The Copenhagen Dependency Treebank (CDT)
Extending syntactic annotation to morphology and semantics

Henrik Høeg Müller
Copenhagen Business School
Copenhagen, Denmark
E-mail: hhm.isv@cbs.dk

Abstract

This paper has two main objectives. The first is to provide an overview of the CDT annotation design with special emphasis on the modeling of the interface between syntactic and morphological structure. Against this background, the second objective is to explain the basic fundamentals of how CDT is marked-up with semantic relations in accordance with the dependency principles governing the annotation on the other levels of CDT. Specifically, focus will be on how Generative Lexicon theory has been incorporated into the unitary theoretical dependency framework of CDT by developing an annotation scheme for lexical semantics which is able to account for the lexico-semantic structure of complex NPs.

1. Introduction

The Copenhagen Dependency Treebank (CDT)[1] is a set of parallel text collections (treebanks) of approx. 60.000 words each for Danish, English, German, Italian and Spanish with a unified annotation of morphology, syntax and discourse, as well as an alignment system of translational equivalences (Kromann, 2003; Buch-Kromann et al., 2009). The treebanks are annotated on the basis of the dependency-based grammar formalism Discontinuous Grammar (Buch-Kromann, 2006) and can be used to train natural language parsers, syntax-based machine translation systems, and other statistically based natural language applications. CDT is unique in creating parallel treebanks for 5 languages and combining this effort with a unitary level of analysis which can provide annotations that span all levels of linguistic analysis, from morphology to discourse, on a principled basis.[2] Here, however, the centre of attention will be morpho-syntax and semantics.

This paper is structured as follows. In Section 2, it is explained how syntactic structure is annotated in CDT. In Section 3, focus is on how morphological structure is marked-up on the basis of an operator notation system. In section 4, building on the insights reached in the previous sections, the annotation principles for lexical-semantic structure are presented, and, finally, Section 5 sums up the most central points.

2. Syntactic annotation

The syntactic annotation of the treebanks is based on the principles accounted for in the dependency theory Discontinuous Grammar (Buch-Kromann, 2006) and in the CDT-manual (Buch-Kromann et al., 2010). In accordance with other dependency theories, it is assumed that the syntactic structure of a sentence or an NP can be represented as directed relations between governors and complements and adjuncts. Complements function as arguments and are lexically licensed by the governor, whereas adjuncts are modifiers that take the governor as argument.

Figure 1 below shows the primary dependency tree for the sentence *Kate is working to earn money* (top arrows), enhanced with secondary subject relations (bottom arrows). The arrows point from governor to dependent, with the relation name written at the arrow tip.

[1] The project is hosted on Google Code – http://code.google.com/p/copenhagen-dependency-treebank/ – and all the sources are freely available.

[2] Many treebank projects focus on annotating a single linguistic level or a single language: The Penn Treebank (Marcus et al., 1993) focuses on syntax; the Penn Discourse Treebank (Prasad et al., 2008ab) and the RST Treebank (Carlson et al., 2001) on discourse, and the GNOME project (Poesio, 2004) on coreference annotation. Others, like the TuBa-D/Z treebank (Hinrichs et al., 2004), include both morphology and coreference annotation, and the Prague Dependency Treebank (Böhmová et al., 2003) comprises Czech, English and Arabic.

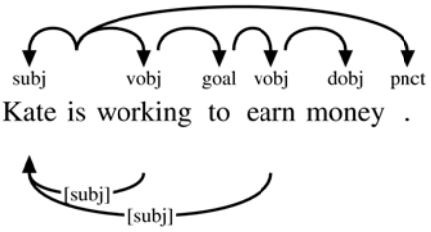

Figure 1. Primary dependency tree (top) augmented with secondary subject dependency relations (bottom).

The finite verb *is* functions as head and top node of the sentence. The arrow from *is* to *Kate* identifies *Kate* as the subject ("subj"), while the arrow from *is* to *working* indicates that *working to earn money* is a verbal object ("vobj") governed by *is*. The finite verb also functions as governor to the punctuation mark ("pnct"), which is categorized as an adjunct relation. The lexical main verb *working* establishes the adjunct relation ("goal") to *to earn money*, and inside the adverbial phrase *earn* functions as verbal object ("vobj") governed by the infinitive marker *to*. Finally, *money* is the direct object ("dobj") of *earn*. So, in the primary dependency structure every word heads a phrase consisting of all words that can be reached from the phrasal head by following the arrows above the text. The arrows below the text specify secondary subject relations [subj] in the sense that *Kate* is the logical subject of both *working* and *earn*.

