Introduction to Linguistics for Natural Language Processing

Ted Briscoe
Computer Laboratory
University of Cambridge

August 28, 2018

Abstract

This handout is a guide to the linguistic theory and techniques of analysis that will be useful for the ACS NLP modules. If you have done some (computational) linguistics, then reading it and attempting the questions interspersed in the text as well as the exercises will help you decide if you need to do any supplementary reading. If not, you will need to do some additional reading and then check your understanding by attempting the exercises. See the end of the handout for suggested readings – this handout is not meant to replace them. I will set additional (ticked) exercises during sessions which will be due in the following week. Ticks will contribute 20% of the final mark assigned for the module. Successful completion of the assessed practicals will require an understanding of much of the material presented, so you are advised to attend all the sessions and do the supplementary exercises and reading.

Contents

1 The Components of Natural Language(s) 3
1.1 Phonetics . . . 3
1.2 Phonology . . . 4
1.3 Morphology . . . 4
1.4 Lexicon . . . 4
1.5 Syntax . . . 5
1.6 Semantics . . . 5
1.7 Pragmatics . . . 5

2 (Unique) Properties of Natural Language(s) 6
2.1 Arbitrariness of the Sign . . . 6
2.2 Productivity . . . 6
2.3 Discreteness / Duality . . . 6

2.4 Syntax 6
2.5 Grammar and Inference 7
2.6 Displacement 8
2.7 Cultural Transmission 8
2.8 Speak / Sign / Write 9
2.9 Variation and Change 9
2.10 Exercises 9

3 Linguistic Methodology **10**
3.1 Descriptive / Empirical 10
3.2 Distributional Analysis 10
3.3 Generative Methodology 11
3.4 Exercises 12

4 Morphology (of English) **12**
4.1 Parts-of-speech 12
4.2 Affixation 14
4.3 Ir/Sub/Regularity 15
4.4 Exercises 15

5 Syntax (of English) **16**
5.1 Constituency 16
5.2 Lexical Features 16
5.3 Phrasal Categories 17
5.4 Clausal Categories 18
5.5 Phrase Marker Trees 18
5.6 Diagnostics for Constituency 19
5.7 Grammatical Relations 21
5.8 Other Relations 23
5.9 Exercises 24

6 Semantics (of English) **25**
6.1 Semantics and Pragmatics 26
6.2 Semantic Diagnostics 26
6.3 Semantic Productivity/Creativity 27
6.4 Truth-Conditional Semantics 27
6.5 Sentences and Utterances 28
6.6 Syntax and Semantics 28
6.7 Semantic Analysis 29
6.8 Sense and Reference 30

6.9 Presupposition 31
6.10 Semantic Features and Relations 31
6.11 Thematic Relations 31
6.12 Exercises 32

7 Pragmatics **33**
7.1 Speech Acts 33
7.2 Deixis & Anaphora 33
7.3 Discourse Structure 35
7.4 Intentionality 35
7.5 Ellipsis 36
7.6 Implicature 36
7.7 Exercises 37

8 Further Reading **37**

1 The Components of Natural Language(s)

1.1 Phonetics

Phonetics is about the acoustic and articulatory properties of the sounds which can be produced by the human vocal tract, particularly those which are utilised in the sound systems of languages. For example, the sound unit (or phone) [b] is a voiced, bilabial plosive; that is a burst of sound is produced by forcing air through a constricted glottis to make the vocal chords vibrate and by releasing it from the oral cavity (mouth) by opening the lips. The phone [p] is the same except that it is unvoiced – the glottis is not constricted and the vocal chords don't vibrate. Say *bun, pun, but, putt* and decide whether [n] and [t] are voiced or unvoiced, by placing a finger gently on your vocal chords. Acoustically, such distinct sounds (mostly) create distinct patterns which we can display via spectral analysis of the waveform they produce (measuring time, frequency & intensity). However, when we look at spectra of utterances containing the same phones it is often difficult to see similar patterns corresponding to each individual phone because of co-articulation: the fact that speech is produced by continuous movement of our vocal apparatus. For example, we hear the /b/ in [bi] and [ba], but the fact that our tongues are moving to different locations – roughly at the top-front and bottom-back of the mouth respectively – to produce the following high-front or low-back vowels, *as* we open our lips, means that it is difficult to see where [b] ends and the vowel begins and difficult to isolate an invariant acoustic component for [b].

1.2 Phonology

Phonology concerns the use of sounds in a particular language. English makes use of about 45 phonemes – contrastive sounds, eg. /p/ and /b/ are contrastive because *pat* and *bat* mean different things. (Note the use of [x] for a phone and /x/ for the related phoneme.) In Vietnamese these two sounds are not contrastive, so a Vietnemese second language learner of English is likely to have trouble producing and hearing the distinction. Phonemes are not always pronounced the same, eg. the [p] in /pat/ is different from that in /tap/ because the former is aspirated, but aspiration or no aspiration /pat/ still means *pat* (in English). Some of this allophonic variation in the pronunciation of phonemes in different contexts is a consequence of co-articulation but some is governed by language or dialect specific rules. For example, in American English, speakers are more likely to collapse *want to* to 'wanna' in an utterance like *who do you want to meet?* than British English speakers, who will tend to say 'wanta'. However neither group will phonologically reduce *want to* when *who* is direct object of *want* e.g. *who do you want to resign?* Phonology goes beyond phonemes and includes syllable structure (the sequence /str/ is a legal syllable onset in English), intonation (rises at the end of questions), accent (some speakers of English pronounce *grass* with a short/long vowel) and so forth.

1.3 Morphology

Morphology concerns the structure and meaning of words. Some words, such as *send*, appear to be 'atomic' or monomorphemic others, such as *sends, sending, resend* appear to be constructed from several atoms or morphemes. We know these 'bits of words' are morphemes because they crop up a lot in other words too – *thinks, thinking, reprogram, rethink*. There is a syntax to the way morphemes can combine – the affixes mentioned so far all combine with verbs to make verbs, others such as *able* combine with verbs to make adjectives – *programable* – and so forth. Sometimes the meaning of a word is a regular, productive combination of the meanings of its morphemes – *unreprogramability*. Frequently, it isn't or isn't completely eg. *react, establishment*.

1.4 Lexicon

The lexicon contains information about about particular idiosyncratic properties of words; eg. what sound or orthography goes with what meaning – *pat* or /pat/ means PAT, irregular morphological forms – *sent* (not *sended*), what part-of-speech a word is, eg. *storm* can be noun or verb, semi-productive meaning extensions and relations, eg. many animal denoting nouns can be used to refer to the edible flesh of the animal (*chicken, haddock* etc) but some can't (easily) *cow, deer, pig* etc., and so forth.

1.5 Syntax

Syntax concerns the way in which words can be combined together to form (grammatical) sentences; eg. *revolutionary new ideas appear infrequently* is grammatical in English, *colourless green ideas sleep furiously* is grammatical but nonsensical, whilst **ideas green furiously colourless sleep* is ungrammatical too. (Linguists use asterisks to indicate 'ungrammaticality', or illegality given the rules of a language.) Words combine syntactically in certain orders in a way which mirrors the meaning conveyed; eg. *John loves Mary* means something different from *Mary loves John.* The ambiguity of *John gave her dog biscuits* stems from whether we treat *her* as an independent pronoun and *dog biscuits* as a compound noun or whether we treat *her* as a demonstrative pronoun modifying *dog.* We can illustrate the difference in terms of possible ways of bracketing the sentence – (john (gave (her) (dog biscuits))) vs. (john (gave (her dog) (biscuits))).

1.6 Semantics

Semantics is about the manner in which lexical meaning is combined morphologically and syntactically to form the meaning of a sentence. Mostly, this is regular, productive and rule-governed; eg. the meaning of *John gave Mary a dog* can be represented as (SOME (X) (DOG X) & (PAST-TIME (GIVE (JOHN, MARY, X)))), but sometimes it is idiomatic as in the meaning of *John kicked the bucket*, which can be (PAST-TIME (DIE (JOHN))). (To make this notation useful we also need to know the meaning of these capitalised words and brackets too.) Because the meaning of a sentence is usually a productive combination of the meaning of its words, syntactic information is important for interpretation – it helps us work out what goes with what – but other information, such as punctuation or intonation, pronoun reference, etc, can also play a crucial part.

