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 ⚙2

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

- ... both in form and in the resulting meaning ⚙3

- 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 ⚙4

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)

5

- 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

6

Language	Family	% ACL 2008	% EACL 2009	Other languages in family
English	Indo-European	63%	55%	French, Welsh, Gujarati
German	Indo-European	4%	7%	Latvian, Ukranian, Farsi
Chinese	Sino-Tibetan	4%	2%	Burmese, Akha
Arabic	Afro-Asiatic	3%	1%	Hebrew, Somali, Coptic

(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 small+est mean+ing+ful unit+s of language

- Form is prototypically a sequence of phones. However:
  - The phones don't have to be contiguous

    #7
  - The form doesn't have to be phones: tonal morphemes, signed languages, non-phone-based writing systems

    #8
  - The form can vary with the linguistic context (cf. morphophonology)

    #9
  - The form can be null (if it contrasts with non-null)

    #10

Example of non-contiguous morphemes

- Semitic root & pattern morphology

Root	Pattern	POS	Word	gloss
ktb	CaCaC	(v)	katav	‘write’
ktb	hiCCiC	(v)	hixtiv	‘dictate’
ktb	miCCaC	(n)	mixtav	‘a letter’
ktb	CCaC	(n)	ktav	‘writing, alphabet’

Hebrew [heb] (Arad, 2005: 27)

Example of tonal morpheme

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

Form	Gloss
àgíkò	‘I stop (something), perfective’
àgíkô	‘I stop (something), habitual’
àgíkkò	‘I stop (something), progressive’

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 #11
  - *Derivational affixes* can change lexical meaning #12
    - But root+derivational affix combinations can also have idiosyncratic meanings
  - *Inflectional affixes* add syntactically or semantically relevant features #13
    - e.g.: case-marking affixes arguably don’t convey meaning directly
  - Morphemes can be ambiguous (alternatively: underspecified)

Examples of inflectional morphemes (English)

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

(O'Grady et al, 2010:132)

Examples of derivational morphemes (English)

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

(O’Grady et al, 2010:124)

What is a ‘word’?

- The notion of ‘word’ can be contentious in many languages. #14
- ... 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’	ai	]	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

#15

- 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)

#16

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

#17

the person standing by the river’s edge/coat

Crosslinguistic variation in morphology

- Analytic v. synthetic: How many morphemes per word #18

- Prefixing v. suffixing: Do most affixes precede or follow the root

- Agglutinating v. fusional: How easily separated are the morphemes within a word

Languge	Index of synthesis	Language	Index of synthesis
Vietnamese	1.06	Swahili	2.55
Yoruba	1.09	Turkish	2.86
English	1.68	Russian	3.33
Old English	2.12	Inuit (Eskimo)	3.72

(Karlsson, 998)

