VP-Fronting in Czech and Polish—A Case Study in Corpus-Oriented Grammar Research *

Roland Meyer

University of Regensburg

Fronting of an infinite VP across a finite main verb—akin to German "VP-topicalization"—can be found also in Czech and Polish. The paper discusses evidence from large corpora for this process and some of its properties, both syntactic and information-structural. Based on this case, criteria for more user-friedly searching and retrieval of corpus data in syntactic research are being developed.

1 Introduction

Word order in Slavic languages is commonly claimed to be free in the sense that many permutations of the words in a sentence are acceptable, given suitable context. While this is most often demonstrated for adjuncts and argument expressions, it holds equally for verbs and verb phrases, cf.[1]

(1) *(Bez względu na to, jaki był cel podjątej*
without regard on that which was aim undertaken-GEN
wyprawy,) prowadzić to będzie do przedłużania
expedition-GEN lead that FUT-AUX to extension
bałkańskiego dramatu.
Balkan-GEN drama

* Thanks are due to the Institute of the Czech National Corpus and to the Institute of Computer Science of the Polish Academy of Sciences which kindly provided the data source for this paper. I also wish to thank Denisa Lenertová and the audiences of the Potsdam Workshop on Heterogenity in Linguistic Databases and of the 13th JungslavistInnen-Treffen for discussion, and the editors for their friendly insistence. All errors are my own.

[1] The glosses are abbreviated as follows: ACC=accusative, AUX=auxiliary, DAT=dative, FUT=future, GEN=genitive, INS=instrumental, MASC=masculine, MP=modal particle, NEG=negation, NOM=nominative, PL=plural, PT=past tense, REFL=reflexive, SBJ=subjunctive, SG=singular.

Interdisciplinary Studies on Information Structure 02 (2005): 87–115
Dipper, S., M. Götze and M. Stede (eds.):
Heterogeneity in Focus: Creating and Using Linguistic Databases

‘(Disregarding what was the goal of the expedition,) it will lead to prolongation of the Balkan drama.’ [Polish]

This type of inversion is reminiscent of the German construction commonly called “VP- topicalization”, as in

(2) a. *Sie hat nicht* [$_{VP}$ *den Peter geküsst*]
she has not the P. kissed
‘She has not kissed Peter.’
b. [$_{VP}$ *den Peter geküsst*]$_j$ *hat sie nicht* t_j
the P. kissed has she not

In both (1) and (2), a non-finite VP or V^0 head has been moved to the left, crossing over the surface position of the governing auxiliary.[2] In (1), only the verbal head has been shifted, leaving behind its directional argument PP, while the whole VP constituent has fronted in (2); but, as (3) and (4) indicate, Polish also allows fronting of a complete VP, and German also allows topicalization of only a partial VP:

(3) *(. . . leży i o Bożym świecie nie wie! . . .) — I jadł kiełbasy*
lies and about god’s world not knows and eat sausage
nie będzie, i gorzałki nie posmakuje?
not FUT-AUX and vodka not tastes
‘(he is lying around and doesn’t know about the world . . .) — He won’t eat sausage, and he won’t try vodka either?’

(4) *Geküsst hat sie den Peter nicht.*
kissed has she the P. not
‘She hasn’t kissed Peter.’ (ex.(2) and (4) after Fanselow (2004))

The two *partial* constituent fronting structures (1) and (4) pose a harder ana-

[2] In (2), the auxiliary is, of course, itself in a derived position, German being a SOV language. In Polish, I assume that the future auxiliary is generated in T^0, from where it has the option of moving up to highfer heads in the extended verbal projection (see Meyer 2004, ch. 4 for explicit argumentation). For the remainder of this paper, it is immaterial whether the auxilary itself moves or not.

lytical problem than the full VP shift illustrated in (2). On one plausible approach, the former actually involve two movement steps, namely (i) the extraction of a subconstituent of the VP—*do przedłużania bałkańskiege dramatu* and *Peter*, respectively—and (ii) the fronting of the rest of the VP, the so-called *remnant* (cf. Müller 1998). This analysis obviously presupposes that the two operations—Scrambling (or extraposition, as argued by Müller 1998) as in step (i) and fronting of a full VP as in step (ii)—exist independently in the language in question, and that their characteristic properties are shared by the partial fronting construction. Fanselow (2004) develops a different approach, in which there is no *remnant* movement. Instead, the subcategorization frames of both the auxiliary and the main verb are merged and their requirements fulfilled on the syntactic level. The moved constituent is not a full VP, but a smaller verbal projection, which does not contain a trace.

The most relevant syntactic properties of VP-fronting in German include the following (cf. Müller 1998, Fanselow 2004):

- A moved VP becomes itself an island for the extraction of one of its subconstituents
- Partial VP-fronting is only possible if the VP targets the SpecC position. Thus, (5-b) is scarcely acceptable in German.

(5) a. *dass [den Peter geküsst] keiner hatte*
that the Peter kissed nobody had
'that nobody had kissed Peter'
b. ?-*dass geküsst keiner den Peter hat

The present paper has two goals: First, it presents the relevant evidence for VP-fronting in Czech and Polish which can be gathered from two large-scale, annotated corpora, namely, the Czech National Corpus (ČNK) and the Corpus of the Institute of Computer Science of the Polish Academy of Sciences (IPI PAN).[3]

[3] See section 4 for information on these corpora. Unless mentioned otherwise, all Czech ex-

Second, it shows *how* this evidence may be accessed and discusses selected design features of these two corpora from the perspective of the user.

2 Clause Structure and Potential Landing Sites in Czech and Polish

There are a few pivotal elements, such as sentential negation, verbal elements, and clitic auxiliaries and pronouns, which can be used to delimit basic clause structure in Czech and Polish. I will briefly present the relevant evidence and a topological overview for orientation.

2.1 Czech

In Czech, sentential negation immediately precedes the main verb, the future auxiliary, and the present or future copula, while it obligatorily follows the clitic past and subjunctive auxiliary, as long as the verb stays below the latter (cf. Junghanns 1999):

(6) a. *To (*ne)jsem/bych (ne)řekl.*
this not-AUX.PT.1SG/AUX.SBJ.1SG not-said
'I didn't/wouldn't say that.'

b. *To vám (ne)budu (*ne)říkat.*
this you-DAT not-AUX-FUT not-say
'This, I won't tell you.'

