introduction

This paper describes the process of preparing Bulgarian lexical databases for the CONCEDE: *Con*sortium for *C*entral *E*uropean *D*ictionary *E*ncoding. *CONCEDE*[1] is a EC project whose aim is to harmonise the methodology, tools and resources for building Lexical Data Bases (LDBs) in a general-purpose document-interchange format, for six Central European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. The Bulgarian partner in CONCEDE is the Institute of Mathematics and Informatics. The project strives to produce lexical resources that respect the guidelines of the Text Encoding Initiative Dictionary Working Group (TEI-DWG), and so are compatible with other TEI-conformant resources, [1].

The input for the CONCEDE dictionaries in each language is the frequency list of Orwell's *1984* corpus, prepared in the MULTEXT-East EC project (see [2] for an overall description of the project). The content of the CONCEDE LDBs entries is based on the information in published dictionaries for each of the six languages. CONCEDE dictionaries development proceeded in two phases, initially, a 500-word pilot

* *CONCEDE* is supported by the European Commission under INCO-Copernicus project No PL96-1142.

phase, and later, the larger-scale phase with up to 2,500 further entries for each language.

In the process, the project must, in addition, validate the guidelines (which were developed primarily with reference to Western European languages) for the Central European languages and propose any extensions or modifications required to accommodate them. The general procedure and the particular modifications used for Bulgarian language are outlined.

2. Headword selection procedure

One of the initial tasks in the project was the selection of a sample of 500 headwords, equal number of lemmas in each language on a common basis. The applied method, proposed by Dan Tufis (see [3]), is statistical and linguistic at the same time. A procedure for selecting the headwords, taking into account word frequency, word class, and the number of words there were in a given word-class and word-frequency band, was developed by the Romanian partner. The point briefly describes a procedure, which can automatically produce Parts of Speech (POS) lists of any length, and then considers the manual modifications that were necessary only for the sample of the first 500 entries. Furthermore, we adopted an approach, involving a generic sampling method for selection of headwords into the lexical database. The texts used were encoded as CES_ANA, [4], which specifies for each word-form its associated lemma and grammatical information. Such parallel corpora were developed in the MULTEXT-East project, [2]. The POS composition of this sample has to reflect the corresponding distribution of the different POS in the corpus.

First, the corpus is divided into sequences of text, which contain 500 different lemmas of different parts of speech. In practice, the whole corpus is reduced to a sequence of <lemma, POS> pairs. Second, a counter is incremented each time a new lemma is encountered. When the counter reaches the value 500, a new text sample starts and the counter is reset to zero. This operation is repeated until the end of the corpus is reached. A statistical formula calculates the number of each POS in the sample.

This method, ensures the following: the POS composition of the sample reflects the corresponding distribution of the different parts of speech in the corpus and to some extent the structural POS distribution of the language; and the number of POS lemmas chosen should not depend on the size of the corpus. The reason behind this advantage is the stylistically coherent text, from which the samples are initially taken.

Lemmas were chosen for the relevant ten grammatical categories identified in the MULTEXT-East project, according to the frequency of their occurrence in corpus. Three frequency ranges are considered: high, medium and low. The high frequency range was assigned the interval [0.5, 1], the medium frequency range the interval [0.25, 0.5] and all the words with frequency range below 0.25 were considered in the low frequency range.

The frequency ranges were computed (for each POS) based on a normalised occurrence ranking of each word form. The normalised ranking of a lemma was computed as the ratio between the number of the occurrences of the respective lemma and the number of the occurrences of the most frequent lemma of that POS. Therefore the normalised ranking of a lemma is a real number less or equal to 1 (it is 1 only for the

most frequent lemma). For each occurrence of an inflected form of a given lemma, the respective lemma was credited with one more occurrence. The frequency range figures were computed for each part of speech, so that we could select for each part of speech high, medium and low frequency words of the respective category.

The proper names and abbreviations were discarded from the selection process (usually, they are not proper items for explanatory dictionaries).