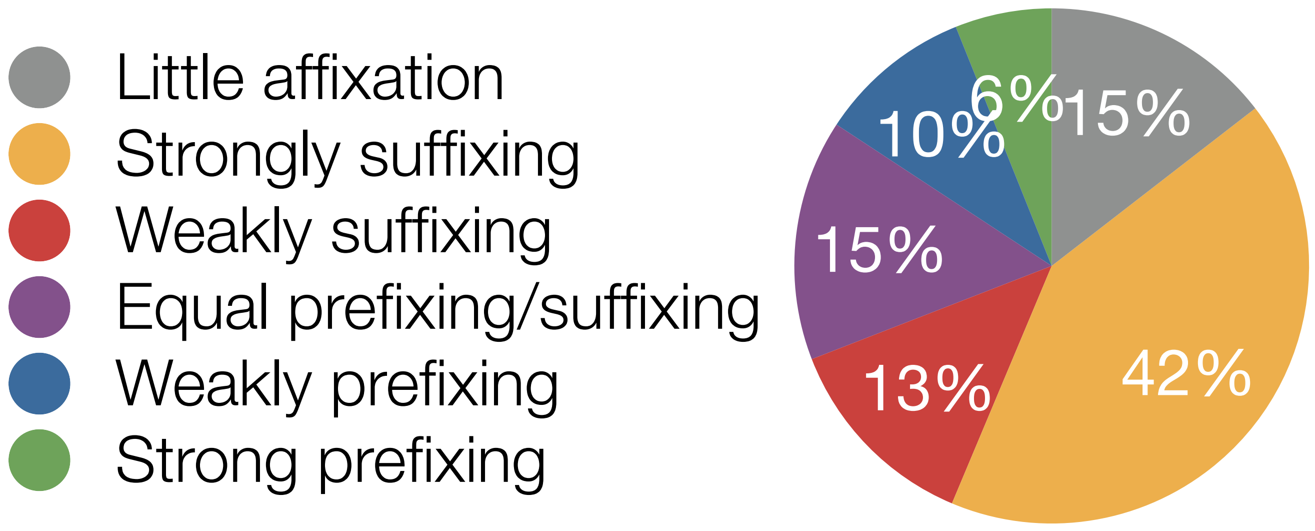
Crosslinguistic variation in morphology

- Analytic v. synthetic: How many morphemes per word
- Prefixing v. suffixing: Do most affixes precede or follow the root

#19

- Agglutinating v. fusional: How easily separated are the morphemes within a word

#20



(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

#21

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

#22

- Suppletion: Wholly different form for stem+affix

#23

    - Words can't always be neatly divided into substrings representing invariant morphemes.

Phonologically conditioned allomorphy example

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

-dAn: ablative

hava-dan	‘from the air’	ev-den	‘from the house’
kız-dan	‘from the girl’	biz-den	‘from us’
yol-dan	‘by the road’	göl-den	‘from the lake’
şun-dan	‘of this’	tür-den	‘of the type’

(Göskel and Kerslake 2005:23)

üz-ül-dü-nüz ‘You became sad.’

Morphologically conditioned allomorphy example

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

	-er	-ir	-re
Infinitival form	manger	choisir	descendre
Gloss	‘eat’	‘choose’	‘descend’
1sg	mang+e	chois+is	descend+s
2sg	mang+es	chois+is	descend+s
3sg	mang+e	chois+it	descend+
1pl	mang+eons	chois+issons	descend+ons
2pl	mang+ez	chois+issez	descend+ez
3pl	mang+ent	chois+issent	descend+ent

Stem changes conditioned by affixes

- Finnish ([fin], Uralic) assibilation 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

      #24

    - Aspect: (Roughly) how the internal temporal structure of the described event is portrayed

      #25

    - Mood: (Roughly) speakers attitude towards sentential content and/or illocutionary force

      #26

- Languages vary in how many values they grammaticize in each of tense/aspect/mood

  #27

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

#28

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

#29

- Gender/noun class: Subcategories of nouns, sometimes related to natural gender, sometimes not
    - m/f, m/f/n, m/f/vegetable/other, ...

#30

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

#31

- More elaborate case systems mark different kinds of adjuncts: genitive, locative, ablative, instrumental, adessive, inessive, ...

#32

Information marked by inflectional morphemes: Other

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

#33

- 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.

#34

- Honorifics: Speaker's relationship to addressee/referent

#35

- Definiteness: Referent's relationship to common ground

#36

- Possessives: Marked on possessor, possessed or both

#37

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

#38

- Verbs commonly agree in person/number/gender with subjects, sometimes other arguments

#39

- Determiners and adjectives commonly agree with nouns in person/number/gender and case

#40

- Agreement can be with a feature that is inherent (e.g., gender, person) or added via inflection (e.g., number)

#41

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.’

