Morphology and Syntax

A Typological Approach

David R. Mortensen

Language Technologies Institute
Carnegie Mellon University

November 1, 2018

1. Morphology
    - What is Morphology?
    - What is a Word?
    - Formal Operations
    - Morphological Functions
    - Traditional Typology of Morphology
    - Improved Typological Features

2. Syntax
    - What is Syntax
    - Constituency
    - Dependency
    - Word Order Typology

3. Conclusion

Morphology

Linguistic Morphology is the study of the structure of words

1 Morphology

- What is Morphology?
- What is a Word?
- Formal Operations
- Morphological Functions
- Traditional Typology of Morphology
- Improved Typological Features

2 Syntax

- What is Syntax
- Constituency
- Dependency
- Word Order Typology

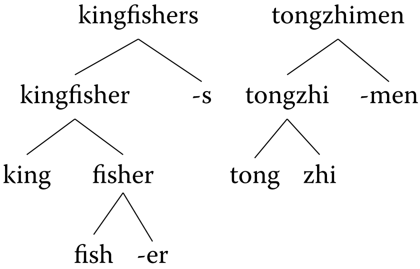
3 Conclusion

Breaking the definition down

- Morphology is the study of the structure of words
- Assumptions
  - There are linguistic units called “words”
  - These units can have internal structure
- Examples
  - *un-dead*
  - *king-fish-er-s*
  - *re-implement-ation-s*
  - 同志们 *tong-zhi-men* same-purpose-PL ‘comrades’
  - 牛肉 *niu-rou* cattle-meat ‘beef’
- The minimal meaningful units of words are called morphemes

Hierarchical structure

- Words are not just sequences of morphemes
- Words have hierarchical structure
- Examples:



1 Morphology

- What is Morphology?
- What is a Word?
- Formal Operations
- Morphological Functions
- Traditional Typology of Morphology
- Improved Typological Features

2 Syntax

- What is Syntax
- Constituency
- Dependency
- Word Order Typology

3 Conclusion

The problem of wordhood

- Perhaps the most difficult aspect of morphology is providing a good, cross-linguistically valid, definition of *word*
- Token separated by whitespace? Many languages don't delimit words with punctuation or whitespace; also, there are clitics like *'s* and *n't*
- Meaning needs to be listed in a dictionary? Many multi-word expression are also idiosyncratic; all of these may be grouped together as *listemes*, but *listemes* are clearly a superset of words
- Follows a different set of combinatorial principles than syntactic units? This is promising, but it is not always possible to tell
- A single phonological domain? Also useful, but not adequate by itself
- Intuitions of speakers? Not always consistent

1 Morphology

- What is Morphology?
- What is a Word?
- Formal Operations
- Morphological Functions
- Traditional Typology of Morphology
- Improved Typological Features

2 Syntax

- What is Syntax
- Constituency
- Dependency
- Word Order Typology

3 Conclusion

Compounding

- Perhaps the most widespread morphological operation is compounding, where two STEMS are combined to form a new stem
- Very common in English, but sometimes not evident because many English compounds are written with spaces (unlike, e.g. German compounds)
  - *dog house*
  - *red head*
  - *figher-bomber*
- Compare German compounds:
  - *Handschuh* hand-shoe 'glove'
  - *Weltschmerz* world-ache 'world-weariness'
  - *Schweinehund* pig-dog 'pig-dog; bastard'
- Chinese also uses compounding extensively:
  - 田鼠 *tianshu* field-mouse 'field mouse'
  - 书包 *shubao* book-container 'sachell'
  - 天地 *tiandi* heaven-earth 'universe'

Affixation

Affixation is the concatenation of a MORPHEME other than a stem to a stem. Affixes can be concatenated after the stem (suffixes) or before the stem (prefixes):

	Present	Perfect	Preterit
1SG	mach-e	ge-mach-t	mach-t-e
2SG	mach-st	ge-mach-t	mach-t-est
3SG	mach-t	ge-mach-t	mach-t-e
1PL	mach-en	ge-mach-t	mach-t-en
2PL	mach-t	ge-mach-t	mach-t-et
3PL	mach-en	ge-mach-t	mach-t-en

Table: German weak verb: MACHEN ‘to make’

Across languages, suffixes are more common than prefixes.