1.7 Pragmatics

Pragmatics is about the use of language in context, where context includes both the linguistic and situational context of an utterance; eg. if I say *Draw the curtains* in a situation where the curtains are open this is likely to be a command to someone present to shut the curtains (and vice versa if they are closed). Not all commands are grammatically in imperative mood; eg. *Could you pass the salt?* is grammatically a question but is likely to be interpreted as a (polite) command or request in most situations. Pragmatic knowledge is also important in determining the referents of pronouns, and filling in missing (elliptical) information in dialogues; eg. *Kim always gives his wife his wages. Sandy does so too.*

General Knowledge also plays an important role in language interpretation; for example, if I say *Lovely day* to you whilst we are both being soaked by heavy rain, you will use knowledge that people don't usually like rain to infer that I am being ironic. Similarly, the referents of names and definite descriptions, if not determined situationally, are determined through general knowledge which may

be widely shared or not; eg. *the prime minister*, *Bill, my friend with red hair*. Pronoun reference can also often only be determined using general knowledge; eg. *Kim looked at the cat on the table. It was furry / white / varnished / fat / china / frisky ...*.

2 (Unique) Properties of Natural Language(s)

2.1 Arbitrariness of the Sign

Words relate sounds (or written equivalents) to referents / meanings. There is no systematicity or semantic motivation to this relationship. Onomatopeia is usually a myth (e.g. *whisper* and French *chuchoter* are both often said to be onomatopeic), though there are sometimes intuitive commonalities of meaning to words that contain similar sound components (maybe related to synaesthesia). What is common to the meaning of many English words beginning with *gl* or with *fl* and can you find some clear exceptions? – look in a dictionary or at text on-line...

2.2 Productivity

Animal communication appears to be restricted to a finite set of calls. Vervet monkeys have 3 alarm calls for 'look out there's a *snake* / *leopard* / *eagle*' which induce different defensive behaviour in the troop (up tree / away from tree / under tree). But human languages allow an infinite range of messages with finite resources. How?

2.3 Discreteness / Duality

Words and **morphemes** are comprised of **phonemes**. Words and morphemes have (referential or grammatical) meanings, but phonemes do not. /pat/ and /bat/ are different words distinguished by the phonemes /p/ and /b/ which also distinguish /pad/ and /bad/ but /p/ and /b/ alone don't have a meaning. The plural morpheme (*+s*) can be suffixed to three of these words, but is realised as either /s/ and /z/ – so-called allomorphs of the plural morpheme. (Can you explain the exception and the difference?) An inventory of 40 or so phonemes provides a much bigger inventory of words, even given phonotactic restrictions on the combination of phonemes into syllables (*/vlim/, */mbok). Once we allow polysyllabic words (e.g. *batter*, *paddle*) is there any restriction on the number of words that can be formed? What is the longest one you know, or can find in a dictionary? What does longest mean in this context?

2.4 Syntax

Human languages are not just bags of words with no further structure – why not? The organisation of words into sentences is conveyed partly by word structure (endings / inflectional suffixes in English) and arrangement / order.

So *Kim loves Sandy* doesn't mean the same thing as *Sandy loves Kim* and **loves Kim Sandy* doesn't convey much at all. In *They love each other*, *love* has a different form because it is agreeing with a plural subject rather than a 3rd person singular subject.

In order to gain further insight into the function of syntax, consider what a language without syntax would be like. Such a language would be just a vocabulary and a sentence would be any set of words from that vocabulary. Now imagine that this language has English as its vocabulary. A 'sentence' in this imaginary language is shown below:

```
      the          hit(s)           a

           with             tramp(s)

  sharp             poor      rock(s)     some

           boys           cruel
```

There is no clue which words should be interpreted with which others in this sentence, so there are many possible interpretations which can be 'translated' into real English, as in (1a,b).

(1) a The cruel boy(s) hit(s) some poor tramp(s) with a sharp rock.

b The cruel, sharp tramp with a rock hit some poor boys.

How many more possible interpretations can you find? Without syntax, sentences would be very ambiguous indeed and, although context might resolve some of these ambiguities in everyday communication, imagine trying to discuss politics, philosophy or to explain the design of a computer in such a language!

2.5 Grammar and Inference

Linguists tend to use the term **grammar** in an extended sense to cover all the structure of human languages: **phonology, morphology, syntax** and their contribution to meaning. However, even if you know the grammar of a language, in this sense, you still need more knowledge to interpret many utterances. All of the following, sentences are underspecified in this sense. Pronouns, ellipsis (incomplete sentences) and other ambiguities of various kinds all requires additional non-grammatical information to select an appropriate interpretation given the (extra)linguistic context.

1. She smiled
2. I didn't

3. Who?
4. Yes
5. The farmer killed the duckling in the barn
6. Everyone in this room speaks one language
7. Every student thinks he is the cleverest person in Cambridge
8. Can you open the gate?

Can you contextualise them to give them different meanings and explain how the context resolves the ambiguities?

Whilst the grammatical knowledge required to encode or decode messages in a particular language is circumscribed, the more general inference required to extract messages from utterances is not. Consider the kinds of knowledge you use to make sense of the following dialogues:

1. A: The phone's ringing. B: I'm in the bath.
2. A: John bought a Porsche. B: His wife left him.
3. A: Pint, please. B: Bitter?

You need to know all sorts of culturally specific and quite arbitrary things like the normal location of phones in houses, the **semiotics** of car brands, and the form of public house transactions, and can make plausible inferences based on these, then these dialogues make sense.

2.6 Displacement

Most animal communication is about the here and now (recall Vervet monkey calls, though the bee dance, indicating direction and distance of food sources, is sometimes said to be a partial exception) but human language allows communication about the past, the future, the distant and the abstract, as well as the here and now and the perceptually manifest.

2.7 Cultural Transmission

Animal communication systems are very largely innate – vervet monkeys are genetically programmed to make 3 calls, although some aspects of the meaning and sound are tuned up by experience. Human language is very largely learnt (that's why there are 6K or so attested languages with widely differing grammatical systems and vocabulary). However, in many ways first language acquisition differs from learning, say, to swim or do sums – it's very reliable under widely differing conditions, does not require overt tuition, and there isn't that much variation in the core grammatical skills of all adult humans. Human children only consistently fail to learn fluent language if entirely denied access

to any sample until they are in their teens. There is much wider variation between individuals and between children and adults in acquisition of passive (understood) and active (produced) vocabulary. Vocabulary learning is an ongoing process throughout life and is supported by teaching aids like dictionaries in literate cultures, whilst first language, grammatical acquisition appears to be largely complete before puberty.

2.8 Speak / Sign / Write

Animal languages always use a single modality: manual gestures, 'dances', oral sounds, clicks, etc. Humans can acquire or even create natural sign languages if denied access to spoken language. Human languages also often have a written form, though the latter is significantly less 'natural' and literacy is only acquired (by most individuals) if explicitly taught over a sustained period.

2.9 Variation and Change

Human languages, unlike animal communication systems, vary considerably through time and space (within-species birdsong being the partial exception). Of the 6K attested languages we know about, 1K are spoken in Papua New Guinea (an area about the size of Texas). There have probably been 100K-500K human languages depending on when language first emerged (mostly undocumented, prehistoric, and extinct, of course). Languages have constantly (dis)appeared as a result of population movements, and the birth and collapse of societies. However, the current rate of language death far exceeds that of creation. Why?

For each language spoken by a population of any size, there are many dialects associated with different regions and/or social classes. New words and novel grammatical constructions are constantly entering languages and old ones are constantly decaying. It is impossible to predict with certainty whether an innovation will spread or decay, although afterwards it is possible to document with some accuracy what did happen (historical linguistics), and some social situations (e.g. creolisation, population movement) cause partly predictable rapid and radical change. Dialectal variation is often a function of social groups' self-identity, so often the explanation of change or variation is in terms of social change, movement or interaction of individuals between groups, etc (sociolinguistics).

2.10 Exercises

What are the similarities and differences between natural human languages and artificial human languages, such as logics or programming languages? – use the properties above as a checklist, but also see if you can think of anything I haven't mentioned. Are natural languages more like species which evolve? (Be succinct!)

3 Linguistic Methodology

3.1 Descriptive / Empirical

Linguists are interested in what people do say (or write), not what they think they say or think they should say. They also have a good line on why prescriptivists are usually misguided and ignorant; for instance, the prescription that thou shalt not split an infinitive is said to derive from a misplaced elevation of Latin grammar, in which there is no infinitive to split. In fact, for English if you accept that *to* is an auxiliary verb like *has* etc, then the simplest rule to internalise as a child would be one which allows adverbs to occur before or in between a sequence of auxiliary verbs, but not between a main verb and its direct object:

1. Captain Kirk (boldly) has (boldly) gone beyond our galaxy
2. Captain Kirk's mission is (boldly) to (boldly) go beyond our galaxy
3. Captain Kirk (boldly) has (boldly) been (boldly) travelling (*boldly) the universe for 30 years

and this is what children do appear to learn... Some linguists deviously argue that such prescriptive rules are intentionally 'unnatural' or arbitrary so that the prestige class, which invents them, can preserve its (linguistic) self-identity. That is, they are like arbitrary rules of etiquette – use knives and forks from the outside in and not the inside out, don't fold your napkin after the meal, etc. (No doubt your experiences of college dining in Cambridge will conclusively disprove this theory.)

