LASSY: LARGE SCALE SYNTACTIC ANNOTATION OF WRITTEN DUTCH

Deliverable 1-2: Corpus Selection LASSY Large

material	size	IPC
SONAR release 1	156M	settled in SONAR
XMLWiki Wikipedia 2008	110M	GFDL
Europarl version 3	38M	public
TwNC newspaper material	531M	ok for research; unclear otherwise
Mediargus newspaper material	1397M	ok for research; unclear otherwise
EMEA European Medicines Agency	14M	public

Table 1: Corpus selection Lassy Large. Information about the copyright status of the material from the European Medicines Agency can be obtained from `http://www.emea.europa.eu/htms/technical/dmp/copyrite.htm#copyright`

1 Background

Lassy Large is the large automatically syntactically annotated corpus. It is not manually verified.
The corpus will contain the same three manually verified annotation layers as Lassy Small, namely POS-tags, Lemma, and syntactic dependency annotation.

In this document, we define in more detail which texts will be included in the Lassy Large corpus.

The word counts listed in the current document list the number of tokens (this includes for instance punctuation). Note that the summed word counts exceed the 500 million words originally promised in the project proposal.

2 Motivation

We have included SONAR data, as originally foreseen. In addition we have included some other large Dutch corpora that were readily available to us.

There is some overlap between some of the corpora listed here. For instance, SONAR contains sections with Wikipedia as well as with Europarl data.

The SONAR data is obtained from the SONAR project.

The EMEA texts are selected, cleaned, and tokenized in collaboration with the Paco-MT STEVIN project.

If further data from the SONAR project is available in time, this data will also be included.