MULTI-WORD EXPRESSIONS:
LINGUISTIC PROPERTIES
AND LEXICAL REPRESENTATION

Manfred Sailer\textsuperscript{1}  Shuly Wintner\textsuperscript{2}

\textsuperscript{1}Institute for English and American Studies
Goethe University Frankfurt
Frankfurt am Main, Germany

\textsuperscript{2}Department of Computer Science
University of Haifa
Haifa, Israel

PARSEME Prague Training School
19-23 January, 2015
1 **Overview: linguistic properties of MWEs**

2 **MWEs in linguistic theory**

3 **MWEs in linguistic theory 2: Beyond English**

4 **Challenges from other languages (Hebrew)**

5 **Lexical Encoding of MWEs**

6 **Applications to other languages**

7 **Summary**
Multi-word expressions (MWEs) are lexical units consisting of more than a single orthographic word.

Orthographically, they are written with intervening spaces.

Morphologically, their behavior is often idiosyncratic.

Syntactically, they may function as words or as phrases and are often more restricted than compositional phrases.

Semantically, their meaning is usually non-compositional (i.e., cannot be established from the meanings of their components).

MWEs blur the boundaries between the lexicon and the grammar.
A heterogeneous class of phenomena with diverse sets of characteristics:

Lexical phrases are chunks of language of varying length, conventionalized structures that occur more frequently and have more idiomatically determined meaning than language that is put together each time (DeCarrico and Nattinger, 1993)

A prefab is a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization (Erman and Warren, 2000)

Idiosyncratic interpretations that cross word boundaries (or spaces) (Sag et al., 2002)
That was not such a strange thing, nor did Alice think it so much out of the way to hear the Rabbit say, “Oh dear! Oh, dear! I shall be late!” But when the Rabbit took a watch out of its pocket, and looked at it and then ran on, Alice started to her feet, for she knew that was the first time she had seen a Rabbit with a watch.
**Definition**

MWEs are expressions consisting of more than one word that have to be stored in the lexicon because they exhibit some idiosyncratic behavior, be it orthographic, morphological, syntactic, or semantic.
MWEs constitute a major part of any language, and the magnitude of this phenomenon is far greater than has traditionally been realized within linguistics.

Jackendoff (1997, page 156) estimates that the number of MWEs in a speaker’s lexicon (in English) is of the same order of magnitude as the number of single words.

In WordNet 1.7 (Fellbaum, 1998), 41% of the entries are multiwords.

Erman and Warren (2000) revealed that over 55% of the tokens in the texts they studied were instances of what they call *prefabs*. 
MULTIWORD EXPRESSIONS

Scale

Though our dear friend kicked the bucket, let us remember to not cry over spilt milk for the grass is always greener on the other side...
Expressions with idiosyncratic features that cannot be predicted on the basis of their component words must be included in language descriptions (such as lexicons) in order to account for actual usage.

Identification of MWEs is an important task for a variety of natural language processing (NLP) applications (Villavicencio et al., 2005):

- **INFORMATION RETRIEVAL** (Doucet and Ahonen-Myka, 2004)
- **TEXT ALIGNMENT** (Venkatapathy and Joshi, 2006)
- **MACHINE TRANSLATION** (Baldwin and Tanaka, 2004; Uchiyama et al., 2005)
MULTIWORD EXPRESSIONS

SIGNIFICANCE
MWEs are a challenge for computational processing of natural languages because they combine properties of words and phrases, and because phonological, morphological and orthographic processes apply to them differently than to ordinary tokens (Sag et al., 2002; Copestake et al., 2002; Villavicencio et al., 2004).

They are even more challenging in languages with complex morphology, because of the unique interaction of morphological and orthographical processes with the lexical specification of MWEs (Oflazer et al., 2004; Alegria et al., 2004; Savary, 2008; Al-Haj et al., 2014).
MWEs and linguistic theory (mostly English, some German)
Challenges in other languages (with examples from Hebrew)
Lexical representation of MWEs
Integrating MWE lexicons in NLP applications
1. **Overview: Linguistic Properties of MWEs**

2. **MWEs in Linguistic Theory**

3. **MWEs in Linguistic Theory 2: Beyond English**

4. **Challenges from Other Languages (Hebrew)**

5. **Lexical Encoding of MWEs**

6. **Applications to Other Languages**

7. **Summary**
**What is an idiom?**

phraseology, phraseme, phraseological unit, multiword expression, ...  

**Definition**

MWEs are expressions consisting of more than one word that have to be stored in the lexicon because they exhibit some idiosyncratic behavior, be it orthographic, morphological, syntactic, or semantic.

**phrasal +**

1. **idiomatic:** non-literal meaning; holistic meaning
2. **lexically fixed:** words cannot be exchanged
3. **syntactically fixed:** restricted syntactic flexibility
4. **lexicalized:** conventionalized combination; represented as one unit

Not considered:

- orthographic: theory
- morphological: see rest of the week
**Some examples**

**Example (kick the bucket (‘die’))**

1. idiomatic: ok
2. lexically fixed: \( \neq \text{kick the pail}; \neq \text{throw the bucket} \)
3. syntactically fixed: *The bucket was kicked.*
4. lexicalized: ok
**Some examples cont.**

**Example (Spill the beans (‘reveal information’))**

1. Idiomatic: ok

2. Lexically fixed: \( \neq \) spilled the vegetables; \( \neq \) dropped the beans

3. Syntactically fixed?:
   
   - The beans were spilled.
   - The beans appeared to be spilled.
   
   * The beans, Pat spilled.

4. Lexicalized: ok
SOME EXAMPLES cont.

**Example (make headway (‘make progress’))**

1. idiomatic: no? (cranberry word/bound/fossil word)

2. lexically fixed: ?? achieve headway

3. syntactically fixed?
   - Considerable headway was made.
   - How much headway did they make on the job?
   - *That much headway I’m sure they made on the job.* (Postal, 1998, p. 31)

4. lexicalized: ok
Some Examples cont.

Example (Brush one’s teeth (‘Clean one’s teeth’))

1. Idiomatic: No? (collocation, idiom of encoding)
2. Lexically fixed?
   - I brushed my choppers.
   - I cleaned/polished my teeth
3. Syntactically fixed?
   - The teeth were brushed.
   - Those teeth he hadn’t brushed in ages.
4. Lexicalized?
**OVERVIEW**

Phase 1: \( \text{syn. fixedness} = \text{sem. fixedness} = \text{fixed} \)

Phase 2: \( \text{synt. fixedness} = \text{sem. fixedness} = \text{fixed} \)
\( \text{synt. fixedness} \neq \text{sem. fixedness} \neq \text{fixed} \)

Phase 3: \( \text{synt. fixedness} = \text{sem. fixedness} = \text{fixed} \)
\( \text{synt. fixedness} \neq \text{sem. fixedness} = \text{fixed} \)
\( \text{synt. fixedness} = \text{sem. fixedness} \neq \text{fixed} \)
Consider, for example, such phrases as ‘take for granted’, which abound in English. From a semantic and distributional point of view, this phrase seems to be a single lexical item, and it therefore must be entered in the lexicon as such, with its unique syntactic and semantic features. On the other hand its behavior with respect to transformations and morphological processes obviously shows that it is some sort of Verb-with-Complement construction. Once again we have a lexical item with a rich internal structure (Chomsky, 1965, p. 190)
**Historical Overview**

- **Chafe 1968**: Four problems of idioms:
  - non-compositional
  - transformationally defective
  - (sometimes) syntactically ill-formed
  - idiomatic reading of a combination is more frequent than literal meaning.

- **Weinreich 1969**:
  - Phrasal lexical entry lists all possible transformations.

- **Fraser 1970**:
  - Idioms inserted with structure in Deep Structure
  - Classification according to syntactic flexibility.

- **Jackendoff 1975**: Phrasal lexical entries with only partial specification, for syntactically regular idioms: structure follows from syntactic rules as lexical redundancy rule.
**Idiom arguments in Principles and Parameters**

As summarized in Nunberg et al. 1994:

- Idiom inserted en bloc at Deep Structure (DS)
- Transformations apply to DS trees, even if of idiomatic origin.
- More recently: Compositional aspects of idioms used to motivate functional projections (*X gave Y the boot* — *Y got the boot from X*, Richards 2001)
Idiom arguments in principles and parameters

As summarized in Nunberg et al. 1994:

Predictions:

P1 Idioms have a regular syntactic structure.

P2 Idioms can have
  - only canonical form,
  - or canonical and transformed form;
  - but never only transformed form

P3 Only the idiom as a whole has a meaning, idiom parts are not assigned meaning.
IMPORTANT PUBLICATIONS TO CHANGE OUR VIEW ON IDIOMS

- Higgins 1974: Critique of en bloc insertion, attempt of a more semantic theory; unpublished
- Ernst 1981: Modifiers inside idioms as argument against monolithic semantics of idioms
- McCawley 1981: Shows a transformational paradox for idioms in relative clauses
- Wasow et al. 1983; Nunberg et al. 1994: Two classes of idioms distinguished by decomposability (also: Langacker 1987)
- Ruwet 1991: Lists arguments against the traditional en bloc insertion view
EVIDENCE AGAINST P1

Chafe 1968; Nunberg et al. 1994: There are idioms that don’t have a regular syntactic structure

**EXAMPLE**

- *trip the light fantastic* (‘dance’)
- *kingdom come* (‘eternity’)
- *easy come easy go* (‘what you get easily, you lose easily’)
- *every which way* (‘in every direction’)
EVIDENCE AGAINST P2

Nunberg et al. 1994: idioms that only occur in non-canonical constructions, i.e. only as “transformed”

**Example**

- passive only: *(be)* cast/carved/set in stone (*‘cannot be changed’*)
- Wh-moved only: the hell (as in *What the hell are you doing?*)
- inverted only: *Is the pope catholic?* (*‘of course’*)
- imperative only: *Break a leg!* (*‘Good luck!’*)
EVIDENCE AGAINST P3

Ernst 1981: Idiom parts that show independent meaning

**Example (Modification (Ernst, 1981))**

External modification:

1. *Pat kicked the social bucket.* ( = *Socially, Pat kicked the bucket.*)
2. *Pat pulled some economic strings.* ( = *Pat pulled some strings in economy.*)

Internal modification:

1. *Katz and I had by then become good friends, having long before buried the old hatchet* (L. Melamed, *Escape to the Future*)
2. *My girls should’ve buried the damn hatchet when they were in their prime.* (www; expressive modifier)
3. *Pat pulled some important strings.* ( = *Pat used some important connections.*)
EVIDENCE AGAINST P3 (CONT.)

