Erschienen in: Konopka, Marek/Kubczak, Jaqueline/Mair, Christian/Štícha, František/Waßner, Ulrich H. (Hrsg.): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.-24.09.2009. - Tübingen: Narr, 2011. S. 63-86. (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, Bd 1)

MIRJAM FRIED

Grammatical analysis and corpus evidence

Abstract

This study explores the interdependence of qualitative and quantitative analysis in articulating empirically plausible and theoretically coherent generalizations about grammatical structure. I will show that the use of large electronic corpora is indispensable to the grammarian's work, serving as a rich source of semantic and contextual information, which turns out to be crucial in categorizing and explaining grammatical forms. These general concerns are illustrated by the patterns of use of Czech relative clauses (RC) with the non-declinable relativizer *co*, by taking a set of existing claims about these RCs and testing their accuracy on corpus material. The relevant analytic categories revolve around the referential type of the relativized noun, the interaction between relativization and deixis, and the semantic relationship between the relativized noun and the proposition expressed by the RC. The analysis demonstrates that some of the existing claims are fully invalid in the face of regularly attested semantic distinctions, while others are more or less on the right track but often not comprehensive or precise enough to capture the full richness of the facts.

1. Introduction

One of the central challenges in articulating empirically grounded generalizations about grammatical patterning is the task of maintaining balance between two sources of pressure: the need to identify the inventory of relatively stable, predictably recurrent patterns that we can collectively refer to as 'grammar', while respecting and capturing the inherently dynamic, variable nature of grammatical structure. It has been increasingly noted and argued in functionally and cognitively oriented research that descriptive and explanatory adequacy in grammatical descriptions requires reference to meaning and to patterns of usage; such an approach, in turn, calls for systematic attention to a sufficiently representative body of authentic linguistic material. The goal of this paper is to show that the use of large electronic corpora is indispensable to the grammarian's work, primarily as a rich source of semantic and contextual information, which is highly relevant in categorizing and explaining grammatical forms. This is in contrast to the traditional methods and approaches, in which grammatical descriptions generally take the form of static, discrete

'rules' that are often formulated on the basis of introspection. Even when textual evidence is taken into account, it is used rather unsystematically and selectively, and any quantificational claims based on such evidence have, at best, very limited informational value. The present study argues for combining qualitative and quantitative evidence as two interdependent dimensions of grammatical analysis which aims at both descriptive and explanatory adequacy.

The theoretical and methodological issues will be illustrated on one particular syntactic form in Czech, concerning the use and classification of relative clauses (RC) with the non-declinable relativizer *co*, shown in (1); the relativizer is often accompanied by a resumptive pronoun, in (1a) exemplified by the personal pronoun *ho* 'him'. The absolutive RCs constitute a relativization strategy that is formally distinct from agreeing RCs, introduced by a fully declinable agreeing relative pronoun *který* 'which'; the examples in (2) are constructed agreeing variants of (1).[1]

(1) a. *Ten člověk, co jste ho za mnou kdysi poslal,*
that man CO AUX.2PL 3SG.ACC.M after me once sent[2]
{viděl jste ho ještě někdy potom?}
"The man [CO] you sent [him] to me a while back, {did you ever see him again later}?"

b. *ta paní, co u nás bydlí, je moc hezká*
that woman CO at us lives is much pretty
"the woman who lives with us is very pretty"

(2) a. *Ten člověk, kterého jste za mnou kdysi poslal,*
that man which.ACC.SG.M AUX.2PL after me once sent
{viděl jste ho ještě někdy potom?}
"The man who[m] you sent to me a while back, {did you ever see him again later}?"

b. *ta paní, která u nás bydlí, je moc hezká*
that woman which.NOM.SG.F at us lives is much pretty
"the woman who lives with us is very pretty"

[1] Unless otherwise noted, the examples all come from the SYN2000 corpus of written Czech.

[2] Abbreviations used in the glosses: AUX 'auxiliary', SG/PL 'singular/plural', NOM 'nominative', ACC 'accusative', DAT 'dative', GEN 'genitive', INS 'instrumental', M 'masculine', F 'feminine', NEG 'negation', PRES 'present', FUT 'future', IMP 'imperative', PST 'past', RF 'reflexive'.

While the agreeing RCs are fairly well understood, the absolutive RCs have so far attracted only sporadic attention among Czech linguists, although some partial studies of their properties and distribution do exist (Zubatý 1918; Poldauf 1955; Svoboda 1967, 1972; Lešnerová / Oliva 2003) and reference grammars or other comprehensive grammatical works may briefly mention them (Trávníček 1951, Kopečný 1962, Šmilauer 1972, *Mluvnice češtiny* 1987, Grepl / Karlík 1998). As a first step toward a more comprehensive examination of the absolutive RCs, this study will take a subset of existing claims about them and the 'rules' for their form, interpretation, and distribution as presented in the Czech grammatical literature, and test their accuracy on corpus material. The relevant analytic categories, with implications for relativization strategies beyond the Czech facts, will revolve around the referential type of the relativized noun (henceforth referred to as the head N), the interaction between relativization and deixis, and the semantic relationship between the head N and the proposition expressed by the RC. The analysis, which takes into account frequency-based quantitative patterns of usage, will demonstrate that some of the existing claims are either fully invalid, or too general to capture relevant semantic distinctions, while others are more or less on target but often too inflexible to truly capture the attested facts. In general, the point of the present work will be to introduce corpus evidence into the task of analyzing the absolutive RCs (or *co*-RCs) in their full, empirically documented complexity.

Thus, on the basis of corpus material, the present study argues for a more dynamic approach to grammatical analysis, one in which grammatical generalizations can be structured in cognitively and communicatively coherent networks of related grammatical patterns. The networks simultaneously provide a tool for (i) identifying points of potential fluctuations within the usage of a particular form and (ii) tracking incipient shifts between the form and / or function of a given grammatical pattern.

2. Background – relative clauses with the relative pronoun *který*

Relative clauses marked by the agreeing relative pronoun *který* cover a broad functional and semantic spectrum. For the purposes of this study, I will take it for granted that we can, at a minimum, identify the interpretations exemplified in Table 1; this taxonomy is a synthesis of two existing and roughly compatible accounts of these clauses (Svoboda 1972: 109, Grepl / Karlík 1998: 184-196) that, taken together, provide a sufficient level of detail to be useful.

