Multi-word Expressions: Linguistic properties and lexical representation

Manfred Sailer[1] Shuly Wintner[2]

[1]Institute for English and American Studies
Goethe University Frankfurt
Frankfurt am Main, Germany

[2]Department of Computer Science
University of Haifa
Haifa, Israel

PARSEME Prague Training School
19-23 January, 2015

Outline

Multiword Expressions

Characterization

- 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

Multiword Expressions

Definitions?

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)

Multiword Expressions

Example

Example

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.

Multiword Expressions

Working definition

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.

Multiword Expressions

Scale

- 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

Multiword Expressions

Significance

- 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

Multiword Expressions

Challenge

- 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)

Structure of This Tutorial

- 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

Outline

What is an idiom?

phraseologism, 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: $\neq$ *kick the pail*; $\neq$ *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: ≠ *spilled the vegetables*; ≠ *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:	syn. fixedness	= sem. fixedness	= fixed
Phase 2:	synt. fixedness	= sem. fixedness	= fixed
	synt. fixedness	= sem. fixedness	$\neq$ fixed
Phase 3:	synt. fixedness	= sem. fixedness	= fixed
	synt. fixedness	$\neq$ sem. fixedness	= fixed
	synt. fixedness	= sem. fixedness	$\neq$ fixed

Phrasal lexical entries in Chomsky 1965

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 hatched* (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:	syn. fixedness	= sem. fixedness	= fixed
Phase 2:	synt. fixedness	= sem. fixedness	= fixed
	synt. fixedness	= sem. fixedness	$\neq$ fixed
Phase 3:	synt. fixedness	= sem. fixedness	= fixed
	synt. fixedness	$\neq$ sem. fixedness	= fixed
	synt. fixedness	= sem. fixedness	$\neq$ 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
- Idiomatically combining expressions (ICE):
  - *spill the beans*, *keep tabs on s.o./s.th.* ('wathc 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

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')

Tests for ICEs 1 *cont.*: relatively language independent

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.*

Tests for ICEs 2: more language dependent

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 hes pulled on our account.* (www)
2. **It was the bucket that John kicked* (Schenk, 1995)

Tests for ICEs 2 *cont.*: more language dependent

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.*

TESTS FOR ICES 2 *cont.*: MORE LANGUAGE DEPENDENT

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:

- $\neq$ 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
- $\neq$ 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: **The**y[**bucket**(y)]([**kick**(x,y)) $\mapsto$ **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))

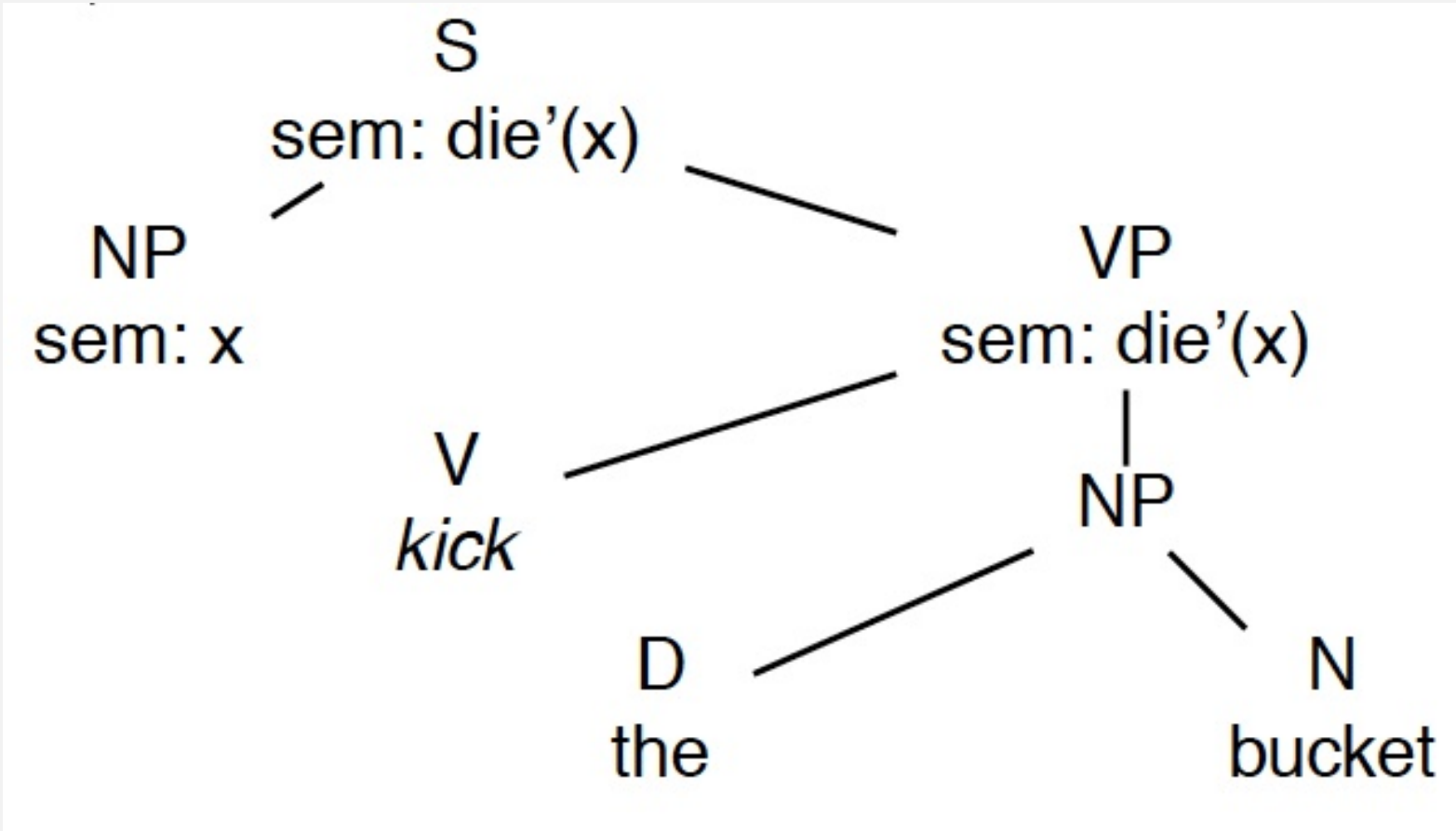
1. *have a bone to pick with s.o.* ('X has s.th. to discuss where Y annoyed X');
   I had a bone to pick with them, but they were so nice that I forgot about it.
2. *shoot the breeze* ('chat')
   **Tony shot the breeze with Junior, and Paulie shot it with Silvio.*

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?