Example (Determiner variation)

1. Pat kicked the/*a bucket.
2. I have buried many hatchets with my parents but this still burns me up. (www)
3. his reputation, however, was as a figure who pulled many strings. (www)

Internal modification and determiner variation are strong evidence for meaningful idiom parts.
Additional Problem: McCawley’s Transformational Paradox

If the idiom *pull strings* must be inserted as one VP unit from the lexicon, there is a paradox:

**Example**

1. The strings that Pat pulled got Chris the job.
   - bad if *strings* originates in the surface position
   - ok if *strings* originates inside the relative clause

2. Pat pulled the necessary strings that got Chris the job.
   - ok if *strings* originates in the surface position
   - bad if *strings* originates inside the relative clause
Summary: Classical en bloc insertion

- Basic idea: Insert idiom as syntactic and semantic unit.
- In TG/GB/P&P/MP: Insertion in canonical form
- Counterarguments:
  - irregular syntactic structure
  - obligatory non-canonical construction ("transformed only" idioms)
  - internal modification
  - transformational paradox
**Overview**

Phase 1: \( \text{syn. fixedness} = \text{sem. fixedness} = \text{fixed} \)

Phase 2: \( \text{synt. fixedness} = \text{sem. fixedness} = \text{fixed} \)
\( \text{synt. fixedness} \neq \text{sem. fixedness} \neq \text{fixed} \)

Phase 3: \( \text{synt. fixedness} = \text{sem. fixedness} = \text{fixed} \)
\( \text{synt. fixedness} \neq \text{sem. fixedness} = \text{fixed} \)
\( \text{synt. fixedness} = \text{sem. fixedness} \neq \text{fixed} \)
TWO CLASSES OF IDIOMS

Wasow et al. 1983; Nunberg et al. 1994:

- **Idiomatic phrases (IPh):**
  - *kick the bucket* (‘die’), *saw logs* (‘snore/sleep’), *trip the light fantastic* (‘dance’)
  - idiom parts cannot occur in positions/constructions that require content
    → syntactically and semantically fixed

- **Idiometrically combining expressions (ICE):**
  - *spill the beans, keep tabs on s.o./s.th.* (‘watch s.o./s.th. closely’),
    *make headway* (‘make progress’), *bury the hatchet* (‘stop arguing/fighting’)
  - idiom parts can occur in positions/constructions that require content
    → syntactically and semantically flexible
MWEs in linguistic theory

Tests for ICEs 1: relatively language independent

If an idiom part can occur in a position/construction that must have some meaning, the idiom is decomposable.

- Internal modification possible
- Determiner change possible

Example (Determiner variation)

1. We’ve made some headway this year.
2. *Alex was sawing many logs last night. (saw logs ‘snore’)

Manfred Sailer and Shuly Wintner

Ling. properties and lexical representation

Prague, January 2015
If an idiom part can occur in a position/construction that must have some meaning, the idiom is decomposable.

**Example (Pronominalization)**

1. Eventually they spilled the beans, but they didn’t spill them deliberately.
2. Kim’s family pulled some strings on her behalf, but they weren’t enough to get her the job. (Nunberg et al., 1994)
3. *Pat kicked the bucket and Chris kicked it too.
4. *Pat tripped the light fantastic but Alex didn’t want to trip it.
If an idiom part can occur in a position/construction that must have some meaning, the idiom is decomposable.

**Example (Movement)**

**Fronting/topicalization**

1. *The strings Pat has pulled.*
2. *The bucket Pat has kicked.*

**It*-cleft:**

1. *...and we react with thanks as if it were some huge string that he pulled on our account.* (www)
2. *It was the bucket that John kicked* (Schenk, 1995)
If an idiom part can occur in a position/construction that must have some meaning, the idiom is decomposable.

**Example (Relative clause)**

1. Partially inside a RelC:
   - *The strings that Pat pulled got Chris the job.*
   - *The bucket that Pat kicked was unexpected.*

2. Internal modification by a RelC:
   - *Pat pulled the strings that got Chris the job*
   - *Pat kicked the bucket that nobody expected.*
If an idiom part can occur in a position/construction that must have some meaning, the idiom is decomposable.

**Example (Passive, raising)**

1. *I’m pleased that sufficient strings have been pulled, . . .* (www)
2. *The bucket has been kicked.* (Makkai, 1972, p. 150)
PROBLEMS WITH THE TESTS

- Internal modification
  - Maybe external? *pull economic strings*
  - conjunction interpretation (Ernst, 1981):
    *They had to tighten their Gucci belts.*
  - eventive nouns:
    *How the CIA waged a secret war against Cuba* (www)

- Language-specific tests: Similar morpho-syntactic processes in different languages may differ with respect to meaningfulness of constituents.
Decomposability problematic/circular?

(See discussion in Nunberg et al. 1994 and Svensson 2008)

Decomposability is a purely semantic notion; not to be confused with:

- ≠ transparency of the expression as a whole (relation between literal and idiomatic meaning):
  - *bury the hatchet* (‘stop the fighting’): transparent, decomposable
  - *saw logs* (‘snore’): transparent, non-decomposable
  - *spill the beans* (‘divulge information’): non-transparent, decomposable
  - *shoot the breeze* (‘chat’): non-transparent, non-decomposable

- ≠ plausible paraphrasability:
  - *bury the hatchet* = *stop the fighting*: paraphrasable, decomposable
  - *kick the bucket* = *lose one’s life*: paraphrasable, non-decomposable
TWO CLASSES

- Decomposability is defined via tests for meaningfulness of idiom components
- An expression that passes some of these tests is decomposable, all others are non-decomposable.
- Nunberg et al. 1994 see a strong connection between semantic decomposability and syntactic flexibility. But see tomorrow’s meeting!
AIMS OF A FORMAL ANALYSIS

What we want:

- Varying syntactic flexibility
- Semantics of the well-formed strings

What we won’t talk about:

- Relation between the literal and the non-literal meaning
- Cognitive basis of idioms
- Word play
- Text-constituting potential of idioms
EXAMPLES OF FORMAL ANALYSES

- Pulman 1993: Inference-based analysis
- Abeillé 1995: Constructional analysis
- Gazdar et al. 1985: Denotational analysis
**Inference-based analysis: Sketch**

- **Representatives:** Pulman 1993, Egan 2008
- Literal parse mapped to idiomatic interpretation:
  - Pulman 1993: sem.repr. $\mapsto$ sem.repr. (special inference rules)
  - Example: $\text{They}[	ext{bucket}(y)][\text{kick}(x, y)] \mapsto \text{die}(x)$
    (applies if the literal reading is inconsistent in the context)
- Syntax non-holistic, semantics holistic
- Idiom is stored as a special inference rule, different from lexical entries.
Inference-based Analysis: Strengths

- no idiomatic words necessary
- literal meaning available; necessary for "extended uses"
- possibly: relation to other cases of figurative language

Example (Extended use):

If you let this cat out of the bag, a lot of people are going to get scratched.
Inference-based Account of Idiom Properties

- **Idiomaticity**: mapping between lexical and idiomatic reading
- **Lexical fixedness**: inference rule can rely on word-specific semantic contributions
- **Syntactic fixedness**: possible, if syntactic structure correlates with different semantic representations
Inference-based analysis: Problems

Problems noted in (Wearing, 2012)

- processing: idiomatic sense sometimes faster than literal sense.
- vague predictions on degree of flexibility

Example (Pronominalization (Egan, 2008))

1. *have a bone to pick with s.o.* ('X has s.th. to discuss where Y annoyed X');
   
   *I had a bone to pick with them, but they were so nice that I forgot about it.*

2. *shoot the breeze* ('chat')
   
   *Tony shot the breeze with Junior, and Paulie shot it with Silvio.*
Inference-based analysis: Problems cont.

- Idioms with bound/cranberry/fossil words? *make headway, the whole (kit and) caboodle* (‘everything’)
- Idioms with syntactic peculiarities? *trip the light fantastic* (‘dance’), *kingdom come* (‘eternity’)
- Pulman 1993: type of inference required elsewhere?
**Constructional analysis: Sketch**

- Representative: Abeillé 1995, Tree Adjoining Grammar
- Idiom is represented as a syntactic tree (elementary tree)
- Nodes in the tree can, but need not have semantic annotation.

![Constructional Analysis Diagram]

```
S
   sem: die'(x)
     NP
       sem: x
       V
         kick
       D
         the
     VP
       sem: die'(x)
       NP
         N
           bucket
```
CONSTRUCTIONAL ANALYSIS: SKETCH

- Representative: Abeillé 1995, Tree Adjoining Grammar
- Idiom is represented as a syntactic tree (elementary tree)
- Nodes in the tree can, but need not have semantic annotation.

\[
S \\
\text{sem: The } y \ [\text{info'}(y)] (\text{reveal'}(x,y)) \\
\text{NP} \\
\text{sem: } x \\
\text{VP} \\
\text{sem: } \lambda x. \text{The } y \ [\text{info'}(y)] (\text{reveal'}(x,y)) \\
\text{V} \\
\text{sem: reveal'} \\
\text{spill} \\
\text{NP} \\
\text{NP} \\
\text{sem: } \lambda P. \text{The } y \ [\text{info'}(y)] (P(y)) \\
\text{D} \\
\text{the} \\
\text{N} \\
\text{beans}
\]
CONSTRUCTIONAL APPROACH: FLEXIBILITY

- Transformations: each elementary tree belongs to a “tree family”, where all possible derived trees are included (such as for passive etc.)
- Modification: Possibility to mark in the structure whether modifiers are possible.
- Internal modification: available if attachment node has meaning
- Pronominalization: unclear
**Constructional approach: Strengths**

- Account of syntactically ill-formed idioms (*trip the light fantastic*), idioms in transformed form only (*Get lost!*), or idioms with bound words (*make* headway).
- All idioms are represented as units.
- Parts of an idiom can have an idiomatic meaning, but only if the rest of the idiom is present.
CONSTRUCTIONAL ACCOUNT OF IDIOM PROPERTIES

- Idiomaticity: done via ambiguity.
- Lexical fixedness: lexical items and word forms are included into the elementary trees.
- Syntactic fixedness: via diacritic marking
CONSTRUCTIONAL APPROACH: PROBLEMS

- Marking for applicable transformations not grounded in semantics
- Analysis of pronominalization not clear
Denotational Approach: Sketch

- Representatives: Gazdar et al. 1985
- Hybrid approach:
  - Idiomatic phrases: fixed tree with meaning is in the lexicon
  - ICE: co-occurrence of idiom parts by special denotations
- Words in idioms are ambiguous:
  - \textit{spill} $\mapsto$ \textit{reveal-idiom}
  - \textit{beans} $\mapsto$ \textit{secret-idiom}
  - Pat spilled the beans:
    - \textit{The $x[\text{secret-idiom}(x)](\text{reveal-idiom}(\text{pat}, x))$}
- Semantic constants as partial functions:
  - $[[\text{reveal-idiom}]]([[\text{beans}]]):$ undefined.
  - $[[\text{spill}]]([[\text{secret-idiom}]]):$ undefined
Denotational Approach: Sketch cont.

- Passive: The beans had been spilled.
  \[ \text{The } x[\text{secret-idiom}(x)](\exists y(\text{reveal-idiom})(y, x)) \]

- Strengths:
  - attempt to encode Nunberg et al. 1994
  - internal modification ok
  - syntactic flexibility related to semantics
DENOTATIONAL ACCOUNT OF IDIOM PROPERTIES

- Idiomaticity: by ambiguity
- Lexical fixedness: via the denotation of special, lexeme-specific predicate-symbols.
- Syntactic fixedness: fixed tree (for IPh) vs. syntactically free combination (for ICE).
Denotational account: Problems

- Phrasal lexical entry for non-decomposable idioms not well defined in Gazdar et al. 1985
- Evidence for lexical ambiguity?
- Complicated underlying denotations (Pulman, 1993)
- Difference between various types of decomposable idioms?

