SURFACE AND DEEP CASES

JARMILA PANEVOVÁ
Institute of Formal and Applied Linguistics
Charles University
Prague, Czechoslovakia

HANA SKOUMALOVÁ
Institute of Theoretical and Computational Linguistics
Charles University
Prague, Czechoslovakia

Abstract

In this paper we show the relation between the "surface (morphological) cases" and "deep cases" (participants), and the possible way to automate the creation of a syntactic dictionary provided with frames containing information about deep cases and their morphemic counterparts of particular lexical items (Czech verbs).

Introduction

In the project **MATRACE**[1] (MAchine TRAnslation between Czech and English) the first aim is to create two parallel text corpora (Czech and English), morphologically and syntactically tagged. Then it will be possible to use these corpora not only for creating an MT system but also for other linguistic research, needed e.g. for systems of NL understanding. For these purposes we try to make the syntactic representation "broader" so that the further work would be easier.

In the syntactic representation of a sentence, based on dependency grammar, we will specify not only the dependency and syntactic roles of the modifications but also their underlying counterparts (i.e. "deep cases"). For this sort of tagging we need a dictionary with morphological and syntactic information, which consists of morphological paradigms of single words and their valency frames containing both syntactic and underlying roles of their members. As there is no such dictionary in machine-readable form we have to create it. Unfortunately we even cannot extract the words with their frames from an existing corpus as we are only creating it. What we have is a morphological dictionary, which is to be enriched by the syntactic information. The linguist adding this information should enter the surface frame and specify its underlying counterpart. We try to help him/her by automating the choice of the appropriate correspondence between "surface" and "deep" cases.

In this paper we will concentrate on the problems of verb and its valency slots. The generalization of our method for nouns and adjectives will not be difficult as in many cases the syntactic frame of these words is just derived from the corresponding verb.

[1] Project MATRACE, a research project of the Institute of Applied and Formal Linguistics and the Institute of Theoretical and Computational Linguistics, is carried out within the IBM Academic Initiative project in Czechoslovakia.

Theoretical background

Using the framework of the functional generative description (FGP, see Sgall et al. 1986), slightly simplified for the purpose of this paper, we distinguish two levels: a level of underlying structure (US, with the participants or "deep cases") and a level of surface structure (SS, morphemic units as parts of this are used here). As for the modifications of verbs we distinguish inner participants and free modifications (see Panevová 1974-5). This can be understood as the paradigmatical classification of all possible verbal modifications. The other dimension of their classification (combinatoric or syntagmatic dimension) concerns their obligatoriness and optionality with the particular lexical item within the verbal frame. The verbal frame contains slots for obligatory and optional inner participants (which will be filled by the labels for "deep cases" and corresponding morphemic forms) and obligatory free modifications. The difference between an obligatory and optional participant is important for a parser, however, we will leave this dichotomy aside in this contribution.

The following operational criteria for distinguishing between inner participants and free modifications are used: If the verbal modification can occur only once with a single verb token and if the governing verbs for a particular modification may be listed, the modification is considered as an "inner participant". There are five participants: Actor, Objective, Addressee, Origin and Effect. The other modifications (Time, Locative, Direction, Aim, Reason, Instrument, Regard, Manner etc.) can reoccur with a single verb token and may modify any verb. With some verbs free modifications can also enter the respective verb frame: either the construction is ungrammatical without them (*to behave HOW, to last HOW LONG, to live WHERE* etc.) or they are semantically obligatory, although they can be omitted on the SS level. This can be tested by a dialogue of the following type:

A. My friend came.
B. Where?
A. *I don't know.

Unacceptability of the answer "I don't know" indicates that the modification *where* is a part of a verbal frame of the verb *to come*.

According to the theory proposed by Panevová (1974-5, esp. § 5) the following consequences are accepted here: If a verb has only one inner participant then this participant is Actor. If a verb has two participants then these are Actor and Objective. As for the 1st and 2nd participant our approach is similar to Tesnière's (1959). However, if three or even more slots of a verbal frame are occupied then semantic considerations are involved. This is different from Tesnière's solution and does not fully coincide with Fillmore's proposals (Fillmore 1968, 1970).