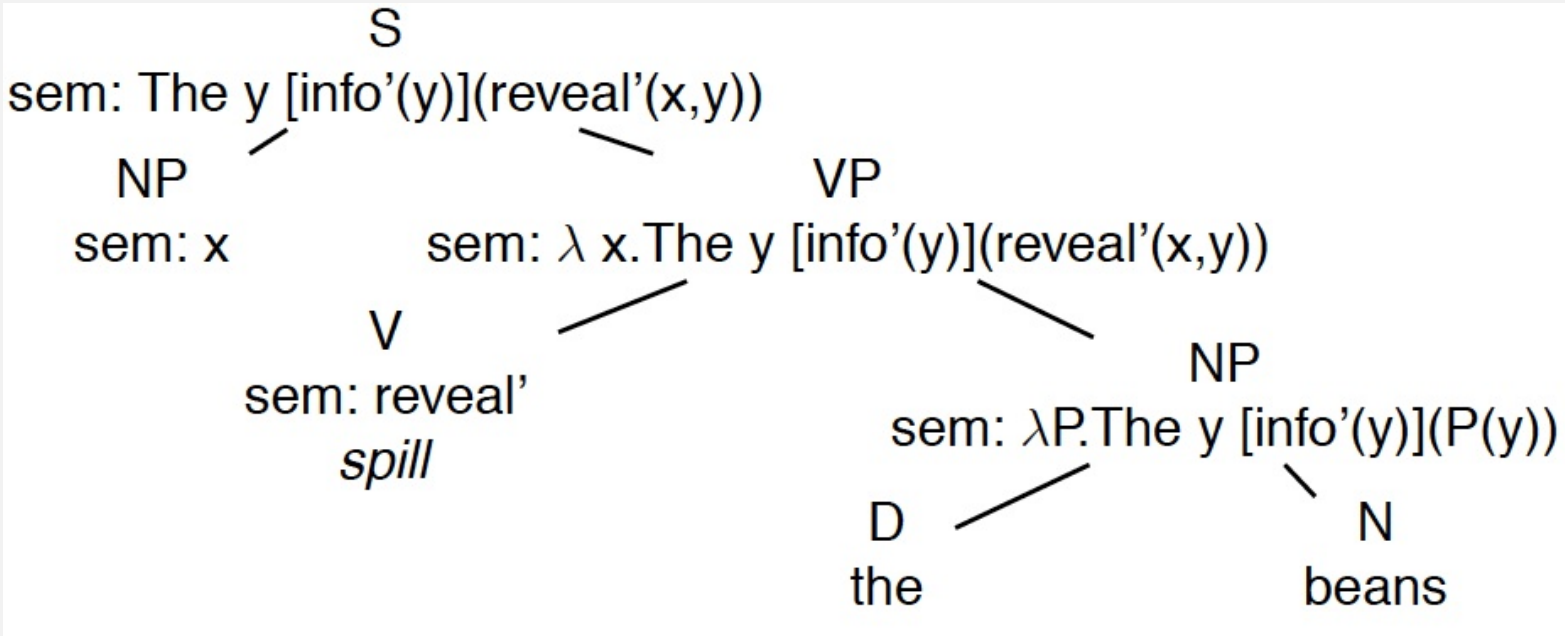
Constructional analysis: Sketch

- 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: 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 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:
  - *spill* $\rightsquigarrow$ **reveal-idiom**
  - *beans* $\rightsquigarrow$ **secret-idiom**
  - *Pat spilled the beans*:
    The x[**secret-idiom**(x)](**reveal-idiom**(**pat**, x))
- semantic constants as partial functions:
  [[**reveal-idiom**]]([[**beans**]]): undefined.
  [[**spill**]]([[**secret-idiom**]]): undefined

Denotational Approach: Sketch *cont.*

- Passive: The beans had been spilled.
  The x[**secret-idiom**(x)]$(\exists y($**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-dimensional, representational analysis: Sketch

- Representatives: Sailer 2003, Soehn 2006, Richter and Sailer 2009, Webelhuth et al. to appear
- 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: $\lambda x.\mathbf{dance}(x)$

EXAMPLE: *trip the light fantastic cont.*

EXAMPLE (SYNTACTIC FLEXIBILITY)

1. *Alex tripped the light fantastic.*
   ok; VP and the head daughter is of the right lexeme.
2. **The light fantastic was tripped*. (passive)
   bad; no VP of the right form
3. **The light fantastic, Alex tripped*. (fronting)
   bad; no VP of the right form

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: **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: **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
   $\textbf{The}y[\textbf{string-id}(y) \wedge \textbf{pull-id}(\textbf{alex}, y)](\textbf{get-C-job}(y))$
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
- Flexibilty captured by stipulation (Sag, 2010)

Limitations of linguistic theorizing

- Focus on English, sometimes making other languages fit the English model (for example Soehn 2006)
  $\Rightarrow$ from Tuesday on: Beyond English and back to English
- Focus on few classes of MWEs, few examples
  $\Rightarrow$ from Wednesday on: Broader overview, lexical resources

Overview

Phase 1:	syn. fixedness	= sem. fixedness	= fixed
Phase 2:	synt. fixedness	= sem. fixedness	= fixed
	synt. fixedness	= sem. fixedness	$\neq$ fixed
Phase 3:	synt. fixedness	= sem. fixedness	= fixed
	synt. fixedness	$\neq$ sem. fixedness	= fixed
	synt. fixedness	= sem. fixedness	$\neq$ fixed

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

Summary from yesterday

- 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 decomposablilty 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:	syn. fixedness	= sem. fixedness	= fixed
Phase 2:	synt. fixedness	= sem. fixedness	= fixed
	synt. fixedness	= sem. fixedness	$\neq$ fixed
Phase 3:	synt. fixedness	= sem. fixedness	= fixed
	synt. fixedness	$\neq$ sem. fixedness	= fixed
	synt. fixedness	= sem. fixedness	$\neq$ 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)

C	Mittelfeld	verbal complex	Nachfeld
dass	*Alex gestern [einen Vogel]*	*mitgebracht hat*	*[der schön singt].*
that	Alex yesterday [a bird]	brought.along has	[that sings well]

Vorfeld	FINIT	Mittelfeld	verbal complex	Nachfeld
	Hat	*A. gestern [einen Vogel]*	*mitgebracht*	*[der schön singt].*
Alex	*hat*	*gestern [einen Vogel]*	*mitgebracht*	*[der schön singt].*
[Einen Vogel]	*hat*	*Alex gestern*	*mitgebracht*	*[der schön singt].*

Scrambling

- 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	*has*	*Alex*	*to.a colleague*	*a joke*	*told*

"Yesterday Alex told a joke to a colleague"

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

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 constitutent
- 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 has a bird [that sings] brought.along [that . . .]
"Alex brought along a bird that sings well."

