Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank

Anna Nedoluzhko
Faculty of Mathematics and Physics, Charles University in Prague
nedoluzko@ufal.mff.cuni.cz

Abstract

This paper discusses the problem of annotating coreference relations with generic expressions in a large scale corpus. We present and analyze some existing theories of genericity, compare them to the approaches to generics that are used in the state-of-the-art coreference annotation guidelines and discuss how coreference of generic expressions is processed in the manual annotation of the Prague Dependency Treebank. After analyzing some typical problematic issues we propose some partial solutions that can be used to enhance the quality and consistency of the annotation.

1 Introduction

One of the most problematic issues of annotating coreference in large scale corpora is processing coreference of generic expressions. The decision to annotate generic noun phrases produces a significant decrease of inter-annotator agreement. On the other hand, neglecting coreference relations between generic expressions causes a significant loss of information on the text coherence that is primordially the reason for annotating coreference relations at all. It also causes the inconsistency of annotation guidelines: due to relatively vague definition of genericity, it is almost impossible to exclude *all* coreference relations between generics from the annotation.

In the Prague Dependency Treebank (henceforth PDT), we tried to distinguish coreference relations between nominal expressions with specific and generic reading. Comparing the inter-annotator agreement for these groups shows that the agreement for noun coreference with specific reading is significantly higher than the agreement for the coreference of generic NPs (F1-measure 0.705 for specific NPs and 0.492 for generics[1]). Moreover, the manual analysis of the cases of disagreement of specific NPs coreference demonstrates that most cases of disagreement are those where NPs in question may be interpreted generically.

Having formulated a set of criteria which help identifying generic expressions, there still remains a wide range of typical examples which can have generic interpretation, though not necessarily. In this paper, we try to delimit the set of generic NPs presenting the overview of some existing theories of genericity (Sections 2 and 3.1) and compare them to the stand-of-the-art coreference annotation guidelines (Section 3.2). Then we present our approach to annotating coreference with generic noun phrases in PDT where we apply the presented theories to coreference and bridging relations annotation (Section 4). We analyze typical problematic issues (Section 5) and discuss some possible solutions (Section 6).

2 What are generics and can they corefer?

Generic reference is a term commonly used in linguistic semantics to describe noun-phrase reference to kinds of things (Carlson 2005). In different languages, generic reference may be expressed by noun phrases with definite and indefinite articles and with determinerless expressions quite generally. In languages without articles, the determinerless form is typically used (Carlson 2005, Hlavsa 1975; Padučeva 1985, etc.).

[1] F1-measure for generics is closer to inter-annotator agreement for bridging relations (0.460 for all annotated data).

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 103–111,
Sofia, Bulgaria, August 8-9, 2013.

Compare some typical examples for generic noun reference (different uses of *a/the dog(s)*) in English, German and Czech:

English: *Dogs bark – The dog has evolved from the Jackal – A dog knows when it is time for his walk*[2].

German: *Hunde beißen*. *Der Hund stammt vom Schakal ab*. *Ein Hund weiß// Hunde wissen, wenn es Zeit für seinen Spaziergang ist.*

Czech (non-article language): *Psi štěkají. – Pes je šelma*.

The examples above demonstrate that generic noun phrases cannot be recognized by their forms (this fact was pointed out in Lyons 1999, Carlson 2005, etc.). While in English the plural form of the definite can only marginally have generic reference, in German, which is closely related to English, the plural definite may imply generic reference quite easily. In Romance languages, the form of bare plural with generics is prohibited (Delfitto 2006) and even in languages without articles, generics with determiners are not so rare (see e.g. common examples with Czech in Nedoluzhko 2003)[3]. This leads to a suggestion that genericity is not a primitive category of semantic or syntactic description.

Theoretical studies like Carlson (1980) appeal to typical examples with noun phrases referring to specific objects. A discussion on his approach (Paducheva 1985, Delfitto 2006, Lyons 1999) concerns theoretical issues that are analyzed in similar typical cases.

When analyzing real corpus examples we encounter a lot of cases indicating that not all generic expressions are generic in the same way. Problems with processing generic expressions arise also from the lack of a universally accepted theory of genericity which would be applicable to the real texts analysis.