Example

Different degree of flexibility among decomposable idioms

1. *It’s the beans that John spilled.* (Müller, 2012)
2. ...if it were some huge string that hes pulled on our account. (www)
**Summary: Recent Approaches**

- **Inference:**
  - captures extended uses
  - useful for metaphoric idioms and diachronic development of idioms
  - but problematic for fixedness

- **Construction:**
  - emphasizes unit-like behavior of idioms
  - but flexibility stipulated

- **Denotation:**
  - emphasizes semantics
  - but complicated model theory
  - but varying degree of syntactic flexibility
**Towards a Two-dimensional, Representational Analysis**

- Non-decomposable idiom: as completely fixed tree
- Decomposable idiom: normal syntactic combination; semantic constants rather than denotations.

1. Idiomatic phrases: Syntactically (almost) frozen idioms, *kick the bucket*
2. Idiomatically combining expressions: Mobile idioms
   - A. Syntactically connected idioms, *spill the beans*
   - B. Semantically connected idioms, *pull strings*

Two dimensions:
1. Construction (for syntactically irregular and fixed expressions)
2. Collocations (for syntactically flexible MWEs)

Syntactically irregular expressions: Via a phrasal lexical entry

Decomposable idioms: Normal syntactic combination; collocational restrictions on semantic representations

Framework: Head-driven Phrase Structure Grammar (HPSG, (Pollard and Sag, 1994))
Example: *trip the light fantastic*

- Only morphological variation:
  
  *She is tripping/tripped the light fantastic*

- Syntactic structure unclear, syntactically fixed

- Phrasal lexical entry:
  
  syn: VP
  
  head-dtr: [lex-id: *trip*]
  
  nonhead-dtrs: [lex-id: *the*], [lex-id: *light*], [lex-id: *fantastic*]
  
  sem: $\lambda x.\text{dance}(x)$
Example: \textit{trip the light fantastic cont.}

\begin{itemize}
  \item [1] \textit{Alex tripped the light fantastic.}
    \begin{itemize}
      \item ok; VP and the head daughter is of the right lexeme.
    \end{itemize}
  \item [2] * \textit{The light fantastic was tripped.} (passive)
    \begin{itemize}
      \item bad; no VP of the right form
    \end{itemize}
  \item [3] * \textit{The light fantastic, Alex tripped.} (fronting)
    \begin{itemize}
      \item bad; no VP of the right form
    \end{itemize}
\end{itemize}
EXAMPLE: *spill the beans*

Syntactically flexible, semantically decomposable
Two collocationally related words, *spill* and *bean*

- **Idiomatic *spill***: phon: spill
  syn: V
  lex-id: spill-i
  sem: \textit{spill-idiom}(x, y)

  Collocational restriction: A phrase with [lex-id spill-i] must be selected by *spill*

- **Idiomatic *bean***: phon: bean
  syn: N
  lex-id: bean
  sem: \textit{bean-idiom}(y)

  Coll. restriction: The phrase headed by this word must be on a valence list of a word with [lex-id spill-i].
**Example: spill the beans cont.**

**Example (Syntactic flexibility)**

1. *How talking parrot spilled beans on owner’s cheating girlfriend* (www)
   ok; verb and noun occur in the right constellation

2. *But the beans were spilled by her pal Britney Spears earlier this year* (www, passive)

3. *It’s the beans that John spilled* (fronting)
   Phrase *the beans* not on a valence list of *spill*

4. *The beans that Alex spilled chocked Chris.* (relative clause)
   Phrase *the beans* . . . not on a valence list of *spill*

5. *Alex pulled his connections
   *I revealed the beans*  
   bad; collocational restriction not satisfied
**Example:** *pull strings*

Syntactically flexible, semantically decomposable
Two collocationally related words, *pull* and *string*

- **Idiomatic *pull***: phon: pull
  syn: V
  lex-id: pull
  sem: *pull-idiom*(x, y)

  Collocational restriction: In overall semantics, the second argument of *pull-idiom* must also occur as argument of *string-idiom*

- **Idiomatic *string***: phon: string
  syn: N
  lex-id: string
  sem: *string-idiom*(y)

  Coll. restriction: In the overall semantics, the argument of *string-idiom* must occur as the second argument of *pull-idiom*. 
Example: *pulled strings* cont.

**Example (Syntactic Flexibility)**

1. *Virginia Tech has pulled some serious strings to allow meal plans to cover purchases made here.* (www)
   ok; verb and noun occur in the right constellation

2. *Some strings were pulled on their behalf.* (passive)

3. *Some influential strings, Alex pulled.* (fronting)

4. *The strings that Alex pulled got Chris the job.* (relative clause)
   ok; semantic, not syntactic constraint on co-occurrence
   \[\text{They}_{y}[\text{string-id}(y) \land \text{pull-id}(\text{alex}, y)](\text{get-C-job}(y))\]

5. *Alex pulled his connections*
   *I am impressed by Alex’s many strings.*
   bad; collocational restriction not satisfied
**STRENGTHS**

- Coverage of neglected phenomena:
  - bound words (Richter and Sailer, 2003): *make headway*
  - phraseological phrases (Richter and Sailer, 2009): *wissen, wo Bartel den Most holt* (lit.: know where Barthel gets the wine, ‘know what’s going on’)

- Integration of collocations
- Captures different degree of flexibility
- Follows the insights of (Nunberg et al., 1994)
PROBLEMS

- Why collocations and constructions if one is enough? (Kay and Sag, ms.)
- Collocation mechanism very/too? powerful (Soehn, 2006; Sag, 2010) and still under development
- For ICEs: no natural sense of a unit-like representation and stipulated lexical ambiguity
- Flexibility captured by stipulation (Sag, 2010)
Limitations of linguistic theorizing

- Focus on English, sometimes making other languages fit the English model (for example Soehn 2006)
  \[\Rightarrow\text{ from Tuesday on: Beyond English and back to English}\]

- Focus on few classes of MWEs, few examples
  \[\Rightarrow\text{ from Wednesday on: Broader overview, lexical resources}\]
### Overview

<table>
<thead>
<tr>
<th>Phase 1:</th>
<th>synt. fixedness $= \text{sem. fixedness} = \text{fixed}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 2:</td>
<td>synt. fixedness $= \text{sem. fixedness} = \text{fixed}$  &lt;br&gt; synt. fixedness $\neq \text{sem. fixedness} \neq \text{fixed}$</td>
</tr>
<tr>
<td>Phase 3:</td>
<td>synt. fixedness $= \text{sem. fixedness} = \text{fixed}$  &lt;br&gt; synt. fixedness $\neq \text{sem. fixedness} = \text{fixed}$  &lt;br&gt; synt. fixedness $= \text{sem. fixedness} \neq \text{fixed}$</td>
</tr>
</tbody>
</table>
THINK ABOUT FOR TOMORROW

1. Think of 3 idioms/MWEs from your language.
2. Test them for decomposability.
3. Can you sketch an analysis within one of the theories (inference, construction, denotation, 2-dimensions)?
4. Syntactic and morphological process in your language that do not exist in English or differ from English in their properties. (German: scrambling, verb second, Vorfeld placement, impersonal passive, . . .)
OUTLINE

1 OVERVIEW: LINGUISTIC PROPERTIES OF MWEs

2 MWEs in linguistic theory

3 MWEs in linguistic theory 2: Beyond English

4 Challenges from other languages (Hebrew)

5 Lexical Encoding of MWEs

6 Applications to other languages

7 Summary
MWEs are a heterogeneous phenomenon.

Generative Grammar: en bloc insertion at Deep Structure

Ernst 1981, Nunberg et al. 1994 and others: some idioms components have identifiable meaning

Nunberg et al. 1994: English: semantic decomposability correlates with syntactic flexibility
**Plan for today**

- General properties of German
- Non-decomposable idioms in German
- German vs. English
- Non-decomposable idioms in other languages
- Sketch of a theory
**OVERVIEW**

Phase 1: \( \text{syn. fixedness} = \text{sem. fixedness} = \text{fixed} \)

Phase 2: \( \text{synt. fixedness} = \text{sem. fixedness} = \text{fixed} \)
\( \text{synt. fixedness} \neq \text{sem. fixedness} \neq \text{fixed} \)

Phase 3: \( \text{synt. fixedness} = \text{sem. fixedness} = \text{fixed} \)
\( \text{synt. fixedness} \neq \text{sem. fixedness} = \text{fixed} \)
\( \text{synt. fixedness} = \text{sem. fixedness} \neq \text{fixed} \)
General properties of German syntax

Topological fields

- Semi-flexible word order
- Höhle 1986: Topological fields
  - Mittelfeld: free order, determined by information structure
  - Verbal complex: only verbs, own order restrictions
  - Nachfeld: prosodically heavy/clausal constituents
  - FINIT: finite verb in root clauses (V1 or V2)
  - Vorfeld: any constituent in root clause (V2)

<table>
<thead>
<tr>
<th>C</th>
<th>Mittelfeld</th>
<th>verbal complex</th>
<th>Nachfeld</th>
</tr>
</thead>
<tbody>
<tr>
<td>dass</td>
<td>Alex gestern [einen Vogel]</td>
<td>mitgebracht hat</td>
<td>[der schön singt].</td>
</tr>
<tr>
<td>that</td>
<td>Alex yesterday [a bird]</td>
<td>brought.along has</td>
<td>[that sings well]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vorfeld</th>
<th>FINIT</th>
<th>Mittelfeld</th>
<th>verbal complex</th>
<th>Nachfeld</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>Hat</td>
<td>A. gestern [einen Vogel]</td>
<td>mitgebracht</td>
<td>[der schön singt].</td>
</tr>
<tr>
<td></td>
<td>hat</td>
<td>Alex gestern</td>
<td>mitgebracht</td>
<td>[der schön singt].</td>
</tr>
</tbody>
</table>
Constituent order in the Mittelfeld is relatively free, determined by argument structure, word class, scope, theme/rheme, . . .

**Example**

Gestern hat Alex [einem Kollegen] [einen Witz] erzählt.


“Yesterday Alex told a joke to a colleague”

Gestern hat Alex [einen Witz] [einem Kollegen] erzählt.

**Verb Second**

- FINIT needs to be filled in root clauses (V1 or V2), “Verb Second”
- Verb second is clause-bound
- With few exceptions: all finite verbs can occur in FINIT and in the verbal complex
- No semantic or pragmatic constraints on the verb
VORFELD PLACEMENT

- Needs to be filled in V2 clauses.
- Vorfeld placement is considered an unbounded dependency (Müller, 2007)
- Vorfeld constituent not necessarily meaningful (Vorfeld-es)

**Example**

*Gestern hat jemand angerufen.*
*yesterday has someone called*

“Someone called yesterday.”

*Es hat gestern jemand angerufen.*
*it has yesterday someone called*

“Someone called yesterday.”
**TOPOLOGICAL FIELDS**

Nachfeld

- Syntactically complex, prosodically heavy constituent
- Preferred for relative clauses, almost obligatory for complement clauses
- Nachfeld placement ("extraposition") is clause-bounded.