Infixation

Infixation is the insertion of an affix into a BASE. It is not the same as "stacking affixes"—the infix can actually interrupt another morpheme.

- Infixation is important to the grammar of many languages, especially languages of the Pacific and North America
- It plays a marginal role in English
- Expletive infixation:
  - *Pennsyl-fuckin'-vania*
  - *im-fuckin'-plausible*
  - *ty-bloody-phoon*
- In a moment, we'll see a less frivolous-looking example of this process, but first…

Reduplication

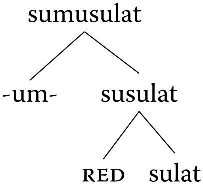
Reduplication is when all or part of a BASE is repeated.

- Reuplication is commonly used to express notions like plurality, diminution, and imperfectivity
- *anak* ‘child’ → *anak-anak* ‘children’
- It may express anything, though

Infixation and reduplication in Tagalog

Tagalog, the basis of Filipino (the national language of the Philippines) makes extensive use of both infixation and reduplication in its grammar:

Stem	Perfective	Contemplative	Imperfective	Gloss
kain	kumain	kakain	kumakain	‘eat’
sulat	sumulat	susulat	sumusulat	‘write’
hanap	humanap	hahanap	humahanap	‘seek’



Internal change

- Morphology may also take the form of changes internal to the base
- English has two types of this kind of process: ablaut and umlaut
  - ABLAUT affects verbs
    - *sing* : *sang* : *sung*
    - *begin* : *began* : *begun*
    - *bleed* : *bled* : *bled*
  - UMLAUT affects nouns
    - *foot* → *feet*
    - *tooth* → *teeth*
    - *goose* → *geese*
- Internal change is common in Indo-European languages including many languages of the Indian subcontinent (e.g. Bengali and Sinhala)

Root-and-pattern morphology

Many Afroasiatic languages, including the Semitic languages Arabic, Amharic, and Hebrew, employ so-called root-and-pattern (or templatic) morphology where a consonantal root combines with a template and a sequence of vowels to form a word. Here is an example with the Arabic root *ktb*, 'pertaining to writing':

	Perfect		Imperfect		Participle	
	Active	Passive	Active	Passive	Active	Passive
I	katab	kutib	ktub	ktab	kaatib	ktuub
II	kattab	kuttib	kattib	kattab	kattib	kattab
III	kaatab	kuutib	kaatib	kaatab	kaatib	kaatab
IV	ʔaktab	ʔuktib	ktib	ktab	ktib	ktab
V	takattab	tukuttib	takattab	takattab	takattib	takattab
VI	takaatab	tukuutib	takaatab	takaatab	takaatib	takaatab
VII	nkatab	nkutib	nkatib	nkatab	nkatib	nkatab
VIII	ktatab	ktutib	ktatib	ktatab	ktatib	ktatab
IX	ktab(a)b	ktab(i)b	ktab(i)b			
X	staktab	stuktib	staktib	staktab	staktib	staktab

1 Morphology

- What is Morphology?
- What is a Word?
- Formal Operations
- Morphological Functions
- Traditional Typology of Morphology
- Improved Typological Features

2 Syntax

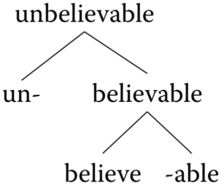
- What is Syntax
- Constituency
- Dependency
- Word Order Typology

3 Conclusion

Derivation

Morphological derivation refers to morphological processes that create new lexemes—that change the meaning and/or part of speech of the base

- English derivational morphology



unbelievable

un- believable

believe -able

- Karok derivational morphology

laːy	‘to pass’	legaːy	‘to really pass’
koʔmoy	‘to hear’	kegoʔmoy	‘to really hear’
trahk	‘to fetch water’	treganhk	‘to really fetch water’