Generic reading is possible not only with referring nouns, but also with mass nouns, group nouns, abstract nouns, quantifiers and deverbatives. Look at the example (1). Everyone should probably agree that *the homeless* is a generic expression, but is the same true about *the homeless population*?

(1) *Your comments implied we had discovered that the principal cause of homelessness is to be found in the large numbers of mentally ill and substance-abusing people in the homeless population. [...] The study shows that nearly 40% of the homeless population is made up of women and children and that only 25% of the homeless exhibits some combination of drug, alcohol and mental problems*[4].

Another relevant question is if generic expressions referring to the same kind can be considered coreferent in the same sense as noun phrases with a specific reading. According to Carlson's (1980) and Lyons' (1999) claim, generics refer to classes in the similar way as proper names refer to unique entities. In this sense, coreference of generic expressions appears to be obvious. On the other hand, Carlson's observations seem to be quite language-specific. Arguing against a quantificational analysis of bare plurals with generic meaning, he claims that the sentence *Miles wants to meet policemen* cannot be assigned a reading according to which "there are certain policemen that Miles wants to meet," whereas this interpretation is naturally available in the case of *Miles wants to meet some policemen*. This is not the case of languages without articles where plural forms can be assigned any reading regardless of the use of the quantifier[5]. Generally, we suppose that quantificational (or predicative) interpretation of generic expressions in different languages is not impossible (see for example almost obligatory predicative reading of *Czech exporters in* (7)). However, this fact does not necessarily exclude the coreference relation between them. Eventually, the discourse deixis as reference to events is also often considered and annotated as coreference.

3 Recent research on generics

We believe that it would not be a strong exaggeration to claim that theoretical and computational linguistics have different goals as concerns their approach to genericity. The challenge of linguistic research is to find out more about the essence of genericity. The aim of annotating is to

[2] However, Carlson – Pelletier (1995) do not consider *a dog* in the last sentence to be generic, because it cannot be combined with kind-level predicates.

[3] It may be possible to determine generics in sentences with so-called "kind-level predicates" (Carlson 2005), they interact with aspectual distinctions in verbs (Lyons 1999) etc, but these approaches are not applicable to real-text data.

[4] The example comes from the Prague English Dependency Treebank (PEDT, Hajič et al. 2009)

[5] Actually, even in English not all bare plurals should necessarily refer to kinds. In modern journalistic texts, the tendency to omit articles appears to be quite strong.

make the group of generics as clear as possible, in order to reach higher agreement and better results of automatic processing.

It is also generally known that the features of an annotation must be adapted to the task it is designed for. However, the existing large-scale annotated corpora (especially those prepared on university basis) are often meant to be multi-purpose. They serve both as train data for (different!) automatic tasks and as a rich manually annotated material for linguistic research.

In what follows, we complete the theoretical overview (started in section 2), present the annotation approach and look for the common points.

3.1 Linguistic research

There is a rich variety of linguistic approaches to genericity. Even as concerns the terminology with generics, it is quite inconsistent and cannot be relied on with much certainty. According to different researchers, generic NPs are considered to be either referring to classes (Carlson – Pelletier 1995, Mendoza 2004) or non-referring (rather predicating) classifications over kinds (Paducheva 1985), beeing able to have specific and non-specific interpretation (Mendoza 2004, Smelev 1996) and divided from non-specific NPs as a separate group (Carlson – Pelletier 2005, Paducheva 1985).

Carlson (1980) represents the most influential approach to genericity that has been elaborated in the framework of formal semantics and generative grammar. Calson's hypothesis is that generics are kind-referring expressions, roughly names for kinds, as opposed to individual-referring expressions that refer to individuals or groups of individuals. In his approach, there is a difference between generic reference and individual non-specific reference, i.e. reference to an open set of individual objects. For example, NP *lions that have toothaches* is not generic, its reference is individual (i.e. non-generic) and non-specific, which can be demonstrated by the fact that it cannot be substituted by the definite NP *the lion that has toothache* (such NP can have only individual reading). However, the problem with this criterion is that it is clearly language-specific (it cannot be applied at all to Czech, for instance).

