DERIVATION OF UNDERLYING VALENCY FRAMES FROM A LEARNER'S DICTIONARY

ALEXANDR ROSEN, EVA HAJIČOVÁ and JAN HAJIČ

Universita Karlova
Praha, Czechoslovakia

ABSTRACT

The authors collect lexical data for a module of English syntactic analysis in the context of a bilingual research project. The computer usable version of OALD (Hornby, 1974) is used as the primary source. The main focus is on the structure and derivation of valency frames for verbal entries in the target lexicon. Illustration of the complex relation between OALD's verb subcategorization codes and the target complementation paradigms is provided, and an approach to the derivation procedure design suggested.

1. INTRODUCTION

The present paper describes a part of a larger project, which should result in the extraction of lexical and structural correspondences between grammatical units in large parallel English and Czech texts. The correspondences will then be used to build a transfer module for an English-to-Czech (and possibly Czech-to-English) machine translation system. Final as well as partial results should also be useful as source data for text-oriented linguistic research, both bi- and monolingual[1].

This task entails the need for tools to analyse unrestricted Czech and English texts. In the first stage of the project the goal is to produce Czech and English lexicons of adequate coverage and implemented analysis grammars, which will later be augmented with tools for preliminary disambiguation. The parser will build annotated dependency structures, usable for tagging word forms, clauses and sentences with morphological and syntactic information. The lexicon and grammars, enriched by feedback from the parsed texts, can later be used within the machine translation system proper.

At present, the primary source of lexical data for the English analysis is a machine readable dictionary, preprocessed to contain only relevant information in a transparent format. This paper focusses on how valency frames for verbal entries are extracted from subcategorization codes in the machine readable dictionary.

2. THE CHOICES

Even though the correspondences between parallel text units can be established at an arbitrary level starting from word forms up to an elaborate logical representation, the practical solution seems to lie somewhere in between. The approach we have chosen is based on the representation of linguistic analysis in terms of *underlying (tectogrammatical)* structures, which are determined by the given language, but void of various irregularities of the surface strings, including the ambiguity of morphemic and surface syntactic units.[2] A "deeper" analysis would increase the risk of errors and introduce more theoretical bias while a very shallow level would require larger amounts of data to arrive at simple facts when parallel text units are compared.

The (underlying) syntactic description is dependency-based (with coordination and apposition as relations of a different type) and the project described here makes it possible (i) to test the basic assumptions of the theory on a large data collection, and (ii) to formulate an implementable relation between the surface string and the underlying representation.

A constrained-based (unification) formalism was selected due to its declarativeness, conciseness and formal rigour, but its other interesting properties were also appreciated: i.a., the important role of the lexicon and the need to treat surface facts within the same rigorous framework as deeper concepts.[3]

3. THE SOURCE

As a shortcut towards a lexicon of reasonable coverage we decided to build upon an available machine readable dictionary, which we intend to augment later by hand and from other sources. Our primary source of English lexical data is now *CUVOALD*, or the *Expanded Computer Usable Version of the Oxford Advanced Learner's Dictionary of Current English, 3rd edition* (*OALD*, Hornby, 1974), which is available from *Oxford Text Archive* (see Mitton, 1986)[4]. *CUVOALD* lists all headwords, headword variants and derivatives with simple codes denoting word classes and inflection patterns, supplemented by verb pattern codes for verbs. Sense distinctions from *OALD* are not retained.

Whereas the derivation of lexical information as needed by the analysis from *CUVOALD* word class codes is relatively straightforward, the *OALD* verb pattern codes, which are crucial for our purpose, present a real challenge. The dictionary classifies verbs according to the number and form of complements into 51 "verb patterns", marked by numbers 1-25, supplemented in some cases by letters (4A,4B,4C,4D,4F). The number of verbs in a single pattern is quite variable: starting from a single item in [VP4F] for *be* followed by an infinitive up to 4855 standard transitive verbs in [VP6A]. A pattern groups together verbs which exhibit the same behaviour in a standard context and are subject to the same set of transformations under specified conditions. So e.g. the class of intransitive verbs [VP2A] can take introductory *there* and postpone the subject if it is indefinite and "heavy": *There comes a time when we feel we must make a protest.* A single pattern is also used for verbs which allow the same morphosyntactic variations of a complement. ([VP1]: *She's dark/in good health/here/a pretty girl.*) A different verb pattern is, however, used if only a subset of the relevant class permits the variation. ([VP6C]: *She enjoys swimming / *to swim.* vs. [VP6D]: *She likes swimming / to swim.*) Some variations may be treated as a different verb pattern. (This is the case of the above example: *She likes swimming.* [VP6D] and *She likes to swim.* [VP7A])