Table 1: Examples of RCs with relative pronoun *který*

I-A. Determinative restrictive	
1. Concept/category membership/ defining feature of head N	*Jsou lidé, které o tomhle nikdy nepřesvědčíš.* are people.NOM which.ACC.PL about this never NEG.convince.FUT.2SG "There are people who you'll never convince." (Svoboda 1972)
2. 'Kind of' specification	*Hledáme manažerku, která umí francouzsky.* seek.PRES.1PL manager.ACC.SG.F which.NOM.SG.F know.PRES.3SG French "We're looking for [a] manager who [can] speak French."
3. Identification	*Podej mi knihu, která leží tam na stolku.* hand.IMP.2SG 1SG.DAT book.ACC.SG.F which.NOM.SG.F lies there on table "Hand me [the] book that's over there on the table."
4. Characterization	*Včera jsem viděl film, který natočil.* yesterday AUX.1SG see.PST.SG.M film.ACC.SG.M which.ACC.SG.M made *{Forman ještě v Československu.}* (Grepl / Karlík 1998) "Yesterday I saw [a] movie that Forman made {when still [working] in Czechoslovakia.}"
I-B. Determinative non-restr.	*{ale nakonec mně bude chybět i}* *ten Zetka, kterým jsme ve třídě všichni opovrhovali.* that Z.NOM.SG.M which.INS.SG.M AUX.1PL in class all.NOM.PL.M look.down.PST.PL "{but in the end I'll be missing even} that [guy] Zetka, who the whole class looked down on"
II. Non-determinative (always non-restrictive)	
II-A. Explicative	*{ale nakonec mně bude chybět i}* *Zetka, kterým jsme ve třídě všichni opovrhovali.* Z.NOM.SG.M which.INS.SG.M AUX.1PL in class all.NOM.PL.M look.down.PST.PL.M "{but in the end I'll be missing even} Zetka, who the whole class looked down on"
II-B. Continuative	*Hledal asi hodinu poštovní schránku, kterou nenašel.* seek.PST.SG.M maybe hour mailbox.ACC.SG.F which.ACC.SG.F NEG.find.PST.SG.M "He[spent] about an hour looking for [a] mailbox, which he didn't find."

Functionally, the RCs form two major classes: +/– determinative (type I vs. II) and +/– restrictive (type I-A vs. the rest). The former captures the RC's status according to its (ir)relevance for identifying, or determining, the referent of the head N, while the latter establishes restrictiveness-based distinctions within the determinative patterns; non-determinative patterns are all non-restrictive. The determinative restrictive clauses come in several semantic flavors. The RC may determine the head N in terms of *category* membership by expressing some fundamental, defining features of the head N (ex. I-A-1) in Table 1; as a possible (and potentially non-existent) token of a *kind* (ex. I-A-2); as a concrete individual that is fully *identified* in a given context by the proposition expressed in the RC (ex. I-A-3); or as a concrete unique referent that is *characterized* as such by the RC but whose identity cannot be fully established in a given context (ex. I-A-4). The determinative non-restrictive RCs (type I-B) co-occur with head Ns that consist of a deictically anchored noun with unique reference (e.g., proper nouns); the obligatory presence of the demonstrative pronoun *ten / ta / to* "that.m / f / n" individuates the referent in context and contributes to its identifiability. Type I-B forms a minimal pair with non-determinative *explicative* RCs (type II-A), in which the head N is also a noun with unique reference but any presence of a demonstrative pronoun is prohibited; the job of these RCs is to provide further commentary about a referent that is already fully identified without the RC. Finally, *continuative* RCs express a proposition that is logically independent of the properties of the head N and is in a coordination relation to the main clause (type II-B).

The meanings and functions exemplified in Table 1 can be organized in a preliminary representational taxonomy sketched in Figure 1. It has been acknowledged (e.g., Svoboda 1972: 109, Grepl / Karlík 1998: 187) that it may not always be easy (or even possible) to categorically differentiate one type from another. Certain semantic overlaps and somewhat fluid transitions between parts of the taxonomy are apparent, particularly among the non-restrictive uses (in the diagram enclosed in the gray area), but potentially also in the characterization RCs since these do not allow explicit deixis and do not ensure full identification of the head N, in contrast to other restrictive RCs; their somewhat special relationship to the individuating function is indicated by the dotted line in Figure 1.

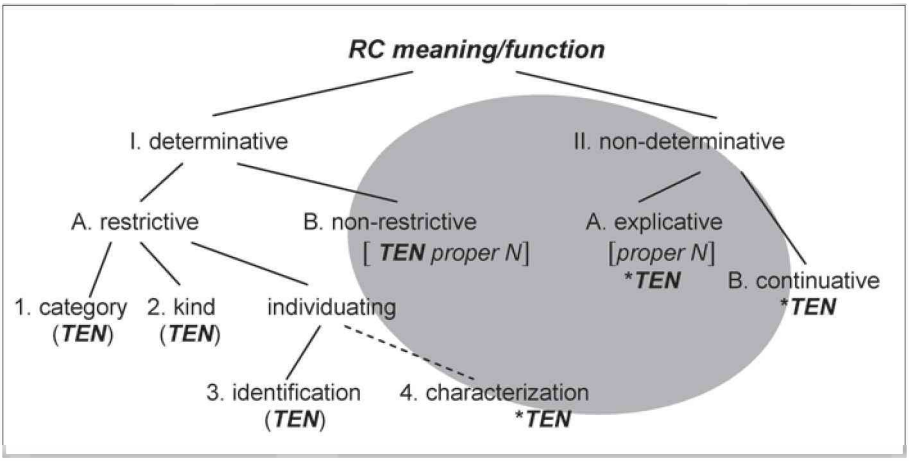

Figure 1: Functional and semantic classification of RCs with relative pronoun *který*

The properties of the agreeing RCs can thus be summarized as follows: (i) They can have both restrictive and non-restrictive interpretation and there is no obligatory marking associated with (non)restrictiveness, whether in written or spoken Czech. (ii) The relative pronoun agrees in number and gender with the relativized noun. (iii) Agreeing RCs are stylistically neutral (in terms of register, genre, text-type). (iv) Different semantic types appear to interact with deixis in different ways, showing distinct collocational preferences regarding the use of the demonstrative pronoun (in the diagram indicated by the pronoun *TEN* "that"); explicit deixis is possible in restrictive clauses of type I-A-1,2,3 (indicated by the parentheses), obligatory in the determinative non-restrictive clauses (type I-B), and prohibited everywhere else (indicated by the asterisk).

