Uncovering the 'lost' structure of translations with parallel treebanks

Matthias Buch-Kromann, Iørn Korzen, and Henrik Høeg Müller

Abstract

Parallel treebanks provide a systematic way of expressing the structural relationships between source and target texts. In this paper, we present the general design principles behind the Copenhagen Dependency Treebanks, a set of parallel treebanks for Danish, English, German, Italian and Spanish with a unified annotation of morphology, syntax, discourse, and translational equivalence. Finally, we suggest some hypotheses about morphology and discourse, and describe how we plan to explore them empirically on the basis of the treebanks.

1. Introduction

Human translation processes are highly complex. For each translation segment, the translator must analyse the source text in terms of morphology, syntax, discourse, semantics and pragmatics, and generate a target text which invokes a meaning which is as close as possible to the meaning or pragmatic function of the source text.[1] To create a complete

[1] The translation segments may be short word spans in the two texts that correspond closely in meaning, long word spans which correspond only loosely in meaning (for instance, when content has been shifted because of cultural adaptations), or word spans in one text which have no natural counterpart in the other text (explicitations and implicitations). Our notion of translation segments even includes translation errors whose corresponding text segments are easy to identify in the two texts, but where the two segments unintentionally differ in meaning or function. Although our annotation scheme makes it possible to annotate explicitations, implicitations, translation errors and shifts in meaning, and the resulting annotation can be used to easily identify these phenomena when they occur, the annotation does not in itself explain the deeper semantic and pragmatic mechanisms that are responsible for these phenomena, and these mechanisms are not the focus of our investigation.

low-level model of human translation processes, one would therefore need to identify the underlying linguistic structures that the translator assigns to the source and target texts in each instance of time, and their relationships to each other – the 'lost' structure of the translation, which may exist only temporarily in the translator's mind.

Uncovering how the underlying structure unfolds in time seems infeasible with current techniques. Key-logging and eye-tracking (Jakobsen 2006, Carl *et al.* 2008) provide one important part of the puzzle, but in order to produce a plausible model of the 'lost' structure of the translation, these process data must be linked to a model of the detailed linguistic analyses that translators assign to the segments during the translation. Since translators' unconscious processes cannot be observed directly, the most promising alternative source of information is to create a linguistically annotated corpus of translations and their source texts, a so-called parallel treebank. The treebank analyses can be viewed as approximated, reconstructed snapshots of an idealised translator's analyses at the end of the translation.[2] By training monolingual and bilingual parsers on the parallel treebank and incorporating them in detailed low-level models of human translation processes, the parallel treebank can contribute to our understanding of human translation processes. The parallel treebank can also be used to answer a wide range of quantitative and qualitative research questions about translations, such as how often particular structures occur in different languages, how they are mapped to other languages, and what the systematic differences between languages are.

In this paper, we briefly present the Copenhagen Dependency Treebank (CDT) project, an ongoing project which seeks to create a parallel treebank for Danish, English, German, Italian, and Spanish with 40,000 words in each language. The CDT treebanks are an extension of the 100,000-word Copenhagen Danish-English Parallel Dependency Treebank (Kromann 2003, Buch-Kromann *et al.* 2007). Like the original Danish-English treebank, the new CDT treebanks are based on the linguistic

[2] Because of memory limitations, the brains of real translators are unlikely to store all the processed parts of a source and target text along with a corresponding bilingual linguistic analysis. However, we do believe that the brains of real translators are likely to store locally coherent fragments of bilingual linguistic analyses at each instance of time, and that, taken together, these locally coherent fragments will usually (but not always) constitute a globally coherent bilingual linguistic analysis, which functions as the cognitive justification for our idealised translator.

principles of the dependency theory Discontinuous Grammar (Buch-Kromann 2006), which will be outlined in Section 2. They differ from the original parallel treebank in that they incorporate a larger number of languages (German, Italian, and Spanish) and a linguistic annotation which includes not only syntax, but morphology, discourse structure, anaphors, and a much more fine-grained inventory of adverbial relations. The underlying text corpus (Keson and Norling-Christensen 1998) consists of a diverse mixture of general-purpose texts, which have been translated from Danish by a single professional translator for each language. In order to separate "pure" translation from paraphrasing and reformulation, the translators were instructed to produce fluent target-language text while staying as close to the Danish original as possible. Our aim was to maximise the usefulness of the treebanks for machine translation and contrastive linguistics, where large translation segments with loosely coupled meanings are not of much value. However, treebanks are useful even in the study of translations without these restrictions since we can train a bitext parser for any two languages in the parallel treebank, and use the parser to automatically produce treebank annotations for any set of translations between the two languages.

The paper is structured as follows. In Section 2, our unified annotation principles for syntax, morphology, discourse and translational equivalences are described. In Section 3, we proceed to demonstrate how information can be extracted from the treebanks. Section 4 presents some hypotheses about morphology and discourse and shows how the treebanks could be used to explore these hypotheses. Finally, our findings are summed up in Section 5.

2. Unified annotation principles

2.1. The underlying theory

The CDT treebanks are based on the theoretical assumption that human language users interpret a text compositionally by linking all the morphemes, words, and discourse segments in a text or spoken discourse by means of a large inventory of dependency relations, including complement and adjunct relations, supplemented by a set of secondary dependency relations and anaphor-antecedent relations. Figure 1 exemplifies our approach by giving a complete CDT analysis of the morphology, syntax, and discourse structure of the English utterance (1) below. The annotation at the different levels will be explained in detail in the following sections.

(1) Mary was furious. John's missing repayment had forced her to give up all she had worked for. She pulled the trigger. It was payback time.

