Multiword Expressions and LMF

Jan Odijk
PARSEME Workshop
Iaşi, 21-22 Sep 2015
Overview

- MWEs
- Lexical Representation of MWEs
- DuELME
- DuELME and LMF
- Extensions
- Summary
Overview

MWEs

• Lexical Representation of MWEs
• DuELME
• DuELME and LMF
• Extensions
• Summary
What is an MWE?

• MWE = Multiword Expression
• Focus is on MWEs in an NLP context
What is an MWE?

• sequence of words

• that has linguistic (lexical, orthographic, phonological, morphological, syntactic, semantic, pragmatic) or translational properties

• not predictable from the individual component words and the normal rules for combining them
What is an MWE?

- sequence of
  - Not necessarily contiguous in a concrete utterance
    - ...omdat hij de plaat wilde poetsen
    - …because he the plate wanted polish
    - ‘…because he bolted’
  - Not necessarily always in the same order in each utterance
    - Hij poetsde gisteren de plaat
    - He polished yesterday the plate
    - ‘he bolted yesterday’
What is an MWE?

- **words**
  - Ambiguity between type and token (intentional)
  - Inflected word form v. lemma (both are needed)
  - Ambiguity between
    - Character sequences separated from other character sequences by spaces and other separators (Narrow interpretation)
      - *Bibliotheekzaal* v. *library hall*
      - compounds in Dutch, German, Norwegian, Swedish are NOT included
      - Compounds in English are included (parts separated by space)
    - Abstract lexical units of the grammar (Broad interpretation)
      - Dutch, German compounds ARE included if they meet the other criteria
What is an MWE?

• that has linguistic (lexical, orthographic, phonological, morphological, syntactic, semantic, pragmatic) or translational properties not predictable from the individual components and the normal rules for combining them
What is an MWE?

• the normal rules for combining them
  – Assumptions about this must be made explicit
    • In some cases they are not known
  – For each concrete NLP-system: the rules of that NLP-system
What is an MWE?

• Whether a word sequence is an MWE is an empirical hypothesis (or, in NLP, a proposed engineering solution)

• Intuitions about the status of expressions as MWEs have limited validity

• MWE-status must be argued for (or against)
  – Using the definition as a guide
Types of MWEs (I)

• Fixed
• Semi-flexible
• Flexible
Fixed MWEs

- Fixed MWEs
  - Words of the MWE in a fixed order
  - No variation in lexical item choice
  - Always contiguous (no other elements in between)
  - No inflectional processes except at the edges
Fixed MWEs

• Fixed MWEs
  – *ad hoc, stante pede, ter plaatse*
  – *Hong Kong, Kuala Lumpur, New York, San Francisco*
  – *credit card, travel agency, real estate agency*

• NOT
  – *in plaats van* (cf. *in plaats daarvan*) (‘instead of’)
  – *carta telefonica* (cf. *carte telefoniche*)
  – *de plaat poetsen* (‘polish the plate’, ‘bolt’)

13
Semi-Flexible MWEs

- Semi-Flexible MWEs
  - MWEs with fixed order of elements
  - That are impenetrable for other words
  - Parts can be inflected
Semi-Flexible MWEs

• Examples:
  – Chambre des représentants
    • House of representatives
  – Patatas fritas
    • French fries
  – Mise au point automatique
    • Autofocus
  – Calculateur analogique
    • Analogue computer
Semi-Flexible MWEs

- Examples:
  - Cité plus haut
    - Above-stated
  - Résistant aux acides
    - Acid-proof
  - Malade en altitude
    - Airsick
Flexible MWEs

- Flexible MWEs
  - Allow or require inflection in multiple parts, and
  - Allow permutations of subphrases, or
  - Allow intrusion by other phrases, or
  - Have controlled variation (bound pronouns)
Flexible MWEs

– *de plaat poetsen* (‘bolt’)
  • Hij heeft gisteren *de plaat ge poetst*
  • ...omdat hij *de plaat wilde poetsen*
  • Hij *poetste* gisteren *de plaat*

– But of course not just anything:
  • *Hij gepoetst plaat de heeft*
  • *..omdat wilde poetsen hij de plaat*

– to lose one’s temper
  • *He lost* his temper
  • *She lost* her temper
Types of MWEs (II)

