100 Things You Always Wanted to Know about Linguistics But Were Afraid to Ask*

Emily M. Bender
University of Washington

NAACL-HLT 2012
3 June 2012
Montréal, Québec

*for fear of being told 1,000 more
Outline

• Introduction

• Morphology

• Basic Syntax

• Syntactic Complications

• Resources
Linguistics ≠ morphology + syntax

- Structure-based subfields:
  - Phonetics
  - Phonology
  - Morphology
  - Syntax
  - Semantics
  - Pragmatics
  - ...

- Language-and subfields:
  - Sociolinguistics
  - Psycholinguistics
  - Language acquisition (1st, 2nd)
  - Historical linguistics
  - Forensic linguistics
  - Lexicography
  - ...

...
What is morphosyntax?

• The difference between a sentence and bag of words

• The constraints that a language puts on how words can be combined

• ... both in form and in the resulting meaning

• In NLP, we often want extract from a sentence (as part of a text) who did what to whom

• The morphosyntax of a language solves the inverse problem: how to indicate the relationship between the different parts of a sentence

• Different languages do this differently, but there are recurring patterns
My goals for this tutorial

- Provide information about the structure of human languages that is useful in creating NLP systems

- Give a sense of the ways in which languages differ from each other, to support more language-independent NLP systems

- Provide pointers to useful resources to find out more
Your goals for this tutorial

- What kind of applications are you currently (considering) using dependency structures, constituent structures or morphological information in?

- What are you hoping to get from them?
Typological preliminaries

- Languages can be classified “genetically” (by family), areally (by region spoken) or typologically (by grammatical properties).

- These dimensions are distinct, but correlated (cf. Daumé III, 2009).

- Ethnologue.com (as of 4/5/12) lists 6,909 known living languages, distributed across 128 language families, with 1-1,532 languages each.

<table>
<thead>
<tr>
<th>Language</th>
<th>Family</th>
<th>% ACL 2008</th>
<th>% EACL 2009</th>
<th>Other languages in family</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Indo-European</td>
<td>63%</td>
<td>55%</td>
<td>French, Welsh, Gujarati</td>
</tr>
<tr>
<td>German</td>
<td>Indo-European</td>
<td>4%</td>
<td>7%</td>
<td>Latvian, Ukrainian, Farsi</td>
</tr>
<tr>
<td>Chinese</td>
<td>Sino-Tibetan</td>
<td>4%</td>
<td>2%</td>
<td>Burmese, Akha</td>
</tr>
<tr>
<td>Arabic</td>
<td>Afro-Asiatic</td>
<td>3%</td>
<td>1%</td>
<td>Hebrew, Somali, Coptic</td>
</tr>
</tbody>
</table>

(Lewis 2009; Bender 2011)
Morphology: Overview

- Morphology: The study of the internal structure of words
- Morphotactics: What morphemes are allowed and in what order
- Morphophonology: How the form of morphemes is conditioned by other morphemes they combine with
- Morphosyntax: How the morphemes in a word affect its combinatoric potential
Morphology

• Morphemes: The smallest meaningful units of language, i.e., smallest pairings of form and meaning

the smallest meaningful units of language

• Form is prototypically a sequence of phones. However:
  
  • The phones don’t have to be contiguous
  
  • The form doesn’t have to be phones: tonal morphemes, signed languages, non-phone-based writing systems
  
  • The form can vary with the linguistic context (cf. morphophonology)
  
  • The form can be null (if it contrasts with non-null)
Example of non-contiguous morphemes

- Semitic root & pattern morphology

<table>
<thead>
<tr>
<th>Root</th>
<th>Pattern</th>
<th>POS</th>
<th>Word</th>
<th>gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ktb</td>
<td>CaCaC</td>
<td>(v)</td>
<td>katav</td>
<td>‘write’</td>
</tr>
<tr>
<td>ktb</td>
<td>hiCCiC</td>
<td>(v)</td>
<td>hixtiv</td>
<td>‘dictate’</td>
</tr>
<tr>
<td>ktb</td>
<td>miCCaC</td>
<td>(n)</td>
<td>mixtav</td>
<td>‘a letter’</td>
</tr>
<tr>
<td>ktb</td>
<td>CCaC</td>
<td>(n)</td>
<td>ktav</td>
<td>‘writing, alphabet’</td>
</tr>
</tbody>
</table>

Hebrew [heb] (Arad, 2005: 27)
Example of tonal morpheme

- Marker of tense/aspect in Lango (Nilo-Saharan, Uganda):

<table>
<thead>
<tr>
<th>Form</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>àgíkò</td>
<td>‘I stop (something), perfective’</td>
</tr>
<tr>
<td>àgíkô</td>
<td>‘I stop (something), habitual’</td>
</tr>
<tr>
<td>àgíkkò</td>
<td>‘I stop (something), progressive’</td>
</tr>
</tbody>
</table>

Lango [laj] (Noonan, 1992: 92)
Morphology

• Morphemes: The smallest meaningful units of language, i.e., smallest pairings of form and meaning

• The meaning part of that form-meaning pairing can also be less than straightforward.
  