Akkerman (1989) lists several shortcomings of the *OALD* verb patterns. As Sampson (1990) noted, some of them are arguable. For our purpose, the most problematic seems to be the treatment of compound verbs (with the resulting loss of information in *CUVOALD*) and too surface-level definition of some verb patterns. These classes are quite a heterogenous collection: by [VP14] are marked verbs in all of the following uses, the only requirement being that the verb is followed by a noun and a prepositional phrase:

They accused him of stealing the book.
I explained my difficulty to him.
Compare the copy with the original.

Another "misbehaved" pattern is VP4A where, depending on the verb, the infinitive can be complement or adjunct:

The swimmer failed to reach the shore.
He came to see that he was mistaken.
She stood up to see better.

Apart from these "systemic" blemishes we expect a number of other inconsistencies and errors to appear during the process of derivation and use of the target lexicon.

4. THE TARGET

The target lexicon contains the following information about the valency of a verb (or its complementation), grouped in an entry as a complementation paradigm:

SUBCATEGORIZATION LIST (SC) gives syntactic and morphological categories for every dependent, i.e. either a *participant* (complement, may be obligatory or optional) or an *obligatory free modification* (obligatory adjunct). An item in the list is in fact an underspecified representation of the corresponding dependent. The ordering of items in the list corresponds to the unmarked word order in a declarative sentence.

SYNTACTIC FRAME (SF), a feature structure with syntactic functions as attributes; values of these attributes are co-indexed with the corresponding items of the *subcategorization list*.

UNDERLYING STRUCTURE (US), a feature structure with tectogrammatical functions as attributes; values of these attributes are identical with *underlying structures* within the corresponding items of *subcategorization list* and *syntactic frame*. The value of the attribute GOV (*governor*) is identical with the value of the *lexeme* attribute of the verb's feature structure.

The analysis will establish index links between saturated frame slots and their fillers in the analysis tree. This will provide easy access to the analysis results at the three levels of description, highlighting the structure of the sentential core.

The following simple example gives complementation paradigm for an intransitive verb. N[*nom*] is shorthand for a feature structure representing noun in the nominative case with saturated subcategorization requirements; the numbers co-index feature structures which are shared as values of some attributes, the small index selects only a part of the structure, namely the nominal equivalent of the *underlying structure*; the attribute GOV gives the lexical value of the verb while ACT stands for *actor/bearer*, the function representing subject of an active verb at the underlying level. Angle brackets enclose lists, square brackets (conjuctions of) feature structures, curly brackets disjunctions. Commas separate members of conjuction, vertical bars members of disjunction.

```
SC < [1] N[nom]2 > ,
SF [ SUBJ [1] ] ,
US [ GOV sleep , ACT [2] ]
```

The same could be expressed in a PATR-like style (Shieber (1986)):

```
< SC first > = N[nom]
< SC rest > = end
< SF SUBJ > = < SC first >
< US GOV > = sleep
< US ACT > = < SC first US >
```

Next, we give two possible complementation paradigms for a transitive verb. (PAT stands for *patient*, V[*prespart*,SC<N$_3$>]$_4$ is abbreviation for present participle form of a verb whose single valency slot for subject in the SC list is co-indexed with the actor/bearer of the matrix verb):

```
SC < [1] N[nom]3 , [2] N[acc]4 > ,
SF [ SUBJ [1] , OBJ [2] ] ,
US [ GOV enjoy , ACT [3] , PAT [4] ]

SC < [1] N[nom]3 ,
     [2] V[prespart,SC<N3>]4 > ,
SF [ SUBJ [1] , OBJ [2] ] ,
US [ GOV enjoy , ACT [3] , PAT [4] ]
```

As the value of the attribute PAT of *enjoy* is shared with the value of the attribute US of the object, the correct value for the dependent verb's ACT attribute is supplied via co-indexing of the subject of *enjoy* with the subject of the non-finite clause within the SC list of *enjoy*:

```
US [ GOV enjoy , ACT [3] ,
     PAT [ GOV swim , ACT [3] ] ]
```

The complementation paradigm, rather than being stated within full-fledged feature structures, is expressed in terms of *templates*, preferrably allowing defaults and multiple inheritance. Accordingly, the above two paradigms will be expressed as follows:

transitive
transitive , 2ing , equi

Two verbal entries can be related by a lexical rule with the effect that one of these two entries need not be explicitly present (the other should then be marked by the rule's name). This will solve phenomena such as *there* preposing, dative alternation, and passivization.

