CSE6390 3.0 Special Topics in AI & Interactive Systems II
Introduction to Computational Linguistics
Insructor: Nick Cercone – 3050 CSEB – nick@cse.yorku.ca
Tuesdays,Thursdays 10:00-11:30 – South Ross 104
Fall Semester, 2010

Morphemes, Parts of Speech, POS Tagging, Corpus Linguistics

Some edited from Wikipedia, the free encyclopedia

In morpheme-based morphology, a morpheme is the smallest linguistic unit that has semantic meaning. In spoken language, morphemes are composed of phonemes (the smallest linguistically distinctive units of sound), and in written language morphemes are composed of graphemes (the smallest units of written language).

The concept *morpheme* differs from the concept *word*, as many morphemes cannot stand as words on their own. A morpheme is *free* if it can stand alone, or *bound* if it is used exclusively alongside a free morpheme. Its actual phonetic representation is the morph, with the different morphs representing the same morpheme being grouped as its allomorphs.

For example, in English the word "unbreakable" has three morphemes: "un-", a bound morpheme; "break", a free morpheme; and "-able", a bound morpheme. "un-" is also a prefix, "-able" is a suffix. Both "un-" and "-able" are affixes.

The morpheme plural-s has the morph "-s", /s/, in *cats* (/kæts/), but "-es", /ɨz/, in *dishes* (/dɪʃɨz/), and even the voiced "-s", /z/, in *dogs* (/dɒgz/). "-s". These are allomorphs.

Types of morphemes

Free morphemes like *town*, and *dog* can appear with other lexemes (as in *town hall* or *dog house*) or they can stand alone, i.e. "free".

Bound morphemes like "un-" appear only together with other morphemes to form a lexeme. Bound morphemes in general tend to be prefixes and suffixes. Unproductive, non-affix morphemes that exist only in bound form are known as "cranberry" morphemes, from the "cran" in that very word.

Derivational morphemes can be added to a word to create (derive) another word: the addition of "-ness" to "happy," for example, to give "happiness." They carry semantic information.

Inflectional morphemes modify a word's tense, number, aspect, and so on, without deriving a new word or a word in a new grammatical category (as in the "dog" morpheme if written with the plural marker morpheme "-s" becomes "dogs"). They carry grammatical information.

Allomorphs are variants of a morpheme, e.g. the plural marker in English is sometimes realized as /-z/, /-s/ or /-ɨz/.

Other variants

A null morpheme is a morpheme that is realized by a phonologically null affix (an empty string of phonological segments). In simpler terms, a null morpheme is an "invisible" affix. It's also called zero

morpheme; the process of adding a null morpheme is called *null affixation*, *null derivation* or *zero derivation*.

The root is the primary lexical unit of a word, which carries the most significant aspects of semantic content and cannot be reduced into smaller constituents. Content words in nearly all languages contain, and may consist only of, root morphemes. However, sometimes the term "root" is also used to describe the word minus its inflectional endings, but with its lexical endings in place. For example, *chatters* has the inflectional root or lemma *chatter*, but the lexical root *chat*. Inflectional roots are often called stems, and a root in the stricter sense may be thought of as a monomorphemic stem.

The traditional definition allows roots to be either free morphemes or bound morphemes. Root morphemes are essential for affixation and compounds. However, in polysynthetic languages with very high levels of inflectional morphology, the term "root" is generally synonymous with "free morpheme". Many such languages have a very restricted number of morphemes that can stand alone as a word: Yup'ik, for instance, has no more than two thousand.

The root of a word is a unit of meaning (morpheme) and, as such, it is an abstraction, though it can usually be represented in writing as a word would be. For example, it can be said that the root of the English verb form *running* is *run*, or the root of the Spanish superlative adjective *amplísimo* is *ampl-*, since those words are clearly derived from the root forms by simple suffixes that do not alter the roots in any way. In particular, English has very little inflection, and hence a tendency to have words that are identical to their roots. But more complicated inflection, as well as other processes, can obscure the root; for example, the root of *mice* is *mouse* (still a valid word), and the root of *interrupt* is, arguably, *rupt*, which is not a word in English and only appears in derivational forms (such as *disrupt*, *corrupt*, *rupture*, etc.). The root *rupt* is written as if it were a word, but it's not.