- Idioms
- Semi-idioms
- Support-verb constructions
Types of MWEs(II)

• Idioms
  – Meaning not predictable from the components
  – The components have no or an unpredictable meaning
  – Fixed (or very limited) lexical item selection
Types of MWEs (II)

• Idioms
  – Non-transparant
    • *de plaat poetsen, kick the bucket, casser sa pipe*
    • Many restrictions on syntactic behavior (see handout example (4))
Types of MWEs(II)

• Idioms
  – Semi-transparant
    • *een bok schieten*
      – Bok (male goat) = blunder (but only with *schieten*)
      – Schieten (shoot) = make (but only with *bok*)
    • *dat varkentje wassen*
      – Varkentje (little pig) = problem (only with *wassen*)
      – Wassen (wash) = address, take care of (only with *varkentje*)

Little restrictions on syntactic behaviour. See handout example (5)
Types of MWEs(II)

• Semi-idioms (collocations)
  – One element occurs in its normal meaning
  – The lexical selection of the other element is fixed or very limited
  – The other element has a special meaning
  – Very little restrictions on syntactic behaviour. See handout example (6)
Types of MWEs(II)

• Semi-idioms (collocations)
  – Examples
    • Zware / * sterke tabak (heavy / *strong tobacco) `strong tobacco’
    • Scherpe kritiek (sharp criticism) `severe criticism’
    • Heavy / *strong smoker
Types of MWEs (II)

• Support verb constructions
  – Type I
    • Direct object + verb
    • Verb idiosyncratically determined by the direct object head noun
    • Arguments of the noun often realized outside the NP in the VP. See handout (8), (9)
Types of MWEs(II)

• Support verb constructions
  – Type I Examples
    • Een poging wagen ‘dare an attempt’
    • Een lezing houden / geven ‘hold / give a lecture’
    • With hebben ‘have’: see handout (7)
    • To pay attention to (aandacht schenken aan)
    • To take advantage of
Types of MWEs (II)

– Type II
  • Predicative complement (AP, PP)
    – often itself idiomatic
    – expressing a state or property
  • Combination with intransitive or transitive verb is idiosyncratic
## Types of MWEs (II)

<table>
<thead>
<tr>
<th>pred</th>
<th>Literal</th>
<th>intransitive</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>In de war</td>
<td>In the tangle</td>
<td>Zijn / raken / * gaan / *komen / *zitten</td>
<td>Confused (of humans)</td>
</tr>
<tr>
<td>In de war</td>
<td>In the tangle</td>
<td>*zijn / * raken / *gaan / komen / zitten</td>
<td>Entangled, mixed-up</td>
</tr>
<tr>
<td>In zijn nopjes DIM</td>
<td>In his studs-</td>
<td>Zijn / raken / *gaan / *komen / *zitten</td>
<td>Delighted,</td>
</tr>
<tr>
<td></td>
<td>DIM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>De pijp uit</td>
<td>The pipe out</td>
<td>Zijn / *raken / gaan / *komen / * zitten</td>
<td>dead</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Be / get / go / come / sit</td>
<td></td>
</tr>
</tbody>
</table>
Overview

• MWEs

➢ Lexical Representation of MWEs
  • DuELME
  • DuELME and LMF
  • Extensions
  • Summary
Lexical representation

• Focus on flexible MWEs
• Lexical representation for (grammar-based) NLP systems;
• NLP:
  – A sequence of words that is an MWE must be parsed / generated
  – A sequence of words that is an MWE must be recognized as an MWE
  And mapped to the appropriate semantics / translation
Lexical representation

• Flexibility
  – Can be accounted for by assuming a syntactic structure for an MWE
  – Is usually identical to the syntactic structure of the literal expression
  – → no problem to parse or generate sequences of strings that are MWEs.
  – Syntactic structure can NOT be determined automatically by an NLP system (ambiguities)
Lexical representation

• Flexibility (cont.)
  – Restrictions on flexibility must follow from general principles or additional MWE-specific properties
  – But: the syntactic structure is of course highly framework/theory/implementation-dependent
Lexical representation

- Examples of syntactic structures in
  - Lexical Functional Grammar (LFG): (1)
  - Tree Adjoining Grammar (TAG): (2)
  - M-Grammar: (3)
Lexical representation

