Ambiguity, language structures and corpora

Vladimír Petkevič

Dans LA LINGUISTIQUE 2014/2 (Vol. 50), pages 63 à 82
Éditions Presses Universitaires de France

ISSN 0075-966X
ISBN 9782130629009
DOI 10.3917/ling.502.0063

Article disponible en ligne à l'adresse
https://www.cairn.info/revue-la-linguistique-2014-2-page-63.htm

Découvrir le sommaire de ce numéro, suivre la revue par email, s'abonner...
Flashez ce QR Code pour accéder à la page de ce numéro sur Cairn.info.
AMBIGUITY, LANGUAGE STRUCTURES AND CORPORA

par Vladimír Petkevič
Charles University, Faculty of Arts
Vladimir.Petkevic@ff.cuni.cz

L'article présente une méthode de désambiguïsation (résolution d’homonymie morphologique) automatique des textes de la langue tchèque par des règles linguistiques et sa portée sur la théorie de la structure de la phrase tchèque. L’auteur donne quelques exemples du domaine du traitement automatique de la langue, plus précisément de la désambiguïsation automatique et de l’analyse syntaxique de la phrase tchèque par des règles morphosyntaxiques et d’autres règles ayant trait par exemple à des particules réfléctives. L’auteur montre que la nécessité de trouver une solution pour la désambiguïsation conduit à la spécification de nouveaux problèmes concernant le sous-système (morpho-)syntaxique de la langue tchèque. Ce sont des problèmes que le structuralisme de l’entre-deux-guerres n’a pas abordés, parce qu’il n’avait pas les moyens suffisants de le faire en mathématique et en informatique. À vrai dire, il n’en avait pas besoin. L’auteur démontre à quel point la (morpho-)syntaxe de la phrase tchèque est intéressante, lorsqu’elle est envisagée du point de vue de la recherche strictement formelle de structures linguistiques fournies par l’utilisation de corpus linguistiques extensifs. Il est clair que ce point de vue sur la structure de la phrase tchèque peut aussi apporter beaucoup d’éléments nouveaux dans le domaine de la recherche théorique, en particulier sur la phrase tchèque.

INTRODUCTION

In recent decades, computer technologies gave rise to considerable progress in natural language processing, which manifested itself, apart from other things, also in developing extensive language corpora and a new methodology of linguistic research: corpus methodology. Today, every language spoken by a relatively high number of speakers has its corpus or even more corpora (synchronic and diachronic
ones, corpora containing texts of various genres, etc.). These corpora\(^1\) are usually lemmatized and morphologically and syntactically (or even semantically) annotated, i.e. individual word forms in a text or groups of such word forms are assigned their lemmas and morphological properties (including part-of-speech), and also syntactic and other information in the form of so-called tags: for example, the part-of-speech ambiguous French word form *devant* is – on the basis of the context in which the word occurs – assigned a tag for either a preposition (*Il s’est assis *devant* moi*. English: *He sat down in front of me.*), or a participle of the verb *devoir* (*... l’homme *devant* à son ami plus de 500 francs... E. the man *owing* his friend more than 500 francs...*).

For many languages of the world it is immensely difficult to annotate ambiguous word forms in huge corpora, because such corpora comprising hundreds of millions of word occurrences cannot be annotated manually, but only automatically by specialized software tools. There exist four main methods of automatic annotation/tagging (i.e. disambiguation, solving the problem of ambiguity) in texts of large corpora:

(a) statistical methods based on machine learning;

(b) rule-based methods consisting in the development of rules automatically inferred from textual data also on the basis of machine learning;

(c) rule-based methods consisting in computerized rules manually developed by linguists who use corpus data, their knowledge of *langue* and methods of introspection;

(d) cooperation of the abovementioned methods, mainly the methods denoted (a) and (c).