3.2 Annotation coreference with generic expression

Let's now have a look on how generic NPs are processed in annotation projects with anaphoric and coreference annotation.

In some projects, e.g. ARRAU and other corpora based on the MATE coreference annotation scheme (Poesio 2004), genericity is marked as a part of lexico-semantical information of the noun (an attribute `generic-yes/no/undersp` is applied to each noun). This information is contemplated in the annotation of identical coreference. Identical coreference for generics is also annotated in AnCora (Recasens 2010) and PDT (Nedoluzhko 2011).

In other projects, annotation of coreference with generic NPs may be excluded from annotation schemes that are geared towards a reliable annotation of large text quantities. For example, generics are not annotated for coreference in Ontonotes (Pradhan et al. 2007), TüBA-DZ (Hinrichs et al. 2004) and PoCoS (Krasavina-Chiarchos 2007).

However, even if an annotation scheme explicitly says that coreference of generic NPs is not annotated, there are some borderline cases where coreference can still be annotated quite systematically. So, TüBA annotates coreference with the nominal expression if it appears repeatedly in the text with the same interpretation. In Ontonotes, the explicit anaphora with *it* in the anaphoric position is commonly annotated for coreference:

> (2) *Still, <u>any change in East Germany</u> has enormous implications, for both East and West. <u>It</u> raises the long-cherished hopes of many Germans for reunification*[6].

Furthermore, systematic exclusion of generic expressions from the annotation will force the coders not to mark the cases like (3) and (4)[7]. From the point of view of applied tasks and automatic coreference resolvers it will lead to the loss of relevant information and to an essential complification of automatic tools.

> (3) *The sterilizing gene is expressed just before the pollen is about to develop and it deactivates the anthers of every flower in <u>the plant</u>. Mr. Leemans said this genetic manipulation doesn't hurt the growth <u>of that plant</u>.*

> (4) *A workshop needs to be planned carefully. Otherwise <u>it</u> may turn in a disaster*.

As far as we know, there are no significant projects for annotating coreference separately for

[6] This example is taken from PEDT, to which the Ontonotes coreference was applied.

[7] Examples come from PEDT.

generic, unspecific non-generic and specific expressions.

4 Coreference annotation in Prague Dependency Treebank

In this section we describe how generic expressions (or more precisely, what we decided to consider generic expressions) are annotated in the Prague Dependency Treebank.

Annotation of coreference and discourse relations is a project related to the Prague Dependency Treebank 2.5 (PDT; Bejček et al. 2011). It represents a new manually annotated layer of language description, above the existing layers of the PDT (morphology, analytic syntax and tectogrammatics) and it captures linguistic phenomena from the perspective of discourse structure and coherence. This special layer of the treebank consists of annotation of nominal coreference and bridging relations (Nedoluzhko et al. 2009), discourse connectives, discourse units linked by them and semantic relations between these units (Mladová 2011).

Considering the fact that Czech has no definite article (hence no formal possibility to exclude non-anaphoric coreference), our annotation is aimed at coreference relations regardless to their anaphoricity.

Coreference relations are marked for noun phrases with specific and generic reference separately – coreference of specific noun phrases – type SPEC, coreference of generic noun phrases – type GEN[8]. Bridging relations, which mark some semantic relations between non-coreferential entities, are also annotated in PDT. The following types of bridging relations are distinguished: PART-OF (e.g. *room - ceiling*), SUBSET (*students - some students*) and FUNCT (*state - president*) traditional relations, CONTRAST for coherence relevant discourse opposites (*this year - last year*), ANAF for explicitly anaphoric relations without coreference or one of the semantic relations mentioned above (*rainbow - that word*) and the further underspecified group REST[9].

As seen from the point of view of the annotated groups, generic NPs are explicitly marked only with the second element of the coreference relation. However, this distinction remains unregistered by bridging relations. Moreover, it appears to be possible (and even not so uncommon) that a coreference relation was annotated between a generic and a non-generic noun phrase. These cases are interpreted as either (linguistically) ambiguous or insufficiently classified by the guidelines. For example, in (5), the specific noun phrase *tento národ* (=*this nation*) is coreferent with generic plural *Romy* (=*the Gipsies*):

(5) *Nic z toho se však nevyrovná míře neštěstí, které Romy postihlo v letech druhé světové války. Spolu se Židy byli označeni za méněcennou rasu a stali se objektem patologických fašistických opatření, jejichž cílem byla úplná genocida tohoto národa*. (= *Nothing of this, however, compares to the misfortune that befell the Gipsies during the Second World War. Together with the Jews, they were called an inferior race and became the object of pathological fascist measures, their purpose being the complete genocide of the nation*.)