• Syntactic structure contains references to lexical items from the lexicon used in the NLP-system
  – Otherwise it cannot be parsed / generated correctly
  – And the lexical properties must be correct!
    • **Inflection**
    • **Syntactic and semantic selection**
  – Extremely framework / grammar / implementation-dependent!
Lexical representation

• Summary: MWE lexical representation
  – Syntactic structure compatible with NLP-system
  – Correct references to lexical items in the NLP-system’s lexicon corresponding to the MWE components
  – Maximally framework / theory / implementation- independent
Overview

• MWEs
• Lexical Representation of MWEs
  ➢ DuELME
• DuELME and LMF
• Extensions
• Summary
DuELME

• Dutch Electronic Lexicon of Multiword Expressions
• App. 5000 entries
• MWEs of different types:
  – Mostly flexible idioms
  – Collocations (semi-idioms)
  – Mostly verbal
• Focus on syntax
DuELME

• Maximally theory-neutral:
  – (parameterized) Equivalence Class Method (ECM):
    • Method to lexically represent MWEs
    • Procedure to incorporate MWEs thus represented into a concrete NLP system

• See
  – Odijk 2004a, 2004b, 2013a, 2013b
  – Grégoire 2010
DuELME Lexical Representation

• Lexical Entries
  – MWEs with the same syntactic structure
    • by means of an MWE pattern id
  – Components: sequence of their lemmas
    • Any order but the same order within one pattern
  – Example sentence
    • Identical syntactic structure for each example in one equivalence class
DuELME Lexical Representation

- MWE Pattern descriptions
  - Mwe pattern id
  - Description (free text)
DuELME Lexical Representation

• DuELME is a *proto*-lexicon
  – Lexical resource from which a lexicon can be derived automatically or semi-automatically
  – By a well-defined procedure

• [Link to DuELME description](#)
• [Search GUI, User Documentation](#)
• [Metadata](#)
• [Product and license](#)
Incorporation Procedure

• Incorporation in some NLP system
• Assumes the NLP system contains a parser
• For each MWE pattern P do
  – Bootstrap part
    • Contains some manual actions
  – Repeat part (for each MWE of pattern P)
    • Fully automatic

• **Procedure and example** (no parameters)
Further properties

- DuELME does contain models for syntactic structures
  - Based on *de facto* standard for Dutch
  - Used in Alpino, LASSY, CGN treebanks
- DuELME assumes the parameterized ECM
- Encodes several lexical properties
  - auxiliary used for perfect tenses (conjugation)
  - Negative and positive polarity (polarity)
  - Gender of nouns in an MWE
  - …
Further properties

• MWEs have been extracted from corpora
  – After automatic parsing with Alpino
  – Using a variety of statistical and (morpho-)syntactic measures

• Corpora statistics have been included in DuELME
  – E.g., for *een rol spelen* ‘play a role’, tuple= rol spelen, freq=1612
    • Number of ‘rol’: mor1: "sg 1563,pl 49,"
    • Dim form of ‘rol’: dim1: "nodim 1612,"
    • Det with ‘rol’: Det1: "een 918,de 311,die 98,zijn 48,NO 44,deze 38,geen 36,hun 31,welk 20,haar 19,"

• Ten example sentences from these corpora have been included for each MWE
Overview

- MWEs
- Lexical Representation of MWEs
- DuELME
  - DuELME and LMF
- Extensions
- Summary
DUELMEME and LMF

• LMF
  – Abstract metamodel for computational lexicons
  – Represented through UML class diagrams
  – Multiple serialisation options

• DuELMEME-LMF
  – UML class model created for DuELMEME
  – Serialized in XML
DuELME Lexicon

• Lexicon
  – **Lexical Entry** 0..*
  – **MWE Pattern** 0..*

• MWE Pattern
  – **MWE Pattern attributes**
  – **MappingList**
  – **MWE Node**

• (see the example MWE and pattern in the handout)
DuELME and LMF

• DuELME-LMF v. LMF
  – Compare DuELME Class Model with LMF Core Package
  – Compare DuELME Class Model with LMF NLP MWE patterns extension (normative)
LMF Core Package
LMF NLP MWE extension
DuELME and LMF

