The Grammatical Analysis of Sentences

Chris Mellish and Graeme Ritchie

Constituents and composition

If we examine the form of English sentences (and comparable observations can be made in other languages) it seems that there are certain regularities in the structure of the sentence, in terms of where words may occur (their *distribution*, in linguistic terminology) and how words and phrases may combine with each other. For example, if we compare quite dissimilar sentences such as:

> Elbert adores spaghetti.
> There was a storm.

there is a general pattern "Subject – Verb – Complement" in which the "Subject" is some sort of self-contained phrase, the "Verb" is one of a particular class of words which behave in certain ways (e.g. varying their endings depending on what the Subject is), and the "Complement" is another phrase of some sort. Such regularities are quite widespread, within phrases as well as in sentence structure, and appear in sentences with quite a wide variety of meanings (as in the two examples above). This has led to the idea that there are regularities which are purely *syntactic* (or *grammatical*), and that some rules can be formulated to describe these patterns in a way that is largely independent of the meanings of the individual sentences. The assumption (or intention) is that the problem of making sense of a sentence can be usefully decomposed into two separate aspects — *syntax* (which treats these broad structural regularities) and *semantics* (which specifies how these groups of items mean something). The advantage of such a set-up would be that the rules describing the syntactic groupings would be very general, and not "domain-specific"; that is, they would apply to sentences regardless of what subject-matter the sentences were describing. Also, having a statement of what "chunks" were in the sentence (phrases, clauses, etc.), would simplify the task of defining what the meaning of the whole sentence was.

In terms of processing a sentence to extract its meaning, this corresponds to the (extremely common) idea that the analysis can be decomposed into two stages. A few NLP programs perform the input translation in a single stage (so-called "conceptual" or "semantic" parsing), but more often the task is split into two phases — "syntactic analysis" (or "parsing") and "semantic interpretation".

The first stage uses grammatical (syntactic) information to perform some structural preprocessing on the input, to simplify the task of the rules which compute a symbolic representation of the meaning. This preprocessing stage is usually known as *parsing*, and could be roughly defined as "grouping and labelling the parts of a sentence in a way that displays their relationships to each other in a useful way". The question then arises —

useful for what? That is, what criteria are relevant to defining what the internal structure of a sentence might be? One common answer to this (and the one which we shall adopt here) is that the structure built by the parser should be a suitable input to the semantic interpretive rules which will compute the "meaning" of the sentence (in some way that will not be considered in this document – see other parts of the course).

That may seem a rather obvious answer, but it is worth noting that within mainstream twentieth-century linguistics, it was quite commonplace to assume that sentences (and phrases) had an internal structure which could be defined and determined in non-semantic terms. It was held that there were purely syntactic relationships between parts of a sentence ("constituents"), and a linguistic technique called *immediate constituent analysis* consisted of trying to segment a sentence into nested parts which reflected (in some intuitive way) this natural grouping. For example, the sentence

The man ate the large biscuit

was typically grouped as:

((The man) (ate (the (large biscuit))))

or sometimes as:

((The man) (ate) (the (large biscuit)))

For more complicated sentences, the "natural grouping" or "intuitive syntactic structure" is more difficult to decide. It could be argued that it is impossible to talk of a natural grouping without considering meaning. When someone segments a sentence as in the above example, perhaps it is the semantic groupings which are being sketched. That is, the bracketting is an attempt to display the fact that the meaning of "the large biscuit" is composed of the meaning of "the" and the meaning of "large biscuit", and that the latter is made up of the meaning of "large" and "biscuit" joined together. Most linguistic research assumes, either explicitly or implicitly, that the meaning of a sentence is composed, in some way, from the meaning of its parts (an idea often attributed to the nineteenth century philosopher Frege), and so it is natural to devise syntactic structures that reflect these groupings of items into larger meaningful units. This idea of *compositional semantics* (i.e. making up the meaning of the whole from the meaning of the parts) is very widespread, and it is one of the guidelines which will be adopted here in deciding on suitable syntactic structures.