3. Relative clauses with absolutive relativizer *co*

Against the background of the agreeing RCs just described, the RCs with the absolutive relativizer *co* can be characterized as follows. As shown in the introductory examples in (1) and (2), the two types of RCs often (though not always) appear interchangeable. The absolutive relativizer is often accompanied by a resumptive (personal) pronoun, which agrees with the head N in number and gender and indicates the head N's grammatical function in the RC by being marked for the appropriate case (in the agreeing RCs, all three categories are expressed by the relative pronoun *který*). Czech grammatical literature al-

most uniformly describes the absolutive RCs as stylistically restricted and, specifically, as just a colloquial variant of the agreeing RCs.[3] However, beyond these general observations, the existing treatment leaves a lot unanswered about the defining features and the distribution of the absolutive relativizer and about the functions of these RCs. In this section, I will examine several specific claims that have been put forward about the *co*-RCs and confront them with what can be found in the *Czech National Corpus*.

The first two claims, somewhat interrelated, are rather general and have to do with the interaction of *co*-RCs with the determinative function and with deixis. The remaining questions to be addressed are more specific and concern concrete distributional constraints that follow from the first two claims. I will turn to the general issues first.

3.1 Absolutive RCs in non-determinative functions

Existing accounts mostly agree that the absolutive RCs do not occur in non-determinative uses, i.e., they cannot replace the agreeing relative pronoun *který* in the type II clauses of the taxonomy (e.g., Trávníček 1951: 1164, Svoboda 1972: 106, *Mluvnice češtiny* 1987: 528). The terminology may vary from author to author but the conceptual consensus is clear, although sometimes it is only implied by the examples used for the *co*-RCs rather than being explicitly stated. Some accounts do hedge their classification by noting that RCs "usually" (e.g., Šmilauer 1972: 262) express determination, but no further commentary is offered as to the conditions under which they may not serve this function. Overall, the functional and distributional constraints on *co*-RCs in Czech grammatical literature can be summarized as a prohibition on non-determinative contexts (type II), with a stronger version formulated by Svoboda (1972: 106), who states it as a prohibition on non-restrictive contexts (both type II and type I-B).

Let us now examine how these hypothesized distributional constraints hold up against corpus evidence. The material on which my observations are based consists of a randomly selected sample of 879 relevant tokens, each of which is coded for the functional / semantic type of RC (according to the taxonomy in

[3] This blanket statement turns out to be an exaggeration. The corpus shows that the textual distribution of *co*-RCs is more complex and any stylistic or genre-based conditioning of *co*-RC usage requires finer-grained semantic analysis. I will ignore this dimension here, leaving it to future research, in which both written and spoken data must be used.

Figure 1); animacy, grammatical number, and semantic type of the head N, and the grammatical function the relativized referent serves in the RC, as reflected in the case form of the resumptive pronoun. The latter is exemplified by the introductory examples: in (1a) we have a resumptive pronoun in the accusative since the head N is the direct object in the RC, while (2a) contains no resumptive pronoun since the head N is the subject of the RC. Some other case options are illustrated below: a dative in (3) and instrumental in (4).[4]

(3) *boji se tlustých líných kocourů, co jim tam pan Bajer*
they.fear RF fat lazy tomcats CO 3PL.DAT there Mr. B.
nosí jídlo
carries food
"[they] are afraid of the fat, lazy tomcats [CO] Mr. Bajer brings food to [them]"

(4) {*Pomyslel jsem na doktora Fausta a*} *na ten starý problém, co jsem*
on that old problem CO AUX.1SG
se jím zabýval {*při nudnějších hodinách náboženství*}
RF 3SG.M.INS be.occupied.PST.SG.M
"{I thought of doctor Faust and} of that old problem [CO] I used to occupy myself with [it] {during somewhat boring lessons in theology}"

The sample also contains numerous tokens of two additional kinds of RCs that are not explicitly included in the *který*-taxonomy in Figure 1 and that require further commentary. One, exemplified in (5), is defined semantically by the type of its head N: the head is a relatively generic temporal expression (typically *doba* 'time period', but also various calendric units) that marks a point on a timeline. The RC identifies the temporal point as having particular properties, and the RC never refers back to the time expression through a resumptive pronoun, regardless of the grammatical function in the RC; I will label these RCs as 'temporal':

(5) a. *Od té doby, co jsem byl u vás doma*
from that time CO AUX.1SG be.PST.SG.M at you home
{*na návštěvě, už utekly asi čtyři roky*}
"{It's been about four years since} the time [CO] I visited you in your home"

[4] It is a well-known fact that the resumptive pronoun is prohibited in the subject function (2a), obligatory in all oblique functions (genitive, dative, instrumental, locative), and appears to be optional as a direct object, at least in some cases. The details of the distribution of the accusative pronoun are not well understood and require further study; I will not be concerned with this issue here.

b. *Tu neděli, co u nás byla paní Bohdalová, {padly všechny světové rekordy}*
that Sunday CO at us was Mrs. B.
"That Sunday [CO] Mrs. Bohdalová came by {all world records were broken}"

The other type, to my knowledge never mentioned in the existing accounts, is structurally and semantically more complex but for our purposes, it is sufficient to note that it is semantically distinct from all other RCs and must be, therefore, categorized as a separate subtype. I will refer to it as a 'quantifying' RC and a typical example is given in (6). Semantically, the RC is a particular – namely, quantifying – semantic subtype of the Explicative RC: the proposition expressed in the RC is presented as applicable to all possible members of the class denoted by the head N and the quantity is presented as exhaustive (notice, for example, the presence of the universal quantifier *všechny* "all" modifying the head N, as an explicit marker of the exhaustiveness). Moreover, the RC always contains a resumptive pronoun in the genitive of quantity (*jich* "of them").