This distinction between the word as a unit of speech and the root as a unit of meaning is even more important in the case of languages where roots have many different forms when used in actual words, as is the case in Semitic languages. In these, roots are formed by consonants alone, and different words (belonging to different parts of speech) are derived from the same root by inserting vowels. For example, in Hebrew, the root *gdl* represents the idea of largeness, and from it we have *gadol* and *gdola* (masculine and feminine forms of the adjective "big"), *gadal* "he grew", *higdil* "he magnified" and *magdelet* "magnifier", along with many other words such as *godel* "size" and *migdal* "tower".

In linguistics, a stem (sometimes also theme) is a part of a word. The term is used with slightly different meanings. In one usage, a stem is a form to which affixes can be attached.[1] Thus, in this usage, the English word *friendships* contains the stem *friend*, to which the derivational suffix *-ship* is attached to form a new stem *friendship*, to which the inflectional suffix *-s* is attached. In a variant of this usage, the root of the word (in the example, *friend*) is not counted as a stem.

In a slightly different usage, which is adopted in the remainder of this article, a word has a single stem, namely the part of the word that is common to all its inflected variants.[2] Thus, in this usage, all derivational affixes are part of the stem. For example, the stem of *friendships* is *friendship*, to which the inflectional suffix *-s* is attached.

Stems may be roots, e.g. *run*, or they may be morphologically complex, as in compound words (cf. the compound nouns *meat ball* or *bottle opener*) or words with derivational morphemes (cf. the derived verbs *black-en* or *standard-ize*). Thus, the stem of the complex English noun *photographer* is *photo·graph·er*, but not *photo*. For another example, the root of the English verb form *destabilized* is *stabil-*, a form of *stable* that does not occur alone; the stem is *de·stabil·ize*, which includes the derivational affixes *de-* and *-ize*, but not the inflectional past tense suffix *-(e)d*. That is, a stem is that part of a word that inflectional affixes attach to.

The exact use of the word 'stem' depends on the morphology of the language is question. In Athabaskan linguistics, for example, a verb stem is a root that cannot appear on its own, and that carries the tone of the word. Athabaskan verbs typically have two stems in this analysis, each preceded by prefixes.

Morphological analysis

In natural language processing for Japanese, Chinese and other languages, morphological analysis is a process of segmenting given sentence into a row of morphemes. It is closely related to Part-of-speech tagging, but word segmentation is required for these languages because word boundaries are not indicated by blank spaces. Famous Japanese morphological analysers include Juman, ChaSen and Mecab.

Parts of Speech Table

This is a summary of the 8 parts of speech*. You can find more detail if you click on each part of speech.

part of speech	function or "job"	example words	example sentences
Verb	action or state	(to) be, have, do, like, work, sing, can, must	EnglishClub.com **is** a web site. I **like** EnglishClub.com.
Noun	thing or person	pen, dog, work, music, town, London, teacher, John	This is my **dog**. He lives in my **house**. We live in **London**.
Adjective	describes a noun	a/an, the, 69, some, good, big, red, well, interesting	My dog is **big**. I like **big** dogs.
Adverb	describes a verb, adjective or adverb	quickly, silently, well, badly, very, really	My dog eats **quickly**. When he is **very** hungry, he eats **really** quickly.
Pronoun	replaces a noun	I, you, he, she, some	Tara is Indian. **She** is beautiful.
Preposition	links a noun to another word	to, at, after, on, but	We went **to** school **on** Monday.
Conjunction	joins clauses or sentences or words	and, but, when	I like dogs **and** I like cats. I like cats **and** dogs. I like dogs **but** I don't like cats.
Interjection	short exclamation, sometimes inserted into a sentence	oh!, ouch!, hi!, well	**Ouch**! That hurts! **Hi**! How are you? **Well**, I don't know.

* Some grammar sources categorize English into **9** or **10** parts of speech. At EnglishClub.com, we use the traditional categorization of **8** parts of speech. Examples of other categorizations are:

Verbs may be treated as two different parts of speech:
Lexical Verbs (*work, like, run*)
Auxiliary Verbs (*be, have, must*)
Determiners may be treated as a separate part of speech, instead of being categorized under Adjectives

Words with More than One Job

Many words in English can have more than one job, or be more than one part of speech. For example, "work" can be a verb and a noun; "but" can be a conjunction and a preposition; "well" can be an adjective, an adverb and an interjection. In addition, many nouns can act as adjectives.

To analyze the part of speech, ask yourself: "What job is this word doing in this sentence?"