  • *Roots* convey core lexical meaning
  
  • *Derivational affixes* can change lexical meaning
    
    • But root+derivational affix combinations can also have idiosyncratic meanings
  
  • *Inflectional affixes* add syntactically or semantically relevant features
    
    • e.g.: case-marking affixes arguably don’t convey meaning directly

• Morphemes can be ambiguous (alternatively: underspecified)
### Examples of inflectional morphemes (English)

<table>
<thead>
<tr>
<th>Affix</th>
<th>morphosyntactic effect</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>-s</td>
<td>NUMBER: plural</td>
<td>cat → cats</td>
</tr>
<tr>
<td>-s</td>
<td>TENSE: present, SUBJ: 3sg</td>
<td>jump → jumps</td>
</tr>
<tr>
<td>-ed</td>
<td>TENSE: past</td>
<td>jump → jumped</td>
</tr>
<tr>
<td>-ed/-en</td>
<td>ASPECT: perfective</td>
<td>eat → eaten</td>
</tr>
<tr>
<td>-ing</td>
<td>ASPECT: progressive</td>
<td>jump → jumping</td>
</tr>
<tr>
<td>-er</td>
<td>comparative</td>
<td>small → smaller</td>
</tr>
<tr>
<td>-est</td>
<td>superlative</td>
<td>small → smallest</td>
</tr>
</tbody>
</table>

(O’Grady et al, 2010:132)
Examples of derivational morphemes (English)

<table>
<thead>
<tr>
<th>Affix</th>
<th>POS change</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>-able</td>
<td>V → A</td>
<td>fixable, doable, understandable</td>
</tr>
<tr>
<td>-ive</td>
<td>V → A</td>
<td>assertive, impressive, restrictive</td>
</tr>
<tr>
<td>-al</td>
<td>V → N</td>
<td>refusal, disposal, recital</td>
</tr>
<tr>
<td>-er</td>
<td>V → N</td>
<td>teacher, worker</td>
</tr>
<tr>
<td>-ment</td>
<td>V → N</td>
<td>adjournment, treatment, amazement</td>
</tr>
<tr>
<td>-dom</td>
<td>N → N</td>
<td>kingdom, fiefdom</td>
</tr>
<tr>
<td>-less</td>
<td>N → A</td>
<td>penniless, brainless</td>
</tr>
<tr>
<td>-ic</td>
<td>N → A</td>
<td>cubic, optimistic</td>
</tr>
<tr>
<td>-ize</td>
<td>N → A</td>
<td>hospitalize, vaporize</td>
</tr>
<tr>
<td>-ize</td>
<td>A → V</td>
<td>modernize, nationalize</td>
</tr>
<tr>
<td>-ness</td>
<td>A → N</td>
<td>happiness, sadness</td>
</tr>
<tr>
<td>anti-</td>
<td>N → N</td>
<td>antihero, antidepressant</td>
</tr>
<tr>
<td>de-</td>
<td>V → V</td>
<td>deactivate, demystify</td>
</tr>
<tr>
<td>un-</td>
<td>V → V</td>
<td>untie, unlock, undo</td>
</tr>
<tr>
<td>un-</td>
<td>A → A</td>
<td>unhappy, unfair, unintelligent</td>
</tr>
</tbody>
</table>

(O’Grady et al, 2010:124)
What is a ‘word’?

• The notion of ‘word’ can be contentious in many languages.

• ... if there isn’t an orthographic tradition establishing one notion of word boundaries (cf. Japanese, Chinese, Thai); and even if there is:

  • Penn Treebank (Marcus et al 1993) segments don’t into do + n’t, but Zwicky & Pullum (1983) show that n’t is an affix

  • Romance languages separate so-called clitics from the verb root with white space, but Miller & Sag (1997) show that they are affixes

  [ je ne te l’aï ] pas dit
  [ 1sg.SUBJ NEG 2sg.IND.OBJ 3sg.DIR.OBJ have ] NEG said

  ‘I haven’t told you it.’ [fra]
What is a ‘word’?

