BRIDGE dictionaries

Hana SKOUMALOVÁ, Praha, Czech Republic

Abstract

BRIDGE dictionaries are a new sort of dictionary for learners of English. They can serve for creating new bilingual or multilingual dictionaries. In my paper I discuss how this aim can be achieved, and what techniques we can use. The final product can be either a new bilingual printed dictionary or an electronic multilingual dictionary.

1 Introduction

BRIDGE dictionaries are a new sort of dictionary for learners of English. They are based on the monolingual COBUILD learners' dictionaries, and they are partly translated. BRIDGE dictionaries can be used as a source for new bilingual dictionaries by putting together the translations in the target languages [Sinclair, unpubl.]. In my paper I want to discuss some problems that can occur, and perspectives of this way of creating dictionaries.

I experimented with two translated versions of the Cobuild dictionary—Czech and Lithuanian, but I believe that my conclusions would also be valid for other languages.

2 The source dictionaries

I worked with the electronic form of the dictionaries, as they were prepared for typesetting. The source contains a markup, whose main purpose is to instruct the printing program, but it also serves for structuring the entries. It is possible to distinguish headwords, pronunciation, English definitions, translated definitions, translated equivalents, examples, etc. (cf. figure 1).

Roughly speaking, every entry (`[EB]` and `[EE]` tags) consists of a lemma part (`[LB]` and `[LE]` tags) and a meaning part (`[MB]` and `[ME]` tags). The lemma part contains the English headword, pronunciation, and inflection information. The meaning part consists of one or more meanings, each of which contains the grammatical information, English definition (`[DT]` tag), a translation of the definition, with the English headword untranslated (`[KD]` tag), and one or more translation equivalents (`[KQ]` tag). The meaning part can be followed by phrases or phrasal verbs, which have again the structure of the whole entries. For the further processing of the dictionary, it was necessary to discover the exact structure, which was not explicitly defined. As a possible way to achieve this, I decided to convert the data to SGML format. It turned out, however, that the structure was sometimes inconsistent, or too free; in such cases I set certain rules to ensure that the structure was consistent.

```
[EB]
[LB]
[HW]abacus
[PR]/*!ab%ek%es/,
[IF]abacuses.
[LE]
[MB]
[GR]COUNT N
[DT]An [HH]abacus [DC]is a frame used for counting.
    It has rods with sliding beads on them.
[KD][HH]Abacus [DC]je rámeček, který se používá pro počítání. Má tyčky, po \
kterých kloužou kuličky.
[KQ]počitadlo.
[ME]
[EE]
```

Figure 1: Extract from the source dictionary

3 SGML format

SGML (standard generalized mark-up language) [Goldfarb, 1990, Colby/Jackson, 1998] is widely used for encoding documents, corpora and lexical databases. There are software tools for processing the SGML data which simplify our task.

The SGML form of the dictionaries brings us the following advantages:

- the correctness of the structure can be checked with the help of an SGML parser;
- the corrected text can be returned to the typesetting markup;
- corresponding parts of the two translated dictionaries can be aligned.

The first two facets are useful for any dictionary making, no matter whether we want to combine translations from two or more dictionaries; the third is used for the combination work.