The annotation scheme for syntax currently includes approx. 20 complement relations and 60 adjunct relations of which ten of the most frequent ones are listed in Table 1.[3]

Complement relations	**Adjunct relations**
nobj (nominal object: *for the$_{nobj}$ child$_{nobj}$*) **subj** (subject: *They$_{subj}$ saw me$_{dobj}$*) **vobj** (verbal object: *He$_{subj}$ had left$_{vobj}$ it$_{dobj}$*) **dobj** (direct object: *He$_{subj}$ left us$_{dobj}$*) **pobj** (prepositional obj.: *one of$_{pobj}$ them$_{nobj}$*) **preds** (subject predic.: *It$_{subj}$ was blue$_{preds}$*) **@loc** (locative object: *living in$_{@loc}$ Rome$_{nobj}$*) **predo** (object predicative: *We$_{subj}$ found it$_{dobj}$ disappointing$_{predo}$*) **iobj** (indirect object: *We$_{subj}$ gave him$_{iobj}$ flowers$_{dobj}$*) **avobj** (adverbial object: *as before$_{avobj}$*)	**time** (time adverbial: *We$_{subj}$ leave now$_{time}$*) **loc** (location adverbial: *I$_{subj}$ fell here$_{loc}$*) **man** (manner adverbial: *I$_{subj}$ read slowly$_{man}$*) **quant** (degree adverbial: *very$_{quant}$ hard*) **neg** (negation: *I$_{subj}$ will not$_{neg}$ leave$_{vobj}$*) **pnct** (punctuation: *It$_{subj}$ is !$_{pnct}$*) **attr** (attribution: *la tarea$_{nobj}$ difícil$_{attr}$*) **appr** (restrictive apposition: *the genius$_{nobj}$ Einstein$_{appr}$*) **appa** (parenthetic apposition: *Einstein, the$_{appa}$ genius$_{nobj}$*) **relr** (restrictive relative clause: *the cat$_{nobj}$ that$_{subj}$ died$_{relr}$*)

Table 1. Ten of the most frequent complement and adjunct relations in the syntactic annotation.

Generally, on the sentence level we do not annotate semantic relations, only syntactic dependency relations. However, in order to achieve a more fine-grained analysis of modifiers, adverbials are annotated according to the semantic relation established between governor and adverbial, cf., e.g., the relations ("time"), ("loc") and ("man") in Table 1. The introduction of semantic relations into the syntactic annotation is, of course, debatable, but was preferred over annotating all adverbial adjuncts as ("mod") relations in accordance with their syntactic function, an alternative adopted in earlier versions of CDT.

With respect to interannotator agreement an experiment has been conducted where two annotators annotated 21 English and Danish texts with a total of 4287 relations. The results were the following:[4]

81% : *Full labeled agreement*, i.e. the probability that another annotator assigns the same label and out-node to the relation.

[3] For a full specification of the relation inventory see CDT manual on http://code.google.com/p/copenhagen-dependency-treebank/.

[4] See CDT manual (op.cit).

93% : *Unlabeled agreement*, the probability that another annotator assigns the same out-node (but not necessarily label) to the relation.

85% : *Label agreement*, the probability that another annotator assigns the same label (but not necessarily out-node) to the relation.

In general, the results are satisfactory and prove the system to be quite solid.

3. Morphological annotation

The morphological annotation in CDT only deals with derivation and composition, since inflectional morphology can be detected and analysed automatically with high precision for the treebank languages.

The internal structure of words is encoded as a dependency tree. However, in order to annotate dependency relations inside solid orthography compounds and derivationally constructed words, which appear as tokens in the automatically produced word tokenisation, an operator notation scheme has been developed (Müller, 2010). The operator notation is an abstract specification of how the dependency tree for a morphologically complex word is constructed from roots, annotated as lemmas or in some cases imperatives, dependent on the specific language, in combination with morphological operators. Examples of this notation form, applied to derived nouns and nominal compounds in Danish, are shown in figure 2 to 5.[5]

Antistof [antibody]:
stof –anti/NEG:contr

Figure 2. Operator notation of the Danish prefixed derivation *antistof* [antibody].

Lancering [launching]:
lancer +ing/DERvn:core

Figure 3. Operator notation of the Danish suffixed derivation *lancering* [launching].

Loftslampe [ceiling lamp]**:**
lampe –[loft]s/LOC

Figure 4. Operator notation of the Danish compound *loftslampe* [ceiling lamp].

Vindmølle [wind mill]**:**
mølle –vind/FUNC

Figure 5. Operator notation of the Danish compound *vindmølle* [wind mill].