• Is this one of those theoretical issues that don’t matter to NLPers?

• Maybe not: Words and morphemes are subject to different ordering principles.

• Generally: Words can be separated from the other words they are ordered with respect to by e.g., modifiers; morphemes appear in a stricter sequence.

• On the other hand, the distinction isn’t clear partially because of language change:
  
  • Words with relatively free position > words with fixed position > clitic > bound morpheme (Hopper and Traugott 2003)

  • Clitic: A linguistic element which is syntactically independent but phonologically dependent. Examples: English the and (possessive) ’s

  the person standing by the river’s edge/coat
Crosslinguistic variation in morphology

- Analytic v. synthetic: How many morphemes per word
- Prefixing v. suffixing: Do most affixes precede or follow the root
- Agglutinating v. fusional: How easily separated are the morphemes within a word

<table>
<thead>
<tr>
<th>Language</th>
<th>Index of synthesis</th>
<th>Language</th>
<th>Index of synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnamese</td>
<td>1.06</td>
<td>Swahili</td>
<td>2.55</td>
</tr>
<tr>
<td>Yoruba</td>
<td>1.09</td>
<td>Turkish</td>
<td>2.86</td>
</tr>
<tr>
<td>English</td>
<td>1.68</td>
<td>Russian</td>
<td>3.33</td>
</tr>
<tr>
<td>Old English</td>
<td>2.12</td>
<td>Inuit (Eskimo)</td>
<td>3.72</td>
</tr>
</tbody>
</table>

(Karlsson, 998)
Crosslinguistic variation in morphology

- Analytic v. synthetic: How many morphemes per word
- Prefixing v. suffixing: Do most affixes precede or follow the root
- Agglutinating v. fusional: How easily separated are the morphemes within a word

(Dryer 2011)
Morphology: Overview

• Morphology: The study of the internal structure of words

• Morphotactics: What morphemes are allowed and in what order

• Morphophonology: How the form of morphemes is conditioned by other morphemes they combine with

• Morphosyntax: How the morphemes in a word affect its combinatoric potential
Morphophonology: changes in form in morphemes in context

- Phonologically conditioned: The triggering context is in the form

- Morphologically conditioned: The triggering context is in lexical identity of some element

- Suppletion: Wholly different form for stem+affix

- Words can’t always be neatly divided into substrings representing invariant morphemes.
Phonologically conditioned allomorphy example

- Vowel harmony in Turkish ([tur], Altaic)

-dAn: ablative

<table>
<thead>
<tr>
<th>Turkish (from Goeksel and Kerslake 2005:23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hava-dan</td>
</tr>
<tr>
<td>kız-dan</td>
</tr>
<tr>
<td>yol-dan</td>
</tr>
<tr>
<td>şun-dan</td>
</tr>
</tbody>
</table>

üz-ül-dü-nüz ‘You became sad.’
Morphologically conditioned allomorphy example

- French verb classes: -er, -ir, -re

<table>
<thead>
<tr>
<th>Infinitival form</th>
<th>-er</th>
<th>-ir</th>
<th>-re</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloss</td>
<td>‘eat’</td>
<td>‘choose’</td>
<td>‘descend’</td>
</tr>
<tr>
<td>1sg</td>
<td>mang+e</td>
<td>chois+is</td>
<td>descend+s</td>
</tr>
<tr>
<td>2sg</td>
<td>mang+es</td>
<td>chois+is</td>
<td>descend+s</td>
</tr>
<tr>
<td>3sg</td>
<td>mang+e</td>
<td>chois+it</td>
<td>descend+</td>
</tr>
<tr>
<td>1pl</td>
<td>mang+eons</td>
<td>chois+issons</td>
<td>descend+ons</td>
</tr>
<tr>
<td>2pl</td>
<td>mang+ez</td>
<td>chois+issez</td>
<td>descend+ez</td>
</tr>
<tr>
<td>3pl</td>
<td>mang+ent</td>
<td>chois+issent</td>
<td>descend+ent</td>
</tr>
</tbody>
</table>
Stem changes conditioned by affixes

- Finnish ([fin], Uralic) assimilation across morpheme boundaries:

  halut-a  ‘want-INF’  halus-i  ‘want-PAST’
  tilat-a  ‘order-INF’  tilas-i  ‘order-PAST’
  äiti  ‘mother’

(Burzio 2011:2092)
Suppletion examples

• English: go/went

• English: good/better/best

• French: aller ‘go’/ir-ai ‘I will go’
Approximations of morphology

• Many NLP systems approximate morphology by creating features from suffix substrings of up to N characters.