The most economical way to grasp these positional restrictions is to assume a structure with fixed slots for clitics, negation, the future auxiliary, and the verb phrase, in this order:

XP?	*že* *'that'*	XP?	aux. > refl. > pron. clitics	XP*	NEG	AUX-FUT	XP*	VP

amples in the remainder are from ČNK and all Polish ones from IPI PAN.

Auxiliary and pronominal clitics[4] principally occupy the "second position" of the clause, following the first constituent. In colloquial Czech, they may also occur clause-initially. In embedded clauses, there is an optional slot between the copula and the clitics, which may be filled by an emphasized, focused or topicalized constituent.

However, this is not the only possible structure. The main verb (in the form of the so-called *l*-participle), including negation, may precede the clitics. This can only happen in matrix clauses, where there is no first constituent XP:

(7) *(Ne)řekl* *(*ne)bych* *to.*
not-said-MASC not-AUX-SBJ.1SG this
'I wouldn't say that.'

The movement of the participle differs fundamentally from German VP-topicalization in that it cannot take along any further material (8); therefore, it has been argued to involve V^0 head movement rather than phrasal movement.

(8) a. **Posílal dopisy jsem* *ti* *pravidelně každý týden.*
sent letters AUX-PT.1SG you regularly every week
(Avgustinova & Oliva 1997, 40)

b. **... že* *nedal* *by* *mu* *to.*
that not-gave SBJ him this
(Veselovská 1995, 149)

Since participle movement is so restricted in Czech, I will concentrate on the movement of *infinitival* VPs—including partial VPs—to the pre-clitic position, which *is* comparable to German VP-topicalization. However—as we will see below—there is a further landing site for this VP-movement between the clitics and the negation slot.[5] I will refer to the former as "high VP-fronting", and to the latter as "low VP-fronting" in the remainder.

[4] These include the past and subjunctive auxiliary, as well as the short forms of pronouns.

[5] Cf. (5-b) above, which shows that *partial* VP-Scrambling to the left edge of the middle field is excluded in German.

A set of examples for VP-fronting to the left of the clitics is mentioned by Avgustinova and Oliva (1997, 40), including infinitives as in (9-a) and passive participles as in (9-b):[6]

(9) a. *Posílat dopisy ti budu pravidelně každý týden.*
send letters you FUT-AUX regularly every week
'Send letters to you I shall regularly every week.'
b. *Srdečně uvítáni domorodým obyvatelstvem jsme*
cordially greeted original-INS inhabitants-INS AUX-PT.1PL
rozhodně nebyli.
certainly not-been
'For sure, we were not greeted by the original inhabitants cordially.'

2.2 Polish

There are less obvious structural markers in the Polish clause than in the Czech one. Clitic pronouns (so-called weak forms), past tense verbal person and number affixes and the subjunctive marker *by* do not obey a strict second-position requirement:

(10) *Do której kategorii pan by się zaliczył?*
in which category sir SBJ REFL counted
'Into which category would you put yourself?' (APTC)[7]

However, there are restrictions on the relative linear order among these clitic elements (cf. Witkoś 1996, 165):

(11) a. *Maria (*go*) spotkała (*go*) w środę.*
M.-NOM him met him on Wednesday-ACC

[6] The authors refer to these as "partial VP-fronting"; actually, they rather involve full VP-fronting (maybe except for the clitic raising in (9-a)). However, truly *partial* VP-fronting is also possible in Czech (see below for examples).

[7] This example stems from Adam Przepiórkowski's "Toy Corpus", an early predecessor of the IPI PAN corpus, which has been disconnected recently.

'Mary met him on Wednesday.'

b. *Maria* by *(*go*) (/*go by*) spotkała (*go) w środę.*
M.-NOM SBJ him him SBJ met him on Wednesday-ACC
'Mary would have met him on Wednesday.'

c. *Maria (*go) spotkałaby (go) w środę.*
M.-NOM him met-SBJ him on Wednesday-ACC

The pattern in (11) is commonly accounted for via two assumptions (Witkoś 1996, Błaszczak 2001): (i) the subjunctive marker is generated above the pronominals and none of them move, and (ii) the main verb may move up to the position of the subjunctive marker. The behaviour of the verbal person and number (PN-) affixes is similar, but not identical, cf. Dornisch (1998) and Błaszczak (2001):

(12) a. *My*śmy *(*go*) (/*gośmy*) widzieli (*go) wczoraj.*
we-PT.1PL him him-PT.1PL saw him yesterday
'We saw him yesterday.'

b. *My (*go*) widzieli*śmy *(*go*) wczoraj.*
we him saw-PT.1PL him yesterday
'Wir sahen ihn gestern.'

Since the past PN-marker can follow the main verb even if the clitic pronominal stays above it (12-b), it seems that it may occupy either of two positions: a high one above the pronominal clitics (but below the subjunctive marker), and a low one next to V^0.[8]

The relative order of verbal elements and negation supports this view: sentential negation follows the subjunctive and past tense PN-markers if they occur in their high position, but it precedes the main verb or the future auxiliary.[9] These considerations lead to the following topological picture of the Polish

[8] In the latter case, the verb has to stay low, because otherwise the excluded sequence [*go* ... *-śmy* ... V] would be predicted again (cf. (12-b)).

[9] If the main verb raises up to the position of the subjunctive or PN-marker as in (11-c), it takes the negation along, resulting in the order Neg+V+*go*.

clause:

że *'that'*	XP*	(verb+) PN- / SBJ-marker	pron. clitics	XP*	NEG	FUT-AUX	XP*	VP (+ PN-marker)

Note that—exactly as in Czech (cf. (8))—no VP, complete or partial, may raise to the slot of the first XP* in this schema:

(13) *[*Poszli do szkoły*]$_k$ *-śmy* t_k
went to school PT.1PL
(Bański 2001, 185)

To be sure, a detailed corpus search using the query in (14-a) yielded at least one example of this kind:[10]

(14) a. `[pos=praet] [] "by" [pos=aglt] within s`
b. *. . . , pochował ja bym go tak, żeby go i na sąd*
hidden I SBJ-1SG him so that him also on court
ostateczny nie znaleźli.
last not found-3PL
'. . . I would hide him so that he would not even be found on judgment day.' (Sienkiewicz 1895, IPI PAN)

While (14-b) may be doubtful (coming from an earlier stage of modern Polish, colloquial in style), there is no problem in the standard language with an XP and the verb raising independently, as in

(15) *Co radziłbyś bliskiemu sobie młodemu człowiekowi,*
what advise-SBJ-2SG close-DAT REFL-DAT young-DAT person-DAT
aby zrobił po ukończeniu szkoły?
that did after finish school
'What would you advise a young person close to you to do after finishing school?'