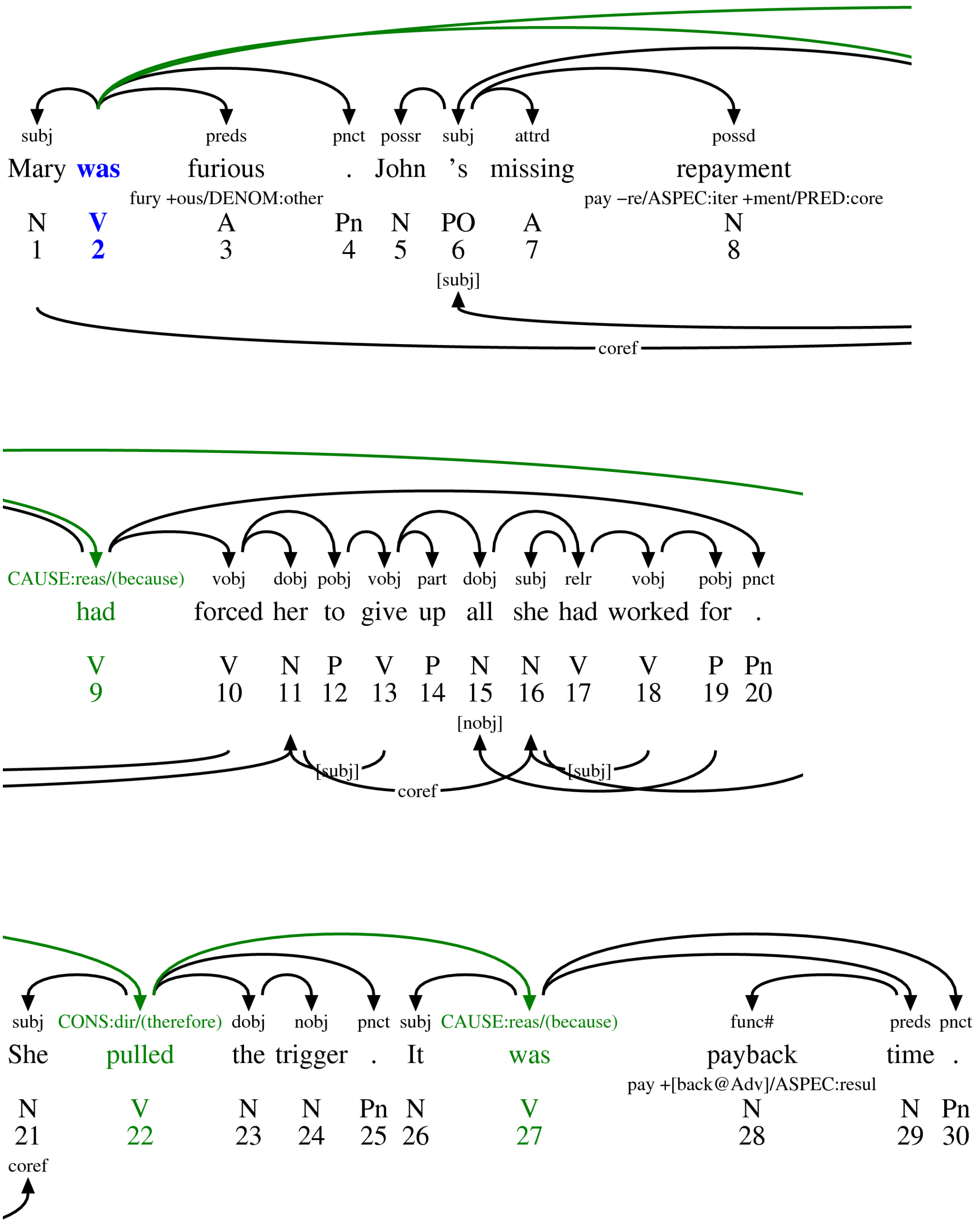

subj preds pnct possr subj attrd possd
Mary was furious . John 's missing repayment
fury +ous/DENOM:other
pay –re/ASPEC:iter +ment/PRED:core
N V A Pn N PO A N
1 2 3 4 5 6 7 8
[subj]
coref
CAUSE:reas/(because) vobj dobj pobj vobj part dobj subj relr vobj pobj pnct
had forced her to give up all she had worked for .
V V N P V P N N V V P Pn
9 10 11 12 13 14 15 16 17 18 19 20
[nobj]
[subj]
coref
[subj]
subj CONS:dir/(therefore) dobj nobj pnct subj CAUSE:reas/(because) func# preds pnct
She pulled the trigger . It was payback time .
pay +[back@Adv]/ASPEC:resul
N V N N Pn N V N N Pn
21 22 23 24 25 26 27 28 29 30
coref

Figure 1. A complete CDT analysis of the entire discourse (1). The labels are explained in the following sections.

The unified annotation distinguishes CDT from other treebank projects, which tend to focus on a single linguistic level. For instance, the Penn Treebank (Marcus *et al.* 1993) and the Prague Dependency Treebank (Böhmová *et al.* 2003) mainly focus on syntax; the Penn Discourse Treebank (Prasad *et al.* 2008a/b) and the RST Treebank (Carlson *et al.* 2001) on discourse, and the GNOME project (Poesio 2004) on coreference annotation. Although a unified annotation scheme is slightly unusual, it offers several advantages.

First, it means that we are able to solve the difficult problem of having to draw a precise boundary between morphology, syntax, and discourse, i.e., we do not have to answer questions such as the following: Are clitic elements viewed as part of syntax or as part of morphology? Are the relations between subordinate clauses and their governing clauses viewed as part of syntax or part of discourse? If the annotation of morphology, syntax and discourse is made independently by different people in different projects, the resulting annotations will almost inevitably have mutually incompatible gaps and overlap at fuzzy boundaries between the different levels. With a unified annotation scheme, we can avoid this problem.

Second, if the annotation is restricted to a single linguistic level, the segmentation will limit the kind of linguistic relations that can be annotated. For example, most discourse theories and discourse treebanks divide the text into segments that correspond to clauses or entire sentences. But this crude segmentation of the text impedes the insight that discourse relations are not always relations between entire clauses or sentences: a long discourse segment consisting of several sentences may actually be better described as an elaboration of a single morpheme, word or short phrase than as an elaboration of the containing clause. An example is given in utterance (2) below, where the second and third sentences are most naturally analysed as elaborations of the NP *their judgment* in the first sentence.

(2) The two convicted executives, chairman Niels Jensen and director Peter Hansen, appealed their judgment with a demand for acquittal. The chairman received a year and a half in jail. The bank's director received 6 months in jail.

The exclusive reliance on a single linguistic level may also blur the insight that many constructions at one linguistic level have almost identical counterparts at other linguistic levels, and that the inventory of linguistic relations should be very similar across the different levels.

Finally, there is a plethora of linguistic theories that could be used to describe the different linguistic levels. Although these theories tend to be based on the same underlying principles, these principles are frequently formulated in ways that may seem almost entirely incompatible with each other, as dictated by the technical machinery that the theory uses to encode complex phenomena such as discontinuous word order and secondary dependencies. While it is obviously impossible to eliminate the reliance on linguistic theory, the use of a unified annotation scheme for morphology, syntax and discourse greatly reduces the number of potential incompatibilities between the different linguistic levels because the annotation is based on a single, coherent underlying linguistic theory.

In the Copenhagen Dependency Treebanks, the smallest segments in the text or spoken discourse are morphemes which are linked into larger units of morphology, syntax and discourse by means of a primary tree structure supplemented by an inventory of secondary relations. Our claim is that this inventory of mechanisms is sufficient to give us a unified account of morphology, syntax and discourse which is theoretically appealing while providing an excellent basis for time efficient large-scale linguistic annotation. In the following sections, we will describe our annotation scheme in more detail.

