Towards a Multilingual Valency Dictionary
German-English-Modern Greek

Wolf Paprotté, Arbeitsbereich Linguistik, WWU Münster
Panagiota Balanga Arbeitsbereich Linguistik, WWU Münster
Lothar Lemnitzer, ZERES GmbH, Bochum

Abstract
Aims:
A corpus-based, multilingual valency dictionary of English, German and Modern Greek verbs is under development at the Arbeitsbereich Linguistik, WWU, Münster. The dictionary is intended for human and machine users.

The lexical acquisition process:
The basis of the project is large corpora in German, English and Modern Greek. In the first stage a set of lexicographic tools (xlex) for segmentation, indexing, concordancing, tagging, term extraction and collocation analysis has been developed for the purposes of this project. These tools will be used to preprocess the data. In the second stage of the lexicographic process, complementation patterns have been extracted interactively; the preprocessed data are categorized and presented in a suitable format for subcategorization frames. A number of dictionary entries, as the results of the automatic analysis, will be presented and the underlying assumptions and principles discussed. In order to achieve a platform-independence, HTML, was chosen as a mark-up language.

1. Introduction
A range of typical mistakes learners even of an advanced level make when using German, English or Modern Greek as a foreign language are due to idiosyncratic features of verbs in these languages. Depending on the linguistic theory, these idiosyncratic features are treated in linguistic and lexicological theory under the labels "valency" or "subcategorization". Idiosyncratic as they are, they can not be treated adequately in a rule system of a grammar. The phenomena to be dealt with are clearly subject to a lexical description of the lexical items. They form part of the lexicon of a language. The traditional treatment of these features is the "transitive" - "intransitive" distinction, which can be found in most bilingual dictionaries. However, this broad distinction is inadequate and therefore of little use for the language learner. Take for example the lexical equations

see – sehen - βλέπω (vlepo)
help – helfen - βοηθώ (voiθo)

remember – (sich) erinnern - θυμάμαι θιμάμετε

In Standard Modern Greek as well as in English there is no structural difference in sentences like

τον βλέπω (ton vlepo) - I see him
τον βοηθώ (ton voitho) - I help him
τον θυμάμαι (ton thimame) - I remember him

This fact contrasts to the German equivalents

Ich sehe ihn (acc)
Ich helfe ihm (dat)
Ich erinnere mich seiner (gen)

In German "sehen" calls for a direct object in the accusative case, "helfen" calls for an indirect object in the dative case, and "erinnern" can take a complement in the genitive case. Obviously, there is more to the picture than "transitivity".

2. Theoretical Background
The concept we use for the description of the phenomena involving complementation of verbal lexemes in English, German and Standard Modern Greek is that of valency. This concept originates in the grammatical theory of Lucien Tesnière (1953, 1959). He introduced dependency grammar as a descriptive device. Valency theory is a central part of the formalism. Tesnière starts from the assumption that the verb is the structural centre of the sentence. It governs the rest of lexical elements, the complements, within the sentence. The valency of a lexical element designates the ability of taking a certain number and types of complements as the nucleus of a well-formed sentence. Consequently, verbal valency is associated with number, syntactic nature and obligatoriness of the complements a verbal lexical entry needs. These properties of a verbal lexical element involving its subcategorization frame2 form the core of a lexical description in a specialised "valency dictionary". In short, a valency dictionary should determine the subcategorization potential of verbal lexical items, including syntactic-categorial, syntactic-functional and semantic features of the complements.