(6) *Všechny politické strany, co jich máme, {zastávají skvělé*
all political parties CO 3PL.GEN have.PRES.1PL
a objevné myšlenky: prosperitu, pořádek, péči o potřebné, morálku, svobodu}.
"All the political parties – [the full number CO] we have [of them] – {advocate splendid, novel ideas: prosperity, order, care for the needy, morality, freedom}."

Both of these special subtypes of *co*-RCs also show signs of formulaicity, both structural and semantic, but those details need not concern us here. Their idiosyncrasies are discussed in Fried (in press).

If we just do a simple count of all the tokens in the sample, the determinative RCs outnumber the non-determinative ones at the rate of about 9:1, or, put differently, the non-determinative tokens represent around 10% of the total; this is shown in actual numbers at the top of Table 2. At this global level, then, the corpus distribution lends support to the intuition (however vaguely stated and left without evidence or argumentation) that *co*-RCs 'usually' express determination; put differently, the cautious formulation offered by some grammarians is closer to reality than any categorically stated prohibition on non-determinative functions (found in most accounts). However, it is very instructive to take a closer look at the distribution of the individual semantic types within each of the two broad categories, as summarized in Table 2. The columns represent the semantic types within the RC taxonomy, while the rows

itemize the RCs according to the grammatical function of the resumptive pronoun (including the null expression in the nominative) and the two special constructions exemplified above in (5-6).

	Total	I. Determinative					II. Non-determinative		
	789[5]	714					75		
		Cat.	Kind	Ident.	Charact.	Unique N	Explic.	Cont.	Other
	Totals	1	17	640	43	13	71	1	3
NOM	306	1	15	242	20	12	12	1	3
ACC	147		2	127	12	1	4		
DAT	24			18	3		3		
GEN	12			6	6				
INS	9			7	1		1		
LOC	5			4			1		
Quant.	50						50		
Temp.	236			236					

Table 2: Distribution of *co*-RCs according to function and semantic type[5]

First of all, it is evident that the totals are somewhat skewed if we include the two special and partially formulaic types, i.e., the quantifying and temporal RCs; they are each disproportionately frequent within their functional category. The temporal RCs are always of the Identification type and the quantitative RCs cannot be anything but a subtype of Explicative RCs. If we leave these two types out of the count, the relative frequency of non-determinative tokens goes up significantly from the overall 9:1 ratio; without the two special constructions, the occurrence of non-determinative RCs in the sample almost doubles, making up about 19% of all tokens. That alone undermines the traditionally held view that *co*-RCs are excluded from non-determinative functions.

At the same time, the distribution suggests an explanation for the traditional – and evidently inaccurate – analysis. The evidence that is typically offered in support of the analysis of *co*-RCs as purely determinative focuses solely on the Continuative type, and the corpus indeed confirms that this function is rather unexpected (although not completely unattested), as shown in Table 2. Instead, the vast majority of the non-determinative uses of *co*-RCs are found in the Explicative category whether we count the quantifying construction or

[5] This number includes only full nouns as the head N, not personal pronouns. Those will be discussed in Section 3.3.

not. An example of an Explicative *co*-RC is the second *co*-clause in (7): the proposition expressed in the RC simply elaborates on the description of Admiral Nelson, whose identity is already fully established without the RC. In contrast, the first *co*-RC in the example is a straightforward case of an Identification function, restricting the referential range of the phrase *ten pán* "the man".

(7) *Ten pán, co stojí nad vámi, je admirál Nelson, co porazil v roce*
that man CO stands above you is Admiral N. CO defeated in year
{1805 zlého Napoleona}
"The man that is standing above you is Admiral Nelson, who in 1805 defeated {the bad [guy] Napoleon}."

However, the existence of Explicative RCs, as a distinct semantic subtype of non-determinative relativization, has not been identified or acknowledged explicitly except in the analysis of *který*-RCs in Grepl / Karlík (1998) and its existence, let alone its specific properties, has not been considered anywhere in the context of *co*-RCs.

It is also worth noting that by far the most common usage of *co*-RCs centers on the Identification function, with Characterization being a distant second. This patterning is consistent with the traditional view that marking restrictiveness might be the core domain of the *co*-RCs, but yet again, the corpus provides tangible evidence that it is merely a tendency, however strong. Identification constitutes the focal point within a wider distributional range and thus cannot be presented in the form of a categorical 'rule' along the lines of, for example, Svoboda's (1972: 106) conclusions. In fact, one preliminary generalization we can draw from the correlations gathered in Table 2 is the following: there is a hierarchy of semantic preferences exhibited by the distribution of *co*-RCs in authentic discourse. Crucially, the hierarchy does not follow a clean determinative / non-determinative distinction as the traditional accounts suggest, but rather follows the finer semantic distinctions. The hierarchy appears to take the shape suggested in (8), which entails that the most common, typical candidate for the use of a *co*-RC is the context of identifying or otherwise describing specific individuals (around the middle portion of the taxonomy in Figure 1); the symbol '*' indicates that Continuative RCs are barely attested in the corpus. I will return to the significance of this hierarchy in Section 4.[6]

[6] Other interesting observations emerge from Table 2 as well, such as, for example, the overwhelming preference of *co*-RCs in which the head Ns serve the subject function inside the RC. Due to space limitations, I have to leave this aspect of the distribution aside for now.

(8) **Hierarchy of semantic preferences:**
identification (type I-A-3) > characterization (I-A-4) > explicative (II-A) > kind-of (I-A-2) > non-restr. determinative (I-B) > category (type I-A-1) > *continuative (II-B)

3.2 Correlation with deixis

To the extent that deixis has been addressed at all in the context of relativization, it has been noted that *co*-RCs are predominantly deictic (Svoboda 1967: 10, 1972: 105-106): their primary function is to point – in space, time, or discourse – to specific entities, thereby uniquely identifying (or individuating) the referent of the head N. A concrete manifestation of this relationship is the collocation of the head N with the demonstrative pronoun *TEN* "that", as illustrated in the introductory examples in (1);[7] this collocational pattern is also hypothesized to be the historical origin of the *co*-RCs (Svoboda 1967: 10). However, actual corpus data call for a substantially more nuanced analysis. For the sake of expediency, I will refer to the head Ns that are modified by *TEN* as 'deictic' and the head Ns without a demonstrative as 'non-deictic'.