The other criterion for deciding on suitable segmentations and labellings of a sentence (when constructing a parser or a set of syntactic rules) is the overall simplicity of the syntactic description. If a particular part of the sentence (e.g. the subject position) seems to allow certain kinds of phrases and another position (e.g. object position) allows the same variations, then it is neater to give a name to this kind of item (e.g. noun phrase), and describe it separately; then the two positions can be specified as allowing that class of item. In programming terms, this is analogous to separating out a self-contained and commonly-occurring section as a named procedure.

This notion of regularity of structure is also a justification for the two-stage approach. Without considering any particular semantic analysis of English, it can be seen that there are certain general patterns in the structure of sentences (e.g. a subject phrase followed by a verb), so it is worthwhile making use of them to sort out the overall layout of the sentence; that is what the "parser" does.

A *grammar* is a set of rules which describes which sequences of words are valid sentences of a language. Usually, the rules will also indicate in some way an *analysis* or *structure* for the sentence; that is, information about what the component parts of the sentence are, and how they are linked together (see comments above about about bracketting parts of a sentence to show its structures). On this course, we shall be studying some very precise notations for grammar rules, which allow grammars to be used computationally in analysing sentences (inside a parser), but first we must clarify the nature of this endeavour, and we will also look at some of the types of words, phrases, and clauses used in analysing English sentences.

Why Syntax?

Newcomers to computational linguistics (or even linguistics) are sometimes suspicious of the proposal that we should consider grammar. With its overtones of "learning to talk properly", the notion of grammar has unfortunate associations for many people. It is worthwhile, therefore, considering why we study at syntax when we are interested in building computer systems that understand language.

Natural languages are infinite - there are infinitely many English sentences that we have never heard but which we will understand immediately if we ever do hear them. How is this possible? Our brains are only of limited size, and so we can't store all possible sentences and their meanings. The only way to handle all the possibilities is to have principles about how longer and longer sentences can be constructed and how their structure can be decoded in a general way to yield meaning. At the heart of this is knowledge of the syntax of the language. There does not seem to be any alternative.

From a practical point of view, in a natural language understanding system there seems to be no alternative to an (implicit or explicit) analysis of the syntactic structure of a sentence taking place before its meaning can be grasped. A syntactic analysis is useful because:

- It provides a hierarchical set of groupings of words and phrases which can be the basis for a general-purpose, finite and *compositional* procedure to extract meaning from any sentence of the language. For instance, if we wish to find the meaning of (1):

  (1) Poetry is displayed with the "verse" environment.

  we need to have some model of how the meanings of the individual words conspire together to produce the meaning of the whole. A syntactic analysis tells us that phrases

like 'Poetry', 'with the "verse" environment' and 'is displayed with the "verse" environment' are meaning-bearing items in their own right (because they fill distinct slots in possible sentence patterns), whereas phrases like 'with the' and 'Poetry is' are not such good candidates for breaking down the meaning into smaller parts.

- Different possible semantic readings of a sentence can often be ascribed to different possible syntactic analyses, and hence syntactic analysis provides an important basis for the enumeration of possible interpretations. For instance, the two possible readings of (2):

  (2) The explosives were found by a security man in a plastic bag.

  (one of which would be most unlikely in most contexts) correspond to the two following (legal) ways to group the words in 'found by a security man in a plastic bag':

  found by (a security man in a plastic bag)
  (found by a security man) in a plastic bag

- A detailed characterisation of the structure of possible sentences can serve to eliminate possible interpretations, syntactically, semantically and pragmatically:

  (3) He saw the rope under the boxes which was just what he needed.
  (4) Never throw your dog a chicken bone.
  (5) Ross looked at him in the mirror.

  The fact that in (3) it is not possible that it was the boxes that were needed can be put down to the unacceptability of the phrase 'the boxes which was ...', and this can be explained by the failure in this case of the principle of *number agreement* between a subject and its verb. In (4), semantically there is always the possibility that we are talking about throwing dogs to bones. A look at the way sentences are built, however, reveals that the pattern 'throw X Y' is related semantically to 'throw Y to X' (a principle sometimes known as *dative movement*), and this observation provides easy disambiguation here. Finally, in (5) the structural relationship between 'Ross' and 'him' prevent both of these phrases referring to the same individual (otherwise the reflexive 'himself' would have been used). This is one of a number of constraints on coreference which can be described in terms of syntactic structure.