**Example**

Alex hat [einen Vogel ([der singt])] mitgebracht ([der ...]).

"Alex brought along a bird that sings well."

Alex hat mir *[dass Chris schnarcht] gesagt [dass ...].

"Alex told me that Chris snored."
GENERAL PROPERTIES OF GERMAN SYNTAX

Passive

- English-like passive:
  
  Alex las das Buch. Das Buch wurde (von A.) gelesen.
  
  Alex read the book the book was by A. read
  
  “Alex read the book.”

- But also: impersonal passive

  Hier wird schwer gearbeitet.
  
  here is heavily worked
  
  “People work hard here.”

  Gestorben wird immer
  
  died is always
  
  “There is always someone dying.”

- German passive “demotes” the active subject.
SUMMARY ON GERMAN SYNTAX

- Certain positions (Vorfeld, FINIT) need to be filled, but there is flexibility on the choice of the filler.
- Certain positions (Nachfeld) impose complexity constraint on filling, no semantic/pragmatic constraint.
- Passive
### Example

<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ins Gras</td>
<td>beißen in. the gras bite</td>
<td>“die”</td>
</tr>
<tr>
<td>den Löffel</td>
<td>abgeben the spoon pass.on</td>
<td>“die”</td>
</tr>
<tr>
<td>den Geist</td>
<td>aufgeben the ghost give.up</td>
<td>“die [primarily for machines]”</td>
</tr>
</tbody>
</table>
NON-DECOMPOSABLE IDIOMS IN GERMAN

Example (No case of internal modification)

*und ihr beißt ins virtuelle Gras*
*and you bite in the virtual gras*
“and you die in the virtual world.”

*Alex gab den *friedvollen Löffel ab*
*Alex passed the peaceful spoon on*
“≠ Alex lost his peaceful life.”

*Das Auto gab den elektrischen Geist auf*
*the car gave the reliable ghost up*
“As far as electricity is concerned, the car broke.”
MOVING PARTS OF NON-DECOMPOSABLE IDIOMS

**Example**

- **Scrambling**
  
  *Alex hatte gerade den Löffel abgegeben, als . . .*
  
  *Alex had just the spoon passed on when*
  
  “Alex had just died when . . .”

  *Alex hatte den Löffel gerade abgegeben, als . . .*

- **Verb second**
  
  *Alex gab den Löffel ab.*
  
  *Alex passed the spoon on*
  
  “Alex died.”

- **Vorfeld placement**
  
  *Den Löffel hat Alex abgegeben.*
  
  *the spoon has Alex on*
  
  “Alex died.”
SYNTACTIC FLEXIBILITY IN NON-DECOMPOSABLE IDIOMS

- Pointed out by (Webelhuth and Ackerman, 1994).
- Discussed in (Nunberg et al., 1994). Analytic options:
  - Word order variation just linearization
  - Lexically encoded non-compositional idioms
No linearization option: Passive

Bei den Grünen wird der politische Löffel schon vor Amtsabschied abgegeben.

“In the Green party, people die politically already before resigning passed.on

No linearization-account possible for German passive
(morphological changes)

Only option: Lexically encoded non-compositional idioms

Consequently: No idiom-based argument for linearization analysis
of scrambling, verb second, Vorfeld placement
**Incompatible with (Nunberg et al., 1994)?**

“We predict that the syntactic flexibility of a particular idiom will ultimately be explained in terms of the compatibility of its semantics with the semantics and pragmatics of various constructions.” (Nunberg et al., 1994, p. 531)

- Scrambling is quite free.
- Verb second is obligatory in German main clauses, no influence on semantics/pragmatics
- Vorfeld placement is obligatory in German main clauses, few restrictions on semantics/pragmatics
- German passive demotes the active subject rather than promotes the object.
- (Nunberg et al., 1994): No construction-specific requirements $\Rightarrow$ flexibility to be expected
Non-decomposable idioms in French

- Abeillé 1995, based on Gross 1989, Gaatone 1993
- Passive: no impersonal passive in French, but passive for verbs with human, agentive subject in the active (Gaatone, 1993)
- Idiom without internal modification but with passive:

**Example** (*mettre un bémol* (lit. put a flat, ‘speak softer’))

*L’opposition démocrate n’a guère été convaincu par les bémols mis à cette decision.*

‘The democratic opposition was hardly convinced by the fact that this decision was taken silently.’
Pronominalization with non-decomposable idioms

Luc a cassé sa pipe et tu te casseras la tienne un jour aussi.

Luc has broken his pipe and you will break yours one day as well.

“Luc kicked the bucket and so will you some day.”

Abeillé 1995: Pronominalization possible, but pronoun is not referential.
MUISCHNEK AND KAALEP 2010: ESTONIAN

- Relatively free word order → non-continuous occurrences of all idioms.
- Impersonal passive → possible for all idioms
- Case alternation on direct object for aspectual marking → also possible for non-decomposable idioms

**Example**

(Muischnek and Kaalep, 2010, p. 129)

- *otsi andma* (end.PL.PART give, ‘die’)
- *andis otsad* (gave end-PL.NOM, ‘die’)

*(Muischnek and Kaalep, 2010, p. 129)*

- *otsi andma* (end.PL.PART give, ‘die’)
- *andis otsad* (gave end-PL.NOM, ‘die’)*
**Summar: Other languages**

- Syntactic flexibility for non-decomposable idioms confirmed.
- Still: syntactic and semantic restrictions.
- Complication: We need understanding of the language to explain syntactic flexibility of MWEs.
LEXICAL APPROACH TO *kick the bucket*

- External modification interpretation (Ernst, 1981)
  - *kick the proverbial/fucking bucket*
  - *Bugsy kicked the social bucket (when s/he committed that faux pas at the party).* (McClure, 2011)
    
    = Socially, Bugsy kicked the bucket . . .

- Analytic options:

  - Rather complicated phrasal lexical entry—with potentially open slots (Sailer, 2004; Richter and Sailer, 2009)
  - Or: evidence for lexical encoding of non-compositional idioms and dissociation of syntactic flexibility and semantic decompositionality (Kay and Sag, ms.)

- Non-passivizability?
NON-PASSIVIZABILITY OF *kick the bucket*

- No impersonal passives in English
- Kay and Sag ms.: idiomatic *kick* assigned a non-passivizable verb class.
- Bargmann and Sailer (in prep.):
  - English: Passive subjects must be topical (Kuno and Takami, 2004)
  - *the bucket* makes a redundant semantic contribution and, therefore, does not satisfy the constraints on passive subjects in English.

**Example**

*When you are dead, you don’t have to worry about death anymore.*

*... The bucket will be kicked.*

# Summary: English Idioms

<table>
<thead>
<tr>
<th>Phase 1: syn. fixedness</th>
<th>= sem. fixedness</th>
<th>= fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>kingdom come</td>
<td></td>
</tr>
<tr>
<td></td>
<td>kick the bucket</td>
<td></td>
</tr>
<tr>
<td></td>
<td>spill the beans</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase 2: synt. fixedness</th>
<th>= sem. fixedness</th>
<th>= fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>kingdom come</td>
<td></td>
</tr>
<tr>
<td></td>
<td>kick the bucket</td>
<td></td>
</tr>
<tr>
<td></td>
<td>spill the beans</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase 3: synt. fixedness</th>
<th>= sem. fixedness</th>
<th>≠ fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>kingdom come</td>
<td></td>
</tr>
<tr>
<td></td>
<td>kick the bucket</td>
<td></td>
</tr>
<tr>
<td></td>
<td>spill the beans</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>synt. fixedness</th>
<th>≠ sem. fixedness</th>
<th>= fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>spill the beans</td>
<td></td>
</tr>
</tbody>
</table>
CHALLENGES FOR A LEXICAL ENCODING

1. How can we represent structurally idiosyncratic idioms?
2. What is the semantic contribution of the idiom parts?
3. How are the parts of an idiom constraint to co-occur?
4. How is the syntactic flexibility correctly restricted?
5. How to avoid massive lexical ambiguity?

Two closely related theories:

1. Structurally Idiosyncratic Idioms?

- Only morphological variation:
  
  *She is tripping/tripped the light fantastic*

- Syntactic structure unclear, syntactically fixed Phrasal lexical entry/idiosyncratic construction
2. **Semantics of Idiom Parts?**

- **Decomposable idioms**: idiom-specific semantics
- **Syntactically idiosyncratic, non-decomposable idioms**:
  - K&S: empty semantics
  - R2DT: semantics can be overwritten by the phrasal lexical entry
- **Syntactically flexible, non-decomposable idioms**:
  - K&S: idiom predictor carries all the semantics, other elements have empty semantics
  - R2DT: semantics can be present at each word; words don’t make “unique/independent” semantic contribution
3. Co-occurrence restrictions on idiom parts?

- K&S: mildly extended selection mechanism
- R2DT: collocation mechanism
Follows from general properties of the constructions in individual languages
5. Avoid lexical ambiguity?

- Syntactic flexibility motivates massive lexical ambiguity
- We lose the intuition that MWEs form a unit
- Vision:
  - Lexical representation: syntactic tree with semantic annotation at individual nodes (if decompositional)
  - Lexicon-to-grammar processing generates
  - Idiomatic lexical entries (with “knowledge about the overall MWE through their collocational specification)
Linguistics representation of MWEs

- Syntactically irregular MWEs: phrasal lexical entry
- Syntactically regular MWEs, non-decomposable: syntactically free combinations, but semantically (partially) redundant contribution
- Syntactically regular, decomposable MWEs: syntactically free combinations, semantically non-redundant contribution
GENERAL VISION

- Importance to look beyond English
  Expect more findings from considering languages with more morphology than English/German and other word order processes
- Linguistics: Development from a monolithic, phrasal treatment of all MWEs to a more and more decompositional, word-level treatment
- From lexicon representation to input for grammar
  - Represent MWEs as unit at some level, much like in construction grammar (with potentially redundant meaning assignment to idiom parts)
  - If syntactically regular $\Rightarrow$ transform into independent, collocationally related lexical entries for HPSG analysis or parser
OUTLINE

1. OVERVIEW: LINGUISTIC PROPERTIES OF MWEs
2. MWEs IN LINGUISTIC THEORY
3. MWEs IN LINGUISTIC THEORY 2: BEYOND ENGLISH
4. CHALLENGES FROM OTHER LANGUAGES (HEBREW)
5. LEXICAL ENCODING OF MWEs
6. APPLICATIONS TO OTHER LANGUAGES
7. SUMMARY
Inflectional morphology is highly productive and consists mostly of suffixes, but sometimes of prefixes or circumfixes

- Nominals (nouns, adjectives and numerals) inflect for number and gender
- In addition, nominals have three phonologically (and orthographically) distinct forms
- Nominals can take pronominal suffixes (possessive pronouns)
- Verbs inflect for number, gender and person and also for a combination of tense and aspect
- Prepositions can combine with pronominal affixes that are interpreted as the object of the preposition
### Nominal Morphology