• DuELME Class Model v. LMF Core Package
  – no Lexical Resource and Global Information
    • This is an error
  – Lexical Entry: no Form Class (but LMF requires one)
    • Not needed for MWEs
    • Not desirable for components of MWEs since DuELME is a *proto-lexicon*
LMF Core Package

Global Information

Lexical Resource

Lexicon

Lexical Entry

Form

Form Representation

Representation

Text Representation

Sense

Definition

Statement
DuELME and LMF

• DuELME Class Model v. LMF NLP MWE Extension
  – Richer but compatible:
    • DataRecords: corpus-derived information
    • ExampleSentence
    • Alternative Components in ComponentList
    • MWE Pattern
LMF MWE Pattern Example

: Lexicon
  language = "eng"
  : List Of Components
    : Lexical Entry
      : Component
      : Component
      : Component
    : Lexical Entry
      : Component
    : MWE Pattern
      id = "VPSomebodyPP"
      comment = "for a pattern, VP somebody IndirectObject"
      : MWE Node
        syntacticConstituent = "VP"
        : MWE Edge
          function = "directObject"
        : MWE Node
          syntacticConstituent = "NP"
          semanticRestriction = "human"
          : MWE Lex
            componentRank = "1"
            graphicalSeparator = "space"
            structureHead = "yes"
        : MWE Lex
          componentRank = "2"
          graphicalSeparator = "space"
      : MWE Edge
        function = "IndirectObject"
      : MWE Node
        syntacticConstituent = "PP"
        grammaticalNumber = "plural"
      : MWE Lex
        componentRank = "3"
        graphicalSeparator = "space"
Overview

- MWEs
- Lexical Representation of MWEs
- DuELME
- DuELME and LMF
- Extensions
- Summary
NOT in DuELME

- Meaning
- Semantic selection restrictions
- Translation
Meaning

- MWEs are described as a special kind of Lexical Entry
- Sense class, and all its dependents, can be used as with single word lexical entries
Meaning

• For collocations and semi-transparent idioms the meaning of each part?
  – Zware shag (lit. heavy tobacco, ‘strong tobacco’) -> zwaar-a-3 shag-n-1
  – Varkentje wassen (lit. pig-DIM wash)-> varkentje-n-1, wassen-v-7
  – Flater slaan (lit. blunder hit)-> flater-n-1 slaan-v-10

(Sense IDs from Cornetto or should be added to Cornetto)
Meaning

• And how they are combined(?)
  – Or maybe this follows from their syntactic manner of combination?
• LMF makes no specific provisions for this
• Perhaps by adding a MWE in the other languages’ lexicons (‘address problem’)

Semantic selection restrictions

• DuELME already specifies
  – Syntactic variables, and syntactic selection restrictions
  – Semantic variables, and semantic selection restrictions
  – Their mutual relation

• But not linked to Sense
  – This should be adapted
Translation

- Supports semantics based translation, possibly interlingual, and transfer
- Relations between entries from lexicons of different languages can be adopted straightforwardly for MWEs in DuELME
Figure I.1 – Multilingual notations model
Overview

• MWEs
• Lexical Representation of MWEs
• DuELME
• DuELME and LMF
• Extensions

➢ Summary
Summary

• DuELME
  – Lexical entries for MWEs
  – With focus on syntax
    • Almost no semantics
    • No translational equivalence
  – Still very incomplete
    • Lacks many syntactic restrictions (e.g. passivisation)
    • Semantic restrictions mostly not specified
Summary

• DuELME
  – Encoded in LMF
    • But some improvements are needed
    • Proposes some deviations
  – Explicit Semantics:
    • only partly (ISOCAT, CLARIN Concept Registry)
    • not formally encoded in the schema yet
Summary

• DuELME
  – highly theory-neutral but
    • Specifically aimed at NLP systems with an explicit grammar
    • Some parts are highly Dutch-specific
THANKS FOR YOUR ATTENTION
References


DO NOT ENTER HERE
DuELME Lexicon

- Lexical Entry (see also the example)
  - Lexical Entry attributes
  - List of Components
  - DataRecords
  - Example Sentence
  - List of SyntacticVariables
  - List of SemanticVariables
  - List of SynSemVar Maps
DuELME Lexicon