In Figure 2, the Danish word *antistof* [antibody] is constructed from the root *stof* [body] by attaching the prefix *anti-* as a "NEG:contr" dependent of the root. The "NEG:contr" relation indicates that *anti-* negates the meaning of *stof* so that the new word acquires the opposite meaning of the base. The minus sign introducing the notation specifies the pre-head position of the prefix. In Figure 3, the word *lancering* [launching] is constructed from *lancer* [launch] by transforming the verbal root into a predicative eventive core noun by means of the transformative suffix *-ing* which takes *lancer* as its dependent. Here, the plus sign indicates the post-head position of the suffix. With respect to dependency, the operator notation follows the convention that transformative affixes take the root as dependent, whereas non-transformative affixes are dependents to the root.

The analyses of the minimally complex Danish compounds in Figure 4 and 5 can be explained in the following way: *Loftslampe* [ceiling lamp] in Figure 4 is composed of the modifier *loft* [ceiling], the head *lampe* [lamp] and the linking consonant or interfix *-s*. The annotation is to be understood as follows: The minus sign specifies the pre-head position of the modifier, the lexical material of the modifier itself occurs in square brackets, then comes the interfix which is a phonetically induced morpheme which only acts as a glue between the head and the modifier, and finally, following the oblique slash, the meaning aspect of the head noun selected by the non-head modifier, in this case a locative meaning relation. The analysis of

[5] In CDT, the three word-classes nouns, adjectives and verbs are marked-up according to the operator notation scheme, but, for matters of space, we only provide examples with nouns. Moreover, CDT has a system for separating linking elements such as thematic vowels, infixes and interfixes, on the one hand, from what is the suffix proper, on the other hand, and it allows CDT to regenerate the word form in question on the basis of the operator instructions. This system is also not detailed here.

vindmølle [wind mill] in Figure 5 follows the same scheme, but here the meaning component activated by the modifier is functional.

Of course, the system must also be able to handle more complex expressions, such as, e.g., the combination of derivation and compounding, cf. Figure 6 below.

Flerbrugersystem [multiple user system]: **system –[[brug@V] +er/DERvn:agent –fler/MOD:quant]/GOAL**

Figure 6. Operator annotation of the Danish compound *flerbrugersystem* [multiple user system].

The head of the compound is the simple lexeme *system* [system], and the non-head is the complex lexeme *flerbruger-* [multiple user]. The operator notation of the complex non-head lexeme, i.e. "–[[brug@V] +er/DERvn:agent –fler/MOD:quant]/GOAL", should be analyzed step by step as follows:

1. the minus sign introducing the square brackets that delineate the non-head indicates the pre-head position of the non-head.
2. "[brug@V] +er/DERvn:agent" specifies that the derivationally complex head *bruger* [user] is an agent nominalization of the verb *bruge* [use] triggered by the suffix *-er*. (The indication of word class in separate square brackets with the specification "*@word-class*" is optional, but it should be indicated when the form is ambiguous, as in this case between a noun and a verb.)
3. "–fler/MOD:quant" indicates via the minus sign the pre-head position of *fler* [multiple] with respect to *bruger* [user], and that the semantic relation established is one of quantificational modification, cf. "MOD:quant".
4. Finally, the last part of the operator, i.e. "/GOAL", specifies that the primary level non-head prompts a semantic ("goal")-relation between the non-head and the head in the sense that the interpretation of *flerbrugersystem* is a system which has the goal/purpose of several people being able to use it.

Summarizing, in the operator annotation the dependency tree for a morphological complex lexeme is annotated as a root – given abstractly by means of its lemma or imperative form – followed by one or more operators "*lemma* op_1 op_2..." applied in order. Each operator encodes an abstract affix and a specification of how the abstract affix combines with the base (root or complex stem) in its scope. Here, *abstract affix* is used to denote either a traditional affix (prefix or suffix) or the non-head constituent of a compound. The operator itself has the form "*pos affix*/*type*". The field *pos* specifies whether the abstract affix is attached to its base in prefix position ("–") or suffix position ("+"), or a combination of these in case of parasynthetic verbs, cf. Table 2 (*adormecer* [lull to sleep]). The field *type* specifies the derivational orientation (e.g., "DERvn", {fig. 3}), either in the form of a categorial shift, or not. Moreover, the field *type* semantically and functionally identifies the type and, where relevant, the subtype, of the semantic relation created between the base and the abstract affix (e.g., "NEG:contr", {fig 2}). The field *affix* specifies the abstract affix and its possibly complex internal structure. The abstract affix may be encoded either as a simple string representing a simple affix or a simple root (e.g., *-er*, "*brug*", {fig. 6}), or as a complex string of the form "[*stem*]" or "[*stem*]*interfix*", where "*stem*" encodes the internal structure of the abstract affix in operator notation (e.g., "–[loft]s/LOC" or "–vind/FUNC" , {fig. 4 and 5}).

As mentioned previously, the abstract affix functions as a dependent of the base when it is non-transformational, whereas if it triggers word class change or a significant change of meaning, the base is assumed to function as a dependent of the abstract affix.