In the table below you can see a few examples. Of course, there are more, even for some of the words in

the table. In fact, if you look in a good dictionary you will see that the word **but** has six jobs to do:

verb, noun, adverb, pronoun, preposition and conjuction!

word	part of speech	example
work	noun	My **work** is easy.
	verb	I **work** in London.
but	conjunction	John came **but** Mary didn't come.
	preposition	Everyone came **but** Mary.
well	adjective	Are you **well**?
	adverb	She speaks **well**.
	interjection	**Well**! That's expensive!
afternoon	noun	We ate in the **afternoon**.
	noun acting as adjective	We had **afternoon** tea.

Part-of-speech tagging

In corpus linguistics, **part-of-speech tagging** (**POS tagging** or **POST**), also called grammatical tagging or word category disambiguation, is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context —ie. relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Once performed by hand, POS tagging is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags.

Principle

Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words can represent more than one part of speech at different times. This is not rare—in natural languages (as opposed to many artificial languages), a large percentage of word-forms are ambiguous. For example, even "dogs", which is usually thought of as a just a plural noun, can also be a verb:

The sailor dogs the hatch.

"Dogged", on the other hand, can be either an adjective or a past-tense verb. Just which parts of speech a word can represent varies greatly.

Schools commonly teach that there are 9 parts of speech in English: noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, and interjection. However, there are clearly many more categories and sub-categories. For nouns, plural, possessive, and singular forms can be distinguished. In many languages words are also marked for their "case" (role as subject, object, etc.), grammatical gender, and so on; while verbs are marked for tense, aspect, and other things.

In part-of-speech tagging by computer, it is typical to distinguish from 50 to 150 separate parts of speech for English, for example, NN for singular common nouns, NNS for plural common nouns, NP for singular proper nouns (see the POS tags used in the Brown Corpus). Work on stochastic methods for tagging Koine Greek (DeRose 1990) has used over 1,000 parts of speech, and found that about as many words were ambiguous there as in English. A morphosyntactic descriptor in the case of morphologically rich

languages can be expressed like **Ncmsan**, which means Category=Noun, Type = common, Gender = masculine, Number = singular, Case = accusative, Animate = no.

History

Research on part-of-speech tagging has been closely tied to corpus linguistics. The first major corpus of English for computer analysis was the Brown Corpus developed at Brown University by Henry Kucera and Nelson Francis, in the mid-1960s. It consists of about 1,000,000 words of running English prose text, made up of 500 samples from randomly chosen publications. Each sample is 2,000 or more words (ending at the first sentence-end after 2,000 words, so that the corpus contains only complete sentences).

The Brown Corpus was painstakingly "tagged" with part-of-speech markers over many years. A first approximation was done with a program by Greene and Rubin, which consisted of a huge handmade list of what categories could co-occur at all. For example, article then noun can occur, but article verb (arguably) cannot. The program got about 70% correct. Its results were repeatedly reviewed and corrected by hand, and later users sent in errata, so that by the late 70s the tagging was nearly perfect (allowing for some cases on which even human speakers might not agree).

This corpus has been used for innumerable studies of word-frequency and of part-of-speech, and inspired the development of similar "tagged" corpora in many other languages. Statistics derived by analyzing it formed the basis for most later part-of-speech tagging systems, such as CLAWS (linguistics) and VOLSUNGA. However, by this time (2005) it has been superseded by larger corpora such as the 100 million word British National Corpus.

For some time, part-of-speech tagging was considered an inseparable part of natural language processing, because there are certain cases where the correct part of speech cannot be decided without understanding the semantics or even the pragmatics of the context. This is extremely expensive, especially because analyzing the higher levels is much harder when multiple part-of-speech possibilities must be considered for each word.

In the mid 1980s, researchers in Europe began to use hidden Markov models (HMMs) to disambiguate parts of speech, when working to tag the Lancaster-Oslo-Bergen Corpus of British English. HMMs involve counting cases (such as from the Brown Corpus), and making a table of the probabilities of certain sequences. For example, once you've seen an article such as 'the', perhaps the next word is a noun 40% of the time, an adjective 40%, and a number 20%. Knowing this, a program can decide that "can" in "the can" is far more likely to be a noun than a verb or a modal. The same method can of course be used to benefit from knowledge about following words.

More advanced ("higher order") HMMs learn the probabilities not only of pairs, but triples or even larger sequences. So, for example, if you've just seen an article and a verb, the next item may be very likely a preposition, article, or noun, but much less likely another verb.

When several ambiguous words occur together, the possibilities multiply. However, it is easy to enumerate every combination and to assign a relative probability to each one, by multiplying together the probabilities of each choice in turn. The combination with highest probability is then chosen. The European group developed CLAWS, a tagging program that did exactly this, and achieved accuracy in the 93-95% range.

