The e-Lex-Multi-word lexicon for Language- and Speech Technology

Richard Piepenbrock
7 April 2005

General information

e-Lex sets out to be a Dutch-language lexical database that is as widely applicable as possible within the fields of language and speech technology. It is both suited for fundamental and applied research.

Owing to the large number of lexical features and its theory-neutral design, the lexicon can be used for numerous language-related research disciplines and the development of language-aware software. These can be said to include automatic speech recognition and synthesis, spelling checking, part-of-speech tagging and grammatical parsing, morphological parsing, multi-word-unit detection, and information extraction and retrieval.

The TST-Multi-word lexicon, with the name `tstmlex`, only comprises multi-word expressions that include blanks. Continuous (single) word forms can be found in a separate TST-Lexicon. The file name of this lexicon is `tstlex`. In this (single word) lexicon, however, all parts of the multi-word expressions listed here have been included as separate entries, alongside those features relevant to each of the parts, such as pronunciation and part-of-speech. Unique reference numbers link the relevant single-word lexical items of the general TST-Lexicon to the multi-words they occur in of the Multi-word lexicon.

The TST-Multi-word lexicon is based on an inventory of all multi-word expressions that occur in a number of sources (CELEX (Centre for Lexical Information) [1], RBN (Referentiebestand Nederlands, Reference File Dutch) [2], Woordenlijst Nederlandse Taal (Groene Boekje, Dutch Spelling Guide, 1995), Corpus Uit den Boogaart [3] and the Van Dale Groot Woordenboek der Nederlandse Taal [4]), complemented with all multi-word expressions that were encountered in the Corpus Gesproken Nederlands (CGN, Spoken Dutch Corpus) [5].

The following multi-word expressions are distinguished:

- discontinuous:
  - verbs that can be separated into particles and verbal parts and can thus occur as discontinuous strings (eg *opnemen - ik neem meteen op*, *ademhalen - zij haalt diep adem*)
- continuous:
  - common (originally) foreign expressions (eg *et cetera*, *wishful thinking*)
  - native and non-native proper names and titles (eg *Berg En Dal*, *Avril Lavigne*, *De Morgen*, *De Pfaffs*)
  - multi-word contractions, i.e. from which a morphologically identifiable part shared by both parts has been omitted (eg *binnen- en buitenland*, *ouderrechten en -plichten*)

- fixed, invariant expressions, one of the parts of which cannot occur outside this expression (eg *op heterdaad* (red-handed), *'s anderendaags* (the other day), *ten enenmale* (once and for all))

Format and contents of the TST-Multi-word lexicon

The lexicon is available in two formats:

1. A standard text file (flat ASCII) with the name `tstmlex.txt`. The backslash ('\') is used as field separator. Letters with diacritics are represented in SGML format. This file can be opened by means of a simple text editor, or a database system such as Access, ORACLE or dBASE.
2. An XML file with the name `tstmlex.xml`. This file can be opened by any XML browser or editor, and then searched for certain values. The associated DTD (Document Type Definition) file `tstmlex.dtd`, which defines the structure of the data file, is also available.

The TST-Multi-word lexicon comprises 10 columns. Both lexicon files have been ordered according to columns *Orthography Multi-word* and then *Id-Number Multi-word Lemma*, *Part-of-speech Multi-word* and *Rank Number Part* (of the members of the multi-word expression).

Number of unique multi-word expressions	77,271
Number of unique multi-word lemmas	26,465
Number of entries for multi-words	180,655

Contents of the lexicon fields

1. TST_MLEXICON.**Orthography Multi-word**

   XML-label:
   <orth>
   Character set:
   ([0-9][A-Z][a-z][&'*-;])+
   Empty fields TXT-version:
   0

   Orthographic representation of the multi-word expression. If the expression can occur in more than one inflected form, often because these occur in the CGN Corpus, all inflected forms have been included with one and the same *Id-Number Multi-word Lemma*. Diacritics are represented in SGML format as follows:

   "&" + capital/small letter + diacritic + ";"

   To be precise:

"&" +	"a" +	"grave"	+ ";"

"c"	"acute"
"e"	"circ" (= circumflex)
"i"	"uml" (= umlaut/diaeresis)
"n"	"cedil" (= cedilla)
"o"	"tilde"
"u"	"ring"
"A"	
"C"	
"E"	
"I"	
"N"	
"O"	
"U"	

eg '&agrave; la carte' for 'à la carte'
and
'Gustaf &Aring;kermans' for 'Gustaf Åkermans'

The SGML symbol '&amp;' is used to represent the ampersand ('&'), as in 'College Van B&amp;W'.