Writing a Grammar

In developing a grammar, one has to devise a suitable set of grammatical categories to classify the words and other constituents which may occur. It is important to understand that the mnemonic names given to these categories (e.g. "noun phrase") are essentially arbitrary, as it is the way that the labels are used in the rules and in the lexicon that gives significance to them. If we labelled all our noun phrases as "aardvarks", our grammar would work just as well, providing that we also used the label "aardvark" in all the appropriate

places in the rules and in the lexicon (dictionary) . (It might be a less readable grammar, of course). The issue is the same as that of using mnemonic symbols when writing computer programs; systematically altering the names of all the user-defined procedures in a program makes no difference to the operation of the program, it merely alters its clarity to the human reader.

You might think that there is a single agreed set of categories for describing English grammar, and perhaps even an agreed "official" grammar. Neither of these are the case. Although there are certain common, traditional terms ("noun", "verb", etc.) the exact usage of these terms is not officially defined or agreed, so it is the responsibility of each grammar-writer to use these terms in a consistent way. It is usually best to use such familiar terms in a way which approximates traditional informal usage, to avoid confusing people, but there are no hard and fast conventions. The set of grammatical categories which used to be taught in schools, and which is used in language-teaching texts, is very rough, informal, and not nearly subtle enough for a large, precise, formal grammar of a natural language, since there are many more distinctions that have to be made in a real parser than can be reflected by a dozen or so (mutually exclusive) classes such as "noun", "verb", "adjective", etc.

It follows from the above remarks that what the grammar-writer has to do is try to work out what sorts of words and other constituents there are in the language, and how they interact with each other. It is this sorting out of the data, and detecting the regularities in it, which is the main task; making up names for the entities thus postulated is the least of the problem.

It is worth knowing about a newer orthodoxy in this area, within generative linguistics. Largely as a result of Chomsky's work on transformational generative grammar, there has been a vast amount of fairly formal descriptive linguistics carried out since about 1960, and a repertoire of terminology has grown up within that work which augments the old-fashioned informal set of terms. That is, as a result of trying to write fairly detailed grammars, academic linguists found various other classes which were useful to describe what was happening. In fact, only a small number of these innovations were labels for syntactic constituents. More often, each of the terms in this jargon was for a particular *construction*; that is, a particular way of organising the structure of a sentence or phrase. We will try to avoid the complications of introducing these more esoteric terms, but we shall rely on a few fairly standard syntactic labels, which are given below.

To many people, the term "grammar" is associated with rules taught at school, prescribing "the correct way to write English". This is not the sense in which "grammar" is used in lingustics and A.I. — we are not concerned with *prescriptive* grammar ("what the speaker/writer ought to do"), but with *descriptive* grammar ("what a typical speaker/writer actually does (subject to certain idealisations)"). That is, we are trying to write down a detailed description of the observed characteristics of English. Notice that this is also slightly different from the use of grammar in the description of programming languages. A programming language is an artificial system which is under our control and which we can define by specifying its grammar. The programming language has no other existence apart from the formal rules which define it. A natural language, on the other

hand, is an existing phenomenon whose workings are not known, and which we attempt to describe as best we can by writing grammars which give close approximations to its behaviour, in roughly the same way that a physicist tries to formulate equations that characterise the observed behaviour of physical phenomena. It is important to bear this in mind — no one knows exactly what the rules of English are.

Thus, when reading a linguistic discussion, it is important to realise that what is often going on is the design of a grammar, and the "decisions" being discussed (e.g. "should we class this as a relative clause or as a prepositional phrase?") are about the rules that would fit the data best.

A formally defined grammar G (i.e. a set of symbolic rules) of a language describes which sentences are possible; this is known as *the language generated by G*, sometimes written "L(G)". The aim of the grammar writer is to make this set of sentences as close as possible to the given natural language, That is, L(G) should "fit" the language as exactly as possible. The grammar is said to be *weakly adequate* if it generates (i.e. defines as well-formed) all the sentences of the natural language, and excludes all non-sentences. However, since we are also interested in constructing structural descriptions of sentences, it is not enough simply to sift out the sentences from the non-sentences — the grammar should, as far as possible, be *strongly adequate*, in the sense that it assigns correct syntactic decompositions to the sentences of the language.