Inflection

Morphological inflection adds syntactically-relevant information (case, number, gender, tense, aspect, modality, etc.) to a word. Consider the following example of the Latin noun *amīca* ‘friend (fem.); girlfriend’:

	SG	PL
NOM	amīca	amīcae
VOC	amīca	amīcae
ACC	amīcam	amīcās
GEN	amīcae	amīcārum
DAT	amīcae	amīcīs
ABL	amīcā	amīcīs

English is poor in inflectional morphology, but has some inflectional suffixes like *-s/-es* ‘plural’, *-s/-es* ‘third person singular non-past’, *-ed* ‘past’, and so on.

1 Morphology

- What is Morphology?
- What is a Word?
- Formal Operations
- Morphological Functions
- Traditional Typology of Morphology
- Improved Typological Features

2 Syntax

- What is Syntax
- Constituency
- Dependency
- Word Order Typology

3 Conclusion

Five types

Traditionally, the morphologies of language have been divided into five types:

- Isolating
- Agglutinating
- Flexional/fusional
- Templatic
- Polysynthetic

Problematically, these categories are not all in the same dimension, but the terms are widely used so we'll cover them anyway.

Isolating and agglutinating

Isolating languages are those where each word, to a great extent, consists of a single morpheme; agglutinating languages are those where words consist of sequences of morphemes, each of which has (roughly speaking) one meaning.

- Isolating languages
  - Parade example: **Chinese**
  - Some compounding, very little affixation
  - Almost all lexemes have a single form
  - English is also relatively isolating
- Agglutinative languages
  - Parade example: **Turkish**
  - Extensive suffixation; each suffix usually carries a single meaning
  - Many forms for a single lexeme
  - *ev* *-ler* *-iniz* *-den*
    house PL POSS2SG ABL

    ‘from your house’

Flexional/fusional and templatic

Flexional languages are those in which there is frequently not a one-to-one relationship between affixes and units of meaning. In a single word, one affix may express multiple meanings or one meaning may be expressed by multiple affixes. Templatic languages are a special case of flexional languages characterized by extensive root-and-pattern morphology.

- Flexional/fusional languages
  - Parade example: **Latin**

	SG	PL
NOM	amīc-a	amīc-ae
VOC	amīc-a	amīc-ae
ACC	amīc-am	amīc-ās
GEN	amīc-ae	amīc-ārum
DAT	amīc-ae	amīc-īs
ABL	amīc-ā	amīc-īs

- Templatic languages
  - Parade example: **Hebrew**

Polysynthetic

Polysynthetic languages are languages in which noun arguments like objects can be expressed as part of a verb, meaning that full sentences can be expressed as a verb alone (not just through agreement with person and number, but through the "incorporation" of the noun into the verb). Take the following example from Nahuatl:

- *ni-c-qua in nacatl*
  I-it-eat the flesh

  'I eat the flesh.'
- *ni-naca-qua*
  I-flesh-eat

  'I eat flesh.'

1 Morphology

- What is Morphology?
- What is a Word?
- Formal Operations
- Morphological Functions
- Traditional Typology of Morphology
- Improved Typological Features

2 Syntax

- What is Syntax
- Constituency
- Dependency
- Word Order Typology

3 Conclusion

Improved typological features: degrees of synthesis and fusion

A simplified framework for morphological typology that better captures variation in morphology is based on DEGREE OF SYNTHESIS and DEGREE OF FUSION, both of which are treated as scales.