It is worth remembering, as Eugene Charniak points out in *Statistical techniques for natural language parsing* [1], that merely assigning the most common tag to each known word and the tag "proper noun" to all unknowns, will approach 90% accuracy because many words are unambiguous.

CLAWS pioneered the field of HMM-based part of speech tagging, but was quite expensive since it enumerated all possibilities. It sometimes had to resort to backup methods when there were simply too many (the Brown Corpus contains a case with 17 ambiguous words in a row, and there are words such as "still" that can represent as many as 7 distinct parts of speech).

In 1987, Steven DeRose and Ken Church independently developed dynamic programming algorithms to solve the same problem in vastly less time. Their methods were similar to the Viterbi algorithm known for some time in other fields. DeRose used a table of pairs, while Church used a table of triples and an ingenious method of estimating the values for triples that were rare or nonexistent in the Brown Corpus (actual measurement of triple probabilities would require a much larger corpus). Both methods achieved accuracy over 95%. DeRose's 1990 dissertation at Brown University included analyses of the specific error types, probabilities, and other related data, and replicated his work for Greek, where it proved similarly effective.

These findings were surprisingly disruptive to the field of natural language processing. The accuracy reported was higher than the typical accuracy of very sophisticated algorithms that integrated part of speech choice with many higher levels of linguistic analysis: syntax, morphology, semantics, and so on. CLAWS, DeRose's and Church's methods did fail for some of the known cases where semantics is required, but those proved negligibly rare. This convinced many in the field that part-of-speech tagging could usefully be separated out from the other levels of processing; this in turn simplified the theory and practice of computerized language analysis, and encouraged researchers to find ways to separate out other pieces as well. Markov Models are now the standard method for part-of-speech assignment.

The methods already discussed involve working from a pre-existing corpus to learn tag probabilities. It is, however, also possible to bootstrap using "unsupervised" tagging. Unsupervised tagging techniques use an untagged corpus for their training data and produce the tagset by induction. That is, they observe patterns in word use, and derive part-of-speech categories themselves. For example, statistics readily reveal that "the", "a", and "an" occur in similar contexts, while "eat" occurs in very different ones. With sufficient iteration, similarity classes of words emerge that are remarkably similar to those human linguists would expect; and the differences themselves sometimes suggest valuable new insights.

These two categories can be further subdivided into rule-based, stochastic, and neural approaches. Some current major algorithms for part-of-speech tagging include the Viterbi algorithm, Brill Tagger, Constraint Grammar, and the Baum-Welch algorithm (also known as the forward-backward algorithm). Hidden Markov model and visible Markov model taggers can both be implemented using the Viterbi algorithm.

References

1. Charniak, Eugene. 1997. "Statistical Techniques for Natural Language Parsing". *AI Magazine* 18(4):33–44.
2. Hans van Halteren, Jakub Zavrel, Walter Daelemans. 2001. Improving Accuracy in NLP Through Combination of Machine Learning Systems. *Computational Linguistics*. 27(2): 199–229. PDF
3. DeRose, Steven J. 1990. "Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages." Ph.D. Dissertation. Providence, RI: Brown University Department of Cognitive and Linguistic Sciences.
4. DeRose, Steven J. 1988. "Grammatical category disambiguation by statistical optimization."¶ Computational Linguistics 14(1): 31–39. [2]

External links

- Overview of available taggers
- Cypher A natural language transcoder that performs POS-tagging, morphological processing, lexical analysis, to produce RDF and SPARQL from natural language

- Resources for Studying English Syntax Online
- CLAWS
- LingPipe Java natural language processing software including trainable part-of-speech taggers with first-best, n-best and per-tag confidence output.
- OpenNLP Tagger LGPL Tagger based on maxent maximum entropy package
- CRFTagger Conditional Random Fields (CRFs) English POS Tagger
- JTextPro A Java-based Text Processing Toolkit
- Citar LGPL C++ Hidden Markov Model trigram POS tagger, a Java port named Jitar is also available
- Ninja-PoST PHP port of GPoSTTL, based on Eric Brill's rule-based tagger

Corpus linguistics

Corpus linguistics is the study of language as expressed in samples *(corpora)* or "real world" text. This method represents a digestive approach to deriving a set of abstract rules by which a natural language is governed or else relates to another language. Originally done by hand, corpora are now largely derived by an automated process, which is corrected.