Determining the Addressee, Origin and Effect is rather difficult and requires taking into account the combination of surface cases in the frame (including the form of the Objective), the animacy of single members of the frame etc. Though there is no one-to-one mapping between "deep cases" and "surface cases", we are able to discover certain regularities and provide some generalization reflected in an algorithm.

Observation

In inflectional languages with (morphological) cases it is apparent that some cases are typical for certain participants. Objective is typically realized

as the Accusative and Addressee as the Dative case. In Czech there are other typical (prepositional) cases. Thus *z+Genitive (out of sb,st)* or *od+Genitive (from sb,st)* are typical for Origin, *na+Accusative (at st)*, *do+Genitive (to st)* or *v+Accusative (into sb,st)* are typical for Effect etc. This well known fact led us to the idea of creating a program as a tool for introducing verbal frames (to be used even by researchers without deep linguistic training) based on correspondences between surface and deep cases. At first we sorted the Czech verbs into four groups:

1. Verbs without Nominative in their frames.

Examples:
prší
[*(it) rains*]
hučí mi(Act(Dat)) v hlavě
[*(it) is buzzing to me in head*]
(my head is buzzing)

This group contains verbs with empty frames but also a few verbs with very untypical frames. If the frame contains only one participant, then this is obviously an Actor. If there are at least two participants in the frame and one of them is Dative, then this is the Actor. If, beside this, only one more participant occurs in the frame, it is necessarily the Objective. All other verbs must be treated individually by a linguist as a kind of exception.

2. Verbs with Nominative and at most one more inner participant.

Examples:
on(Act(Nom)) zemřel
[*he died*]
Jan(Act(Nom)) vidí Marii(Obj(Acc))
[*John sees Mary*]
ze semene(Obj(Prep(z)+Gen)) vyrostl strom(Act(Nom))
[*from a seed grew a tree*]
to(Obj(Nom)) se mi(Act(Dat)) líbí
[*it to me appeals*] (I like it)

According to the theory, if the frame contains only one participant, it is Actor, if it contains two participants, one of them is Actor and the other is Objective. Nominative usually represents the Actor but there is an exception to this rule: if the other participant is in Dative, then this participant is the Actor and the Nominative represents the Objective. Reasonability of this exception can be proved by translating particular verbs into other languages, in which the surface frames are different while there is no obvious reason why the deep frames should differ. Thus e.g. the verb *líbit se* has Nominative/Clause and Dative in its surface frame while in the frame of the corresponding English verb *to like* there are Subject and Object/Clause, where Subject corresponds to Czech Dative and Object to Nominative.

3. Verbs with Nominative and two or more other inner participants, which occur only in "typical" cases (i.e. Accusative, Dative, z+Genitive, od+Genitive, na+Accusative, do+Accusative, v+Accusative). A verb belongs to this group even if some of the slots for inner participants can be occupied either by a typical case or any other (prepositional) case or a clause or infinitive.

Examples:
Jan(Act(Nom)) dal Marii(Addr(Dat)) knihu(Obj(Acc))
[*John gave Mary a book*]
Otec(Act(Nom)) udělal dětem(Addr(Dat)) ze dřeva(Orig(Prep(z)+Gen)) panáčka(Obj(Acc))
[*father made to children out of wood a puppet*]

The verbs of the third group behave "typically", which means that Nominative represents the Actor, Accusative the Objective, Dative the Addressee etc.

4. Other, i.e. verbs with Nominative and two or more other

inner participants, which occur not only in typical cases.