[10] The above query would read "sequence of a past tense form, an arbitrary token, *by*, and a PN-marker within one clause" in natural language (cf. Przepiórkowski 2004). The syntax is obviously similar to the one of the CQP query language (Christ, 1994).

I conclude that participles generally raise as heads, not as VPs, in modern Polish (like in Czech), but the motivation for their movement has to be completely independent of the requirement to support the clitics (other than in Czech). Relevant constructions for the purposes of this paper mainly include infinitival VPs raised to the position before the future auxiliary ("low VP-fronting" in the remainder) or to the left of the preverbal PN- and subjunctive marker ("high VP-fronting").

3 Results of the Corpus Query

In this section, I will show some results of a corpus-oriented investigation of infinitival VP-fronting constructions in Czech and Polish, based on the Czech National Corpus (Český Národní Korpus, ČNK), and the IPI PAN corpus of Polish (IPI PAN, Przepiórkowski 2004). Both corpora have been lemmatized and annotated for morphosyntactic categories using a stochastic tagger. Nevertheless, there are important differences in the design of the annotation (see section 4). I use corpus evidence in a purely qualitative manner here, as an indication of what constructions can be found with some basic frequency in the two corpora, and what contexts they occur in. Needless to say, something which is not in a corpus, however large it may be, can still be part of the language and its grammar. But we can at least challenge restrictive intuitive judgments by counterevidence from the corpus, or support an intuitive restriction by the lack of the latter. Given that the VP-fronting data are very context-dependent and not always easy to judge for informants, this is already of some help.

3.1 Czech

3.1.1 High VP-fronting

High VP-fronting of an infinitive in Czech may easily be searched in the ČNK using an expression like[11]

(16)
```
[tag="Vf."] [tag!="Z.*" & word!="a"]*
"((js[iemt][me]?)|(sis)|(ses))" within s
```

We find that the infinitive may target the slot between the complementizer and the clitic auxiliaries (17), a configuration which is known to be disallowed with participle raising (8-b):

(17) *Samozřejmě uznávám, že ohánět jsem se po něm*
certainly acknowledge-1.SG that beat AUX-1SG REFL for him
neměl.
not-should
'Of course I acknowledge that I shouldn't have beaten him up.'

Second, there are many cases of a *complete* VP being fronted across the clitics, as in Avgustinova & Oliva's (1997) examples mentioned above:

(18) [*zahodit ji a vydat se pěšky na útěk k východním*
throw-away her and start-out REFL on-foot on flight to Eastern
hranicím směr domů]$_k$ *jsi nemohl* t_k
borders direction home AUX-PT.2SG not-could
'You couldn't throw her away and start out on foot, fleeing towards the Eastern borders.'

Third, we also encounter some clear cases of *partial* VP-fronting, as in

(19) *Usadit nastálo jsem se chtěl v Patagonii*
settle-down constantly AUX-PT.1SG REFL wanted in P.

[11] In ordinary language, "a sequence of an infinitive, any number of non-punctuation and non-*a* '*and*', and a past auxiliary, within one sentence".

'I wanted to settle down constantly in Patagonia.'

Interestingly, the search hits show a linear order restriction such that the infinitival verb always *precedes* its objects. The only cases in which another, sentence-initial word preceded the infinitive consisted of stacked infinitives, as in

(20) [$_{VP}$ *Pomoci objasnit celý případ*] *by mohli taxikáři, kteří*
help explain whole case SBJ could taxi-drivers who
napadení turistů viděli.
attack tourists-GEN.PL saw
'The taxi drivers, who saw the attack on the tourists, could help to clarify the whole case.'

Intuitive judgments support this impression, cf.

(21) **celý případ objasnit by mohli taxikáři*
whole case explain SBJ could taxi-drivers

Obviously, the base order within VP (this time, verb final) has to be preserved also in analogous German examples:

(22) a. *Den ganzen Fall aufklären könnten die Taxifahrer, ...*
the whole case clarify could the taxi-drivers
'The taxi drivers could clarify the whole case ...'
b. **Aufklären den ganzen Fall könnten die Taxifahrer, ...*
clarify the whole case could the taxi-drivers

An explanation for this pattern could build on the idea that in the ungrammatical (22-b) and (21), two constituents move independently, while there is only one landing site available. Under the view of scrambling as A-movement to some specifier in the Agr- or T-domain (Zybatow and Junghanns, 1998), the derivation of (21) would have to involve an intermediate step in which the object and the main verb do not form a constitutent any more, so they cannot move as one VP.

As concerns information structuring, the corpus examples display a pattern

of contrastive topic plus sentence-final focus throughout:

(23) *(O kolej žádá asi šestnáct set lidí.)* [$_{\text{TOP}}$
for dormitory ask probably sixteen hundred people
Nabídnout] jsme mohli pouze [$_{\text{FOC}}$ *pět set sedmdesát*
offer AUX-PT.1PL could only five hundred seventy
míst, která uvolnili letošní absolventi.]
places which freed this-year's alumni
'(About 1600 persons apply for the dormitory.) We could offer only 570 places, (which opened after this year's alumni left.)'

3.1.2 Low VP-fronting

VP-fronting to a position between the clitics and the finite verb was searched using a query like[12]

(24)
```
"((js[iemt][me]?)|(sis)|(ses))" [tag!="Z.*" &
lemma!="a"]* [tag="Vf.*"] [tag!="Z.*" &
lemma!="a"]+ [tag="Vp."] within s
```

The example in (25) illustrates the fronting of a complete VP to this area, with the base order *verb*>*object* preserved:

(25) *(… pak je to nejen proto,) že jsem* [$_{\text{VP}}$ *udělat kariéru*
then is this not-only because that AUX-PT.1SG make career
] *chtěla a chci, ale také proto, že …*
wanted and want-1SG but also because that
'(…then this is the case not only) because I wanted and want to make a career, (but also because …)'

The low landing site is below the subject position, as (26) indicates:

[12] In natural language, "a sequence of a clitic auxiliary, any number of non-punctuation and non-*a* 'and', an infinitive, at least one token which is not punctuation and not *a* 'and', and a finite verb, within one sentence."