3.1 DTD

For discovering and checking the structure of the dictionary, the best way is to create its DTD (document type definition). Unlike text corpora, the structure of lexical corpus is much richer. The following list contains all elements of the dictionary and shows their hierarchy.

```
<!ELEMENT BRIDGE - O ( entry+ )>
<!ELEMENT entry - - ( lemma, sense*, phrases?, phrasal-verbs? )>
<!ELEMENT lemma - - ( ( headword | spell ),
        (pronun | infl | register-note )*, ( x-ref | lemma-note )* )+>
<!ELEMENT sense - - ( ( grammar*, (definition+ | def-reg-note)+, x-ref*,
        pic-ref*, (def-transl+ | transl-reg-note)+, translation,
        (x-ref+ | pic-ref+ | examples | run-on+ |
        syntax-change | phrase )*) | (phrase | x-ref)+ )>
<!ELEMENT headword - O ( body, punct? )>
<!ELEMENT infl - O ( body, punct? )>
<!ELEMENT spell - O ( body, punct? )>
<!ELEMENT register-note - O ( body, punct? )>
<!ELEMENT lemma-note - O ( body, punct? )>
<!ELEMENT pronun - O ( body, punct? )>
<!ELEMENT x-ref - O ( body, punct? )>
<!ELEMENT pic-ref - O ( body, punct? )>
<!ELEMENT grammar - O ( body, punct? )?>
<!ELEMENT definition - O ( body, punct? )?>
<!ELEMENT def-reg-note - O ( body, punct? )?>
<!ELEMENT def-transl - O ( body, punct? )?>
<!ELEMENT transl-reg-note - O ( body, punct? )?>
<!ELEMENT translation - O ( body, punct? )?>
<!ELEMENT syntax-change - - ( grammar*, (definition+ | def-reg-note)*,
        (def-transl+ | transl-reg-note)*, translation*, examples* )>
<!ELEMENT run-on - - ( headword*, pronun*, grammar*, examples? )>
<!ELEMENT examples - - ( example+ )>
<!ELEMENT example - O ( body, punct? )>
<!ELEMENT phrasal-verbs - - ( phrasal-verb+ )>
<!ELEMENT phrasal-verb - - ( headword+, grammar+, sense+ )>
<!ELEMENT phrases - - ( register-note?, (phrase, x-ref*)+ )>
<!ELEMENT phrase - - ( grammar?, definition+, def-reg-note?,
        def-transl+, transl-reg-note?, translation, examples?, pic-ref* )>
<!ELEMENT body - O (#PCDATA)>
<!ELEMENT punct - O (#PCDATA)>
```

Figure 2: The element list of the dictionary

We can see that the structure is rather complex. The mutual order of some elements is quite free (e.g. definitions and register notes), certain elements can occur as members of several different super-elements (e.g. a headword can occur in a lemma, phrasal verb or run-on rubric; the translation occurs in a phrase or in the sense element of an entry or phrasal verb).

3.2 Data encoded in SGML

As a result of converting the source data into SGML we obtain the structure in which important elements are identified by identifiers. The identifier is constructed from the dictionary letter, entry identifier, and other identifiers distinguishing different parts of the entry.

The conversion consists of two steps. The first step just transforms the source data to SGML, adds missing parts (e.g. empty tags for missing translations), and adds the identifiers. In the second step the program goes through the SGML data and looks for cross-references (attribute

`target` in `x-ref` tag). When the program finds the attribute `target`, it tries to find its value as the content of a headword element of another entry. If the search succeeds, the value of the attribute `target` is replaced by the identifier of the first meaning of the found headword.

```
<entry flat="EB/EE" id=a.e2>
 <lemma flat="LB/LE">
  <headword type=basic flat=HW id=a.e2.h1><body>aback</body>
  <pronun type=basic flat=PR><body>/%eb*!ak/</body><punct>.</punct>
 </lemma>
 <sense type=basic flat="MB/ME" id=a.e2.m0>
  <x-ref type=rubric flat=QS><body>See</body>
  <x-ref type=target flat=QH target=t.e28.ph1.m0><body>take aback</body>
  <punct>.</punct>
 </sense>
</entry>
```

The above example shows the final format of the entry *aback*. In the printed dictionary, this entry looks as follows:

aback /əbæk/. See **take aback**.

and in the source electronic dictionary it has this format:

```
[EB]
[LB]
[HW]aback
[PR]/%eb*!ak/.
[LE]
[MB]
[QQ][QS]See [QH]take aback.
[ME]
[EE]
```

After the first step of conversion, the sense part of the entry has this format:

```
 <sense type=basic flat="MB/ME" id=a.e2.m0>
  <x-ref type=rubric flat=QS><body>See</body>
  <x-ref type=target flat=QH target="take aback"><body>take aback</body>
  <punct>.</punct>
 </sense>
```

In the second step, the first phrasal verb in the entry *take* is identified as the target of the cross-reference and its identifier `t.e28.ph1.m0` replaces the value of the `target` attribute in the entry *aback*.

