# Representing Multiword Expressions in Lexical and Terminological Resources

## An Analysis for Natural Language Processing Purposes

Carla Parra Escartín, Gyri Smørdal Losnegaard

University of Bergen
(Norway)

PARSEME General Meeting
IPIPAN, Warsaw, Poland
September 17, 2013

# Introduction

- Lexical Language Resources and Tools (LRTs) = key element of advanced NLP

# Introduction

- Lexical Language Resources and Tools (LRTs) = key element of advanced NLP
- Computational lexicography has fostered the creation of reusable and interoperable lexical resources

# Introduction

- Lexical Language Resources and Tools (LRTs) = key element of advanced NLP
- Computational lexicography has fostered the creation of reusable and interoperable lexical resources
  - Electronic and machine readable dictionaries

# Introduction

- Lexical Language Resources and Tools (LRTs) = key element of advanced NLP
- Computational lexicography has fostered the creation of reusable and interoperable lexical resources
  - Electronic and machine readable dictionaries
  - Lexical databases

# Introduction

- Lexical Language Resources and Tools (LRTs) = key element of advanced NLP
- Computational lexicography has fostered the creation of reusable and interoperable lexical resources
  - Electronic and machine readable dictionaries
  - Lexical databases
  - Terminological databases, etc.

# Introduction

- Lexical Language Resources and Tools (LRTs) = key element of advanced NLP
- Computational lexicography has fostered the creation of reusable and interoperable lexical resources
  - Electronic and machine readable dictionaries
  - Lexical databases
  - Terminological databases, etc.
- The lexis of a language is more than just single words!

# Introduction

- Lexical Language Resources and Tools (LRTs) = key element of advanced NLP
- Computational lexicography has fostered the creation of reusable and interoperable lexical resources
  - Electronic and machine readable dictionaries
  - Lexical databases
  - Terminological databases, etc.
- The lexis of a language is more than just single words!
  - *Fit as a fiddle*

# Introduction

- Lexical Language Resources and Tools (LRTs) = key element of advanced NLP
- Computational lexicography has fostered the creation of reusable and interoperable lexical resources
  - Electronic and machine readable dictionaries
  - Lexical databases
  - Terminological databases, etc.
- The lexis of a language is more than just single words!
  - *Fit as a fiddle*
  - *Give in*

# Introduction

- Lexical Language Resources and Tools (LRTs) = key element of advanced NLP
- Computational lexicography has fostered the creation of reusable and interoperable lexical resources
  - Electronic and machine readable dictionaries
  - Lexical databases
  - Terminological databases, etc.
- The lexis of a language is more than just single words!
  - *Fit as a fiddle*
  - *Give in*
  - *Pose a problem*

# Introduction

- Lexical Language Resources and Tools (LRTs) = key element of advanced NLP
- Computational lexicography has fostered the creation of reusable and interoperable lexical resources
  - Electronic and machine readable dictionaries
  - Lexical databases
  - Terminological databases, etc.
- The lexis of a language is more than just single words!
  - *Fit as a fiddle*
  - *Give in*
  - *Pose a problem*
  - *As a matter of fact*

# MWEs are a *pain in the neck* and not only for NLP

- MWEs exceed word boundaries

# MWEs are a *pain in the neck* and not only for NLP

- MWEs exceed word boundaries
- MWEs have unpredictable properties

# MWEs are a *pain in the neck* and not only for NLP

- MWEs exceed word boundaries
- MWEs have unpredictable properties
- MWEs are semantically non-transparent and non-compositional

# MWEs are a *pain in the neck* and not only for NLP

- MWEs exceed word boundaries
- MWEs have unpredictable properties
- MWEs are semantically non-transparent and non-compositional
- There is no agreed upon typology of MWEs

# Our research question

**What information should be recorded when including MWEs in lexical and terminological resources?**

# Some preliminary remarks:

- The intended usage of a resource may condition its layout and the information recorded in it.

# Some preliminary remarks:

- The intended usage of a resource may condition its layout and the information recorded in it.

# Some preliminary remarks:

- The intended usage of a resource may condition its layout and the information recorded in it.
  In the case of MWEs:

# Some preliminary remarks:

- The intended usage of a resource may condition its layout and the information recorded in it.
  In the case of MWEs:
  - Purpose of the resource?

# Some preliminary remarks:

- The intended usage of a resource may condition its layout and the information recorded in it.
  In the case of MWEs:
  - Purpose of the resource?
  - Type(s) of MWE?