• Under what circumstances will this work okay?

• Why/when might it not work so well?
Morphology: Overview

• Morphology: The study of the internal structure of words

• Morphotactics: What morphemes are allowed and in what order

• Morphophonology: How the form of morphemes is conditioned by other morphemes they combine with

• Morphosyntax: How the morphemes in a word affect its combinatoric potential
Information provided by inflectional morphemes: Tense, Aspect, Mood (on verbs, adjectives)

- Tense/aspect/mood on verbs (and sometimes adjectives): Temporal information about events
  - Tense: (Roughly) how the time of the described event relates to the speech time
  - Aspect: (Roughly) how the internal temporal structure of the described event is portrayed
  - Mood: (Roughly) speakers attitude towards sentential content and/or illocutionary force
- Languages vary in how many values they grammaticize in each of tense/aspect/mood
Sample systems/values

- Tense: past/non-past, future/non-future, past/present/future, also remote past, remote future, and varying degrees of same

- Aspect: perfect/imperfect, also: habitual, inceptive, inchoative, cessative, resumptive, punctual, iterative, experiential, ...

- Tense+aspect: perfective (completion of event prior to some reference time)

- Mood: indicative, conditional, optative, imperative, irrealis, ...
Information marked by inflectional morphemes: Person, number, gender (on nouns)

- Person: Relationship of referent to speech act: speaker, addressee, other
  - 1st, 2nd, 3rd; sometimes also 4th (!); inclusive/exclusive distinction on 1st person non-singular

- Number: (Roughly) cardinality of set of referents of referring expression
  - sg/pl; sg/dual/pl; sg/dual/paucal/pl

- Gender/noun class: Subcategories of nouns, sometimes related to natural gender, sometimes not
  - m/f, m/f/n, m/f/vegetable/other, ...
Information marked by inflectional morphemes: Case (on nouns)

- Case: Role of NP within a sentence

- Distinctions among core grammatical functions: nominative/accusative; nominative/accusative/dative; ergative/absolutive

- More elaborate case systems mark different kinds of adjuncts: genitive, locative, ablative, instrumental, adessive, inessive, ...
Information marked by inflectional morphemes:

Other

- Negation: 396/1159 (34%) languages sampled by Dryer (2011) mark sentential negation with an affix.

- Evidentiality: Speaker’s confidence in a statement and source of evidence; de Haan (2011) finds some grammaticized marking of evidentiality in 237/418 (57%) of languages sampled. Most use affixes for this purpose.

- Honorifics: Speaker’s relationship to addressee/referent

- Definiteness: Referent’s relationship to common ground

- Possessives: Marked on possessor, possessed or both
Information marked by inflectional morphemes: Agreement

- Inflectional categories can be marked on multiple elements of a sentence.
- Usually considered to belong to one element; marking on others is agreement.
  - Category might not be marked on the word it belongs to.
- Verbs commonly agree in person/number/gender with subjects, sometimes other arguments.
- Determiners and adjectives commonly agree with nouns in person/number/gender and case.
- Agreement can be with a feature that is inherent (e.g., gender, person) or added via inflection (e.g., number).
Agreement example

- Bantu languages have many noun classes, and both verbs and nominal dependents agree with nouns in those classes:

Swahili [swa]:

Wa-tu wa-zuri wa-wili wa-le wa-me-anguka.
NC1p-person NC1p-good NC1p-two NC1p-those NC1p-PastP-fall

‘Those two good people have fallen.’

(Hargus, class notes)
Why might we care?
Hohensee & Bender 2012 preview

• Previous work incorporating morphology into language-independent dependency parsing algorithms didn’t model agreement

• Hohensee & Bender propose a series of features that capture agreement between head & dependent in any morphological feature, discarding the actual value

  • Serves as a kind of natural (nearly) non-lossy back-off

  • Improves performance across languages/treebanks with any morphological information, with far fewer features than baseline (MSTParser)
Why might we care?
Hohensee & Bender 2012 preview

- Error reduction wrt to no morphological features original (MSTParser; McDonald et al, 2006) configuration, new agreement features and both:
Morphology: Overview

- **Morphology**: The study of the internal structure of words
- **Morphotactics**: What morphemes are allowed and in what order
- **Morphophonology**: How the form of morphemes is conditioned by other morphemes they combine with
- **Morphosyntax**: How the morphemes in a word affect its combinatoric potential
Outline

• Introduction

• Morphology

• Basic Syntax

• Syntactic Complications

• Resources
Functions of syntax

• Constraints on possible sentences (grammaticality)

• Scaffolding for semantic composition

• Both together: modeling grammaticality constrains ambiguity
Syntax: Overview

- What’s syntax for?