2.2. Syntax

The syntactic annotation in the CDT treebanks is based on the linguistic principles outlined in the dependency theory Discontinuous Grammar (Buch-Kromann 2006) and the syntactic annotation principles of the Copenhagen Danish-English Parallel Dependency Treebank (Kromann 2003, Buch-Kromann *et al.* 2007). As in other dependency theories, all linguistic relations in the CDT treebanks are represented as directed relations between words or morphemes. We presuppose that the compositional semantics of a sentence is determined by a primary dependency structure in which each word or morpheme is assumed to act as a complement or adjunct to another word, called the governor. Complements function as *arguments* to their governors (i.e., they are lexically licensed by the governor, and the governor's lexical entry specifies how the meaning of the complements should be integrated into the meaning of the governor in the compositional semantics), whereas adjuncts function as *modifiers* (i.e., the governor is lexically licensed by the adjunct, and the adjunct's lexical entry specifies how the meaning of the adjunct should be integrated into the joint meaning of the governor, its complements, and its lower-scoped adjuncts). As we will see later, this distinction between arguments and modifiers applies to dependency relations in morphology and discourse as well.

Figure 2 shows the primary dependency tree associated with the sentence "It had forced her to give up all she had worked for" (the arrows at the top), along with arrows representing secondary dependencies and anaphor-antecedent relations (the arrows at the bottom). The root word in the sentence – the word "had$_2$", which is the only word without any incoming arrows – is shown in bold.

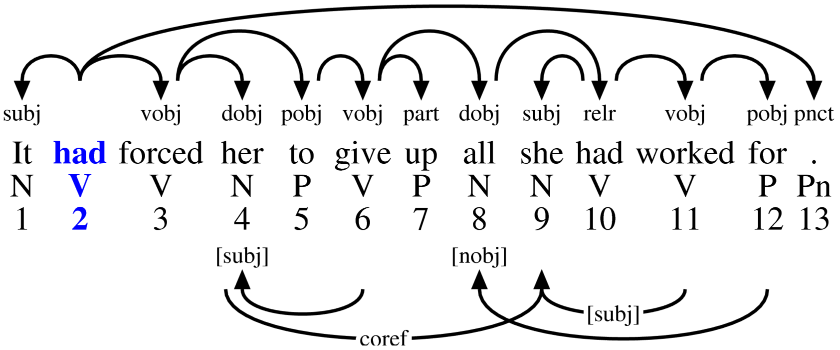


Figure 2. A primary dependency tree (top) augmented with secondary dependencies and antecedent-anaphor relations (below)

The arrows point from governor to dependent, with the relation name written either at the arrow tip, or in the middle of the arrow if a word has more than one incoming arrow. For example, the arrow from "had$_2$" to "It$_1$" identifies "It" as the subject ("subj") of "had", and the arrow from "forced$_3$" to "to$_5$" identifies the phrase headed by "to" as the prepositional object ("pobj") of "forced". The primary dependency relations are organised in a relation hierarchy in which the relations are classified as either complement or adjunct relations. The syntactic annotation scheme currently includes approximately 20 complement relations and 60 adjunct relations. The fifteen most frequent complement and adjunct relations are listed in Table 1.

The primary dependency structure induces a phrase structure in which every word heads an associated phrase consisting of all words that can be reached from the word by following the arrows in the primary dependency structure. For example, in Figure 2, “worked$_{11}$” heads the phrase “worked$_{11}$ for$_{12}$”; “had$_{10}$” heads the phrase “she$_{9}$ had$_{10}$ worked$_{11}$ for$_{12}$”; and “It$_{1}$” heads the phrase “It$_{1}$”.

Table 1. The fifteen most frequent complement and adjunct relations in the syntactic annotation with examples

Complement relations - top 15	**Adjunct relations - top 15**
nobj (nominal object: *for the*$_{nobj}$ *child*$_{nobj}$)	**pnct** (punctuation: *It*$_{subj}$ *is !*$_{pnct}$)
subj (subject: *They*$_{subj}$ *saw me*$_{dobj}$)	**attrr** (restrictive attributive: *la tarea*$_{nobj}$ *difícil*$_{attrr}$)
vobj (verbal object: *He*$_{subj}$ *had left*$_{vobj}$ *it*$_{dobj}$)	**conj** (conjunct: *John and*$_{coord}$ *Mary*$_{conj}$)
dobj (direct object: *He*$_{subj}$ *left us*$_{dobj}$)	**coord** (coordinator: *Tea or*$_{coord}$ *coffee*$_{conj}$)
pobj (prepositional obj.: *one of*$_{pobj}$ *them*$_{nobj}$)	**attrd** (descriptive attributive: *la difícil*$_{attrd}$ *tarea*$_{nobj}$)
preds (subject predic.: *It*$_{subj}$ *was blue*$_{preds}$)	**time** (time adverbial: *We*$_{subj}$ *leave now*$_{time}$)
possd (possessed: *His hat*$_{possd}$)	**loc** (location adverbial: *I*$_{subj}$ *fell here*$_{loc}$)
possr (possessor: *Peter*$_{possr}$ *’s hat*$_{possd}$)	**man** (manner adverbial: *I*$_{subj}$ *read slowly*$_{man}$)
lobj (locative object: *living in*$_{lobj}$ *Rome*$_{nobj}$)	**degr** (degree adverbial: *very*$_{degr}$ *hard*)
qobj (quotational object: *He*$_{subj}$ *said: “*$_{pnct}$ *No*$_{qobj}$ *”*$_{pnct}$)	**neg** (negation: *I*$_{subj}$ *will not*$_{neg}$ *leave*$_{vobj}$)
expl (expletive: *There*$_{expl}$ *arises one*$_{dobj}$ *question*$_{nobj}$)	**namef** (first name: *Igor*$_{namef}$ *Stravinsky*)
predo (object predicative: *We*$_{subj}$ *found it*$_{dobj}$ *disappointing*$_{predo}$)	**relr** (restrictive relative clause: *the cat*$_{nobj}$ *that*$_{subj}$ *died*$_{relr}$)
iobj (indirect object: *We*$_{subj}$ *gave him*$_{iobj}$ *flowers*$_{dobj}$)	**appr** (restrictive apposition: *the genius*$_{nobj}$ *Einstein*$_{appr}$)
avobj (adverbial object: *as before*$_{avobj}$)	**appa** (parenthetic apposition: *Einstein, the*$_{appa}$ *genius*$_{nobj}$)
part (verbal particle: *break up*$_{part}$)	**list** (list sequence: *Johnson, Oklahoma*$_{list}$)

The arrows in the primary dependency structure are allowed to cross, so discontinuous word orders such as topicalisations and extrapositions do not require special treatment, as exemplified by the discontinuous dependency tree in Figure 3 in which the relative clause headed by “was$_{7}$” has been extraposed from the direct object and placed after the time adverbial "today$_5$".