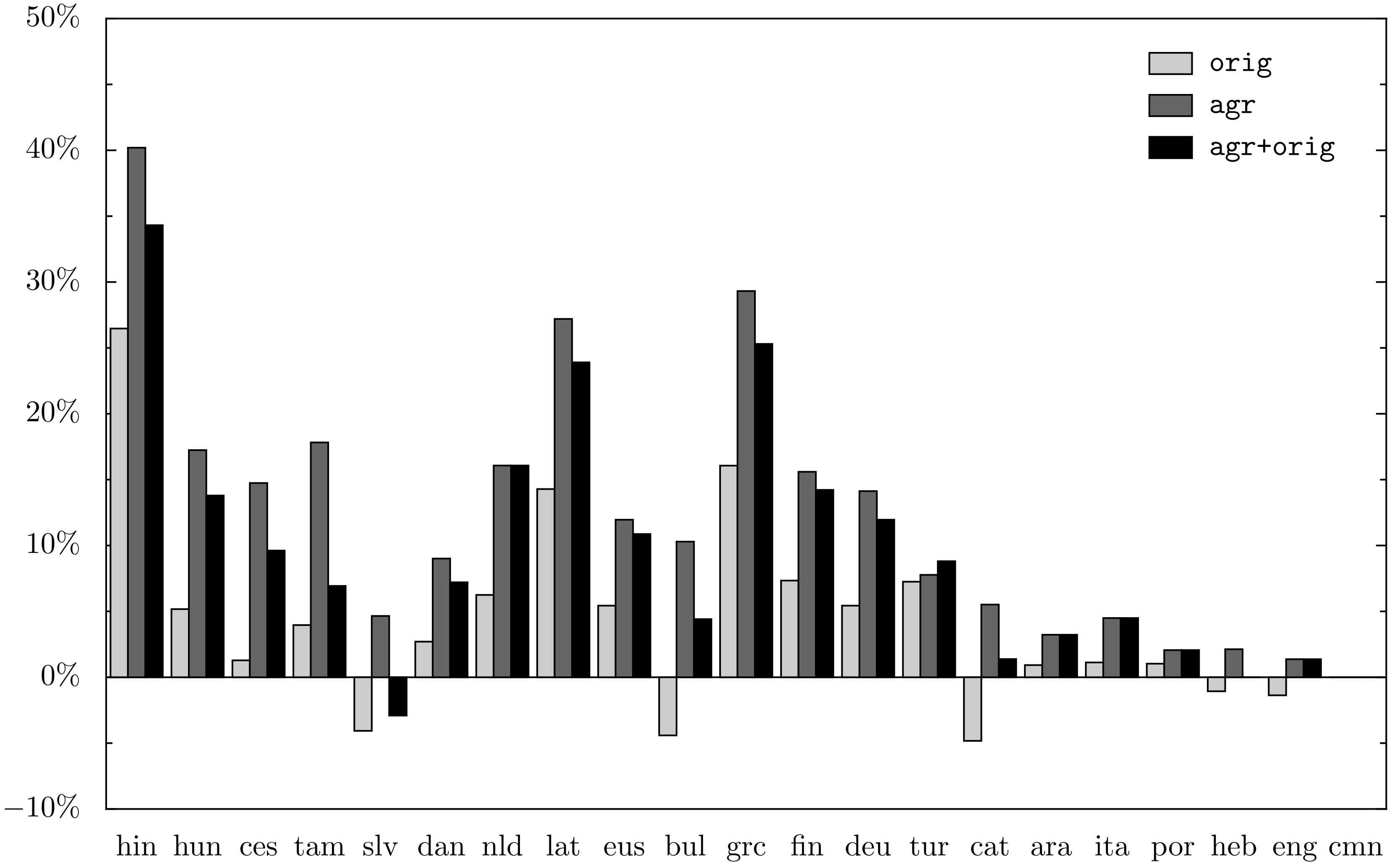
(Hargus, class notes)

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 morhpological 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

#42

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

#43

- Major categories: noun, verb, adjective, adverb

#44

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

#45

- No one universal set, even among the major categories

#46

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

#47

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)

#48

- Dependents (after Kay 2005):
  - Arguments: Required by the head; complete the meaning of a predicate
  - Adjuncts: Optional; refine the meaning of a complete predication

