COGNITIVE STUDIES | ÉTUDES COGNITIVES, 13: 57–65
SOW Publishing House, Warsaw 2013

DOI: 10.11649/cs.2013.003

SVETLA KOEVA[1] & **DIANA BLAGOEVA**[1]
[1]Institute for Bulgarian Language. Bulgarian Academy of Sciences

THE DICTIONARY WRITING SYSTEM LEXIT AND ITS APPLICATION IN BILINGUAL LEXICOGRAPHY[1]

Abstract

The paper presents *LexIt* — a web-based Dictionary Writing System designed for the creation and enlargement of dictionaries. A wide range of dictionaries — monolingual, specialised, multilingual, thesauri, can be configured with the system, as it enables users to model, copy, and save templates describing the structure of any dictionary. The paper describes a model for a bilingual dictionary containing phonetic, grammatical, stylistic, semantic, and collocational information and its configuration within LexIt. A bilingual dictionary is chosen for a demonstration, as its entries have a complex structure that includes various elements reflecting the correspondences between source and target languages.
Keywords: Dictionary Writing System, bilingual lexicography.

1. The Dictionary Writing System *LexIt*

The Dictionary Writing System *LexIt*[2] is a web-based system designed for the creation and development of dictionaries. The system design (developed in close collaboration between lexicographers, computational linguists and computer scientists) allows it to be used for the compilation of wide range of dictionaries — monolingual, specialised, multilingual, thesauri, etc.

The key features of the system are as follows:

1.1. Making a new dictionary. New dictionaries can be configured in two ways:

a) By defining the structure of the dictionary and importing a word list. The structure of the dictionary entries is either predefined, or the collaborators can build it dynamically, "dragging and dropping" predefined building blocks.

b) Through integration of existing dictionaries created with *LexIt* and further filtering using particular parameters (the latter is possible because each dictionary

[1] *LexIt* is developed with the financial support of the Fund for scientific research at the Bulgarian Ministry of Education, Youth and Science, contract DTP 02-53/2009.

[2] `http://dcl.bas.bg/LexIt/`

component has a clearly defined structure). For example, a dictionary can be compiled according to a specific thematic domain or to a given stylistic label.

1.2. Working with dictionaries. *LexIt* helps dictionary developers to create, change, copy, delete, save, and print content from dictionaries and their components, and to search for various types of information and references, while it also provides simultaneous access for multiple users and a collaboration medium for them. Lexicographers can compare, copy and combine dictionary entries from one or more dictionaries, share examples, notes, and other information. Access rights determine the different scopes of actions available for lexicographers, editors and viewers. The dictionary compilation process can be managed by assigning different tasks and distributing workload among different users, as well as by monitoring the completion of the work and the consistency of the data.

1.3. Information retrieval filters. Different types of information can be extracted from a dictionary with multiple filtering parameters: author, creation date, content, lexicographic labels, linguistic and extralinguistics relations, and many others, including regular expression patterns. For example, all headwords that begin with the letter sequence "ac", or definitions containing a given word, may be extracted. Various statistics can be compiled — number of headwords, usage examples, synonyms, obsolete words, and many others. The history of entry is saved and can be retrieved if necessary.

1.4. Corpus support. The system allows import of large corpora that can be grouped into directories. Lexicographers can thus have at their disposal suitable texts for extracting usage examples and for validation of semantic granularity. Multi-file, multi-directory regular expression searches are possible. There is also a notebook for users to share comments and notes.

1.5. Dictionary publishing. Data can be exported in various formats for publication: PDF, XML or HTML. Thus the *LexIt* meets the requirements for traditional publishing and web publishing, as well as for data exchange between different Dictionary Writing Systems.

1.6. Interface. *LexIt*'s user interface is simple but functional, based on a three-window display. In dictionary building mode, the leftmost window contains the wordlist, while the rightmost window serves for managing corpora and notes. These two windows are open only when the tasks to be performed call for them. The central window is used for creating and editing dictionary content. It consists of several panes: a) for dynamic design of entries through dragging and dropping components; b) for creating and editing the content of a dictionary entry; c) for display of a print preview of the entry, including formatting of individual elements.

2. Defining the structure of a dictionary with LexIt

LexIt allows defining the dictionary structure at several levels — dictionary entry, dictionary section or dictionary itself. Dictionary entries usually have a complex

structure, and LexIt handles this through smaller components: building blocks, modules, and templates.

