The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks

Copenhagen Business School
HANDELSHØJSKOLEN

Matthias Buch-Kromann <mbk.isv@cbs.dk> and Iørn Korzen <ik.ikk@cbs.dk>. Center for Research and Innovation in Translation and Translation Technology, Copenhagen Business School, Denmark.

GOAL: Unified model of syntax, discourse, coreference, morphology, semantics, and word alignment

The annotation is based on the dependency theory Discontinuous Grammar (Buch-Kromann 2006). The annotation provides coherent, unified analyses of all six linguistic levels according to the theory. The theory provides a formal interpretation of the annotation, including a compositional semantics, a word order theory, and a dependency-based theory of the underlying lexicon and grammar. The annotation is manual and semi-automatic.

Current annotation status

	Syn.	Disc .	Coref.
Relations	83	50	16
Agreement	83%	(50%)	(97%)
Ann. tokens	85k	45k	45k

Parenthesized agreements are preliminary. See www.treebank.dk for latest releases, guidelines, status, agreement statistics, and DTAG ann. tool. Completion date: early 2011.

Segmentation

The text is segmented into *lexical elements* such as words, multi-word units, and morphemes (not shown). Complex segments are defined by means of phrases.

Phrases

Each lexical element heads a *phrase* consisting of all nodes that can be reached from it by following the arcs in the primary tree. Eg, “executives” → “executives of the July 6 Bank”.

Phrase-structure trees

Dependency trees can be viewed as (possibly discontinuous) phrase-structure trees in which every phrase has a lexical head. Conversion can be automatic (not shown).

Text corpus

Balanced written mixed-genre 100k corpus for Danish consisting of 200-250 word excerpts. Native translation into English (100k words) and German-Italian-Spanish (50k words).

Dependency graphs

All linguistic relations are modelled as binary directed labeled relations between lexical elements. Arcs go from head to dependent element. Relation names written at arrow tip.

Primary dependency tree

Top arcs. Link head word to its complements/adjuncts. Determine compositional semantics. May cross.

Discourse relations

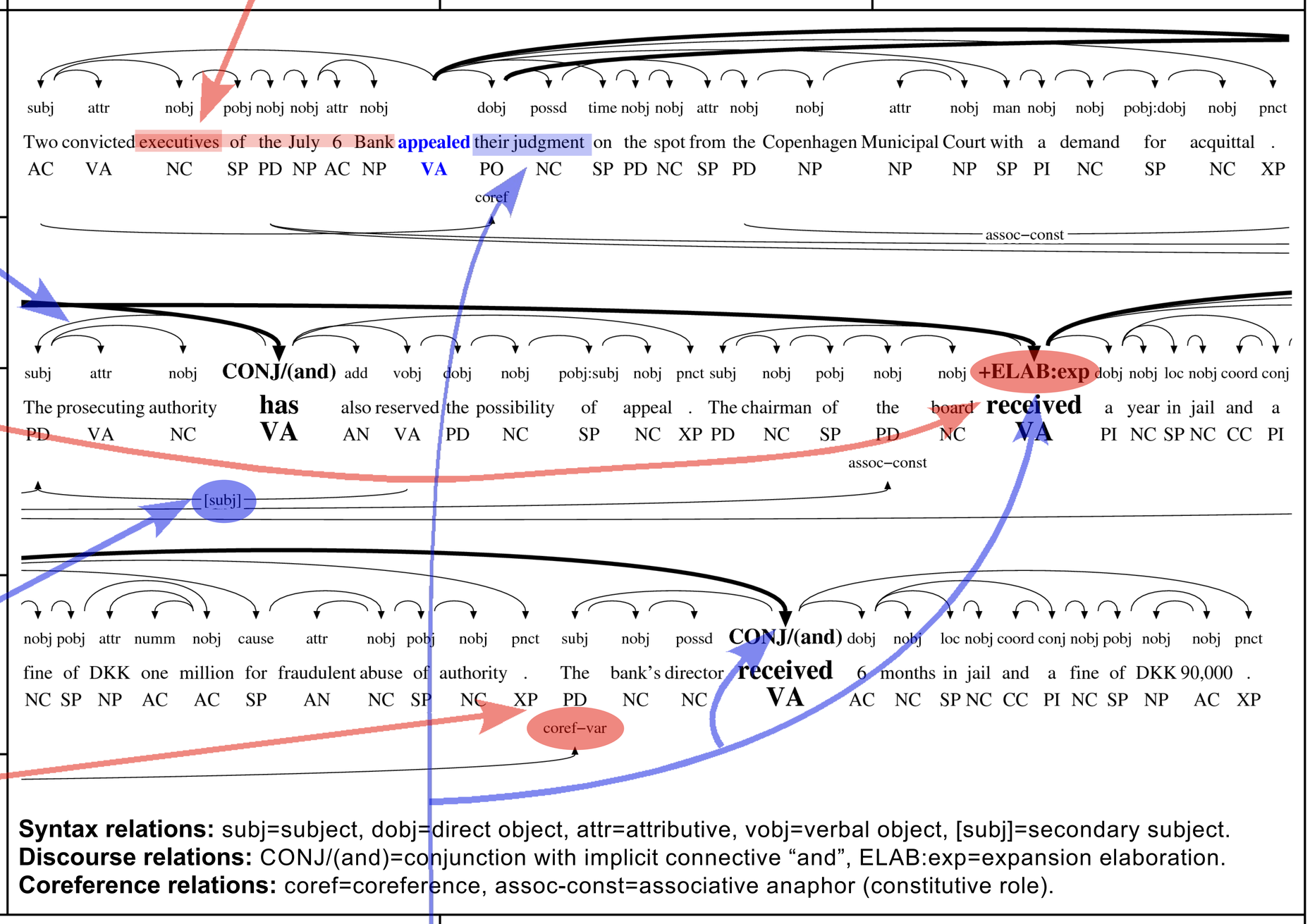
Top arcs. Viewed as primary dependency relations linking sentences into discourses by adjunction. Discourse phrases may function as syntactic complements (eg, in attribution).

