Corpus-Derived Profiles: A new framework for the analysis of word relations in corpora

Gregory Garretson
Boston University / Mittuniversitet
gregory@bu.edu

Semantics seminars
Department of English
Stockholm University
March 31, 2009

Overview

- Prologue: introductory remarks
- Theoretical background
  - Syntagmatic relations and meaning
  - The notion of collocation
    - Aspects of collocation
    - Existing definitions
    - Collocation in the CDP framework
- The CDP framework
  - The form and content of the framework
  - An implementation of the framework: CenDiPede
  - Demo: using CenDiPede to answer some research questions
- Epilogue: Conclusions, questions and discussion

My background

- Doing PhD in Applied Linguistics at Boston University
  - Previously did MA there
- Now teach linguistics and English here in Sweden
  - Most recently at Mittuniversitetet (Mid-Sweden University)
- Shift in approach over the years: generative linguistics → empirical linguistics

Overlapping areas of linguistics

- This thesis involves various fields of linguistics:
  - Corpus linguistics, computational linguistics, semantics, lexical semantics, discourse/pragmatics
- Some relevant theoretical questions:
  - What is a word? What is a lexical item? What does it *mean* for a word to have meaning?
  - What aspects of a word's meaning have to do with its co-occurrence relations with other words? How much of a text's meaning come from groups of words chosen as one unit?
  - How can the choice of a word affect not just the meaning of the sentence, but the meaning of the discourse?
- Note: This is a work in progress! All aspects are subject to change, and all input is welcome.

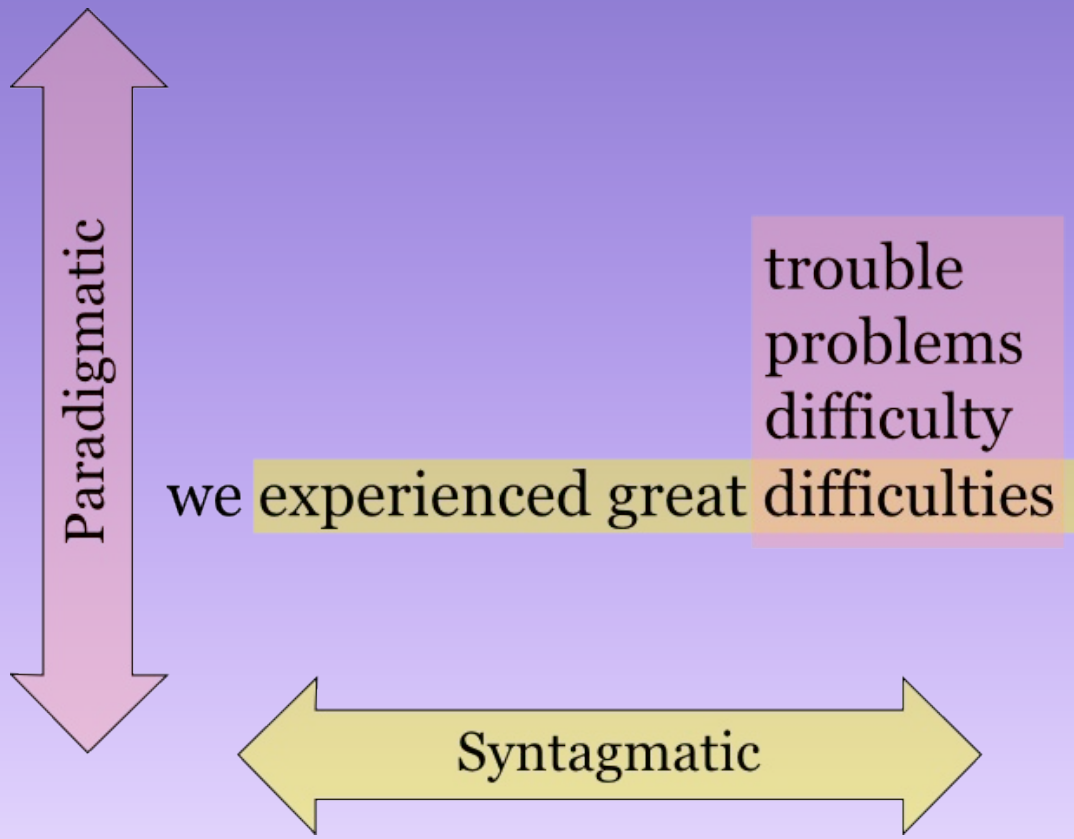
Theoretical background

Firth: meaning by collocation

- This project owes a great deal to the research tradition of scholars such as J. R. Firth, John Sinclair, Michael Hoey, etc.
- For Firth, **meaning** was central to linguistics:
  - The disciplines and techniques of linguistics are directed to assist us in making statements of meaning. Indeed, the main concern of descriptive linguistics is to make statements of meaning. (Firth 1957:190)
- A key part of his theory of meaning was **collocation:**
  - You shall know a word by the company it keeps! (Firth [1957] 1968:179)
  - One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, colocation with *night*. (Firth [1951] 1957:196)