This is possible trough an abstract representation of dictionary structure. Each dictionary entry consists of objects (building blocks) characterised by a) content properties, b) structural properties, and c) formal properties.

Content properties indicate the type of object depending on the values that are attributed to it, and are obligatory. Values of objects can be variables (e.g., headword, text note: definition, example, note) or constants (e.g., part of speech, stylistic label, etc.).

The structural properties of objects are used to define the place of the object in the structure of the dictionary and are obligatory — for example, a unique name, an object type description, etc.

The formal properties are optional and they are related to the typographic formatting of the dictionary entry, such as: font, character size, bold, italic, underline, subscript, superscript, capital letters, paired quotation marks, paired brackets, etc.

The objects can be linked with different relations, both in the scope of a particular dictionary entry and among different entries. Relations are semantic (synonyms, antonyms, etc.), morphological (lemma to word forms; lemma to part of speech, etc.), interpretative (lemma to word sense, lemma to definition, etc.), extralinguistic (word sense to thematic domain, lemma to category domain, etc.), derivative (noun to verb, noun to adjective, etc.), morpho-semantic (Agent to Action, Cause to Result, etc.), syntactic (predicate to animate object; predicate to human subject, etc.), phonetic (word to pronunciation), structural (word sense to unique sense number, headword to interlanguage index, etc.). Each relation is specified with a name and indication of its specific properties, for example that hyperonomy is a transitive relation.

Each dictionary entry has a structure, determined by the principles accepted for a particular dictionary and the properties of the headword. When creating a new dictionary entry, lexicographers have the option to choose a template from a repository or to develop a new template that represents the most appropriate entry structure. Objects (building blocks) usually come in a specific order within a dictionary entry, while some of them can be repeated, and in addition there exists between some of them relations such as simultaneous appearance, dependent appearance or exclusion. Objects and modules are characterised by the following obligatory properties:

a) succession: specifies the order of elements (e.g., the building block Headword obligatorily precedes the building block Homograph index);

b) reusability: indicates whether an item can be repeated in a given module (e.g., the building block Example can be repeated in the module Illustrative Examples);

c) mandatory: indicates when an element is required for a module (e.g., the building block Headword is a compulsory element while the building block Homograph index is not).

The following relations can be defined between the elements of a module:

a) concurrency: a given element appears only if another element appears or a particular value is chosen (e.g., if the selected value for the building block Part of speech is Verb, then the building block Verb aspect is obligatory);

b) dependency: the set of values for a given element depends on the selected value for another element (e.g., if the selected value for the building block Part of speech is Noun, then the values of the building block Number are singular, plural, and count form);

c) exclusion: a given element appears only if another element is absent (e.g., there is no building block Number when the building block Grammatical label has the value Singularia tantum).

A module that represents an entire dictionary entry is called a template. For example, the following template can be used for creating and editing a dictionary entry for headwords that are verbal nouns, abstract nouns, diminutive nouns, adjectives, past passive participles, adverbs, etc:

Module Headword
    Building block Headword (mandatory)
    Building block Homonym index
Module Grammatical feature
    Building block Part of speech (mandatory)
    Building block Grammatical feature (the set of its values is dependent on the building block Part of speech)
Module Stylistic and Grammatical note
    Building block Stylistic label
    Building block Thematic domain label
    Building block Grammar usage label
    Module Definition (mandatory, multiple)
    Building block Structural definition (mandatory)
    Building block Structural definition index (mandatory)
    Module Illustrative Examples (mandatory, multiple)
        Building block Example (mandatory, multiple)

3. Application of LexIt in bilingual lexicography

The successful development of any kind of dictionary is determined by the formulation of clear and consistent principles for the macro-, micro- and medio-[3] structure of a dictionary entry and adherence to these principles. In this paper we focus on the modelling and representation of a bilingual dictionary entry and present the application of this model within LexIt. The proposed model is abstract and language-independent with respect to both source and target language. We shall illustrate its application in the construction of a bilingual dictionary where the source language is Bulgarian.[4]

A dictionary entry in a bilingual dictionary has a complex structure. In order to answer the needs of the specific target group of dictionary users it includes elements

[3]The terms macro-and microstructure are well known in the field of theoretical lexicography (Hartmann, 2000), while the term mediostructure needs some clarification. This term, introduced by H. E. Wiegand (1996), refers to various cross-references between elements of different dictionary entries or between the elements of a dictionary entry and another specific part of the macrostructure of the dictionary (e.g., list of abbreviations, list of excerptions, etc.).