The multilingual valency dictionary which we are compiling in Münster is intended first and foremost for the human user, particularly for learners of these languages. However, it is also targeted at machine "users", that is programmes for natural language processing: applications such as machine

---

2 See e.g. the theoretic framework of HPSG laid out in Pollard/Sag (1987).
translations, syntactic parsing, information retrieval etc. Consequently, there are two principles which guide the compilation of the multilingual valency dictionary, the empirical and the monolingual orientation. As far as the empirical experience is concerned, the basis of our descriptive work are large corpora of all three languages (by large, we mean a size of 20 million to 200 million running words) restriction to verbal lexical elements. Moreover, we restrict ourselves to the 5000 most frequent verbal lexical elements. Frequency is measured by occurrence of all inflected forms of a verbal lexical element in the text corpus. All three dictionaries English, German, Standard Modern Greek - are compiled with a strict monolingual orientation in mind. The linking of all three dictionaries will be done in a second step.

The information programme for each lexical entry consists of a formal description of the valency or subcategorization frames, targeted primarily to machine users. It also consists of more easy to read descriptions with text variables which are well known to the human user from his knowledge of bilingual dictionaries; both levels of abstract lexical description are illustrated by citations taken from the corpora. Morphosyntactic information - e.g. about the perfect auxiliary - complete the information programme. From the users' point of view, the dictionary comprises three monolingual parts with a monolingual morphosyntactic description of the entries with pointers to entries of the other two dictionaries. The user who follows these pointers is referred the complete morphosyntactic description of the target entry. This multilingual valency dictionary will therefore be a lexical help for the production of texts in the foreign language.

3. Course of development
The development of the dictionary\textsuperscript{3} is carried out in several steps:
- Accumulation of corpora
- Development of a conceptual scheme which is common for the representation of the valency information in all three monolingual dictionaries
- Development of three monolingual dictionaries
- Linking of the three monolingual sources
- Implementation of a lexical database

3.1 Accumulation of corpora
The basis of the project are primarily large corpora in German (100 million running words), English (200 mill. rw) and Modern Greek (20 mill. rw); secondary bases of the work are lexicological descriptions, monographies etc., as well as the intuition of the lexicographers. The latter are used in the

\textsuperscript{3} For more information see Storrer (1992).
case of lexical gaps in the corpora. The corpora are the source for the examples of authentic usage.

3.2 Development of a conceptual scheme
The next step is the preprocess of the data by using a set of lexicographic tools developed at the Arbeitsbereich Linguistik in Münster (called xlex). xlex comprises tools for index building, segmentation, concordance generation and co-occurrence analysis. For the lexical description, we start with the concordance lines for each lexical entry. In the following a part of the results of an automatic analysis, some concordance lines with the English verb "spoil", are presented. The information, given within the pointed brackets < > are the source src, e.g. L1, and the example, e.g. no. 176:

...<src=L1/176> "I ain't the man to [spoil] a bit of sport.
<src=L1/266> "It was the voice of the grown-up brother admonishing
the baby about his gravy-stained bib, but Pauli didn't allow the
condescension to [spoil] things.
<src=L2/195> I could breathe a tropical bouquet from the woven velvet
flowers on the wall, I was near a swamp where butterflies and tropical
birds went fanning up - and over the [spoil] of animals who looked for
flesh - and floated on the air which rose from vegetation growing in the
damp and drowning in the wet ...

The next step within this stage is to order these lines, which give useful hints to the number and type of complements the keyword typically subcategorizes for. The contextual elements of the keyword are ordered by the significance with which they co-occur with the keyword. For this analysis we consider three words before and three words after the key-word. Function words like "and" or "or" were eliminated. According to a statistical method we extract those words which co-occur significantly with the key-word and note the frequency (f) of the collocate (y) together with the key-word (the collocate x) in the pattern 0/0/0/X/0/0/0. The registration of the collocate (y) results from to the frequency (f/cy).