Alex hat mir ?[dass Chris schnarcht] gesagt [dass . . .]*
Alex has to.me [that Chris snores] said
"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

Non-decomposable idioms in German

Example

ins *Gras* *beißen*
in.the *gras* *bite*
"die"

den *Löffel* *abgeben*
the *spoon* *pass.on*
"die"

den *Geist* *aufgeben*
the *ghost* *give.up*
"die [primarily for machines]"

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*

"$\neq$ 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.
with the Green is the political spoon already before resigning passed.on
"In the Green party, people die politically already before resigning from office."

- 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

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')

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.
(J. Pascha & M. Louis, *The Single Man*, iUniverse. p. 195)

Summary: English idioms

Phase 1:	syn. fixedness	= sem. fixedness	= fixed	*kingdom come*
				kick the bucket
				spill the beans
Phase 2:	synt. fixedness	= sem. fixedness	= fixed	*kingdom come*
				kick the bucket
	synt. fixedness	= sem. fixedness	$\neq$ fixed	*spill the beans*
Phase 3:	synt. fixedness	= sem. fixedness	= fixed	*kingdom come*
	synt. fixedness	$\neq$ sem. fixedness	= fixed	*kick the bucket*
	synt. fixedness	= sem. fixedness	$\neq$ fixed	*spill the beans*

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:

- Kay and Sag ms. (K&S):
  Sign-based Construction Grammar (SBCG, Sag 2012)
  Minimal Recursion Semantics (MRS, Copestake et al. 2003)
- Representational two-dimensional theory (R2DT):
  Head-driven Phrase Structure Grammar (HPSG, Pollard and Sag 1994)
  Lexical Resource Semantics (LRS, Richter and Sailer 2004)

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

4. Restrict syntactic flexibility?

- Follows from general properties of the constructions in individual languages

5. Avoid lexical ambiguity?

- Syntactic flexibility motivates massive lexical ambiguity
- We loose the intuition that MWEs form a unit
- Vision:
  - Lexical representation: syntactic tree with semantic annotation at individual nodes (if decompositional)

  $\Rightarrow$ 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

Hebrew Morphology

- 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)

	Singular	Plural
Masculine	*šwmr*	*šwmrim*
Feminine	*šwmrt*	*šwmrwt*
	"guard"	

	Singular	Plural
Masculine	*kxwl*	*kxwlim*
Feminine	*kxwlh*	*kxwlwt*
	"blue"	

Nominal morphology

Example (Nominal status)

	Absolute	Definite	Construct
Masculine Singular	*klb*	*hklb*	*klb*
Feminine Singular	*klbh*	*hklbh*	*klbt*
Masculine Plural	*klbim*	*hklbim*	*klbi*
Feminine Plural	*klbwt*	*hklbwt*	*klbwt*

"dog"

	Absolute	Definite	Construct
Masculine Singular	*kxwl*	*hkxwl*	*kxwl*
Feminine Singular	*kxwlh*	*hkxwlh*	*kxwlt*
Masculine Plural	*kxwlim*	*hkxwlim*	*kxwli*
Feminine Plural	*kxwlwt*	*hkxwlwt*	*kxwlwt*

"blue"

Nominal morphology

Example (Promonimal suffixes)

	1st	2nd	3rd
Masculine Singular	*klbi*	*klbk*	*klbw*
Feminine Singular	*klbi*	*klbk*	*klbh*
Masculine Plural	*klbnw*	*klbkm*	*klbm*
Feminine Plural	*klbnw*	*klbkn*	*klbn*
		"dog"	

Verbal morphology

Example (Verb inflections)

	1st	2nd	3rd
Masculine Singular	*hlkti*	*hlkt*	*hlk*
Feminine Singular	*hlkti*	*hlkt*	*hlkh*
Masculine Plural	*hlknw*	*hlktm*	*hlkw*
Feminine Plural	*hlknw*	*hlktn*	*hlkw*

	1st	2nd	3rd
Masculine Singular	*alk*	*tlk*	*ylk*
Feminine Singular	*alk*	*tlki*	*tlk*
Masculine Plural	*nlk*	*tlkw*	*ylkw*
Feminine Plural	*nlk*	*tlknh*	*ylkw*

"walk"

Prepositional morphology

Example (Promonimal suffixes)

	1st	2nd	3rd
Masculine Singular	*lidi*	*lidk*	*lidw*
Feminine Singular	*lidi*	*lidk*	*lidh*
Masculine Plural	*lidnw*	*lidkm*	*lidm*
Feminine Plural	*lidnw*	*lidkn*	*lidn*
		"near"	

Hebrew Orthography

- 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)

Hebrew Orthography

Example (Segmentation)

Possible readings of *šbth*:

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

▸ Skip to summary

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* *bdbr*
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 inflection

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

Example (Partial inflection)

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

Partial inflection

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

Example (Partial inflection)

npl *'l* *hraš*
fell *on* *the-head*
"fell on his head" ⇒ lose one's mind

**pwl* *'l* *hraš*
fall *on* *the-head*
"fall on your head" ⇒ lose your mind

Non-standard inflection

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

Non-standard inflection

Example (Non-standard inflection)

idit *dlt*
*handle.*CONST *door*
"door handle"

idit *hdlt*
*handle.*CONST *the-door*
"the door handle"

**hidit* *dlt*
*the-handle.*CONST *door*
"the door handle"

iwšb *raš*
*sitter.*CONST *head*
"sitting head" ⇒ chairman

hiwšb *raš*
*the-sitter.*CONST *head*
"the sitting head" ⇒ the chairman

Fossil words

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

▸ Skip to summary

Syntactic properties

Open slots

Example (Open slots)

akl	*at*	*X*	*bli*	*mlx*
ate	ACC		*without*	*salt*

"ate someone without salt" ⇒ easily defeat

išb	*'l*	*X*	*šb'h*
sat	*on*		*seven*

"sat seven (days) on someone" ⇒ mourne

Syntactic properties

Word order

Example (Verb argument structure)

hwa *ica* *mbito* *l'bwdh*
he *went-out* *from-house-his* *to-the-work*
"he left home for work" ⇒ he left home for work

mbitw *hwa* *ich* *l'bwdh*
from-house-his *he* *went-out* *to-the-work*
"he left home for work" ⇒ from home he left for work

hwa *ica* *mhqlim* *b'bwdh*
he *went-out* *from-the-tools* *in-the-work*
"he left the tools at work" ⇒ he was furious at work