- Parts of speech:
  - combinatoric potential of words

- Grammatical functions:
  - scaffolding

- Deep dependencies v. surface syntax:
  - more elaborate aspects of scaffolding
Parts of speech

- Grammatical notion, defined in terms of distributional characteristics or functionally

- Group words according to substitution classes (syntax) and affix sets (morphology)

- Major categories: noun, verb, adjective, adverb

- Other categories: adposition, determiner/article, conjunction, number names, numeral classifier, ‘particle’, ...

- No one universal set, even among the major categories
Functional generalizations (Hengeveld 1992)

- Noun: Head (non-optional element) of a referring expression
- Verb: Can only be used predicatively
- Adjective: Non-head (modifier, optional) element of a referring expression
- Adverb: Non-head (modifier) of predicate
What is part of speech useful for?

- Coarse-grained WSD
- Default lexical properties for unknown words in parsing
- Other?
Grammatical functions

• Heads v. dependents

• Arguments v. adjuncts

• Different types of arguments (grammatical roles)
Heads v. dependents

• Heads:
  
  • Required element of a constituent
  
  • Determine its internal structure (what else is required)
  
  • Determine its external distribution (where it can appear)

• Dependents (after Kay 2005):
  
  • Arguments: Required by the head; complete the meaning of a predicate
  
  • Adjuncts: Optional; refine the meaning of a complete predication
Heads v. dependents examples: N as head of NP

• Required element:
  
  [The cat on the mat] is sleeping.  
  The cat is sleeping.  
  *The on the mat is sleeping.  
  *Cat on the mat is sleeping.

• Determines what else can appear:
  
  The book about syntax is heavy.  
  *The cat about syntax is heavy.

• Determines external distribution:
  
  The book about cat/cats is heavy.  
  *The book about cat/cats are heavy.
Arguments v. adjuncts

- Arguments can in principle be predicted from the lexical identity of the head
  - In many (all?) languages, (some) arguments can be left unexpressed
  - The number of semantic arguments provided for by a head is a fundamental semantic property

- Adjuncts
  - Not required by heads
  - Generally can iterate
Syntax/semantics mismatches

• Syntactically, modifiers are dependents

• Semantically, they introduce predicates which take the heads as arguments

The book about syntax is heavy.
∃x book(x) about(x,y) syntax(y) heavy(x)
Tests distinguishing arguments from adjuncts

- Obligatoriness: If it’s required, it’s an argument

- Entailment: If X Ved (NP) PP does not entail X did something PP, then the PP is a complement

  - Pat relied on Chris does not entail Pat did something on Chris

  - Pat put nuts in a cup does not entail Pat did something in a cup

  - Pat slept until noon does entail Pat did something until noon

  - Pat ate lunch in Montreal does entail Pat did something in Montreal
Arguments v. adjuncts in PropBank

• Framing guidelines take a pragmatic approach to distinguishing arguments (ArgNs) from adjuncts (ArgMs):

  • “A semantic role is being marked as an argument, if it frequently occurs in a corpus and is specific to a particular class of verbs.” (http://verbs.colorado.edu/~mpalmer/projects/ace/FramingGuidelines.pdf)
Types of adjuncts

- Single words: yesterday, blue, very
- Phrasal constituents: on the bus, very elaborate
- Clausal modifiers: while Kim was reading a book
Types of adjuncts (syntactic)

- Adnominal modifiers: adjectives, adpositional phrases (PPs), relative clauses

- Adverbial modifiers: adpositional phrases (PPs), adverbs, subordinate clauses, discourse markers
Types of adverbial adjuncts (semantic; PropBank)
(http://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf)

- Directional: to the store
- Locative: at the store
- Manner: with haste
- Temporal: yesterday, frequently
- Extent: more, further, 25%
- Reciprocal: together, jointly, both
- Secondary predicates: as a director
- Purpose clauses: in order to
- Cause clauses: as a result of
- Discourse markers: but, vocatives
- Negation: not, never, no longer
- Other: only, even, possibly, fortunately
One and the same phrase can be adjunct or argument, depending on the context

- The potential to be a modifier is inherent to the syntax of a constituent
  
  Kim swam Tuesday/for two days/*two days.