562 lexical entries from the Bulgarian Explanatory Dictionary (BED) [5], covering the word list produced according to the above-mentioned procedure, were selected. The number is slightly greater than 500 because the dictionary contained multiple entries for homographs. It includes some reference entries as well. These 562 lexical entries contain information for 591 lemmas, because some of the entries contain more than one lemma (for instance, masculine and feminine forms for some nouns). As to the breakdown of lemmas to parts of speech, the CONCEDE consortium agreed upon the following principal breakdown: open parts of speech (nouns, verbs, adjectives, adverbs) – no more than 90 %, closed parts of speech (numerals, pronouns, conjunctions, prepositions, particles and interjections) – minimum10 % out of the whole set of lemmas chosen.

The chosen entries are divided in the following POS:

Noun	200	33.84%
Verb	130	21.99%
Adjective	74	12.52%
Adverb	68	11.51%
Total (open)	472	79.86%
Numeral	9	1.52%
Pronoun	31	5.24%
Conjunction	24	4.06%
Preposition	21	3.55%
Particle	26	4.40%
Interjection	8	1.35%
Total (closed)	119	20.13%
Total	**591**	**100%**

3. Encoding scheme

The CONCEDE languages use different character sets and the dictionaries contain symbols not present in ASCII. All LDBs use 8-bit encoding defined in one of the ISO 8859 standards: Bulgarian LDBs uses ISO 8859-5 (Cyrillic), Czech, Hungarian, Romanian – ISO 8859-2 (Latin 2).

In phase 1 of CONCEDE project each of the 500-headword samples uses a different input formalism – Document Type Definition (DTD) – for producing well-structured SGML-document (Standard Generalized Markup Language-document). Bulgarian sample used own DTD – BG DTD, Estonian – TEI-type, Czech, Hungarian, Romanian used TEI-type DTD with local extensions. The phase 2 must ensure a harmonisation of the resources, and produced a portable uniform SGML resource. Hence, an up-translation step was evaluated. Up-translation is a process of transforming of a dictionary in given format into a more useful/ richer format by program and, if neces-

sary, by human intervention. The dictionaries are large and complex (many types of information, cases, different structures, etc.) In the first phase of CONCEDE project a 500-headword sample of each language was up-translated into TEI-based SGML-document. In parallel with the up-translation, after a long discourse, a formal grammar for CONCEDE and other lexical databases have been prepared, in the form of an SGML Document Type Definition, the CONCEDE DTD. It has taken forward some of the ideas discussed in the TEI Dictionaries Working Group, which were not implemented in TEI guidelines owing to the demand for those guidelines to be highly permissive. In CONCEDE, all dictionaries use common tags, all were encoded according to the TEI. The CONCEDE DTD aims to be a language-neutral, dictionary-neutral framework for presenting lexical information which has not been compromised in its generality by the characteristics of any of the CONCEDE dictionaries. (***See Appendix 1*** *for more information on CONCEDE DTD.*)

The starting point of the Bulgarian LDBs was the Bulgarian Explanatory Dictionary (BED) which is available in electronic form (MS Word for DOS). However, there turned out to be major differences between the preliminary phase DTD and the structure of the dictionary that reflects the language specifics. Here is a list of the discrepancies that had to be accounted for in the DTD:

- Multiple headwords: BED contains a lot of lexical entries that have more than one headword. These are usually derivational forms (masculine and feminine forms for nouns and perfect and imperfect forms for verbs) and paradigm members for irregular lexemes. To any of these headwords it is possible to find some grammatical or stylistic information.

- Subheadwords: In some entries of BED (usually verbs) some derivational forms or different uses of the headword are given. Such forms are followed by a list of senses.

- Introductory note: Common notes like "As a preposition" precedes some times a set of senses.

- The etymology is at the end of the entry.

- Some senses in BED are marked with small or capital letters.

- Sometimes inside a definition there are some grammatical remarks.

- Examples in BED entries contain two types of information - the example itself and the source of the example.

- Phraseology in BED is a list of phrases with some grammatical, stylistic information, definition and examples.