The corpus approach runs counter to Noam Chomsky's view that real language is riddled with performance-related errors, thus requiring careful analysis of small speech samples obtained in a highly controlled laboratory setting[*citation needed*].

The problem of laboratory-selected sentences is similar to that facing lab-based psychology: researchers do not have any measure of the ethnographic representativity of their data.

Corpus linguistics does away with Chomsky's *competence/performance* split; adherents believe that reliable language analysis best occurs on field-collected samples, in natural contexts and with minimal experimental interference. Within CL there are divergent views as to the value of corpus annotation, from John Sinclair[1] advocating minimal annotation and allowing texts to 'speak for themselves', to others, such as the Survey of English Usage team (based in University College, London)[2] advocating annotation as a path to greater linguistic understanding and rigour.

History

A landmark in modern corpus linguistics was the publication by Henry Kucera and Nelson Francis of *Computational Analysis of Present-Day American English* in 1967, a work based on the analysis of the Brown Corpus, a carefully compiled selection of current American English, totalling about a million words drawn from a wide variety of sources. Kucera and Francis subjected it to a variety of computational analyses, from which they compiled a rich and variegated opus, combining elements of linguistics, language teaching, psychology, statistics, and sociology. A further key publication was Randolph Quirk's 'Towards a description of English Usage' (1960)[3] in which he introduced The Survey of English Usage.

Shortly thereafter, Boston publisher Houghton-Mifflin approached Kucera to supply a million word, three-line citation base for its new *American Heritage Dictionary*, the first dictionary to be compiled using corpus linguistics. The AHD made the innovative step of combining prescriptive elements (how language *should* be used) with descriptive information (how it actually *is* used).

Other publishers followed suit. The British publisher Collins' COBUILD monolingual learner's dictionary, designed for users learning English as a foreign language, was compiled using the Bank of English. The Survey of English Usage Corpus was used in the development of one of the most important Corpus-based Grammars, the *Comprehensive Grammar of English* (Quirk *et al.* 1985)[4].

The Brown Corpus has also spawned a number of similarly structured corpora: the LOB Corpus (1960s British English), Kolhapur (Indian English), Wellington (New Zealand English), Australian Corpus of English (Australian English), the Frown Corpus (early 1990s American English), and the FLOB Corpus (1990s British English). Other corpora represent many languages, varieties and modes, and include the International Corpus of English, and the British National Corpus, a 100 million word collection of a range of spoken and written texts, created in the 1990s by a consortium of publishers, universities (Oxford and Lancaster) and the British Library. For contemporary American English, work has stalled on the American National Corpus, but the 400+ million word Corpus of Contemporary American English (1990-present) is

now available through a web interface.

The first computerized corpus of transcribed spoken language was constructed in 1971 by the Montreal French Project,[5] containing one million words, which inspired Shana Poplack's much larger corpus of spoken French in the Ottawa-Hull area.[6]

Besides these corpora of living languages, computerized corpora have also been made of collections of texts in ancient languages. An example is the Andersen-Forbes database of the Hebrew Bible,[7] developed since the 1970s,[8] in which every clause is parsed using graphs representing up to seven levels of syntax,[9] and every segment tagged with seven fields of information.[10]

Methods

Corpus Linguistics has generated a number of research methods, attempting to trace a path from data to theory. Wallis and Nelson (2001)[11] first introduced what they called the 3A perspective: Annotation, Abstraction and Analysis.

Annotation consists of the application of a scheme to texts. Annotations may include structural markup, part-of-speech tagging, parsing, and numerous other representations.

Abstraction consists of the translation (mapping) of terms in the scheme to terms in a theoretically motivated model or dataset. Abstraction typically includes linguist-directed search but may include e.g., rule-learning for parsers.

Analysis consists of statistically probing, manipulating and generalising from the dataset. Analysis might include statistical evaluations, optimisation of rule-bases or knowledge discovery methods.

Most lexical corpora today are part-of-speech-tagged (POS-tagged). However even corpus linguists who work with 'unannotated plain text' inevitably apply some method to isolate terms that they are interested in from surrounding words. In such situations annotation and abstraction are combined in a lexical search.

The advantage of publishing an annotated corpus is that other users can then perform experiments on the corpus. Linguists with other interests and differing perspectives than the originators can exploit this work. By sharing data, corpus linguists are able to treat the corpus as a locus of linguistic debate, rather than as an exhaustive fount of knowledge.