- Just about anything can be an argument, for some head
  
  Kim put the book on the table.
  *Kim put the book.
  Kim found the book on the table.
  Kim found the book.

  *That doesn’t bode.
  That doesn’t bode well.
Types of arguments

• Subject v. complements

  • Whether subject exists as a GR in all languages is a matter of debate

  • Subjects = distinguished argument, which may be the only one to display properties related to agreement, relativization, control, coordination, word order

• Obliqueness: Arguments can generally be arranged in order of centrality to the event

  Subject > direct object > indirect/2nd object > oblique
Types of arguments

- Clauses can also be arguments (subjects or complements)

  - Finite, closed clausal arguments
    
    Kim believes [(that) Sandy left.]
    [That Sandy left] surprised Kim.

  - Non-finite, controlled clausal arguments
    
    Kim expects Sandy [to leave]
    Kim tried [to leave]

  - Non-finite, non-controlled clausal arguments
    
    To leave now would be a bad idea.
Argument types in the Penn Treebank

- **SBJ (surface subject):** *Kim went to the store.*
- **LGS (logical subject):** *The picture was taken by Kim.*
- **PRD (non-verbal predicate):** *Kim left and Sandy did so too.*
- **PUT (locative complement of put):** *Kim put the book on the table.*
- **TPC (“topicalized”):** *Bagels we think Kim likes.*
- **VOC (vocatives):** *Kim, you should put the book on the table.*

(http://bulba.sdsu.edu/jeanette/thesis/PennTags.html)
Argument types in the Stanford dependency format (de Marneff and Manning, 2011)

- nsubj (nominal subject): *Kim took the picture.*

- nsubjpass (passive nsubj): *The picture was taken by Kim.*

- csubj (clausal subject): *What she said makes sense.*

- csubjpass (passive csubj): *That she lied was suspected by everyone.*

- xsubj (controlling subject): *Kim likes to take pictures.*

- agent (in passives): *The picture was taken by Kim.*

- expl (existential there): *There is a ghost in the room.*

- dobj (direct object): *They win the lottery.*

- ccomp (clausal complement): *He says that you like to swim.*

- xcomp (controlled clause): *You like to swim.*

- iobj (indirect object): *She gave me a raise.*

- pcomp (prep’s comp): *They heard about you missing class.*

- pobj (obj of P): *The sat on the chair.*
Syntactic v. semantic arguments

- Syntactic and semantic arguments aren’t the same
- ... though they often stand in regular relations to each other
- For many applications, it’s not the surface (syntactic) relations, but the deep (semantic) dependencies that matter.
  - Examples?
What are grammatical functions good for?

• Syntactic phenomena differentiating arguments are sensitive to grammatical function

• Lexical items map semantic roles to grammatical functions

The dog scared Kim.
Kim feared the dog.
Kim loaded the wagon with hay.
Kim loaded the hay onto the wagon.
What are grammatical functions good for?

• There can be mismatches:
  
  • Some syntactic phenomena rearrange the mapping (e.g., passive)

        Kim took the picture./The picture was taken by Kim.

  • Some syntactic dependents don’t fill a semantic role

        Kim expects it to bother Sandy that Pat left.
        Kim expects Pat to leave.

  • Some syntactic dependents aren’t realized locally

        Kim continues to be likely to be easy to talk to.
What are grammatical functions good for?

• The mapping of syntactic constituents to semantic argument positions is mediated by both:
  • Grammatical functions
  • The lexical properties of the selecting predicate

• Identifying the grammatical function of a constituent can help us understand its semantic role with respect to the head, provided we also know:
  • The mapping provided by the head
    • and any intervening heads (e.g., raising predicates)
  • Whether the clause is passive, etc
Grammatical function identifying phenomena

• Word order (in fixed word order languages): In prototypical English clauses, the subject is the only argument preceding the verb

• Agreement (head marking): Morphological marking on the head reflecting properties of the constituent(s) filling particular argument slots

• Case (dependent marking): Morphological marking on the dependent indicating what role it plays in the sentence
Grammatical function identifying phenomena examples

- **Word order (English):**
  
  Kim saw Sandy $\neq$ Sandy saw Kim

- **Agreement (Swahili [swa; Niger-Congo]):**
  
  Jana ni-li-mw-on-a m-levi
  yesterday SA.1S-PAST-OA.NCL1-see-IND NCL1-drunkard
  
  ‘Yesterday I saw (that) drunkard.’ (Ud Deen, 2006:233)

- **Case (Wambaya [wmb; Australian]):**
  
  Ngaragana-nguja ngiy-a gujinganjanga-ni jiyawu ngabulu.
grog-PROP.IV.ACC 3.SG.NM.A-PST mother.II.ERG give milk.IV.ACC
  
  ‘(His) mother gave (him) milk with grog in it.’ (Nordlinger 1998:223)
Grammatical function identifying phenomena

- Word order (in fixed word order languages): In prototypical English clauses, the subject is the only argument preceding the verb.