First of all, the full sample splits down the middle between the deictic (396) and non-deictic (393) tokens; this alone contradicts Svoboda's assertion quite robustly. Moreover, the assumed dominance of deictic contexts in the distribution of *co*-RCs becomes even less convincing when we consider correlations between deixis and other criteria concerning the nature of the head Ns, namely, animacy and number. All three parameters – deixis, animacy, and number – are known to correlate with differences in degrees of referentiality or individuation and it is therefore relevant to examine how they interact in the context of the RCs as well, since they all can be expected to bear on the question of determination and restrictiveness.

Before we address these correlations, though, let us note that there is one domain in which the dominance of deictically marked head Ns appears to be confirmed. In the temporal RCs, the collocation with *TEN* is more than twice as likely as the use of a bare N: out of the total of 236 tokens, a demonstrative phrase as the head NP occurs in 159 cases, in contrast to 77 cases without *TEN*. Considering that the temporal usage of the *co*-RCs is often the only one that the

[7] The use of capital letters (*TEN*) is a typographical indication that I am only referring to a lexeme, without making explicit reference to its morphological shape, particularly the formal differences in gender and number.

existing accounts consider as accepted in the literary language (Trávníček 1951: 1165) and therefore on a par with *který*-RCs, it is not surprising that the temporal clauses may simply be taken as the only (or at least the primary) example of *co*-RCs. Nevertheless, even here the use of *TEN* is far from obligatory.

Let us now turn to the correlation between deixis and animacy. In order to remove any bias contributed by the two special constructions, i.e., the temporal RCs, which are overwhelmingly deictic and necessarily with inanimate head Ns only, and the quantifying RCs, which are necessarily non-deictic, I will exclude those tokens from the counts in the rest of this section. Deixis also seems irrelevant in the oblique grammatical functions (DAT, GEN, INS, LOC). The actual token frequencies of these forms are included in the counts and shown in Table 3, but their contribution to the analysis is marginal. After all, these case forms are rather rare to begin with compared to the direct cases (NOM, ACC), as we saw in Table 2.

Total	**Non-deictic**			**Deictic**		
503	Total	266 (=53%)		Total	237 (=47%)	
		animate	**inanimate**		**animate**	**inanimate**
		46%	**54%**		**56%**	44%
NOM	173	**61%**	39%	133	**78%**	22%
ACC	73	12%	**88%**	74	19%	**81%**
DAT	10	8	2	14	11	3
GEN	4	–	4	8	4	4
INS	4	–	4	5	–	5
LOC	2	–	2	3	–	3

Table 3: Deixis and animacy

The general pattern captured in the top portion of Table 3 shows two things: (i) the overall distribution favors non-deictic over deictic context, albeit not in a dramatic way (53% over 47%), and (ii) there is a general asymmetry between inanimate and animate head Ns: inanimate Ns are more frequent in the non-deictic contexts, while animate Ns outnumber inanimate Ns in the deictic contexts. The relative frequencies do not provide an overwhelming contrast but are sufficiently suggestive of the potential correlation between animacy and an explicitly marked determination. This potential comes into relief when considered in relation to the grammatical functions played by the head N in the RC.

Subjects (NOM) and indirect objects (DAT) generally attract animate referents more than inanimates, but it is interesting that in the nominative, the likeli-

hood of animate head Ns increases significantly in the deictic contexts (78%) compared to the non-deictic contexts (61%). This is particularly striking in light of the fact that in the actual number of tokens, animate head Ns are about equally distributed across deictic and non-deictic contexts (104 vs. 106, respectively). This asymmetry is not contradicted by the dative pattern and is further confirmed by the accusative pattern, where the vast majority of head Ns are inanimate entities (again, not surprisingly) but their distribution in deictic vs. non-deictic contexts displays a comparable correlation between animacy and deixis. Their distribution with respect to deixis is about even (64 tokens in non-deictic contexts vs. 60 in deictic ones), but inanimate non-deictic contexts are somewhat more likely (88%) than deictic ones (81%).

Given these patterns, we may explore further the hypothesis that the use of the demonstrative pronoun has to do with the degree of referentiality of the head N, rather than being an inherent property of the *co*-RCs. To test this possibility further, we can probe the distribution of grammatical number (singular vs. plural) as another relevant parameter. Overall, singular head Ns are more frequent than plural Ns (59% vs. 41%, respectively) and this distributional asymmetry becomes even more pronounced when we track the correlations with deixis and animacy. The relative frequencies are summarized in Table 4. We can see that there is about the same number of singular tokens in both non-deictic and deictic contexts (151 vs. 149), but the likelihood of a singular head N goes up in deictic contexts; the ratio singular : plural in the sample is roughly 5:3 (63% over 37%) in favor of the singular, while the difference between singular and plural in non-deictic contexts is less pronounced (57% over 43%).

Total	**Non-deictic**			**Deictic – *TEN***		
503	Total	266 (=53%)		Total	237 (=47%)	
		animate	inanimate		animate	inanimate
		123	143		133	104
Sg.	151 (57%)	51%	**61%**	**149** (63%)	**65%**	55%
Pl.	115 (43%)	49%	39%	**88** (37%)	35%	45%

Table 4: Distribution of the demonstrative *TEN* relative to animacy and number

The distributions in our sample thus suggest that the presence of the demonstrative cannot be attributed to the *co*-RCs as their inherent feature but, rather,

depends on the properties of the head N, particularly number and animacy. The prototypical constellation that attracts deixis appears to be a singular animate N. It also follows from the frequencies, though, that number ranks higher than animacy in determining preferential co-occurrence with *TEN*. We can propose a hierarchy of deictic contexts (i.e., the structure [*TEN* N, *co*]) as follows; note that animacy plays a role in the singular NPs but does not seem to make any difference in the plural (the numbers are percentages of a given configuration in [*TEN* N] occurrences):

(9) head N = Anim. sg > Inanim. sg > (Anim. pl, Inanim. pl)
37% 23% 20% 20%

The correlations in (9) are consistent with treating the usage of *co*-RCs as an issue of individuation or high referentiality, rather than simply deixis. The preferred head N tends to be a highly individuated / referential entity, at the expense of less individuated / referential ones.