In the sequel, we shall deal exclusively with the (c) method, i.e. with rules capturing the system of language, more precisely, its syntactic and other structure patterns, and we shall demonstrate this method on Czech, an inflectional language that is morphologically and syntactically very complex. Our main objective will be to show that the necessity of formulating mathematically exact disambiguation rules which assign

Ambiguity, language structures and corpora

proper tags to corpus texts leads to the emergence of new questions and problems which were not formulated by the structuralists of the Prague school. It is because they could not anticipate that in the 1940s and 1950s computer science will emerge and that Chomskyan revolution will introduce formal, mathematical view to the study of language, and even its computer processing (mathematical and logical principles of computer science emerge in the 1930s). The formal approach to language gave rise to many linguistic (and also mathematical and computer science) problems whose solutions substantially amplify our knowledge of language and lead also to the revisions of the then existing ideas about language: the linguists in the pre-corpus epoch neither had at their disposal necessary empirical data in sufficient amount, nor necessary software tools (they had to be satisfied with card files).

AMBIGUITY AND DISAMBIGUATION

Many problems presented by natural language processing are very intricate – they cannot be solved with 100% accuracy. This primarily concerns those levels of language description where large number of elements are interrelated in complex and potentially limitless relations: syntax and semantics. There are many problems which have not been solved in a fully satisfactory way yet: comprehensive exact description of syntax of clauses and phrases based on an appropriate theory and on large amounts of language data, in semantics it is the formalization of language meaning, let alone, e.g., pragmatics. In the sphere of linguistic applications that have a lot of shortcomings due to the complexity of algorithmic descriptions of natural languages there belong, apart from other things: a grammar checker of a given language, morphological disambiguation and syntactic analysis of corpus texts, word sense disambiguation, machine translation, programs for automatic understanding of language and other. In all these automated systems it is the problems of ambiguity that must be solved, i.e. the cases where one form has, due to Karcevskij’s asymmetric
dualism, more functions: on all levels of linguistic description. The solutions to theoretical and application problems require a new perspective of the language structure that necessitates using very extensive language data being available now and a new kind of thinking about language.

We shall deal with some problems of automatic morphological and (morpho)syntactic disambiguation of Czech and we will demonstrate – on several cases of ambiguity – our main claim: we want to show how a solution to the problem of disambiguation as one of the main problems of natural language processing leads – by formulating new questions and finding answers to them – to the amplification and extension of our knowledge of the system of language. We shall be concerned with the langue of Czech, its (morpho)syntactic subsystem in particular, some of whose regularities will be identified by virtue of solving disambiguation problems. We ask, for instance, the following seemingly simple questions:

- how to automatically disambiguate the word form se (reflexive particle or preposition)?
- how to associate a particular occurrence of a reflexive particle se, si (se, E. approximately accusative reflexive; si, E. approximately dative reflexive) with its “base” word?
- do transitive reflexive only verbs exist?

Disambiguation of the word form se (see Oliva, 2003: 299 – 314)

The reflexive word form se is the second most frequent part-of-speech ambiguous word form in contemporary Czech (the most frequent word is the conjunction / particle a, E. and). The word form se is either:
A. the vocalized form of the preposition s (E. with or from) requiring the instrumental or genitive case. This vocalized form can stand immediately only in front of the words beginning with a sibilant s, z (optionally in front of words beginning with s a ž) or in front of words beginning with a complex consonant cluster not containing (very roughly speaking) neither a liquid r, nor l.
B. (reflexive) particle / reflexive pronoun.
Ambiguity, language structures and corpora

For example in sentence:
(1) Byl tam se prep starým Adj instr přítel N instr 
  He was there with old friend
  He was there with his old friend.

the word form se is a preposition, because the following word starým is in the instrumental case and it starts with the sibilant s. However, in sentence:
(2) Lékaři rozhodli, že se refl musím Verb vrátit. 
  Doctors decided that I must return.

The doctors decided that I had to return.

the word form se is a reflexive.

The word se as a non-preposition has several meanings in Czech and their description can be further refined (Skoumalová: 2001):

(B1) The word se is a (reflexive) particle with no independent meaning, it is only a tag coforming reflexive only verb, reflexive only adjectives and reflexive deverbal nouns. For instance, the form se has this meaning in sentence:
(3) Snažil V-Refl jsem V-Aux se refl part to V pochopit. 
  Tried I-am it understand.
  I tried to understand it.

In (3), the verb form snažil (E. tried) is a form of the reflexive only verb snažit, se being its particle.

(B2) The word se is a reflexive pronoun denoting a reflexive object in accusative coreferential with the subject of a given clause or with the subject of an infinitive in the clause:
(4) Petr se refl umyl prep řece. 
  Peter himself washed in river.