Annotation rules for generics in PDT are described in detail in sections 4.1-4.3.

4.1 Type coreference of generic NPs

Coreference relations between the same types are annotated as coreference of generic NPs (attribute `coref_text, type GEN`). Cf. (6) where antecedent generic *drug* is pronominalized in the anaphoric position:

(6) *Droga je tedy tak účinná, že ten, kdo ji užívá, se snadno dostane do „pohody" kouřením nebo šňupáním*. (= *The drug is so effective that the person who takes it can easily achieve the state of "coolness" by smoking or snorting*.)

The "generic coreference" is more frequent for plural forms (7):

(7) *Nová striktní omezení vlády SR proti českým exportérům. Již několik dnů je všeobecně známo, že ochranářská opatření slovenské vlády proti českým exportérům se dotýkají zejména oblasti obchodu s potravinami a zemědělskými produkty*. (= *The new Slovak government's strict restrictions on Czech exporters. It's commonly known for several days that protective measures of Slovakia's government against Czech exporters apply mostly to the trade of food and agricultural products*.)

[8] The reason for this decision is the lack of semantic information assigned to nouns themselves, as it is done e.g. for Gnome in MATE sceme (Poesio 2004).

[9] For detailed classification of identity coreference and bridging relations used in PDT, see e.g. Nedoluzhko et al. 2011.

Textual coreference of type GEN is also annotated for the majority of abstract nouns (see more detail in Section 5.5), cf. (8):

(8) *Tímto faktorem je podnikatel-inovátor, který se snaží o zisk, a proto logicky nemůže existovat ve stavu statiky, která nezná ani zisk, ani ztrátu.* (= *This factor is the enterpreneur-innovator, who is trying to gain profit, and hence, logically, cannot exist in a static state, where there is no profit or loss.*)

4.2. Classes and subclasses

The relation "category – sub-category" is marked as a bridging relation of the SUBSET type. Cf. (9).

(9) *I když konzervativní Anglie jeho čin odsoudila, ... Británie se pro žvýkačku stala bránou do Evropy. Ještě jeden milník si zaslouží zmínku – zrod bublinové žvýkačky* (= *Although conservative England did not accept it, ... for the gum, Britain has become the gateway to Europe. Another milestone is worth mentioning, that is the birth of a bubble gum.)*

Annotating the SUBSET relation with generic expressions appears to be quite a serious problem. This relation has a different meaning compared to the SUBSET relation of noun phrases with specific reading. However, such relations may be quite relevant for cohesion.

4.3 The relation "type – entity"

If a specific mention is used in the text after a generic mention (or the contrary), the relation between them is annotated as a bridging relation of the SUBSET type. Cf. (10):

(10) *Nový VW Golf je vybaven motorem o síle... Dostali jsme možnost se novým golfem projet.* (= *The new VW Golf is equipped with an engine power ... We had an opportunity to ride a new golf.*)

Similar, but not the same is the relation between a set of specific objects and a non-specific element in (11):

(11) *[volontéři] Absolvovali školení v první pomoci pro člověka v nouzi . […]Když dítě zavolá, dostane buď radu hned, nebo si s ním volontér domluví další hovor.* (=*The volunteers have been trained in first aid for people in need. [...] When a child calls, it will get get an advice immediately, or a volunteer will arrange a meeting with him.*)

5 Problem cases with generics in PDT

Although the cases presented in sections 4.1-4.3 do not look very reliable, they are still considered to be relatively clear as compared to what follows in 5.1 -5.6. The decisions made in annotation guidelines for these cases are often case-sensitive, might be in some cases contra-intuitive, and they result in high inter-annotator disagreement.