3.2 Distributional Analysis

Linguists have attempted to develop a **methodology** for discovering the grammars of languages by empirical and objective (replicable, scientific) means. The heart of this method is distributional analysis. You have already seen some examples of this method above with /p/ and /b/ and *boldly*. The basic idea is that we can create templates, perform substitutions, and test for grammaticality either by using our intuition or that of an informant. For example, the following template could be used to find more examples of English (animate) common nouns:

The —- can run.

where possible answers are *children*, *sheep*, *teacher* and nonanswers are *quickly*, *hallucinate*, *Fred* because all these result in ungrammatical (asterisked (*) sentences). What about *grass*, *table* or *tortoise*? – these don't result in ungrammaticality so much as varying degrees of (semantic) implausibility or oddness. Linguists usually put a question mark or two in front of such examples. Telling the difference between ungrammatical and nonsensical / implausible / odd is surprisingly tricky – we'll return to this below.

The next stage is to take a template like:

— can run.

and discover that *it*, *the car*, *the old car*, *the old car with red seats* and infinitely more (multiword) units or constituents can be substituted into this slot. Thus, we are led to a hierarchical structure in which words are grouped into larger phrases or clauses, all called **constituents**, which have the same distribution: hence, **immediate constituent analysis**, the dominant methodology of American Structural Linguistics from the publication of Leonard Bloomfield's *Language* in 1933 until the 1960s when generative linguistics became influential.

Taken to its logical conclusion, distributional analysis should provide a 'discovery procedure' for grammars, so mechanical and so objective that it would be possible to start from nothing and develop a complete grammar simply by rigorously following the method – Charles Fries' *The Structure of English* published in 1954 tried this and classified constituents into 'type1', 'type2' etc instead of the more traditional and mnemonic noun, verb (phrase) etc. These days, we could try to automate the method, as we have a lot of text in electronic form – how well would this work?

The same process works as well for phonology or morphology: /— a t/ or sell+ — what can go in these slots?

In Europe, the emphasis in grammatical analysis was, and to some extent still is, on relations rather than constituents. That is, instead of trying to classify words and phrases into categories like noun (phrase), verb (phrase), etc., linguists preferred to emphasise that *the car* is **subject of** *run* in *The car can run* and *can* and *the* are **dependents of** the **heads** *car* and *run*. To a large extent, the insights of this tradition have been integrated with many modern generative grammatical theories, as derivative from the more basic and fundamental notion of constituency.

3.3 Generative Methodology

Noam Chomsky published *Syntactic Structures* in 1957 ushering in the generative era of linguistic theory. The essential paradigm shift or methodological innovation was that linguistic analysis was no longer an entirely 'bottom-up', data-driven purely empirical process, but rather, generative linguists started out with a metatheory of what grammars of human languages look like and attempted to express specific grammars within this metatheory. Such grammars are **generative** because they consist of finite sets of rules which should predict all and only the infinite grammatical sentences of a given human language (and what is conveyed about their meaning by their grammatical structure). Thus generative grammars define well-formed sets or mappings between sentences and (part of their) meanings.

Generative grammar got going at much the same time that theoretical computer science, and much of the theory of parsing and compiling programming languages has its antecedents in early generative linguistics (**The Chomsky**

Hierarchy, etc). For example **context-free grammars** and Backus-Naur notation are weakly equivalent formalisms, generating the same class of context-free languages, which seem quite appropriate for capturing the hierarchical structure that emerges from immediate constituent analysis. However once formulated this way, the analysis becomes predictive because the rules of the grammar generate further sentences paired with hierarchical structure.

Generative theory is thus good for capturing the productivity of human language(s). However, even with a metatheory, we still need methods to choose between analyses and to choose the metatheory. So we'll focus primarily on linguistic analysis (and terminology) for now.

3.4 Exercises

Demonstrate by distributional analysis that a specific class of English words can appear in the following slot:

Kim put the book — the shelf

What's the name for this class? Does it combine first with *the shelf*? Can you define (generative) rules that will ensure that this class of words combines with the right kind of constituent?

4 Morphology (of English)

It is very useful to be able to analyse words into morphemes and determine their part-of-speech. What follows is a brief outline of how to do this.

4.1 Parts-of-speech

Words can be analysed into parts-of-speech: major lexical syntactic categories, such as N(oun), V(erb), A(djective), P(reposition), or more minor categories, such as Comp(lementizer), Det(erminer), Deg(ree intensifier), and so forth:

```
N: car, cars; woman, women...
V: thinks, thinking; sold, selling...
A: old, older, oldest; pedantic...

P: in, on, with(out), although...
Comp: that, if...
Det: the, a, those, that, some...
Deg: so, very...
```

N,V,A are the categories of the contentful or open-class vocabulary. Membership of these categories is large (as a glance at any dictionary will tell you)

and open-ended (people invent new words (neologisms) like *fax*, *biro*) and often open-class words belong to more than one category (e.g. *storm* can be a noun or verb, and morphologically-related *stormy* is an adjective); that is, they are ambiguous in terms of lexical syntactic category. (Some words are ambiguous at the level of lexical semantics though not in terms of lexical syntactic category e.g. *match*, N: game vs. lighter) Adverbs also form a large open-ended class, but they are highly related to adjectives and often formed by adding the suffix *+ly* to adjectives (*badly*, *stormily*, etc) so we won't give them a separate category but treat them as A[+Adv].

The other categories are those of functional or closed-class words, which typically play a more 'grammatical' role with more abstract meaning. Membership of these categories is smaller and changes infrequently. For example, prepositions convey some meaning but often this meaning would be indicated by case endings or inflection on words in other languages and sometimes there are English paraphrases which dispense with the preposition: *Kim gave a cat to Sandy / Kim gave Sandy a cat.* Degree intensifiers in adjectival or adverbial phrases *very beautiful(ly)* convey a meaning closely related to the comparative suffix *more beautiful / taller.* Determiners, such as the (in)definite articles (*the, a*), demonstrative pronouns (e.g. *this, that*) or quantifiers (e.g. *some, all*) help determine the reference of a noun (phrase) – quite frequently articles are absent or indicated morphologically in other languages (hence the common non-native speaker error of the form *please, where is train station?*).

The complete set of lexical syntactic categories (for English) depends on the syntactic theory, but the smallest sets contain around 20 categories (almost corresponding to traditional Greek/Latin-derived parts-of-speech) and the largest thousands. For the moment the set introduced above will do us, but see e.g. the frontpiece (opening pages) of Jurafsky and Martin for one popular part-of-speech tagset.

Often words are ambiguous between different lexical categories. What are the possibilities for *broken*, *purchase*, *that* and *can*? There are diagnostic rules for determining the category appropriate for a given word in context; e.g.s: if a word follows a determiner, it is a noun: *the song was a hit*; if a word precedes a noun, is not a determiner and modifies the noun's meaning, it is an adjective: *the smiling boy laughed* – can you think of an exception to the last rule?

These rules and categorial distinctions can be justified by doing distributional analysis both at the level of words in sentences. The process is more long-winded, though. The following template schemata are enough to get you to the rules above which are abstractions based on identifying the classes like noun, determiner, and adjective

1. — boy(s) can run
2. — older boy(s) can run
3. The — boy(s) can run
4. The older — can run

There are other ways to make these distinctions too. For example, nouns often refer to fairly permanent properties of individuals or objects, *boy*, *car*, etc., verbs often denote transitory events or actions, *smile*, *kiss*, etc. However, there are many exceptions, *(a) storm*, *philosophy*, *weigh*, *believe*, etc. Linguists have striven to keep syntax and semantics separate and justify syntactic categories on distributional grounds, but there are many interactions between meaning and syntactic behaviour.

4.2 Affixation

Affixes can be added to word **stems** (lemmas or headwords with some abstraction to account for spelling / sound change modifications). Combining free and bound (allomorphs of) morphemes (stems and affixes) usually involves spelling changes – *able* → *ability*, *change* → *changing*.

Inflectional suffixes like *+s*, *+ed* or *+ing* create variants of the same part-of-speech as the stem / headword, e.g. *boy+s* N-sg| pl, *think+s* V-not3sg| 3sg, *think+ing* V-bse| prog, etc. The change in meaning associated with inflectional suffixes relates to the syntactic context in which they occur – they affect agreement, tense etc which are properties of sentences and not (just) words. Derivational affixes affect the inherent meaning of words and often change the part-of-speech too, e.g. *teach(er)* V| N, *old(er)* A| A-comp(arative). There are productive rules about the combination of morphemes to make words and their consequent meaning:

```
((un   ((re  program) able)) ity)
((A/A  ((V/V V)        A\V)) N\A)

((un ((re program) able)) ity)
'the-property-of not being-able to-program (x) again'
```

where X/Y means a prefix combines with a Y to make a word of category X and X\ Y is the analogue for suffixes. What is the final category of the word? What is the bracketing indicating? How do the affixes pair up with the meaning elements in the gloss?