[4]A model for a Foreign-Bulgarian dictionary is proposed by S. Pavlova as well (Павлова, 2010).

that reflect various information concerning lexical correspondences between source and target language. Atkins (2000, p. 2–3) notes:

A systematic approach to the study of what a bilingual dictionary does and how it does it must take account of the following aspects of the entry:

- function of that information (what the user can use it for);
- mode of expression (how it is expressed);
- type of user: Source Language speaker or Target Language speaker;
- purpose of use (encoding into a foreign language, or decoding into one's own language).

We offer a model for the microstructure of bilingual dictionaries that contains phonetic, grammatical, stylistic, semantic, and collocational information. Each type of information is recorded in a specific manner within the individual components of the lexical entry, here called zones. The outlined zones in the dictionary entry structure are represented in Fig. 1.

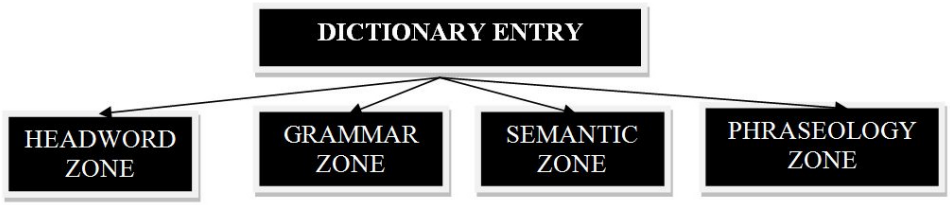

Figure 1: Dictionary entry zones in a bilingual dictionary.

The Headword zone contains the following elements: a headword with marked stress, homonym index (if applicable), headword pronunciation idiosyncrasy (transcription), variant of the headword (e.g., a derivative variant) or second headword (e.g., in some dictionaries the members of verb aspect pairs share one dictionary entry).

The Grammar zone includes particular word forms of the headword (e.g., exemplifying vowel alternations), label for part of speech, and labels for grammatical categories.

The Semantic zone is the main component of the dictionary entry. It reflects the semantic structure of the headword with respect to its translational equivalents in the target language. Different meanings are distinguished through simple illustrations (e.g., with a synonym, with typical collocations, etc.) or through illustrative usage examples.

The term Semantic zone is conventional, since not only semantic information is encoded in this zone. There is also stylistic and pragmatic information — functional domain of the headword or one of its senses, e.g., the label "Biol." for terms belonging to the domain of biology, "Poet." for expressive words and word senses in poetry, "Iron." for words and word senses that have ironic connotation, etc., grammatical information (grammatical labels indicating usage restrictions for a particular word sense, e.g., "Imf. only" to specify that a given sense is manifested only with imperfective verbs), information about the selectional preferences of words by means of

typical collocations, etc.

The complex structure of this zone stems from the fact that in it, the parts that in theoretical lexicography are traditionally called "left" and "right" (headwords in the source language and their translations in the target language) intersect.

This zone gains further complexity if the authors decide to make a bidirectional dictionary; that is, not in the sense of a two-way dictionary, but in the sense of a bilingual dictionary that aims at native speakers of both languages. In this case, the semantic zone has to provide a certain amount of phonological, phonetic, morphological, syntactic, and/or stylistic information about the translations.

The Phraseology zone reflects the set expressions, idioms, and proverbs in which the headword participates, together with their translational equivalents in the target language. The stylistic features of the phraseological units are indicated by stylistic labels.

The proposed structure of a bilingual dictionary entry for verbs can be represented as follows:

Module Headword
- Building block Headword (mandatory)
- Building block Homonym index
- Module Paradigmatic features
  - Building block Inflectional forms (mandatory, multiple)
  - Building block Verb aspect (mandatory)

Module Headword variant
- Building block Headword variant or second headword
- Building block Headword
- Building block Homonym index
- Module Paradigmatic
  - Building block Inflectional forms (mandatory, multiple)
  - Building block Verb aspect (mandatory)
- Module Part of speech
  - Building block Part of speech (mandatory)
  - Building block Grammatical features (mandatory, multiple)
  - Building block Labels of language register/style/areas of specialisation (multiple)