#49

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 #50
  - The number of semantic arguments provided for by a head is a fundamental semantic property #51
- Adjuncts
  - Not required by heads #52
  - Generally can iterate #53

Syntax/semantics mismatches

- Syntactically, modifiers are dependents

#54

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

#55

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

#56

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

#57

    - 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* #58
- Phrasal constituents: *on the bus, very elaborate* #59
- Clausal modifiers: *while Kim was reading a book* #60

Types of adjuncts (syntactic)

#61

- Adnominal modifiers: adjectives, adpositional phrases (PPs), relative clauses
- Adverbial modifiers: adpositional phrases (PPs), adverbs, subordinate clauses, discourse markers

#62

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.

#63

- 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.

#64

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

#65

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

Subject > direct object > indirect/2nd object > oblique

#66

Types of arguments

- Clauses can also be arguments (subjects or complements)
    - Finite, closed clausal arguments

#67

Kim believes [(that) Sandy left.]
[That Sandy left] surprised Kim.

    - Non-finite, controlled clausal arguments

Kim expects Sandy [to leave]
Kim tried [to leave]

#68

    - Non-finite, non-controlled clausal arguments

To leave now would be a bad idea.

#69

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

#70

- ... though they often stand in regular relations to each other

#71

- For many applications, it’s not the surface (syntactic) relations, but the deep (semantic) dependencies that matter.

#72

  - Examples?

What are grammatical functions good for?

- Syntactic phenomena differentiating arguments are sensitive to grammatical function

#73

- Lexical items map semantic roles to grammatical functions

#74

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) #75

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

  - Some syntactic dependents don’t fill a semantic role #76

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

  - Some syntactic dependents aren’t realized locally #77

    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

#78

- 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

#79

Grammatical function identifying phenomena

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

#80

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

#81

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

#82

Grammatical function identifying phenomena examples

- Word order (English):

  Kim saw Sandy ≠ 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)

#83

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

#84

- 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)

#85

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.

#86

- Interacts with passive:

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

#87

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)

  #88

- Syntactic arguments without any local semantic role: raising

  Sandy expected Kim to laugh.
  Sandy continued to laugh.

  #89

- Syntactic arguments which play two roles: control

  Sandy persuaded Kim to leave.
  Sandy tried to laugh.