2. TST_MLEXICON.**Rank Number Part**

XML-label:
<(part) n>
Character set:
[0-9]+
Empty fields TXT-version:
0

This number indicates the position of the word form in the sentence relative to the other parts of the multi-word expression.

3. TST_Lexicon.**Id-Number Word Form**

XML-label:
<wid>
Character set:
[0-9]+
Empty fields TXT-version:
0

Unique rank number (*Id* = 'identification') for each word form as part of the multi-word expression. This Id-Number refers to the field bearing the same name in the TST-(Single Word) Lexicon `tstlex`. In this lexicon, you will find the orthography of the word form, alongside other relevant information, such as the pronunciation of the word form, its part-of-speech, usage and its associated lemma. The orthography of the word form and its part-of-speech, even though these are, strictly speaking, redundant to this lexicon, have been included in the multi-word lexicon as well to facilitate selections on the data.

4. TST_MLEXICON.**Orthography Word Form**

XML-label:
<worth>
Character set:
([0-9][A-Z][a-z][&'-;])+
Empty fields TXT-version:
0

Orthographic representation of the word form, i.e. the individual parts of the multi-word expression. Diacritics are represented as described above for *Orthography Multi-word*.

5. TST_MLEXICON.**Part-of-speech Word Form**

XML-label:
<pos> (part of speech)
Character set:
([123][A-Z][a-z] | "(" | ")" | "," | "-")+
Empty fields TXT-version:
0

The part of speech of the word form, i.e. the individual parts of the multi-word expression. Values for the word classes conform to the notation used in the CGN (Spoken Dutch Corpus), as described in the document *Part of Speech Tagging en Lemmatisering van het Corpus Gesproken Nederlands* (Van Eynde 2004).

A shorter English version of this document is also available as the LREC-2000 paper *Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus* (Van Eynde et al. 2000).

```
"ADJ(" value ("," value)* ")" |
"BW("")" |
"LID(" value ("," value)* ") |
"N(" value ("," value)* ")" |
"SPEC(afgebr)" |
"SPEC(deeleigen)" |
"SPEC(meta)" |
"SPEC(onverst)" |
"SPEC(vreemd)" |
"TSW()" |
"TW(" value ("," value)* ")" |
"VG(" value ")" |
"VNW(" value ("," value)* ")" |
"VZ(" value ")" |
```

"WW(" value ("," value)* ")"

Description of the codes:

ADJ
adjectief (= adjective)
BW
bijwoord (= adverb)
LID
lidwoord (= article)
N
substantief (= noun)
SPEC(afgebr)
code that is used exclusively in the lexicon for parts of contracted multi-word expressions (eg 'in- en uitvoer')
SPEC(deeleigen)
code for part of a multi-word proper name
SPEC(meta)
code for a word mention
SPEC(onverst)
code for an incomprehensible utterance
SPEC(vreemd)
code for an utterance in a foreign language, or for an originally foreign part of a multi-word expression that in itself has not been assimilated
TSW
tussenwerpsel (= interjection)
TW
telwoord (= numeral)
VG
voegwoord (= conjunction)
VNW
voornaamwoord (= pronoun)
VZ
voorzetsel (= preposition)
WW
werkwoord (= verb)

6. TST_MLEXICON.**Part-of-speech Multi-word**

XML-label:
<mpos>
Character set:
([A-Z][a-z] | "(" | ")" | "," | "-")+
Empty fields TXT-version:
0