The collection of three "levels" of description within a single complementation paradigm provides a means to express rather subtle differences. Let us take as an example four superficially identical constructions:

(a) *I wanted him to see the monster.*
(b) *I expected him to see the monster.*
(c) *I elected him to see the monster.*
(d) *I told him to see the monster.*

Following Quirk et al. (1985, p.1216), the verb is monotransitive in (a), complex-transitive in (b) and (c), and ditransitive in (d). The example (b) is closer to the monotransitive type while (c) is closer to the ditransitive type.

If we have the *subcategorization list*

SC < [1] N[*nom*]$_4$, [2] N[*acc*]$_5$,
     [3] V[*inf*,SC<N$_5$>]$_6$ >

to express the superficial identity of all the four cases, we can assume the above verbs to have the following *syntactic frames*:

(a) SF [SUBJ[1], OBJ[3]]
(b) SF [SUBJ[1], OBJ[2], OBJCOMPL[3]]
(c) SF [SUBJ[1], OBJ[2], OBJCOMPL[3]]
(d) SF [SUBJ[1], OBJ[3], OBJ2[2]]

The difference between the types (a) and (b) vs. (c) and (d) is that between the Raising and Equi types. Therefore, (b) will have only two participants at the level of *underlying structure* while (c) will have three:

(a) US [ACT [4], PAT [6]]
(b) US [ACT [4], PAT [6]]
(c) US [ACT [4], PAT [5], EFF [6]]
(d) US [ACT [4], PAT [6], ADDR [5]]

The respective templates will be:

(a) *transitive, 3inf, raising*
(b) *complex-transitive, 3inf, raising*
(c) *complex-transitive, 3inf, equi*
(d) *ditransitive, 3inf, equi*

A problem remains how to derive such information from *OALD*'s verb patterns.

5. THE DERIVATION

CUVOALD was not primarily intended for use with a syntactic parser, so a few modifications were necessary. First, the pronunciation field was deleted and homograph entries with different pronunciations merged. (In *CUVOALD*, each word, or word form, has only one entry, unless it has two different pronunciations.) Second, entries headed by regular forms within irregular paradigms as headwords were also deleted. And finally, reference to base forms was provided in entries of all the remaining nonbase (irregular) forms. Base forms of irregular paradigms were marked by a code specifying the paradigm type. After that, we tried to find a way how to derive the complementation paradigms.

Ideally, templates of the sort described in Section 4 should correspond to OALD verb patterns while lexical rules would account for structures listed in Hornby (1975) as variants of the same verb pattern. Although this idea works in the case of the most frequent patterns ([VP2A], [VP6A]), there are many patterns where the relation between pattern and paradigm can be 1:n, n:1, or even n:n (n > 1) (see Section 3).

The case of *n* patterns : *1* paradigm reduces the number of paradigms and as such is a welcome situation. The case of *1:n* can mean (i) ambiguity for all verbs listed under the pattern (and can possibly be accounted for by lexical rules), (ii) the possibility to subdivide the verbs of this class into *n* subclasses, or (iii) a combination of the two. For (i), the derivation of complementation paradigm from a verb pattern will yield a disjunction. For (ii), verbs with different complementation paradigms should be distinguished. Boguraev and Briscoe (1989) used valency codes in LDOCE (*Longman Dictionary of Contemporary English*) to automatically extract the (explicitly unmarked) distinction between Equi and Raising verbs. Similar approach can be used to make this and other distinctions in *OALD* by taking into account co-occurences of verb patterns. Our situation is simpler in that we, as yet, make no attempt to treat distinct word senses, and more difficult in that the blurred sense distinctions can have negative effect on any derivation procedure. It remains to be seen whether such a method will lead to results of sufficient reliability. However, at the same time we have to supply more information to some classes of verbs, for which any possibility of automatic treatment is exluded. The current efforts include the specification of lexical values of particles and prepositions for compound verbs and assigning verbs marked by verb pattern codes such as VP14 to relevant subclasses.