5.1 Non-generic non-specific NPs

In case of non-generic non-specific noun phrases, when antecedent and anaphoric noun phrases have the same t-lemmas and the same scope, but anaphoric NP does not have a determiner, coreference of type GEN is annotated. Although this kind of relation does not contribute much to text coherence, we still tend to mark this relation, also for the reason that the border between what should be annotated and what should not is not always easy to determine.

(12) *Když si dítě bude přát, aby se o jeho problému nikdo z rodiny nebo školy nedozvěděl, musíme to respektovat, vysvětluje Jana Drtilová . […] Většinou se stává, že dítě ani nechce, aby se rodina dozvěděla, že se nám ozval. Linka by neměla rodinu nahrazovat, ale doplňovat*. (= *If a child desires that no one from the family or school would find out about his problems, we have to respect that, says Jana Drtilova. […] It is usually the case that the child does not even want for the family to know that he contacted us. The hotline should not replace the family, but to supplement it.*)

There are also cases of non-specific non-generic NPs the referential value of which is provided by syntactic factors. These are so-called contexts with removed assertiveness, e.g. sentences with modal verbs (*can, want, need*), imperative sentences, future tense, questions, negations, disjunctions, irreality, uncertainty and so on. Non-specific NPs are often used with performative verbs, propositional attitudes (*want, think, consider*) and some constructions as e.g. in English *such as*, in Czech *jde o* (=lit. *It is about*), *takový X* (=*such* X), etc. These contexts can give a non-specific reading to an expression, even if it actually has a specific meaning. Cf (13), where

(13) *Ale jedna věc je jistá - palác bude stavebně předáván letos na podzim.* […] *Provoz tak obrovské budovy přijde ročně na desítky milionů korun*. (=lit. *But one thing is certain – the reconstruction of the palace will be finished this fall*. [...] *It will cost tens of millions crowns, to run such a huge building*.)

5.2 Borderline cases between coreference of specific and generic NPs

In some cases, it is hard to decide if a noun phrase has a specific or a generic reading. Mostly, both interpretations are possible. There are no firm rules for an unambiguous assignment of the types in those cases; the type is chosen on the basis of the available context and the annotator's consideration. Uncertainty of the choice between generic and specific reference is common with some typical groups of noun phrases, first of all with those that have or may have modifications. Cf. *pořad* (=*TV show*) in (14) that may have a temporal modification. The obligatoriness of this modification influences the annotator's decision if (s)he should read it as a generic or a specific NP. For this case, the specific reading was chosen.

(14) *K tématu pořadu TV NOVA TABU "Zrak za bílou hůl" byl přizván ke konzultaci Oldřich Čálek. Kateřina Hamrová, dramaturgyně pořadu, TV NOVA.* (= *To consult the topic of the TV NOVA show TABU "Vision for a white cane", Ulrich Čálek was invited. Catherine Hamrová, the dramatist of the show, TV NOVA*)

Also, for example for (15), *the detergent Toto* can be understood as a specific (a name for a detergent brand) or generic (the type of the detergent of such brand). Also in this case, the specific reference is preferred in PDT:

(15) *U detergentu Toto jsme například řešili problém s udržením stálé kvality, protože jednotlivé partie byly nevyvážené. Investovali jsme dva miliony korun do nákupu pásových vah, zpřesnili dávkování a jakost pracího prášku stabilizovali*. (=*For example, with the Toto detergent we face problems with maintaining consistent quality... We invested two million crowns... and stabilized the quality of the detergent.*)

5.3 Borderline cases between coreference of generic NPs and zero relation

There is also a borderline between the cases of coreference of the generic NPs and the cases where it makes no sense to mark a coreferential relation. We do not annotate "generic coreference" if noun phrases have different scope (i.e. they refer to different sets of objects), e.g. *ženy* (= *women*) *– ženy v 19. století* (= *women in 19th century*). In this case, the bridging relation of the type SUBSET is annotated instead. In other problematic cases, annotators usually apply to their intuition and the text coherence. If both say no, no coreference is annotated.