Examples:
šéf(Act(Nom)) jmenoval Jana(Obj(Acc)) zástupcem(Eff(Instr))
[*boss appointed John a deputy*]
Jan(Act(Nom)) obklopil Marii(Addr(Acc)) péčí(Obj(Instr))
[*John surrounded Mary with care*]

In this group Nominative always represents Actor but for determining other participants it is necessary to take into account an additional aspect, namely the prototypical character of the animacy of the participants; this enables us to distinguish the difference between deep frames of the two last examples *jmenovat* and *obklopit*. The surface frames are identical: Nominative, Accusative and Instrumental, but while the verb *jmenovat* has Accusative standing for the Objective and Instrumental for the Effect, the verb *obklopit* has Accusative standing for the function of Addressee and Instrumental for the function of Objective.

Algorithmization

The algorithms for the verbs of the first two groups were described in the previous paragraph.

The possible algorithmization of determining the correspondences between "surface" and "deep" cases of the verbs of the last two groups can be seen from the following table of several Czech verbs with different frames:

	Pat	Addr	Orig	Eff	
udělat	Acc		z+Gen		make
vzít	Acc	(Dat)	(od+Gen)		take
dostat	Acc		od+Gen		get
požadovat	Acc/Cl		(od+Gen)		ask (for)
měnit	Acc	(Dat)	na+Acc		change
zaplatit	Acc /za+Acc	Dat			pay
dědit	Acc		(po+Loc)		inherit
vyprávět	Acc/Cl	(Dat)		o+Loc	talk
vědět	Acc/Cl			o+Loc	know
spojit	s+Instr	Acc			connect
blahopřát	k+Dat/Cl	Dat			congratulate
obklopit	Instr	Acc			surround
stát se	Instr		z+Gen		become
jmenovat	Acc			Instr	appoint
žádat	o+Acc	Acc			ask (for)
hovořit	o+Loc	(s+Instr)			speak
pomáhat	s+Instr /INF	Dat			help
ptát se	na+Acc /Cl	Acc			ask
říkat	o+Acc	Dat			ask (for)
vsadit se	o+Acc	s+Instr			bet

We can see that the prepositional cases "typical" for Origin occur only in the position of Origin, and Dative occurs only in the position of Addressee. After these members of the surface frame are determined, in most cases only one undetermined participant remains, which must be Objective. If two or three participants are remaining we have to take into account the animacy

(typical for Addressee) and inanimacy of the participants and the set of prepositional cases which are typical for Effect.

This algorithm is used in a program which reads Czech verbs from an input file and asks a linguist (in the interactive regime) to fill in the surface verbal frame.

Conclusions

Some general linguistic statements concerning relations between "centre" (prototypes) and "periphery" (marginality) in the domain of verb and its valency could be inferred from an application of the rules presented in our paper. In "nominative" languages the verbal frame **Act Obj Addr** can be considered as central (while e.g. **Act (Obj) Addr** is not typical). Moreover, the correspondences between US and SS as **Act -> Nom, Obj -> Acc, Addr -> Dat** can be treated as prototypes (while e.g. correspondences **Act -> Dat, Addr -> Acc, Obj -> Instr** occur in Czech as marginal). The strategy of our algorithm is based principally on an observation of this type. We assume that this method can be easily adapted for any other inflectional language and perhaps also for such languages as English. Languages may differ as to correspondences between a particular deep case (US) and its surface (morphemic form), but the idea of prototypical and marginal relations seems to be valid and is supported by the algorithmic procedure for determining these correspondences.

References:

Fillmore, Ch. (1968): The Case for Case, In: Universals of Linguistic Theory (ed. E. Bach, T. Haims), New York, pp. 1-88.

Fillmore, Ch. (1970): Subjects, Speakers and Roles. Synthese, Vol. 21, pp. 251-274.

Panevová, J. (1974-5): On verbal Frames in Functional Generative Description, Part I, Prague Bulletin of Mathematical Linguistics, Vol. 22, 1974, pp. 3-40, Part II, ibid, Vol. 23, 1975, pp. 17-37.

Sgall, P. - Hajičová, E. - Panevová, J. (1986): The Meaning of the Sentence in Its Semantic and Pragmatic Aspects, Prague - Dordrecht.

Tesnière, L. (1959): Eléments de syntaxe structurale, Paris.