• List of Components
  – {Component}
  – Component attributes to express the parameters
  – Lemma with attributes for the writtenform and the (separable) particle
DuELME Class Model
DuELME Lexicon

• Example Sentence
  – Full sentence and a tokenized version
DuELME Lexicon

• DataRecords
  – For tuples identified as candidate MWEs
  – Contains statistics on occurring arguments, modifiers, determiners, morphosyntactic properties, etc
  – Formally structured but not in the class model hence not in XML
  – Tuple $\neq$ MWE
DuELME Class Model
DuELME Lexicon

• List of Syntactic Variables
  – syntactic open slots and restrictions
  – Restrictions: syntactic selection
  – E.g. HETVP, VP, NOHETSSSUB, …

• List of Semantic Variables
  – semantic open slots and restrictions
  – Restrictions: limited number semantic selection restrictions
  – E.g. ANIM, NONANIM, FEM PL, …
DuELME Lexicon

• List of SynSemVar Maps
  – relates syntactic and semantic open slots

• Analogous to the NLP syntax and NLP Semantics extensions [ISO 08, pp 32, 38]
DuELME Class Model
DuELME Lexicon

• Lexical Entry attributes
  – Expression (text)
  – **PatternId** (text)
  – Type: collocation or unspecified
  – [Conjugation]: H (*have*), Z (*be*) or B (*both*)
  – [Comments] (text)
  – [Polarity]: NPI or PPI
DuELME Class Model
DuELME Lexicon

• MWE Pattern attributes
  – ID
  – Description
  – [comments]

• MappingList
  – Needed to relate actual example to tree model

• MWE Node
  – Used to define the syntactic tree model
DuELME Class Model

Lexicon

Lexical Entry

- expression
- patternID
- type
- conjugation
- polarity
- comments

DataRecords

DataRecord

- tuple
- frequency
- corpus size

Example Sentence

- sentence

Example Word

- value

SynSemVarMaps

List of Syntactic Variables

- syntacticFeature
- semanticFeature

Semantic Variable

- label
- restriction

SynSemVarMap

Lexicon

Data

Corpus Examples

Data Entry

- key
- value

Alternative Component

- parameters
- form

Component

- parameters
- form

Lemma

- written Form
- particle

Mapping List

- mapping
- {ordered}

Mapping Number

- value

MWE Lex

- componentRank
- graphicalSeparator

MWE Node

- syntacticCategory

MWE Edge

- syntacticFunction
Lexical

• Lexical
  – De plaat poetsen ‘the plate polish’

• NOT any synonym:
  – Poetsen: afnemen-v-4, doen-v-8, kuisen-v-2 reinigen-v-1, schoonmaken-v-1
  – Plaat: afbeelding-n-1, plaatje-n-4, plaatje-n-6, draaischijf-n-1, grammofoonplaat-n-1, bank-n-3, schol-n-3
  – Een poging wagen / doen / *maken
  – *dare / *do / make an attempt
  – Perdre la tête/ la boule / *la cervelle
  – Se creuser la tête / * la boule / la cervelle
Orthographic

• Orthographic
  – viz. , Bijv., i.v.m., [http://www.uilots.nl](http://www.uilots.nl)
  – Yahoo! , Groen!
  – Aujourd’hui (v. l’homme)
  – ‘s (avonds/morgens/middags)
    • D-gen evening-gen / morning-gen / afternoon-gen
    • In the evenings / mornings / afternoons

• Is dependent on the tokenization rules (cf. the normal rules of combining them)
Optional Intervocalic /d/ deletion obligatory in some MWEs [Zonneveld 1978]

<table>
<thead>
<tr>
<th>expression</th>
<th>literal</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over de rooie / *rode (gaan/zijn/raken)</td>
<td>Over the red / red (go/be/get)</td>
<td>Lose one’s cool</td>
</tr>
<tr>
<td>Om de dooie / *dode donder niet</td>
<td>For the dead / dead thunder not</td>
<td>Absolutely not</td>
</tr>
<tr>
<td>Je niet in de kouwe / *koude kleren gaan zitten</td>
<td>You not in the cold cloths go sit</td>
<td>Affect you seriously</td>
</tr>
<tr>
<td>Een gouwe /* gouden ouwe / *oude</td>
<td>A gold old</td>
<td>A classical music hit</td>
</tr>
</tbody>
</table>
## Morphological