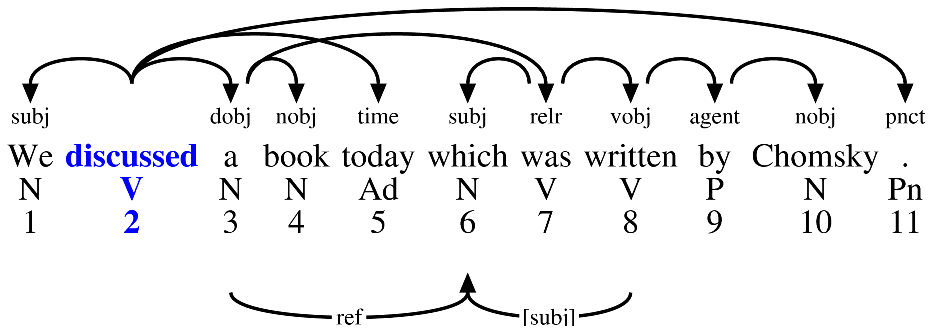
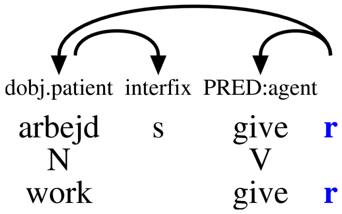


Figure 3. A discontinuous (non-projective) dependency tree with crossing arrows

In addition to the primary dependency structure, the CDT analyses encode secondary dependencies (shown in the bottom part of the analyses along with anaphoric links). In a secondary dependency, a word (called the *licensor*) generates a phonetically empty copy of a second word (the *secondary dependent*) which can then function as a dependent of a third word (the *secondary governor*); loosely speaking, the licensor allows the word to function as a secondary dependent of the secondary governor. These secondary relations are encoded by means of an arrow from the secondary governor to the secondary dependent, with the relation name enclosed in square brackets. For example, the relative clause verb "had$_{10}$" in Figure 2 allows the phrase "all$_8$" to function as a secondary nominal object ("[nobj]") of the preposition "for$_{12}$"; it also allows the word "she$_9$" to function as a secondary subject ("[subj]") of "worked$_{11}$".

2.3. Morphology

The morphological annotation in the CDT treebanks focuses exclusively on derivation and composition, since inflectional morphology can be detected and analysed automatically with high precision for the languages in the treebank. The complex internal structure of words and word-like phrases is encoded as a dependency tree which can be specified in two different ways: either as an ordinary dependency tree (the *dependency notation* in Figure 4, left) or by means of an abstract specification of how the dependency tree for a morphologically complex word is constructed from roots annotated as lemmas in combination with morphological operators (the *operator* *notation* in Figure 4, right). In other words, the dependency notation specifies the tree directly, whereas the operator notation indicates how the tree can be constructed from a set of operators. In the treebank annotation, we use the dependency notation to encode dependency structure between tokens in the automatically produced word tokenisation, while the operator notation is employed to encode dependency structure within tokens.



arbejdsgiver
give +r/PRED:agent –[arbejde@N]s/dobj.patient
employer

Figure 4. Morphological analysis annotated in dependency notation (left) and operator notation (right)

In the operator notation, the dependency tree for a stem is annotated as a root given abstractly by means of its lemma followed by one or more morphological operators: "*lemma op_1 op_2...*" The operators are applied in order, and 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 or the non-head constituent in a compound. A root or stem may be followed by an optional word-class specification "@*wordclass*", indicating a non-standard word class of the stem or modifier (e.g., "pay@V +[back@Adv]/ASPEC: resul @N" as an extreme example).

An operator has the form "*pos affix*/*type*". The field *pos* specifies whether the abstract affix is attached to its base in prefix position ("–"), infix position ("~"), suffix position ("+"), or a combination of these (e.g. "–~", "~+"). The field *type* semantically and functionally identifies the type and, where relevant, the subtype, of the dependency relation that links the base with the abstract affix (e.g., "PRED:agent"). Finally, 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", "arbejd"), 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., "[arbejd@N]s" or "[arbejde@V +r/PRED:agent]"). The interfix is a phonetically induced morpheme whose only function is to act as a glue between the base and the abstract affix. The abstract affix is assumed to function as a dependent of the base, unless it is transformational in terms of triggering word class change or a significant change of meaning, in which case the headedness is reversed so that the base is assumed to function as a dependent of the abstract affix.

The most important relation types in the morphological annotation are listed in Table 2 below (the examples are not always fully analysed, i.e., the real treebank annotation sometimes includes other derivational structures in addition to the exemplified relation). In the table, relation types with head-switching are italicised; the distinction between argument and modifier relations is not shown. The inventory of relations is inspired by Varela & Martín García (1999), Rainer (1999), and Müller (2006).

Examples (3) to (6) below show how the operator notation can be used to encode the dependency structure of complex words.

(3) antihéroe 'antihero' héroe −anti/NEG:oppo

In (3), the Spanish word "antihéroe" is constructed from the root "héroe" by attaching the prefix "anti" as a "NEG:oppo" dependent of the root. The NEG:oppo relation indicates that "anti" negates the meaning of "héroe" so that the new word acquires the opposite meaning of the base.

(4) repayment pay −re/ASPEC:iter +ment/PRED:core

In (4), the word "repayment" is constructed from the root "pay" by first indicating iterative aspect by attaching the prefix "re" to "pay", and then transforming "repay" into a predicative eventive core noun by means of the transformative suffix "ment" which takes "repay" as its dependent.

Table 2. The main relation types in the morphological annotation (relation types with head-switching are italicised)

Relations that typically appear with prefixes

LOC:pos (position: *intramural = mural −intra/LOC:pos*)

LOC:dir (direction/origin: *deverbal = verbal −de/LOC:dir*)

TIME:prec (temporal precedence: *prehistorical = historical −pre/TIME:prec*)

TIME:succ (temporal succession: *postmodernism = modernism −post/TIME:succ*)

NEG:oppo (opposition: *antihero = hero −anti/NEG:oppo*)

NEG:priv (privation: *desalt = salt −de/NEG:priv*)

GRAD:size (size/quantity: *minibar = bar −mini/GRAD:size*)

GRAD:qual (quality: *supercomputer = computer −super/GRAD:qual*)

ASPEC:rev (reversion: *deactivate = activate −de/ASPEC:rev*)

ASPEC:iter (iterative: *redefine = define −re/ASPEC:iter*)

ASPEC:cause (causative: *acallar ‘silence’ = callar −a/ASPEC:cause*)

ASPEC:reflex (reflexive: *autopilot = pilot −auto/ASPEC:reflex*)

ASPEC:term (terminative: *oplåse ‘open’ = låse −op/ASPEC:term*)

ASPEC:resul (resultative: *fastnagle ‘rivet’ = nagle −fast/ASPEC:resul*)

MOD:quant (quantification: *multicultural = cultural −multi/MOD:quant*)

MOD:man (manner: *uneducated = educated −un/MOD:man*)

MOD:qual (qualification: *paleochristian = christian −paleo/MOD:qual*)

TRANS (transitivising: *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*)

PEJ (pejorative: *vinacho ’bad vine’ = vino +acho/PEJ*)

DER:nv (noun→verb derivation: *salar 'to salt' = sal +ar/DER:nv*)

DER:av (adjective→verb derivation: *darken = dark +en/DER:av*)

DER:vv (verb→verb derivation: *adormecer ’lull to sleep’ = dormir −+[a] [ecer]/DER:vv*)

PRED:agent (agent derivation: *singer = sing +er/PRED:agent*)

PRED:core (core derivation: *exploitation = exploit@V +ation/PRED:core*)