For example, the entry in the paper Bulgarian Explanatory Dictionary:

Без *предл*. Означава: **1**. Лишеност от нещо, липса на нещо. *Мъж без пари и къща без жени огън да ги гори.* Посл. *Без дъно крина – празен хамбар*. Посл. Излезе без шапка и горна дреха. **2**. Отделяне, откъсване, изваждане, отнемане. *Дружината без трета рота излезе на позиция. Десет без три е седем*. **Без време** – преждевременно, не навреме, много бързо. *Без време осърна, без време олете*. П.Р.Сл. **Без да** *сз* – подчинителен обстоятелствен съюз за начин, който показва, че действието в главното изречение се извършва при отсъствие на действие от подчиненото. *Заминал, без да се обади.* **Без друго** – непременно, положително, сигурно; бездруго. *Без друго ще дойда.* **Без малко**.

The corresponding entry in the LDBs:

```
<entry><hw>без</hw>
<pos>предл.</pos>
<struc type=”Sense” n=”1">
        <def>Лишеност от нещо, липса на нещо.</def>
        <eg><q>Мъж без пари и къща без жени огън да ги гори.</q><source>Посл.</source></eg>
        <eg><q>Без дъно крина - празен хамбар.</q><source>Посл.</source></eg>
        <eg><q>Излезе без шапка и горна дреха.</q></eg>
</struc>
<struc type=”Sense” n=”2">
        <def>Отделяне, откъсване, изваждане, отнемане.</def>
        <eg><q>Дружината без трета рота излезе на позиция.</q></eg>
        <eg><q>Десет без три е седем.</q></eg>
</struc>
<struc type=”Phrases”>
        <struc type=”Phrase” n=”1"><orth>Без време.</orth>
                <def>Преждевременно, не навреме, много бързо.</def>
                <eg><q>Без време осърна, без време олете.</q><source>П.Р.Сл.</source></eg>
        </struc>
        <struc type=”Phrase” n=”2"><orth>Без да.</orth><pos>сз.</pos>
                <def>Подчинителен обстоятелствен съюз за начин, който показва, че действието в
                главното изречение се извършва при отсъствие на действие от подчиненото.</def>
                <eg><q>Заминал, без да се обади.</q></eg>
        </struc>
        <struc type=”Phrase” n=”3"><orth>Без друго</orth>
                <def>Непременно, положително, сигурно; бездруго.</def>
                <eg><q>Без друго ще дойда.</q></eg></struc>
        <struc type=”Phrase” n=”4">
                <orth>Без малко.</orth>
        </struc>
</struc>
</entry>
```

4. Validation process

The validation process of CONCEDE LDBs has two aspects: a *formal* validation and a *content* one. The formal validation ensures that the each LDBs is valid SGML document. The SGML declaration defines the concrete SGML syntax used on a system, so the process of validation involved harmonisation of the resources, and produced a portable SGML resource, comprising the large (500) and small (10) samples. Each of them is an SGML document and contains an SGML declaration, SGML DTDs, as well as the complete TEI and SGML character entity sets.

The formal validation consisted of checking that all the samples are valid applications of ISO 8879: Standard Generalized Markup Language (SGML). The objects of the validation procedure for each language were: a document containing the 500 entries, and a DTD giving the SGML structure of the document.

The formal validation ensured that all the documents could be parsed with a validating SGML parser (*SP/nsgmls by James Clark*), using a common SGML declaration, and that the DTD used by the document is constructed according to the TEI guidelines. Additionally, all the dictionary samples, together with their complete SGML environment (entities, DTDs) are available in distribution format.

The content validation required the human examination to produce an dictionary encoded in SGLM-format according CONCEDE DTD without errors in senses and in used tagsets. (***See Appendix 2***.)

Here is a short description of the validation process of the dictionary entries in the Bulgarian dictionary for CONCEDE LDBs. The dictionary contains 2700 selected lexical entries, automatically extracted from the Bulgarian Explanatory Dictionary available in electronic form (MS Word for DOS). The lexical items were converted from MS Word to SGML-format by a program developed especially to this aim. The program uses the typography structure of the MS Word files and marks all unrecognised elements of the structure in a special tag. The entries in Bulgarian dictionary for CONCEDE LDBs save as much as possible the structure of the original paper dictionary.