In the following you can see the co-occurrence pattern for the verb "spoil":

collocator (x): spoil, spoils, spoilt, spoiling, spoiled
frequency: 630
corpus size: 51332448
Formula: "maximum likelihood ratio"
<table>
<thead>
<tr>
<th>collocate (y)</th>
<th>F(xy)</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>choice</td>
<td>32</td>
<td>0/0/0/0/X/0/32/0/</td>
</tr>
<tr>
<td>party</td>
<td>18</td>
<td>0/1/0/X/0/13/4/</td>
</tr>
<tr>
<td>children</td>
<td>13</td>
<td>1/0/0/X/9/3/0/</td>
</tr>
<tr>
<td>fight</td>
<td>12</td>
<td>0/0/0/X/0/0/12/</td>
</tr>
<tr>
<td>fun</td>
<td>9</td>
<td>0/0/0/X/0/9/0/</td>
</tr>
<tr>
<td>things</td>
<td>8</td>
<td>0/0/0/X/7/1/0/</td>
</tr>
<tr>
<td>game</td>
<td>8</td>
<td>3/0/0/X/2/2/1/</td>
</tr>
<tr>
<td>view</td>
<td>8</td>
<td>0/0/0/X/1/6/1/</td>
</tr>
<tr>
<td>child</td>
<td>8</td>
<td>0/0/0/X/5/3/0/</td>
</tr>
<tr>
<td>brat</td>
<td>6</td>
<td>0/0/0/X/6/0/0/</td>
</tr>
<tr>
<td>record</td>
<td>5</td>
<td>0/0/0/X/0/4/1/</td>
</tr>
<tr>
<td>case</td>
<td>5</td>
<td>0/0/0/X/0/4/1/</td>
</tr>
<tr>
<td>people</td>
<td>5</td>
<td>1/1/0/X/2/0/1/</td>
</tr>
<tr>
<td>surprise</td>
<td>4</td>
<td>0/0/0/X/0/3/1/</td>
</tr>
</tbody>
</table>

The resulting ordered list reveals typical lexical fillings of the complements which accompany the keyword. "Significance" is expressed as a numeric value which is generated by statistical computations (Lemnitzer, 1997). The resulting structures are known as "collocations" in the lexicological literature and form an important part of our information programme. This automatic preprocessing of the data is the first stage of the lexicographic process. With these sources – concordance, co-occurrence patterns, secondary lexicographic and lexicological descriptions – at hand, we fill the relevant lexical information into the following scheme which is common to all three dictionaries:

- **<LEMMA>**
  - key word
- **<LESART>**
  - reading (a separate information structure will be given for each reading)
- **<MUSTER>**
  - subcategorization frame
- **<AUXILIAR>**
  - perfect auxiliary for each reading
- **<BEISP>**</BEISP>
  - example(s) of use
- **<KOLL>**
  - collocation(s)
- **<ANM>**
  - comments

### 3.3 Development of three monolingual dictionaries

According to this scheme we get, for instance, the entry for the English "spoil"

```
***
<LEMMA> spoil
<LESART> spoil sb/sth
<MUSTER> subj # dir0
<AUXILIAR> have
```
Any substantial quantity of moisture forced out of the cheese when it is subjected to heat may [spoil] the consistency of a dish, making it watery.

I didn't want to [spoil] Christmas.

You don't think I want to [spoil] that memory by smashing about with you?

He [spoiled] twice as much fruit as he ate by crushing the berries beneath his clumsy feet in a mad rush from bush to bush.

His voice was shrill, and that [spoiled] his speech sometimes, when he would get excited, and would raise it at the end.

the Greek "εξετάζω"

or the German "arbeiten"

<LESART> and <MUSTER> give informations concerning the subcategorization frame of a reading of a certain verb. The subcategorization frame is given both in a formal way (e.g. NP(nom)/Subj # für(acc)/PP for the verb arbeiten) and in an informal manner (e.g. jmd
arbeitet für jmld/etw). As the informal description shows, at least a minimal information of the semantic type of a complement is given: [+hum], that is jmld, and [-hum], that is etw. In many cases, this information is more precise, taking into account further restrictions to the semantic type of complement (e.g. Kleidungsstück for the Direct Object of one reading of the verb anziehen). Additionally the syntactic reading informs about the obligatoriness of each complement: complements which can be left out in the surface structure of a sentence are put in brackets. The subcategorization frame determines the syntactic functional (subject, direct object) as well as the categorial (NP(nom), NP(acc)) features as can be seen in the German example above. However, the set of descriptors varies in the different languages ¹. For English and Standard Modern Greek we can obviously not apply a case distinction to the complements. Each entry is divided into different readings according to differences in subcategorization frame and meaning distinction (polysemy).