These rules can be motivated by distributional analysis using templates like the following:

1. The — +able computer
2. They re+ — the computer
3. The un+ — computer
4. — +ity is not a good feature

English is relatively isolating (not much inflectional morphology), languages like Hungarian, Finnish and Turkish have many variants (often 100s sometimes 1000s) of each verb. Others, like Arabic, use infixation rather than suffixation (or prefixation): *ktb*, *katab* etc. - not much in English but *abso+bloody+lutely* etc. However, English has a lot of derivational affixes and many words are morphologically complex. In the limit, the set of words in English is not finitely specifiable because of iterative / recursive derivational affixes, e.g. *great-great-great grandmother*, *anti-anti-missile*, *rereprogram*, etc. This also means that in the limit a lexicon cannot be organised like a conventional dictionary but must be more 'active' / generative, integrating (semi-)productive lexical processes.

Another important lexical process is **conversion** or **zero-derivation** in which words change class or gain extended meanings by systematic means, e.g. *purchase*, *cry* V can become nouns denoting the result of the V act, *butter*, *oil* N can become verbs denoting the act of applying N, and as mentioned above a lot of animal nouns can also denote the edible flesh of the animal – a semi-productive sense extension i.e. conversion process.

4.3 Ir/Sub/Regularity

Few morphological/lexical rules are fully-productive or regular because not every headword/stem in a lexical class undergoes them and/or the resulting meaning is not always fully systematic and predictable. **Blocking**, that is, preemption by synonymy or by lexical form, is a big source of semi-productivity – avoiding unnecessary redundancy (synonymy) or ambiguity in the lexicon:

```
teach/teacher, buy/buyer, smoke/smoker   (agentive)
dry/dryer, freeze/freezer  (instrumental, subregular)
stick / sticker  (only result not agent, irregular?)
station/?stationer (newsagent), lie/?lyer (liar)
steal/?stealer (thief) but 'a stealer of Porsches' (synonymy)
hammer/?hammerer (lexical form)
grammaticality / ?grammaticalness
curiosity / ?curiousness
but 'The curiousness (?curiosity) of the phenomenon intrigued him'
```

As these e.g.s suggest, this is a complex topic about which a lot more can be said. What problems does semi-productivity raise for automated analysis of words?

4.4 Exercises

Identify three English derivational affixes and do a distributional analysis for them. (Look in a book or newspaper for examples.)

Construct an analysis for three morphologically complex words like that done above for *unreprogramability*.

Is the possessive marker *+'s* as in *Bill's car* an English suffix? If so is it inflectional or derivational? Can you think of examples that might suggest that it is a special kind of 'semi-bound' morpheme called a clitic, and functions syntactically rather than morphologically – more like *+(n)'t* – see (2)

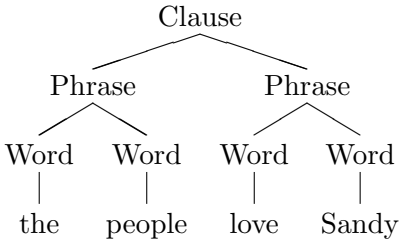
(2) a He didn't come / He did not come
   b ?A good Christian can't attend church and still be saved
   c A good Christian can not attend church and still be saved
   d I love my granny, not! / *I love my granny, n't!

5 Syntax (of English)

In this section we'll develop a basic framework for doing syntactic analysis.

5.1 Constituency

Words go together to form syntactic units of various kinds called constituents. We will recognise the following types of constituents: words, phrases and clauses. These constituents form a hierarchy:



Words, phrases and clauses can be of different types depending on their constituents. Constituents tend to represent coherent units of meaning in some sense. For example, *The people* seems to be interpretable independently of the rest of this sentence. However, what exactly is meant by coherent unit of meaning is not clear, so we will try to find more precise ways of defining constituents.

5.2 Lexical Features

Traditionally, words are categorised according to parts-of-speech. More recently, parts-of-speech have been absorbed into the more general concept of a syntactic category. The major lexical categories are noun, verb, preposition and adjective. There are a variety of minor categories, such as the determiners, intensifiers, complementisers, and so forth (see above). However, we also need to be able to make distinctions within parts-of-speech or major lexical categories.

We've seen some already in terms of morphological variants like sg/pl etc. Here is a preliminary list with some e.g.s:

```
Num(ber): Sg / Pl -- boy(+s)
N-type: Mass,Count, Name: -- boy, information *information+s, Fred
Per(son): 1,2,3 -- I (1sg), you (2sg) (s)he (3sg)
Case: Nom, Acc --  he (nom), him (acc)
Valence: Intrans, Trans, Ditrans, Scomp,... smile, kiss, give, believe,...
A-type: base / comparative / superlative -- old, older, oldest
```

Some of these features affect more than one major category – number, person. Others are category-specific – case, valence. These and similar fine-grained within category or subcategory distinctions can all be justified distributionally, as we'll see below. Can you come up with a distributional argument for the count and case distinctions on nouns?

5.3 Phrasal Categories

Each of the major lexical categories is associated with a corresponding (noun, verb, adjectival or prepositional) phrase in which the major lexical category, or head, is obligatory, as illustrated:

```
NP (eg. boys, the boy, an old castle, kings of England)

VP (eg. run, kiss Sandy, give me a present)

AP (eg. old, very old, quite pretty, difficult to understand)

PP (eg. up, to the house, without me)
```

Can you think of sentences which contain these constituents? Where do these constituents occur in relation to each other? If we look at the noun phrase (NP) in a bit more detail, we can see that the head noun is the only obligatory element of every NP and that this noun can co-occur with Det(erminers), APs and PPs, as illustrated in (3).

(3) a The castle is old
    b *The is old
    c The big castle is old
    d *The big is old
    e The castle by the hill is old
    f *The by the hill is old
    g Castles are interesting

5.4 Clausal Categories

Clauses can be independent sentences or embedded inside other sentences. There are various types which can be distinguished by syntactic features:

```
S[decl(arative)] (eg. They kiss Sandy, God exists)

S[interog(ative)]  (eg. Did he kiss Sandy, Who did he kiss)

S[imp(erative)] (eg. kiss Sandy, get up)

S[rel(ative)] (eg. who he kissed, who likes me)

S[comp(lement)] (eg. that Kim kissed Sandy)

S[passive] (eg. Sandy got/was kissed [by Kim])
```

What is the head or obligatory element in a clause? – try to use distributional analysis to work it out.

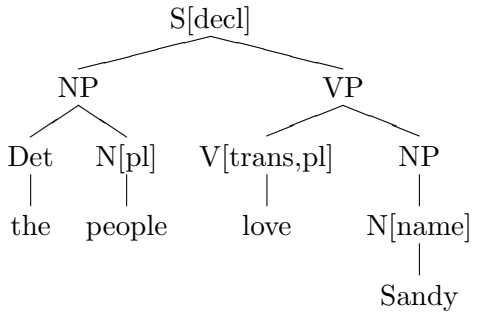
A couple of examples of sentences with further clauses embedded inside them are given in (4). Can you recognise which types of clause they are?

(4) a Kim thinks that castles are interesting.
    b Kim likes the person who she met yesterday.

Once again try to use distributional analysis to work out the constituency of these sentences. This will help you understand how embedded clauses work.

5.5 Phrase Marker Trees

The constituency of a particular sentence can be shown in a phrase marker tree. For instance, we can now show the types of the constituents for *The people love Sandy*:



5.6 Diagnostics for Constituency

Grammaticality is an acquired and quite sophisticated intuition about the correctness of a sentence considered in isolation or the 'null context'. For example (5)

(5) *Reagan thinks bananas

is not a complete grammatical sentence considered independently. Therefore, as syntacticians, we would reject it (hence the asterisk). However, it would be quite possible for this sequence of words to occur in a conversation or textual corpus, as in (6)

(6) a What kind of fruit does Bush like?
    b Reagan thinks bananas

In this context, the missing constituents are 'understood' and the sequence is perfectly acceptable as an elliptical form of (7).

(7) Reagan thinks that Bush likes bananas.

(In addition, there is the issue of nonsensicality vs. ungrammaticality discussed above.)