The formal validation has been done by means of a validating SGML-parser (nsgmls) and the content of the entries was validated manually. The more detailed validation on Bulgarian data was done during the check of the result after the conversion of the entries from MS Word format into SGML format based on the CONCEDE DTD. The marked-up entries were compared with the structure of the paper entries. Special attention was paid to the entries that contained some unrecognised elements of the typography structure of the original dictionary. Main problems were found in the embedded subentries, such as derivational forms and special usage of the headword, grammatical descriptions in the definitions, or phraseology part of entries. In phraseology each phrase can be regarded as a subentry and some time it has a very complicated structure including several senses and examples. In some cases the entries were restructured in order not to introduce new tags only for one or two entries.

CONCEDE evaluated the preliminary phase databases in two ways: first formally, by validating them with an SGML parser, which ensured whether they complied with the formal grammar, or DTD that they were associated with; second, by determining whether equivalent types of data in dictionaries for different languages had been encoded in corresponding ways via a manual examination of a sample of entries. Where discrepancies were discovered, they were examined closely and strategies were developed, aiming at maximum consistency.

5. Conclusion

The combined results of the projects CONCEDE and MULTEXT-East will constitute an integrated multilingual resource of great value to researchers and application developers for the CONCEDE languages.

For more information on the MULTEXT-East project visit http://nl.ijs.si/ME/

For more information on the CONCEDE project visit
http://www.itri.brighton.ac.uk/projects/concede/

For more information on the TEI visit http://www.hcu.ox.ac.uk/TEI/

For more information on the CES visit http://www.cs.vassar.edu/CES/

Acknowledgements: The authors express their gratitude to all partners of the projects CONCEDE and MULTEXT-East.

References

1. Text Encoding Initiative. Background and Context. (Nancy Ide, Jean Veronis (eds.)), Dordrecht, Boston, London, Kluwer Academic Publishers, 1995.
2. Dimitrova, L., T. Erjavec, N. Ide, H.-J. Kaalep, V. Petkevic, D. Tufis. Multext_East: parallel and comparable corpora and lexicons for six Central and Eastern European languages. – In COLING-ACL'98, 315-319, Montreal, Quebec, Canada, 1998.
3. Tufis, D., G. Rotariu, A.-M. Barbu. Data sampling, lemma selection and a core explanatory dictionary of Romanian. – In COMPLEX'99, Pecs, Hungary, 1999, 219-228.
4. Ide, N. Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. – In: First International Conference on Language Resources and Evaluation, LREC'98, Granada, ELRA, 463-470, 1998.
5. Andreichin, L. et al. Bulgarian Explanatory Dictionary. 4th revised edition. Dimitar G. Popov editor. Sofia, Nauka i izkuvstvo Publishing House, 1994.