(B3) The word form se is a reflexive particle associated with verbs, active deverbal adjectives (those derived from transgressives, we shall refer to them as deverbal adjectives in the sequel) and deverbal nouns. These verbs, adjectives and nouns are not reflexive only verbs, adjectives and nouns, respectively, i.e. the verbs and adjectives can be used in a non-reflexive meaning,
mainly the transitive one. If the particle *se* is present, the respective verb or adjective is used in a non-transitional meaning, i.e. with no object as its complement, and *se* is related to the subject of the given clause or to the subject of an infinitive, cf. sentence:

(5a) *Svou otázkou novinář rozčilil toho politika.*

  With his question journalist annoyed that politician.

where the verb *rozčilil, E. annoy* is used transitively, i.e. with the direct object *politika (E. politician)*, whereas in sentence:

(5b) *Ten politik se refl rozčilil.*

  That politician became annoyed.

it is used intransitively, with the reflexive *se* relating to the subject *politik (E. politician).*

(6a) *Vrátil jsem mu knihu.*

  I-returned to-him book.

(6b) *Vrátil jsem se refl domů.*

  I-returned home.

The situation is similar in sentences (6a) and (6b): in (6a), the verb *vrátil (E. returned)* is used transitively, i.e. with the direct object *knihu (E. book)*, whereas in (6b) the verb is used intransitively, *se* being coreferential with the pro-dropped subject.

(B4) The word *se* can be used to coform a passive construction – with the 3rd pers. active voice of a transitive finite verb (in (7) the verb is *připravují, E. prepare*), e.g.:

(7) *Tady se pass připravujíV3rd-pl agentiN-nom-Sb*

  Here pass they-prepare agents.

  The agents are being prepared here.

In these constructions, the subject and the verbal predicate can be in singular or plural.
Ambiguity, language structures and corpora

(B5) The word *se* can be used to coform a deagentive construction – with the 3rd pers. sg. of a finite verb (in (8) the verb is *pracuje*, E. *works*), e.g.:

(8)  
\[ Tady \; se_{\text{deagent}} \; \text{pracuje}_{\text{V-3rd-sg}} \; do \; šesti \; hodin. \]  

Here one *works* till six o’clock.  
Office hours are till six o’clock here. / One works till six o’clock here.

In these constructions a subject is missing and the finite verb can neither be in plural, nor in the 1st or 2nd person.

(B6) The word *se* expresses a modality of disposition, usually with an evaluative adverb/adverbial:

(9)  
\[ V \; té \; troubě \; se_{\text{refl}} \; \text{pečě}_{\text{V-3rd-sg}} \; dobře_{\text{Adv}} \]  

In that oven one roasts *well.*  
That oven is good for roasting.

Similarly as in (B5) above, in these constructions, a subject is missing and the finite verb cannot be in plural, nor in 1st or 2nd person.

(B7) The word *se* expresses reciprocity of the verbal meaning:

(10)  
\[ Martin \; a \; Jana \; se_{\text{refl-recip}} \; \text{mají}_{\text{V}} \; rádi_{\text{Adj}}. \]  

Martin and Jane each other love.  
Martin and Jane love each other.

In (10), reciprocal *se* is used with the multi-word verb *mají rádi* (E. *love*). The sentence (10) is, in fact, ambiguous: (a) either *Martin* and *Jana* love each other, or (b) *Martin* loves himself and *Jana* loves herself (the haploglogized word *se* with an object meaning refers to respective subjects *Martin* and *Jana*).

Let us specify several facts (primarily morphosyntactic ones) about the word *se* from the viewpoint of a coarse-grained disambiguation: we shall endeavour to distinguish *se* in the:

(a) prepositional meaning
(b) non-prepositional meaning, i.e. as a (reflexive) particle / reflexive pronoun.

These facts must be formulated in a very exact manner, so that they could be implemented as computer programs (some
of the facts were extracted from Oliva’s article (2003) and Kučera’s article (1984) and adequately reinterpreted).

**Phonetic fact Ph1.** Generally: if a preposition s has a vocalized form se besides its non-vocalized form s, there can be three alternatives: in front of an immediately following word

- either the preposition obligatorily vocalizes,
- or the preposition never vocalizes
- or the preposition can vocalize but need not, both options being possible.