QUAL (deadjectival noun: *bitterness = bitter +ness/QUAL*)

NOPRED:agent (agent derivation: *miller = mill +er/NOPRED:agent*)

NOPRED:cont (container derivation: *azucarero ‘sugar bowl’ = azucar +ero/NOPRED:cont*)

DEVERB:pas.poten (deverbal adjective: *transportable = transport +able/DEVERB:pas.poten*)

DENOM:rel.norm **(denominal adjective):** *presidential = president +ial/DENOM:rel.norm*)

Relations that typically appear with compounds

CONST (constitutive: *træbord ‘wooden table’ = bord −træ/CONST*)

AGENT (agent: *politikontrol ‘police control’ = kontrol −politi/AGENT*)

ORIGIN (origin: *rørsukker ‘cane sugar’ = sukker −rør/ORIGIN*)

FUNC (function: *krigsskib ‘war ship’ = skib −[krig]s/FUNC*)

POS (position: *loftlampe ‘ceiling lamp’ = lampe −loft/POS*)

TIME (time: *oktoberregn ‘October rain’ = regn −oktober/TIME*)

ABOUT (theme: *skattelov ‘tax law’ = lov −[skat]te/ABOUT*)

(5) arbejdsgiver ’employer’ give +r/PRED:agent −[arbejde@N]s/dobj.patient

In (5), the Danish word "arbejdsgiver" is constructed from the verb "give" ("give") by turning it into the agent nominalisation (PRED:agent) "giver", headed by the transformative suffix "r", and then specifying the direct object (patient) role of "giver" by means of the prefix dependent "arbejde@N" with interfix "s", where "arbejde@N" denotes the noun reading of "arbejde". (In Danish "arbejde" exists both as a noun and as a verb.)

(6) arbejderkrav 'worker demand' krav −[arbejde@V +r/PRED:agent]/AGENT

In (6), the Danish word "arbejderkrav" ("worker's demand") is constructed from the root noun "krav" ("demand") by combining it with the complex stem "arbejder" by means of an AGENT composition relation.

2.4. Discourse

Just like sentence structures can be seen as dependency trees linking sentence parts such as morphemes and words, discourse structures can be seen as dependency trees linking discourse parts such as sentences and fragments of sentences separated by full stops. The CDT discourse annotation is strongly inspired by Rhetorical Structure Theory (e.g. Mann & Thompson 1987, Matthiessen & Thompson 1988) and the discourse annotation in the Penn Discourse Treebank (Prasad *et al.* 2008a/b; Webber 2004, 2006). A multi-level approach to discourse annotation that resembles our annotation scheme in many respects is found in Mladová *et al.* (2008).

In our view, the single-level approach followed by Rhetorical Structure Theory (RST) and the Penn Discourse Treebank (PDT) results in an overly restrictive view of discourse structure. For example, RST and PDT only consider relations between entire discourse segments, typically clauses, and the PDT annotation only allows for relations between adjacent segments. In contrast, the CDT annotation allows for more precise analyses owing to (i) the unification with the syntax level, which permits annotations of relations between a text segment and a single word or phrase, cf. example (2); (ii) the possibility of discontinuous dependencies, which permits analyses of relations between non-adjacent segments, as in example (2) or in direct or indirect speech separated by attribution text spans.

Furthermore, neither RST nor PDT distinguishes between sentence internal and sentence external relations, but generally include in their analyses clauses that belong to the same rhetorical text sequence. As a consequence of the above, many of the relations investigated by RST and PDT are dealt with at the syntax level in CDT, typically as adverbial adjuncts, our model operating with 32 different types. The CDT discourse annotation therefore focuses exclusively on sentence external relations, including main clauses separated by a colon or semicolon, see Section 4.2. This also means that the inventory of CDT discourse relations differs from the mutually dissimilar RST and DPT inventories. At this point, we have established the primary discourse relations listed in Table 3 below. Those marked with ":*" are abstract types that have a total of 21 subtypes, which are not shown in the table, e.g. "ELAB:exem" for exemplification and "CAUSE:goal" for goal.

Most of the relations have a natural choice of head (nucleus) and dependent (satellite). However, the last five relations are essentially multi-nuclear, i.e., there is no natural choice of head. In the case of CONJ, CONTR, DISJ and JOINT, the two nuclei are merely coordinated. In the case of QUEST, the two segments presuppose each other (Halliday & Hasan 1976: 4). In these relations, the arrows by convention point from the first occurring nucleus to the second one, and "dependency" should not be understood in the sense of a hierarchical relation, but in the sense of a linear relation in which the second segment depends on the first segment simply because it comes last in the textual order. All discourse relations are assumed to be modifier relations.

In order to clearly distinguish between the discourse and syntax levels, the main type in a discourse relation is written with capital letters, whereas subtypes are written with lower-case letters following a colon, e.g. "CAUSE:reas" (reason). In several mono-nuclear relations, the labels are identical to those of adverbial adjuncts at the syntax level. This link between syntax and discourse is by no means coincidental: in fact, mono-nuclear satellites which are textualised as independent sentences in a Germanic language are frequently incorporated into a single large sentence in a Romance language.

Table 3. The main relation types in the discourse annotation with examples (dependent sentence is underlined)

CAUSE:* (cause: *We should be ashamed of the state of our schools. We are one of the world's richest countries.*)
CONC (concession: *I know we haven't known each other for long. Will you marry me?*)
COND (condition: *OK, I will help you. On the condition that this never happens again.*)
CONS:* (consequence: *In three years many of the DSB ferries will have to seek new waters. An exciting story will be history.*)
CONSOL:* (consolidation: *One of the country's legal experts in cooperative housing is Svend Trangeled. According to him the greatest problems lie in partnership arrangements.*)
DESCR:* (description: *Let's go to your place, she said. She was different than I thought.*)
ELAB:* (elaboration: *She was different than I thought. In the taxi she put her head on my shoulder.*)
INTACT:* (interactional signals: – *There is something I would like to ask you.* – *Yes?*)
PREPAR (preparation, headline: *A permanent job. I am so happy that I do not have to go on welfare again, says Lisa…*)
TIME:* (temporal: *The Ministry of the Interior expects this part of the job to be concluded before 2010. Then the Ministry and the national board of health will consider whether the work is to be conducted further.*)
CONJ:* (conjunction: *The bank's director was sentenced to six months in prison. Two board members were acquitted.*)
CONTR:* (contrast: *The societal savings are not immediately visible. But it is a fact that new drugs will minimise the overall health cost.*)
DISJ (disjunction: *Are you coming up now? Or are you going to stay in bed all day?*)
JOINT (no clear discourse relation: *We are hungry and are refused by all restaurants. On a side street a façade is blinking in green neon.*)
QUEST:* (question/answer sequence: *Why did he say that? I have no idea.*)

Inspired by the Penn Discourse Treebank (especially Webber 2004, 2006 and Prasad *et al.* 2006), we annotate explicit and implicit connectives as well as attribution. Explicit connectives are specified in the dependency label after a slash sign, e.g. "CAUSE:reas/because". Implicit but inferable connectives are specified between parentheses, e.g. "CAUSE:reas/(because)". The attribution of direct speech, indirect speech, thoughts, hopes, ideas, etc. is encoded with an "/ATTR" following the relation label and, in case of more than one speaker, with "/ATTR1", "/ATTR2", etc. Although less detailed than the attribution system in PDT, this system captures much of the same information.