(26) *Pokud by zákonodárci schválit misi nestihli, ...*
if SBJ legislators approve mission not-managed
'If the legislators would not manage to approve of the mission ...'

The moved VP can be partial—e. g., if one of its subconstituents goes into topic position:

(27) *Zadarmo jsem se jí [vzdát]$_l$ ale nechtěla t_l.*
for-free AUX-PT.1SG REFL her give-up but not-wanted
'But I didn't want to give her up for free.'

As opposed to the pre-clitic, high landing site, the low landing site imposes no linear restrictions on the fronted VP—e. g., the object can precede its governing infinite verb:

(28) *... ale dál jsem se politice* [*věnovat*]$_k$ *skutečně*
but further AUX-PT.1SG REFL politics devote really
nechtěl t_k.
not-wanted
'...but I really did not want to devote myself to politics any further.'

This is not very surprising, given that scrambling to the front of the VP is iterable in Czech anyway, so that there are always enough structural positions available to front a VP and one of its elements independently. As far as information structural distinctions are concerned, low VP-fronting seems to serve mainly one purpose: to move background material out of the focus domain, which then consists only of the right-peripheral—mostly negated—finite verb. All of the 106 relevant examples extracted from the ČNK conformed to this pattern.

(29) *"...Za druhou půli jsme však* [$_{VP}$ *vyhrát]$_i$ určitě* [$_{FOC}$
during second half AUX-1PL MP win certainly
zasloužili t_i]," řekl trenér
deserved said coach
' "But in the second half we surely deserved to win", said the coach.'

3.2 Polish

3.2.1 High VP-fronting

The IPI PAN query for Polish VP-fronting illustrates some pecularities of the annotation scheme used by Przepiórkowski (2004):[13]

(30)
```
[pos=inf] [pos!=interp & pos!=praet &
base!="((i)|(lub)|(albo))"]* "by" [pos=aglt]
within s
```

Some common word classes are being replaced by more fine-grained distinctions (`inf, praet`), which correspond more closely to differences in inflectional type. Furthermore, the PN-marker is always analyzed as a seperate token (of word class `aglt`), even if it occurs immediately after the main verb. These changes result from a careful and linguistically motivated tagset design (cf. Przepiórkowski and Woliński 2003): As mentioned above, the subjunctive marker and the PN-marker, although intuitively part of the verbal inflectional paradigm, may attach to various constituents phonetically, and also orthographically. It is therefore easiest to treat these items as tokens of their own, ignoring orthographic word boundaries at the level of morphosyntax (cf. section 4 for further remarks on the tagset).

The query in (30) mainly uses the subjunctive marker to delimit the high landing site of fronted infinite VPs. Since the analysis involving raising of the finite verb up to the subjunctive marker may be controversial, I will rely on subjunctive markers occurring separately as topological markers. Fronted VPs may be either complete (31) or partial (32):

[13] The query roughly reads as “a sequence of an infinitive, an arbitrary number of tokens which are neither punctuation nor a past tense verb nor one of *i* ‘and’, *lub* ‘or (incl.)’ *albo* ‘or (excl.)’, *by*, and an agglutinative affix (*-śmy, -ście*, etc.), all within one sentence”.

(31) *W każdym bądź razie bodaj na części zasobów muzealnych*
in every possible case MP on part stock-GEN.PL museum
uwłaszczyć byśmy się mogli, . . .
acquire SBJ-AUX-1PL REFL could
'But every single time we could have acquired at least part of the museum stocks, . . . '

(32) *Żyć bym bez nich nie potrafił.*
live SBJ-AUX-1SG without them not managed
'I would not have managed to live without them.'

As (31) shows, there is no constraint on linear order within the fronted VP, as opposed to the analogous case in Czech. In fact, more than one constituent may precede the fronted VP, cf.

(33) *. . . , ale ja niczego absolutnie wykluczać bym w tej materii*
but I nothing absolutely exclude SBJ-AUX-1SG in that issue
nie chciał.
not wanted
'. . . but I would not want to exclude anything in this matter.'

In the light of (33), it is not surprising that several subconstituents of the infinite VP may move independently to several stacked specifiers or adjunction positions, resulting in a linear order as in (31).

Regarding information structure, in the corpus examples either the whole fronted VP (34) or at least a subconstituent of it (35) functions as a contrastive topic:

(34) *. . . bez sieci nie potrafię już żyć, a pewnie* [$_{\text{TOP}}$ *umrzeć*] *też*
without net not be-able already live and surely die also
bym nie potrafił.
SBJ-AUX-1SG not be-able
'. . . without a net(work) I cannot live any more, and probably also could not die (without it).'

(35) *(Ponieważ jednak los tej ustawy w chwili obecnej jest wysoce*
because however fate this law-GEN in time present is highly
hipotetyczny,) dalej w tej kwestii posunać bym się
hypothetical further in this question move SBJ-AUX-1SG REFL
nie mógl.
not could
'(But since the fate of this law is highly hypothetical at present,) I could not go any further in this issue.'

3.2.2 Low VP-fronting

The query for VP-fronting to the lower landing site looks very similar to (30), except for the infinite VP *following* the subjunctive and PN-marker. A partial VP-fronting example from the IPI PAN corpus would be

(36) *. . . , ale też długo bym leżeć nie chciała, (bo*
but also long SBJ-AUX-1SG lie not wanted because
bym chyba nie wytrzymała.)
SBJ-AUX-1SG probably not stand
'. . . but I wouldn't want to lie around for long, (because I probably couldn't stand it).'

As in Czech, the fronted VP mostly moves out of the domain of focus, leaving behind a minimally focused main verb. However, there are also cases in which the most plausible focus would be on the whole finite VP or even on the whole clause:

(37) *("Miasto Lozanna zresztą dość nudne. . . . byłoby nam tu*
city L. by-the-way rather boring SBJ-3SG us here
dobrze,) gdybyśmy przywyknąć mogli do cudzej ziemi!"
good if-SBJ-AUX-1PL adjust could to foreign country
'("By the way, the city of Lausanne is rather boring . . . we would feel good here,) if we could adjust to the foreign country." '

The construction also occurs with enumerations of non-minimal focus domains,

as in (3) above. At least in these two cases, there is no obvious information structural side effect of VP-fronting in Polish.