  #90

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

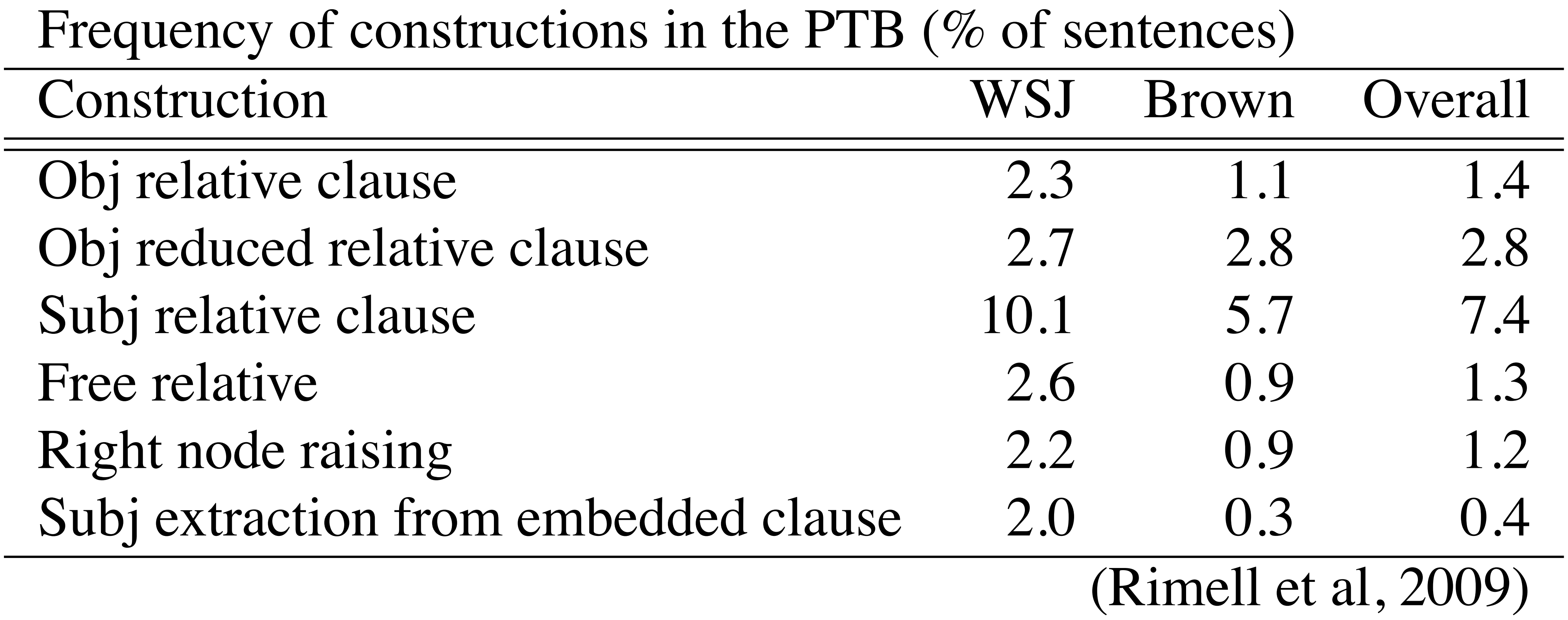
Long-distance dependencies

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

#91

- 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!



Frequency of constructions in the PTB (% of sentences)

Construction	WSJ	Brown	Overall
Obj relative clause	2.3	1.1	1.4
Obj reduced relative clause	2.7	2.8	2.8
Subj relative clause	10.1	5.7	7.4
Free relative	2.6	0.9	1.3
Right node raising	2.2	0.9	1.2
Subj extraction from embedded clause	2.0	0.3	0.4

(Rimell et al, 2009)

Semantically empty words

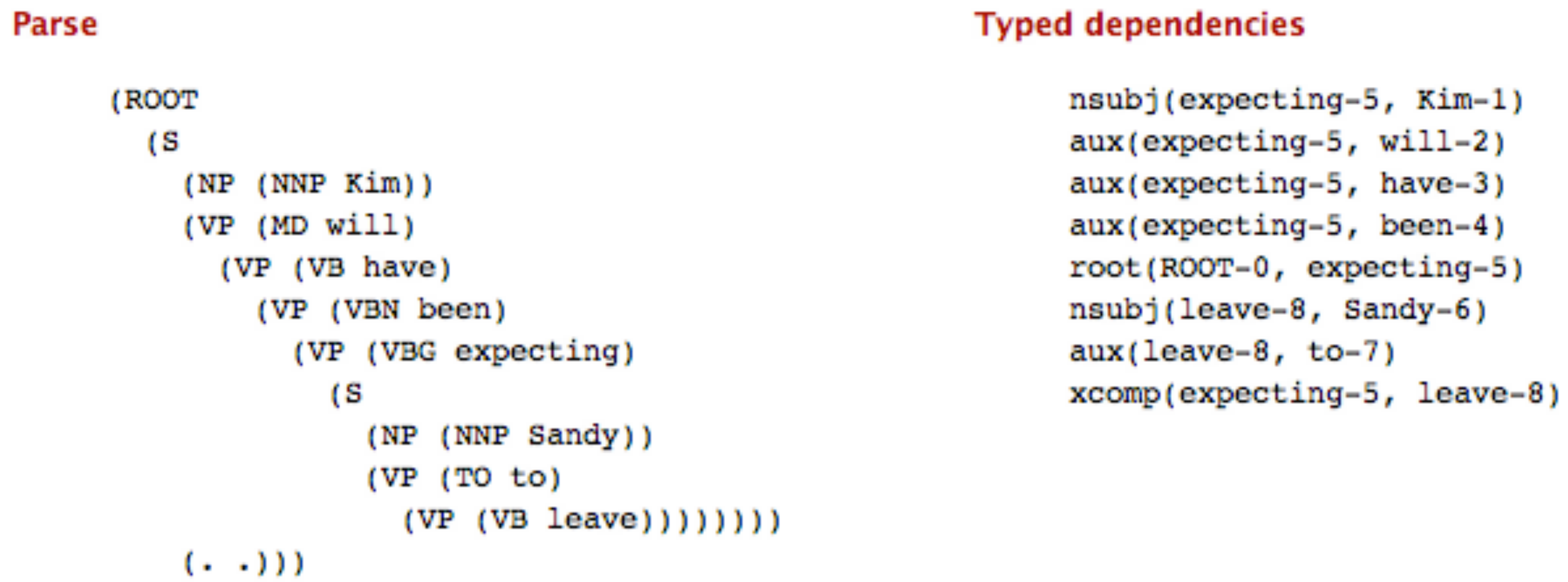
- Don’t contribute lexical content
- Do serve as syntactic “glue”
- Sometimes contribute features to the semantics
- Vary across languages
- Give rise to mis-matches in aligned bitexts