A further constraint on the grammar is what is sometimes called "simplicity" or "elegance". There have been attempts to make this notion precise and formal (e.g. suggestions that some way of counting the numbers of rules and symbols in a grammar would give a measure of how "simple" it was), but these have generally not been very successful. Normally, linguists employ an intuitive notion of "elegance" in assessing alternative grammars, in a way rather similar to that used by programmers to compare possible programming solutions.

The concern with having an adequate grammar may seem excessively pedantic for those who are not primarily concerned with the grammar as a theory of how the language works, but there are practical reasons for wanting to get the grammar right. If the grammar does not assign correct labels and structures to items, it may cause problems for later semantic processing:

- by causing incorrect meanings to be assigned to sentences;
- by accepting and assigning structures to sentences which are in fact ungrammatical (i.e. not in the language);
- by assigning extra (incorrect) possible structures to sentences, thereby creating spurious ambiguity.

Capturing Regularities

What does all this imply for the person who has to construct a working NL processing system? There are various computer-based grammars around, which may or may not be

suitable for a particular application. If you have to write your own grammar, your design may have to be influenced by *two* sorts of factor: syntactic patterns (such as the fact that the typical English sentence consists of a subject phrase of some sort followed by a verbal group and possibly other material) and semantic regularities (for example, if two radically distinct meanings are possible for a construction, you may have to allow two different syntactic analyses for it – see discussion elsewhere in the course on Ambiguity). You will also want the grammar to be as short and elegant as possible, whilst describing as much as possible of the language. For this, the grammar will have to reflect regularities in the language where they exist. There are a number of guidelines that can be useful for the grammar-writer in producing something that is useful and extensible, rather than complex and ad-hoc. These include the following:

- **Substitutability**. Consider what happens if you take part of a complex phrase and substitute something else in its place. If the result is still an acceptable phrase then this suggests there is some similarity between the original and its substitute. If the substitution can be made in many different contexts, then one might hypothesise that the two phrases can be described by the same category. Thus, for instance, one could "define" a noun phrase as being any phrase which, when substituted for "John" in an acceptable sentence, yields another acceptable sentence. Usually this kind of argumentation only works up to a point - for instance the result of substituting "John's friends" for "John" in "John was really mad" is not as acceptable as the original, even though one would like to say that "John's friends" is a noun phrase.

- **Conjoinability**. It is generally thought that two constituents can most naturally be joined with "and" (or "or") if they are of the same type. That is two Noun Phrases will conjoin very naturally, but a Noun Phrase and a Prepositional Phrase will not. Hence we could argue that "smoking" and "bad diet" are of the same type (probably Noun Phrases) in:

  > Bad diet and smoking were his downfall.

  On the other hand, a slightly odd or humorous effect is caused by conjoining two dissimilar phrases:

  > She arrived in a hurry and a long silk dress.

  This is a rather difficult criterion to apply, as the oddity may result not from mixing different types of constituents, but from mixing different "roles" that the constituents play in the sentence semantically.

- **Semantics**. If two phrases have the same kind of meaning (e.g. both refer to physical objects, or actions) then it is plausible to give them the same syntactic category. If the semantic analysis will involve the meaning of one sequence of words modifying or augmenting the meaning of another then it is plausible to regard them as separate

phrases that are joined together at some point in the phrase structure tree. Sometimes one would like to explain semantic ambiguity in terms of there being multiple syntactic possibilities. There are many ways in which semantic considerations can affect the way one designs a grammar.

Grammatical relations

In describing the parts of an English sentence, it is traditional and often useful to label the roles which various phrases (or clauses) play in the overall structure (as opposed to saying what sort of shape they themselves have internally). The commonest labels used in this way (which we shall use *very informally* on this course when indicating portions of text), are as follows.

Subject . At the front of an English sentence, there can be a self-contained phrase or clause, such as:

The president opened the building.
He ran up the stairs.