# Some preliminary remarks:

- The intended usage of a resource may condition its layout and the information recorded in it.
  In the case of MWEs:
  - Purpose of the resource?
  - Type(s) of MWE?
  - Intrinsic features?

# Some preliminary remarks:

- The intended usage of a resource may condition its layout and the information recorded in it.
  In the case of MWEs:
  - Purpose of the resource?
  - Type(s) of MWE?
  - Intrinsic features?
- A hybrid user scenario should be envisaged, i.e. the resource may be used both for human purposes and NLP

# N-grams: project overview (I/III)

**Summary** Empirical study applying statistical association measures (AMs) to extract bigram and trigram term candidates from the Norwegian Newspaper Corpus (NNC).

**Aim** Determine which AMs are better for different MWE types:

- Identify efficient tools to detect different MWEs automatically
- Identify recurrent collocational patterns for term extraction

**AMs used** 9 AMs for bigrams, 4 AMs for trigrams

# N-grams: project overview (II/III)

**Categories** anglicism MWE, foreign MWE, grammatical MWE, idiomatic phrase, term candidate, concept structure appositional phrase

**Representation** Project specific representation needs include:

- Standardised way of expressing statistical information about the rank of an item
- Part-of-Speech (PoS) tagging
- Lemma information
- Meaning of foreign expressions and the language they are written in
- Inflectional features/paradigms

# N-grams: project overview (III/III)

Examples of high-ranked collocations in the study

| **Multiword unit** | **English translation** | **Suggested classification** |
|---|---|---|
| consumer confidence | - | anglicism MWE |
| annus horribilis | (Lat.) horrible year | foreign MWE |
| etter hvert | gradually | grammatical MWE |
| grøss og gru | shiver and horror | idiomatic phrase |
| alternative energikilder | alternative energy sources | term candidate |

# Specialised collocations: project overview (I/II)

**Summary** Study of collocations in Free Trade Agreements (i.e. specialised legal and economics texts)

**Aim** Formal representation of specialised collocations in English and Spanish

**Representation** Project specific representation needs include:

- Node of the collocation
- Specialised subject field
- All collocates that the node may take in the respective subject fields
- Morphosyntactic information
- Dialectal aspects

# Specialised collocations: project overview (II/II)

Specialised collocations in English and their Spanish equivalents

| **English** | **Spanish** |
|---|---|
| accord favorable treatment | otorgar trato favorable |
| labor or environmental law enforcement | cumplimiento de la legislación laboral o ambiental |
| prescribe a conformity assessment procedure | exigir un procedimiento de evaluación de conformidad |
| cross-border financial service suppliers | proveedores transfronterizos de servicios financieros |
| prepare \| adopt \| apply a technical specification | preparar \| adoptar \| aplicar una especificación técnica |

[Source: FTA Corpus]

# Translational correspondences: project overview (I/II)

**Summary** Study of German nominal compounds and their phraseological correspondences in Spanish.

**Aim** Improve 1:n word alignments within Germanic and Romance languages and automatic extraction of compound dictionaries.

**Representation** Project specific representation needs include:

- Translational correspondences 1:n
- Splitting of the compound in German and tagging of the head
- Compound internal structure
- Elements which may be inflected in Spanish
- Fixed and semi-fixed elements in Spanish
- Modifier information

# Translational correspondences: project overview (II/II)

Examples of German compounds and their correspondences into Spanish

| **German Compound** | **Compound constituents** | **Spanish Correspondence** |
|---|---|---|
| Wohnungsförderungsverordnung | Wohnung·s·förderung·s·verordnung | Ley de promoción de vivendas |
| Warmwasserbereitung | Warm·wasser·bereitung | preparación de agua caliente |
| Wärmepumpeanlagenförderung | Wärme·pumpe·anlagen·förderung | promoción de instalaciones de bombas de calor |

English = *Housing Promotion Act / Water heating / Promotion of heat pumping systems*

[Source: TRIS Corpus]

# Norwegian MWEs: project overview (I/III)

**Summary** Extensive study of Norwegian MWEs.

**Aim** Build the first extensive inventory of MWEs for Norwegian.

- Basis for a typology of Norwegian MWEs
- Basis for the integration of different types of MWE into NorGram, a computational grammar for Norwegian.