- Degree of synthesis
    - **The number of units of meaning per word**
    - "Agglutinating" languages have a high degree of synthesis
    - "Isolating" or "analytic" languages have a low degree of synthesis
    - "Fusional" or "flexional" languages may have a high or low degree of synthesis; English is arguably flexional, but has a low degree of synthesis
- Degree of fusion
    - **The number of units of meaning per formative (root or affix)**
    - "Fusional" or "flexional" languages have a high degree of fusion
    - "Agglutinating" languages have a low degree of fusion
    - "Isolating" languages would typically have a low degree of fusion
- Two dimensional space, with every language occupying some point in that space, instead of a system of prototypes more-or-less like actual languages

Syntax

Syntax is the structure of phrases and sentences

1 Morphology

- What is Morphology?
- What is a Word?
- Formal Operations
- Morphological Functions
- Traditional Typology of Morphology
- Improved Typological Features

2 Syntax

- What is Syntax
- Constituency
- Dependency
- Word Order Typology

3 Conclusion

1 Morphology

- What is Morphology?
- What is a Word?
- Formal Operations
- Morphological Functions
- Traditional Typology of Morphology
- Improved Typological Features

2 Syntax

- What is Syntax
- Constituency
- Dependency
- Word Order Typology

3 Conclusion

Context-free grammars

Most linguists do not use context free grammars to model natural language—they are not expressive enough (the grammars, not the linguists). However, a lot of NLP work assumes CFGs or PCFGs, so we will use them as an example of constituency grammars.

The mathematical definition of a context free grammar, or CFG:

- Vocabulary of terminal symbols, Σ
- Set of non-terminal symbols, N
- Special start symbols, $S \in N$
- Production rules of the form $X \rightarrow \alpha$ where
  $X \in N$
  $\alpha \in (N \cup \Sigma)^*$

A context-free grammar

Here is a simple context-free grammar. S is the start symbol; you can think of it meaning either “start” or “sentence”:

- S → NP VP
- NP → Det Noun
- VP → Verb NP
- Det → the, a
- Noun → boy, girl, hotdogs
- Verb → likes, hates, eats

What sentences does this grammar recognize? Which of these are ungrammatical?

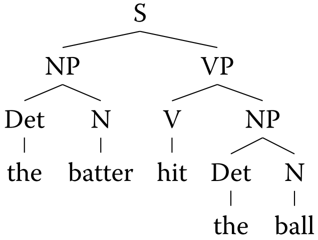
What is a constituent?

In terms of a context-free grammar, a constituent is a sequence of terminal nodes that are dominated by a single node. The node must dominate all of the terminals and must dominate no other terminals. In theoretical terms, a constituent is a sequence of words/tokens that pass certain tests. Some of these are specific to English:

- Coordination
- Substitution
  - General substitution
  - Pro-form substitution
  - Do-so substitution
  - One substitution
- Ellipsis
  - Answer ellipsis
  - VP-ellipsis
  - Pseudoclefting
  - Passivization
  - Deletion
  - Intrusion
  - Wh-fronting
- Topicalization
- Right-node raising

Constituency

Take, for example, the following parse tree, illustrating the constituency of the sentence *The batter hit the ball*:



- We can tell that *the batter* should be a constituent, and therefore should be dominated by a single non-terminal
  - GENERAL SUBSTITUTION: *Batters hit the ball.*
  - PRO-FORM SUBSTITUTION: *She hit the ball.*
  - COORDINATION: *The batter and the bat hit the ball.*

1 Morphology

- What is Morphology?
- What is a Word?
- Formal Operations
- Morphological Functions
- Traditional Typology of Morphology
- Improved Typological Features

2 Syntax

- What is Syntax
- Constituency
- Dependency
- Word Order Typology

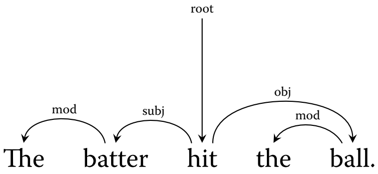
3 Conclusion

Dependency

- Constituency is only one way of looking at syntactic structure
- Another, equally valid, way of looking at syntax is through the lens of dependencies
- In fact, some syntactic frameworks like LFG (Lexical Functional Grammar) use constituency and dependency as simultaneous and mutually-constraining representations
- While constituency grammars look at sentences as trees of nested constituents, each consisting of one or more terminal nodes, dependency grammars look at sentences as graphs of BILEXICAL DEPENDENCIES
  - By "bilexical," we mean that the relations are between two words
  - One of these words is (typically) the head and the other word is the dependent; that is, it depends on the head
  - A head is the more syntactically central word
  - It is difficult to come up with universally agreed-upon tests for this, thus there are many conventions for making dependency trees/graphs

