Predicting the unpredictable
Developing a lexicon model for Norwegian MWEs

Gyri Smørdal Losnegaard
Department of Linguistic, Literary and Aesthetic Studies, University of Bergen, Norway

Multiword expressions (MWEs): linguistic anomalies

MWEs are word combinations (lexical units) characterized by having some degree of **unpredictable** or **idiosyncratic** linguistic properties. They are problematic because they break with linguistic convention—deviating from what is productive and derivable in a language. MWEs are challenging for different disciplines, and for tasks such as **foreign language learning**, **(machine) translation**, **parsing**, **spelling and grammar checking** and **information retrieval**. Consider the following example where MWEs cannot be translated word for word:

Til syvende og sist måtte et eller annet en gang ha blitt til av null og niks. (*Sofies verden* "Sophie's World" by Jostein Gaarder)
[Ultimately] At some point, something must have come from nothing. (English translation of the novel)

Main objectives

The main objective of this project is to create a broad-scope, multi-purpose and reusable **lexical resource** of Norwegian multiword expressions (MWEs). This involves developing a linguistically informed and largely language-independent **lexicon model** and **methodology** for identifying, classifying and representing linguistic properties of MWEs.

Subtasks

Identification is the task of distinguishing MWEs from free combinations. In this project, this involves operationalizing the notions of idiosyncrasy and productivity. The aim is to develop a method that can serve as a practical tool in dictionary and grammar development. **Classification** is the task of distinguishing types of MWEs. This involves applying a range of criteria with the aim of arriving at a holistic, broad, extendable and linguistically motivated classification model. **Delimitation** of the MWE lemma concerns distinguishing between MWE variants and new MWE lemmas. Finally, **MWE description** involves deciding *what* are the necessary and sufficient properties to be represented for each MWE lemma in the lexical resource, and *how* to represent this information.

MWE identification

Operationalizing linguistic idiosyncrasy:

- institutionalization (usage, statistics)
- lexical and grammatical fixedness (lexicon and morphosyntax)
- extragrammaticality (lexicon and morphosyntax)
- figurative meanings (compositional semantics)
- non-productive uses of single words (lexical semantics)
- specialized meanings and discourse functions (semantics, pragmatics)

Operationalizing productivity: when and how does a lexical item deviate from its "usual" lexical and syntactic behaviour?

MWE lexicon model

Model development is based on a linguistically informed and largely language-independent methodology for identifying, classifying and representing linguistic properties of MWEs. It builds on existing knowledge about MWE properties and principles of word classification.
Three complementary perspectives guide the lexicon model development:

- **automatic analysis** *(the NLP perspective)*
- **lexicon development** *(the lexicographic perspective)*
- **foreign language acquisition and use** *(the language learning perspective)*

Tools and materials

The data are approx. 2000 MWE candidates compiled during the construction of **NorGramBank**, a large LFG treebank hosted by the **INESS** infrastructure, which is part of the **CLARINO Bergen Centre**. Supplementary data is retrieved from the treebank using INESS search, a querying system for treebanks in a variety of formats.

Frameworks

The framework **Lexical Functional Grammar (LFG)** is used for linguistic analysis and **Lexical Markup Framework (LMF)** for MWE description [1, 3, 8].

MWEs as linguistic signs

MWEs are complex form-function units, cf. **constructions** [2, 4] or **linguistic signs** [6]. In this project, idiosyncrasy is considered their main distinctive feature. Type(s) of idiosyncrasy distinguish MWEs from free combinations and types of MWEs from each other.

MWE classification

Many existing classification models are not easily applicable to large and diverse data sets because they:

- are too *narrow* or *purpose-specific*,
- rely on only a *few criteria*, or
- are *difficult to operationalize*.

This leads to the following requirements for the current model:

- *broad-scope*
- *extendable*
- *holistic and linguistically founded*

The model draws on principles of word classification, with an emphasis on **form**, **function** and **idiosyncrasy**.

The MWE lemma

MWEs are lexical units that may deserve status as lemmas in lexical resources (LRs) [5]. Since MWEs may vary at the semantic and syntactic level [7, 9], the representation of MWEs in LRs requires principled **delimitation** of the MWE lemma.
An important task is thus to determine its possible **variation scope**: at which point does deviation from the canonical form violate the mutual dependency between MWE form and meaning, to the extent that the MWE meaning fails to be evoked?
What is the **necessary and sufficient** information to be represented for a MWE lemma if the LR is to be useful for NLP, lexicographic, and language learning purposes?

References

[1] Dalrymple, Mary (2001). *Lexical Functional Grammar*.
[2] Fillmore, Charles, Paul Kay and Catherine O'Connor (1988). *Regularity and Idiomaticity in Grammatical Constructions: The Case of let alone*.
[3] Francopoulo, Gil (ed.). *LMF. Lexical Markup Framework*.
[4] Goldberg, Adele (2006). *Constructions at Work: The Nature of Generalization in Language*.
[5] Jónsson, Jón Hilmar (2009). *Lemmatisation of Multiword Lexical Units: Motivation and Benefits.*
[6] Mel'čuk, Igor (1998). *Collocations and lexical functions*.
[7] Moon, Rosamund (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*.
[8] Odijk, Jan (2013). *Identification and Lexical Representation of Multiword Expressions*.
[9] Sköldberg, Emma (2004). *Korten på bordet. Innehålls- och uttrycksmässig variation hos svenska idiom.*