Danish English Treebank – CDT2

1 BASIC INFORMATION

1.1 Resource composition

Parallel treebank

1.2 Representation of the resource (flat files, database, markup)

Flat files with markup

1.3 Character encoding

ISO

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Name: Matthias Buch-Kromann,

E-mail: Matthias@Buch-Kromann.dk

2.2 Copyright statement and information on IPR

GNU GPL v3.0

3 TECHNICAL INFORMATION

3.1 Data structure of an entry

Dependency-annotated aligned sentences

3.2 Resource size (num. of rules)

95.000 word tokens

4 CONTENT INFORMATION

4.1 Type of the resource (language (in(dependent)

Parallel treebank, created on the basis of the dependency-based grammar formalism Discontinuous Grammar (Buch-Kromann 2009). Texts are analyzed as a single dependency structure that includes morphology and syntactic dependency.

4.2 The natural language(s) for the resource is applicable (if language dependent)

Danish – English

4. 3 Domain(s)/register(s) of the resource

Danish – English PAROLE corpus

4.4 Annotations in the resource (if an annotated resource)

4.4.1 Types of annotations

Part-of-speech, syntax (dependency relations)

4.4.2 Tags (if POS/MSD/TIME/discourse/etc –tagged or parsed),

POS-tagged with the PAROLE tag set. The list of dependency relations is contained in annotation manual: http://code.google.com/p/copenhagen-dependency-treebank/wiki/CDT

4.5 Intended application of the resource

Training of natural language parsers, syntax-based machine translation systems, and other statistically based natural language applications

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Semiautomatic annotation and alignment

5 RELEVANT REFERENCES AND OTHER INFORMATION

Matthias Buch-Kromann, Jürgen Wedekind, and Jakob Elming, 2007. *The Copenhagen Danish-English Dependency Treebank v. 2.0.* Parallel dependency treebank for Danish-English with 100,000 words based on the Danish Dependency Treebank.