In Czech, if the preposition s stands in front of a word neither beginning with a sibilant s, z, š, ž, nor with a complex consonant cluster (both Oliva and Kučera specify the sets of such clusters in their articles), it does not vocalize. Then it means that the word form se is not a vocalized preposition, i.e. it is a (reflexive) particle / reflexive pronoun.

**Morphosyntactic fact M1.** Let the word form se be followed, in a clause, by a group of words, none of which is in the genitive or instrumental case. Furthermore, let this group be followed by a verbal form or by a word in a case different from the genitive and instrumental (i.e. nominative, dative, accusative, locative in Czech) or by a punctuation mark. Then the word se is a (reflexive) particle / reflexive pronoun.

**Morphosyntactic fact M2.** Let a clause/sentence contain only one occurrence of the word se and let there be, in the same clause/sentence, only one reflexive only verb or reflexive only adjective. Then the word se is not a preposition: it is a reflexive particle (associated with this verb or adjective).

**Morphosyntactic fact M3.** Let the word se immediately follow the first reflexive only verb form in a clause/sentence. Then the word se is not a preposition: it is a reflexive particle.

It is relatively easy to identify the first verb in a sentence by an automatic morphological analysis (this task is much more difficult in a clause because first the clause must be identified as such!) in case the verbal form in question is part-of-speech unambiguous (the problems concern part-of-speech ambiguous verb forms).

**Morphosyntactic fact M4.** The (reflexive) particle / reflexive pronoun se stands on the syntactically second position in
a clause, i.e. on Wackernagel’s position. This well-known fact can be trivially used as follows: reflexive \textit{se} cannot be the first word in a clause. If the fact M4 is to be used in a non-trivial way, the problem of the identification of the first syntactic position should be solved: it is necessary to exactly specify which words in the clause belong to the phrase that takes up the first syntactic position in the clause – a very difficult problem a (not only) rule-based automatic morphological disambiguation and syntactic analysis must solve. Precise rules must be specified that identify a nominal phrase, i.e. at least the set of words belonging to it. For instance, in sentence:

(11a) \textit{Ten muž svou manželku se prep třemi dětmi opustil.}
\textit{That man his wife with three children abandoned.}
\textit{That man abandoned his wife leaving her with three children.}

there are two syntactic positions occupied with two nominal phrases: (a) \textit{ten muž (E. that man)} and (b) \textit{svou manželku (E. his wife)}. It follows from the fact M4 that \textit{se} cannot be a (reflexive) particle / reflexive pronoun, i.e. it is a preposition. However, if the word \textit{se} occupies the second syntactic position in a clause, it can be a (reflexive) particle / reflexive pronoun as in sentence

(11b) \textit{Ten muž se refl part manželku se prep třemi dětmi snažil opustit.}
\textit{That man wife se prep with three children tried to abandon.}
\textit{That man tried to abandon his wife leaving her with three children.}

\textbf{Morphosyntactic fact M5.} Let there be only one verbal form \textit{V} in a sentence/clause and let there be neither a deverbal adjective, nor a verbal noun which can be reflexivized. Let it be possible for \textit{V} to reflexivize. Let there be the word form \textit{se} with an already disambiguated non-prepositional meaning. Then \textit{se} associates as a (reflexive) particle / reflexive pronoun with this verb \textit{V}. Moreover, if there are some other occurrences of the word form \textit{se}, they are vocalized prepositions.
The following example will elucidate this fact:

(12) **Rozběhla se** refl **zcela opačným směrem k pánovi**

*She-rushed reflexive entirely in-opposite direction to a man with a dog.*

In sentence (12) **se** refl is a reflexive associated with the verbal form **rozběhla** (E. *rushed*), the second occurrence of **se**, i.e. **se** prep, is a vocalized preposition because there is no word in the sentence this **se** can associate with as a (reflexive) particle / reflexive pronoun. The fact M5 is based on the following morphosyntactic observations:

(a) one “base word” (verb, adjective, noun) can associate with one non-prepositional occurrence of **se** at most;

(b) an occurrence of a non-prepositional **se** is associated with only one “base word” (verb, adjective, noun) with the notable exception of haplology where an occurrence of a non-prepositional **se** can be associated with two base words:

(13) **Snažil se** refl hapl **smějící dívku odvézt.**

*He-tried laughing girl to take away.*

Here we see that **se** refl hapl associates with two base words belonging to different parts of speech (!):

- the verbal form **snažil** (E. *tried*)
- the deverbal adjective **smějící** (E. *laughing*)

i.e. it is a clear case of haplology where base words even belong to different parts of speech.