Dependency parses

Here is a dependency graph:



- The head of the whole sentence is the verb *hit*
- The direct dependents of *hit* are the SUBJECT *batter* and the OBJECT *ball*
- Because this is a labeled dependency graph, the arcs are labeled with the corresponding relation ("subj," "obj," and "mod")
- "Batter" and "ball" are both modified by definite articles (*the*)

Dependency versus constituency

If you have to choose, should you use dependency or constituency representations in your work? Which is better?

- Dependency graphs (particularly labeled dependency graphs) have a more direct representation of certain aspects of grammatical encoding
  - It is easier to tell what is subject and what is object
  - It is therefore easier to tell what is agent and what is patient
  - Dependency trees can be better for semantic role labeling (SRL)
- Constituency trees have a better alignment with model-theoretic semantics—constituents line up with semantic units
- Dependency graphs are simpler and more compact
- Constituency trees contain information that is not in dependency graphs, while the reverse is not necessarily true
- There are widely agreed-upon tests for constituency; there are not such tests for headedness/dependency

1 Morphology

- What is Morphology?
- What is a Word?
- Formal Operations
- Morphological Functions
- Traditional Typology of Morphology
- Improved Typological Features

2 Syntax

- What is Syntax
- Constituency
- Dependency
- Word Order Typology

3 Conclusion

Subject, verb, and object

One way in which main-clause word-order has been characterized is in terms of subject (S), object (O), and verb (V). Listed in order of frequency, here are the permutations of S, O, and V:

- **SOV**: Japanese, Korean, Turkish, Hindi, Tamil
- **SVO**: English, Spanish, Chinese, Vietnamese, Swahili
- **VSO**: Tagalog, Irish, Maori, Mixtec
- **VOS**: Malagasy, Tzotzil, Seediq, Nicobarese
- **OVS**: Hixkaryana, Tuvaluan, Urarina
- **OSV**: Kxoe, Nadëb, Tobati

Head-initial and head-final word order

There are a great many other ways that the word order of languages can vary:

- object and *verb* (separate from subject)
- adjectival modifier and *noun*
- *adposition* (preposition or postposition) and noun phrase
- possessor and *head noun*
- relative clause and *head noun*

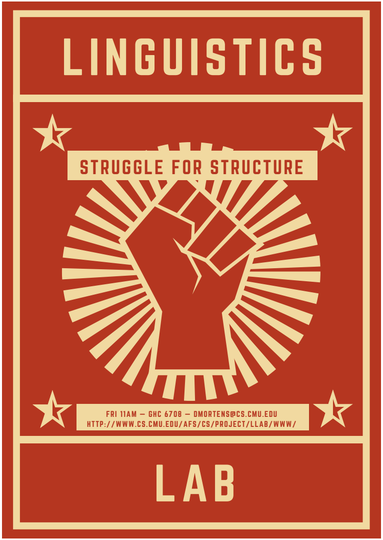
The constituents given in italics are “heads”; the others are “dependents”. There is a interesting correlation between these variables:

- In languages with V-O order, heads occur before dependents at well above chance frequency; these languages are called HEAD-INITIAL
- In languages with O-V order, heads occur after dependents at well above chance frequency; these languages are called HEAD-FINAL

Conclusion

Both morphology and syntax are important areas of research that touch on many aspects of language technologies including machine translation. The point of this lecture has been to provide a relevant introduction to these fields rather than to tie them directly to NLP or MT. I hope you will have learned something that you can apply in this course and to your future research.

Shameless Plug


LINGUISTICS
STRUGGLE FOR STRUCTURE
FRI 11AM — GHC 6708 — DMORTENS@CS.CMU.EDU
HTTP://WWW.CS.CMU.EDU/AFS/CS/PROJECT/LLAB/WWW/
LAB

Questions?