Finally, it is important to keep in mind that the operator notation is merely an abstract specification of a dependency tree, not an autonomous annotation system which follows individual rules.

A sample of morphological relation types is listed in Table 2 below.[6] The system is flexible in the sense that all relations can be annotated as either prefixes or suffixes, or non-head roots in case of compounds; here they are just listed as they typically appear in the CDT languages.

[6] The different relation types have taken inspiration from the works on morphological categories by Rainer (1999) and Varela and Martín García (1999). The total number of morphological relation types in CDT is 70, out of which 57 are derivational relations (17 prefix; 40 suffix) and 13 compositional relations (see CDT-manual, cf. footnote 3).

Relations that typically appear with prefixes
SPACE:loc (location: *intramural = mural −intra/SPACE:loc*)
TIME:pre (precedency: *prehistorical = historical −pre/TIME:pre*)
NEG:contr (contrast: *antihero = hero −anti/NEG:contr*)
AGENT (causative: *acallar 'silence' = callar −a/AGENT*)
TELIC (telic: *oplåse 'open' = låse −op/TELIC*)
MOD:quant (quantification: *multicultural = cultural −multi/MOD:quant*)
TRANS (transitivity: *påsejle 'colide' = sejle −på/TRANS*)

Relations that typically appear with suffixes
AUG (augmentative: *perrazo 'big dog' = perro +azo/AUG*)
DIM (diminutive: *viejecito 'little old man' = viejo +ecito/DIM*)

Verb derivation
DERnv (noun→verb derivation: *salar 'to salt' = sal +ar/DERnv*)
DERav (adjective→verb derivation: *darken = dark +en/DERav*)
DERvv (verb→verb derivation: *adormecer 'lull to sleep' = dormir −+[a][ecer]/DERvv*)

Noun derivation
DERvn:agent (verb→noun derivation: *singer = sing +er/DERvn:agent*)
DERvn:core (verb→noun derivation: *exploitation= [exploit@V] +ation/DERvn:core*)
DERnn:cont (noun→noun derivation: *azucarero 'sugar bowl' = azucar +ero/DERnn:cont*)

Adjective derivation
DERva:pas.epi (deverbal adjective: *transportable = transport +able/DERva:pas.epi*)
DERna:rel (denominal adjective: *presidential = president +ial/DERna:rel*)

Relations that typically appear with compounds
CONST (constitutive: *træbord 'wooden table' = bord −træ/CONST*)
AGENT (agent: *politivold 'police violence' = kontrol −politi/AGENT*)
SOURCE (source: *rørsukker 'cane sugar' = sukker −rør/SOURCE*)
GOAL (goal: *krigsskib 'war ship' = skib −[krig]s/GOAL*)
FUNC (function: vindmølle 'wind mill' = *mølle −vind/FUNC*)
LOC (location: *loftlampe 'ceiling lamp' = lampe –[loft]s/LOC*)

Table 2. Relation types in the morphological notation system.

4. The semantic dimension

4.1 Basic annotation of NPs

A number of semantic annotation projects have developed over the years.[7] In CDT, the dependency structure has been enhanced with semantic annotation with respect to sentence level adverbials, derivations and different kinds of NPs. In this context, we limit ourselves to focusing on the description of how Generative Lexicon theory (GL) has been integrated into the current dependency framework in order to account for the lexical semantics of certain NPs.

GL (Pustejovsky, 1991, 1995, 2001) is based on the assumption that any lexeme can be defined by the four qualia, FORMAL, CONSTITUTIVE, TELIC and AGENTIVE, which constitute the fundamental rules according to which the integration of mental representations of entity types is produced. In other words, Qualia can be described as a template representing the relational force of a lexical item, a system of constructive understanding and inference.

Below, we exemplify the integration of lexical semantic knowledge in the dependency-based multilevel CDT annotation scheme by describing the annotational challenges posed by one single type of NPs, viz. Spanish N+PP constructions.

In N+PP constructions like *taza de café* [coffee cup] and *taza de porcelana* [china cup], the PP-modifiers *de café* and *de porcelana* are syntactic dependents of the head *taza*, but they select different sub-senses of *taza*, Telic and Constitutive, respectively, and act semantically as governors (Johnston and Busa, 1999).[8] The relationship between syntactic and semantic dependencies is implemented in terms of annotation in the following way.

[7] *PropBank* (Palmer et al., 2005) is a corpus semantically annotated with verbal propositions and their arguments; *NomBank* (Meyers et al., 2004ab) marks up the sets of arguments that co-occur with nouns; *VerbNet* marks up the sets of syntactic frames a verb can appear in to reflect underlying semantic components constraining allowable arguments; and *FrameNet* (Ruppenhofer et al., 2006) is an on-line lexical resource for English based on frame semantics and supported by corpus evidence.