- Examples from English:
  - *complementizers: that, to*
  - *expletives: there, it*
  - *auxiliaries: do, be, will, have*
- Examples from Japanese:
  - *case particles:* が、を、に

#92

Dependency parsers and semantically empty words

- Stanford:



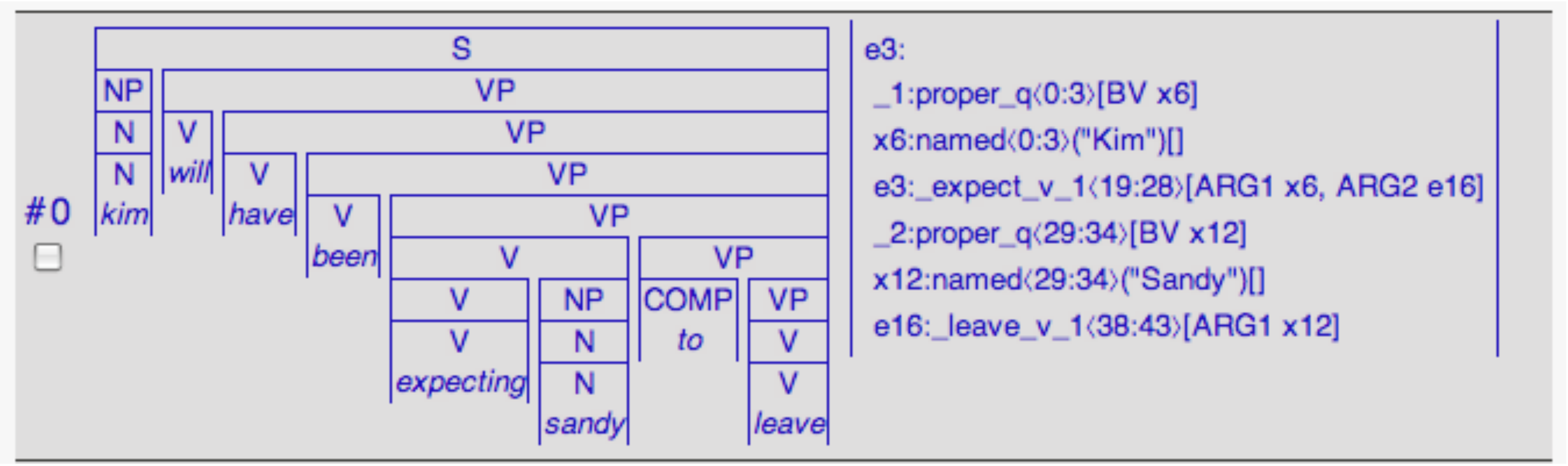
Parse

```
(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

