

Elaborating the Parameterized ECM for Dutch

Nicole Grégoire

Uil-OTS, University of Utrecht

The parameterized Equivalence Class Method for Dutch is an approach developed to incorporate standard lexical representations for Dutch idioms into representations required by any specific NLP system with as minimal manual work as possible. The purpose of the paper is to give an overview of parameters applicable to Dutch, which are determined by examining a large set of data and two Dutch NLP systems. The effects of the introduced parameters are evaluated and the results presented.

Introduction

MultiWord Expressions (MWEs) form a serious problem for many areas of language technology. Their unpredictable meaning and restrictions on syntactic variability make them unsuitable for literal treatment. For successful handling of MWEs, both the grammar and the lexicon of an NLP system must be extended.

Aim

Creating a database of 5,000 expressions that meets the criterion of being highly theory- and implementation-independent, and which can be used in various NLP systems.

What are idioms?

In this research idioms are defined as MWEs headed by a verb (non-finite in the canonical form) with a fixed (or very limited) item selection and which meaning cannot be obtained compositionally from the meaning of its parts when used in isolation.

Some examples of Dutch idioms:

- (1) het licht zien
the light see
'see the light'
- (2) over lijken gaan
across dead-bodies go
lit. 'go across dead bodies'
id. 'show no mercy'
- (3) naast zijn schoenen lopen
next-to his shoes walk
lit. 'walk next to his shoes'
id. 'be full of conceit'

Problems with idioms in NLP

- Idioms need the same syntactic structure as their literal counterpart,
- BUT highly specific representations are undesirable, because:
 - it requires effort to make such representations for each NLP system again, and
 - reuse of significant effort is not possible.

No de facto standard for the lexical representation of MWEs currently exists. The parameterized ECM should offer a theory- and implementation-independent solution to the problem of lexical representation of idioms.

The Equivalence Class Method for Dutch

Instead of describing the structure of an idiom, the ECM requires that it is specified which idioms have the same structure.

Ingredients of an idiom description:

- An idiom pattern name: an identifier that uniquely identifies the structure of the idiom.
- A list of idiom components (Idiom Component List: ICL).
- An example sentence that contains the idiom.

Ingredients of an idiom pattern description:

- An idiom pattern.
- Comments, i.e. free text in which the uniqueness of the pattern is described.

ECM: conversion procedure

System independent idiom lexicon

A set of idiom pattern descriptions

Pattern	Comments
IDp1	Expressions headed by a verb taking a direct object NP that consists of a determiner and a singular noun.
IDp2	...

+

A set of idiom descriptions: idioms with the same pattern belong to the same equivalence class

Pattern name	ICL	Example
IDp1	de plaat poetsen id. 'to clear off'	Hij heeft de plaat gepoetst
IDp1	de boot missen id. 'to miss the boat'	Hij heeft de boot gemist
IDp1	de kar trekken id. 'to carry the load'	Hij heeft de kar getrokken

convert ↓↓ into

System specific idiom lexicon



A potential problem of the ECM as proposed is the risk that the number of equivalence classes will run into thousands of which the majority contains only a small number of idioms.

Parameterized ECM

The central idea behind the parameterized ECM is to group idioms that are for a large part identical and only differ e.g. in the number of the noun it requires: *de plaat-SG poetsen* vs. *de benen-PL nemen*

- parameter = ⟨parameter category (PC), parameter value (PV)⟩
- PC refers to the aspect we parameterize.
- PV refers to the value a PC takes.
- We distinguish 7 PCs and 23 PVs.
- E.g.: ⟨nnum,SG⟩, ⟨nnum,PL⟩, ⟨afm,SUP⟩

The ICLs of some idioms extended with parameters.

Expression	ICL
<i>de plaat poetsen</i> id. 'to clear off'	de plaat[DE][SG][POS] poetsen
<i>de benen nemen</i> id. 'to escape'	de been[HET][PL][POS] nemen
<i>de pijp uitgaan</i> id. 'kick the bucket'	uit[POST] de pijp[DE][SG][POS] gaan
<i>op de fles gaan</i> id. 'to go broke'	op[PREP] de fles[DE][SG][POS] gaan
<i>zijn draai vinden</i> id. 'feel comfortable'	zijn[SB] draai[DE][SG][POS] vinden

Reducing the number of idiom patterns means reducing the number of equivalence classes. As a result, the number of idioms that have to be dealt with manually minimizes, whereas the number of idioms that can be incorporated into an NLP system in a fully automatic manner increases.