[8] In practice, CDT operates with an expanded set of qualia-roles. For instance, the Telic-role can manifest itself either as Goal or Function (see Table 2), dependent on the specific interpretation.

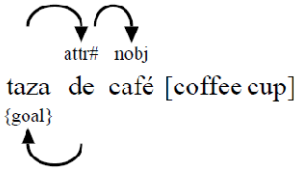

Figure 7. Syntactic and semantic annotation of the Spanish phrasal NP-compound *taza de café* [coffee cup].

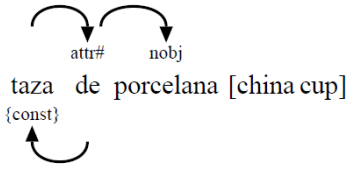

Figure 8. Syntactic and semantic annotation of the Spanish phrasal NP-compound *taza de porcelana* [china cup].

The arrows above the text from the head *taza* [cup] to the PPs *de café* [of coffee] and *de porcelana* [of china] in Figure 7 and 8, respectively, indicate that the relation is non-argumental, i.e. what we understand as one of contribution ("attr") – basically because the head is non-predicative or non-relational. In other words, the non-head is not lexically licensed by the governing head. The hash symbols following the ("attr") label stipulate that the phrases in question show composite structure (see later discussion). The nouns *café* and *porcelana* are syntactically governed by the preposition *de* and function as noun objects ("nobj"). The "reversed" arrows below the text indicate semantic structure. The non-heads activate the Telic quale – we refer to it as a ("goal") relation – and the Constitutive quale of the head, respectively, being the general assumption that the qualia of the head can be triggered by different modifiers, in these cases PPs.[9]

Moreover, *taza de café* is ambiguous as it allows yet another interpretation equivalent to cup of coffee, where *taza* functions as a specifier of quantity. In these cases it is the complement *café* which has to respect the selectional restrictions imposed by, e.g., the predicate, and, consequently, the construction must be re-analyzed as yielding a specifier+head structure, i.e. a case of head switching, cf. Figure 9 below.

[9] Of course, the preposition *de* in itself is purely syntactic, but we have chosen to see the whole PP as the unit which activates the semantic relation between head and non-head.

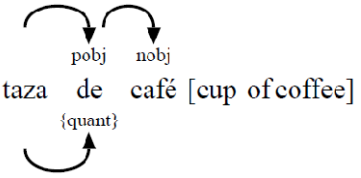

Figure 9. Syntactic and semantic annotation of Spanish NP expressing quantification.

In terms of annotation the difference between Figure 7 and 9 is that in Figure 9 the noun *taza* is relational and thus selects the PP *de café* as a dependent. Therefore *de café* functions as an argument to the head, which is made clear by the fact that the relation name written at the arrow tip is ("pobj"), a lexically governed prepositional object. Consequently, the syntactic labels ("pobj") and ("nobj") indicate that the modifying noun or PP is lexically governed by the head, whereas the ("attr")-label indicates that this is not the case. The label ("nobj") is also used more widely when a noun is governed by an article or a preposition. The arrow below the text indicates that *taza* does not function as a semantic head, but as a specifier which imposes a quantificational reading on the PP. Therefore the arrows showing syntactic and semantic dependency, respectively, are oriented in the same direction in this case.

Apart from the Qualia inspired inventory of semantic relations, CDT also operates with a set of "standard" semantic roles in the form of Agent, Patient, Recipient, etc. These roles are used when the head noun is deverbal or deadjectival and thus projects an argument structure, cf. Figure 10.

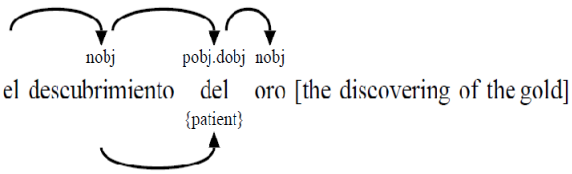

Figure 10. Full syntactic and semantic annotation of Spanish NP with deverbal head.

In Figure 10, the bottom arrow specifies that the PP *del oro* [of-the gold] functions as Patient with respect to the deverbal head noun *descubrimiento* [discovering]. The top arrow from head noun to PP demonstrates that the PP is a syntactically governed ("pobj") with the function of direct object ("dobj").

Generally, the qualia-structure has been a guiding principle for the organization of the semantic inventory of CDT on all levels, i.e. with respect to adverbial adjuncts, NPs and derivational morphology.[10] This attempt to unify the inventory through the qualia-structure, which provides a rather general template for structuring semantic relations, is theoretically appealing because it accommodates the fact that similar semantic relations are found on different linguistic levels. However, this does not mean that any semantic relation can be accounted for with point of departure in the qualia-structure. For instance, the nature of the arguments to a predicate (semantic labeling), cf. Figure 10, or certain adverbial adjunct relations, such as condition, concession, contrast, etc., fall outside the explanatory frame of the qualia-structure.