- Agreement (head marking): Morphological marking on the head reflecting properties of the constituent(s) filling particular argument slots.

- Case (dependent marking): Morphological marking on the dependent indicating what role it plays in the sentence.

- Languages tend to prefer one or the other of these; agreement (head-marking) is more common in the world’s languages (Nichols 1986).
Mismatches: Passive

• Passive is a grammatical process which demotes the subject to oblique status, making room for the next most prominent argument to appear as the subject.

• Note that this changes the semantic role associated with subject position for a given verb.

  Kim saw Sandy.
  Sandy was seen (by Kim).

• In English, only transitive verbs allow passive (and even then, not all transitives).

• Other languages (including German, Dutch, Turkish, Shona [sna; Niger-Congo]) allow passives of intransitives, too. (Keenan and Dryer, 2007)
Mismatches: Dative shift

• Another example of a grammatical phenomenon affecting the mapping between syntactic and semantic arguments

Kim gave a book to Sandy.
Kim gave Sandy a book.

• Interacts with passive:

A book was given to Sandy (by Kim).
Sandy was given a book (by Kim).
Mismatches in arity

• Syntactic arguments without any semantic role: expletives
  It seems that Sandy left.
  It turns out that Sandy was right.
  I take it that Sandy left.
  Sandy is living it up.
  Kim and Sandy battled it out.
  (Postal and Pullum, 1988)

• Syntactic arguments without any local semantic role: raising
  Sandy expected Kim to laugh.
  Sandy continued to laugh.

• Syntactic arguments which play two roles: control
  Sandy persuaded Kim to leave.
  Sandy tried to laugh.
Deep dependencies v. surface syntax: Putting it all together

- Which NP refers to the patient (undergoer, deep object) of *interview*?

  Sandy appeared to have been persuaded by Kim to be interviewed by the reporter.

- What syntactic processes are involved?

- Which lexical items have arity mismatches?
Syntax: Overview

• What’s syntax for?

• Parts of speech:
  • combinatoric potential of words

• Grammatical functions:
  • scaffolding

• Deep dependencies v. surface syntax:
  • more elaborate aspects of scaffolding
Outline

• Introduction

• Morphology

• Basic Syntax

• Syntactic Complications

• Resources
Syntactic complications: Overview

• Long-distance dependencies

• Semantically empty words

• Argument drop
Long-distance dependencies

• Some languages allow arguments and/or adjuncts to appear separated from their selecting head, even in a different clause

• Typical examples:

  • *wh* questions: *What does Sandy think Kim likes to eat _?*

  • *relative clauses: This is the dish which Sandy thinks Kim likes to eat _.*

  • “topicalization”: *This dish Sandy thinks Kim likes to eat _.*

  • *easy-adjectives: This dish is easy to imagine Kim likes to eat _.*
Not just a fun corner case for linguists!

<table>
<thead>
<tr>
<th>Construction</th>
<th>WSJ</th>
<th>Brown</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obj relative clause</td>
<td>2.3</td>
<td>1.1</td>
<td>1.4</td>
</tr>
<tr>
<td>Obj reduced relative clause</td>
<td>2.7</td>
<td>2.8</td>
<td>2.8</td>
</tr>
<tr>
<td>Subj relative clause</td>
<td>10.1</td>
<td>5.7</td>
<td>7.4</td>
</tr>
<tr>
<td>Free relative</td>
<td>2.6</td>
<td>0.9</td>
<td>1.3</td>
</tr>
<tr>
<td>Right node raising</td>
<td>2.2</td>
<td>0.9</td>
<td>1.2</td>
</tr>
<tr>
<td>Subj extraction from embedded clause</td>
<td>2.0</td>
<td>0.3</td>
<td>0.4</td>
</tr>
</tbody>
</table>