4 Corpora of Slavic Languages as a Source for Syntactic Research

In this section, I will take a step back and consider the usability of the available Czech and Polish corpora for research into syntax and information structure of the kind presented in the preceding section. The intention is not a confrontative comparison of the ČNK and the IPI PAN corpus, but rather a collection of ideas for more user-friendliness and search power.

4.1 Corpus Annotation (POS-Tagging and Morphosyntactic Analysis)

The two main features we used in the queries were regular expressions over word forms and part-of-speech (POS) tags. The query language of the ČNK is essentially the CQP language (cf. Christ 1994); the query language of the IPI PAN corpus is very similar, but offers some interesting modifications (see below). Both corpora are fully lemmatized.

The ČNK (or rather, its publicly accessible part SYN 2000) consists of about 100 million tokens, with about 60 % taken from the public press, 20 % from non-fictional literature, and 15 % from fictional literature. The mixture of texts has been carefully considered, and SYN 2000 nowadays functions as a stable reference corpus. Around this core, two larger corpora of spoken Czech and several other collections have been made available.

Rather than mere POS-tagging, a full morphosyntactic analysis was conducted on SYN 2000, using a set of more than 2000 different theoretically possible tags. Accuracy of the stochastic tagging was estimated at about 93 % in 1998 (Hajič and Hladká 2000). The remaining errors, partly accidental, partly systematic, have evoked harsh criticism by members of the Institute of the Czech National Corpus and their colleagues. Efforts are being made to improve on the

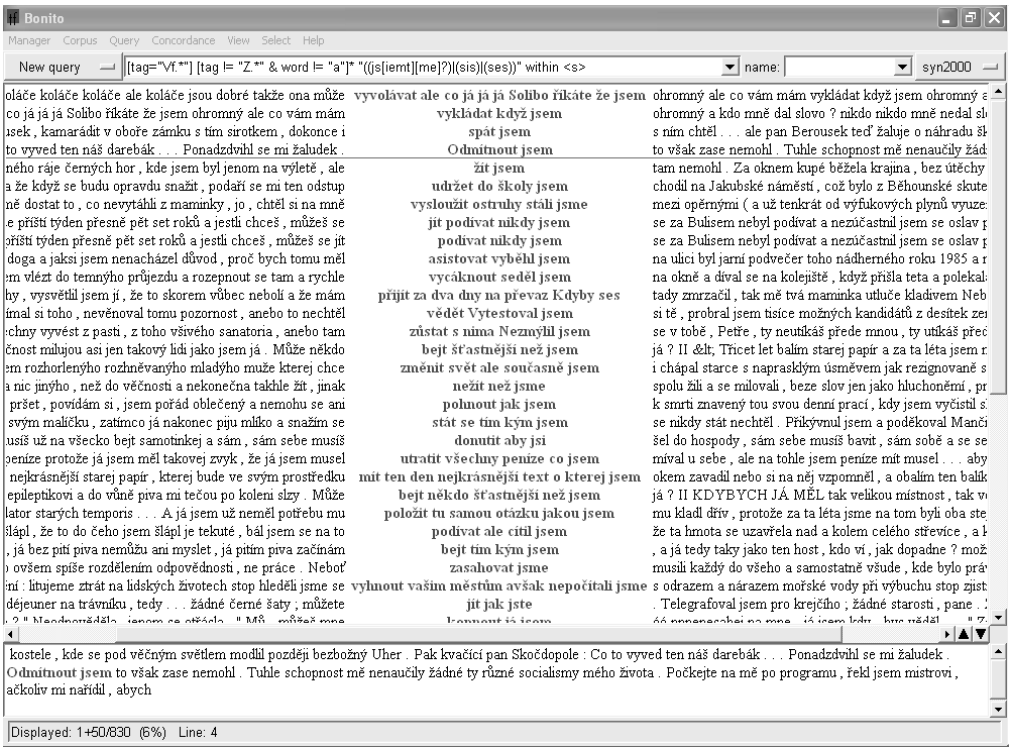

Figure 1: The Czech National Corpus and its query interface Bonito

tagging by using a mixture of stochastic and rule-based methods ((Hajič et al., 2001)). The ČNK, just as the IPI PAN corpus or the MULTEXT-East corpora (Erjavec and Monachini 1997), uses a positional tagset, i. e., the value of each grammatical category is encoded by one character in a fixed position in the string which makes up a tag.

The IPI PAN corpus was released by the Institute of Computer Science of the Polish Academy of Sciences as a first version in June 2004. It currently consists of about 70 million tokens, or rather “segments” (Przepiórkowski 2004): the clitic PN- and subjunctive marker are regarded as units of their own right and are written seperately from their prosodic hosts. Half of the materials contained in the corpus are newspaper texts, 20% are fictional (prose), about 10% scientific writing, 5% law texts, and 15% session protocols from the Polish parliament. As Przepiórkowski (2004) states himself, this collection is rather opportunistic

than balanced and will have to be improved upon at a later stage.

Figure 2: The IPI PAN Corpus and its query interface Poliqarp

The POS- and morphosyntactic tagging was done by stochastic methods; however, at least two new ideas set this corpus apart from other stochastically tagged sources like the British or the Czech National Corpus:

First, the set of POS tags has been designed in a somewhat non-traditional way which relies as little as possible on lexical semantic groups, and thus could make word class recognition easier. Tokens were grouped into different word classes whenever they differed systematically with respect to their sets of possible grammatical categories (i. e., when they belonged to different *flexemes* in the terminology of Przepiórkowski and Woliński (2003), originally attributed to Janusz Bień.). Thus, instead of a mixed bag like “adjective”, there are the classes *adjective*, *ad-adjectival adjective* (e. g., *‘polsko-niemiecki’ Polish-German*), and *post-prepositional adjective* (e. g., *(po) polsku ‘(in) Polish’*), of which only the

first one shows a complete inflectional paradigm (Przepiórkowski 2004, 28). The number of possible tags is reduced dramatically by disregarding all the lexical semantic sub-distinctions of pronominals (interrogative, relative, demonstrative, etc.), grouping them with adjectives or nouns as far as possible, according to their inflectional behaviour. Most of the lexical semantic distinctions can be regained anyway by doing a lemma search; and there is virtually no hope that any automatic tagger will be able to detect the correct tags for homonymous pronouns belonging to different semantic classes. To give an example, a search of interrogative vs. (free) relative pronouns in the ČNK yields words from both categories without any obvious pattern.[14] A similar rationale as in the case of IPI PAN seems to have played a role in the design of the Stuttgart-Tübingen Tagset for German. Unfortunately, success rates of the IPI PAN tagging are not yet avaliable.