**Representation (I)** Project specific representation needs include:

- Lexical information: PoS, definition, canonical form
- Source information (corpus, dictionary, etc.)
- Source text information (type, genre, publication date, author, etc.)

# Norwegian MWEs: project overview (II/III)

**Representation (II)** Further requirements:

- MWE extraction method
- MWE surface form and context
- MWE frequency
- Linguistic levels of anomalous behaviour
- Degree of semantic transparency
  (from transparent to opaque)
- Syntactic flexibility
  (fixed, semi-fixed, syntactically flexible)
- Internal structure
- Morphosyntactic restrictions

# Norwegian MWEs: project overview (III/III)

MWEs in *Sofies verden* (Sophie's World)

| **Norwegian** | **Literal translation** | **Idiomatic translation** |
|---|---|---|
| snakke om | talk about | talk about |
| stå igjen | stand again | be left, remain |
| gjøre lekser | do homework | do (one's) homework |
| skille lag | part team | split, part |
| komme rekende på en fjøl | come drifting on a board | come from nowhere (with origin unknown) |
| sikker på | sure on | sure that, sure of/about |
| et eller annet | one or other | something |

## What information should a standard allow us to represent?

- Semantic and morphosyntactic properties of the overall expression and of the component words
- Internal structure of the MWE and dependencies
- Syntactic variation
- Regional varieties
- Translational correspondences

## What information should a standard allow us to represent?

- Semantic and morphosyntactic properties of the overall expression and of the component words
- Internal structure of the MWE and dependencies
- Syntactic variation
- Regional varieties
- Translational correspondences

## Other considerations:

Input / Output formats of the resources and tools that will exploit the resource being created!

## Projects and initiatives aiming at unifying the coding of computational lexicons and terminologies:

- GENELEX
- MULTEXT
- EAGLES
- SIMPLE
- ISLE

## Projects and initiatives aiming at unifying the coding of computational lexicons and terminologies:

- GENELEX
- MULTEXT
- EAGLES
- SIMPLE
- ISLE

## The standards being fostered are:

- TermBase eXchange format (TBX)
- Lexical Markup Framework (LMF)

* We will also have a brief look to the Text Encoding Initiative (TEI)

# TBX

- Its DTD is extremely flexible
- The attribute names and values can be customised by the user: interoperability concerns!
- MWEs can only be registered as strings: not possible to process non-fixed MWEs successfully
  - *it's raining cats and dogs*
  - *it's* **certainly** *raining cats and dogs*
  - *it* **is/was/will be/has been** *raining cats and dogs*
- Not adequate for monolingual representation
- Allows for the specification of language varieties

# Requirements for representing MWEs in TBX:

- Attribute and values should be restricted and agreed upon
- Granularity up to token level should be integrated
- It should be possible to encode paradigm and inflection information
- Features required for NLP applications should be studied and integrated
- Monolingual lexicons should be enabled

# LMF

- Developed combining the best designs and methods from many NLP lexicons
- Not intended for human users
- Extension for bilingual and multilingual dictionaries
- Designed to express equivalence relations in MT
- Module for the representation of MWEs:
  The NLP MWE Pattern allows for the representation of the internal structure of lexical units which are
  - Fixed
  - Semi-fixed
  - Flexible

# TEI

- Specific module for encoding dictionaries
- Extensive and detailed documentation
- Encoding of compounds is enabled and exemplified
- Encoding of other NLP relevant properties is possible:
  - Part of Speech
  - Geographical area
  - Etymological information
  - Links and cross-references to other entries in the same resource
- Drawbacks: very flexible and not fostered

# Overview of the three standards considered

| Standard | Monolingual | Bilingual | Encoding of morphosyntactic features | |
|---|---|---|---|---|
| | | | MWE level | Token level |
| TBX | No | Yes | Yes | No |
| LMF | Yes | Yes | Yes | Yes |
| TEI | Yes | Yes | Yes | Yes |

# Common requirements

- Part-of-Speech (PoS) (1,3,4)
- Lemma information for component words (1)
- Lemma (base form/canonical form) (4)
- Inflectional features/paradigms for component words (1)
- Morphosyntactic information for component words (2)
- Morphosyntactic restrictions (4)
- Internal structure of the MWE and dependencies (4)
- Syntactic variation (4)
- Meaning of foreign expressions, language (1)
- Meaning/definition (3,4)