The correspondences between the *OALD* patterns and complementation paradigms are stated in the simple cases by rules relating one or more patterns to one or more paradigms - templates. Where possible, frequently co-occurring verb patterns are collapsed into a single paradigm with local disjuction, e.g. [VP6D] and [VP7A] for *like (swimming / to swim)* give the following template: [*transitive*, { *2ing* ¦ *2inf* }, *equi*], which expands into:

```
SC < [1] N[nom]3 ,
     [2] V[{prespart¦inf},SC<N3>]4 > ,
SF [ SUBJ [1] , OBJ [2] ] ,
US [ GOV like , ACT [3] , PAT [4] ]
```

Now there are two possible strategies representing two extremes. The first strategy disregards the actual distribution of verb patterns in the dictionary and attempts to combine results of rule application into a compact and meaningful complementation paradigm. The second strategy starts from a list of all combinations of verb patterns within the dictionary and assigns a rule to every combination. Let us look how the first approach works.

The process of derivation of a complementation paradigm for a verb entry consists of the following steps:

1. Application of rules rewriting a verb pattern code (or more verb pattern codes if the resulting paradigms can be related by a lexical rule) by a template or a sequence of templates connected by logical operators "and" and "or", the result may be marked by one or more lexical rule names. Rules rewriting more patterns are preferred to those rewriting fewer patterns. A rule may be supplemented by a condition stipulating the presence or absence of other paradigms within the same entry. A rule whose condition is satisfied is preferred to a rule without condition. Verbs with patterns which do not correspond to a single complementation paradigm while co-occurring verb patterns do not indicate a preference for one paradigm or the other have to be treated manually.

2. Simplification of the sequence of templates by making all disjunctions as local as possible.

3. Consistency check performed by expansion of the sequence of templates into feature structures.

E.g.: *believe* 3A 6A 9 10 25

step 1:

```
rules applied:
3A  ->  transitive, prepositional
6A  ->  transitive, 2n
9   ->  transitive, 2cls, 2that
10  ->  transitive, 2cls, 2wh-
25  ->  complex_transitive, 3inf,
        raising
    / ^ { 12A¦12B¦12C¦13A¦13B } a)
```

```
after application of the rules:
{ transitive, 2prep ¦  b)
  transitive, 2n ¦
  transitive, 2cls, 2that ¦
  transitive, 2cls, 2wh- ¦
  complex_transitive, 3inf, raising }
```

after step 2:

```
{ transitive, { 2prep ¦
                2n ¦
                2cls, {2that¦2wh-} } ¦
  complex_transitive, 3inf, raising }
```

after step 3:

US [ACT[3], PAT[4]],
SF [SUBJ[1], OBJ[2]],
SC < [1] N[*nom*]$_3$,
[2] { N[{*prep*| *acc*}] |
V[*cls*,{*that*| *wh*}] }$_4$ >

US [ACT[4], PAT[6]],
SF [SUBJ[1], OBJ[2], OBJCOMPL[3]],
SC < [1] N[*nom*]$_4$, [2] N[*acc*]$_5$,
[3] V[*inf*,SC<N$_5$>]$_6$ >

a) This is a condition stipulating that neither of the patterns should be present; the character ^ stands for negation.
b) This is the template of a prepositional verb. The lexical value of the preposition should be supplied.

This looks like a principled solution, but step 1 can be a source of unforeseen complexities with the result that too many entries will have to be handled manually. The second strategy is much safer: if there are not too many different combinations of verb patterns it might not be too difficult to state rewriting rules for all of them, thus eliminating steps 2 and 3 from the above procedure. However, to make a decision, some statistical analysis is necessary.

CUVOALD lists 5695 verbs with 633 different combinations of verb patterns.[5] 4853 verbs (85.2%) are marked by one of the 56 most frequent combinations (each occurring seven and more times). The first ten most frequent combinations are given below:

verb patterns	frequency
6A	1971
2A,6A	575
2A	338
6A,14	331
2A,2C	165
2A,3A	137
6A,15B	101
2A,2C,6A	100
2A,2C,6A,15B	81
3A	64

At the other end, there are 442 combinations occurring only once, 191 two and more times, 119 three and more times and 77 five and more times.

Another survey was aimed at finding most frequent combinations as proper subsets of the full combinations treated above. E.g. the combination of three patterns 2A,3A,6A occurs alone in 54 entries, but as a proper subset of a larger combination already in 566 entries.

From the above data it seems that a compromise between the treatment of individual verb patterns and of entire combinations would be most efficient. 119 combinations can already be treated by individual rules quite comfortably while the rest can be composed from results of rules applied independently, where more alert supervision is required. It also seems feasible to use the rules for combinations to treat parts of the remaining lists of verb patterns, and perhaps add a few more, selected according to the second statistics.