3.3 Head Ns with unique reference

Perhaps the least explored domain within the proposed taxonomy of Czech relativization are the segments in the middle, at the hypothesized boundary between determinative non-restrictive clauses (I-B) and the non-determinative Explicative clauses (II-A). Both of these segments involve the same type of head N (nominals with unique reference) and the crucial difference between them is the obligatory presence of *TEN* with the former and obligatory absence of *TEN* with the latter. It is the demonstrative that contributes the determinative function (I-B), thereby invoking an interpretation that is based on some sort of contrast, whether explicitly stated or just implied; in the absence of the demonstrative (II-A), no contrastive reading is available. We thus obtain the interpretive distinction between (7) above and (10) below. While in (7), the communicative objective is to offer further commentary about Admiral Nelson, in (10), the speaker's goal is to establish the identity of a guy named Vantoch:

(10) *nejste vy ten Vantoch, co se se mnou v Jevíčku prával,*
aren't 1PL.NOM that V. CO RF with me in J. fight.PST.SG.M
{*když jsme byli kluci*?}
"are you that [guy] Vantoch, who used to have fights with me in Jevíčko,
{when we were little boys}?"

On the one hand, the use of the person's last name suggests unique reference and thus complete identification. The context, however, places this person in contrast to other schoolmates among which the speaker is trying to single out just one, by offering a description that might set Vantoch apart from other potential candidates. We cannot classify the reading as restrictive (there is only one person named Vantoch that the speaker went to school with), but the demonstrative creates a distinctly different setting from the bare noun structures illustrated in (7); in (10), the RC is relevant for the head N's precise identification.

If we take seriously the blanket prohibition on non-determinative usage, discussed in Section 3.2, we should expect no attestations of the kind in either (7) or (10); Svoboda (1967: 7, 1972: 106) states this condition directly. In reality, we find both, as has already been noted and quantified in Table 1, although not in any overwhelming numbers (25 tokens in the sample). The existing accounts thus overstate the case by making a categorical judgment, but the basic insight about the limited compatibility of *co*-RCs with unique-reference Ns is on the right track. The deictic usage (I-B, example 10) is essentially consistent with the patterning discussed in the preceding section in that the majority of tokens involve proper nouns denoting human individuals (i.e., animate singular). As expected, the deictic usage of unique-reference Ns outnumbers the non-deictic usage, but only at the ratio of about 3:2, which indicates that non-deictic usage (specifically the Explicative) is not only possible but is not even all that exotic within the domain of unique-reference head Ns. Overall, then, the corpus contradicts both of the two general claims: the requirement of deixis on the head N as an inherent feature of *co*-RCs and the expectation that *co*-RCs cannot serve non-determinative functions or co-occur with *TEN*.

Sorting out the issue of unique reference also extends to one particular subtype of head nominals, namely, personal pronouns. The full range of such pronouns can occur in the agreeing RCs with the relative pronoun *který* 'which', e.g., the structures *já, který* "I who", *my, kteří* "we who", etc. It follows from the assumptions about *co*-RCs being necessarily determinative that *co*-RCs cannot be headed by pronouns that necessarily mark unique reference, such as *já* "I" and *ty* "you-sg." (yielding **já, co* "I who" or **ty, co* "you-sg. who"). The reasoning, explicated in Svoboda (1967: 6) goes as follows: the speaker and the hearer are fully and uniquely identified by the pronoun itself and cannot, therefore, be modified ('determined') by an RC whose semantic range is limited to indicating restrictiveness or at least determination. If we

take the speaker as an example, it should not be possible to restrict reference to a specific ego in contrast to the same ego. The presence / absence of unique-reference personal pronouns thus, again, speaks to the issue of (non)restrictiveness and (non)determinativeness.

The sample contains 90 unambiguous tokens of personal pronouns as the head N of *co*-RCs and the distribution of the 1st and 2nd pers. sg. confirms Svoboda's insight that the singular pronouns *já* "I" and *ty* "you-sg." are incompatible with the determinative function: they are very rare in the sample (altogether only six tokens among all the personal pronouns) and only one of those, found in a dialog of a theatrical play and shown in (11), can be classified as helping establish the referent's identity.

(11)

{FANKA:	*To sou voni, milostpane?*				
LOUPEŽNÍK:	*Ne, to jsem já, Fany.*}				
FANKA:	*Kterej já?*				
LOUPEŽNÍK:	*Já, co*	*tu*	*byl*	*ráno.*	
	I [CO]	here	was	morning	
FANKA:	*Ten zabitej? To už běhaj?*				
{FANKA:	'Is that you, sir?				
LOUPEŽNÍK:	'No, Fanny, it's me.'}				
FANKA:	'Which me?'				
LOUPEŽNÍK:	'I who was here this morning.'				
FANKA:	'The dead one? You're on your feet again?'				

The utterance in line 3 explicitly presents a setting that presupposes multiple referents, by posing the question "which [one] I?"; but the full context is also conducive to this shift since Fanka is evidently faced with the task of choosing between two distinct individuals: one that she expects and addresses in the first line (her master), and another, who shows up, unexpectedly, instead (the master's daughter's young admirer). Note also that the contrastive context, necessary for the determinative reading, is not marked explicitly by anything in the sentence itself (e.g., by using a demonstrative, as was the case in (10) above) but merely follows from the broader context. This observation further supports the generalization that the determinative function of *co*-RCs need not be encoded directly as an inherent feature of these RCs.

Aside from this clearly shifted reading, however, the remaining tokens, exemplified in (12) below, are all cases of non-determinative usage. They all fit the Explicative category, in simply adding an informative comment about the

speaker or hearer, with no identificational relevance; in (12) the RC actually suggests the flavor of a *because*-clause: not just "... I who needs it more" but "... I, since I need it more".

(12) *proč jsem nevyhrála já, co to víc potřebuji?*
why AUX.1SG NEG.win.PST.SG.F 1SG.NOM CO it more need.PRES.1SG
"Why wasn't the winner me, who needs it more?"