Secondary dependencies

Bottom arcs. Used when lexical elements satisfy more than one dependency role (eg, double subjects “[subj]” in verbal chains, relatives).

Coreference relations

Bottom arcs. Go from antecedent to anaphoric element. Associative anaphors annotated as well.



Syntax relations: subj=subject, dobj=direct object, attr=attributive, vobj=verbal object, [subj]=secondary subject.
Discourse relations: CONJ/(and)=conjunction with implicit connective “and”, ELAB:exp=expansion elaboration.
Coreference relations: coref=coreference, assoc-const=associative anaphor (constitutive role).

Problems and solutions

PROBLEM: Fuzzy syntax-discourse boundary?

(a) Discourse segments hard to define: eg, continuum from “He laughed. That annoyed me” to “His laugh annoyed me.” (b) An entire discourse may function as a syntactic complement (eg, attribution). (c) An entire discourse may elaborate an NP rather than a clause, eg, the two last sentences in our example are best described as elaborations of “their judgment”.

SOLUTION: Unified syntax-discourse model

The boundary problems disappear if we use a unified syntax-discourse annotation model: all elementary segments represent lexical elements, and all other segments are induced as phrases. This results in a minimal, principled, and clearly defined set of segments: the artificial boundary between syntax and discourse disappears, and we can easily encode relations between lexical segments and discourse segments (=multi-sentence phrases).

PROBLEM: Annotation based on syntax or semantics?

Syntax-based annotation often diverges from semantics-based annotation. Eg, discourse connectives are analyzed as conjunctions or markers in nearly all syntactic theories, but are frequently analyzed as functors (heads) in discourse frameworks (eg, PDT). The three analyses are summarized below.

	Functor analysis	Conjunction anal.	Marker analysis
Syntax	comp comp X **C** Y	mod comp **X** C Y	mod mod **X** C Y
Semantics	C'(X',Y')	[C'(Y')] (X')	[Y'(C')] (X')

SOLUTION: Explicit syntax, derived semantics

Syntactic theories make sharp distinction between constituent structure (syntax) and functor-argument structure (semantics). CDT discourse annotation is syntactic, using marker or conjunction analysis for discourse connectives. Semantic structure (functor-argument structure) can be derived uniquely from primary tree given modifier scope; its semantics agrees with PTB analysis.

PROBLEM: Discourse structures: trees or graphs?

The structure of discourse is debated. Is it best described as trees or graphs (eg, Wolf-Gibson vs. Marcu)? Are crossing branches allowed or not?

SOLUTION: Trees with secondary relations

Syntactic structure modelled as tree + other relations in all syntactic theories, where tree provides interface to compositional semantics and word order. No reason why syntax stops at the sentence boundary, so CDT simply extends it to entire discourse. CDT treebanks demonstrate that this works in practice.

PROBLEM: What do the annotations mean?

Text is observable data, but linguistic annotations are unobservable theoretical constructs that only make sense if linked to a theory that relates the constructs to some observable properties of human languages. Without explicit theory, users and annotators resort to ad-hoc/post-hoc interpretation.

SOLUTION: Theory-based annotation

Discontinuous Grammar provides the link to linguistic theory (compositional semantics, word order, lexicon) needed to make sense of the annotations.

Poster presented at The Fourth Linguistic Annotation Workshop (LAW IV), ACL-2010, Uppsala, Sweden, 15-16 July 2010.