Module Word meaning
- Building block Word meaning index (mandatory, multiple)
- Building block Grammatical label (multiple)
- Building block Semantic label (multiple)
- Building block Labels of language register/style/areas of specialisation (multiple)
- Building block Explanation of the meaning (definition, synonyms, typical subjects or objects of the entry) (mandatory)
- Building block Translation (translation equivalent) (mandatory, multiple)
- Building block Linguistic features (multiple)
- Building block Example (mandatory, multiple)

Module Subheadword

- Building block Headword (mandatory)
- Building block Homonym index
  - Module Paradigmatic features
    - Building block Inflectional forms (mandatory, multiple)
    - Building block Verb aspect (mandatory)

Module Subheadword variant
- Building block Headword variant or second headword
- Building block Headword
- Building block Homonym index
  - Module Paradigmatic features
    - Building block Inflectional forms (mandatory, multiple)
    - Building block Verb aspect (mandatory)
- Module Part of speech
  - Building block Part of speech (mandatory)
  - Building block Grammatical features (mandatory, multiple)
  - Building block Labels of language register/style/areas of specialisation (multiple)

Module Word meaning
- Building block Word meaning index (mandatory, multiple)
- Building block Grammatical label (multiple)
- Building block Semantic label (multiple)
- Building block Labels of language register/style/areas of specialisation (multiple)
- Building block Explanation of the meaning (by definition, synonyms, typical subjects or objects of the entry) (mandatory)
- Building block Translation (word equivalent) (mandatory, multiple)
- Building block Linguistic features (multiple)
- Building block Example (mandatory, multiple)

Module Phraseology
- Building block Phraseological unit
- Building block Labels of language register/style/areas of specialisation (multiple, depends on Building block Phraseological unit)
- Building block Explanation of the meaning (mandatory, depends on Building block Phraseological unit)
- Building block Translation (word equivalent) (mandatory, multiple, depends on Building block Phraseological unit)
- Building block Linguistic features (multiple, depends on Building block Phraseological unit)
- Building block Example (mandatory, multiple, depends on Building block Phraseological unit)
- Building block Example (mandatory, multiple)

Some of the elements in the entry tree structure (e.g., Headword, Part of speech, Translation) are mandatory, while others (e.g., Homograph number, Labels of language register / style / areas of specialisation, Subheadword) are optional. Some

elements are recurrent (e.g., Labels of language register / style / areas of specialisation). Certain relations and dependencies between elements are encoded — their description is important for building a coherent model of dictionary microstructure.

An example of dictionary entry from a Bulgarian-English dictionary built on the proposed model is presented below:

забѝва|м[1] (ш) imf., забѝ|я[1] (еш, past ˜х, p.p.p. ˜т) pf., v.t. 1. (зачуквам) drive, hammer; з. нож в гърдите на нкг plunge a knife into sb's breast; 2. (нокти, зъби) dig, plunge; (fig.) fix, fasten; з. поглед в нщ fix one's eyes on sth; 3. (топка) strike; з. топката във вратата score a goal.

■ ˜м се impf., ˜я се pf., v.i. 1. (проникам) dig, lodge; думите се забиха в съзнанието ми these words stuck into my mind; 2. (fig.) (coll.) (загубвам се) stray, get lost.

◊ з. нос (падам по лице) fall flat on one's face; (съсредоточавам се) pore over.

Example 1: An example of a dictionary entry from a Bulgarian-English dictionary.

Building of the same dictionary entry within LexIt is illustrated in the Fig. 2.