**mhqlim* *hwa* *ica* *b'bwdh*
from-the-tools *he* *went-out* *in-the-work*
"he left the tools at work" ⇒ he was furious at work

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 *bšn* *w'in*
went-out *in-tooth* *and-eye*
"went out in tooth and eye" ⇒ be injured

**ica* *b'in* *wšn*
went-out *in-eye* *and-tooth*
"went out in tooth and eye" ⇒ be injured

pxwt *aw* *iwtr*
less *or* *more*
"less or more" ⇒ more or less

**iwtr* *aw* *pxwt*
more *or* *less*
"more or less" ⇒ more or less

Syntactic properties

Limited paraphrasing

Example (Genitive constructions)

‘wrk *h‘itwn*
*editor-.*CONST *the-journal*
“the journal editor” ⇒ the journal editor

h‘wrk *šl* *h‘itwn*
the-editor *of* *the-journal*
“the editor of the journal” ⇒ the journal editor

‘wrk *hdin*
*editor-.*CONST *the-law*
“the law editor” ⇒ the lawyer

**h‘wrk* *šl* *hdin*
the-editor *of* *the-law*
“the editor of the law” ⇒ the lawyer

Syntactic properties

Limited paraphrasing

Example (Pronominal suffixes)

ica *mbitw*
went-out *from-house-his*
"went out from his house" ⇒ left home

ica *mhbit* *šlw*
went-out *from-the-house* *his*
"went out from his house" ⇒ left home

ica *md'tw*
went-out *from-mind-his*
"went out from his mind" ⇒ lost his mind

**ica* *mhd't* *šlw*
went-out *from-the-mind* *his*
"went out from his mind" ⇒ lost his mind

Syntactic properties

Limited reference

Example (Limited reference)

qiblti *awr* *irwq*
I-received *light* *green*
"I received green light" ⇒ they gave me the green light

hawr *šqiblti* *...*
the-light *that-I-received* *...*
"the light I received..." ⇒ the light they gave me...

Syntactic properties

Limited reference

Example (Category change)

hxlwnwt *hgbwhim*
the-windows *the-high*
"The high windows" ⇒ the powers that be

hgwbh *šl* *hxlwnwt*
the-height *of* *the-windows*
"the height of the windows" ⇒ ???

Syntactic properties

Violated agreement

Example (Violated agreement)

'in *hr'*

eye-INDEF *evil*-DEF

"evil eye" $\Rightarrow$ 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" $\Rightarrow$ 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)

bit	*spr*
house-.CONST	*book*

"house of book" ⇒ school

sprwt	*iph*
literature	*pretty*

"beautiful literature" ⇒ belles-lettres

ab	*bit*	*din*
father-.CONST	*house-*.CONST	*law*

"father of house of law" ⇒ President of the Court

Example (Noun Phrase)

hawmwt	*hmawxdwt*

Common MWE constructions

- Noun-noun constructs
- Adjective-noun constructs
- Support verbs
- Proper names

Common MWE constructions

Example (Noun-noun constructs)

bit	*spr*	
*house-.*CONST	*book*	

"house of book" ⇒ school

ab	*bit*	*din*
*father-.*CONST	*house-.*CONST	*law*

"father of house of law" ⇒ President of the Court

Common MWE constructions

Example (Adjective-noun constructs)

ql	*d't*
light-.CONST	*mind*

"light minded" ⇒ hasty

kxwlt	*'inim*
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)

ap *'l* *pi* *š*
even *on* *mouth-.*CONST *that*
"???" ⇒ although

ap *'l* *pi* *kn*
even *on* *mouth-.*CONST *thus*
"???" ⇒ nevertheless

iwca *dwpn*
go-out *side,bank*
"leaving through the membrabe" ⇒ exceptional

lxm *xwq*
*bread-.*CONST *law*
"bread of law" ⇒ routine

Lexical fixedness

Example

akl *at* *hkwb'*
eat ACC *the-hat*
"eat the hat" ⇒ eat one's hat (regret)

**Trp* *at* *hkwb'*
devour ACC *the-hat*
"devour the hat" ⇒ ???

**akl* *at* *hmgb't*
eat ACC *the-bowler*
"eat the bowler" ⇒ ???

Lexical fixedness

Example

šwlxn *'bwdh*
*table-.*CONST *work*
"working table" ⇒ desk

**kisa* *'bwdh*
*chair-.*CONST *work*
"working chair" ⇒ ???

**šwlxn* *mlakh*
*table-.*CONST *work*
"working table" ⇒ ???

Translational equivalents

Example

bit *spr*
*house-.*CONST *book*
"house of book" ⇒ school

ap *'l* *pi* *kn*
even *on* *mouth-.*CONST *thus*
"???" ⇒ nevertheless

b'l *š'wr* *qwmh*
*owner-.*CONST *measure-.*CONST *height*
"owner of a measure of height" ⇒ honorable