#### Example (Number and Gender)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Singular</th>
<th>Plural</th>
<th>Gender</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine</td>
<td>Šwmr</td>
<td>Šwmrim</td>
<td>Masculine</td>
<td>Kxwl</td>
<td>Kxwlim</td>
</tr>
<tr>
<td>Feminine</td>
<td>Šwmrt</td>
<td>Šwmrwt</td>
<td>Feminine</td>
<td>Kxwlh</td>
<td>Kxwlwt</td>
</tr>
<tr>
<td></td>
<td>“guard”</td>
<td></td>
<td></td>
<td>“blue”</td>
<td></td>
</tr>
</tbody>
</table>
### Nominal Morphology

#### Example (Nominal Status)

<table>
<thead>
<tr>
<th></th>
<th>Absolute</th>
<th>Definite</th>
<th>Construct</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Masculine Singular</strong></td>
<td>klb</td>
<td>hklb</td>
<td>klb</td>
</tr>
<tr>
<td><strong>Feminine Singular</strong></td>
<td>klbh</td>
<td>hklbh</td>
<td>klbt</td>
</tr>
<tr>
<td><strong>Masculine Plural</strong></td>
<td>klbim</td>
<td>hklbim</td>
<td>klbi</td>
</tr>
<tr>
<td><strong>Feminine Plural</strong></td>
<td>klbwt</td>
<td>hklbwt</td>
<td>klbwt</td>
</tr>
<tr>
<td>&quot;dog&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Masculine Singular</strong></td>
<td>kxwl</td>
<td>hkxwl</td>
<td>kxwl</td>
</tr>
<tr>
<td><strong>Feminine Singular</strong></td>
<td>kxwlt</td>
<td>hkxwlt</td>
<td>kxwli</td>
</tr>
<tr>
<td><strong>Masculine Plural</strong></td>
<td>kxwlim</td>
<td>hkxwlim</td>
<td>kxwli</td>
</tr>
<tr>
<td><strong>Feminine Plural</strong></td>
<td>kxwlwt</td>
<td>hkxwlwt</td>
<td>kxwlwt</td>
</tr>
<tr>
<td>&quot;blue&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Nominal Morphology

### Example (Promonimal Suffixes)

<table>
<thead>
<tr>
<th>Gender</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine Singular</td>
<td>klbi</td>
<td>klbk</td>
<td>klbw</td>
</tr>
<tr>
<td>Feminine Singular</td>
<td>klbi</td>
<td>klbk</td>
<td>klbh</td>
</tr>
<tr>
<td>Masculine Plural</td>
<td>klbnw</td>
<td>klbkm</td>
<td>klbm</td>
</tr>
<tr>
<td>Feminine Plural</td>
<td>klbnw</td>
<td>klbkn</td>
<td>klbn</td>
</tr>
</tbody>
</table>

“dog”
**Example (Verb inflections)**

<table>
<thead>
<tr>
<th>Type</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine Singular</td>
<td>hlkti</td>
<td>hlkt</td>
<td>hlk</td>
</tr>
<tr>
<td>Feminine Singular</td>
<td>hlkti</td>
<td>hlkt</td>
<td>hlkh</td>
</tr>
<tr>
<td>Masculine Plural</td>
<td>hlknw</td>
<td>hlktm</td>
<td>hlkw</td>
</tr>
<tr>
<td>Feminine Plural</td>
<td>hlknw</td>
<td>hlktm</td>
<td>hlkw</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine Singular</td>
<td>alk</td>
<td>tlk</td>
<td>ylk</td>
</tr>
<tr>
<td>Feminine Singular</td>
<td>alk</td>
<td>tlki</td>
<td>tlk</td>
</tr>
<tr>
<td>Masculine Plural</td>
<td>nlk</td>
<td>tlkw</td>
<td>ylkw</td>
</tr>
<tr>
<td>Feminine Plural</td>
<td>nlk</td>
<td>tlknh</td>
<td>ylkw</td>
</tr>
</tbody>
</table>

“walk”
### Example (Promonimal suffixes)

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine Singular</td>
<td>$\text{lidi}$</td>
<td>$\text{lidk}$</td>
<td>$\text{lidw}$</td>
</tr>
<tr>
<td>Feminine Singular</td>
<td>$\text{lidi}$</td>
<td>$\text{lidk}$</td>
<td>$\text{lidh}$</td>
</tr>
<tr>
<td>Masculine Plural</td>
<td>$\text{lidnw}$</td>
<td>$\text{lidkm}$</td>
<td>$\text{lidm}$</td>
</tr>
<tr>
<td>Feminine Plural</td>
<td>$\text{lidnw}$</td>
<td>$\text{lidkn}$</td>
<td>$\text{lidn}$</td>
</tr>
</tbody>
</table>

“near”
Hebrew orthography poses several problems for computational processing

- Most vowels are not explicit
- Many particles, including four of the most frequent prepositions (b “in”, k “as”, l “to” and m “from”), the coordinating conjunction w “and” and some subordinating conjunctions (such as š “that” and kš “when”) attach to the following word
- When a definite nominal is prefixed by one of the prepositions b, k or l, the definite article h is assimilated with the preposition and the resulting form is ambiguous with respect to definiteness
- The rules that govern the combination of Hebrew prefix particles with the words they attach to are syntactic (and, hence, are constrained by the category of the entire MWE)
Possible readings of şbth:

- šbth: “capture”, third person singular feminine past
- šbth: “go on strike”, third person singular feminine past
- š+bth: “that+field”
- š+bt+h: “that+her+daughter”
- š+b+th: “that+in+tea”
- š+b+h+th: “that+in+the+tea”
Linguistic properties of Hebrew MWEs

- Morphological properties
- Syntactic properties
- Semantic properties
Morphological properties

- Fixed form
- Partial inflection
- Non-standard inflection
- Fossil words
FIXED FORM

Constituents can appear in a fixed, frozen form. This form can be the citation form or a frozen inflected form.

**Example (Fixed Form)**

```
kptwr  wprx
button  and-flower
“a button and a flower” ⇒ fantastic

ain    lw    id    bdbfr
there-isn’t  to-him  hand  in-the-thing
“not have a hand in the thing” ⇒ be uninvolved

hxlwnwt    hgbwhim
the-windows  the-high
“The high windows” ⇒ the powers that be
```
**PARTIAL INFLLECTION**

In some cases, constituents undergo a (strict) subset of the full inflections that they would undergo in isolation.

**EXAMPLE (PARTIAL INFLCTION)**

- **hlk axri libw**  
  *walked after heart-his*  
  “followed his heart” ⇒ follow one’s heart

- **hlkw axri libm**  
  *walked after heart-their*  
  “followed their heart” ⇒ follow one’s heart

- **hlkw axri lbbwtihm**  
  *walked after hearts-their*  
  “followed their hearts” ⇒ follow one’s heart
In some cases, constituents undergo a (strict) subset of the full inflections that they would undergo in isolation.

**Example (Partial Inflection)**

- `npl ʿl hraš`
  - fell on the-head
  - “fell on his head” ⇒ lose one’s mind

- `*pwl ʿl hraš`
  - fall on the-head
  - “fall on your head” ⇒ lose your mind
Constituents can also undergo non-standard morphological inflections that they would not undergo in isolation.

**Example (Non-standard inflection)**

bdltiim sgwrwt

*in-doors*-DUAL closed

“in two closed doors” ⇒ behind closed doors
### Example (Non-standard Inflection)

<table>
<thead>
<tr>
<th>idit</th>
<th>dlt</th>
<th>idit</th>
<th>hdlt</th>
</tr>
</thead>
<tbody>
<tr>
<td>handle. CONST</td>
<td>door</td>
<td>handle. CONST</td>
<td>the-door</td>
</tr>
<tr>
<td>“door handle”</td>
<td></td>
<td>“the door handle”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>*hidit</th>
<th>dlt</th>
<th>hiwšb</th>
<th>raš</th>
</tr>
</thead>
<tbody>
<tr>
<td>the-handle. CONST</td>
<td>door</td>
<td>the-sitter. CONST</td>
<td>head</td>
</tr>
<tr>
<td>“the door handle”</td>
<td></td>
<td>“the sitting head”</td>
<td>⇒ the chairman</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>iwšb</th>
<th>raš</th>
<th>hiwšb</th>
<th>raš</th>
</tr>
</thead>
<tbody>
<tr>
<td>sitter. CONST</td>
<td>head</td>
<td>the-sitter. CONST</td>
<td>head</td>
</tr>
<tr>
<td>“sitting head” ⇒ chairman</td>
<td></td>
<td>“the sitting head” ⇒ the chairman</td>
<td></td>
</tr>
</tbody>
</table>
Sometimes, MWE constituents have no other usage or literal meaning outside the expression they appear in.

**Example (Fossil Words)**

- **kmTxwi** kšt
  - *like-??* bow
  - “like a bow’s ???” ⇒ not far away

- **abd 'liw qlx**
  - *lost on-him ???*
  - “???” ⇒ outdated

- **lit man dplig**
  - *there-isn’t who that-disagrees*
  - “no-one disagrees” ⇒ without dispute
SYNTACTIC PROPERTIES

- Open slots
- Word order
- Limited transformations
- Limited paraphrasing
- Limited reference
- Violated agreement
- Irregular agreement
- Syntactic idiosyncrasies
- The syntactic category of MWEs
- Some common constructions
SYNTACTIC PROPERTIES

OPEN SLOTS

EXAMPLE (OPEN SLOTS)

\[ akl \quad at \quad X \quad bli \quad mlx \]
\[ ate \quad ACC \quad without \quad salt \]

“ate someone without salt” \( \Rightarrow \) easily defeat

\[ išb \quad ‘l \quad X \quad šb‘h \]
\[ sat \quad on \quad seven \]

“sat seven (days) on someone” \( \Rightarrow \) mourn
## Syntactic Properties

### Word Order

<table>
<thead>
<tr>
<th>Example (Verb argument structure)</th>
<th>Hebrew</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>hwa ica mbito l’bwdh</code></td>
<td>he went-out from-house-his to-the-work</td>
<td>“he left home for work” ⇒ he left home for work</td>
</tr>
<tr>
<td><code>mbitw hwa ich l’bwdh</code></td>
<td>from-house-his he went-out to-the-work</td>
<td>“he left home for work” ⇒ from home he left for work</td>
</tr>
<tr>
<td><code>hwa ica mhqlim b’bwdh</code></td>
<td>he went-out from-the-tools in-the-work</td>
<td>“he left the tools at work” ⇒ he was furious at work</td>
</tr>
<tr>
<td><em><code>mhqlim hwa ica b’bwdh</code></em></td>
<td>from-the-tools he went-out in-the-work</td>
<td>“he left the tools at work” ⇒ he was furious at work</td>
</tr>
</tbody>
</table>
SYNTACTIC PROPERTIES

WORD ORDER

Example (Adverbial Locations)

Tmn mktb bargz
hid letter in-the-box
“hid a letter in the box”

Tmn bargz mktb
hid in-the-box letter
“hid a letter in the box”

Tmn at rašw bxwl
hid ACC head-his in-the-sand
“buried his head in the sand” ⇒ bury one’s head in sand

*Tmn bxwl at rašw
hid in-the-sand ACC head-his
“buried his head in the sand” ⇒ bury one’s head in sand
SYNTACTIC PROPERTIES
LIMITED TRANSFORMATIONS