Informally, this is in some sense the entity about which the sentence is saying something, but that is difficult to characterise precisely in the case of sentences like:

It is raining.

where there is certainly a grammatical subject "it", even though it is unclear what it refers to.

Object After certain kinds of verbs (known as *transitive* verbs), there can be a phrase (or clause) usually describing the entity acted upon, or created, or directly affected, by the process described by the verb, such as:

The president opened *the building*.
He imagined *what might happen*.

This is often called the *direct object* to emphasise the difference from the indirect object (below).

Indirect Object Again occurring after the verb this phrase or clause is also some fairly central participant in the process described by the verb, but more obliquely than the direct object:

The president presented the prize *to the athlete*.
The president gave *the athlete* the prize.
The president gave *to charities*.

Verbs which take both a direct and an indirect object are sometimes called "ditransitive".

Complement Although this phrase sometimes has a precise technical meaning within particular linguistic theories, here we shall use it very roughly to mean "an additional item which forms part of some central construction". For example :

> He is *extremely silly.*
> They want *to go home.*
> Being happy *to live in peace* is *very desirable.*

In the last example, the phrase "to live in peace" might be said to be a complement to "happy", and "very desirable" could be classed as the complement of "is".

Modifier This even vaguer term refers to words, phrases or clauses attached to other words, etc., to refine the meaning.

> The man *on the roof* had a *badly injured* hand.

Both the emphasised phrases in this example are modifiers (of "the man" and "hand" respectively; also "badly" is a modifier of "injured".

The four main Lexical Categories

The four main *lexical* categories in English (that is, those categories that are filled by individual words) are noun, verb, adjective and preposition.

Nouns

Traditionally, a noun is "a naming word", but this doesn't tell you much since the entity "named" can be fairly abstract, and some names don't count as (common) nouns. Examples: 'snow', 'unicorn', 'sideboard', 'measles', 'beauty'. Common nouns typically have distinct singular and plural forms (e.g. 'carton', 'cartons'), although certain nouns don't show both possibilities, (e.g. 'sheep', 'deer'). These *count* nouns refer to entities or sets where the objects are regarded as separate, but there is also a class of *mass* nouns, e.g. 'furniture', 'toast', which look grammatically singular but describe some collection of stuff which is not separated into objects you can count. Confusingly, the latter can often be used as common nouns with the meaning "type of" - e.g. "there are three silver paints available today".

Adjectives

Traditionally, an adjective is a "describing word". It can attach to a noun to modify its meaning, or be used to assert some attribute of the subject of a sentence. Examples: 'blue', 'large', 'fake', 'main'. Adjectives typically have a base form, an adverb form, a comparative and a superlative, e.g 'great', 'greatly', 'greater' and 'greatest', though many adjectives are defective (i.e. lack some of these possible forms). For example, some adjectives don't have the latter two forms, e.g. 'unique', 'uniquely'. Also, some adjectives build the comparative

Description	*be*	*write*	*bake*
base form	be	write	bake
infinitive	to be	to write	to bake
finite present 1sing	am	write	bake
finite present 3sing	is	writes	bakes
finite present other	are	write	bake
finite past sing	was	wrote	baked
finite past plur	were	wrote	baked
past participle	been	written	baked
present participle	being	writing	baking
passive participle	???	written	baked

Figure 1: Different forms of verbs

and superlative forms using 'more' and 'most', e.g. 'beautiful', 'more beautiful', 'most beautiful'.

Verbs

Traditionally, a verb is a "doing word", but this doesn't help much, since some verbs describe some rather passive forms of doing. Examples: "run", "know", "be", "have". We can distinguish between ten different forms that a verb can take (Figure 1). In practice, verbs have at most eight distinct forms - for the verb 'be' there are eight, for 'write' there are six, and for 'bake' there are only five.