The part-of-speech of a multi-word expression where from a grammatical point of view the full expression can be regarded as one word. Values are the same as those found with the word form. In addition we find

COMB(eigen)
code for compound proper name or title to which no specific attributes like gender or number have been assigned
SPEC(samentr)
code for the complex (multi-word) contraction *da's* (= that's) which cannot be assigned to any of the common parts-of-speech

7. TST_LEXICON.**Id-Number Multi-word Lemma**

   XML-label:
   <lid>
   Character set:
   [0-9]+
   Empty fields TXT-version:
   0

   Rank number (*Id* = 'identification') which indicates which multi-word expressions belong to one and the same paradigm. It refers to the field of the same name in the TST-(Single Word) Lexicon `tstlex`, where you can find information pertinent to the lemma, such as the representation of the lemma, its morphology and semantics.

   The distinction is only relevant for possibly discontinuous verbs, which here haven been assigned ID numbers between 500,000 and 600,000 to indicate that they are multi-word lemmas rather than single-word lemmas. Lemma ID numbers higher than 600,000 denote multi-word units other than separable verbs, viz. proper names, titles, multi-word contractions, some opaque fixed expressions and loan words fully integrated into the Dutch language.

   Where orthographically identical (multi-word) lemmas occur with different ID numbers this implies that lemmas are involved with different morpho-syntactic (eg strong or weak declension) or phonetic (eg stress) characteristics, in combination with a difference in meaning. The difference in meaning is indicated in the field *Definition Lemma* in the TST-Lexicon `tstlex`.

8. TST_MLEXICON.**Multi-word Lemma**

   XML-label:
   <lem>
   Character set:
   ([0-9][A-Z][a-z][&'*-;])+
   Empty fields TXT-version:
   0

   The lemma of multi-word expressions, such as 'uitademen' for a multi-word instance like '(ik) adem uit'. With continuous multi-word expressions, viz. compound proper names and titles, multi-word contractions, some opaque fixed expressions and loan words fully integrated into the Dutch language, a dummy lemma form is postulated which is identical to the value in the field *Orthography Multi-word*.

9. TST_MLEXICON.**Optional Part**

XML-label:
<opt>
Character set:
("J" | "N")
Empty fields TXT-version:
0

If the word form in question is an optional part of a multi-word expression, then the value of this field is 'J' (*ja*, 'yes'). If the word form is an obligatory part of a multi-word expression, then the value of the field is 'N' ('no'). Thus 'ademt' as part of 'inademen' and 'uitademen' has the value 'J', while 'apen' as part of 'na-apen' receives the value 'N'.

10. TST_MLEXICON.**Continuous Multi-word**

    XML-label:
    <cont>
    Character set:
    ("J" | "N")
    Empty fields TXT-version:
    0

    If the multi-word expression cannot be interrupted (by constituents other than hesitations or interjections), as for example 'Tien Voor Taal' or 'per se', the multi-word expression as a whole is seen as continuous and given the value 'J', or else 'N', as in the case of discontinuous verbs.

---

[1] CELEX Dutch database, version 3.2 (1998). Centre for Lexical Information. Interfaculty Research Group Language and Speech, University of Nijmegen & Max Planck Institute for Psycholinguistics, Nijmegen.

[2] Referentiebestand Nederlands (1998). Research Group Lexicology, Free University of Amsterdam & Institute for Dutch Lexicology, Leiden & Department of Linguistics, Catholic University of Louvain & Department of Dutch, University of Utrecht.

[3] Boogaart, P.C. Uit den (1975). Woordfrequenties: in Geschreven en Gesproken Nederlands (Word frequencies in Written and Spoken Dutch). Oosthoek, Scheltema & Holkema, Utrecht. Electronic version available as part of the *Eindhoven Corpus*.

[4] Van Dale Groot Woordenboek der Nederlandse Taal (Van Dale Large (Unabridged) Dictionary of the Dutch Language). Thirteenth impression (1999). Van Dale Lexicografie, Utrecht/Antwerpen.

[5] Spoken Dutch Corpus (CGN), version 1.0 (2004). Dutch Language Union, The Hague.