Vertical vs. horizontal meaning

- Two dimensions of meaning:
  - **Paradigmatic** or “vertical” → the “**open-choice principle**”
  - **Syntagmatic** or “horizontal” → the “**idiom principle**”



Sinclair: meaning and form

- John Sinclair:
  - “[…] the tradition of linguistic theory has been massively biased in favour of the *paradigmatic* rather than the *syntagmatic* dimension.” (Sinclair 2004:140)
  - “The problem is that a text is a unique deployment of meaningful units, and its particular meaning is not adequately accounted for by any organized concatenation of the fixed meanings of each unit. This is because some aspects of textual meaning arise from the particular combinations of choices at one place in the text, and there is no place in the lexicon-grammar model where such meanings can be assigned.” (Sinclair 2004:134)
  - “It is asserted that form and meaning cannot be separated because they are the same thing. Considered in relation to other forms, a lexical item is a form; considered in relation to other meanings, it is a meaning.” (Sinclair 2004:139)

Sinclair's "lexical item"

- Sinclair laid out a model of how a **lexical item** might be described, "a model which reconciles the paradigmatic and syntagmatic dimensions of choice at each choice point" (Sinclair 2004:141). This model of a lexical item contains several components:
  - These begin with *collocation*, the co-occurrence of words, and go on to *colligation*, […] defined as the co-occurrence of words with grammatical choices, then *semantic preference*, which is the co-occurrence of words with semantic choices, and *semantic prosody*. The semantic prosodies express attitudinal and pragmatic meaning; they are the junction of form and function. The reason why we choose to express ourselves in one way rather than another is coded in the prosody, which is an obligatory component of a lexical item. (Sinclair 2004:174)
- In this talk we are going to confine our focus to **collocation**.

The notion of collocation

- Basic idea behind collocation: words tend to co-occur with certain other words.
  - Examples: What can you say about the word *vim*? What about the word *solve*? What about the word *nagging*?
- But: different scholars use different definitions of collocation. The concept is more complex than it might seem.
- I have identified the following “aspects of a theory of collocation”:
  - Ontology
  - Symmetry
  - Paradigmaticity
  - Significance
  - Scope

Ontology

- Question behind ontology:
  - What sort of phenomenon is collocation? Is it a *textual* phenomenon or a *psychological* phenomenon?
- One view:
  - Collocation is an observable property of words in text.
- Another view:
  - Collocation is a relation between words in the mental lexicon that leads to observable patterns in texts.

Symmetry (1)

- Question behind symmetry:
  - Is collocation a one-way or a two-way relation? That is, are these statements equivalent?
    - Word A is a collocate of Word B
    - Word B is a collocate of Word A
  - This question can be concretized in three ways...

Symmetry (2)

- **First**, does it matter whether the collocate is more common than the node or less common than the node?
  - One view:
    - *Vim* and *vigor* are collocates of each other.
  - Another view:
    - *Vim* is a “downward collocate” of *vigor*, since it is less common, while *vigor* is an “upward collocate” of *vim* (see Sinclair 1991).
- **Second**, does it matter whether the collocate precedes or follows the node?
  - Does it matter whether “dark” comes before “night”?
- **Third**, does it matter whether the collocate grammatically modifies the node or vice-versa?
  - Does it matter that “dark” modifies/is a dependent of “night”?

Paradigmaticity

- Question behind paradigmaticity:
  - Is collocation a relation between *surface forms* or *underlying abstractions*?
  - More concretely, is collocation a relation between *word forms* (e.g., follow, follows, followed) or a relation between *lemmas* (e.g., FOLLOW)?
- One view:
  - The word *people* can be expected to have different collocations from the word *person*, and should be studied separately.
- Another view:
  - The form *follows* should be studied together with the form *follow*, since the morphological differences are fairly arbitrary.