The most important diagnostic is the possibility of **substitution** or replacement of a possible constituent by another form (particularly a proform, such as *the*, *that*, *do so*, and so forth). (This is the diagnostic I have used exclusively up to this point.) If the replacement can be made without altering the grammaticality of the sentence, then this suggests that the replaced words form a constituent of the same type as those which replaced them. For example, the NP *The people* can be replaced by a wide variety of material:

The people	love Sandy
They	
*He	
Some friends of hers	
The men who she met	
*The old woman with grey hair	
*the	
*quick	
*hit	
*with the man	

This shows us that all of the unasterisked sequences can be NPs. The same technique can be used to work out what words are (transitive) verbs:

```
The people loved         Sandy
           liked
           hit
           chased
           talked to
           looked at
           ...
         *gave
         *likes
         *think
         *pretty
         *girl
         ...
```

What is a transitive verb? What makes it different from other types of verb, such as *give* or *think*?

Constituents, but not partial constituents can be moved around in a sentence, for example (8).

(8) a The old man has come to dinner.
   b Has the old man come to dinner?
   c *The has old man come to dinner

In this case *the old man* is a NP and *has* is an auxiliary verb, but *old man* is only part of the NP. So **movement** is also a diagnostic for constituency.

Parentheticals and other 'extra' constituents can be inserted between some phrasal constituents, but not within them; for example (9).

(9) a The President of America, Ronald Reagan, is over 70.
   b *The President, Ronald Reagan, of America is over 70.
   c *The President of America is, Ronald Reagan, over 70.

So **insertion** is also a diagnostic for constituency.

The **omissibility** of a potential constituent, either because of its optionality or because it is 'understood' in some context, is a sign of constituenthood. For example, the PP *of the old man* can be omitted from (10).

(10) Some friends of the old man came to dinner.

but *of the old*, which is not a constituent in this sentence, cannot.

If two sequences can be coordinated with a conjunction (eg. *and*), they may be constituents of the same category, as illustrated in (11)

(11) a Kim and Sandy kissed each other
b The old men and women came to dinner
c The old man and his young nephew came to dinner
d Kim and Sandy divorced and remarried each other
e Kim kissed Sandy and remarried her
f That rather old and very unreliable car belongs to Kim
g Kim washed up and Sandy watched the TV

Can you name the constituents coordinated in each case?

All these diagnostics are fallible, **coordination** is particularly controversial though widely used by generative as opposed to 'old school distributionalists' What problems do the examples in (12) raise?

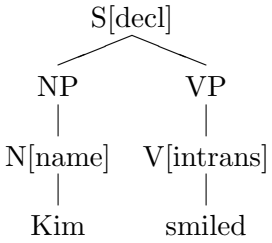
(12) a Kim is a conservative and proud of it
b Kim became a conservative and arrogant
c Kim enjoys chess and watching football
d Kim gave Sandy a pen and Fido a bone
e 'To hell with them and be dammed', he said.

The diagnostics have gone from least theory-laden to most theory-laden in that the implicit metatheory about what can and cannot happen in grammars has got stronger and more constraining. However, even **substitution** assumes that there is such a thing as constituency and this has not gone unchallenged. Some linguists believe that grammatical relations are primary and constituents are derivative. On the whole we will assume the opposite.

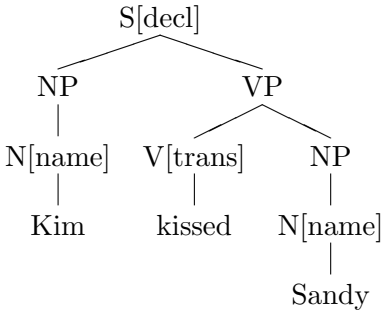
5.7 Grammatical Relations

Traditional grammatical relations like **subject-of**, **direct-object-of** can be reconstructed from the hierarchical constituent structure of a sentence. For example, if we assign the examples in (13) the analyses indicated by the phrase marker trees below, then we can define these relations in terms of the notions of (immediate) dominance and (immediate) precedence. The subject of each verb in each sentence is the NP immediately dominated by S which in turn dominates the verb. The direct object of each verb is the NP immediately dominated by VP and immediately preceded by the verb. The second object of a ditransitive verb is the NP immediately dominated by the VP immediately preceded by NP and preceded by the verb. This definition doesn't capture the traditional notion of 'indirect object' – can you see why not? Finally, an 'oblique object' introduced by a preposition can be defined in similar terms but additionally specifying the PP and preposition type required – can you see how to do this for (13d)?

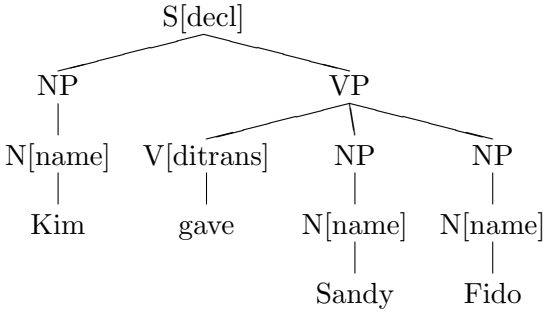
(13) a Kim smiled



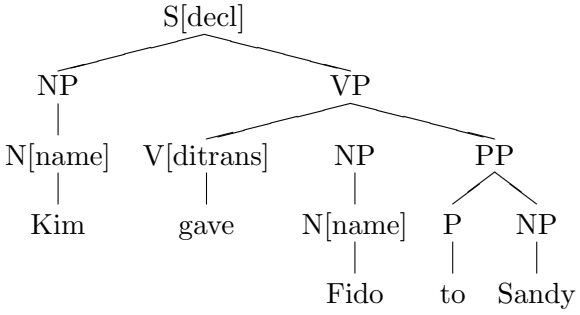
b Kim kissed Sandy



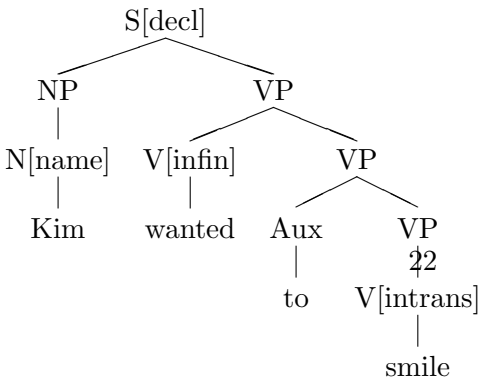
c Kim gave Sandy Fido



d Kim gave Fido to Sandy



e Kim wanted to smile



One way of representing this information is as a set of bilexical head-dependent relations – for instance the relations for (13b) would be:

subject(kiss+ed Kim)
direct-object(kiss+ed Sandy)

These relations form a connected tree in which nodes are labelled with word tokens and edges are labelled with relation types. Many theories allow graphs of grammatical relations – for instance the 'understood subject' of *smile* in (14e) is *Kim* and this can be represented by having the node for *Kim* participate in two subject relations:

subject(wanted+ed Kim)
subject(smile Kim)
infinitive-complement(want+ed smile)

In this case the graphs are directed, connected and may or may not be acyclic. (Draw the graph to convince yourself it is one. Is it cyclic?)

Recently, dependency trees (with one head per dependent) – see next section for more discussion of the notion of head – or dependency graphs, which can represent 'understood' relations have become the de facto standard output for parsers, supplanting constituency trees. There are a number of competing schemes (e.g. Stanford Dependencies, Universal Dependencies, etc.). We will mostly consider (RASP-style) full GR graphs as these provide the most direct mapping to compositional semantics.

5.8 Other Relations

In (13) the verbs are the head (daughters) of the VPs and Ss and the phrases within the VP are dependent complements of the verb. In Dependency Grammar, the subject outside the VP would also be called a dependent of the verb. In the NP-VP analysis of clauses this is not so clear, though if the verb or VP is the head (daughter) of the clause it can be maintained. Nevertheless, all linguists would still call the subject and complements of a verb its arguments. As we'll see below verbs denote predicates which ascribe properties or relations to individuals and thus require a certain number of arguments given their inherent meaning. On the other hand, there are other optional elements to clauses and phrases called variously **specifiers**, and **modifiers** / adjuncts. All of these terms are also relational – a constituent has to be a modifier or specifier of some other constituent. (14) gives some examples where the italicised constituents are of the type indicated in brackets.

(14) a *Those* boys can run (specifier)
b *Bill's* boys can run (specifier)
c He is a *very* proud father (specifier)
d He fell *right* out of the window (specifier)
e He is a *very proud* father (nominal premodifier)
f Those boys can run *this morning* (verbal postmodifier)
g Those boys *definitely* can run (sentential modifier)

Can you name the syntactic categories of each specifier or modifier constituent above? If not, can you work them out by distributional analysis?

There is more to the distinction between heads and other daughters within phrases than just predicates and their obligatory arguments. In fact, many linguists might argue that this is far too 'semanticky' a way of thinking about a syntactic distinction. Heads are not only the only obligatory element of a phrase of any given type (see diagnostics section above), but also grammatical features of phrases are determined (mostly) by grammatical features of the head daughter. For instance, the Per and Num features of the VP are determined by the morphology of the verb, and the Tense of S by the Tense of V(P). The Num of a NP is determined by the morphology (and sometimes semantics) of the noun, and so forth.