**Example (Passivization)**

špk  at  lbw
spilled   ACC  heart-his
“spilled his heart”  ⇒  confess

*lbw  nšpk
heart-his  spilled
“his heart spilled”  ⇒  confess

bnh  mgdlim  bawwir
built  towers  in-the-air
“built towers in the air”  ⇒  build castles in the air

*mgdlim  nbnw  bawwir
towers  were-built  in-the-air
“towers were built in the air”  ⇒  build castles in the air
SYNTACTIC PROPERTIES

LIMITED TRANSFORMATIONS

Example (Coordination)

\[ ica \ b\šn \ w\‘in \]
\[ \text{went-out} \ in-tooth \ and-eye \]
“went out in tooth and eye” $\Rightarrow$ be injured

\[ \ast ica \ b\‘in \ w\šn \]
\[ \text{went-out} \ in-eye \ and-tooth \]
“went out in tooth and eye” $\Rightarrow$ be injured

\[ pxwt \ aw \ iwtr \]
\[ \text{less or more} \]
“less or more” $\Rightarrow$ more or less

\[ \ast iwtr \ aw \ pxwt \]
\[ \text{more or less} \]
“more or less” $\Rightarrow$ more or less
**SYNTACTIC PROPERTIES**

**LIMITED PARAPHRASING**

**Example (Genitive Constructions)**

<table>
<thead>
<tr>
<th>‘wrk</th>
<th>h‘itwn</th>
</tr>
</thead>
<tbody>
<tr>
<td>editor- .CONST</td>
<td>the-journal</td>
</tr>
<tr>
<td>“the journal editor” ⇒ the journal editor</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>h‘wrk</th>
<th>šl</th>
<th>h‘itwn</th>
</tr>
</thead>
<tbody>
<tr>
<td>the-editor of the-journal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>“the editor of the journal” ⇒ the journal editor</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>‘wrk</th>
<th>hdin</th>
</tr>
</thead>
<tbody>
<tr>
<td>editor- .CONST</td>
<td>the-law</td>
</tr>
<tr>
<td>“the law editor” ⇒ the lawyer</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>*h‘wrk</th>
<th>šl</th>
<th>hdin</th>
</tr>
</thead>
<tbody>
<tr>
<td>the-editor of the-law</td>
<td></td>
<td></td>
</tr>
<tr>
<td>“the editor of the law” ⇒ the lawyer</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SYNTACTIC PROPERTIES

LIMITED PARAPHRASING

EXAMPLE (PRONOMINAL SUFFIXES)

\[ \text{ica mbitw} \quad \text{went-out from-house-his} \]
“went out from his house” \(\Rightarrow\) left home

\[ \text{ica mhbit šlw} \quad \text{went-out from-the-house his} \]
“went out from his house” \(\Rightarrow\) left home

\[ \text{ica md′tw} \quad \text{went-out from-mind-his} \]
“went out from his mind” \(\Rightarrow\) lost his mind

\[ *\text{ica mhd′t šlw} \quad \text{went-out from-the-mind his} \]
“went out from his mind” \(\Rightarrow\) lost his mind

SYNTACTIC PROPERTIES

LIMITED REFERENCE

EXAMPLE (LIMITED REFERENCE)

$qiblti$ $awr$ $irwq$

$l$-received light green

“I received green light” ⇒ they gave me the green light

$hawr$ $\tilde{q}iblti$ ...

the-light that-$l$-received ...

“the light I received...” ⇒ the light they gave me...
**Syntactic properties**

**Limited reference**

---

**Example (Category change)**

\[ hxlwnwt \quad hgbwhim \]
\[ the-windows \quad the-high \]

“The high windows” \( \Rightarrow \) the powers that be

\[ hgwbb \quad šl \quad hxlwnwt \]
\[ the-height \quad of \quad the-windows \]

“The height of the windows” \( \Rightarrow \) ???
SYNTACTIC PROPERTIES

VIOLATED AGREEMENT

EXAMPLE (VIOLATED AGREEMENT)

'in'  
hr'

eye-INDEF  evil-DEF

“evil eye”  ⇒  evil eye
SYNTACTIC PROPERTIES

IRREGULAR AGREEMENT

Example (Irregular Agreement)

'md  'l d’tw
stand-3.m.sg on mind-3.m.sg
“stand on his mind” ⇒ insist
SYNTACTIC PROPERTIES
SYNTACTIC IDIOSYNCRASIES

EXAMPLE (SYNTACTIC IDIOSYNCRASIES)

*bxwr*  *wTwb*

*young-man*  *and-good*

“???” ⇒ an outstanding young man

*‘šh  xwšbim*

*did  think*

“???” ⇒ hold on and think
THE SYNTACTIC CATEGORY OF MWEs

**Example (Noun)**

<table>
<thead>
<tr>
<th>bit</th>
<th>spr</th>
</tr>
</thead>
<tbody>
<tr>
<td>house-</td>
<td>book</td>
</tr>
</tbody>
</table>

“house of book” ⇒ school

<table>
<thead>
<tr>
<th>sprwt</th>
<th>iph</th>
</tr>
</thead>
<tbody>
<tr>
<td>literature</td>
<td>pretty</td>
</tr>
</tbody>
</table>

“beautiful literature” ⇒ belles-lettres

<table>
<thead>
<tr>
<th>ab</th>
<th>bit</th>
<th>din</th>
</tr>
</thead>
<tbody>
<tr>
<td>father-</td>
<td>house-</td>
<td>law</td>
</tr>
</tbody>
</table>

“father of house of law” ⇒ President of the Court

**Example (Noun Phrase)**

<table>
<thead>
<tr>
<th>hawmwt</th>
<th>hmauxdwt</th>
</tr>
</thead>
</table>

Manfred Sailer and Shuly Wintner ()   Ling. properties and lexical representation   Prague, January 2015   134 / 190
Common MWE constructions

- Noun-noun constructs
- Adjective-noun constructs
- Support verbs
- Proper names
## Common MWE Constructions

### Example (Noun-noun constructs)

<table>
<thead>
<tr>
<th>bit</th>
<th>spr</th>
</tr>
</thead>
<tbody>
<tr>
<td>house-.CONST</td>
<td>book</td>
</tr>
</tbody>
</table>

“house of book” ⇒ school

<table>
<thead>
<tr>
<th>ab</th>
<th>bit</th>
<th>din</th>
</tr>
</thead>
<tbody>
<tr>
<td>father-.CONST</td>
<td>house-.CONST</td>
<td>law</td>
</tr>
</tbody>
</table>

“father of house of law” ⇒ President of the Court
COMMON MWE CONSTRUCTIONS

**Example (Adjective-Noun Constructs)**

- ql  
  light-.CONST mind  
  “light minded” ⇒ hasty

- kxwl’t  
  blue-.CONST eyes  
  “blue eyed” ⇒ blue eyed
COMMON MWE CONSTRUCTIONS

EXAMPLE (SUPPORT VERBS)

- qibl  hxlTh
  receive  decision
  “received a decision” ⇒ decide

- ntn  sTirh
  gave  slap
  “gave a slap” ⇒ slap

- ‘šh  mqlxt
  made  shower
  “made a shower” ⇒ take a shower
**Semantic properties**

**Semantic compositionality**  “MWEs do not fall cleanly into the binary classes of compositional and non-compositional expressions, but populate a continuum between the two extremes” (Bannard et al., 2003)

**Lexical fixedness** An expression is lexically fixed if replacing any of its constituents by a semantically (and syntactically) similar word results in an invalid or literal meaning

**Translation equivalents** MWEs can be translated to *single* words in other languages (or paraphrased as single words in the same language)
**Semantic Compositionality**

**Example (Opaque)**

\[
\begin{align*}
\text{ap} & \quad \text{‘l} & \quad \text{pi} & \quad \ddot{s} \\
\text{even} & \quad \text{on} & \quad \text{mouth}-.\text{CONST} & \quad \text{that} \\
\text{“???”} & \quad \Rightarrow & \quad \text{although} \\
\text{ap} & \quad \text{‘l} & \quad \text{pi} & \quad \text{kn} \\
\text{even} & \quad \text{on} & \quad \text{mouth}-.\text{CONST} & \quad \text{thus} \\
\text{“???”} & \quad \Rightarrow & \quad \text{nevertheless} \\
\text{iwca} & \quad \text{dwpn} \\
\text{go-out} & \quad \text{side, bank} \\
\text{“leaving through the membrane”} & \quad \Rightarrow & \quad \text{exceptional} \\
\text{lxm} & \quad \text{xwq} \\
\text{bread}-.\text{CONST} & \quad \text{law} \\
\text{“bread of law”} & \quad \Rightarrow & \quad \text{routine}
\end{align*}
\]
Lexical fixedness

Example

\begin{align*}
\text{akl at hkw} & \text{b'} \\
\text{eat ACC the-hat} \\
\text{“eat the hat” } \Rightarrow \text{ eat one’s hat (regret)}
\end{align*}

\begin{align*}
*\text{Trp at hkw} & \text{b'} \\
\text{devour ACC the-hat} \\
\text{“devour the hat” } \Rightarrow \text{ ???}
\end{align*}

\begin{align*}
*\text{akl at hmgb’t} \\
\text{eat ACC the-bowler} \\
\text{“eat the bowler” } \Rightarrow \text{ ???}
\end{align*}
Lexical fixedness

Example

\[ \text{šwlxn} \quad \text{‘bwdh} \]
\[ \text{table-.CONST} \quad \text{work} \]

“working table” ⇒ desk

\[ *\text{kisa} \quad \text{‘bwdh} \]
\[ \text{chair-.CONST} \quad \text{work} \]

“working chair” ⇒ ???

\[ *\text{šwlxn} \quad \text{mlakh} \]
\[ \text{table-.CONST} \quad \text{work} \]

“working table” ⇒ ???
## Translational Equivalents

### Example

<table>
<thead>
<tr>
<th>$bit$</th>
<th>$spr$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$house$.CONST</td>
<td>$book$</td>
</tr>
<tr>
<td>&quot;house of book&quot;</td>
<td>⇒ school</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$ap$</th>
<th>$'l$</th>
<th>$pi$</th>
<th>$kn$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$even$</td>
<td>$on$</td>
<td>$mouth$.CONST</td>
<td>$thus$</td>
</tr>
<tr>
<td>&quot;???”</td>
<td>⇒ nevertheless</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$b'l$</th>
<th>$\ddash'wr$</th>
<th>$qwmh$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$owner$.CONST</td>
<td>$measure$.CONST</td>
<td>$height$</td>
</tr>
<tr>
<td>&quot;owner of a measure of height”</td>
<td>⇒ honorable</td>
<td></td>
</tr>
</tbody>
</table>
**Properties of MWEs: Summary**

<table>
<thead>
<tr>
<th>Morph.</th>
<th>‘wrk din “lawyer”</th>
<th>‘wrk ‘itwn “journal editor”</th>
<th>mkwnt qphca mhxa “coffee machine”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed form</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Partial inflection</td>
<td>+</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Non-standard inflection</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Fossil words</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Syntax</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Word order:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– Rigid argument structure</td>
<td>−</td>
<td></td>
<td>−</td>
</tr>
<tr>
<td>– No passivization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– Adverbial locations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– No coordination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limited paraphrasing</td>
<td>+</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Limited reference</td>
<td>+</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Violated agreement</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Syntactic idiosyncrasies</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sem.</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Compositionality</td>
<td>Semi</td>
<td>Transparent</td>
<td>Transparent</td>
</tr>
<tr>
<td>Lexical fixedness</td>
<td>+</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Translation equivalents</td>
<td>+</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>
OUTLINE

1. OVERVIEW: LINGUISTIC PROPERTIES OF MWEs

2. MWEs IN LINGUISTIC THEORY

3. MWEs IN LINGUISTIC THEORY 2: BEYOND ENGLISH

4. CHALLENGES FROM OTHER LANGUAGES (HEBREW)

5. LEXICAL ENCODING OF MWEs

6. APPLICATIONS TO OTHER LANGUAGES

7. SUMMARY
**Goals**

- Develop an architecture for lexical specification of MWEs in Hebrew, and extend an existing lexicon of the language with capabilities to store them.
- Develop techniques for morphological processing of MWEs in Hebrew, and extend an existing morphological processor (anaylzer/generator) with capabilities to process them.
- Develop techniques to extract MWEs from monolingual and bilingual corpora, and populate the lexicon with MWEs acquired automatically.
Welcome to MILA!