Since discourse is generally extremely open to doubts, misunderstandings and ambiguity, there is often more than one interpretation of the discourse, and hence also more than one annotation of the relation. The CDT annotation therefore allows relations to be combined by means of "|" and "&", to signal doubt between two relations and double meaning, respectively.

In addition to the primary discourse relations, we also annotate anaphoric links as arrows pointing from the antecedent to the anaphor (or from the postcedent to the cataphor, in the case of cataphors). Like secondary dependency relations in syntax, these relations are shown in the bottom part of the analyses, as illustrated in Figure 1. The inventory of anaphoric relations is shown in Table 4. We distinguish between coreferential and associative anaphors. Coreferential anaphors (coref-*) refer to the same entity as their antecedent, whereas associative anaphors (assoc-*) refer to an entity which is associated with the antecedent by means of a lexico-semantic or pragmatic relation.

Table 4. Anaphoric relations in the CDT discourse annotation

ref (syntactically bound coreference)

coref (coreferential pronouns)
coref-id (NP with lexically identical antecedent)
coref-var (NP with lexical variation: *a car → the vehicle*)
coref-prg (NP indicating speech act: *I shall be back tomorrow → the threat / the promise / the warning*)
coref-res (resumptive anaphors: *John sold his car → that / that sale*)
coref-part (partial coreference: *a bouquet of flowers → the roses*)

assoc (associative pronouns: *We went to a restaurant. → They [= the waiters] were very quick in serving us*)
assoc-const (refers to constitutive part of antecedent: *a car → the engine / the windscreen*)
assoc-agent (refers to agentive entity: *a car → the factory / the manufacturer*)
assoc-form (refers to form of antecedent: *a car → the shape / the size*)
assoc-scope (refers to scope or purpose entity: *a car → the chauffeur / the passengers*)
assoc-loc (the anaphor is located in the antecedent: *a village → the church / the inn*)
assoc-cause (anaphors such as *for this reason, on this basis,* where the antecedent is the preceding text segment)

2.5. Translational equivalence

In order to extend the CDT annotation system to translations, we need to specify the translational equivalences between the source and target texts, i.e., the minimal word groups in the source and target language that correspond to each other with respect to meaning or function. In a dependency-based theory, these equivalences can be encoded straightforwardly by means of a many-to-many word alignment, in which a group of (possibly non-contiguous) source words is aligned to a group of target words. Special properties of the alignment are encoded with an inventory of approximately 20 alignment relations. The most important ones are listed in Table 5.

Table 5. The most important alignment relations

none (normal alignment)
d (change in determination: *bread* ↔ *the bread*)
e (translation error: *1978* ↔ *1987*)
f (free/fuzzy translation: *that day* ↔ *that morning*)
lv (lexical variation – change of lexeme with same referent: *Maria* ↔ *the girl*)
mv (morphological variation – change in inflection: see section 4.2)
mc (morph. change – change in word class: *Upon his arrival* ↔ *When he arrived*)
n (change in number: *his contribution* ↔ *his contributions*)
s (syntactic alignment – syntactic marker aligns with subsegment: *“dagen (day-the)”*↔ *“the day”: normal alignment “dagen”* ↔ *“day”, s-alignment “dagen”* ↔ *“the”*)
wo (marked change in word order: *John arrived early that morning* ↔ *That morning, John arrived early*: see section 4.2)

Deletions (implicitations) are encoded as alignments in which a group of source words is aligned with itself, and similarly for added text (explicitations). In this way, a complete CDT analysis of a source text and its translation can be given by means of monolingual CDT analyses of the source and target texts, coupled with a word alignment that links the two monolingual analyses. Figure 5 gives an example of this kind of analysis, in which the English source text is shown at the top, the Danish translation at the bottom, and the word alignment in between the two analyses.

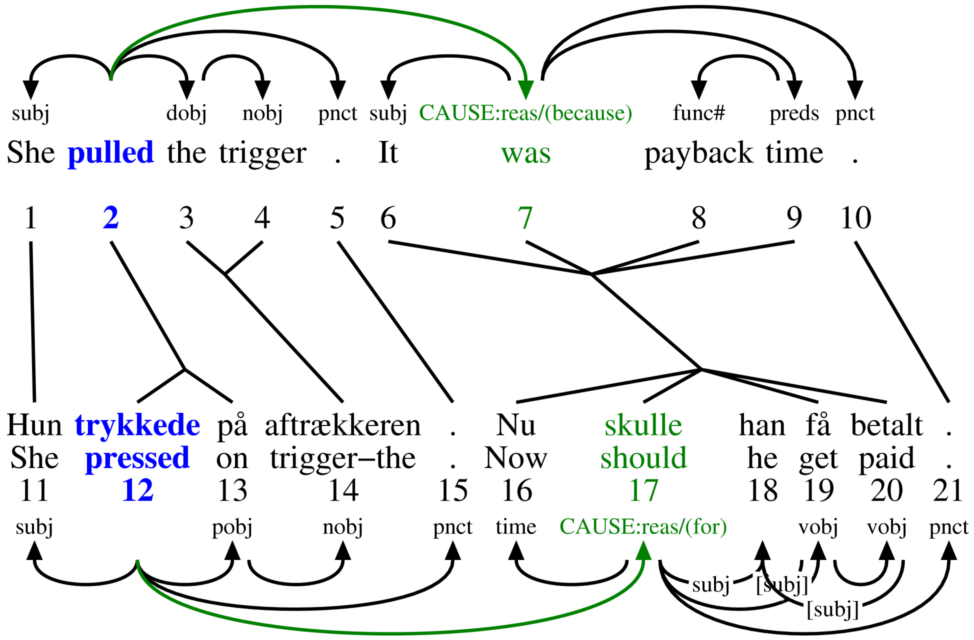


Figure 5. A complete CDT analysis of a translation with source analysis (top), target analysis (bottom), and word alignment (middle)

3. Searching the treebanks

The treebanks can be searched for linguistic constructions that match a given set of criteria by means of the DTAG dependency treebank annotation tool (Kroman 2003). Nodes in the graph are represented by variables of the form "$node" (alignment edges are encoded as nodes as well), and queries express a set of conditions that the node variables must satisfy. The most important query types are explained below.

- $node_1$ *@* $node_2$ and $node_1$ *@ pattern* $node_2$ hold if $node_1$ is aligned to $node_2$ by means of an alignment edge whose type matches *pattern*. Patterns are either names which must be matched exactly, or regular expressions.
- $node_1$ *pattern* $node_2$ holds if there is an arc from $node_1$ to $node_2$ whose type matches *pattern*.
- *feature =~ pattern* holds if *feature* matches *pattern*. A feature of the form *node* represents the position associated with the node in numerical contexts and the string associated with *node* in string contexts, and the feature *node[name]* represents the value of attribute *name* at *node*.