Appendix 1: CONCEDE DTD

```
<!— CONCEDE project - Deliverable DR2.1: concede.dtd                    →
<!— copyright CONCEDE project consortium, 1999                          →

<!— ENTITY DECLARATIONS                                                 →

<!ENTITY % a.global '
        id                      ID              #IMPLIED
        n                       CDATA           #IMPLIED
        lang                    IDREF           #IMPLIED'               >

<!ENTITY % a.text '%a.global
        rend                    CDATA           #IMPLIED
        wsd                     CDATA           #IMPLIED'               >

<!—ENTITY % ces.header PUBLIC "-//CES//ENTITIES Header//EN"             →
<!ENTITY % ces.header SYSTEM "header.elt"                               >
%ces.header;

<!ENTITY % basetags
        '( orth | pron | hyph | syll | stress | pos | gen | case |
        number | gram | tns | mood | q | source | gloss |
        usg | def | per | aspect | degree |voice |
        eg | etym | xr | trans | itype |subc)'                          >

<!ENTITY % dictbase.seq '(#PCDATA | na)*'                               >

<!— STRUCTURAL ELEMENTS                                                 →

<!ELEMENT cesDic - - (cesHeader, body)                                  >
<!ATTLIST cesDic  %a.global;
        type                    CDATA           #IMPLIED
        version                 CDATA           #REQUIRED
        TEIform                 CDATA           'teiCorpus.2'           >

<!ELEMENT body - - (entry+)                                             >
<!ATTLIST body   %a.global;
        type                    CDATA           #IMPLIED                >
```

```
<!ELEMENT entry - -
                (hw, (%basetags; | struc | alt | brack)*)                    >
<!ATTLIST entry         %a.global;
        type                    CDATA           #IMPLIED                    >
<!ELEMENT struc - -     (%basetags; | struc | alt | brack)*                  >
<!ELEMENT trans - -     (%basetags; | struc | alt | brack)*                  >
<!ELEMENT alt   - -     (%basetags; | brack )*                               >
<!ELEMENT brack         - -     (%basetags;)*                                >

<!ATTLIST (struc, trans, alt, brack) %a.global;
        type                    CDATA           #IMPLIED                    >

<!— CONTENT ELEMENTS                                                         →
<!ELEMENT       ( hw| hyph | syll | stress | pos | gen
                | case | number | gram | source | lang | m
                | itype | tns | mood | subc | na | per | aspect
                | degree |voice )
                                - -                     (%dictbase.seq;)    >
<!ATTLIST        ( hw| hyph | syll | stress | pos | gen
                | case | number | gram | source | lang | m
                | itype | tns | mood | subc | na | per | aspect
                | degree |voice )
%a.text;                >

<!ELEMENT eg                - -             (source | q | gloss)*           >
<!ATTLIST eg            %a.global;                                           >

<!ELEMENT pron              - -             (%dictbase.seq;)                >
<!ELEMENT q                 - -
                (%dictbase.seq; | gloss |ptr |xptr | oref)*                  >
<!ELEMENT etym - -
                (%dictbase.seq; | gloss | lang | m |ptr |xptr | oref)*       >
<!ELEMENT xr    - -     (%dictbase.seq; | ptr |xptr )*                       >
<!ELEMENT (def | gloss)
                - -     (%dictbase.seq; | ptr |xptr |oref )*                 >
<!ATTLIST (pron | q | etym | xr |def | gloss )
%a.text;
      type              CDATA                           #IMPLIED              >

<!ELEMENT orth  - -     (%dictbase.seq; | ptr |xptr |oref )*                  >
<!ATTLIST orth %a.text;
                expansion       NMTOKEN         #IMPLIED
                extent  (full | pref | suff | part ) full
                type            CDATA           #IMPLIED                    >

<!ELEMENT       usg             - -             (%dictbase.seq;)            >
<!ATTLIST       usg     %a.text;
                type    (syn | hyper | colloc | comp | plev
                | acc | lang | gram | obj | subj | verb
                | hint | geo | dom |register | time
                | style | hyponym | antonym | other) other                   >

<!ELEMENT       oref    - O     EMPTY                                        >
<!ATTLIST       oref            &a.text;
                target          IDREF           #IMPLIED
                fullform        NMTOKEN         #IMPLIED                    >
```

```
<!ELEMENT    ptr        - O    EMPTY                                  >
<!ATTLIST    ptr
             corresp           IDREFS          #IMPLIED
             next              IDREF           #IMPLIED
             prev              IDREF           #IMPLIED
             type              CDATA           #IMPLIED
             resp              CDATA           #IMPLIED
             crdate            CDATA           #IMPLIED
             targType          NAMES           #IMPLIED
             targOrder         (y | n | u)     u
             evaluate          (all | one | none)  #IMPLIED
             target            IDREFS          #REQUIRED              >

<!ELEMENT    xptr       - O    EMPTY                                  >
<!ATTLIST    xptr
             corresp           IDREFS          #IMPLIED
             next              IDREF           #IMPLIED
             prev              IDREF           #IMPLIED
             type              CDATA           #IMPLIED
             resp              CDATA           #IMPLIED
             crdate            CDATA           #IMPLIED
             targType          NAMES           #IMPLIED
             targOrder         (y | n | u)     u
             evaluate          (all | one | none)  #IMPLIED
             target            NMTOKEN         #REQUIRED              >
```