# Project specific requirements (I/II)

- Splitting of the compound in German and tagging of the head (3)
- Compound internal structure (3)
- MWE elements which may be inflected in Spanish (3)
- Fixed and semi-fixed elements in Spanish (3)
- Modifier information (3)
- Statistical information about the rank of an item (1)
- MWE frequency (4)
- Collocation node (head word) (2)
- Specialised subject field (2)
- All collocates that the node may take in the respective subject fields (2)

# Project specific requirements (II/II)

- Dialectal aspects (2)
- Regional varieties (2,3,4)
- Translational correspondences (2)
- Translational correspondences 1:n (3)
- MWE type (1,4)
- Levels of linguistic or statistic deviance (semantic, syntactic, pragmatic...) (4)
- Semantic transparency (degree) (4)
- Syntactic flexibility (degree) (4)
- MWE extraction method (4)
- MWE surface form and context (4)
- Semantic and morphosyntactic properties of the overall expression and of the component words (4)

# META-SHARE

The META-SHARE schema (`http://www.meta-net.eu/meta-share/metadata-schema`)

- Metadata schema for LT resources and tools
- In order to accommodate flexibility, the elements in the META-SHARE schema belong to two basic levels of description:
  - an initial level providing the basic elements for the description of a resource (minimal schema)
  - a second level with a higher degree of granularity (maximal schema), providing more detailed information on each resource

# META-SHARE and CMDI (I/II)

The CMDI schema (`http://www.clarin.eu/node/3219`)

- XML-based schemas
- MS has recently been integrated as a CMDI module
- The CMDI metadata framework:
  - higher abstraction level than META-SHARE
  - flexible, allows for definition of your own metadata components
  - extend/adapt metadata model according to need

# META-SHARE and CMDI (II/II)

- META-SHARE and CMDI are models for the representation of metadata for language resources, and not for the encoding of such resources, but:
- A good starting point for a representation model for lexical resources (especially CMDI):
  - modularised
  - flexible: allows for customisation and definition of new modules

# Our suggestion:

- Modular representation schema
- Designed after the representation model envisaged by META-SHARE /CMDI

# Encoding *profiles* or *modules*

Three levels of detail:

- One mandatory profile (minimal schema, general)
- Two optional but recommended profiles (extended schema, general)
- An extendable set of optional type and purpose dependent profiles (extended schema, specialised)

# Module 1: minimal schema (mandatory, type level)

a) PoS
b) PoS standard
c) Meaning
d) The number of component words

# Module 2: extended schema (optional, type level)

a) Canonical (base) form
b) Level(s) of idiosyncracy
c) Translational correspondences
d) Language variety

# Module 3: extended schema (optional, token level)

a) Part of Speech (PoS)

b) Lemma

c) Grammatical features

# Modules 4-*n* (optional)

- Classification
- Morphosyntactic profile
- Metadata profile
- Organisational data
- Semantic profile
- Terminology
- Multilinguality
- Named Entity
- ...

# Summing up:

- Modules 1 and 2 represent general properties relevant for the description of the overall expression
  - Module 1: basic information
  - Module 2: extension of module 1, targets more advanced users and usages
- Module 3: represents general information about the component words
- Modules 4-: *specialised* profiles

# Discussion

A modular and flexible schema for LRTs such as the one discussed here could ensure the scalability and interoperability of LRT if:

- Feature names
- Values
- Formats

were agreed, standardised and correspondences between different standards were provided along to allow merging of resources.
* Eg. The N-grams project could be used as a starting point for the Norwegian MWEs project

# Future work

- Assess the appropriateness of the different standards for encoding lexical and terminological resources using data from our projects
- Determine to which extent standards allow for the encoding of the features proposed in the modular structure
- Test whether merging of resources including different kinds of information is actually possible
- Mapping of the different encoding formats to enable merging, exchanging and enlarging resources

# Conclusion

We have:

- Identified requirements for the representation of MWEs based on four NLP oriented but otherwise different projects
- Supported the assumption that MWEs form a heterogenous group of linguistic units, whose representation needs vary with:
  - the type of MWE
  - the nature of the project compiling the MWEs
  - the intended application or use of the lexical resource
- Proposed a modularized and flexible model for representation of MWEs in lexical and terminological resources, based on existing schemas for metadata description

# Aknowledgements


- The EU under FP7, Marie Curie Actions, SP3 People ITN, grant agreement 238405 (project CLARA)
- The University of Bergen
- The Norwegian School of Economics