- *number$_1$ numop number$_2$* holds if the numerical operator *numop* holds between *number$_1$* and *number$_2$*. Numerical operators include equality, less-than, greater-than, etc. as well as the adjacency operators "<<" (one before) and ">>" (one after). Numbers can be given as integers or node features.
- Complex queries can be formed by means of the logical operators "and", "or", "not", "exists", "all".

For example, the DTAG query:

```
(6)   find ( $X[lemma] =~ /^there$/ ) and ( $X expl $Y ) and ( $Y < $X )
```

finds all occurrences of the expletive word $X with lemma "there" which has been analysed as an "expl" dependent of a preceding word $Y. The DTAG query language makes it easy to search a parallel treebank for particular phenomena, allowing linguists to find quantitative and qualitative corpus-based evidence for linguistic hypotheses that would be difficult to investigate otherwise.

4. Examining linguistic hypotheses

By means of the DTAG query language, the CDT treebanks can be used to prove or disprove a wide range of linguistic hypotheses. The following two sections outline such hypotheses, one for NP structure and one for discourse.

4.1. NP structure

A frequently observed difference between Germanic and Romance languages is that Germanic languages often use compounding to express what Romance languages convey by a derivational strategy (Bally 1932, Rainer & Varela 1992). This means that many simple and derived words in Spanish have compounds as their translational equivalents in Danish, cf. (7) and (8) below.

	Derivation → compound	
(7a)	escritorio – skrivebord	‘writing desk’ (escribir: ‘write’)
(7b)	dentadura – tandsæt	‘set of teeth’ (diente: ‘tooth’)
(7c)	petrolero – olietankskib/oliehandler	‘oil dealer/oil tanker’ (petróleo: ‘oil’)

	Simple noun → compound	
(8a)	berberecho – hjertemusling	‘cockle’
(8b)	búho – hornugle	‘horned owl’
(8c)	púlpito – prædikestol	‘pulpit’

We believe that we can account for this cross-linguistic contrast by means of the following lexical-typological hypothesis.

Romance (exocentric) mainly artefact-denoting nouns are in general more contentful and precise with respect to their lexical meaning than the corresponding Germanic (endocentric) ones. (For an account of the typological theory of endocentric and exocentric languages, see e.g. Herslund & Baron 2005.) Romance artefact-denoting nouns tend to lexicalise the semantic component figure, i.e. the shape, dimensionality and structure of the object. By contrast, Germanic artefact-denoting nouns tend to lexicalise only the component function, which is an inherent abstract feature of any artefact-denoting noun. The exclusive focus in the Germanic languages on the purpose of the object, or non-focus on its form, means that Germanic simple nouns are in many cases semantically vague. This allows them to function as denominations on a generic prototype level, i.e. a general hyperonymic level, which does not have any corresponding form in the Romance languages.

In particular, when Danish speakers need a level below the general hyperonymic one, they usually achieve it by means of nominal compounds as shown in (9). The Danish noun *tæppe* (‘carpet’) represents lexicalisation on the family level, whereas Spanish nouns must denote subtypes, i.e. objects on a hyponymic level (for Italian, see e.g. Korzen 2008).

(9)	**Danish**	**Spanish**	**lit. transl.**
	tæppe	[Ø]	–
	*senge**tæppe***	*colcha*	‘bed –’
	*væg**tæppe***	*tapiz*	‘wall –’
	ægte ***tæppe***	*alfombra*	‘genuine –’
	væg-til-væg ***tæppe***	*moqueta*	‘wall to wall –’
	*teater**tæppe**/ scene**tæppe***	*telón*	‘theater/ stage –’
	*slumre**tæppe***	*manta*	‘slumber –’

We assume that Danish has to use the compounding system in order to designate entities on a hyponymic level because of the semantic vagueness of Danish simple nouns and their subsequent lexicalisation on a hyperonymic family level. In other words, composition should be very frequent in Danish as compared with Spanish, and thus we would expect it to be incorporated into the grammatical system as a highly automated morphological word-formation process.

By contrast, Spanish simple nouns are already saturated in a semantic sense, so the Spanish language does not need, and has not developed, a full morphological system to deal with this information-packaging task. Either the semantic components are already encapsulated in the simple noun or an alternative strategy is used, namely derivation. Although phrasal composition of the [N prep. N]-type is also often a prerequisite for creating subtype-denoting lexical expressions in the Romance languages, it can be regarded as an addition to the derivational system and semantically contentful nouns. Therefore composition in the Romance languages has not been routinised as part of a morphological system.

This hypothesis is debatable for a number of reasons; we shall restrict ourselves to mentioning a few general ones: (i) the general characterisation of Danish nouns as semantically weak is postulated on the basis of a specific sub-set of nouns, namely those that denote artefacts and typically appear to be equivalent to Spanish simple or derived forms; (ii) it is not immediately evident that a generalisation obtained from this sub-set is valid for all Danish compounds, which in many cases can have heads with lexically strong meanings; (iii) the status of Romance [N prep. N]-constructions is unclear as to whether and in which cases they should be regarded as compounds or free syntactic phrase formations, as the criteria for compounding are not universal or language independent.

In CDT, the combination of morphological annotation and alignment of parallel texts allows us to make specific inquiries into the assumed morphological correlation so that we can either substantiate our hypothesis statistically, reject it, or provide qualitative data that can lead to its refinement and, generally, new insights into morphological cross-linguistic contrasts. Also, the CDT annotation of several Romance and Germanic languages provides data for a more detailed and differentiated assessment of typological variations in the morphological structure of the languages in question. Finally, in order to address the specific problem of Romance phrase formation vs. compounding, the CDT marks all word-like phrases with a hash symbol and decomposes all solid compounds, which will provide us with a firm foundation for "solving the puzzle" of Romance word formation by means of free morphemes.

4.2. Discourse

Romance syntax and discourse generally appears to be more complex and hypotactic than Scandinavian syntax and discourse. Romance sentences tend to be longer and to contain more information, and Romance discourse structure is more hierarchical, i.e., more characterised by structural foreground-background distinctions rendered for example by finite vs. non-finite verb forms.

These differences have been documented in various linguistic studies (e.g. Korzen 2003, 2006), but we expect the CDT combination of discourse annotation and alignment of parallel texts to be able to provide us with a much more precise picture of the situation. For example, the CDT annotation can be used to examine the higher frequency of non-finite verb forms in Romance because it can be used to detect all changes from finite to non-finite verbs and vice versa.