MILA develops, maintains, and distributes open-source resources and tools for computational processing of Hebrew.

The Hebrew language poses special challenges to developers of natural language processing systems, due to its deficient orthography and rich morphology. A solid software infrastructure based on linguistic knowledge is required for natural language applications such as automatic translation, speech-to-text conversion, automatic document summarization, spelling and style checking, and many more.

MILA develops, maintains, and distributes the resources and tools needed for Hebrew natural language processing: corpora of text and speech documenting how people use the language; lexicons (dictionaries); standards for data representation; tools to tokenize, morphologically analyze, and morphologically disambiguate text; and more.

All of MILA’s tools and resources are freely available to the non-commercial research community; see the License section for more details. Commercial entities are invited to contact MILA to inquire about terms.
RESULTS

AL-HAJ ET AL. (2014)

- An architecture for lexical specification of MWEs in morphologically-complex languages
- An implementation for (Modern) Hebrew
  - A solution for storing MWEs in an existing large-scale lexicon (Itai et al., 2006)
  - A protocol for integrating MWEs in an existing morphological processing system (Itai and Wintner, 2008)
- A survey of the variety and diversity of Hebrew MWEs within a computational setup
Hebrew Orthography

- Most vowels are not explicit
- Many particles (the definite article ֶה “the”, the frequent prepositions ֶב “in”, ֶכ “as”, ֶל “to” and ֶמ “from”, the coordinating conjunction ֶו “and” and the subordinating conjunctions ֶש “that” and ֶכש “when”) attach to the word which immediately follows them

Example (The preposition ֶמ “from”)

- Combines with nouns but not with adverbs
- The same rules govern the combination of Hebrew prefix particles with MWEs, but these combinations are constrained by the syntactic category of the whole expression, rather than its first word
  - ֶמ+ֶפ “from mouth”
  - ֶמ+ֶפ ֶל+ֶאצֶד “from mouth to ear” ⇒ “through the grapevine”
  - *ֶמ “from” + ֶפ ֶאצ “mouth one” ⇒ “unanimously”
Lexical Encoding of MWEs

Hebrew Morphology

Nominals (nouns, adjectives and numerals) inflect for number and gender

Nominals have three distinct states: the absolute (citation) state; the definite state, which is indicated by the prefix \( h \) “the”; and the construct state, which is typically used in genitive (possessive) constructions

Nominals (in the construct state) take pronominal suffixes, sometimes referred to as clitics, which are interpreted as possessives

**Example**

* sirh “boat”; \( h+sirh \) “the+boat”; *sirt+i “my boat”; *sirt mnwy “boat-of engine” ⇒ “speedboat”*
Hebrew Morphology

Verbs inflect for number, gender and person, and also for a combination of tense/aspect and mood.

A single verb lemma can yield dozens of inflected forms.

Prepositions can combine with pronominal affixes that are interpreted as the objects of the preposition, and inflect for number, gender, and person.

Example:

lid “near”; lid+i “near me”
The standard constituent order of Hebrew is Subject–Verb–Object, although many other orders are possible, and some are highly frequent.

Within the noun phrase, constituents tend to occur in a fixed order.

Various elements of a noun phrase may be marked as definite; all elements of the noun phrase must agree with respect to definiteness.

**Example**

```

h+ sirh h+ iph h+ zaw  “the+ boat the+ nice the+ this”  ⇒  “this nice boat”
```

Hebrew has three different possessive constructions.
Lexical Encoding of MWEs

The Hebrew morphological processor

THE HEBREW MORPHOLOGICAL PROCESSOR
(WINTNER AND YONA, 2003; ITAI AND WINTNER, 2008)

Raw text input

Tokenizer

Morphological analyzer

XML wrapper

Output XML document

Lexicon

Morphological generator

Inflected words

Tokenizer

Morphological analyzer

XML wrapper

Output XML document

Lexicon

Morphological generator

Inflected words
**Example (šbi)**

```xml
<sentence id="1">
  <token id="1" surface="šbi">
    <analysis id="1">
      <base lexiconPointer="1541" transliterated="šbh">
        <verb gender="feminine" number="singular" person="2" tense="imperative"/>
      </base>
    </analysis>
    <analysis id="2">
      <base lexiconPointer="1636" transliterated="išb">
        <verb gender="feminine" number="singular" person="2" tense="imperative"/>
      </base>
    </analysis>
    ...  
    <analysis id="5">
      <base lexiconPointer="7863" transliterated="šbi">
        <noun definiteness="false" gender="masculine" number="singular" state="absolute"/>
      </base>
    </analysis>
    <analysis id="6">
      <prefix function="relativizer/subordinatingConjunction" id="1" surface="š" />
      <base lexiconPointer="26553" transliterated="b">
        <preposition />
      </base>
      <suffix function="pronominal" gender="masculine and feminine" number="singular" person="1"/>
    </analysis>
  </token>
</sentence>
```
The Hebrew morphological processor operates on a token-by-token basis. Tokens are acquired from the tokenizer which uses only blanks and punctuation for segmentation. The tokenizer is completely independent of the lexicon. The lexicon includes single-word tokens only, and the morphological analyzer is completely unaware of MWEs.
Each MWE is represented as an item in the lexicon, which encodes its morphological and syntactic properties.

A MWE lexical entry includes an element that specifies that the item is a MWE, followed by its POS.

**Example** *(niw iwrq “NEW YORK”)*

```xml
<item id="28498" transliterated="niw iwrq">
  <MWE pos="properName" type="city" gender="feminine"/>
</item>
```
**Word-level properties**

**Frozen form** Constituents can appear in one fixed form, disallowing all inflections. This form can be the citation form:

**Example**

\[ ain \ lw \ id \ bdbr \ “does \ not \ have \ a \ hand \ in \ the \ thing” \Rightarrow “is \ uninvolved” \]
\[ kptwr \ wprx \ “button \ and \ flower” \Rightarrow “fantastic” \]

It can also be some inflected form:

**Example**

\[ hxlwnwt \ hgbwhim \ “the+windows \ the+high” \Rightarrow “upper \ echelon” \]
**Partial Inflection** In some cases, constituents undergo a (strict) subset of the full set of inflections that they would undergo in isolation.

**Example**

- `hlk axri lbw` "walk after his+heart" ⇒ "follow one’s heart"
- `hlkw axri lbm` "they followed their heart"
- *`hlkw axri lbbwtihm` "they followed their hearts"

**Example**

- `bit xwlim` "house-of sick-people" ⇒ "hospital"
- *`bit xwik` "house-of sick-people+your"
Representing MWE Constituents

- Each MWE constituent is realized as an *atom*
- Atoms represent morphemes, rather than words
- To support partial inflections (including frozen forms):
  - **ATOM** Defines a constituent along with all its possible inflected forms. Atoms have the following optional sub-elements:
  - **PREFIX** Specifies that the constituent is a prefix that is an inherent part of the MWE
  - **INFLECT** Restricts the possible inflections of the constituent to those specified
  - **SUFFIX** Specifies that the constituent is a pronominal suffix that attaches to the previous atom
EXAMPLE (mcd šni “FROM SIDE SECOND” ⇒ “ON THE OTHER HAND”)

<item id="29000 transliterated="mcd šni">
  <MWE pos="adverb"/>
  <atom id="1" lexiconPointer="10418"> <!-- m -->
    <prefix/>
  </atom>
  <atom id="2" lexiconPointer="20473"> <!-- cd -->
    <inflect state="absolute" definiteness="false" number="singular"/>
  </atom>
  <atom id="3" lexiconPointer="3561"> <!-- šni -->
    <inflect state="absolute" definiteness="false" number="singular"
      gender="masculine"/>
  </atom>
</item>
PARTIAL MORPHOLOGICAL INFLECTIONS

Example (The lexical entry of ywrk din “lawyer”)

```xml
<item id="28579" transliterated="ywrk din" hprefix="true">
  <MWE pos="noun"/>
  <atom id="1" lexiconPointer="8174">
    <inflect state="construct"/>
  </atom>
  <atom id="2" lexiconPointer="5208">
    <inflect number="singular"/>
  </atom>
  <atom id="3" lexiconPointer="0">
    <suffix/>
  </atom>
</item>
```

\[\text{Pronominal suffix}\]
**WORD-LEVEL PROPERTIES**

**FOSSIL WORDS** Constituents that *only* occur in MWEs

**EXAMPLE**

- \( kmTxwwi \) *qšt* “a stone’s throw”
- \( abd \) *yliw* \( hklx \) “outdated”
- \( lit \) *man* \( dplig \) “without dispute”
FOSSIL WORDS

EXAMPLE (kmTxwwi qšt “STONE’S THROW”)

```xml
<item id="27000" transliterated="kmTxwwi">
    <fossil/>
</item>

<item id="23999 transliterated="kmTxwwi qšt">
    <MWE pos="adverb"/>
    <atom id="1" lexiconPointer="27000"> <!-- kmTxwwi -->
    </atom>
    <atom id="2" lexiconPointer="3507"> <!-- qšt -->
        <inflect definiteness="false" state="absolute"
            number="singular"/>
    </atom>
</item>
```
**Morpho-syntactic properties**

Retrieving morphological features Often, MWEs inherit some of their morphological features from those of their constituents

<table>
<thead>
<tr>
<th>Example</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ywrk din</td>
<td>“editor-of law” ⇒ “a lawyer”</td>
<td></td>
</tr>
<tr>
<td>ywrkt din</td>
<td>“editor-of law” ⇒ “a female lawyer”</td>
<td></td>
</tr>
<tr>
<td>ywrki din</td>
<td>“editors-of law” ⇒ “lawyers”</td>
<td></td>
</tr>
<tr>
<td>ywrk hdin</td>
<td>“editor-of the+law” ⇒ “the lawyer”</td>
<td></td>
</tr>
<tr>
<td>hywrk din</td>
<td>“the+editor-of law” ⇒ “the lawyer”</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lexical Encoding of MWEs