(Rimell et al, 2009)
Semantically empty words

- Don’t contribute lexical content
- Do serve as syntactic “glue”
- Sometimes contribute features to the semantics
- Vary across languages
- Give rise to mis-matches in aligned bitexts

Examples from English:
- complementizers: that, to
- expletives: there, it
- auxiliaries: do, be, will, have

Examples from Japanese:
- case particles: が、を、に
Dependency parsers and semantically empty words

- Stanford:

```
(ROOT
 (S
  (NP (NNP Kim))
  (VP (MD will)
    (VP (VB have)
      (VP (VBN been)
        (VP (VBG expecting)
          (S
            (NP (NNP Sandy))
            (VP (TO to)
              (VP (VB leave)))))))))

Typed dependencies
```

- ERG:
Argument drop

- Lexical predicates introduce expectations for a certain (fixed, given a word sense) number of arguments

- Those arguments aren’t always overtly realized

- Permissible argument drop varies by word class and by language
Argument drop, aka null instantiation (Fillmore 1986)

- Definite null instantiation: Referent is recoverable from discourse context
- Indefinite null instantiation: Referent is non-specific/not recoverable from discourse context
- Constructional null instantiation: Referent is determined by syntactic context (imperatives, control)
Argument drop, aka null instantiation (Fillmore 1986)

• Definite null instantiation:

  She promised.        They agreed.
  I tried.             She found out.
  When did she leave?  I forgot.

• Indefinite null instantiation:

  I spent the afternoon baking.
  We already ate.
  What happened to my sandwich? *Fido ate.
Argument drop, aka null instantiation (Fillmore 1986)

- Lexically licensed: Possibility of an argument going missing depends on the lexical identity of the head (eat v. devour)

  Fido ate.  *Fido devoured.
  She promised.  *She pledged/vowed/guaranteed.
  They accepted.  *They authorized.
  She found out.  *She discovered.
  He lost the race/his wallet.  He lost.

- Systematic: Subjects (e.g., in Spanish) or any argument (e.g., Japanese) can be dropped, if supported by the discourse context
Argument drop: Why does it matter?

• MT: Identifying dropped arguments in the source language that should be overt pronouns in the target

• Reference resolution: Dropped arguments participate in coreference chains; a sufficiently salient argument can be “mentioned” via dropped arguments in successive clauses

• Dependency triples: Dropped arguments participate in dependencies, and (when resolved via their antecedents) can add valuable information to co-occurrence patterns
Syntactic complications: Overview

• Long-distance dependencies

• Semantically empty words

• Argument drop
Outline

- Introduction
- Morphology
- Basic Syntax
- Syntactic Complications
- Resources
Resources: Typology


- Typological properties of languages: 76,492 data points, 2,678 languages, 192 properties

- Adapt NLP systems to languages based on typological properties

- Expand NLP systems to handle more languages based on understanding of features
Resources: Morphological analyzers

- Map surface forms (e.g., standard orthography) to regularized strings of morphemes or morphological features

- Useful for:
  - Machine translation into morphologically complex languages (Toutanova et al 2008)
  - Handling morphologically-induced data sparsity (e.g., through Factored Language Models, Bilmes and Kirchhoff 2003)
Resources: Syntax, beyond the well-known parsers

• The English Resource Grammar (Flickinger, 2011), used with DELPH-IN parsing algorithms (www.delph-in.net), provides linguistically-motivated parses mapping to deep dependencies


• The Grammar Matrix (Bender et al 2002, 2010) supports the creation of new grammars in the style of the ERG
Summary/reflection:
My goals for this tutorial

• Provide information about the structure of human languages that is useful in creating NLP systems

• Give a sense of the ways in which languages differ from each other, to support more language-independent NLP systems

• Provide pointers to useful resources to find out more
Summary/reflection

• Topics covered today:
  • Morphology (incl morphotactics, morphophonology, morphosyntax)
  • Basic syntax
  • Some syntactic complications

• In what ways will this information be useful for NLP?

• What (if anything) was the most surprising thing (of the 100)?

• What do you want to know more about?
References


Haan, F. de. (2011). Coding of evidentiality. In M. S. Dryer & M. Haspelmath (Eds.),
The world atlas of language structures online. Munich: Max Planck Digital Library. Available from http://wals.info/chapter/78


eration models to machine translation. In *Proceedings of acl-08: Hlt* (pp. 514–522). Columbus, Ohio: Association for Computational Linguistics. Available from [http://www.aclweb.org/anthology/P/P08/P08-1059](http://www.aclweb.org/anthology/P/P08/P08-1059)