5.4 Coreference with measure NPs and other NPs with a 'container' meaning

In PDT, a special group of numerals and nouns with a 'container' meaning is singled out. They have the modification in their valency frames denoting the content (people, things, substance etc.) of a container expressed by the governing noun. These 'container' expressions are e.g. nouns and numerals denoting groups, number or amount, sets, collections, portions, etc. (*skupina lidí* (=*group of people*), *počet akcií* (=*number of stocks*), *stádo krav* (=*herd of cows*), *dostatek financí* (=*abundance of finance*), *milióny židů* (=*millions of Jews*), *sklenice piva* (=*glass of beer*), *deset procent obyvatel* (=*ten percent of population*)).

The PDT convention on annotating coreference by NPs with a 'container' meaning follows the maximum-scope rule, i.e., if possible, the governing ('container') node is linked by a coreference link (16). The modifications of containers may be coreferential themselves independently of the 'containers' (17)

(16) *Absolutní většina lidí závislých na heroinu je příliš mladá na to, aby si #PersPron pamatovala rozklad a zesláblost generace sedmdesátých let, takže odvrácenou stránku „fantastického" života si #PersPron mnohdy vůbec neuvědomí*. (=*Absolute majority of people addicted to heroin is too young to remember the decomposition and enfeeblement of the generation of seventies, so they* (lit. '*she'* referring to '*majority'*) *do not realize the downside of the "fantastic" life*.)

(17) *V běžném vzorku sedmdesátých let byla pouze 3–4 procenta čisté suroviny. b. Nyní jsou k dostání balíčky obsahující až 80 procent čistého heroinu*. (=*In an average sam-*

ple from the seventies, there were only 3-4 percent of pure raw material. Currently, one can get packages containing up to 80 percent of pure heroin.)

Coreference of 'containers' can be problematic from the point of view of their generic or specific interpretation. Nouns referring to groups may refer generically to the elements belonging to that group or specifically to the group itself. In the following example, there has been a disagreement between annotators concerning the generic/specific reading of the NP *skupina* (=*group*). We believe that this kind of disagreement could be solved by separating the group of non-specific non-generic references.

(18) *Podle výzkumů ve vyspělých zemích se ukazuje, že lidí, kteří potřebují speciální služby, je daleko víc. U nás by tuto skupinu tvořilo asi tak 70000 osob. Jsou to hlavně starší lidé se zbytky zraku a slabozrací. Tato skupina stojí úplně mimo a má tak život ještě více ztížený, protože mnozí o těchto službách ani nevědí*. (=*According to the research in the developed countries, there are many more people who need special services. In our country, the group of such people would count about 70,000 individuals. They are mainly older people sighted and visually impaired. This group is completely off, their life being even more difficult, because they don't even know about many of these services*.)

More complicated are the cases where coreference chains for 'containers' and their modifications intersect. In (19), a coreference link for *the strikers* in b. should lead to *three and a half thousand workers* but in c., the number of strikers changes, so the container modification *workers* should be marked as coreferent with *the strikers* in b. For such cases, coreference of type GEN is used in PDT.

(19) a. *Tří a půl tisíce dělníků vyhlásili stávku.* b. *Stávkující žádají zvýšení platů o šest procent.* c. *Do 8. března se počet stávkujících může zdvojnásobit*. (a. *Three and a half thousand workers went on strike*. b. *The strikers demand six percent of salary increase.* c. *By 8 March, the number of strikers may double.*)

However, in this case, the problem is rather specific. Here, *počet stávkujících* (=*the number of strikers*) does not actually refer to the strikers (as it would e.g. in *tisíc stávkujících* (=*thousand strikers*) but to the number itself and that is the reason for coreference annotation to *strikers*. In such cases, *the number* does not serve as a 'container' in proper sense.

5.5 Coreference with abstract nouns

Processing coreference of abstract nouns seems to be in some respects close to that of generics. Abstract nouns do not refer to a type, but to a notion. However, this notion is unique in the same way as type is unique to the generic expression which refers to it. Moreover, abstract nouns are close to predicative and quantificational interpretation and there are no formal rules distinguishing them from concrete NPs and deverbatives. They also result in high ambiguity when annotated for coreference.

There have been several changes in the guidelines for the annotation of coreference and bridging relations with abstract nouns. Finally, we decided to distinguish between "specific" and "generic" abstracts. If subjects to annotation have complements with specific reference, or they have unambiguously specific reference themselves, coreference between them is annotated as textual coreference, type SPEC (20). In case of even a little doubt, we annotate textual coreference, type GEN (8).