4 Alignment

The basic alignment is made at the level of entries, using the content of the element `headword`. The two original dictionaries are expected to be identical, but it may happen that some entries are missing in one of them. The basic alignment ensures that we identify all such gaps. To align the translations we have to parse the entries so that can combine the translations from corresponding parts of the entries.

4.1 Combining the information

Once we have two dictionaries in SGML format and alignment has been achieved, we can create a new bilingual dictionary, or more precisely speaking, we can create a list of possible correspondencies.

We extract the corresponding translations, then divide multiple left-hand sides to single entries, sort the entries alphabetically and offer the result to lexicographers for further editing. Such a dictionary is imperfect in many respects; for detailed discussion see [Skoumalová, in press]. In this paper we will concentrate on flaws that can be removed by more detailed tagging of the source dictionary:

- On one or both sides we get an explanation, rather than a translation (e.g. for definite and indefinite articles). Such "translations" should be marked by a special tag:
  `[KQ][KE]neurčitý člen`.
- If we get a multi-word expression on the left-hand side, the sorting program cannot identify the headword and can place the entry incorrectly. For example, *ženské lékařství—gynaecology* should be sorted under *L* rather than under *Ž*. If the headword is not at the first position, a special tag should be used:
  `[KQ]ženské [KS]lékařství`.
- Variant attributes belonging to one headword are separated by commas and they are indistinguishable from variant translations. For example the verb, *gallop* is translated to Czech as *běžet tryskem, (rychlým) cvalem, cválat*, where the words *tryskem* and *(rychlým) cvalem* are variant modifications of the verb *běžet*, and the verb *cválat* is a translation variant. Again, a special tag should be used:
  `[KQ]běžet tryskem, (rychlým) cvalem, [KV]cválat`.
- The meaning of the headword can be restricted by an expression in parentheses, which may be in English or in the target language. Examples:
  galloping—rychlý, prudký, (inflace) pádivý
  gardener—zahradník, (allotment, hobby) zahrádkář
  The English expressions should be, again, marked by a special tag:
  `[KQ]zahradník, [KV][KX](allotment, hobby) [KC]zahrádkář`.

These additions to the original tagging would improve the output and help the lexicographers in their work on the new dictionary.

5 Further perspectives

Though the dictionary obtained by combining the two BRIDGE dictionaries is far from perfect, it can be used as a starting point for the deeper lexicographical work.

Another possibility for utilizing the data is to make an electronic dictionary browser. The browser allows one to go through English entries and displays the corresponding translations in both target languages. It also enables one to go through translation equivalents and display the original English entry, as well as the corresponding equivalents in the other language. A great advantage of this approach is that adding a new language is quite easy. The only condition is that it is based on the same version of Cobuild dictionary. This way we can create multilingual electronic dictionaries which can be useful especially for "small" languages, for which a printed bilingual dictionary would probably not be made.

6 Acknowledgment

I would like to express my thanks to František Čermák, who inspired me to do this work and who provided me with the electronic version of the Czech dictionary. I would also like to thank Rūta Marcinkevičienė and Andrius Utka, who provided me with the Lithuanian part of the dictionary.

The work described in this paper was partly supported by grant GAČR 405/99/0540.

References

[Blatná/Čermák, 1995] Renata Blatná and František Čermák, editors: *Manuál lexikografie*. H & H, Prague, 1995.

[Colby/Jackson, 1998] Martin Colby and David S. Jackson: *Using SGML*. Que, 1998.

[Dictionary, 1998] *Anglicko-český výkladový slovník.* Nakladatelství Lidové Noviny, Prague, 1998.

[Goldfarb, 1990] Charles F. Goldfarb: *The SGML Handbook.* Oxford University Press, Oxford, 1990.

[Sinclair, 1987] John M. Sinclair, editor: *Looking Up—An account of the COBUILD Project in lexical computing.* Collins ELT, London and Glasgow, 1987.

[Sinclair, unpubl.] John M. Sinclair: "Bridge club", unpublished manuscript.

[Skoumalová, in press] Hana Skoumalová: "BRIDGE Dictionaries as Bridges Between Languages", in: *Proceedings of the 4th TELRI European Seminar.* TELRI, in press.