5.9 Exercises

1. Pick three sentences of 10 or so words from a novel or newspaper and assign each word a part-of-speech / lexical syntactic category.

2. Justify the distinction between NP and VP in S – i.e. why not adopt an analysis like ((NP V) NP) or (NP V NP)? The examples in (15) should help you get going but see if you can 'fill in' the missing steps in this distributional argument.

   (15) a Passionately Kim kissed Sandy
   b Kim passionately kissed Sandy
   c Kim kissed Sandy passionately
   d *Kim kissed passionately Sandy
   e Kim kissed Sandy and Robin did so too
   f A: Who kissed Sandy? B: Kim did.
   g Kiss Sandy!

   Can you think of any counter examples and arguments based on the distributional diagnostics that would tend to point to one of the alternative analyses? The asymmetry of the NP-VP helps define grammatical relations in terms of constituents – how?

3. The examples in (16) all contain auxiliary verbs, such as variants of *have*, *be*, or *do*, modals like *may* or *can*, and the infinitive auxiliary *to*.

   (16) a Kim has kissed Sandy
   b Kim may have kissed Sandy
   c Kim can have kissed Sandy
   d Kim did kiss Sandy
   e Kim was kissed by Sandy
   f Kim was kissing Sandy
   g Kim has to kiss Sandy

   Work out the order in which the various types of auxiliary verb can occur in verb groups and what requirements they place on the morphological form / grammatical features of the main verb. Justify the decision by giving more complex grammatical examples and some ungrammatical examples. You may also be able to work out the constituent structure of sentences containing such verb groups, building on what's been covered above. Think about how the features which ensure that the next (aux) verb has the right morphological form will need to pass round the phrase marker tree, i.e. what is the head?

4. The approach to grammatical relations defined in section 5.7 above can be extended to other complements, for verbs taking sentential complements and VP complements of various types. Can you see how to define the various relations involved in (17) by drawing the phrase markers and applying and/or extending the definitions?

   (17) a Kim believes that Sandy kissed Fido
   b Kim persuaded Sandy to kiss Fido
   c Kim enjoyed kissing Sandy
   d Kim bet Sandy 5 pounds that Fido kissed Felix

6 Semantics (of English)

Early work on semantics in generative linguistics concentrated on specifying translation procedures between syntactic and semantic structures. However, the meaning of these 'semantic' structures (usually capitalised English words) was never defined. This process just pushed the problem one level further down – rather as though I translate an English sentence into Swahili (or some language you do not understand) and then tell you that is the meaning of the English sentence. Recent work on semantics in generative grammar has been based on 'logical' truth-conditional semantics. This approach avoids the problem by relating linguistic expressions to actual states of affairs in the world by means of the concept of truth. Furthermore, logics usually have a model-theory, and associated proof-theory, which can support automated inference.

6.1 Semantics and Pragmatics

Semantics and Pragmatics are both concerned with 'meaning' and a great deal of ink has been spilt trying to define the boundaries between them. We will adopt the position that **Pragmatics = Meaning – Truth Conditions**. For the most part we will be concerned with the meaning of sentences, rather than the meaning of utterances. That is, we will not be concerned with the *use* of sentences in actual discourse, the speech acts they can be used to perform, and so forth. From this perspective, the three sentences in (18) will all have the same propositional meaning because they all 'involve' the same state of affairs.

(18) a Open the window
b The window is open
c Is the window open

The fact that a) is most likely to convey an assertion, b) a command and c) a question is, according to this approach, a pragmatic fact about the type of speech act language users will typically associate with the declarative, imperative and interrogative syntactic constructions. We will say that all the sentences of (18) convey the same *proposition* – the semantic 'value' of a sentence.

6.2 Semantic Diagnostics

Just as with syntax we use intuitions about 'grammaticality' to judge whether syntactic rules were correct, we will use our semantic intuitions to decide on the correctness of semantic rules. The closest parallel to ungrammaticality is nonsensicality or semantic anomaly. The propositions in (19) are all grammatical but nonsensical.

(19) a Colourless green ideas sleep furiously
b Kim frightened sincerity
c Thirteen is very crooked

Other propositions are contradictions, as in (20).

(20) a It is raining and it is not raining
b A bachelor is a married man
c Kim killed Sandy but she walked away

The assertion of some propositions implies the truth of other propositions; for example (21a) implies b) and c) implies d).

(21) a Kim walked slowly
b Kim walked
c Kim sold Sandy the book
d Sandy bought the book from Kim

This relation is called **entailment** and is the most important of all semantic intuitions to capture in a semantic theory since it is the basis of many of the inferences we make in language comprehension, and most other semantic notions can be reduced to entailment. For example, two propositions can be synonymous, as in (22), but the notion of synonymy reduces to the notion of identity of entailments.

(22) a Kim is a bachelor
b Kim is an unmarried man

That is, if (22a) and (22b) mean the same then the same conclusions follow from their assertion. We also have intuitions about the (semantic) ambiguity of certain sentences; that is they can convey more than one proposition, for example, those in (23).

(23) a Competent women and men go far
b He fed her dog biscuits
c Everyone knows one language

6.3 Semantic Productivity/Creativity

Another important aspect of meaning that we would like our semantic theory to explain is its productivity. We are able to interpret a potentially infinite number of sentences that convey different propositions. Therefore, just as with syntax, we will need to specify a finite set of rules which are able to (recursively) define/interpret an infinite set of propositions.

6.4 Truth-Conditional Semantics

There are two aspects to semantics. The first is the inferences that language users make when they hear linguistic expressions. We are all aware that we do this and may feel that this is what understanding and meaning are. But there is also the question of how language relates to the world, because meaning is more than just a mental phenomenon – the inferences that we make are (often) about the external world around us and not just about our inner states. We would like our semantic theory to explain both the 'internal' and 'external' nature of meaning.

Truth-conditional semantics attempts to do this by taking the external aspect of meaning as basic. According to this approach, a proposition is true or false depending on the state of affairs that obtain in the world and the meaning of a proposition is its truth conditions. For example, *Kim is clever* conveys a true proposition if and only if Kim is clever. Of course, we are not interested in verifying the truth or falsity of propositions – we would get into trouble with examples like *God exists* if we tried to equate meaning with verification. Rather knowing the meaning of a proposition is to know what the world would need to be like for the sentence to be true (not knowing what the world actually is like).

The idea is that the inferences that we make or equivalently the entailments between propositions can be made to follow from such a theory.

Most formal approaches to the semantics are truth-conditional and model-theoretic; that is, the meaning of a sentence is taken to be a proposition which will be true or false relative to some model of the world. The meanings of referring expressions are taken to be individual entities in the model and predicates are functions from individual entities to truth-values (ie. the meanings of propositions). These functions can also be characterised in an 'external' way in terms of sets in the model – this extended notion of reference is usually called denotation. However, we will mostly focus on doing semantics in a proof-theoretic way by 'translating' sentences into formulas of predicate / first-order logic (FOL, as much as possible) and then passing these to a theorem prover, since our eventual goal is automated language understanding.

6.5 Sentences and Utterances

An utterance conveys far more than a propositional content. Utterances are social acts by speakers intended to bring about some effect (on hearers).

Locutionary Act: the utterance of sentence (linguistic expression?) with determinate sense and reference (propositional content)
Illocutionary Act (Force): the making of an assertion, request, promise, etc., by virtue of the conventional force associated with it (how associated?)
Perlocutionary Act (Effect): the bringing about of effects on audiences by means of the locutionary act

Natural languages do not 'wear their meaning on their sleeve'. Discourse processing is about recovering/conveying speaker intentions and the context-dependent aspects of propositional content. We argue that there is a logical truth-conditional substrate to the meaning of natural language utterances (semantics). Sentences have propositional content, utterances achieve effects.

Context-dependent aspects of a proposition include reference resolution, especially with indexicals, such as some uses of personal pronouns, *here*, *this*, time of utterance, speaker etc., so we talk about the propositional content conveyed by a sentence to indicate that this may underspecify a proposition in many ways. We'll often use the term **logical form** to mean (usually) the proposition / propositional content which can be determined from the lexical and compositional semantics of a sentence reperesented in a given logic.

6.6 Syntax and Semantics

As the ambiguous examples above made clear, syntax affects interpretation because syntactic ambiguity leads to semantic ambiguity. For this reason semantic rules must be sensitive to syntactic structure. Most semantic theories pair syntactic and semantic rules so that the application of a syntactic rule automatically leads to the application of a semantic rule. So if two or more syntactic rules can be applied at some point, it follows that a sentence will be

semantically ambiguous.