(20) *Ve specifických podmínkách české ekonomiky růst nezaměstnanosti v letech 1991–1993 značně zaostal za poklesem HDP. […] Nejméně dvouprocentní růst české ekonomiky již letos*. (=*In the specific conditions of the Czech economy the growth of unemployment... This year at least a two percent growth of the Czech economy*.)

5.6 Coreference with verbal nouns

With verbal nouns, both specifying and generic reference are possible as well. Textual coreference with verbal nouns is annotated according to the following strategy:

- If both verbal nouns are specific, they refer to a specific situation and their possible arguments are coreferential, the relation between them is annotated as textual coreference, type SPEC, cf. (21);
- If both verbal nouns are generic, or rather if their arguments are generic, the relation between them is annotated as textual coreference, type GEN. Cf. (22);
- If both verbal nouns are specific, but their arguments are not coreferential, coreferen-

tial relation between them is not annotated.;

- If one verbal noun is specific and the other is generic, coreferential relation between them is not annotated.

(21) *Vedení Pojišťovny Investiční a Poštovní banky nás upozornilo, že jejich pojišťovna nebyla zařazena mezi ty, které umožňují úrazové připojištění, ač tuto službu poskytují. Omlouváme se za toto nedopatření, dotyčná redaktorka byla pokutována*. (=*The Insurance Investment and the Post Bank management has notified us that their insurance company was not included among those that allow casualty insurance, although it provides this service. We apologize for this oversight, the editor who made the mistake was fined.*)

(22) *Rychlé, avšak i bezpečné vypořádání. Rychlost vypořádání burzovních obchodů v čase odpovídá podle Jiřího Béra potřebám*. (= *Fast, yet safe transaction. According to Jiřího Bér's opinion, the speed of transaction corresponds to the needs*.)

However, such instructions are quite ambiguous themselves, because, firstly, it is not always clear, what a specific verbal noun means and, secondly and most importantly, verbal nouns may have more than one argument, one of them being generic and other – specific (Pergler 2010). Moreover, deverbatives themselves may refer to specific events that has already happened (thus tending to type SPEC if coreferent) or to hypothetic or typical ones (then, in case of coreference, marked as GEN).

6 Discussion

Processing coreference of generic expressions, even in manual annotation, raises a number of problems, both theoretical and the applied, like complification of coreference resolving. As we have seen, the problem of generics is very language-specific. Each resolving system trying to process coreference for generics will have to be oriented towards the specific linguistic description of the language in question. But even so, there are many possibilities of expressing generic expressions in every language, thus making the formal problem of extracting generics even in one separate language extremely difficult.

Generic expressions are analyzed relatively in more detail for English (Carlson 1980, Carlson - Pelletier 1995). However, this research relies heavily on language forms, it is not based on a large-scale corpus and it seems to be too theoretical to be easily adapted to a large corpus (manual or automatic) processing. On the other hand, Carlson's classification of the reference reading of nouns could be used in practice for the distinction between generic and non-specific non-generic NPs. Using our experience, we believe that it would make the annotation more consistent: there would be less ambiguity between specific and generic readings. However, being helpful in resolving the cases from section 5.1, this decision would not resolve the majority of the remaining problematic cases. There still remain borderline cases with specific noun expressions with possible valency frames (see 5.2), coreference with abstract and verbal nouns and so on. Separating the group of NPs with non-specific reading, the coders should concentrate on quite specific semantic issues when annotating. Moreover, annotating more groups of nouns is always a costly and time-consuming task.

From the theoretical point of view, one could imagine a scale: from noun expressions with concrete meaning and specific reading (say named entities) up to abstract nouns and deverbatives with generic reading. However, such an approach will not help to process generic NPs in large-scale corpora.

7 Conclusion

In this paper, we discussed the problem of annotating coreference with generic expressions. Considering theoretical approaches has revealed that they tend to be very language specific. State-of-the-art in annotating coreference relations for generic NPs needs unification but this is complicated, as the formal representation of genericity differs dramatically from language to language and can be hardly unified. We have presented an approach to annotation of generic expressions in PDT and analyzed some typical problematic examples. We consider this issue to be far from being solved. Both, theoretical research and large data approaches should be further investigated.