**Morphosyntactic fact M6.** The non-prepositional word form **se** is an enclitic, and therefore it is placed in a clitic cluster in the Wackernagel’s position (the cluster may be trivial, i.e. composed of the word **se** only). This fact can be used for the disambiguation, too (Oliva, 2003: 308–309).

We have shown several exact morphosyntactic facts which can be formally expressed by rules that can be computationally...
implemented and thus their correctness can be strictly verified. Moreover, a detailed corpus inspection showed the following distribution of the part-of-speech ambiguous word *se*:

- *se* as a preposition: 5–6 %
- *se* as a (reflexive) particle / reflexive pronoun: 94–95 %

In the current rule-based system of the Czech morphological (including part-of-speech) disambiguation there are 30 (!) special rules dealing with this special kind of disambiguation; their accuracy is better than 99.9%. However, no finer disambiguation of the (reflexive) particle / reflexive pronoun *se* in our disambiguation of Czech is available yet.

The abovementioned considerations are based on the following valuable factors (1)–(4) concerning the morphosyntactic system of Czech which were not available for the Praguian structuralists in the heyday of the Prague school between the wars:

(1) Precise analysis of phonetic conditions under which the preposition *se*

- obligatorily vocalizes
- never vocalizes
- optionally vocalizes.

This analysis based on corpora of synchronic Czech should result is the list of corresponding consonant clusters in front of which the preposition *se* always vocalizes, or never vocalizes, or can vocalize. Such lists were prepared by K. Oliva (2003: 302) and K. Kučera (1984: 225–237) and subsequently implemented in disambiguation rules.

(2) The availability of the complete (or at least the most exhaustive) list of reflexive only verbs of the *se* reflexive type (the list contains, e.g., the verb *snažit se*, *E. to try*, and *ošívat se*, *E. to fidget*) (Bartoň, Cvrček, Čermák, Jelínek, Petkević, 2009: 175nn.), the complete list of reflexive only deverbal adjectives of the *se* reflexive type (the list contains, e.g., the adjectives *vyskytující se*, *E. occurring*) and also their counterparts for the *si* reflexive type (*si* is, roughly said, the dative reflexive counterpart of *se* ambiguous with the colloquial form of the 2nd pers. sg. pres. of the verb *být*, *E. to be*). If similar lists were prepared in the past, they could have hardly been exhaustive.
(3) Lists of verbs and deverbal adjectives that never reflexivize (e.g. the verb *souviset*, E. *be associated*; e.g. the adjective *souviset*ějící, E. *associated*).

(4) Precise analysis of the structure of a syntactic phrase, the nominal one in particular, which takes up one syntactic position, and thus its delimitation as a structure of a set of word forms, i.e. its internal and external left and right boundary and the material within the boundaries, the (non)contiguous property of such a phrase in the sentence, etc.

All the abovementioned results concerning syntactic structure of Czech were motivated by the necessity to automatically disambiguate one ambiguous word form, i.e. *se*, in Czech sentence by precisely formulated and implementable rules. The case in point is – with respect to the entire task of a full-fledged rule-base disambiguation – a very tiny, but important fraction of the whole syntactic system of Czech: there are very many individual disambiguation types, in fact, the disambiguation of every ambiguous part-of-speech word form and otherwise morphologically ambiguous word form gives rise to specific problems to be solved. The structuralists may have considered such problems at their time but they did not and could not analyze language data in such a complex way as it is possible nowadays.