PROPERTIES OF MWES: SUMMARY

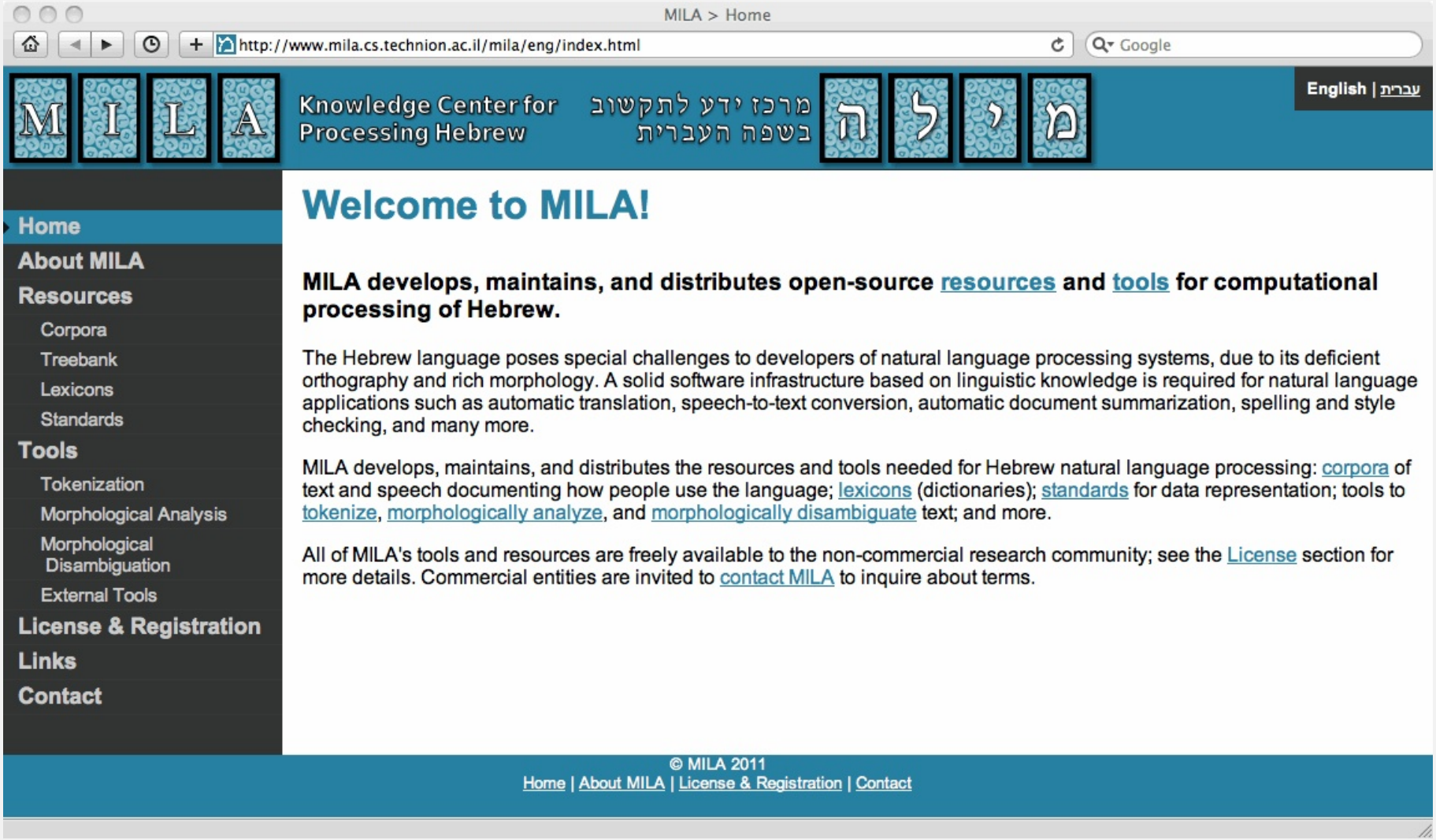
		'wrk din "lawyer"	*'wrk 'itwn* "journal editor"	*mkwnt qpy* "coffee machine"	*hdr hmtnh* "lounge"
Morph.	Fixed form	−	−	−	−
	Partial inflection	+	−	−	−
	Non-standard inflection	+	−	+	−
	Fossil words	−	−	−	−
Syntax	Word order:				
	– Rigid argument structure				−
	– No passivization				
	– Adverbial locations				
	– No coordination				−
	Limited paraphrasing	+	−	−	−
	Limited reference	+	−	−	−
	Violated agreement	−	−	−	−
	Syntactic idiosyncrasies	−	−	−	−
Sem.	Compositionality	Semi	Transparent	Transparent	Transparent
	Lexical fixedness	+	−	−	−
	Translation equivalents	+	−	−	−

Outline

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

Hebrew Language Resources



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 *h* "the", the frequent prepositions *b* "in", *k* "as", *l* "to" and *m* "from", the coordinating conjunction *w* "and" and the subordinating conjunctions *š* "that" and *kš* "when") attach to the word which immediately follows them

Example (The preposition *m* "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
- *m+ph* "from mouth"
- *m+ph l+awzn* "from mouth to ear" ⇒ "through the grapevine"
- **m* "from" + *ph axd* "mouth one" ⇒ "unanimously"

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"