Overall, the corpus confirms that determinative readings are quite marginal with the 1st and 2nd pers. sg. pronouns, but does not substantiate any absolute prohibition on non-determinative usage of the *co*-RCs. When these pronouns do appear they of course have to be non-determinative, which follows from their inherent nature as unique-reference nominals.

4. Functional and semantic range of absolutive relativization in Czech

Based on the attested frequencies in the corpus sample, certain properties emerge that can be seen as *prototypically* associated with *co*-RCs; they are listed in (13):

(13) **Prototypical features of *co*-RCc**

Function:	determinative restrictive
Semantics:	individuation of head referent
Syntax:	relativized N is the subject in RC
Head N:	concrete, animate, singular entity

The functional and semantic features may appear, on the whole, to conform to the traditionally posited constraints. There is one important difference, though: the corpus shows them to be mere tendencies to start with and a closer look at the specific semantic subtypes helps us piece together a much more nuanced picture that leads to a deeper understanding of the nature of this relativization strategy.

Let us now recall the proposed taxonomy of relativization in Figure 1, which reflects the current state of knowledge in Czech grammatical literature and which we took as the starting point for our analysis. However, rather than a strict taxonomy with discrete boundaries, we can view the diagram as delimiting a particular functional or conceptual space (in the spirit of typological se-

mantic maps, e.g., Croft / Shyldkrot / Kemmer 1987; Haspelmath 1997, 2003; Croft 2001 or constructional maps that have been proposed for the purpose of capturing grammatical patterning in a single language, e.g., Fried 2005, 2009) within which attested meanings of RCs can be coherently organized. Such a space presupposes fluid transitions between individual nodes, which is also more consistent with the often observed difficulty in classifying individual tokens as categorically belonging to one type or another, particularly across the gray domain in the middle.

Figure 1 represents the space that is fully covered by the agreeing *který*-RCs and if we were to incorporate the existing accounts of the *co*-RCs, it would amount to essentially admitting *co*-RCs as coinciding with the determinative node (type I) and all its subtypes (with some disagreement left open concerning subtype I-B) and as being excluded from the non-determinative node (type II) and its subtypes. However, if we map the corpus distribution onto this space, we can not only establish points of similarity and dissimilarity in relation to the agreeing *který*-relativization, but can also begin to articulate an empirically grounded and descriptively much more accurate account of the *co*-RCs as a distinct grammatical pattern. Figure 2 summarizes our findings in a preliminary representation of the relevant conceptual space; the dashed-line oval delimits the core domain in which *co*-RCs are attested.

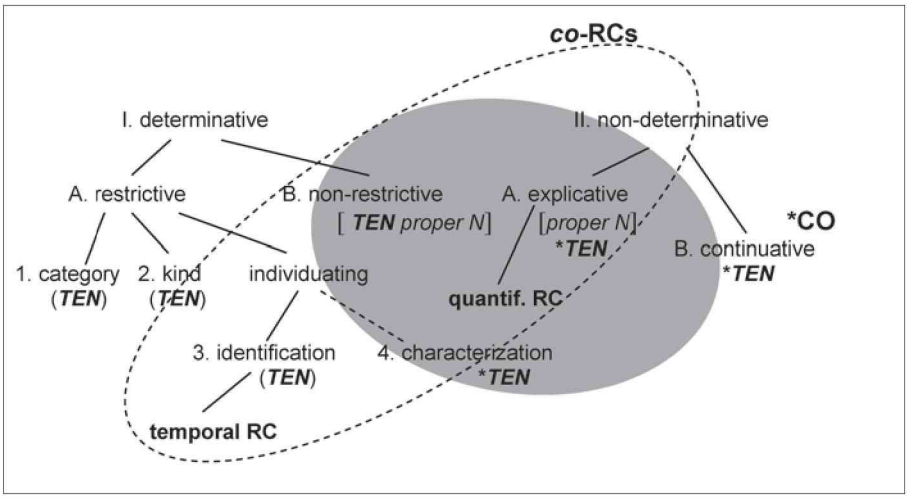

Figure 2: Distribution of absolutive RCs in the corpus

The map in Figure 2 captures several important (and newly established) facts about the nature of the *co*-RCs. In general, their primary function is centered on individuating the referent of the head N in a presupposed contrastive context; the individuation may lead (and overwhelming does so) to pure identificational meaning, but need not (yielding a less definite characterization meaning instead). This general functional preference remains very strong and manifests itself in several ways:

i) The focal point of the functional range is the Identification function (type I-A-3), within which the temporal RCs occupy a prominent position of a highly entrenched, semantically distinct, and formally partially formulaic subtype of identificational RCs.

ii) In contrast, *co*-RCs are very rare in the remaining, less central restrictive functions (determining kinds and categories or class membership). This may not be a very surprising outcome; establishing generic reference or class membership involves diminished individuation and/or a lower degree of referentiality, which is conceptually incompatible with the preferentially individuating function of *co*-RCs.

iii) To the extent that *co*-RCs extend out of the identificational range, they cover primarily the domain of unique-reference head Ns, whether determinative or non-determinative. The latter, moreover, includes a special construction (quantifying RCs) which represents a formally and semantically distinct, well-entrenched, and partially formulaic subtype of the Explicative RCs. The pull toward the unique-reference head Ns can again be motivated by the general affinity toward individuation: these head Ns, by definition, mark highly individuated, highly referential entities.

iv) The individuating character of the *co*-RCs correlates with certain semantic properties of the head Ns: the corpus demonstrates a preference for singular animate entities. Consequently, and contrary to certain existing analyses, these RCs are much less dependent on the presence of demonstrative pronouns to ensure an individuating interpretation. The most we can conclude about explicitly marked deixis is that inherently highly individuating nouns (singular, animate) show stronger co-occurrence patterns with deixis than less individuated ones (plural, inanimate).

v) Finally, we can hypothesize that the *co*-RCs, by pointing to highly individuated entities, form a relatively tight conceptual unit with their head Ns, which they either help identify or at least add some contextually salient information about them. This hypothesis can easily accommodate the Explicative non-determinative usage as well: the relationship between the RC and the head N in this pattern is reminiscent of a subordinating relation – one in which the embedded clause bears signs of conceptual dependence on the head N. In contrast, the same conceptual closeness cannot be expected in the essentially coordinating relation characteristic of the Continuative readings of relativization patterns (type II-B) since the relationship between the head N and the RC is very loose here; the two clauses (main clause and RC) express two conceptually independent propositions, just like other, formally explicit, coordinating structures.