Second, the set of possible morphosyntactic tags before stochastic disambiguation is retained and may be searched. This can be useful and sometimes superior to a forced disambiguation. To give an example, a search was conducted for discontinuous NPs in dative case, which are split into the attributive adjective and the seperate head noun (a so-called *Left-Branch Extraction*). Since the dative case can be homonymous to other cases, depending on inflectional class, a forced stochastic disambiguation—as in most present-day corpora, including the ČNK—will inevitably yield errors:

(38) *ke každé otázce z přízemí*
to every-DAT question-DAT from stalls-*DAT/$\surd$GEN

In this example from ČNK, the genitive on *přízemí* has been wrongly tagged as a (homonymous) dative. All these examples will have to be reconsidered

[14] This is not very surprising, given the huge number of homonyms: In the case of Russian, a traditional, semantically based classification pronouns, as it was considered for some time for the corpora of the Tübingen Sonderforschungsbereich 441, would lead to about 600 different morphosyntactic tags only for this category, which is almost half the size of the whole tagset.

"manually" by the user, in order to filter out the true datives. With the ambiguous tagging of IPI PAN, however, the query may be limited to only those hits in which initial morphological analysis has already yielded a single, *non-ambiguous* form—disregarding all the potentially wrong tags.

Figure 3: Searching for non-ambiguous forms in the IPI PAN corpus

From a user's point of view, keeping track of ambiguous and unambiguous tags is thus certainly an advantage. At the same time, however, the need for better success rates of automatic disambiguation is obvious and has been acknowledged repeatedly by the corpus designers.

4.2 Comfortable Querying

Given the range of corpus users (a part of the ČNK, SYNEK, is even being propagated in secondary schools), user friendliness has become an increas-

ingly important issue.[15] The ČNK has been distributed together with a comfortable and easily understandable search tool from the beginning; this tool, called G(raphical)CQP, which was programmed by Pavel Rychlý of the NLP laboratory of Brno University, has been developed further into the corpus server *Manatee* and the client viewer *Bonito*, offering even more possibilities. A purely web-browser based version has been announced for 2005. To name just a few of the many interesting features of Bonito, there is (i) a graphical representation and construction of queries; (ii) Bonito offers a stepwise refinement of the queries by imposing positive and negative filters on search results. Nicely, the user can always return to a previous set of hits in case a filter did not give the intended result. (iii) Bonito contains enhanced statistic functionality for research into collocations, (iv) handling of user-defined subcorpora, and (v) export functions for search results. These possibilities greatly increase accessibility and would definitely be desirable for the corpora of other languages as well. Compared to the search tools of the British National Corpus, the recently initiated Russian National Corpus, and the German COSMAS corpus, the search tool of the Czech National Corpus definitely sets new standards.

Neither of (i)-(iv) have yet been realized in Poliqarp, the first version of the search tool coming with the IPI PAN corpus. However, at least one idea which increases searching comfort in Poliqarp deserves to be mentioned: the inclusion of tag aliases. E. g., instead of learning that the fourth position of each tag encodes number and the fifth position case, thinking up a regular expression over tags, and then typing in `[tag="...P3.*"]`, the user may simply state `[case=dat & num=pl]`. Given the large number of theoretically possible tags and their sometimes counterintuitive positional encoding (e. g. `[tag="Vf.*"]` for verbs in the *infinitive* in ČNK), this is extremely helpful. It would be even nicer if the user had the alternative to input the values of gram-

[15] The ČNK and the IPI PAN textual source data is not handed over to the user, so (s)he has to rely on the search tools provided by the corpus designers.

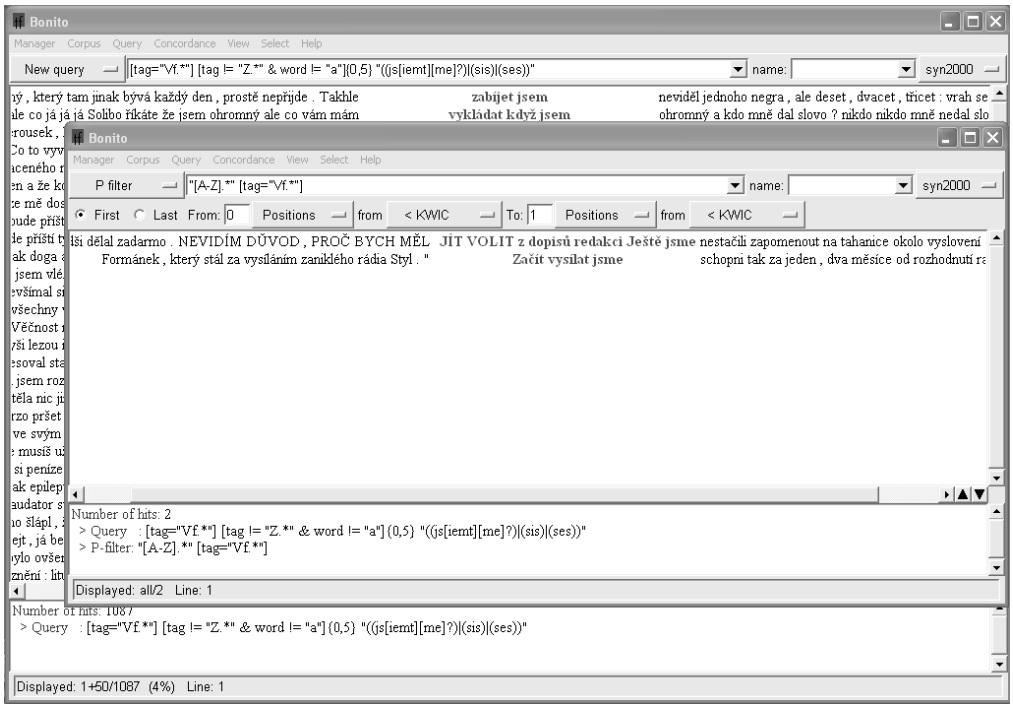

Figure 4: Stepwise filtering of query results in Bonito

matical categories by selecting from menues or simply checking boxes (as in the new Russian National Corpus).