According to the character of the dictionary, valency variation is the primary criterion for the stipulation of readings. Polysemy is only a secondary criterion. The elements of the subcategorization patterns are separated by the number sign (#) and ordered in line with the linear order of the elements in a typical sentence. Non-obligatory complements are enclosed in round brackets, alternative instances of a complement are separated by a slash (/). A fixed vocabulary of functional and categorial descriptors is used for the formal description of the complements. After the semantic description examples are given, which demonstrate one of the most important aspects of the valency description. These examples could be used as a guide to the translation of certain verbs and constructions.

With the scheme illustrated above, the entries for 3.000-4.000 highly frequent verbal lexical entries in German and English have already been described in full detail. Standard Modern Greek will follow.

3.4 Linking of the three monolingual sources
The resulting three monolingual valency dictionaries will be networked by a <UEBERS> / <Equivalent> feature which is still to be added to the scheme. It is to be expected that the source of the link will be a certain reading of an entry, whereas the target of the link will be the whole entry in the target language.

3.5 Implementation of a lexical database
In the second stage of the project, the last step of the course of development, the results of the lexicographic description will be presented in a suitable format. One format for data exchange will be HTML. The data will also be stored in a lexical database.

¹ See Fischer (1997)
4. Perspectives and open issues
There are at least two systematic problems which are not yet treated conclusively:

4.1 Position of the translation equivalents: The translation equivalents are addressed to a certain reading of the source language verb. They can be read as recommendations for the production of a target language sentence. However, we observed that the translation equivalents given to not fit the probable translation of one or more of the example sentences given. Often the translations are semantic equivalents, and this depends on the context. This problem is presented in the following example:

"LEMMA" work
...
"LESART" 16. Work sth # "UEBERS" wirken, bedienen, arbeiten lassen
"BEISP"
In order to clarify the procedure we |work| the 'general case' first and only then a couple of concrete examples. (German translation: durchnehmen)
And he did not |work| many miracles. (German translation: bewirken)
Not only did they not know how to |work| the dogs, but they did not know how to work themselves. (German translation: umgehen mit / behandeln)
"BEISP"

A native speaker of German would translate the English "work" in the first example with "durchnehmen", in the second one with "bewirken" and in the last one with "umgehen mit" or "behandeln" rather with one of the given equivalents under "UEBERS". Obviously, the concept of a reading is already an idealization which could contrast to the authentic use of the verb in the given reading. We therefore intend to mark "particular" target language equivalents directly with the example. However, we are not sure whether this is an approach which will be honoured by the user. Further investigation in the relation between "reading" and "example of usage" is necessary.

4.2 Lexical rules: Some features of verbs, e.g. passivization, are regular to large groups of verbs and should therefore be described by lexical rules. Consequently, the dictionaries should contain a set of lexical rules. The individual entries should be linked with these rules. There will be language specific and language independent lexical rules. The format of these rules is still to be determined.
References

Wolf Paprotté, Arbeitsbereich Linguistik, Westfälische Wilhelms-Universität Münster, Hüfferstr. 27, D-48149 Münster
e-mail: wolfp@marley.uni-muenster.de

Panagiota Balanga, Arbeitsbereich Linguistik, Westfälische Wilhelms-Universität Münster, Hüfferstr. 27, D-48149 Münster
e-mail: balanga@marley.uni-muenster.de

Lothar Lemnitzer, ZERES GmbH, Universitätsstr. 142, D-44799 Bochum
e-mail: Lothar.Lemnitzer@zeres.de