It is perhaps also worth noting that the space in which the *co*-RCs most commonly operate coincides with the 'gray' area in the middle of the map, namely, the part that is generally considered the fuzziest domain, with the least distinct boundaries, which tend to be most dependent on actual discourse context. I hope to have shown that the fuzziness may become much less intractable with the use of corpus data, through which semantic and contextual aspects of grammatical patterning can be readily available and aid in accurate analysis.

5. Conclusions

The goal of this study was to explore the potential of integrating qualitative analysis with frequency-based evidence provided by an electronic corpus, confronted with grammatical descriptions that have been formulated without the use of any large corpora. The observations and results reported in the case study concerning Czech absolutive RCs should be taken as no more than the very first step in a more thorough investigation of this particular grammatical pattern, which has not yet received a truly systematic and comprehensive treatment. However, certain partial generalizations emerge, including potential implications for the study of RCs beyond just the Czech patterns.

In order to fully understand the use and distribution of the *co*-RCs, also in contrast to the formally agreeing relativization pattern, we must take into account finer semantic and contextual distinctions than traditionally applied.

These RCs appear to form a distinct cluster of relativization functions and meanings that all have to do with individuating the head N. The corpus material suggests that while the core domain of the *co*-RCs resides in identification functions, the clauses have spread into other, non-restrictive and non-determinative functions well beyond what traditional analyses admit possible. At the same time, the spread into the non-determinative territory is not random, but follows a conceptually coherent path within a relativization network that organizes all the attested functions and meanings of Czech relative clauses. The path, moreover, suggests a particular direction in the development of RCs, namely, gradual erosion of restrictiveness as a linguistically explicitly marked distinction. While the Czech *co*-RCs can be considered preferentially (though by no means universally) restrictive, the corpus reveals quite clearly that the absolutive relativizer *co* by itself cannot be taken as a reliable marker of restrictiveness. The diachronic dimension of this spread and the details of the *co*-RC development from the hypothesized deictic origins to what the synchronic corpus documents will require much more research. Nevertheless, even this preliminary analysis has some value for broader theoretical and typological studies concerning the status of restrictiveness as a relevant notion in classifying the inventory of RCs. The Czech facts appear to confirm the cross-linguistic observation that restrictiveness is not a highly salient linguistic category that requires explicit marking and, therefore, should not be used as a fundamentally important criterion for analyzing RCs. Instead, the salient notions, which would deserve further testing in other languages as well, seem to include the referential type of the relativized noun, the interaction between relativization and deixis, and the semantic relationship between the head N and the proposition expressed by the RC.

Finally, the present work also shows that the use of large electronic corpora enriches grammatical descriptions in several respects. Corpus material serves as an important source of semantic and contextual information, which turns out to be crucial in categorizing and explaining grammatical forms; forces us to acknowledge and directly address the dynamic nature of language; helps identify specific usage-based factors that affect variability in linguistic 'rules' and categorization; and offers greater reliability of quantificational evidence, provided we exercise a necessary dose of skepticism about its infallibility and apply adequate controls. It is clear that on the basis of corpus evidence, we can arrive not only at sufficiently dynamic, multi-faceted, and, hence, more accu-

rate generalizations about a given form itself, but also capture subtle shifts in its distribution, depending on specific, well-defined criteria. The use of corpus material has the potential of bringing the grammarian's work to a new and more realistic level of analysis.

References

Croft, William (2001): Radical Construction Grammar. Syntactic theory in typological perspective. Oxford: Oxford University Press.

Croft, William / Shyldkrot, Hava Bat-Zeev / Kemmer, Susanne (1987): Diachronic semantic processes in the middle voice. In: Giacalone Ramat, A. / Carruba, Onofrio / Bernini, Giuliano (eds.): Papers from the Seventh International Conference on Historical Linguistics. Amsterdam / Philadelphia: Benjamins, 179-192.

Fried, Mirjam (2005): Constructing grammatical meaning: isomorphism and polysemy in Czech reflexivization. In: Studies in language 31, 4: 721-764.

Fried, Mirjam (2009): Plain vs. situated possession in a network of grammatical constructions. In: McGregor, William (ed.): Expression of possession. Berlin: de Gruyter, 213-248.

Fried, Mirjam (in press): Vztažné věty s nesklonným *co*. In: Štícha, František (ed.): Kapitoly z české gramatiky. Prague: Academia.

Grepl, Miroslav / Karlík, Petr (1998): Skladba češtiny. Olomouc: Votobia.

Haspelmath, Martin (1997): Indefinite pronouns. Oxford: Clarendon Press.

Haspelmath, Martin (2003): The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In: Tomasello, Michael (ed.): The new psychology of language. Mahwah, NJ: Lawrence Erlbaum Publishers, 155-175.

Kopečný, František (1962): Základy české skladby. Praha.

Lešnerová, Šárka / Karel, Oliva (2003): Česká vztažná souvětí s nestandardní strukturou. In: Slovo a slovesnost LXIV, 4: 241-252.

Mluvnice češtiny III – Skladba (1987). Praha: Academia.

Poldauf, Ivan (1955): Vztažné věty v angličtině a v češtině. In: Sborník VŠP, Jazyk a literatura II: 159-194.

Svoboda, Karel (1967): Vztažné věty s nesklonným *co*. In: Naše řeč 50,1: 1-12.

Svoboda, Karel (1972): Souvětí spisovné češtiny. (= Acta Universitatis Carolinae, Philologica XLIII). Praha: Universita Karlova.

Šmilauer, Vladimír (1972): Nauka o českém jazyku. Praha: SPN.

Trávníček, František (1951): Mluvnice spisovné češtiny II – Skladba. Praha: Slovanské nakladatelství.

Zubatý, Josef (1918): Jenž, který, kdo, co atp. In: Naše řeč II, 37: 37-44.

Source of data

ČNK – SYN2000: Czech National Corpus (ČNK). Internet: http://www.korpus.cz. Ústav Českého národního korpusu FF UK, Praha 2000.