Syntactically, the important distinction is between the finite forms (present 3rd person singular 'writes', other present 'write', past 'wrote') and the non-finite forms (bare infinitive 'write', infinitive 'to write', present participle 'writing', past participle 'written', passive participle 'written'). Basically, finite verbs are what you would expect to see in a simple, single-verb sentence; the other forms all combine with other verbs to make more complex constructions, or appear in noun-phraser-like constructions referring to the action. The distinction between the use of a word as a past participle or passive participle reflects the type of construction it appears in ('has baked', vs. 'was baked') — in English these different functions are not distinguished in the form of the verb.

Auxiliary verbs are those odd little verbs that behave differently, and which can be put in a sequence at the front of a verb phrase: 'be', 'have', 'do', 'can', 'will', 'may', 'might', 'could', 'must', 'shall', 'should'. The last 8 are known as modal verbs. The word 'to' can also be considered an auxiliary verb in phrases like "to run" (though it can also behave as a preposition when in front of a noun phrase).

Prepositions

These can attach to the front of a noun phrase to form a prepositional phrase. Examples: 'in', 'by', 'of', 'to'. Note that 'to' can also behave as an auxiliary verb, as in the complex

verb phrase 'to have done'.

Syntactic Categories

We have highlighted the categories noun, verb, adjective and preposition because each of these generally forms the focus of a larger *phrasal category* having a distinctive structure. Thus, from nouns we can build noun phrases, from verbs, verb phrases, and so on. We can often recognise categories intermediate between the lexical categories and the phrasal categories they correspond to. To describe these phenomena, some modern syntactic theories use the notion of *bar level*. Lexical categories (e.g. noun) have bar level 0, phrasal categories (e.g. noun phrase) have bar level 2 or sometimes 3, and intermediate categories are notated accordingly (this approach is called "X-bar theory"). One advantage of this sort of notation is that it makes clear the regularities in the structure of different phrases – rules can be devised which roughly say that "a phrase of type X and bar level N consists of an X at bar level (N-1) plus certain other items"; this general rule might then cover a wide variety of phrasal types.

A *specifier* is a part of a phrase which (if it appears at all) appears only once, at the very beginning. As one works inwards from the outside of the phrase, one next encounters optional modifiers (of which there may be several), and then finally the lexical category and the complements it requires. Modifiers and complements are discussed in more detail later.

Noun Phrases

A noun phrase is any phrase which can act as a complete subject, object, etc. in a sentence; e.g. "The big red block", "Most of the first three coaches". Noun phrases are typically used to refer to objects, but note the use of the dummy NPs 'there' and 'it', as in (1) and (2).

(1) There is a dog howling in the yard
(2) It is impossible for me to see you now

Pronouns are usually abbreviated references to objects that have recently been mentioned or are somehow available from the context: 'it', 'he', 'she', 'I', 'them'. The form of a pronoun is generally determined by the role it is playing in the sentence, often known as its *case* (Figure 2). **Proper nouns/proper names** name specific items, whereas most nouns (common nouns) are viewed as naming some generic class of items or substance. Example proper nouns are 'John', 'Kate Bush', 'Scotland', 'AI2'.

Determiners are the usual specifiers of noun phrases. A determiner can be an **article** ('the', 'a' or 'an'), a possessive pronoun ('his', etc), a quantifier ('many', 'some'), a possessive noun phrase ('my father's'), or a demonstrative ('this' , 'that').

Modifiers of noun phrases take the form of pre-modifiers (e.g. adjectives) and post-modifiers (e.g. prepositional phrases and relative clauses). A **Relative clause** comes after a noun phrase; e.g. as in 'The man *who you saw* is here'. It can start with or without

Role	Traditional case name	Form	Example
subject	nominative	he	*He* came to tea.
(direct) object	accusative	him	John saw *him*.
indirect object	dative	(to) him	John gave *him* biscuits.
possessor	genitive	his	John saw *his* mother.

Figure 2: Examples of pronoun case - he/him/his

a relative pronoun ('who', 'which', 'that', 'where', 'when'). A relative clause is in form similar to a sentence, but with a noun phrase missing at some point.

Verb Phrases

A verb phrase is basically a verb plus its complement(s); e.g. 'gave the parcel to the clerk', 'runs'.

Verb phrases have no obvious specifiers. As modifiers they can have adverbs and prepositional phrases.