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Example</th>
<th>Literal</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obl. diminutive</td>
<td>Het lood*(je) leggen</td>
<td>The lead-DIM lay</td>
<td>‘die’</td>
</tr>
<tr>
<td>Obl. diminutive</td>
<td>Dat varken*(tje) wassen</td>
<td>That pig-DIM wash</td>
<td>‘address that problem’</td>
</tr>
<tr>
<td>Obl. plural</td>
<td>De *raap is / rapen zijn gaar</td>
<td>The turnip is / turnips are cooked</td>
<td>‘there is trouble’</td>
</tr>
<tr>
<td>Exceptional morphology</td>
<td>Van goede(n) huize</td>
<td>Of good-EN house-E</td>
<td>From good homes</td>
</tr>
<tr>
<td>Exceptional morphology</td>
<td>Zonder aanzien des personen</td>
<td>Without regard the-GEN person-GEN</td>
<td>Without respect of persons</td>
</tr>
</tbody>
</table>
### Syntactic

<table>
<thead>
<tr>
<th>Syntax</th>
<th>Example</th>
<th>Literal</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obl. indefinite</td>
<td>(*de) rekening houden met</td>
<td>(*the) count keep with</td>
<td>‘take into account’</td>
</tr>
<tr>
<td>Oblig no –e suffix</td>
<td>Het bijvoeglijk(*e) naamwoord</td>
<td>The adjectival nominal</td>
<td>‘the adjective’</td>
</tr>
<tr>
<td></td>
<td>(v. het klein*(e) meisje</td>
<td>The little girl</td>
<td>‘The little girl’</td>
</tr>
<tr>
<td>Exceptional government</td>
<td>Ten gevolg*(e) van v.</td>
<td>To consequence of</td>
<td>‘as a consequence of’</td>
</tr>
<tr>
<td></td>
<td>Als gevolg(*e) van</td>
<td>As consequence of</td>
<td></td>
</tr>
</tbody>
</table>
## Semantic

<table>
<thead>
<tr>
<th>Expression</th>
<th>Literal</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>De plaat poetsen</td>
<td>Polish the plate</td>
<td>‘bolt’</td>
</tr>
<tr>
<td>Dat varkentje wassen</td>
<td>Wash that little pig</td>
<td>‘address that problem’</td>
</tr>
<tr>
<td>Een bok schieten</td>
<td>Shoot a goat</td>
<td>‘make a blunder’</td>
</tr>
<tr>
<td>Een flater slaan</td>
<td>Hit a blunder</td>
<td>‘make a blunder’</td>
</tr>
</tbody>
</table>
Pragmatic

• Pragmatic
  – Ladies and Gentlemen
  – Ik heb gezegd. (lit. I have said)
  – Eet smakelijk! (Bon appétit!, Enjoy!)
  – Sincerely yours
## Translational

### Translational properties

<table>
<thead>
<tr>
<th>Expression</th>
<th>Literal</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laten zien</td>
<td>Let see</td>
<td>E. show, F. montrer</td>
</tr>
<tr>
<td>Witte wijn</td>
<td>White wine</td>
<td>P. vinho verde</td>
</tr>
<tr>
<td>Nuclear power plant</td>
<td></td>
<td>D. atoomcentrale, G. Kernkraftwerk</td>
</tr>
<tr>
<td>Space probe</td>
<td></td>
<td>F. Sonde spatiale</td>
</tr>
<tr>
<td>Iemand iets laten weten</td>
<td>Someone something let know</td>
<td>E. inform someone of something</td>
</tr>
</tbody>
</table>
The normal rules

• Example: MWE?
  – iemand een zoen geven
  – Someone a kiss give
  – Give someone a kiss

• Productively related
  – van iemand een zoen krijgen
  – From someone a kiss get
  – ‘be kissed by someone’
The normal rules

• Instead of *zoen-n-1* one can also have other words meaning ‘body touch’

• *kus-n-1* and its hyponyms
  – lik-n-4, smak-n-3, smok-n-1, afscheidskus-n-1, kushandje-n-1, french kiss-n-1, tongkus-n-1, tongzoen-n-1, doodskus-n-1, nachtkus-n-1, nachtzoen-n-1, klapzoen-n-1, smakker-n-1, voetkus-n-1, vredekus-n-1, vredeskus-n-1, handkus-n-1, judaskus-n-1, zuigzoen-n-1