4.2 Compounding

As mentioned before, we use the hash symbol to indicate when a phrasal constellation of words should be regarded as a compound. Of course, in the non-English Germanic languages it is not a problem as they have unitary stress (e.g. in Danish, head nouns are reduced prosodically and pronounced with secondary stress) and solid orthography, which means that in CDT they are tackled in accordance with the so-called operator notation scheme. However, when a word constellation should be regarded as a free syntactic phrase formation or a compound is not an uncontroversial issue, which can be appreciated, for instance, in the Spanish grammatical literature about the subject.

Briefly, the problem is that the criteria for compounding in Spanish, and other Romance languages for that matter, are often based on the notion of degree of lexicalization – the more lexicalized the more compound status – which seems to be difficult to deal with both empirically and theoretically in a setting of annotation.

In the standard approach (e.g., Escandell Vidal, 1995; Val Alvaro, 1999), degree of lexicalization is measured by the parameters of internal solidity, i.e. cohesion between the constitutive elements, and, secondarily, possibility of substitution of elements, and finally, as an effect of these criteria, degree of semantic transparency.[11]

According to this approach, good examples of phrasal compounds would be such as the ones in (1) and (2). They have a solid internal structure, and, moreover, the foot-examples in (2) are not semantically transparent. They are exocentrically structured, and they are metaphoric extensions of some original meaning of which we have more or less lost track.

(1) *un punto de vista*
[a point of view]

un punto* ***agudo* *de vista*
[a point **sharp** of view]

un ***agudo*** *punto de vista/un punto de vista* ***agudo***
[a **sharp** point of view/a point of view **sharp**]

(2) *pie de liebre*
[foot-of-hare] ‟sort of clover‟

pie de atleta
[foot-of-athlete] ‟sort of skin desease‟

pie de gallina
[foot-of-chicken] ‟sort of knot‟

However, the examples in (3) and (4) below are not so good phrasal compounds. They do not show a solid internal structure, and the ones in (4) are even headed by the event denoting deverbal noun *venta* [sale], which means that

[10] This also goes for the CDT annotation of anaphoric relations and discourse structure, which, however, has not been the topic of this paper.

[11] Other authors (see, e.g., Corpas Pastor, 1997; Ferrando Aramo, 2002; Ruiz Gurillo, 2002; Alonso Ramos, 2009) intend to establish more or less solid distinctions between compounds, locutions/idiomatic expressions, and collocations on the basis of a wide range of syntactic, semantic and denotative criteria, such as cohesion, transparency and unity of meaning. Although a continuum, rather than an attempt to make clear delimitations, probably is the more adequate way to represent these types, there is no doubt that important phraseological distinction can be identified between different N+PP constructions. However, the point deserving emphasis here is that, contrary to the current discussion in the Spanish literature, the definition of compounding in the non-English Germanic languages, such as Danish, does not hinge on the extent to which a certain construction fulfils an array of criteria, but is solely based on the criterion of unitary stress and, consequently, solid orthography. Therefore, although Germanic compounds can show all kinds of semantic "peculiarities", Germanic compounding is well-defined, while Romance N+PP compounding is a fuzzy edged phenomenon.

they are completely productive and that they resample the corresponding "free" sentence structure.

(3) *lazo de luto*
[bow of grief/mourning]

bolsa de viaje
[bag-of-travel/travel bag]

*lazo **negro** de luto*
[bow **black** of grief]

*bolsa **negra** de viaje*
[bag **black** of travel]

(4) *venta de carne/ trigo/ caballos/ teléfonos*
[sale of meat/ wheat/ horses/ telephones]

It is not the intention here to enter into a theoretical discussion about compounding, but it must be acknowledge that in general the understanding of compounding in Spanish and other Romance languages deviates substantially from a Germanic understanding of the "same" phenomenon, cf. also footnote 11.

In order to cope with these interlingual discrepancies in CDT we have chosen a very liberal approach to Romance compounding in the sense that if the constellation of words in question can be said to designate a single entity or type of entity, we add a hash symbol indicating that the relevant construction shows some kind of tendency towards being a lexical unit. Good signs of such a status is, of course, if the modifying noun, N2, is naked, i.e. appears without determiners, or if an analogous expression in German or Danish manifests itself as a compound (with respect to Germanic compounding see, e.g., Mellenius, 1997; ten Hacken, 1999; Müller, 2001, 2003).