4.3 A Parsed Corpus for Czech: the Prague Dependency Treebank

Both VP-fronting constructions involve displacement from an assumed base position to the left; in partial VP-fronting, even more cases of discontinuous constituency occur. We have seen above that instances of these constructions can be found by relying on morphosyntactic tags and some post-editing and filtering by hand. However, it would be more comfortable to run a search on discontinuous syntactic constituents directly. Queries over syntactic structures can be conducted for Czech using the Prague Dependency Treebank (PDT), a pseudo-random selection of about 55 000 authentic sentences from the ČNK which have been hand-annotated according to the theory of Functional-Generative Descrip-

tion (Sgall et al. 1986). This corpus may be searched on-line, using the search tool Netgraph (developers: R. Ondruška and J. Mírovský). The syntactic structures in the PDT are trees which encode dependency relations and relative linear order between words. A sample query for discontinuous VP constituents and one of its hits is the following:

(39) a.
```
[]([tag=Vf*,ord=1]([ord>=3]),
[afun=AuxV,ord=2])
```

b. *Pomoci by mu v tom měli i noví hráči.*
help AUX-SBJ him in this should also new players
'Also the new players should help him with this.'

The values of the feature `ord` in this case encode that the VP is supposed to be split, i. e., the query concerns an infinitive (first word) with an auxiliary verb as its sister (second word) and a daughter of the infinitive as third or later word.

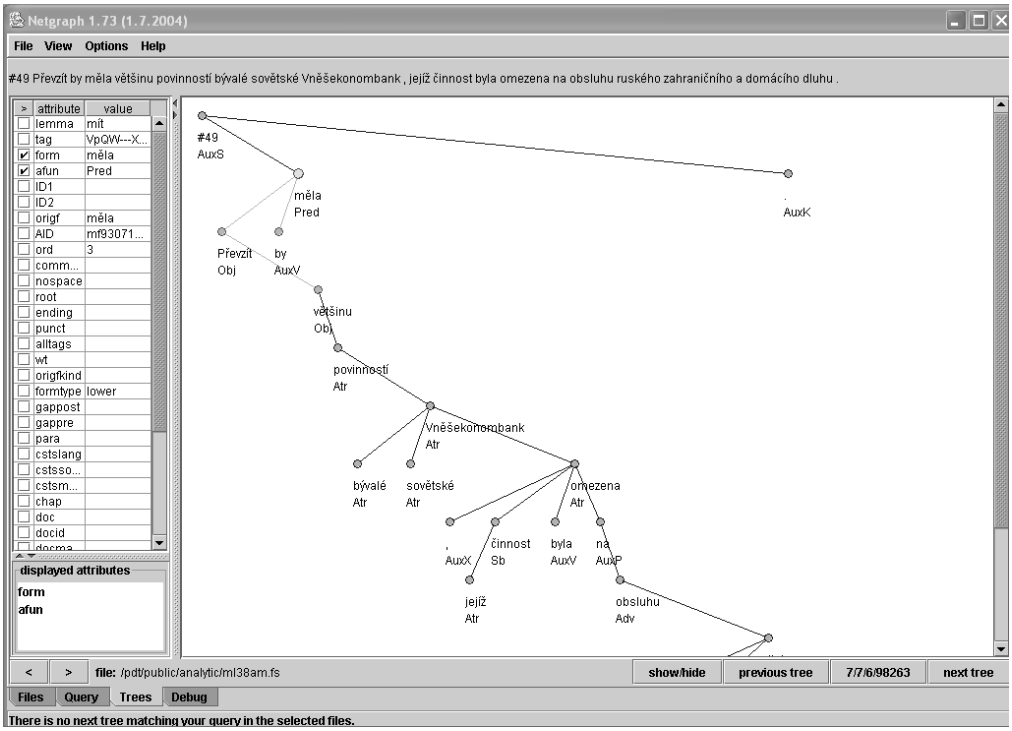

Figure 5: Discontinuous VP-fronting in the Prague Dependency Treebank

A second version of the treebank, including annotation for topic, focus, and contrast (in the sense of Hajičová et al. 2000), has been announced for 2005. It may then serve as a starting point for studies of information structure, but the user still has to go back to the original ČNK and perform a second search to see the context of a retrieved sentence; the search tools of PDT and ČNK are not integrated at present. Although the PDT is definitely valuable for a large scale of natural language processing tasks, it seems that empirical linguistic studies will still have to rely on searching truly large corpora by lower-level annotation such as lemmata and morphosyntactic tags, as well as by forms and collocations.

4.4 Further Freely Accessible Corpora of Slavic Languages

Many modern Slavic languages can already be investigated on the basis of large corpora with free online access. Among those offering at least a search by regular expressions over word forms are the Oslo Corpus of Bosnian Texts, the Croatian National Corpus, the Serbian National Corpus, the Russian Corpora of the Sonderforschungsbereich 441 (University of Tübingen), the Slovak National Corpus, and the Polish Corpus of the PWN publishing house. POS annotation has been provided for the Czech National Corpus, the Polish IPI PAN Corpus, a small portion of the Tübingen Russian Corpora, and, increasingly, for the Russian National Corpus. The MULTEXT-East project has produced POS-annotated versions of Orwell's "1984" for Czech, Slovenian, and Bulgarian. Syntactically analyzed corpora are currently available for Bulgarian (HPSG treebank) and Czech (Prague Dependency Treebank).[16] Work on *diachronic* corpora of Slavic languages has been started at Charles University, Prague (Old Church Slavonic and Czech), the University of Sofia (Bulgarian), and the University of Regensburg (Russian). This short overview is necessar-

[16] IPI PAN has developed a "test suite" of HPSG-parsed Polish sentences, which is, however, rather intended as an overview of possible sentence structures than as a corpus of natural language.

ily incomplete, updated collections and links to web pages may be found e. g. at `www.uni-tuebingen.de/uni/nss/docs/Korpora.html` and at `www-slavistik.uni-regensburg.de/Corpus`.