```
nsubj(expecting-5, Kim-1)
aux(expecting-5, will-2)
aux(expecting-5, have-3)
aux(expecting-5, been-4)
root(ROOT-0, expecting-5)
nsubj(leave-8, Sandy-6)
aux(leave-8, to-7)
xcomp(expecting-5, leave-8)
```

- ERG:



```
e3:
 _1:proper_q⟨0:3⟩[BV x6]
 x6:named⟨0:3⟩("Kim")[]
 e3:_expect_v_1⟨19:28⟩[ARG1 x6, ARG2 e16]
 _2:proper_q⟨29:34⟩[BV x12]
 x12:named⟨29:34⟩("Sandy")[]
 e16:_leave_v_1⟨38:43⟩[ARG1 x12]
```

Argument drop

- Lexical predicates introduce expectations for a certain (fixed, given a word sense) number of arguments
- Those arguments aren’t always overtly realized

#93

- Permissible argument drop varies by word class and by language

#94

Argument drop, aka null instantiation (Fillmore 1986)

- Definite null instantiation: Referent is recoverable from discourse context

#95

- Indefinite null instantiation: Referent is non-specific/not recoverable from discourse context

#96

- Constructional null instantiation: Referent is determined by syntactic context (imperatives, control)

#97

Argument drop, aka null instantiation (Fillmore 1986)

- Definite null instantiation:

She promised.
I tried.
When did she leave?

They agreed.
She found out.
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.

#98

- Systematic: Subjects (e.g., in Spanish) or any argument (e.g., Japanese) can be dropped, if supported by the discourse context

#99

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

#100

- 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

- WALS: World Atlas of Language Structures Online (wals.info; Dryer and Haspelmath, 2011)
    - 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 WikiWoods Treebank (www.delph-in.net/wikiwoods; Flickinger et al 2010): ERG-based Treebank (with automatic parse selection) over Wikipedia snapshot from July 2008

- 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

Arad, M. (2005). *Roots and patterns: Hebrew morphosyntax*. Dordrecht: Springer.

Bender, E. M. (2011). On achieving and evaluating language independence in NLP. *Linguistic Issues in Language Technology*, *6*, 1–26.

Bender, E. M., Drellishak, S., Fokkens, A., Poulson, L., & Saleem, S. (2010). Grammar customization. *Research on Language & Computation*, 1-50. Available from `http://dx.doi.org/10.1007/s11168-010-9070-1` (10.1007/s11168-010-9070-1)

Bender, E. M., Flickinger, D., & Oepen, S. (2002). The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In J. Carroll, N. Oostdijk, & R. Sutcliffe (Eds.), *Proceedings of the workshop on grammar engineering and evaluation at the 19th international conference on computational linguistics* (pp. 8–14). Taipei, Taiwan.

Bilmes, J. A., & Kirchhoff, K. (2003). Factored language models and generalized parallel backoff. In *in proceedings of hlt/naccl, 2003* (pp. 4–6).

Burzio, L. (2011). Derived environment effects. In M. Van Oostendorp, C. J. Ewen, E. V. Hume, & K. Rice (Eds.), *The Blackwell companion to phonology, vol IV: Phonological interfaces* (pp. 2089–2114). Blackwell.

Daume III, H. (2009, June). Non-parametric Bayesian areal linguistics. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics* (pp. 593–601). Boulder, Colorado: Association for Computational Linguistics. Available from `http://www.aclweb.org/anthology/N/N09/N09-1067`

Dryer, M. S. (2011a). Negative morphemes. 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/112`

Dryer, M. S. (2011b). Prefixing vs. suffixing in inflectional morphology. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Munich: Max Planck Digital Library. Available from `http://wals.info/feature/26A`

Dryer, M. S., & Haspelmath, M. (Eds.). (2011). *The world atlas of language structures online* (2011st ed.). Munich: Max Planck Digital Library. Available from `http://wals.info/`