Another problem of compounding is coreless (exocentric) compounds, cf. what with Sanskrit terms is referred to as "bahuvrihi" (e.g., *redskin*), "dvandva" (e.g., *marxism-leninism*) and "imperavitic" (e.g., *forgetmenot*). These constructions are not especially productive, but they do not fit in so neatly in a dependency framework which builds on the assumption that every expression must have a head. This issue also concerns a number of synthetic compounds such as *darkhaired* and *blueeyed*, where it is difficult to decide which element is the head.

With respect to the headedness problem, the CDT, by stipulation, follows the general principle that the element which carries the inflectional endings also is considered the head. However, one exception to this standard is the issue of verbo-nominal compounds illustrated in (5) and (6) below and annotated according to the operator scheme. In these cases, we follow the principle that the verbal part is the head, and the nominal part, although it carries the inflectional endings, is a modifier, very often in the form of a direct object. The problem arises because there is a discrepancy between the inner dependency structure of the compound, which follows the corresponding sentence structure, and its instantiation in syntax, which dictates an inflectional declension of the modifier, when relevant.

(5) *un tocadiscos* [a play-records/record player]:
tocar ! +discos/DOBJ.patient

(6) *un guardapolvo* [a protect-dust/working coat]:
guardar ! +polvo/GOAL

4.3 Semantic agreement figures

Interannotator agreement has also been calculated for semantic relation. This has been done on the basis of the same 21 English and Danish texts that were used for the syntax annotation task, and in this case with a total of 358 semantic relations. The results were the following:[12]

48% : *Full labeled agreement*, i.e. the probability that another annotator assigns the same label and out-node to the relation.
96% : *Unlabeled agreement*, the probability that another annotator assigns the same out-node (but not necessarily label) to the relation.
50% : *Label agreement*, the probability that another annotator assigns the same label (but not necessarily out-node) to the relation.

Obviously, the scores with respect to semantic annotation are rather low in comparison with the syntactic level. A specific analysis of the major disagreement cases has not been conducted yet, but it seems reasonable to suspect that at least

[12] See CDT manual (op.cit).

some of the explanation lies in the fact that the semantic annotation of CDT covers both NPs and derivational morphology, as well as adverbial adjuncts. This makes the system fairly complex and, perhaps, in some respects too detailed. Specifically, informal investigations of compound annotation show that the annotators in many cases tend to disagree on which semantic label should be assigned to the relation between head and non-head. However, we expect to be able to improve the system by introducing a more hierarchical ordering of relations and a higher degree of label specificity.

5. Conclusion

This paper has explained how the basic dependency principles behind the sentence level syntactic analyses, through an operator notation, has been transferred to the morphological level to account for the inner structure of tokens in the form of derivations and compounds. There is a clear analogy between syntactic and morphological annotation in CDT. On both levels we depart from the basic assumption that coherent linguistic units, in the form of either sentences or words, are determined by a dependency structure in which each word or morpheme is assumed to function as complement or adjunct to another word or morpheme, called the governor. In the last part of the paper, we show from a limited subset of examples how GL semantics has been incorporated into a coherent annotation scheme compatible with the CDT dependency principles on the other descriptive levels.

It is expected that the enhancement of CDT with morphological and semantic annotation will enable inquiries into interface issues between different linguistic layers, cross-linguistic contrasts and typological variations between the languages involved in CDT, thereby supporting CDT's applicability in multilingual language processing systems. Of course, these aspects have not been dealt with in the paper, which only introduces the system.

Finally, we have seen that interannotator agreement scores confirm that the system functions robustly with respect to syntax, whereas the annotation of semantic relations is not sufficiently performent yet. Larger scale analyses of the functionality of the morphological annotation system have not been conducted so far, but preliminary studies are generally positive in terms of the user friendliness of the system, despite its obvious complexity. However, on the critical side the annotators find the system time-consuming to get familiar with.

References

Alonso Ramos, M. (2009). Delimitando la intersección entre composición y fraseología. *Lingüística española actual* (LEA), 31(2). 5-37.

Böhmová, A., Hajič, J., Hajičová, E. & Hladká, B. (2003). The Prague Dependency Treebank: a three-level annotation scenario. In A. Abeillé (ed.). *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic Publishers.

Buch-Kromann, M. (2006). *Discontinuous Grammar. A dependency-based model of human parsing and language learning*. Doctoral dissertation. Copenhagen: Copenhagen Business School.

Buch-Kromann, M., Korzen, I. & Müller, H.H. (2009). Uncovering the 'lost' structure of translations with parallel treebanks. In I.M. Mees, F. Alves, & S. Göpferich (eds). *Methodology, Technology and Innovation in Translation Process Research. Copenhagen Studies in Language* 38: 199-224.