Hebrew Syntax

- 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

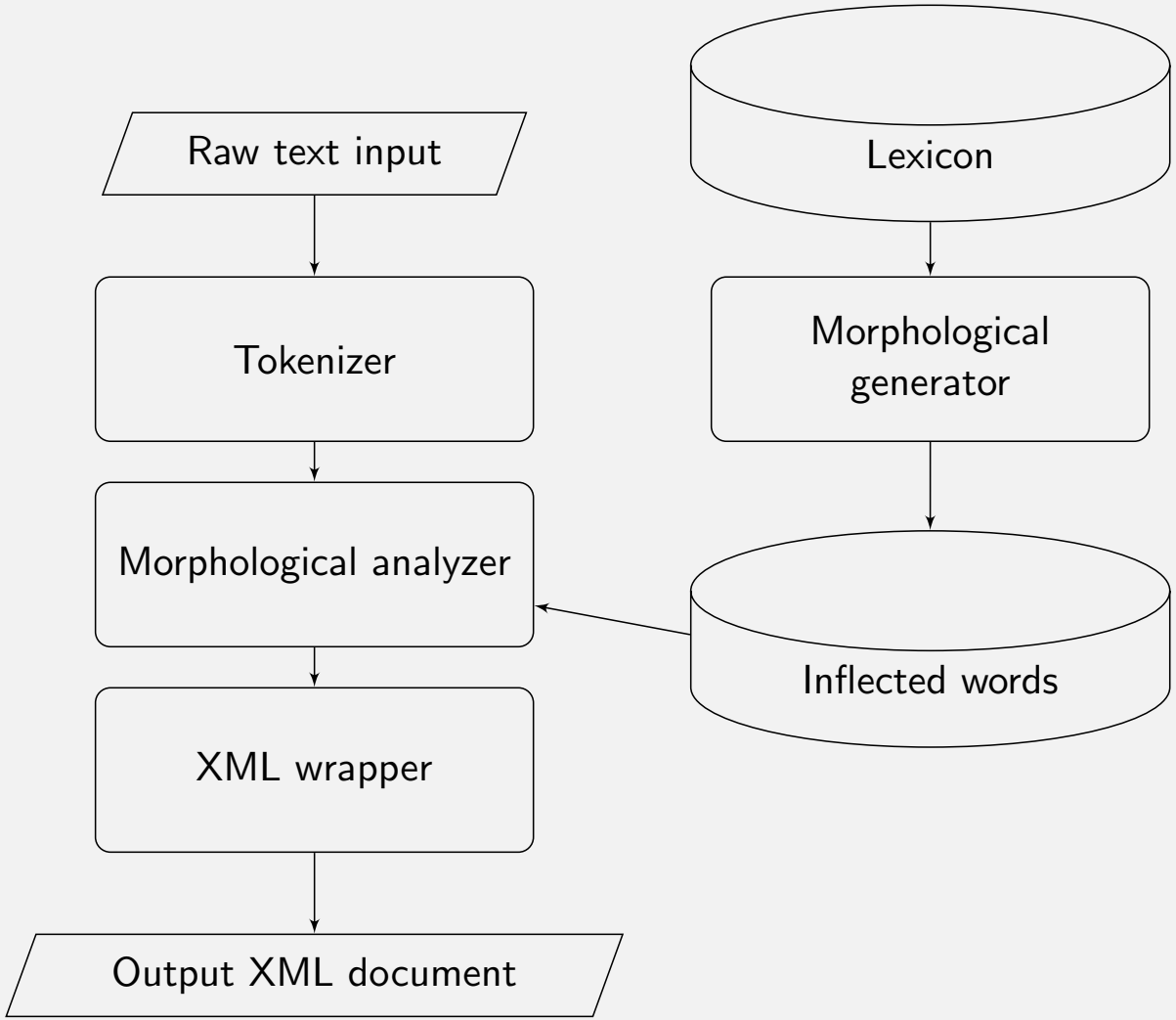
Example

h+sirh h+iph h+zaw "the+boat the+nice the+this" ⇒ "this nice boat"

- Hebrew has three different possessive constructions

The Hebrew morphological processor

(Wintner and Yona, 2003; Itai and Wintner, 2008)



Morphologically analyzed text

Example (*šbi*)

```
    <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

- The 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

Basics

- 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")

```
<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" ⇒ "is uninvolved"
kptwr wprx "button and flower" ⇒ "fantastic"

It can also be some inflected form:

Example

hxlwnwt hgbwhim "the+windows the+high" ⇒ "upper echelon"

Word-level properties

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 xwlik* "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

REPRESENTING MWE CONSTITUENTS

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")

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

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")

```
<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

Example

ywrk din "editor-of law" ⇒ "a lawyer"
ywrkt din "editor-of law" ⇒ "a female lawyer"
ywrki din "editors-of law" ⇒ "lawyers"
ywrk hdin "editor-of the+law" ⇒ "the lawyer"
hywrk din "the+editor-of law" ⇒ "the lawyer"
...

RETRIEVING MORPHOLOGICAL FEATURES

EXAMPLE (*ywrk din* "LAWYER")

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

Morpho-syntactic properties

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

Example

milh nrdpt "word chased" ⇒ "synonym"

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

Example

yin hry "eye the+evil" ⇒ "evil eye"

AGREEMENT AMONG CONSTITUENTS

EXAMPLE (*milh nrdpt* "WORD CHASED" ⇒ "SYNONYM")

```
<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)" ⇒ "mourn"
išb yliw šbyh "sat on-him seven" ⇒ "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 (*yšh imim klilwt* "made days like-nights" ⇒ "work intensively")

```
<item id="39991" transliterated="yšh imim klilwt">
  <MWE pos="VP" tense="1" person="1" number="1" gender="1"/>
  <atom id="1" lexiconPointer="376"><inflect/></atom>     <!-- yšh -->
  <atom id="2" lexiconPointer="9475">                      <!-- iwm -->
    <inflect state="absolute" definiteness="false" number="plural"/>
  </atom>
  <atom id="3" lexiconPointer="20001"><prefix/></atom>      <!-- k -->
  <atom id="4" lexiconPointer="8024">                    <!-- lilh -->
    <inflect state="absolute" definiteness="false" number="plural"/>
  </atom>
  <perms>
    <perm id="1" order="1 2 3 4"/>
    <perm id="2" order="1 4 3 2"/>
  </perms>
</item>
```

Open slots

Example (*akl at X bli mlx* "eat X without salt" ⇒ "defeat")

```
<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/>                                    <!-- akl -->
  </atom>
  <atom id="2" lexiconPointer="3382"/>            <!-- at -->
  <atom id="3" lexiconPointer="0">   <!-- pronominal suffix -->
    <suffix/>
  </atom>
  <atom id="4" lexiconPointer="21542"/>           <!-- bli -->
  <atom id="5" lexiconPointer="608"/>             <!-- mlx -->
```

Open slots

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>
</item>
```

Proper names

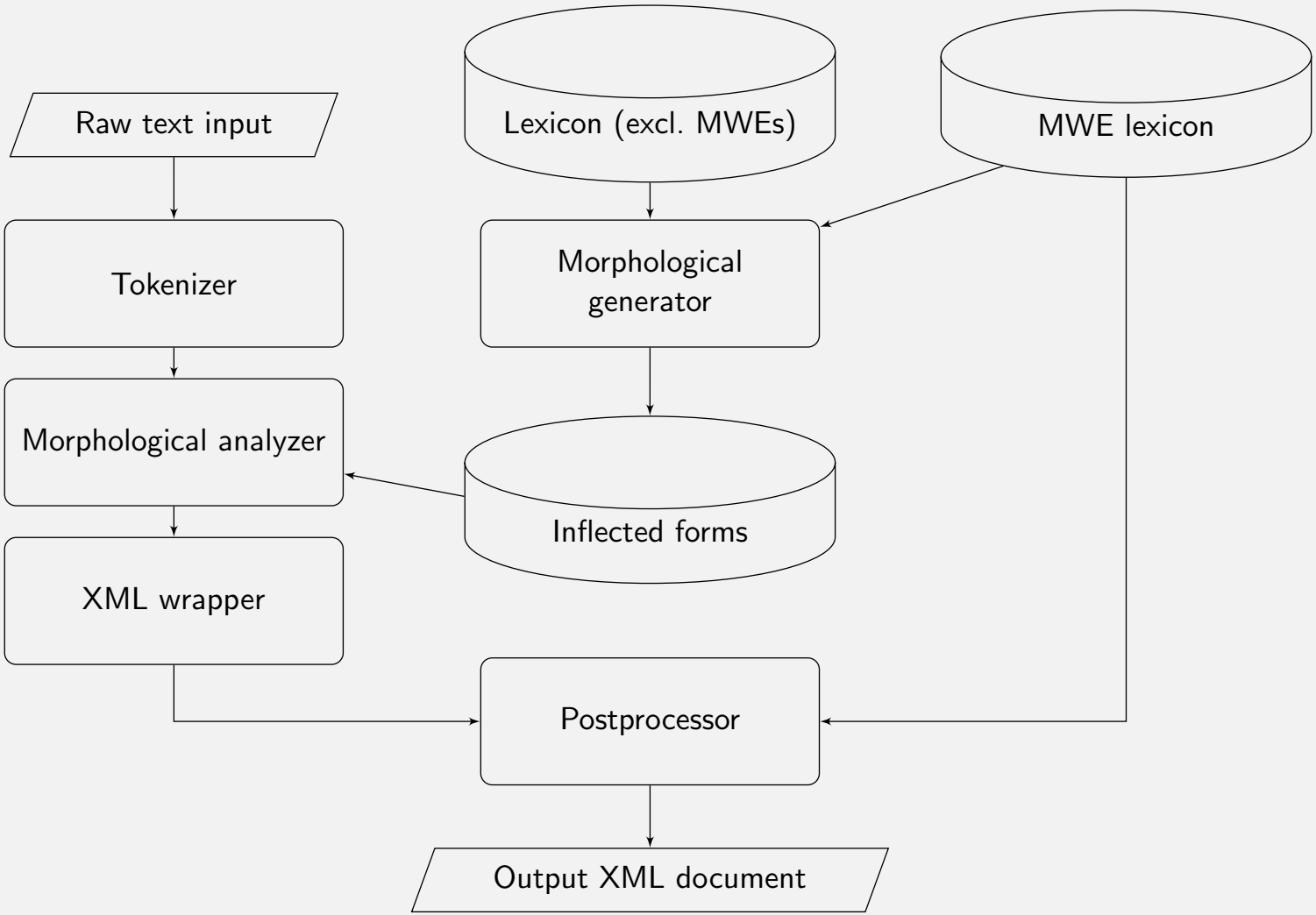
Example (*hnri wiliam pwrd* "Henry William Ford")