Fillmore, C. J. (1986). Pragmatically controlled zero anaphora. In *Proceedings of the twelfth annual meeting of the Berkeley Linguistics Society* (pp. 95–107).

Flickinger, D. (2011). Accuracy v. robustness in grammar engineering. In E. M. Bender & J. E. Arnold (Eds.), *Language from a cognitive perspective: Grammar, usage and processing* (pp. 31–50). Stanford, CA: CSLI Publications.

Flickinger, D., Oepen, S., & Ytrestøl, G. (2010). WikiWoods. Syntacto-semantic annotation for English Wikipedia. In *Proc. lrec 2010.* Valletta, Malta.

Göksel, A., & Kerslake, C. (2005). *Turkish: A comprehensive grammar.* London and New York: Routledge.

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`

Hengeveld, K. (1992). Parts of speech. In M. Fortescue, P. Harder, & L. Kristoffersen (Eds.), *Layered structure and reference in a functional perspective* (pp. 29–55). Amsterdam: Benjamins.

Hohensee, M., & Bender, E. M. (2012, June). Getting more from morphology in multilingual dependency parsing. In *Proc. NAACL 2012.* Montréal, Québec: Association for Computational Linguistics.

Hopper, P. J., & Traugott, E. (2003). *Grammaticalization.* Cambridge: Cambridge University Press.

Karlsson, F. (1998). *Yleinen kielitiede [general linguistics].* Helsinki.

Kay, P. (2005). Argument structure constructions and the argument-adjunct distinction. In M. Fried & H. C. Boas (Eds.), *Grammatical construtions: Back to the roots* (pp. 71–98).

Keenan, E. L., & Dryer, M. S. (2007). Passive in the world's languages. In T. Shopen (Ed.), *Language typology and syntactic description, vol 1: Clause structure* (pp. 325–361).

Kingsbury, P., & Palmer, M. (2002). From treebank to PropBank. In *Proceedings of the 3rd international confernce on language resources and evaluation (LREC2002).* Las Palmas, Spain.

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (Sixteenth ed.). Dallas, TX: SIL International. (Online version: http://www.ethnologue.com)

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*, 313–330.

Marneff, M.-C. de, & Manning, C. D. (2011). *Stanford typed dependencies manual.* Available from `http://nlp.stanford.edu/software/dependencies_manual.pdf` (Revised for Stanford Parser v. 1.6.9)

McDonald, R., Lerman, K., & Pereira, F. (2006). Multilingual dependency analysis with a two-stage discriminative parser. In *Proc. of the tenth conference on computational natural language learning (CoNLL-X)* (pp. 216–220).

Miller, P. H., & Sag, I. A. (1997). French clitic movment without clitics or movement. *Natural Language and Linguistic Theory*, *15*, 573–639.

Nichols, J. (1986). Head-marking and dependent-marking grammar. *Language*, *62*, 56–119.

Noonan, M. (1992). *A grammar of Lango.* Berlin: Mouton de Gruyter.

Nordlinger, R. (1998). *A grammar of Wambaya, Northern Australia.* Canberra: Research School of Pacific and Asian Studies, The Australian National University.

O'Grady, W., Archibald, J., Aronoff, M., & Rees-Miller, J. (2010). *Contemporary linguistics: An introduction* (Sixth ed.). Boston: Bedford/St. Martin's.

Postal, P. M., & Pullum, G. K. (1988). Expletive noun phrases in subcategorized positions. *Linguistic Inquiry*, *19*, 635–670.

Rimell, L., Clark, S., & Steedman, M. (2009). Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 813–821). Singapore: Association for Computational Linguistics.

Toutanova, K., Suzuki, H., & Ruopp, A. (2008, June). Applying morphology generation 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

Ud Deen, K. (2006). Object agreement and specificity in early Swahili. *Journal of Child Language*, *33*, 223–246.

Zwicky, A. M., & Pullum, G. K. (1983). Cliticization vs. inflection: English *n't*. *Language*, *59*, 502–13.