Buch-Kromann, M., Gylling, M., Knudsen, L.J., Korzen, I. & Müller, H.H. (2010). *The inventory of linguistic relations used in the Copenhagen Dependency Treebanks.* Technical report. Copenhagen: Copenhagen Business School. Available at: http://code.google.com/p/copenhagen-dependency-treebank/.

Carlson, L., Marcu, D. & Okurowski, M. E. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue.*

Corpas Pastor, G. (1997). *Manual de fraseología española*. Madrid: Gredos.

Escandell Vidal, M.V. (1995). *Los complementos del nombre*. Madrid: Arco Libros.

Ferrando Aramo, V. (2002). Colocaciones y compuestos sintagmáticos. In A. Veiga Rodríguez, M. González Pereira & M. Souto Gómez (eds.). *Léxico y Gramática*. TrisTram, Lugo. 99-107.

Hinrichs, E., Kubler, S., Naumann, K., Telljohann H. & Trushkina, J. (2004). Recent developments in linguistic annotations of the TuBa-D/Z treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*. Tübingen, Germany. 51-62.

Johnston, M. & Busa, F. (1999). The compositional interpretation of compounds, In E. Viegas (ed.). *Breadth and Depth of Semantics Lexicons*. Dordrecht: Kluwer Academic. 167-87.

Kromann, M. T. (2003). The Danish Dependency Treebank and the DTAG treebank tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003), 14-15 November,* Växjö. 217–220.

Marcus, M. P., Marcinkiewicz, M. A. & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330.

Mellenius, I. (1997). *The Acquisition of Nominal Compounding in Swedish*. Lund: Lund University Press.

Meyers, A. et al. (2004a). The NomBank Project: An interim report. In *Proceedings of the HLTNAACL Workshop on Frontiers in Corpus Annotation*, Boston, MA. 24-31.

Meyers, A. et al. (2004b). Annotating noun argument structure for NomBank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

Mladová, L., Š. Zikánová & Hajičová, E. (2008). From sentence to discourse: building an annotation scheme for discourse based on Prague Dependency Treebank. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LCREC 2008)*. 2564–2570.

Müller, H.H. (2001). Spanish N *de* N-structures from a cognitive perspective. In I. Baron, M. Herslund, & F. Sørensen (eds.). *Dimensions of Possession*. Amsterdam/Philadelphia: Benjamins. 169-186.

Müller, H.H. (2003). Strategies de lexicalisation des noms composés en espagnol. In M. Herslund (éd.). *Aspects linguistiques de la traduction.* Bordeaux: Presses Universitaires de Bordeaux. 55-84.

Müller, H.H. (2010). Annotation of Morphology and NP Structure in the Copenhagen Dependency Treebanks. In M. Dickinson, K. Müürisep, & M. Passarotti, (eds.). *Proceeding of the Ninth International Workshop on Treebanks and Linguistic Theories*. (NEALT Proceedings Series). 151-162.

Palmer, M., Gildea, D. & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1). 71–106.

Poesio, M. (2004). Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the ACL Workshop on Discourse Annotation.*

Prasad, R., Miltsakaki, E., Dinesh, A, Lee, A., Joshi, A., Robaldo L. & Webber, B. (2008a). *The Penn Discourse Treebank 2.0. Annotation Manual*. (IRCS Technical Report IRCS-08-01). University of Pennsylvania: Institute for Research in Cognitive Science.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber, B. (2008b). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).*

Pustejovsky, J. (1991). The Generative Lexicon. *Computational Linguistics* 17. 409-441.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge (Mass.). London, England: MIT Press.

Pustejovsky, J. (2001). Generativity and Explanation in Semantics: A Reply to Fodor and Lepore. In P. Bouillon & F. Busa (eds.). *The Language of Word Meaning*. Cambridge University Press. 51-74.

Rainer, F. (1999). La derivación adjectival. In I. Bosque. & V. Demonte (eds). *Gramática Descriptiva de la Lengua Española*. Madrid: Espasa Calpe. 4595–4643.

Ruiz Gurillo, L. (2002). Compuestos, colocaciones, locuciones: intent de delimitación. In A. Veiga Rodríguez, M. González Pereira & M. Souto Gómez (eds.). *Léxico y Gramática*. TrisTram, Lugo. 327-339.

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. & Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*.

ten Hacken, P. (1999). Motivated Tests for Compounding. *Acta Linguistica Hafniensa* 31. 27-58.

Val Álvaro, J.F. (1999). La composición. In I. Bosque & V. Demonte (eds.). *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe. 4757-4841.

Varela, S. & Martín García, J. (1999). La prefijación. In I. Bosque. & V. Demonte (eds.). *Gramática Descriptiva de la Lengua Española*. Madrid: Espasa Calpe. 4993–5040.