A Dutch Treat for Healthcare Terminology

Ronald Cornet[1,2]

[1] *Academic Medical Center - University of Amsterdam, Department of Medical Informatics, P.O. Box 22700, 1100 DE Amsterdam, The Netherlands*
[2] *Linköping University, Department of Biomedical Engineering, University Hospital, S-581 85 Linköping, Sweden*
r.cornet@amc.uva.nl, ronald.cornet@liu.se

Structured and encoded information are important to maximize the meaningful (re)use of the Electronic Health Record (EHR). SNOMED CT is generally regarded as the preferred terminology system for encoding, but it has been shown that manual encoding (i.e., fully structured data entry) has issues with data quality and usability. Therefore, automated SNOMED CT encoding of free-text clinical narratives needs to be explored, which involves both post-hoc processing of yet unstructured records and ad-hoc processing of text being entered into a record.

Processing requires thesauri and tools which are apt for the clinical language being used. This poses a problem, as tools are to a large extent language dependent, and thesauri may not be available for the language of interest.

Therefore, we have created an inventory of components for processing Dutch natural language, enabling to encode Dutch text as structured SNOMED CT output. This inventory distinguishes language-independent and language-dependent components, according to the pipeline depicted in Figure 1.

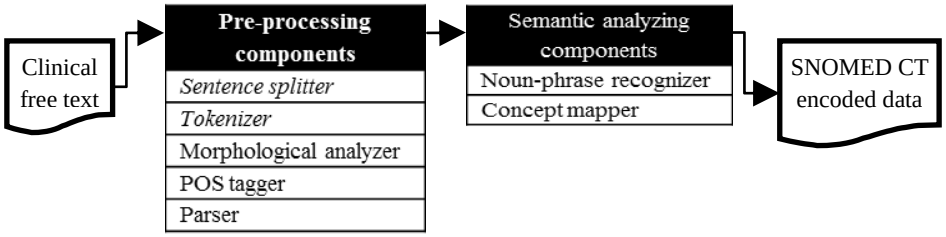

Figure 1. Components of an NLP pipeline. Language-independent components are in italic font.

Table 1 below summarizes the available tools for processing Dutch natural language.

Table 1. Overview of tools suitable for processing Dutch natural language.
(int): tool provides the functionality only internally, i.e., the results cannot be retrieved;
+: the tool offers the functionality.

	Language independent		Language dependent				
	Sentence Splitter	Tokenizer	Morphological Analyzer	POS Tagger	Parser	Noun phrase finder	Concept mapper
Apache Lucene[1]		+	Dutch stemmer and analyzer				
TermTreffer[2]		+	Morphological Analyzer; Stopwords; Named entity recognizer; Negation finder	+		Multi-word recognizer	
Alpino[3]	(int)	+	(int)	(int)	+		

Second, we investigated the possibilities of creating a concept mapper (to map Dutch terms to concepts in SNOMED CT) based on the UMLS. To this end, we assessed the extent to which concepts in the CORE

[1] http://lucene.apache.org/
[2] http://www.inl.nl/tst-centrale/nl/over-de-tst-centrale/projecten/termtreffer
[3] http://www.let.rug.nl/~vannoord/alp/Alpino/

subset, which consists of 5965 SNOMED CT concepts useful for documenting reasons of encounter (RoE), have a Dutch term in one of the source vocabularies of the UMLS. A total of 4236 concepts have a direct translation to Dutch, which is 71% of the concepts in the CORE subset. Furthermore, we attempted to map a set of 3930 free-text reasons of encounter, using Dutch translations of SNOMED CT in the UMLS. Table 2 depicts the extent to which RoE's could be mapped in full, partially, or not at all, to a SNOMED CT concept.

Table 2. Mappings from RoEs to SNOMED CT concepts in numbers and percentages.

	# of matches to SNOMED CT concepts	**Percentage**
Full match	79	2.0%
Partial match	2927	74.5%
Non match	924	23.5%
Total	3930	-

One solution for such a concept mapper would be the complete translation of SNOMED CT, as has been undertaken for Danish, Spanish, and Swedish, but which involves significant effort and resources. However, the increasing interest to map coding systems used in the Netherlands to SNOMED CT provides an opportunity to collect Dutch entry terms for SNOMED CT concepts as a derivative of the mapping process.

A variety of Dutch coding systems is currently either being mapped or planned to be mapped. These systems include the following domains :

- Diagnoses, based on the diagnosis lists of two University Medical Centers
- Procedures, based on the list maintained by Stichting CBV (a national organization)
- Optometry, based on a SNOMED CT subset defined by optometrists.
- Pathology, based on the PALGA thesaurus (a national thesaurus based on SNOMED II)

The mapping-based approached for creating Dutch entry terms for SNOMED CT provides a number of advantages. First, those parts of SNOMED CT are first addressed which have the most value for Dutch users. Second, the terms provided are those with which the users are already familiar.

Once a significant part of the mappings has been performed, evaluation needs to be performed to ensure that the use of the Dutch terms for processing clinical narratives results in sufficiently high precision and recall, i.e., that the SNOMED CT concepts to which the narratives are matched are maximally complete and correct.