Pairing syntactic and semantic rules and guiding the application of semantic rules on the basis of the syntactic analysis of the sentence also leads naturally to an explanation of semantic productivity, because if the syntactic rule system is recursive and finite, so will the semantic rule system be too. This organisation of grammar incorporates the principle that the meaning of a sentence (its propositional content) will be a productive, rule-governed combination of the meaning of its constituents. So to get the meaning of a sentence we combine words, syntactically and semantically to form phrases, phrases to form clauses, and so on. This is known as the Principle of **Compositionality**. If language is not (mostly) compositional in this way, then we cannot explain semantic productivity.

Occasionally, we may have problems deciding whether a particular fact about language should be accounted for syntactically or semantically (just as we may have problems deciding whether it belongs to semantics or pragmatics). In this situation, we can use the syntactic framework to make a decision. For example, consider the ambiguous examples in (23). Can you decide whether their ambiguity should be accounted for in the syntactic or semantic rule system?

6.7 Semantic Analysis

We argued that the semantic value of a sentence is (ultimately) a proposition which is true or false (of some state of affairs in some world). What then are the semantic values of other constituent types such as N(P)s, V(P)s, and so forth? If we are going to account for semantic productivity we must show how the semantic values of words are combined to produce phrases, which are in turn combined to produce propositions. It is not enough to just specify the semantic value of sentences.

One obvious place to start is with proper names, like *Kim* or *Sandy* because the meaning of a proper name seems to be intimately connected to the individual it picks out in the world (ie. the individual it refers to). So now we have the semantic values of proper names and propositions but we still need to know the semantic values of verbs before we can construct the meaning of even the simplest propositions. So what is the 'link' between verbs and the world? Intransitive verbs combine with proper names to form propositions – so intransitive verbs pick out properties of individuals. But how can we describe a 'property' in terms of a semantic theory which attempts to reduce all meaning to the external, referential aspect of meaning? One answer is to say that the semantic value of an intransitive verb is the set of individuals which have that property in a particular model. For example, the semantic value of *snore* might be {kim1, fido1}. Now we are in a position to say specify the meaning of (24) in a compositional fashion.

(24) Kim snores

First find the referent of *Kim* and then check to see whether that individual,

say kim1, is in the set of individuals who snore. Now we have specified the truth-conditions of the proposition conveyed by (24).

Developing a truth-conditional semantics is a question of working out the appropriate 'links' between all the different types of linguistic expression and the world in such a way that they combine together to build propositions. To distinguish this extended notion of reference from its more general use, we call this relation **denotation**. Thus the denotation of an intransitive verb will be a set of individuals and of a proper name, an individual. What is the denotation of a transitive verb? What is the denotation of a definite description, such as *the dog*? (If you have studied FOL and model-theories for FOL or other logics you may still be following. If not it is time to read Jurafsky and Martin, ch14, or Cann.)

At this point we should consider more carefully what sentences denote. So far we have assumed that the semantic value of a sentence is a proposition and that propositions are true or false. But what is the link with the world? How is this to be described in external, referential terms? One answer is to say that sentences denote their truth-value (ie. true or false) in a particular world, since this is the semantic value of a proposition. So we add the 'individuals' true and false to the world and let sentences denote these 'individuals'. However, there is an immediate problem with this idea – all true sentences will mean the same thing, because truth-conditional semantics claims in effect that denotation exhausts the non-pragmatic aspects of meaning. This appears to be a problem because *Mr. Blair was prime minister* and *Mr. Bush was president* are both true but don't mean the same thing.

6.8 Sense and Reference

The problem of the denotation of sentences brings us back to the internal and external aspects of meaning again. What we want to say is that there is more to the meaning of a sentence than the truth-value it denotes in order to distinguish between different true (or false) sentences. There are other problems to; consider, for example, the sentence in (25)

(25) The morning star is the evening star.

It was a great astronomical discovery when someone worked out that a star seen at a certain position in the sky in the morning and one seen at another position in the evening were both in fact Venus. Yet according to our theory of semantics this ought to be a tautologous or logically true statement analogous to (26) because the meaning of a definite description or a proper name is just the individual (object) it denotes.

(26) Venus is Venus.

Traditionally, linguistic expressions are said to have both a sense and a reference, so the meaning of *the morning star* is both its referent (Venus) and the concept it conveys (star seen in morning).

At this point you might feel that it is time to give up truth-conditional semantics, because we started out by saying that the whole idea was to explain the internal aspect of meaning in terms of the external, referential part. In fact things are not so bad because it is possible to deal with those aspects of meaning that cannot be reduced to reference in model-theoretic, truth-conditional semantics based on an intensional 'possible worlds' logic. The bad news is though that such logics use higher-order constructs in ways which are harder to reduce to first-order terms for the purposes of automated theorem proving.

6.9 Presupposition

A related issue for truth-conditional semantics is that some referring expressions (NPs) don't seem to refer.

(27) a The King of France is (not) bald
b Have / Haven't you stopped cheating in exams yet?

Given that there is no King of France is the (negated) proposition in (27a) true or false? Similarly either version of (27b) puts the addressee on the spot by presupposing that they have cheated at some point in the past. In order to preserve the idea that propositions are true or false it is necessary to treat presuppositions as propositions which form part of the context of utterance and determine the appropriateness of an utterance to a context, much like felicity conditions for speech acts (see below).

6.10 Semantic Features and Relations

In many books, you will see a lot of 'notation without denotation (i.e. any model-theory or associated proof-theory) like *man* (main sense) = HUMAN+, MALE+, ADULT+ where word meanings are defined in terms of sets of semantic primitives or features. The problem with this from our perspective is what does HUMAN+ mean? Similarly, there is a tradition of defining word meanings in terms of relations like hyponymy (is-a, superordinate-of). For instance, *man* is a hyponym of *human* which is in turn a hyponym of *animal*. It turns out that all of this can be represented in a logic and used to grind out valid entailments, so long as we have the expressive power to represent general rules or **meaning postulates** like 'if any individual has the property of being a man then that individual has the property of being human' or 'any individual that is male and human and an adult is also a man'. If you know some logic, can you express these glosses as well-formed formulas of FOL?

6.11 Thematic Relations

Another kind of semantic 'notation without denotation' you'll come across is the use of terms like 'agent' to label certain arguments of predicates, as in (28).

(28) a Kim (agent) kissed Sandy (patient/theme)
b Sandy (experiencer) enjoyed being kissed
c Sandy (agent) gave Kim (goal/benefactive) a pen (theme)
d Sandy (agent) flew the plane (patient/theme) from London (locative/source) to Paris (locative/goal)

The set of labels is not entirely agreed or consistent, so you may see others and they are variously also called theta-roles, semantic cases, roles or preferences, etc. However, the crucial issue is whether such labels are anything more than convenient ways of referring to stereoptypical inferences that follow from grammatical relations to some extent independently of verbs, or whether there are actual entailments associated with the labels. Agents are usually subjects of verbs denoting events and often cause these events to come about. In cases like (28b), where this is clearly not the case, a different label like 'experiencer' is often used, but there is an extensive middle ground of unclear cases between (28a) and (28b), such as *Sandy flew from London to New York*. An alternative, extreme, school of thought argues that the inferences that can be made are entirely dependent on the predicate sense involved. A middle position is that there are some default entailments that follow from labels like 'agent' true of most verbs in most contexts of use. Can we represent default entailment in first-order logic?

6.12 Exercises

We've touched on the fact that verbs are semantically predicates with one or more arguments. How many arguments can a verb have? Can you think of some examples of verbs whose inherent meaning requires 3 or even 4 arguments. (If you've followed closely, we've seen one e.g. of a 4-place predicate above.) You might want to do a bit of distributional analysis to prove to yourself that your e.g.s really are all arguments and not verbal modifiers.

See if you can figure out the predicate-argument structure of the following sentences by following and extending the reasoning of section 6.7.

(29) a Kim kissed Sandy
b Sandy gave Kim a pen
c The female cat smiled this morning

Now write down one or more well-formed formuli of FOL which most accurately express the meaning of the following examples:

(30) a Competent women and men are successful
b Kim fed her dog biscuits
c Everyone knows one language

7 Pragmatics

Pragmatics is about the use of language in context, where context includes both the linguistic and/or situational context of an utterance / text sentence.

7.1 Speech Acts

Speech acts have felicity conditions not truth-conditions and the former can't be reduced to the latter. Felicity conditions are constitutive for speech acts (ie. they are essential preconditions for an act to take place). For example, you can't promise to do something unless you intend to do it, believe you can do it, wouldn't do it anyway, are being sincere, etc. Otherwise your act will be something other than a promising act.