References

1. Sinclair, J. 'The automatic analysis of corpora', in Svartvik, J. (ed.) *Directions in Corpus Linguistics (Proceedings of Nobel Symposium 82)*. Berlin: Mouton de Gruyter. 1992.
2. Wallis, S. 'Annotation, Retrieval and Experimentation', in Meurman-Solin, A. & Nurmi, A.A. (ed.) Annotating Variation and Change. Helsinki: Varieng, [University of Helsinki]. 2007.e-Published
3. Quirk, R. 'Towards a description of English Usage', *Transactions of the Philological Society*. 1960. 40-61.
4. Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. *A Comprehensive Grammar of the English Language* London: Longman. 1985.
5. Sankoff, D. & Sankoff, G. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In Darnell R. (ed.) *Canadian Languages in their Social Context* Edmonton: Linguistic Research Incorporated. 1973. 7-64.
6. Poplack, S. The care and handling of a mega-corpus. In Fasold, R. & Schiffrin D. (eds.) *Language Change and Variation*, Amsterdam: Benjamins. 1989. 411-451.
7. ^ Gons, Phil (2008-08-14), *Tutorial Videos for Andersen-Forbes Syntax Resources*, logos.com, retrieved 2009-04-04
8. ^ Eyland, E. Ann (1987), "Revelations from Word Counts", in Newing, Edward G.; Conrad, Edgar W., *Perspectives on Language and Text: Essays and Poems in Honor of Francis I. Andersen's Sixtieth Birthday, July 28, 1985*, Winona Lake, IN: Eisenbrauns, p. 51, ISBN 0-931464-26-9
9. ^ Evans, Eli (2005-11-16), *Syntax: Andersen-Forbes Introduction*, logos.com, retrieved 2009-04-01
10. ^ Andersen, Francis I.; Forbes, A. Dean (2003), "Hebrew Grammar Visualized: I. Syntax", *Ancient Near Eastern Studies* **40**: 43–61
11. ^ Wallis, S. and Nelson G. 'Knowledge discovery in grammatically analysed corpora'. *Data Mining and Knowledge Discovery*, **5**: 307-340. 2001.

Journals

There are several international peer-reviewed journals dedicated to corpus linguistics, for example, Corpora, Corpus Linguistics and Linguistic Theory, ICAME Journal and the International Journal of

Corpus Linguistics.

Book series
Book series in this field include Language and Computers, Studies in Corpus Linguistics and English Corpus Linguistics

Other

Biber, D., Conrad, S., Reppen R. *Corpus Linguistics, Investigating Language Structure and Use*, Cambridge: Cambridge UP, 1998. ISBN 0-521-49957-7

McCarthy, D., and Sampson G. *Corpus Linguistics: Readings in a Widening Discipline*, Continuum, 2005. ISBN 0-826-48803-X

Facchinetti, R. *Theoretical Description and Practical Applications of Linguistic Corpora*. Verona: QuiEdit, 2007 ISBN 978-88-89480-37-3

Facchinetti, R. (ed.) *Corpus Linguistics 25 Years on*. New York/Amsterdam: Rodopi, 2007 ISBN 978-90-420-2195-2

Facchinetti, R. and Rissanen M. (eds.) *Corpus-based Studies of Diachronic English*. Bern: Peter Lang, 2006 ISBN 3-03910-851-4

External links

- AskOxford.com the composition and use of the Oxford Corpus
- Bookmarks for Corpus-based Linguists -- very comprehensive site with categorized and annotated links to language corpora, software, references, etc.
- Corpora discussion list
- Freely-available, web-based corpora (100 million - 400 million words each): American (COCA), British (BNC), TIME, Spanish, Portuguese
- Manuel Barbera's overview site
- Przemek Kaszubski's list of references
- DMCBC.com
- Datum Multilanguage Corpora Based on chinese free sample download
- Corpus4u Community a Chinese online forum for corpus linguistics
- McEnery and Wilson's Corpus Linguistics Page
- Corpus Linguistics with R mailing list
- Research and Development Unit for English Studies
- Survey of English Usage
- The Centre for Corpus Linguistics at Birmingham University
- Gateway to Corpus Linguistics on the Internet: an annotated guide to corpus resources on the web
- Biomedical corpora
- Linguistic Data Consortium, a major distributor of corpora
- Penn Parsed Corpora of Historical English
- Corsis: (formerly Tenka Text) an open-source (GPLed) corpus analysis tool
- ICECUP and Fuzzy Tree Fragments
- Research and Development Unit for English Studies
- Discussion group text mining