6. PERSPECTIVES

Lexicon and grammar together form the basis for the extraction of lexical and structural correspondences. Other tools are necessary, however, and we are currently designing specifications for such tools.

Besides the non-trivial task of text cleanup, for which no special tools will be used, two major needs remain: text unit alignment and data extraction methods.

Automatic text unit alignment (on word, phrase, and sentence level) is also non-trivial. On the sentence level, we will employ a method for alignment based on sentence length (Gale 1991), for which we have developed a flexible front-end for recognizing sentence boundaries. We are considering an extension of Church's algorithm taking into account lexicon-based elementary word correspondences (as in Kay (1988) and Catizone et al. (1991)) for better accuracy, but this extension has not been implemented yet.

Methods for data extraction are still under development. However, it is clear what such data should look like. As our output representation is far from the interlingua ideal, the data will basically be transfer data in a form fitting the structural transfer model, following the ideas of Kaplan et al. (1989). The actual implementation, however, will follow the pattern of the transfer module in the experimental machine translation system ELU (Russel et al. (1991)).

NOTES

[1] This project, called *MATRACE* (from MAchine TRAnslation between Czech and English, is one of the projects carried out within the IBM Academic Initiative in Czechoslovakia.

[2] It is not the aim of this paper to discuss and substantiate the repertoire of valency relations and their classification. The interested reader can find a detailed analysis of these issues and a comparison with other theories of deep (underlying) structure in Sgall, Hajičová and Panevová (1986, esp. Ch.2).

[3] As we are involved in the development of a *practical* constraint-based system, we are aware of the necessity to include some control or *dynamic* information in addition to the *static* description supported by traditional constraint-based formalisms. We expect to deal with this issue seriously in later stages of the project, when partial results will be available.

[4] *CUVOALD* comes in two versions: one lists base forms plus all forms of irregular words while the other contains all inflected forms explicitly. As we intend to have a morphological component, we are using the base forms version.

[5] These and following numbers include base forms only, as well as 876 verbs which were not marked by any pattern and for which defaults were used: 6A for transitive verbs, 2A for intransitive verbs.

REFERENCES

Akkerman, E. (1989) "An independent analysis of the LDOCE grammar coding system", in B. Boguraev and T. Briscoe (eds.) *Computational Lexicography for Natural Language Processing*, Longman, London and New York

Boguraev, B. and T. Briscoe (1989) "Utilising the LDOCE grammar codes", *ibidem*

Catizone, R., Russell, G. and S. Warwick (1991) "Deriving Translation Data from Bilingual Texts", in Zernik (ed.) *Lexical Acquisition: Using on-line Resources to Build a Lexicon*, Lawrence Erlbaum.

Gale, W. A. and K. W. Church (in prep.) *A Program for Aligning Sentences in Bilingual Corpora*, submitted to CL (1991).

Hornby, A.S. (1974) *Oxford Advanced Learner's Dictionary of Current English*, 3rd edition, Oxford University Press, London.

Hornby, A.S. (1975) *Guide to Patterns and Usage in English*, 2nd Edition, Oxford University Press, London.

Kaplan, R. M., K. Netter, J. Wedekind and A. Zaenen (1989) "Translation by Structural Correspondences", in *Proceedings of the 4th EACL*, ACL, Manchester, UK.

Kay, M. and M. Röscheisen (1988) *Text-Translation Alignment*, unpublished manuscript, Xerox Palo Alto Research Center.

Mitton, R. (1986) "A partial dictionary of English in Computer-Usable Form", in *Literary and Linguistic Computing* 1:214-215.

Quirk, R., S. Greenbaum, G. Leech, J. Svartvik (1985) *A Comprehensive Grammar of the English Language*, Longman, London and New York.

Russell, G., A. Ballim, D. Estival and S. Warwick-Armstrong (1991) "A Language for the Statement of Binary Relations over Feature Structures", in *Proceedings of the 5th EACL*, ACL, Berlin, Germany.

Sampson, G. (1990) a review of B. Boguraev and Ted Briscoe (eds.) Computational Lexicography for Natural Language Processing, in *Computational Linguistics* 2:113-116

Sgall, P., E. Hajičová and J. Panevová (1986) *The Meaning of the Sentence in its Semantic and Pragmatic Aspects* (Edited by J. Mey), Reidel, Dordrecht / Academia, Praha.

Shieber, S.M. (1986) *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes No. 4, Stanford: Center for the Study of Language and Information.