Utterances have a 'force' as opposed to just a propositional content, and there are 'indirect' speech acts in which the force of an utterance is not that conventionally indicated by the grammatical mood (declarative (statement), imperative (command), interrogative (question)) of the sentence. Can you construct contexts in which the utterance of the examples in (31) would constitute an indirect speech act?

(31) a Would you pass the salt?
b Nuclear power is an ecological disaster.
c Shoot her!

Computation of the speech act intended by a speaker will be highly context-dependent, but essential to recovery of meaning in discourse.

7.2 Deixis & Anaphora

Utterances often do not contain enough information to allow some determinate proposition to be recovered from them, independently of context. Deictic or indexical expressions are one reason for the need for a theory of pragmatics – a theory which by necessity must refer to language use and context. The examples in (32) exhibit person, place, and time deixis, respectively.

(32) a I am hungry
b Will you shut the window
c That's the shop
d You catch the bus over there
e I didn't have a PhD then.
f I'll see you on Wednesday

In each case the propositional content is unclear until it is fixed by the extra-linguistic context of utterance. Most deixis is reducible to truth-conditional meaning, so linguists have proposed reformulations of possible worlds semantics which treat propositions as functions from possible worlds *and contexts* to truth-

values, or alternatively sentences as functions from contexts to propositions and propositions as functions from possible worlds to truth-values. Context is treated as a set of indices, coordinates or reference points, for speakers, addressees, times, places, and so forth.

Anaphora occurs when a linguistic expression (usually a pronoun) is coreferential with a previous expression and where the semantic content of the anaphor (pronoun, definite NP, etc) is sufficiently vague that it can only select a referent by virtue of the preceding linguistic context (ie. the extra information specified by the antecedent), as in (33).

(33) a Kim thinks that he is clever (he=Kim vs. he=Stephen Hawking)

b Sandy likes cars. She bought a Maserati last week (She=Sandy)

c Volvo drivers who wear hats believe that they own the road (they=Vds+hts)

d The house was empty. The front door was broken. (front door = front door of house)

The class of linguistic expressions which can function deictically or anaphorically overlaps substantially (creating ambiguity). Definite NPs, as well as pronouns, often function anaphorically linking back to previously introduced discourse referents which are either less accessible (e.g. 'further back') in the discourse or require some additional inference to make the link. For example, in (33d) *The front door* is coreferentially linked to *The house* via a so-called 'bridging' inference that houses (mostly) have front doors. Less frequently, anaphors precede there 'antecedents' usually in 'marked' circumstances, as in (34). (The traditional (but largely unused) term for this is cataphora.)

(34) a He was tough. He was good. He was handsome. Superman was going to save Hollywood.

b After she had thought it through, Sandy decided to do linguistics.

Determining antecedents for anaphors appears to require general knowledge (prejudice!), as (35) shows.

(35) a The men allowed the women to found their club. (their = women)

b The men allowed the women to join their club. (their = men)

Is a bridging inference a logical entailment, ie. a deductive inference?

7.3 Discourse Structure

Discourse has an information structure, discourses are about a topic, many phenomena like anaphora are resolved via this information structure. For example, below the discourse topic is initially Kim and Sandy and then switches to their transport arrangements.

a) Kim and Sandy are schoolteachers. b) They live in Carshalton Beeches and work at the same school. c) She drives him there every day, but d) he is taken home by a colleague when Sandy has to take games. e) On winter mornings, Sandy's car often will not start. f) She owns an old Volvo estate, but g) she frequently borrows her mother-in-law's Metro. h) It was her mother-in-law who sold her the Volvo, i) so she feels guilty when it doesn't work.

The unmarked organisation of a discourse is as a set of sentences with given information preceding new information – b), c), d) above. Given information is naturally pronominalized, ellipsed, etc.

Theme/Rheme or Topic/Comment are terms often used to talk about this level of linguistic organisation. These terms are distinct from Subject/Predicate(ie. syntactic NP/VP) and Given/New. For example, theme is defined as 'the communicative point of departure for the rest of the clause'. In b) *They* is grammatical subject, theme, and given information. In e) *On winter mornings* is new information, not grammatical subject, and therefore a 'marked' theme. Passives can function to 'thematise' a NP which cannot occur as subject otherwise; eg. in d) *he* is patient ('takee' not 'taker') of *take* and would therefore normally be the object. This is natural here because *he* is given and *a colleague* is new information.

Focus is a term used to refer to the linguistic expression which conveys the information which is the focus of attention. This can be signalled by prosodic stress in speech, or by particular syntactic constructions; for example, in h) focus is on *her mother-in-law* in this so-called it-cleft construction. This is a marked situation, in normal cases focus is often on all the VP/new information – wide vs. narrow focus. Focus extends 'backwards' from nuclear stress up to but not including the theme. Nuclear (roughly strongest) stress usually occurs on the last contentful (stressable) word of the sentence.

7.4 Intentionality

Not all discourses exhibit the type of discourse structure exemplified above. For example, below is a perfectly coherent discourse which contains no explicit anaphoric links. Its coherence derives from recognising the intentions of the participants:

A: Pint, please. B: Bitter? A: Tetleys. B: 1 pound 80 please. A: Thanks.

Therefore, other researchers have argued that structure is a side-effect rather than essential clue to discourse coherence and have explored the possible dis-

course 'moves' which can be made – rhetorical/discourse coherence relations. For example, (36b) is intended as an elaboration of a) and we resolve the links between the two sentences because we recognise it as such (not because of structural clues such as focus).

(36) a Kim can open the safe.
b He knows the combination.

Other relations include *narrative* – the default, *explanation*, *contrast*, etc. There are between about 12 and 60 depending on whose theory you adopt and whether these are just useful labels or have (default) entailments (as with thematic relations) is controversial.

7.5 Ellipsis

People prefer to make their utterances short and exploit context and what they think their interlocutors know about the context to achieve this. Ellipsis goes one step further than anaphora in that constituents are simply left out and assumed 'understood' given the context, as in (37):

(37) a A: Would you like to go for lunch? B: Yes (I would)
b A: How many students are there? B: 21
c A: Would gave what to whom? B: Well, Kim, a pen to Sandy and Sandy, a bone to Fido, I think
d A: Who got married last weekend? B: Well, Kim didn't

What is left out in each case?

7.6 Implicature

Interlocutors do more inference than deductive entailment on the basis of what is actually said in discourse interpretation:

Grice's Maxims of Conservation:

Cooperative Principle: make your contribution helpful given the purpose(s) of the conversation
Quality: make it true
Quantity: make it informative enough, but not more
Relevance: make it relevant
Manner: avoid obscurity and ambiguity

Apparent failure to follow, these maxims (conventions) leads to conversational implicature, as in (38)

(38) a A: Where's Sandy B: Her car is gone
b A: Do you know the way? B: Here's a map

The inference that these are relevant answers is driven by A's assumption that B is being cooperative.

7.7 Exercises

Take a short paragraph from a newspaper, novel or texbook and for each sentence in the paragraph, identify the speech act(s) conveyed, given/new information, the topic and focus, any anaphoric or deictic constituents and ellipsis.

8 Further Reading

Jurafsky, D. and Martin, J. *Speech and Language Processing*, Prentice-Hall / Pearson International, 2009
is the core book for the NLP modules and contains short introductions to relevant areas of linguistics – my references are to the 2009 edition but the others contain substantially the same material – see http://www.cs.colorado.edu/ martin/slp.html,
https://web.stanford.edu/ jurafsky/slp3/

There are many introductory linguistics texts. One good one is:
Yule, G. *The Study of Language*, Cambridge University Press, any ed.

A good place to look up linguistic terms you don't know or have forgotten is:
Trask, R.L. *A Dictionary of Grammatical Terms in Linguistics*, Routledge, 1999 – still available, or try *Wikipedia*, or more recently and aimed at NLPers / Computational Linguists:
Bender, E. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*, Morgan & Claypool, 2013.

A good book amongst many on distributional, generative syntactic analysis is:
Burton-Roberts, N. *Analysing Sentences*, Longman, 1998
We won't adopt the same analysis of every construction discussed in this book but it teaches you how to *do* syntactic analysis (if you do some of the exercises).

A good first gentler introduction to semantics is:
Kearns, K. *Semantics*, MacMillan Press, 2000.

A better but harder introduction to semantics is:
Cann, R. *Formal Semantics*, Cambridge University Press, 1993.

A very good book on pragmatics is:
Levinson, S. *Pragmatics*, Cambridge University Press, 2000.

A more up-to-date textbook which covers the integration of semantics with discourse interpretation and word meaning is:
Cann, R. Kempson, R. and Gregoromichelaki, E. *Semantics: an introduction to meaning in language*, Cambridge University Press, 2009.

The best light read on linguistic theory is:
Pinker, S. *The Language Instinct*, Penguin, 1994 / 2007.