```
<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>
```

Morphological processing of MWEs

- 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



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

POST-PROCESSING

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

```
<token id="1" surface="ywrkwt">
  <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)

```
<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>
  </analysis>
</token>
```

Post-processing

- 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

POST-PROCESSING

EXAMPLE (A PARTIAL ANALYSIS OF *ywrkwt hdin* "THE (FEMALE) LAWYERS" *after* POST-PROCESSING)

```
<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:

POS	Noun	Adj	Prep	Adv	Intrjct	PropN	Other	Total
Count	1950	105	23	248	38	1215	139	3718

- 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

Outline

Outline

Bibliography I

Anne Abeillé. The flexibility of French idioms: A representation with Lexical Tree Adjoining Grammar. In Martin Everaert, Eric-Jan van der Linden, André Schenk, and Ron Schreuder, editors, *Idioms. Structural and Psychological Perspectives*, pages 15–42. Lawrence Erlbaum Associates, Hillsdale, 1995.

Hassan Al-Haj, Alon Itai, and Shuly Wintner. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2):130–170, June 2014.

Iñaki Alegria, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. Representation and treatment of multiword expressions in Basque. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 48–55, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Timothy Baldwin and Takaaki Tanaka. Translation by machine of complex nominals: Getting it right. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. A statistical approach to the semantics of verb-particles. In Diana McCarthy Francis Bond, Anna Korhonen and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72. Association for Computational Linguistics, 2003. URL `http://www.aclweb.org/anthology/W03-1809.pdf`.

Wallace Chafe. Idiomaticity as an anomaly in the Chomskyan paradigm. *Foundations of Language*, 4:109–127, 1968.

Noam Chomsky. *Aspects of the theory of syntax*. MIT Press, 1965.

Ann Copestake, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan Sag, and Dan Flickinger. Multiword expressions: Linguistic precision and reusability. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1941–7, Las Palmas, Canary Islands, 2002.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. Minimal Recursion Semantics: An introduction. Manuscript, November 2003.

Jeanette S. DeCarrico and James R. Nattinger. Lexical phrases and strategic interaction. In *Georgetown University Round Table on Languages and Linguistics*, pages 558–567. Georgetown University Press, Washington, DC, 1993.

Bibliography II

Antoine Doucet and Helana Ahonen-Myka. Non-contiguous word sequences for information retrieval. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 88–95, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Andy Egan. Pretense for the complete idiom. *Noûs*, 42(3):381–409, 2008.

Britt Erman and Beatrice Warren. The idiom principle and the open choice principle. *Text*, 20(1):29–62, 2000.

Thomas Ernst. Grist for the linguistic mill: Idioms and 'extra' adjectives. *Journal of Linguistic Research*, 1:51–68, 1981.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press, 1998.

Bruce Fraser. Idioms within a transformational grammar. *Foundations of Language*, 6:22–42, 1970.

David Gaatone. Les locutions verbales et les deux passifs du français. *Langage*, 27(109):37–52, 1993.

Gerald Gazdar, Ewan Klein, Geoffrey Pullum, and Ivan Sag. *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge, Massachusetts, 1985.

Maurice Gross. La grammaire transformationelle du français: Les expression figées. Technical report, LADL, University Paris VII, 1989. Technical report.

Francis Roger Higgins. On the use of idioms as evidence for movement. a cautionary note. Unpublished manuscript of a talk given at LSA 1974, New York, 1974.

Tilman N. Höhle. Der Begriff "Mittelfeld". Anmerkungen über die Theorie der topologischen Felder. In A. Schöne, editor, *Kontroversen alte und neue*, Akten des 7. Internationalen Germanistenkongresses, Göttingen 1985, pages 329–340, 1986.

Alon Itai and Shuly Wintner. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98, March 2008.

Alon Itai, Shuly Wintner, and Shlomo Yona. A computational lexicon of contemporary Hebrew. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, May 2006.

Ray Jackendoff. Morphological and semantic regularities in the lexicon. *Language*, 51(3):639–671, 1975.

Ray Jackendoff. *The Architecture of the Language Faculty*. MIT Press, Cambridge, USA, 1997.

Bibliography III

Paul Kay and Ivan A. Sag. A lexical theory of phrasal idioms. Available at: www1.icsi.berkeley.edu/∼kay/idioms-submitted.pdf, ms.

Susumu Kuno and Ken-ichi Takami. *Functional Constraints in Grammar. On the Unergative-unaccusative Distinction*. Benjamins, Amsterdam, Philadelphia, 2004.

Ronald W. Langacker. *Foundations of Ccognitive Grammar*. Stanford University Press, Stanford, 1987.

Adam Makkai. *Idiom Structure in English*. Mouton, 1972.

James D. McCawley. The syntax and semantics of English relative clauses. *Lingua*, 53:99–149, 1981.

Scott McClure. Modification in non-combining idioms. *Semantics and Pragmatics*, 4:Article 7, 1–7, 2011.

Kadri Muischnek and Heiki-Jan Kaalep. The variability of multi-word verbal expressions in Estonian. *Language Resources and Evaluation*, 44:115–135, 2010. doi: 10.1007/s10579-009-9096-x.