• *liefkozing-n-1*, ‘caress’

• Words meaning ‘kick’, ‘slap’ and other forms of ‘body touching’

• schop-n-1, trap-n-2, fleer-n-1, haal-n-2, klap-n-2, muilpeer-n-1, opflikker-n-1, peer-n-4, klets-n-3, mep-n-1, pats-n-2, pets-n-1, tik-n-1, tikje-n-2, duw-n-1, zet-n-1, zetje-n-1, por-n-1, stoot-n-1, schouderduw-n-1, kontje-n-2, kniecheck-n-1, schop-n-1, trap-n-2, doodschop-n-1, hakje-n-1, kukkel-n-1, knietje
The normal rules

• But not:
  – aanraking-n-2, contact-n-1, gefriemel-n-1, gefrunnik-n-1, gepriegel-n-1, aanslag-n-5, steek-n-1, touche-n-3, betasting-n-1, kneep-n-1, handtastelijkheid-n-2, aanraking-n-1, beroering-n-2, gewelddadigheid-n-1, geweldpleging-n-1, molest-n-1, molestatie-n-1, bal-n-7, schot-n-2,
  – (meaning ‘touch’, ‘contact’, etc.)

• And unclear:
  – lik-n-1, aai-n-1, streling-n-1
  – (‘lick’, ‘caress’, ‘caress’)
The normal rules

• describe such constructions by means of properties of the verbs *geven* and *krijgen*?
  – preferable given its productive nature
  – Only if we can characterize the relevant words by means of independently required properties

• NLP context
  – We might invent an ad-hoc feature
  – But are there resources with this feature? (not Dutch Wordnet ([Cornetto]))
Reflexive Verbs

• Example
  – *Hij schaamt *(zich)
  – He ashamed  REFL
  – ‘he is ashamed’

• Analysis
  – Schamen: reflexivity=true
  – Rule that spells out right reflexive pronoun
Verb Particle Combinations

• Example
  – *Houden* = ‘keep’, transitive
  – *Op + houden* = ‘stop’, intransitive

• Analysis
  – Op + houden:
    • *houden*: particle = *op*, intransitive
    • Rule to introduce / check presence of the right particle
  – Houden: particle = _, transitive
Prepositional Complements

• Example
  – *Houden* ‘keep’ v.
  – *Houden van* (lit. keep of, ‘love’)

• Analysis
  – *houden van*, intransitive, takes PCOMP
    • *houden* with property: complprep = *van*
    • Rule to introduce / check presence of *van*
  – *Houden*: complprep = _, transitive
Inflection

• *Plegen 1*, regular conjugation (pleegde) ‘commit’
• *Plegen 2*, irregular conjugation (placht) ‘do usually’
• Hij pleegde een moord => regular conjugation
• He committed a murder
• *Hij placht een moord*
Selection

• Example 1
  – *Nemen* 1 subcat=[subj/NP, obj/NP] ‘take’
  – *Nemen* 2 subcat=[subj/NP, obj/NP, compl/PP] ‘accept as’
  – *Iets in acht nemen*
  – *something in attention take*, ‘obey’ (of rules etc.)
  – Requires *nemen_2*
Selection

• Example 2
  – *Geven* `give’ semantically takes 3 arguments
  – Syntactically: subj/NP, obj/NP, iobj/NP or PP
  – Indirect object optional
  – Absent indirect object still leads to an interpretation with 3 arguments

• But MWE *een gil geven* lit. a cry give, `give a shout’ requires 2 syntactic arguments,
  *Idem*: de geest geven (the ghost give) ‘die’
Selection

• Example 3
  – *Heten* `be called’ 2 arguments
  – Syntactically: subj/NP, predc/NP
  – Ik heet Jan
  – I am-called Jan
  – But MWE *iemand welkom heten* lit. someone welcome be-called, `welcome someone’ requires 3 syntactic arguments, subj, obj, predc
Selection

- Many such cases with support-verb constructions
  - Aandacht hebben voor, etc.
  - See handout (5)
  - These require special treatment
SEQCI