Lexical representation of Hebrew MWEs

Retrieving Morphological Features

Example (*ywrk din* “LAWYER”)

```xml
<item id="28579" transliterated="ywrk din" hprefix="optional">
   <MWE pos="noun">
      <definiteness="2" state="2" number="1" gender="1"/>
   </MWE>
   <atom id="1" lexiconPointer="8174">
      <inflect state="construct"/>
   </atom>
   <atom id="2" lexiconPointer="5208">
      <inflect number="singular"/>
   </atom>
   <atom id="3" lexiconPointer="0">
      <suffix/>
   </atom>
</item>
```

Manfred Sailer and Shuly Wintner ()
Ling. properties and lexical representation
Prague, January 2015 165 / 190
**Morpho-syntactic properties**

**Agreement among constituents**  Some MWEs require agreement between the morphological features of some of their constituents

<table>
<thead>
<tr>
<th>Example</th>
<th>milh nrdpt  “word chased” ⇒ “synonym”</th>
</tr>
</thead>
</table>

**Violated agreement**  In some MWEs, constituents that generally agree in some morphological features violate the agreement constraints

<table>
<thead>
<tr>
<th>Example</th>
<th>yin hry  “eye the+evil” ⇒ “evil eye”</th>
</tr>
</thead>
</table>
AGREEMENT AMONG CONSTITUENTS

EXAMPLE (milh nrdpt “WORD CHASED” ⇒ “SYNONYM”)

```xml
=item id="39991" transliterated="milh nrdpt"
  <MWE pos="noun" state="absolute" gender="feminine"
   definiteness="1" number="1"/>
  <atom id="1" lexiconPointer="3265">
    <!-- milh -->
    <inflect state="absolute"/>
  </atom>
  <atom id="2" lexiconPointer="10097">
    <!-- nrdpt -->
    <inflect tense="participle" type="adjective"
      state="absolute" gender="feminine"
      definiteness="1" number="1"/>
  </atom>
</item>
```
ACCOUNTING FOR SYNTACTIC FLEXIBILITY

**Compositionality** Some MWEs contain *open slots*, which can be filled by a variety of complements.

**Example**

\[
išb\ yl\ X\ šbyh\ "sit\ on\ X\ seven\ (days)"\ \Rightarrow\ "mourn"
\]

\[
išb\ yliw\ šbyh\ "sat\ on-him\ seven"\ \Rightarrow\ "mourn\ him"
\]

**Constituent order** The order of the constituents in most MWEs tends to be fixed, but some, especially verb phrases, still exhibit some flexibility.

**Example**

\[
išb\ šbyh\ yl\ abiw\ "sat\ seven\ on\ his-father"
\]
ACCOUNTING FOR SYNTACTIC FLEXIBILITY

- We add a set of attributes and elements in order to account for syntactic variability.
- By default, all the constituents must appear consecutively in the order determined by the *atoms*.
- If other orders are possible, all the allowed permutations are prescribed within *perm* items.
ACCOUNTING FOR SYNTACTIC FLEXIBILITY

EXAMPLE (יֶשׁ יִמְמֶיָּלּוֹת "MADE DAYS LIKE-NIGHTS" ⇒
"WORK INTENSIVELY")

\[
\text{<item id="39991" transliterated="יֶשׁ יִמְמֶיָּלּוֹת">}
\text{<MWE pos="VP" tense="1" person="1" number="1" gender="1"/>}
\text{<atom id="1" lexiconPointer="376"><inflect/> <!-- yֶשׁ -->}
\text{<atom id="2" lexiconPointer="9475"> <!-- iַמְמֶיָּלּוֹת -->}
\text{<inflect state="absolute" definiteness="false" number="plural"/>}
\text{</atom>}
\text{<atom id="3" lexiconPointer="20001"><prefix/> <!-- k -->}
\text{<atom id="4" lexiconPointer="8024"> <!-- lַיּוֹת -->}
\text{<inflect state="absolute" definiteness="false" number="plural"/>}
\text{</atom>}
\text{<perms>}
\text{<perm id="1" order="1 2 3 4"/>}
\text{<perm id="2" order="1 4 3 2"/>}
\text{</perms>}
\text{</item>}
\]
OPEN SLOTS

EXAMPLE ("akl at X bli mlx" "EAT X WITHOUT SALT" ⇒ "DEFEAT")

```xml
<item id="23986" transliterated="akl at + bli mlx">
  <MWE pos="VP" person="1" number="1" gender="1" tense="1"/>
  <atom id="1" lexiconPointer="8442">
    <inflect/>
  </atom>
  <atom id="2" lexiconPointer="3382"/>
  <atom id="3" lexiconPointer="0">
    <suffix/>
  </atom>
  <atom id="4" lexiconPointer="21542"/>
  <atom id="5" lexiconPointer="608"/>
</item>
```

<!-- akl -->
<!-- at -->
<!-- pronominal suffix -->
<!-- bli -->
<!-- mlx -->
EXAMPLE (akl at X bli mlx “EAT X WITHOUT SALT” ⇒ “DEFEAT”)

<perms>
  <perm id="1" order="1 2 3 4 5"/> <!-- akl awtw bli mlx -->
  <perm id="2" order="2 3 1 4 5"/> <!-- awtw akl bli mlx -->
  <perm id="3" order="2 3 4 5 1"/> <!-- awtw bli mlx akl -->
  <perm id="4" order="4 5 1 2 3"/> <!-- bli mlx akl awtw -->
  <perm id="5" order="4 5 2 3 1"/> <!-- bli mlx awtw akl -->
  <perm id="6" order="1 2 + 4 5"/> <!-- akl at + bli mlx -->
  <perm id="7" order="2 + 1 4 5"/> <!-- at + akl bli mlx -->
  <perm id="8" order="2 + 4 5 1"/> <!-- at + bli mlx akl -->
  <perm id="9" order="4 5 1 2 +"/> <!-- bli mlx akl at + -->
  <perm id="10" order="4 5 2 + 1"/> <!-- bli mlx at + akl -->
</perms>
**PROPER NAMES**

**Example** (*hnri wiliam pwrd “HENRY WILLIAM FORD”*)

```xml
<item id="28605" transliterated="hnri wiliam pwrd">
  <MWE pos="pName" type="person"
       number="singular" gender="masculine"/>
  <atom id="1" lexiconPointer="7356"/> <!-- Henry -->
  <atom id="2" lexiconPointer="2266"/> <!-- William -->
  <atom id="3" lexiconPointer="222"/> <!-- W. -->
  <atom id="4" lexiconPointer="8544"/> <!-- Ford -->
  <perms>
    <perm id="1" order="1 2 4"/> <!-- Henry William Ford -->
    <perm id="2" order="1 3 4"/> <!-- Henry W. Ford -->
    <perm id="3" order="1 4"/> <!-- Henry Ford -->
    <perm id="4" order="3"/> <!-- Ford -->
  </perms>
</item>
```
The morphological generator embodies vast linguistic knowledge which is applicable to MWEs and to single words alike.

But the analyzer operates on a token-by-token basis.

We therefore decided not to interfere with the generator and analyzer, and instead to add a post-processing layer.

First, the existing morphological analyzer is applied to all the tokens of a sentence.

Then, the post-processor identifies MWEs in the analyzed output using information derived from the MWE lexicon.
The extended morphological system

- Raw text input
  - Tokenizer
    - Morphological analyzer
      - XML wrapper
  - Lexicon (excl. MWEs)
    - Morphological generator
      - Inflected forms
      - MWE lexicon
      - Postprocessor
        - Output XML document
Morphological processing of MWEs

The *MWE lexicon* reflects all the information associated with MWEs.

For each MWE we choose an *anchor* word which helps identify it in the text.

When applied to the anchor, the generator produces not only all the inflected forms of that word, but also an additional analysis, as a component of the MWE that this word anchors.

This additional analysis is associated with the ID of the MWE.
EXAMPLE (A PARTIAL ANALYSIS OF *ywrkw* *h* *din* “THE (FEMALE) LAWYERS” *before* POST-PROCESSING)

```xml
<token id="1" surface="ywrkw">
  <analysis id="1">
    <base lexiconPointer="8174" transliterated="ywrk">
      <noun state="absolute" definiteness="false"
        gender="feminine" number="plural"/>
    </base>
  </analysis>
  <analysis id="2">
    <base lexiconPointer="8174" transliterated="ywrk">
      <noun state="construct" definiteness="false"
        gender="feminine" number="plural"/>
    </base>
  </analysis>
</token>
...
POST-PROCESSING

EXAMPLE (A PARTIAL ANALYSIS OF *ywrkwt hdin* “THE (FEMALE) LAWYERS” *before* POST-PROCESSING)

```xml
<token id="2" surface="hdin">
  <analysis id="1">
    <base lexiconPointer="5208" transliterated="din">
      <noun state="absolute" definiteness="true"
             gender="masculine" number="singular"/>
    </base>
  </analysis>
  <analysis id="2"/>
  <base lexiconPointer="28579" transliterated="din">
    <MWE lexiconPointer="28579" atom="2" definiteness="true"/>
  </base>
</token>
```
The post processor works on a sentence-by-sentence basis. It checks the analyses of the tokens in the sentence to find analyses as anchors of MWEs. For each such anchor the post-processor retrieves the entry of the corresponding MWE from the MWE lexicon. This record contains the IDs of the remaining constituents, thereby enabling the post-processor to search for them in the sentence and verify that they satisfy the agreement and order requirements of the MWE. Thus only one database search is needed for each anchor.
Example (A partial analysis of *ywrkwt hdin* "the (female) lawyers" after post-processing)

```xml
<token id="2" surface="hdin">
  ...
  <analysis id="2" />
    <base lexiconPointer="28579" transliterated="ywrk din">
      <MWE pos="noun" definiteness="true" number="plural"
           gender="feminine"/>
    </base>
  </analysis>
</token>
```
IMPLEMENTATION

- These modifications were implemented as part of the MILA tools (Itai and Wintner, 2008), and are currently part of the lexicon and the morphological processor.

- The current MWE lexicon includes a total of 3718 MWEs:

<table>
<thead>
<tr>
<th>POS</th>
<th>Noun</th>
<th>Adj</th>
<th>Prep</th>
<th>Adv</th>
<th>Intrjct</th>
<th>PropN</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>1950</td>
<td>105</td>
<td>23</td>
<td>248</td>
<td>38</td>
<td>1215</td>
<td>139</td>
<td>3718</td>
</tr>
</tbody>
</table>

- More entries are constantly added.
CONCLUSION

- We focus on the special needs of Hebrew, but this architecture is in principle appropriate for several morphologically interesting languages.
- The architecture satisfies many of the properties listed by Savary (2008).
- But not all MWEs can be represented:
  - Constraints on the syntactic structure of potential fillers of the open slot.
  - More intricate interactions of MWEs with productive syntactic structure.
Overview: linguistic properties of MWEs

MWEs in linguistic theory

MWEs in linguistic theory 2: Beyond English

Challenges from other languages (Hebrew)

Lexical Encoding of MWEs

Applications to other languages

Summary


BIBLIOGRAPHY II


Bibliography IV