Acknowledgments

We gratefully acknowledge support from the Grant Agency of the Czech Republic (grants P406/12/0658 and P406/2010/0875).

References

Eduard Bejček, Jan Hajič, Jarmila Panevová, Jan Popelka, Lenka Smejkalová, Pavel Straňák, Magda Ševčíková, Jan Štěpánek, Josef Toman and Zdeněk Žabokrtský. 2011. *Prague Dependency Treebank 2.5*. Data/software, Charles University in Prague, MFF, ÚFAL, Praha, Czech Republic (http://ufal.mff.cuni.cz/pdt2.5/).

Greg Carlson. 1980. *Reference to kinds in English*. New York: Garland.

Greg Carlson. 2005. Generic Reference. In *The Encyclopedia of Language and Linguistics, 2nd Ed. Elsevier*.

Greg Carlson and F.J. Pelletier (eds.). 1995. *The Generic Book*. Chicago: University of Chicago Press.

Denis Delfitto. 2006. Bare plurals. In Martin Everaert and Henk van Riemsdijk (eds.) *The Blackwell Companion to Syntax*. Blackwell Publishing, pp. 214-259.

Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, Julia Trushkina und Heike Zinsmeister. 2004. Recent Developments in Linguistic Annotations of the TüBa–D/Z Treebank. In *Proceedings of the third workshop on treebanks and linguistic theories (TLT 2004)*. Tübingen.

Jan Hajič, Silvie Cinková, Kristýna Čermáková, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jiří Semecký, Jana Šindlerová, Josef Toman, Kristýna Tomšů, Matěj Korvas, Magdaléna Rysová, Kateřina Veselovská, Zdeněk Žabokrtský. 2009. *Prague English Dependency Treebank 1.0*. Institute of Formal and Applied Linguistics. Charles University in Prague.

Zdeněk Hlavsa. 1975. *Denotace objektu a její prostředky v současné češtině*. (*Object denotation and its means in current Czech*). Prague, Czech Republic.

Olga Krasavina and Christian Chiarcos. 2007. PoCoS – Potsdam Coreference Scheme. In *Proceedings of ACL 2007*, Prague, Czech Republic.

Christopher Lyons. 1999. *Definiteness*. Cambridge: Cambridge University Press.

Lucie Mladová. 2011. Annotating Discourse in Prague Dependency Treebank. In *Workshop of Annotation of Discourse Relations in Large Corpora at the conference Corpus Linguistics 2011 (CL 2011)*. Birmingham, Great Britain, July 2011.

Anna Nedoluzhko, Jiří Mírovský, Radek Ocelák, Jiří Pergler. 2009. Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*. Goa, India, 2009, pp. 1–16.

Anna Nedoluzhko. 2003. Ukazovací zájmeno "ten" a generické jmenné fráze v češtině. In *IV. mezinárodní setkání mladých lingvistů Olomouc 2003: Jazyky v kontaktu, jazyky v konfliktu*. Olomouc: Univerzita Palackého v Olomouci, pp. 85 – 96.

Anna Nedoluzhko. 2011. *Rozšířená textová koreference a asociační anafora. Koncepce anotace českých dat v Pražském závislostním korpusu.* Prague, ÚFAL.

Elena V. Paducheva. 1985. *Vyskazyvanie i ego sootnesennost s dejstviteľnosťju*. Moskva.

Jiří Pergler. 2010. *Koreferenční řetězce s nespecifickou a generickou referencí v češtině (Coreferential chains with non-specific and generic reference in Czech)*. Unpublished bachelor thesis. Prague.

Massimo Poesio. 2004. The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. In *Proceedings of SIGDIAL*.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *Proceedings of the International Conference on Se-mantic Computing (ICSC-07)*. Washington, DC, pp. 517–526.

Marta Recasens and Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. In *Language Resources and Evaluation*.

Uriel Weinreich. 1966. On the Semantic Structure of Language. In *Universals of Language*, 2nd ed. Cambridge, Mass.