• Example:
  – Idiom Descriptions
    • Idp30;De pijp uit gaan;Hij is de pijp uit gegaan
    • Idp30;De boot in gaan;Hij is de boot in gegaan
    • Idp30;Het schip in gaan;Hij is het schip in gegaan
  – Idiom pattern definition
    • Idp30
    • Idiom headed by a verb taking a postpositional PP containing a definite singular NP and one free argument as subject
SEQCI

• Incorporation Method
  – Bootstrap part, once for each idiom pattern
  – Repeat Part, for each idiom description
SEQCI

- Bootstrap part (‘hij is de pijp uit gegaan’)
  1. Parse the example sentence of an idiom description with idiom pattern P, yielding the Reference Parse
  2. Define a transformation to turn the reference parse into the idiom structure (Parse Transformation, PT)
  3. Determine the list of unique IDs of the lexical items in the idiom structure for the system derived from the reference parse (Idiom Component ID List, ICIL)
  4. Define a transformation to relate ICL and ICIL (Idiom Component Transformation, ICT)
  5. Apply the ICT to the ICL, yielding the transformed ICL (TICL) and check that each item in it equals the base form of the corresponding element on the ICIL
Repeat part, for each idiom description I
(‘hij is de boot in gegaan’)
1. Parse example sentence (Syntactic Structure)
2. Apply IPT and check identity with idiom structure modulo the lexical items
3. Select the component IDs from the parse tree, in order to obtain the ICIL
4. Apply ICT to the ICL of I, yielding the TICL
5. Check that <bf(c1),…bf(cn)>=TICL
where ICIL = <c1, …cn> (TICL check)
SEQCI

• **Advantages**
  – Technically Simple
  – As theory/grammar/implementation-independent as possible
  – No need for prescribing syntactic structures
  – System-specific aspects are derived from the NLP-system itself
SEQCI: Reference Parse

Rdecl[Rperf
  [Rsubst(j)
   [Rsent
    [Rsubst(i)
     [RVP[$aV_00_ga,
       RPPpost
         [$s_prep1286700,
           VAR_i
         ]
       ]
     ]
   ]
RNPdef [$aN_00_pijp]
],
VAR_j
],
RNP[$hij_PRON]
]
SEQCI: Idiom Structure

- **IPT:** IPT: Delete Rdecl, Rperf, Rsubj(j), RNP[$hij_Pron]

- **D-tree for vpid30 (simplified):**

  ```
  Rsubst,i
  [RVP  [$aV_00_ga,
           RPPpost
           [$s_prep1286700,
            VAR_i
           ]
  ],
  RNPdef [$aN_00_pijp]
  ]
  ```
ICIL

<$aV_00_ga, $prep1286700, $aN_00_pijp >
ICT

ICL: <de, pijp, uit, gaan>
Must be turned into: <gaan, uit, pijp>

ICT: 1 2 3 4 => 4 3 2
TICL

TICL = ICT(ICL) =
ICT(<de, pijp, uit, gaan>) =
<gaan, uit, pijp> =
< Bf($aV_00_ga), Bf($prep1286700),
  Bf($aN_00_pijp) >
Syntactic Structure

Rdecl[Rperf
  [Rsubst(j)
    [Rsent
      [Rsubst(i)
        [RVP[$aV_00_ga,
          RPPpost
            [$s_prep1286800,
              VAR_i
            ]
        ],
        RNPdef [$aN_00_boot]
      ],
      VAR_j
    ],
    RNP[$hij_PRON]
  ]
]
Apply IPT

Rsubst, i

[RVP

[$aV_00_ga,
RPPpost

[$s_prep1286800,

VAR_i

],

],

RNPdef [$aN_00_boot]

]
ICIL

ICIL=<$aV_00_ga, $s_prep1286800, $aN_00_boot$>
TICL

ICT(ICL) =

ICT(<de, boot, in, gaan>) =

<gaan, in, boot>
TICL check

\(<bf($aV_00_ga), bf($s_prep1286800), bf($aN_00_boot) \geq\>

TICL =

<gaan, in, boot>
The normal rules

• Fixed combinations of open class word and closed class word
  – Reflexive verbs
  – Verb particle combinations
  – Prepositional complements

• Described by means of a property of the open class word + special rules
  ➔ no MWEs in these systems