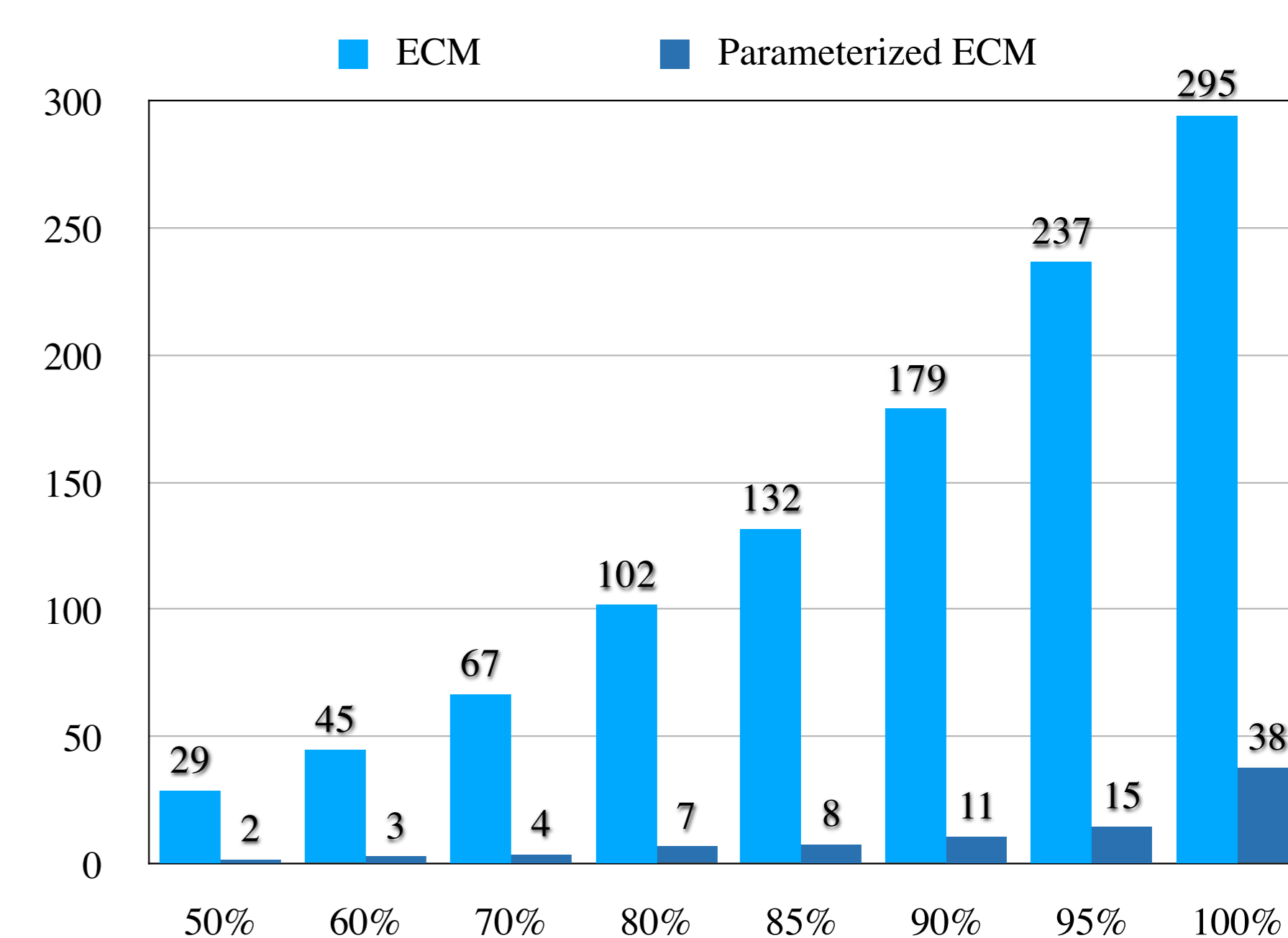
Evaluation

The successfulness of introducing parameters depends on:

1. how many different equivalence classes are distinguished (the less the better), and
2. how many instances each equivalence class contains (the more the better).

- Source: Electronic version of the Van Dale Idiom dictionary for Dutch.
- Data: 1,167 three- and four-word idioms.
- The Alpino parser assigned a part-of-speech tag to the components of each expression.
- The patterns of the idioms, i.e. the ECs, were determined on the basis of these part-of-speech tags.
- The Alpino parser semi-automatically assigned the parameters.

Coverage of Equivalence Classes (ECs). The x-axis shows the coverage of the # of idioms (100% = 1,167 idioms). The y-axis shows the number of ECs needed to cover the number of idioms.



Conclusion

- The introduction of parameters decreases the number of equivalence classes needed with almost 90% with respect to the numbers of equivalence classes needed in the original ECM.
- A total of 15 parameterized equivalence classes are needed to cover 95% (or 1,109) of the three- and four-word idioms.

The use of parameters reduces the number of equivalence classes and increases the number of idioms in each class, supporting the task of converting the standard format into the structure required in the target NLP system.

Future work

- Representing the patterns using a combination of part-of-speech and dependency labels.
- Extend the method to other types of MWEs.
- Test the method in two Dutch parsers: Alpino and Rosetta MT.

Acknowledgements

This research is carried out as part of the IRME project, financed by STEVIN (<http://taaluniversum.org/stevin>). The data were kindly provided by Van Dale Lexicografie.