Adverbs are attached to a verb to qualify its meaning. However, this sometimes tends to become something of an all-purpose class where difficult words (e.g. "only") can be put. Examples: "beautifully", "quickly".

Prepositional Phrases

A prepositional phrase may be required (for instance, by a verb that it comes after) to contain a particular preposition (see below on *subcategorisation*). There are not many possible forms for PPs in English, though adverbs can act as modifiers to PPs, as in 'directly above the window'.

Adjective Phrases

Adjective phrases usually consist of single adjectives, but it is possible for these to be accompanied by an indication of degree and some number of adverbs as modifiers, as in 'very commonly used'.

Complementation and Modification

The general intuition is that a particular lexical item licenses (i.e. is able to combine with) certain complements, which are more tightly bound to it than modifiers. The latter can typically occur in any number with any word of the class. We say that a certain item *subcategorises* (for) certain complements, and these have to be associated with that item in the lexicon in some way. For example, there are the traditional categories of verbs — an 'intransitive' verb (e.g. 'dream') expects no complements, whereas a 'transitive' verb

(e.g. 'hit') expects a single object noun phrase. But there are many more possibilities for what a verb, noun, adjective or preposition can require to follow it.

Another important characteristic of complements is that their head (i.e the item which they are attached to as complements) imposes *selectional restrictions* upon them. That is, a particular head makes sense only with a particular type of complement, *semantically* speaking. For example, 'kill' selects for a living object, but is not particular about its subject. 'Murder' requires as its subject an entity that can be considered responsible for its actions, while 'assassinate' requires the rather specialised semantic property "political figure" to be true of its object.

A given lexical item may occur in several different subcategorisation patterns. For instance, there is a class of verbs like 'give' that occur in the patterns "give y x" and "give x to y". It is a simple matter to enter these alternatives in the lexicon, or at least to have some process going on in the lexicon that produces them both from some single specification of the properties of the word.

Inadequacies of this grammatical survey

So far, we have summarised some of the basic aspects of a fragment of English. There are several areas in which the coverage of this grammar is totally deficient, however. We may cover approaches to some of these in later lectures, but for completeness, we will mention the more notable deficiencies in our review. These issues are being actively researched.

Morphology

We have just assumed that we get fully inflected forms out of the lexicon. That is, we have assumed implicitly that there are separate lexical entries for 'bake', 'bakes', 'baking', etc. and that all the possible properties are listed for each. There are general principles that describe how the different forms of regular verbs vary (e.g. to form plurals, nouns usually add 's'), and a realistic grammar implementation would need to use these directly in order to avoid a huge lexicon.

We have also not treated *derivational morphology*, that is, where one word is derived in a systematic way from one or several others. For instance, the word 'computability' can be analysed as being derived from 'compute'+'able'+'ity'.

Word Order

Many grammatical notations, including those we shall be using on this course, assume that a rule specifies a set of constituents to be concatenated in a strict order. This is not a major problem in writing grammars for English, a language with remarkably rigid word order compared with many others. However, it is not an apt assumption for the description of languages with freer word order. In Slavic languages, for instance, the major constituents of a sentence verb, subject, object, etc. may generally be freely permuted to carry discourse information of the sort associated with English articles ('a' vs. 'the', etc.).

Other languages have even freer word order, allowing what would normally be considered a constituent to be discontinuous. Examples of these so-called $w*$ languages include Classical Latin, and many of those spoken in Australia.

Subcategorisation alternations

There is no reason why a verb could not be entered in the lexicon several times with different indications of possible complements (subcategorisation). For instance, we could have alternative entries for 'give' to handle "give a dog a bone" and "give a bone to a dog".

However, these are alternations which are extremely predictable, and perhaps there should be some way of making use of this regularity to simplify the rules.

Unbounded Dependencies

There is a class of phenomena that are not purely local, but seem to involve a dependency between elements of the sentence that are separated by an unlimited number of clause boundaries. Examples are topicalisation (i.e. having a phrase at the front to indicate that it is the topic), relativisation (i.e. a relative clause in which an embedded clause supplies more details about a noun phrase), questions (using words like "who", "what" to ask about some described item) and "tough-movement" (an obscure construction which relies on using certain words like "easy", "hard"), as in (1)-(4). In these sentences, a constituent (indicated by *italics* here) seems to function (both grammatically and in terms of meaning) as if it were in a different position (marked here by $\diamond$). (These related items need not be at the start or end of the sentence.)