Significance (1)

- Question behind significance:
  - How frequently do two words need to co-occur to be considered collocates?
- Less commonly:
  - more than once in a text
  - more than ten times in the corpus
- More commonly:
  - significantly more often than predicted by chance

Significance (2)

- Why refer to **chance**?
  - Frequency of co-occurrence is compared to the frequency we could expect if the words were **randomly distributed.**
  - Difference must meet some criterion of **significance.**
  - This is a well-established method in the use of statistical measures.
  - The assumption that words are randomly distributed in text is rather daft, but so far it's the best we have.
- The next question is then:
  - What test do we use to determine whether the frequency is greater than chance?
  - Many tests have been used: *t*-test, mutual information, chi-square, log-likelihood, z-score, etc.
    - These all give somewhat different results.

Scope

- Question behind scope:
  - How far apart in the text can two words be and still be considered collocates?
- There are actually two dimensions to scope:
  - **Distance**
    - *Collocates may be separated by* 1 word, 5 words, 10 words, 100 words, etc.
  - **Structure**
    - *Collocates must co-occur within the same* text OR paragraph OR sentence OR clause OR phrase OR direct syntactic relation.

Review

- Aspects of a theory of collocation:
  - Ontology
  - Symmetry
  - Paradigmaticity
  - Significance
  - Scope
- Now let’s see how these are realized in some definitions in the literature…

Collocation: neighbors in the text

- John Sinclair:
  - Collocation is the co-occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening. (Sinclair 1991:170).
  - Collocation (at present) is the co-occurrence of words with no more than four intervening words [...]. (Sinclair 2004:141)

Collocation: a grammatical relation

- Göran Kjellmer:
  - A collocation is a sequence of words that occurs more than once in identical form [in a corpus] and which is grammatically well-structured (Kjellmer 1987:133).
  - Note that this definition includes collocations such as:
    - *to be, one of, had been, United States, in a moment,* etc.

Collocation: psycholinguistic phenomenon

- Michael Hoey:
  - […] collocation is […] a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution. This definition is intended to pick up on the fact that collocation is a psycholinguistic phenomenon, the evidence for which can be found statistically in computer corpora. (Hoey 2005:5)

Collocation: semantically opaque

- Sabine Bartsch:
  - Collocations are lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other. (Bartsch 2004:76)
- For Bartsch, collocates must:
  - be within a **span** of 3 (or 5) words
  - co-occur **significantly** according to one of three tests:
    - Mutual information, *t*-score, chi-square
  - be in a **direct syntactic relation** with each other
  - display lexically and/or pragmatically **constrained lexical selection**
    - OR have an element of **semantic opacity** such that the meaning of the collocation cannot be said to be deducible from the meanings of the constituents

Collocation in the CDP framework (1)

- Ontology
  - Collocation is seen as a property of **words in texts**. This yields the following advantages:
    - Collocation becomes measurable and quantifiable, and may be stated exactly.
    - Precise statements may be made about the differences between corpora, genres, and text types.
    - It ameliorates the problem of making statements about individuals (i.e., about the mental lexicon) based on aggregate data drawn from many individuals.
  - Certainly there exists a related psychological association between words in the mental lexicon, but I would prefer to call this **lexical association**. It is not directly measurable by corpus methods.
    - It is not even clear that experimental methods (i.e., elicitation studies) measure this directly.

Collocation in the CDP framework (2)

- Symmetry
  - Collocation is seen as a **symmetric** relation unless grammatical dependency is involved, in which case it is **asymmetric**. Information is collected on sequencing and relative frequencies but does not have direct theoretical importance.
- Paradigmaticity
  - Collocation is considered to be a relation between **forms**, not between lemmas, though groups of forms may be considered together when appropriate.

Collocation in the CDP framework (3)

- Significance
  - Three statistical tests are used:
    - *t*-score, mutual information, log-likelihood
  - The results of the three tests are combined into a **composite score**, by which the collocates are ranked.
  - Advantages:
    - Counteracts the idiosyncrasies of the different tests.
    - Composite score has an integer value between 0 and 100—easy to work with.
  - Note: All three **component scores** plus the composite score are recorded in a lexical profile.

Collocation in the CDP framework (4)

- Scope
  - The CDP framework defines **four different types of collocates**:
    - **Paragraph collocate:** within the same paragraph as the node
    - **Sentence collocate:** within the same sentence as the node
    - **Neighborhood collocate:** within 5 words of the node
    - **Dependency collocate:** in direct grammatical relation with node
  - Every profile presents data on *all four types of collocates.*
    - Overlap is considerable, but not total.
    - Provide slightly different pictures of the word's context
    - Researcher is free to focus on the kind s/he is interested in.