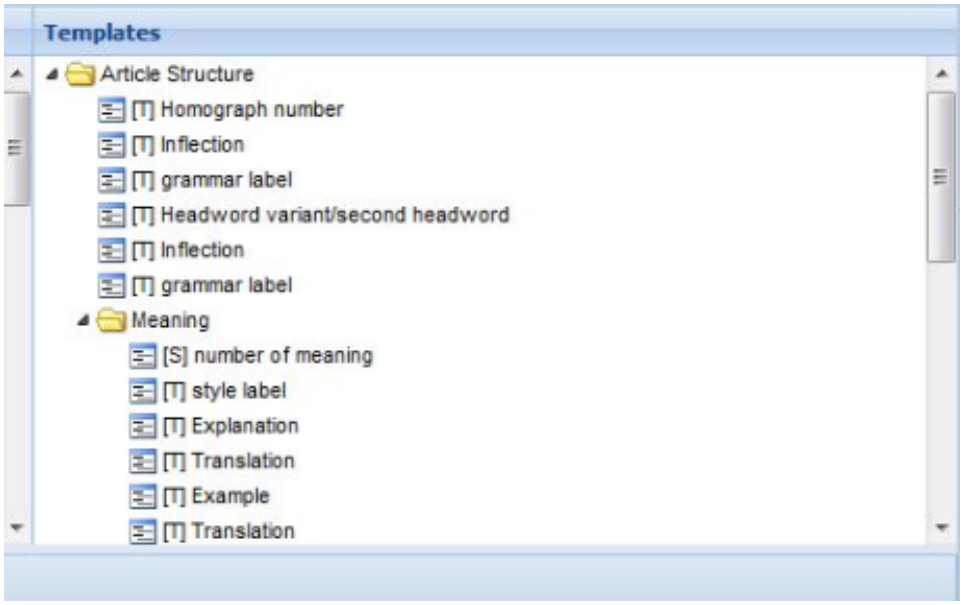

Figure 2: Bilingual dictionary entry created with LexIt.

LexIt allows for already defined modules to be grouped into larger entities (bigger modules) which are structural elements of different dictionary zones. For example, the following complex modules which form the structure of the bilingual dictionary entry are defined: Meaning, Subheadword, Phraseology and others.

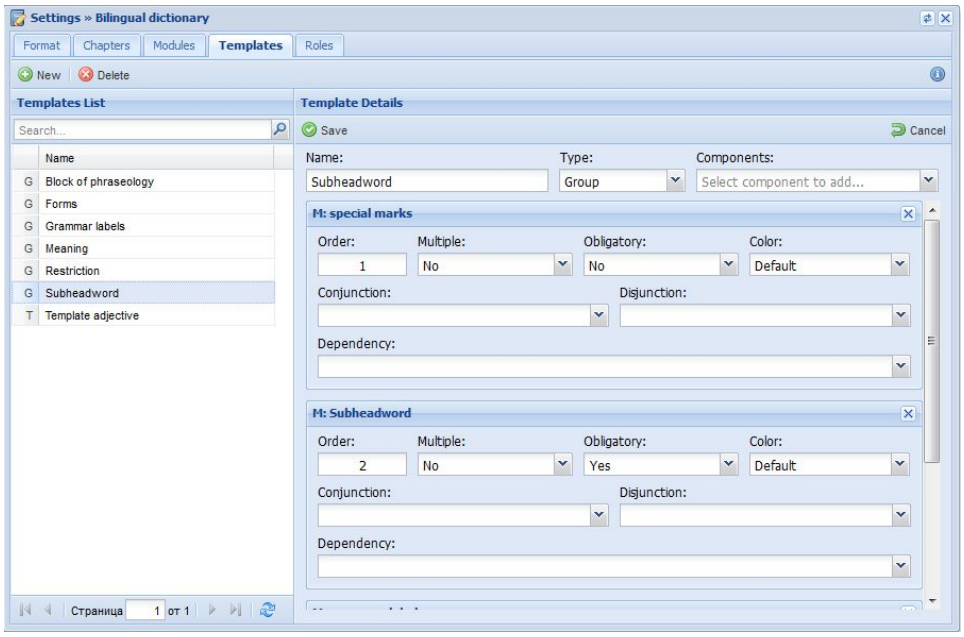

Figure 3: Module Subheadword in LexIt.

Templates for standard dictionary entries (e.g., for different parts of speech) can be built by grouping the relevant modules and complex modules. This facilitates dictionary development and prevents violation of the accepted principles.

4. Conclusion

The diverse functionalities and intuitive interface of LexIt ensure reliable processing and storage of information and professional preparation for publishing. Moreover LexIt increases productivity and reduces the risk of random errors. At present the main functionalities of the system are completed and an active testing phase is in progress in order to eliminate bugs and improve the system.

References

Atkins, B. T. S. (2002). Bilingual Dictionaries — Past, Present and Future. In: *Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins.* M.-H. Corr ard. EURALEX, p. 1–29.

Hartmann, R. R. K. (2000). Structural Perpectives in Dictionary Research. In: За думите и речниците. София: Диос, p. 15–22.

Wiegand, H. E. (1996). Über die Mediostrukturen bei gedruckten Wörterbüchern. In: A. Zettersten, V. Hjørnager Pedersen (Eds.), *Symposium on Lexicography VII.* Tübingen: Niemeyer, p. 11–43.

Павлова, С. (2010). Речниковата статия в двуезичен речник. In: *Лексикографията в европейското културно пространство.* Велико Търново: Знак'94, p. 185–196.