The types of reflexives specified above in (B1) – (B7) were investigated by Skoumalová (2001: 21–59) in a detailed way, but she was not concerned with the criteria for an automatic identification of a particular reflexive type of the reflexive word *se* and *si* (we do not know of any study of the kind, nor is there any study elaborated within automatic morphological disambiguation of Czech). However, every such study will, invariably, lead to the emergence of new and invaluable paradigms of the corresponding types of verbs: reciprocals (*koulovat se*, E. *be snowballing*), transitive (mainly causative) verbs in intransitive reflexive use (e.g. transitive *posadit* vs. reflexive *posadit se*, E. *seat someone vs. sit down*; transitive *přehlédnout* vs. reflexive *přehlédnout se*, E. *glance over vs. fail to see*, or transitive *splést* vs. reflexivní *splést se*, E. *braid vs. make a mistake*), where reflexive variants can differ from their transitive counterparts in meaning. Such studies will also have to deal with properties of individual types of reflexive verbs or adjectives, they will not only consist in preparing lists of verbs having
the given property. For instance, an automatic identification of the reflexive se having the function of passivization and of deagentivization has a fundamental importance both for syntactic theory of Czech and for many applications, such as syntactic analysis, machine translation, etc.

Reflexive particles se, si and their base word

Another interesting morphosyntactic problem associated with automatic morphological disambiguation and syntactic analysis is the relation of the reflexive se, si in Czech (but also in other languages in which there exist reflexive particles as free morphemes, not just bound ones: French, German, Slovak, Polish, but not, e.g., Russian). The problem consists in finding, for a given reflexive, its base word (verb, adjective, noun) in case there are more candidates in a sentence. Let us suppose, for the sake of simplicity, that reflexive words se (E. approximately accusative reflexive) and si (E. approximately dative reflexive) have already been correctly disambiguated as reflexives in the process of disambiguation. The problem of the identification of the base word for a given reflexive se or si can be illustrated by a compound sentence (14):

(14) Mistrovský tým se\textsubscript{refl1} poté, co se\textsubscript{refl2} přípravující\textsubscript{3} Crack team itself after the being-prepared

se\textsubscript{refl3} výluka v NHL stala\textsubscript{2} nevyhnutelnou, zpočátku lockout in NHL became inevitable, at-first

hledal\textsubscript{1}, narychlo se\textsubscript{refl4} tvořila\textsubscript{4} optimální looked-for, hastily was being formed optimum

sestava, ale klíčovým problémem se\textsubscript{refl5} stalo, line-up, but key problem became

zranění gólmana Lašáka.

injury of the goalkeeper Lašák.

After the lockout in NHL being prepared became inevitable, the crack team tried to make sense of its play and an optimum line-up was being formed, but the key problem became an injury of the goalkeeper Lašák.
In sentence (14), the indices indicate the identification of five occurrences of the reflexive *se* with their base words. We will simplify the whole problem of this identification by not considering haplology. We will apply the theory of formal languages developed by N. Chomsky and others to Czech, especially his hierarchy of formal languages, grammars and automata. In Czech (and similarly in many other languages), the following two statements, which appear to be morphosyntactic universals, hold:

St1. The base word as well as its reflexive belong to the same clause in a compound sentence. (Obvious statement)

St2. The parts of discontinuous clauses in a compound sentence form a context-free structure.

The statement St2 means, in other words, that parts of different clauses are – with respect to the word order – properly embedded, i.e. if a clause C1 is split into two or more parts and at least one part of another clause C2 is placed between some two parts C1a and C1b of C1, then all the parts of C2 are placed between C1a and C1b. This means that C2 is entirely embedded between C1 and C2, e.g:

```
Main1 a Sub1 a Sub2 Sub1 b Main1 b
```

where *Main1 a* and *Main1 b* denote two discontinuous parts of the main clause *Main1*, because both parts of the *Sub1* clause, i.e. *Sub1 a* and *Sub1 b*, are embedded in the *Main1* clause. Moreover, there is another clause, *Sub2*, embedded in *Sub1* and thus splitting *Sub1* into two discontinuous parts *Sub1 a* and *Sub2 b*. In accordance with this labeling, we can specify the structure of sentence (14) as follows:

```
Main1 a Sub1 a Main1 b Main2 Main3
```

where each of the five subparts is constituted by the following elements, respectively:

**Main1 a:**

- *Mistrovský tým*
- *se refl1*
- *poté*

**Sub1 a:**

- *co se refl2 připravující*
- *se refl3*
- *výluka v NHL stala*
- *nevyhnutelnou*

**Main2:**

- *the being-prepared*

**Main3:**

- *lockout in NHL became inevitable,*
Ambiguity, language structures and corpora

Main1.b:

zpočátku hledal
at-first looked-for,

Main2:

narycho se refl₁ tvořila optimální sestava
hastily was being formed optimum line-up,

Main3:

ale klíčovým problémem se refl₃ stav₃ zranění gólmanka
but key problem became injury of the goalkeeper

Lasáka.