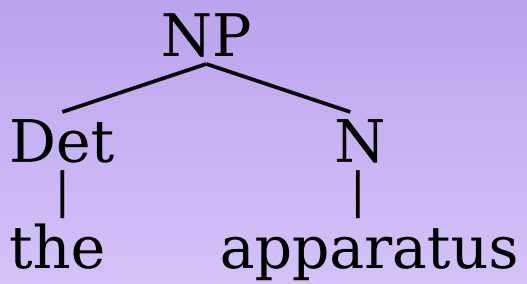
The Corpus-Derived Profiles framework

The CDP framework is...

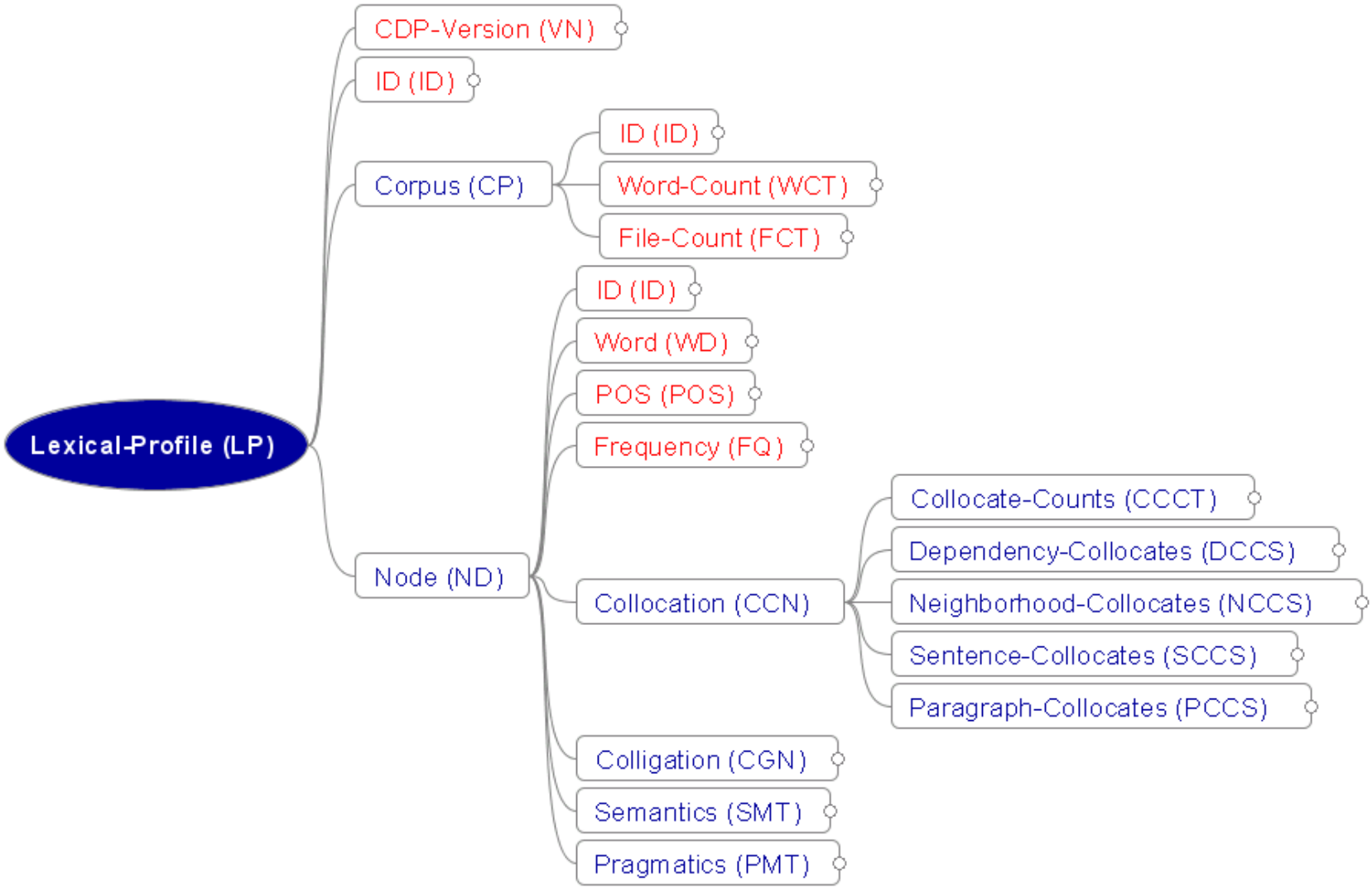
- A framework for studying **syntagmatic word relations**
- A system for **profiling** a given word in a given corpus
- A system for **comparing** different words' profiles
- A system for collecting information of the following types:
  - **Collocation:** the words that collocate with the node
    - dependency, neighborhood, sentence, paragraph collocates
  - **Colligation:** the syntactic categories the node associates with
  - **Semantic preference:** the semantic fields the node associates with
  - **Semantic prosody:** information on the favorability of the node, based on the favorability of its collocates
- A system that may be **fully implemented** in a computer program.