(1) *This book*, I could never manage to persuade my students to read $\diamond$.
(2) *The college* that I expected John to want Mary to attend $\diamond$ has closed.
(3) *What* do you believe Mary told Bill that John had said $\diamond$?
(4) *This film* is easy for me to persuade the children not to see $\diamond$.

However, the trend in modern linguistics is to account for these by means of a series of purely local dependencies, in which some information is passed step-by-step through the parts of the sentence. That is, the description of each clause within the sentence will include information about related items in any immediately enclosing (and hence nearby) clauses, and in this way a "chain" of any length can be constructed, from the "gap" to the "misplaced" item.

References

For background reading on English grammar, it is useful to consult Allen (1987), Ch.2 and Winograd (1983), Appendix B. Burton-Roberts (1986) is an excellent source of intuitions about grammar. Gazdar and Mellish(1989), Section 4.1, is also useful introductory material on what linguistics is about.

- Allen, J. (1987), *Natural Language Understanding*, Benjamin Cummings.
- Burton-Roberts, N. (1986), *Analysing Sentences: An Introduction to English Syntax*, Harlow Longman.
- Gazdar, G. and Mellish, C.(1989), *Natural Language Processing in PROLOG*. Addison-Wesley.
- Winograd, T. (1983), *Language as a Cognitive Process. Volume 1: Syntax*, Addison-Wesley.

Exercises

1. Give examples, in English, of a noun, a verb, an adjective, a preposition and an auxiliary verb.

2. In the sentence "John saw Mary", which phrase is the subject and which the object?

3. Give an example of where a verb subcategorises for a given complement. Do the same for a noun and an adjective.

4. (From Ex.2.4 of Winograd(1983)). Work out what the subcategorisations of the verbs:

   asked, preferred, condescended, promised, tried, considered, accepted, forced, expected , wanted, believed, hoped

   are, by trying to fit them into sentences which you think distinguish between them, such as:

   John ... to succeed.
   John ... his friend to be careful.
   Nobody was ... to be there.

5. The following fragments are all usually classed as "Noun Phrases":

   the many hooligans
   some of the hundred voters
   more than seven people
   many loyal voters

   but the following are not:

   *many the hooligans
   of the hundred many voters
   more voters loyal

Devise a list of all the valid configurations for Noun Phrases, using the following lexical categories:

- Adjectives: "loyal, green, happy, ..."
- Nouns: "voters, hooligans, people, apples, ..."
- Quantifiers: "some, many, all, ..."
- Articles: "the, a"
- Determiners: "the, a, his, her, its, ..."
- Preposition: "of"

For example, "Determiner-Adjective-Noun" is a valid sequence. You may find that the above classes are too broad, and need refinement. For instance, there may be sequences where 'the' is acceptable, but not 'a'; also, 'many' may behave slightly differently from 'some'. Refine the classification as you find it necessary.

In terms of processing a sentence to extract its meaning, this corresponds to the (ex-tremely common) idea that the analysis can be decomposed into two stages. A few NLP programs perform the input translation in a single stage (so-called â€œconceptualâ€ or â€œse-manticâ€ parsing), but more often the task is split into two phases â€” â€œsyntactic analysisâ€ (or â€œparsingâ€) and â€œsemantic interpretationâ€. The rst stage uses grammatical (syntactic) information to perform some structural preprocessing on the input, to simplify the task of the rules which compute a symbolic representation of the meaning. This p Transformational (generative) grammar. Noam Chomsky. In-built ability to master any grammatical structures and to generate an endless variety of grammatically correct sentences. Transforms. Paradigmatic approach: oppositions of sentences â€“ syntactic paradigm. Syntactic derivation â€“ a process consisting of elementary transformational steps: - morphological arrangement (predicate â€“ tense, voice, aspect, mood, number, case; subject â€“ number, case).