5 Conclusion

Corpus evidence from ČNK and IPI PAN indicates that in Czech, high VP-fronting is the non-iterable movement of a single—complete or partial—constituent, which does not allow for internal word order variation. In the other three cases considered, i. e. in Czech low VP-fronting and in Polish in general, there are no such restrictions: VP-movement can be accompanied by further fronting operations to the same area, and VP-internal word order is basically free. Theoretically, high VP-fronting in Czech should thus be analyzed as movement to an A'-specifier (SpecC), while the other three operations end up in an adjunction position—either above T (Polish high VP-fronting) or between T and the main VP (Polish and Czech low VP-fronting). In terms of contextual conditions and information structure, high VP-fronting in our Czech examples always involves a contrastive topic, while low VP-fronting is movement of background material to the left; both combine with a minimimal, right-peripheral focus. In Polish, the high fronting also favors an interpretation as a contrastive topic, while low VP-scrambling seems to be compatible with various information structural partitions of the sentence.

To arrive at these generalizations, we made use of corpus data in the following ways: (i) Single intuitive judgments were supported (i. e., not falsified) by a large body of original data. This holds for most of the evidence on clitic placement in section 2. (ii) Naturally occurring contexts were evaluated, resulting in statements about their relative frequency. This is the case for the information structural effects of VP-fronting reported above. (iii) Of two theoretically possible constructional variants, one could be found with a certain basic frequency,

but the other did not occur at all. The corpus thus indicates a candidate for a restriction, which has to be checked against intuitive judgments. This was done, e. g., for the linear order restriction with Czech high VP-fronting. Technically, we simply searched by sequences of regular expressions over words, lemmata, and morphosyntactic descriptions. The user-friendliness for this process could be enhanced considerably if the following features, partly realized already in ČNK and IPI PAN, became standard: (i) retrieval of (un)ambiguous tags before automatic disambiguation; (ii) stepwise refinement of search hits, handling of user-defined subcorpora, easy export of hits; (iii) statistic functions for research into collocations; (iv) aliases or a graphic interface for the input of tags.

Bibliography

T. Avgustinova and K. Oliva. On the Nature of the Wackernagel Position in Czech. In U. Junghanns and G. Zybatow, editors, *Formale Slavistik*, volume 7 of *Leipziger Schriften Zur Kultur-, Literatur-, Sprach- und Übersetzungswissenschaft*, pages 25–57. Vervuert Verlag, Frankfurt/Main, 1997.

P. Bański. Last Resort Prosodic Support in Polish. In G. Zybatow, U. Junghanns, G. Mehlhorn, and L. Szucsich, editors, *Current Issues in Formal Slavic Linguistics*, volume 5 of *Linguistik International*, pages 179–186. Peter Lang, Frankfurt, 2001.

J. Błaszczak. *Investigation into the Interaction between the Indefinites and Negation*, volume 51 of *studia grammatica*. Akademie-Verlag, Berlin, 2001.

O. Christ. A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*. Budapest, 1994.

E. Dornisch. Auxiliaries and Functional Projections in Polish. In Ž. Bošković, S. Franks, and W. Snyder, editors, *Annual Workshop on Formal Approaches to Slavic Linguistics: The Connecticut Meeting, 1997*, volume 43 of *Michigan Slavic Materials*, pages 183–209. Michigan Slavic Publications, Ann Arbor, 1998.

T. Erjavec and M. Monachini. *Specifications and Notation for Lexical Encoding (= COP Project 106 MULTEXT-East Final Report)*, 1997. http://nl.ijs.si/ME/CD/docs/mte-d11f/.

G. Fanselow. Against Remnant VP-Movement. ms., Universität Potsdam, 2004.

J. Hajič, P. Krbec, P. Květoň, and Vl. Petkevič. Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In *Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference*, pages 260–267, Toulouse, 2001. Morgan Kaufman Publishers.

E. Hajičová, J. Panevová, and P. Sgall. A manual for tectogrammatic tagging of the prague dependency treebank. Technical report, Institute of Formal and Applied Linguistics / Center for Computational Linguistics, Prague, 2000.

J. Hajič and B. Hladká. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *COLING-ACL'98*, pages 483–490. Morgan Kaufman, San Francisco, 2000.

U. Junghanns. Generative Beschreibung periphrastischer Konstruktionen des Tschechischen. In T. Anstatt, R. Meyer, and E. Seitz, editors, *Linguistische Beiträge zur Slavistik aus Deutschland und Österreich. VII. JungslavisInnen-Treffen, Tübingen/Blaubeuren 1998*, pages 133–165. Sagner, München, 1999.

R. Meyer. *Syntax der Ergänzungsfrage. Empirische Untersuchungen am Russischen, Polnischen und Tschechischen*, volume 436 of *Slavistische Beiträge*. Otto Sagner Verlag, München, 2004.

G. Müller. *Incomplete Category Fronting*. Kluwer, Dordrecht, 1998.

A. Przepiórkowski. *The IPIPAN Corpus. Preliminary Version*. Institute of Computer Science, Polish Academy of Sciences, Warszawa, 2004.

A. Przepiórkowski and M. Woliński. A morphosyntactic tagset for polish. In P. Kosta, J. Błaszczak, J. Frasek, L. Geist, and M. Żygis, editors, *Investigations into Formal Slavic Linguistics. Contributions of the Fourth European Conference on Formal Description of Slavic Languages – FDSL IV*, volume 1, pages 349–362. Peter Lang, Frankfurt, 2003.

P. Sgall, E. Hajičová, and J. Panevová. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Academia, Praha, 1986.

L. Veselovská. *Phrasal movement and X^0-morphology: word order parallels in Czech and English nominal and verbal projections*. PhD dissertation, Univerzita Palackého, Olomouc, 1995.

J. Witkoś. Pronominal Argument Placement in Polish. *Wiener Linguistische Gazette*, 57-59:147–194, 1996.

G. Zybatow and U. Junghanns. Topiks im Russischen. *Sprache und Pragmatik*, 47:1–57, 1998.

Roland Meyer
Universität Regensburg
Institut für Slavistik
Universitätsstr. 27
93040 Regensburg
Germany
roland.meyer@sprachlit.uni-regensburg.de
http://www-slavistik.uni-r.de/institut/meyer