The examples below are typical of translations from Danish into Italian (10) to (12) and from Italian into Danish (13):

(10) Den sidste der kommer – L'ultimo a *venire*… [infinitive]
'The last one who is coming' – 'The last one *to come*'

(11) Han kom og bad om en is – Arrivò *chiedendo* un gelato. [gerund]
'He came and asked for an ice cream' – 'He came *asking* for an ice cream'

(12) Da han kom til Rom, blev han syg – *Arrivato* a Roma, si ammalò. [participle]
'When he came to Rome, he became ill' – '[Having] *arrived* in Rome, he became ill.

(13) Nato nel 1947, ... – Han *blev født* i 1947, og … [finite verb]
'Born in 1947, …' – 'He *was born* in 1947, and …'.

The CDT annotation will provide very precise answers regarding the frequency of such changes between verb forms (as well as changes between word classes) and the contexts in which they occur, and will also inform us whether particular verb types, text types or discourse relations occur more frequently than others in specific morphological data.

Another important difference between the Scandinavian and Romance languages, linked to differences in sentence length, is the different use of paragraph change and punctuation marks. Together with longer sentences, there seems to be a tendency, at least in some text types, also to use longer paragraphs in Romance languages while Scandinavian texts tend to subdivide the content into smaller units. Furthermore, the higher proportion of long sentences can be linked – among many other things – to a much more frequent use (especially in French and Italian) of the colon and semicolon in cases where a Danish text would select a full stop. But these phenomena need much more investigation, and therefore the CDT discourse annotation distinguishes between different kinds of sentence and main clause adjoining. If a full stop coincides with a change of paragraph, the sign + is added to the name of the discourse relation. A very frequent relation is the simple conjunction, cf. Table 3, so that a relatively common discourse relation label is "+CONJ/(and)". Usually a connective is not rendered explicit in these cases, hence the parentheses around "and".

Differences in the use of punctuation marks will automatically be detected in CDT, as the system includes an individual annotation of punctuation marks at the syntactic level, cf. Figure 1 (units 4, 20, 25 and 30), and the cross-linguistic alignment in Figure 5 (units 5/15 and 10/21). But in addition to that, since the use of colon and semicolon varies particularly between the Scandinavian and Romance languages, CDT discourse annotation adds a colon or semicolon to the relation in cases where two main clauses are separated by a colon or a semicolon, e.g. ":CONJ/(and)", ";CONJ/(and)".

A final issue concerns differences in word order, where a particularly interesting phenomenon is the Romance tendency to begin sentences with a time, place or modal adverbial vs. the Danish tendency to begin sentences with the subject – a phenomenon which to our knowledge has not yet been investigated. Such cases can be detected effortlessly by means of the CDT annotation and will therefore be easy to search for and analyse both quantitatively and qualitatively once the treebank annotation is complete.

5. Conclusion

In this paper, we have described the main aspects of the annotation scheme in the Copenhagen Dependency Treebanks, and demonstrated that a unified scheme solves many of the consistency problems that arise if the annotations are performed separately. We have argued that treebanks are an important resource for research in linguistics, psycholinguistics and translation studies, and outlined several linguistic hypotheses about translations and contrastive linguistics that can be explored with the CDT treebanks.

6. Acknowledgments

This work was supported by grants from the Danish Research Council for the Humanities and the Copenhagen Business School. The translation work has been kindly co-sponsored by the Danish translation company Lingtech. We thank Lotte Jelsbech Knudsen, Morten Gylling, and the other CDT annotators for many fruitful discussions.

References

Bally, C. 1932. *Linguistique générale et linguistique française*. Berne: Francke.

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., Wedekind, J. & Elming, J. 2007. The Copenhagen Danish-English Dependency Treebank v. 2.0. URL http://buch-kromann.dk/matthias/cdt2.0.

Carl, M., Jakobsen, A. L. & Jensen, K. T. H. 2008. Modelling human translator behaviour with user-activity data. In *Proceedings of the 12th EAMT Conference. 22–23 September 2008, Hamburg, Germany*. 2008. 21–26.

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.*

Halliday, M.A.K. & Hasan, R. 1976. *Cohesion in English*. London: Longman.

Herslund, M. & Baron, I. 2005. Langues endocentriques et langues exocentriques. Approche typologique du danois, du français et de l'anglais. In M. Herslund & I. Baron (eds). *Le Génie de la Langue Française. Perspectives typologiques et contrastives*. (Langue française 145) Paris: Larousse. 35–53.

Jakobsen, A. L. 2006. Research methods in translation – Translog. In K. P. H. Sullivan and E. Lindgren (eds). *Computer Keystroke Logging and Writing: Methods and Applications*. Oxford: Elsevier. 95–105.

Keson, B. and Norling-Christensen, O. 1998. PAROLE-DK. The Danish Society for Language and Literature.

Korzen, I. 2003. Hierarchy vs. linearity. Some considerations on the relation between context and text with evidence from Italian and Danish. In I. Baron (ed.). *Language and Culture*. (Copenhagen Studies in Language 29). Copenhagen: Samfundslitteratur. 97–109.

Korzen, I. 2006. Endocentric and exocentric languages in translation. *Perspectives: Studies in Translatology* 13(1): 21–37.

Korzen, I. 2008. Determination in endocentric and exocentric languages. With evidence primarily from Danish and Italian. In H. H. Müller & A. Klinge (eds). *Essays on Nominal Determination. From Morphology to Discourse Management*. (Studies in Language Companion Series (SLCS) 99). Amsterdam/Philadelphia: John Benjamins. 69–100.

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.

Mann, W. C. & S. A. Thompson 1987. *Rhetorical Structure Theory. A Theory of Text Organization*. Los Angeles (CA): ISI Information Sciences Institute, ISI/RS-87-190. 1–81.

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.

Matthiessen, C. & Thompson, S. A. 1988. The structure of discourse and 'subordination'. In J. Haiman & S. A. Thompson (eds). *Clause Combining in Grammar and Discourse*. Amsterdam/Philadelphia, John Benjamins. 275–329.

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. 2006. Nominalkomposition i moderne spansk. En teori om betydningsdannelse. [Nominal composition in modern Spanish. A theory of meaning construction] Doctoral thesis. Copenhagen: Copenhagen Business School. 512 pages.

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

Prasad, R., Dinesh, N., Lee, A., Joshi, A. & Webber, B. 2006. Annotating attribution in the Penn Discourse TreeBank. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text.* 31–38.

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).*

Rainer, F. 1999. La derivación adjectival. In I. Bosque. & V. Demonte (eds). *Gramática Descriptiva de la Lengua Española*. Madrid: Espasa. Volume 3, chapter 70, 4595–4643.

Rainer, F. & Varela, S. 1992. Compounding in Spanish. *Rivista di Linguistica* 4(1): 117–142.

Varela, S. & Martín García, J. 1999. La prefijación. In I. Bosque. & V. Demonte (eds.). *Gramática Descriptiva de la Lengua Española*. Madrid: Espasa, volume 3, chapter 76, 4993–5040.

Webber, B. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science* 28: 751–779.

Webber, B. 2006. Accounting for discourse relation: constituency and dependency. In M. Dalrymple (ed.). Festschrift for Ron Kaplan. CSLI Publications.