The pattern of sentence (14) clearly shows that sentence (14) has a context-free structure where four occurrences of the reflexive se, viz. se refl₁ se refl₂ se refl₄ se refl₅, are associated with verbs, and reflexive se refl₃ is associated with the deverbal adjective připravující (E. being-prepared).

The following structure would violate proper embedding:

Main₁a Sub₁a Main₁b Sub₁b...

Here neither Sub₁ is entirely embedded in Main₁, nor Main₁ is entirely embedded in Sub₁, thus the structure is not context-free.

On the basis of statement st2 we can state, for instance, that se refl₁ is associated with the verb form hledal₁. Or vice versa: as the reflexives se refl₂, se refl₃, se refl₄ and se refl₅ are associated, roughly speaking, with the closest verb or adjective, se refl₁ must be assigned to the only remaining candidate of reflexivization, i.e. the verbal form hledal₁. It follows from this, in accordance with st1, that both Main₁a and Main₁b are really parts of the clause Main₁. Thus, reflexives help us to analyze the structure of compound sentences.

Let us now study the mutual position of the reflexive se and si and its base word (we will limit our investigation to reflexive se only, because the reflexive si behaves similarly in this respect).

A. Verb as the base word: If the verb precedes “its” se, it precedes it either immediately, or there is, between the two words, only the clitic conjunction li (E. if), conditional verbal clitic by
(E. would) in all of its inflected forms or a clitic particle from a very small set (e.g. už, E. already). If the verb follows its reflexive se, it can stand in an arbitrary distance from this reflexive in the same clause, but – as our corpora show – the closer the reflexive se is to the verb, the more numerous such structures are. We can even state that there is a certain regularity in the decrease in frequency with an increasing distance between the two words. As the reflexive se is a clitic, it stands, in the clitic cluster, on the second syntactic position in a clause.

B. Deverbal adjective as the base word: If the adjective precedes its se, it precedes it either immediately, or there is, between the two words, only the clitic conjunction li (E. if), or a clitic particle from a very small set (e.g. už, E. already). If the adjective follows its reflexive se, it stands – as real usage being reflected in large corpora of Czech shows – immediately following se, or there is, between the two words, only one adverb/particle or a very short prepositional phrase formed by a preposition and a form of a personal pronoun. Theoretically, however, the reflexive can stand arbitrarily far from its adjective in a sentence. Thus, adjectives tend to be more closely associated with their reflexives than verbs. Generally, the position of an occurrence of the reflexive se immediately following its adjective is an unmarked one, i.e. the reflexive typically occupies the second syntactic position in the corresponding adjective phrase. The reflexive can stand in front of an adjective only if it does not occupy the first position in the adjective phrase, cf.:

(15a)* Všechny ty rozesmáté tváře, si_{refl} radostně
_all the laughing faces, cheerfully
hrající děti, přátelské dospělí, všechno mu
playing kids, friendly adults, all to-him
to připadalo přehnané.

it seemed exaggerated.

All the laughing faces, cheerfully playing kids, friendly adults, all of it seemed to him as exaggerated.

(15b) Všechny ty rozesmáté tváře, radostně si_{refl} hrající děti, přátelské dospělí, všechno mu to připadalo přehnané.

The gloss is the same as with (15a) above.
The two sentences differ only in the mutual word order of the words _si_ refl1 and _radostně_ (E. _cheerfully_). In the ungrammatical sentence (15a) the reflexive _si_ refl1 occupies the inadmissible first position in the adjective phrase _si radostně hrající_ (E. _cheerfully playing_); in the grammatical sentence (15b) the reflexive _si_ refl1 takes up the second position in the adjective phrase _radostně si hrající_ (E. _cheerfully playing_). We see that the Wackernagel’s rule about the syntactically second position holds not only for a clause headed by a finite verb, but also for an adjective phrase headed by an adjective.

C. Deverbal noun as the base word. A deverbal noun is most closely associated with its reflexive _se_. _Se_ always immediately follows the noun.