Appendix 2: Basetag Descriptions

BASETAGS

case contains grammatical case information given by a dictionary for a given form.
def directly contains the text of the definition
domain domain
Redundant 21.12.99: contains an example text containing at least one occurrence of the word form, used in the sense being described; examples may be quoted from (named) au thors or contrived. Content model allows **source**, **q** and **gloss**, not PCDATA. **Change**
eg **at 21-12-99**: now that **brack** has been introduced for bracketing, **eg** is no longer required but can always be replaced by brack. **source**, **q** and **gloss** now included the basetags so we no longer need an extra layer of structure between **struc** and **q**. I've left **eg** in for back wards-compatibility. AK.
marks a block of etymological information. Etymologies are not a priority in
eg CONCEDE and the only further structure allowed is that lang and m are permitted content as well as PCDATA.
gen identifies the morphological gender of a lexical item, as given in the dictionary.
geo geographic area
gloss a gloss explains an example *not TEI*
contains grammatical information relating to a term, word, or form **other than** gender, number, case, person, tense, mood, itype — as these all have their own element.
hw the citation form; the headword. Used for alphabeticising and primary means of indexing, access. CONCEDE requires that two entries do nto share a **hw**. *not TEI*
hyph contains a hyphenated form of a dictionary headword, or hyphenation information in some other form.
itype indicates the inflectional class associated with a lexical item.
lang Language; for use in etymologies.
m represents a grammatical morpheme (in the context of etymology – which is very lightly marked up)

mood contains information about the grammatical mood of verbs (e.g. “indicative”, “subjunc tive”, “imperative”)
number indicates grammatical number associated with a form, as given in a dictionary.
orth gives the orthographic form of a dictionary headword.
pos ndicates the part of speech assigned to a dictionary headword (noun, verb, adjective, etc.)
pron contains the pronunciation(s) of the word.
ptr Empty element with attribute oref for crossreferences.
q contains a quotation or apparent quotation, eg the PCDATA of an example.
register Register (‘upgraded’ from the TEI value `reg’ for type attribute on **usg**)
source Bibliographic source for a quotation. *Not in TEI*
stress contains the stress pattern for a dictionary headword, if given separately.
style figurative, literal, etc. (Promoted from value for type attribute on USG in TEI)
subc contains subcategorization information (transitive/intransitive, countable/non-count, etc.)
syll contains the syllabification of the headword.
time temporal, historical era (“archaic”, “old”, etc.) (Promoted from value for type attribute on USG in TEI)
syll contains the syllabification of the headword.
tns ndicates the grammatical tense associated with a given inflected form in a dictionary.
trans contains translation text and related information (within an entry in a multilingual diction ary) so may contain any of the basetags. The principle is that everything under **trans** realtes to the target language.
usg contains usage information in a dictionary entry. **other than** time, dom, register, style — as these all have their own element. Other items specified in TEI as suitable values for the type attribute are retained, viz: plev (preference level -“chiefly”, “usually”, etc.), acc (ac ceptability), lang (language for foreign words, spellings, pronunciations, etc.), syn (syno- nym given to show use), hyper (hypernym given to show usage), colloc (collocation given to show usage), comp (typical complement), obj (typical object), subj (typical subject), verb (typical verb), hint (unclassifiable piece of information to guide sense choice). These items are often given in bilinugal dictiaonries to guide the choice of an appropriate translation.
xr used to group all the text relating to a cross reference (PCDATA) with the pointer (empty PTR element with attribute oref)

Structure tags (relation to TEI)

entry Essentially as in TEI
struc Not in TEI
alt There is a TEI tag ‘alt’, also meaning alternation, though generally for use in quite different contexts (chapter 14 of P4)
and Not in TEI

Българският речник в многоезикова база данни

Людмила Димитрова, Радослав Павлов**, Кирил Симов****

** Институт по математика и информатика, 1113 София*
*** Институт по информационни технологии, 1113 София*
**** Централна лаборатория за паралелна обработка, 1113 София*

(Р е з ю м е)

Статията описва накратко процеса на създаване на българския речник за проекта CONCEDE, чиято цел е хармонизиране на методологията, средствата и ресурсите за създаване на лексическа база данни за шест езика: български, чешки, естонски, унгарски, румънски и словенски.