Gereon Müller. Phrasen und Konstruktion: Einführung. Handout for a seminar meeting. Available at http://www.uni-leipzig.de/∼muellerg/phrasem1.pdf, 2012.

Stefan Müller. *Head-Driven Phrase Structure Grammar. Eine Einfürung*. Stauffenburg, Tübingen, 2007.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. Idioms. *Language*, 70:491–538, 1994.

Kemal Oflazer, Özlem Çetinoğlu, and Bilge Say. Integrating morphology with multi-word expression processing in Turkish. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 64–71, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago and London, 1994.

Paul M. Postal. *Three Investigations of Extraction*. MIT Press, Cambridge, Mass., 1998.

Stephen G. Pulman. The recognition and interpretation of idioms. In Cristina Cacciari and Patrizia Tabossi, editors, *Idioms: Processing, Structure, and Interpretation*, chapter 11, pages 249–270. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1993.

Bibliography IV

Norvin Richards. An idiomatic argument for lexical decomposition. *Linguistic Inquiry*, 32(1):183–192, 2001.

Frank Richter and Manfred Sailer. Cranberry words in formal grammar. In Claire Beyssade, Olivier Bonami, Patricia Cabredo Hofherr, and Francis Corblin, editors, *Empirical Issues in Formal Syntax and Semantics*, volume 4, pages 155–171. Presses Universitaires de Paris-Sorbonne, Paris, 2003.

Frank Richter and Manfred Sailer. Basic concepts of lexical resource semantics. In Arne Beckmann and Norbert Preining, editors, *ESSLLI 2003 – Course Material I*, volume 5 of *Collegium Logicum*, pages 87–143. Kurt Gödel Society Wien, Vienna, 2004. ISBN 3-901546-00-6.

Frank Richter and Manfred Sailer. Phraseological clauses in constructional HPSG. In Stefan Müller, editor, *Proceedings of the 16th International Conference on Head-Driven Phrase Structure Grammar, Göttingen 2009*, pages 297–317, Stanford, 2009. CSLI Publications. `cslipublications.stanford.edu/HPSG/2009/richter-sailer.pdf`.

Nicolas Ruwet. On the use and abuse of idioms in syntactic argumentation. In *Syntax and Human Experience*, pages 171–251. University of Chicago Press, Chicago, London, 1991. Edited and translated by John Goldsmith.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico, 2002.

Ivan A. Sag. Feature geometry and predictions of locality. In Greville Corbett and Anna Kibort, editors, *Features. Perspectives on a Key Notion in Linguistics*, pages 236–271, Oxford, 2010. Oxford University Press.

Ivan A. Sag. Sign-Based Construction Grammar: An informal synopsis. In Hans C. Boas and Ivan A. Sag, editors, *Sign-Based Construction Grammar*, pages 69–202. CSLI Publications, Stanford, 2012.

Manfred Sailer. Combinatorial semantics and idiomatic expressions in Head-Driven Phrase Structure Grammar. Phil. Dissertation (2000). Arbeitspapiere des SFB 340. 161, Universität Tübingen, 2003. URL `www.sfs.uni-tuebingen.de/hpsg/archive/bibliography/papers/sailer_dissertation03.ps`.

Manfred Sailer. Distributionsidiosynkrasien: Korpuslinguistische Erfassung und grammatiktheoretische Deutung. In Kathrin Steyer, editor, *Wortverbindungen — mehr oder weniger fest*, Institut für Deutsche Sprache, Jahrbuch 2003, pages 194–221. de Gruyter, Berlin and New York, 2004.

Bibliography V

Agata Savary. Computational inflection of multi-word units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1:1–53, 2008.

André Schenk. The syntactic behavior of idioms. In Martin Everaert, Eric-Jan van der Linden, André Schenk, and Ron Schreuder, editors, *Idioms. Structural and Psychological Perspectives*, pages 253–271. Lawrence Erlbaum Associates, Hillsdale, 1995.

Jan-Philipp Soehn. *Über Bärendienste und erstaunte Bauklötze. Idiome ohne freie Lesart in der HPSG*. Peter Lang, Frankfurt am Main, 2006. Ph.D. thesis, Friedrich-Schiller-Universität Jena.

Maria Helena Svensson. A very complex criterion of fixedness: Non-compositionality. In Sylviane Granger and Fanny Meunier, editors, *Phraseology: An Interdisciplinary Perspective*, pages 81–93. John Benjamins, Amsterdam, 2008.

Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. Disambiguating Japanese compound verbs. *Computer Speech & Language*, 19(4):497–512, October 2005.

Sriram Venkatapathy and Aravind Joshi. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 20–27, Sydney, Australia, July 2006.

Aline Villavicencio, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. Lexical encoding of MWEs. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 80–87, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, 19(4):365–377, 2005.

Thomas Wasow, Ivan A. Sag, and Geoffrey Nunberg. Idioms: An interim report. In S. Hattori and K. Inoue, editors, *Proceedings of the XIIIth International Congress of Linguistics*, pages 102–115, 1983.

Catherine Wearing. Metaphor, idiom, and pretense. *Noûs*, 46(3):499–522, 2012.

Gert Webelhuth and Farrell Ackerman. German idioms: An empirical approach. *Studies in the Linguistic Sciences*, 24: 455–471, 1994.

Bibliography VI

Gert Webelhuth, Sascha Bargmann, and Christopher Götze. Idioms as evidence for the proper analysis of relative clauses. In Manfred Krifka, Rainer Ludwig, and Mathias Schenner, editors, *Reconstruction Effects in Relative Clauses*. Akademie Verlag, Berlin, to appear.

Uriel Weinreich. Problems in the analysis of idioms. In (Weinreich, 1980), pp. 208–264, 1969.

Uriel Weinreich. *On Semantics*. University of Pennsylvania Press, 1980.

Shuly Wintner and Shlomo Yona. Resources for processing Hebrew. In *Proceedings of the MT-Summit IX workshop on Machine Translation for Semitic Languages*, pages 53–60, New Orleans, September 2003.