Let us summarize: The reflexive _se_ stands in a syntactically second position in a clause / adjective phrase. Moreover, the “more nominal” the base word is, the closer “its _se_” stands.

Thus, on the basis of corpus data we can identify interesting relations between a reflexive _se_ and its base word. These relations can help us to identify also some other properties in the system of Czech syntax.

*Transitive reflexive only verbs*

Another interesting problem associated with reflexivity is a relation of reflexive verbs (and the corresponding deverbal adjectives) and a direct object in accusative case. First, it is necessary to study functions of accusative and their properties. A noun in accusative expresses the following surface-syntactic functions (for instance Jelínek, 2012):

(a) object of a transitive verb or of a deverbal adjective:

(16) _Tam jsem našel_ V-trans _svůj_ plač _Obj-Acc_.

There I found _my regiment_.

(b) temporal adverbial:

(17) _Tam jsem nějaký čas_ Adv-temp-Acc _bydlel_ v _hotelu._

I have been living there _in a hotel for some time_.

(c) adverbial of measure:

(18) Byli už jen kilometr od ranče.

They were already only one kilometer away from the ranch.

Furthermore, the accusative case has the following properties:

(d) Syntactic noun in accusative case is a head noun in a prepositional phrase, governed by a preposition with accusative valency:

(19) Naříbet ruky si nasypala hromádku prášku.

On the back of the hand she sprinkled a pile of dust.

(e) Syntactic noun in accusative is part of a collocation expressing the adverbial of manner:

(20) Procházeli se ruč v ruce až k mlýnu.

They hand in hand as far as the mill.

In (20) the collocation is in bold type.

It is extremely difficult to identify functions of a certain case in textual data: syntactic rules that are to be developed are intricate: even the first step, i.e. the identification of a particular case occurrence (here accusative), is generally difficult because of case syncretism in Czech, let alone its function. For instance, one of the main problems is to identify a syntactic noun as a direct object of a verb/adjective if there are more such verbs in a sentence/clause. However, these difficult tasks and their (at least) partial solution lead to invaluable pieces of knowledge about syntactic structure of Czech which older structuralism could not have. In the course of the development of appropriate disambiguation rules, a linguist will identify – in connection with the functions of a noun in accusative –
several important facts based on corpus research. Apart from other things, the linguist will:
(a) make an exhaustive list of adverbials of time and measure headed by nouns in the accusative case
(b) specify sophisticated rules stating whether a syntactic noun is a head of an (accusative) prepositional phrase, or not
(c) make an exhaustive list of phrasemes formed by adverbials in accusative case
(d) make an exhaustive list of intransitive verbs and deverbal adjectives derived from transgressives of these verbs
(e) find out to what extent the presence of a reflexive *se* blocks the presence of an accusative object in a clause and in this connection he will make an exhaustive list of the reflexive only verbs of the *se* type that are also transitive. The short list of such verbs extracted from the corpora of synchronic Czech (2000–2010) follows:

\[\begin{align*}
& \text{domnívat se, E. } \text{suppose, presume} \\
& \text{domýšlet se, E. } \text{imagine} \\
& \text{doslechnout se, E. } \text{hear} \\
& \text{doučit se, E. } \text{catch on one’s schoolwork} \\
& \text{odnaučit se, E. } \text{unlearn, give up} \\
& \text{naučit se, E. } \text{learn} \\
& \text{učit se, E. } \text{learn, study} \\
& \text{do(z)vědět se, E. } \text{get to know} \\
& \text{dovtípit se, E. } \text{guess} \\
& \text{(po)modlit se, E. } \text{pray} \\
& \text{(na)biflovat se E. } \text{bone up} \\
& \text{(na)šprtat se E. } \text{swot up}
\end{align*}\]

The invaluable paradigms such as this one and corresponding syntactic structures can – in their entirety – be achieved only by introspection substantially supported by data and modern tools of corpus linguistics.

---

Conclusion

On several examples from Czech morphology and syntax, the author demonstrated in this paper how a strictly formal approach to the study of language structure can extend and deepen our knowledge of the system of natural language, especially its syntactic subsystem. Generally, this approach is made possible by the existence of natural language corpora containing extensive morphologically and syntactically annotated textual data that allow a linguist to identify new facts about language, the facts that had been impossible to find for prewar structuralists.

Bibliographical references