Anatomy of a corpus-derived profile

- A CDP (or “lexical profile”) is essentially a collection of attributes.
- Each CDP has a *tree structure*. Every node is one of three things:
  - A **category**
  - An **attribute name**
  - An **attribute value**
- This is not unlike a syntactic tree:



Overview of a CDP



Querying CDPs

- Most interesting use of CDPs: comparing them
  - Allows us to see differences between two or more words' syntagmatic behavior.
- The CDP framework defines a *query language* for comparing profiles.
  - Given one or more profiles, a query will yield answers to questions about those profiles.
- For example:
  - Which of these nouns has the greatest propensity to occur as the subject of a verb?
  - Which of these verbs collocate with nouns related to pain?

The CDP Query language

- Here are some example queries; anything in italics is a well-formed query:
  - *find every Dependency-Collocate where Joint-Frequency is greater than 20 and sort descending by Joint-Frequency*
    - **OR:** *find all DCC where JFQ > 20 and sort by JFQ*
  - *find all Neighborhood-Collocate where MI-score rank is less than 101 and show LL-Score, T-Score and sort by MI-score*
    - **OR:** *find all NCC where MI < 101 and show LL, TS and sort by MI*
  - *find all non-overlap between Paragraph-Collocate and Neighborhood-Collocate where POS is SUBST and display max 50*
    - **OR:** *find all non-overlap between PCC and NCC where POS = SUBST and display max 50*

Comparisons along two axes

1. It's possible to compare **two or more words in the same corpus**:
   - How do the words "problems" and "trouble" differ?
   - How does the *noun* "deal" differ from the *verb* "deal"?
   - What nouns collocate most strongly with "of"?
2. It's also possible to compare **the *same* word in different corpora/subcorpora**:
   - How is "thing" used differently in the spoken and written parts of the BNC?
   - How is the noun "light" used differently in British and American English?
   - How are adverbs used differently in student writing (say, the USE corpus) and in the academic component of the BNC?

Advantages of such a framework

1. It makes it easier to **describe** syntagmatic relations rigorously and unambiguously.
   - The CDP framework attempts to *unify* the description of collocation, colligation, semantic preference, and semantic prosody.
2. It makes it easier to **communicate** research findings by establishing a common terminology of description.
   - "Using version 1 of the CDP framework, I studied all neighborhood collocates of part_SUBST in the BNC with a composite score of 50 or greater."
3. The framework is designed to be **implemented computationally**, thereby facilitating research greatly.

CenDiPede: an implementation

- The program **CenDiPede** is *an implementation* of the CDP framework in Java.
- Features:
  - Stupid name
  - Creates CDPs for chosen corpus (currently only BNC)
  - Can create batches of profiles all at once
  - KWIC and Paragraph concordance views
  - Query language allows queries of profiles
  - Outputs data in different formats
  - Pleasing GUI
  - Cross-platform
  - Free and open-source
  - Will (eventually) work with any corpus, including your data

The tools behind CenDiPede

- CenDiPede is currently made up of 13,000 lines of Java code (that's 300 pages when printed). But beyond this, it makes use of a lot of other publicly available tools:
  - **CLAWS** (Garside & Smith 1997) part-of-speech tagger
  - **WordNet** database (Fellbaum 1998) + JWI Java interface
  - Gerold Schneider's **Pro3Gres** dependency-grammar parser (Schneider 2007), which in turn relies upon:
    - a chunker
    - a tagger
    - a lemmatizer
    - a head-extractor

Concessions and limitations

- The state of the art still requires certain compromises:
  - Division of words into parts of speech:
    - Every word must fall into a category in the CLAWS tagset.
    - But note: We use the simplified tagset.
  - Certain levels of error:
    - In POS-tagging, in parsing, in assigning of dependency relations
    - Even in original corpus compilation
  - Limited categories of WordNet
    - Only nouns, verbs, adjectives, and adverbs
    - Only (!) 155,287 word forms covered
  - Inherent difficulty of assigning semantic prosody

Demo of CenDiPede

- PLEASE SEE THE SCREENCAST ON THE SAME WEBPAGE
- Here are some research questions...
  - What does the noun *apparatus* mean? How many senses does it have? What sort of semantic prosody does it exhibit? How is it used differently in academic writing and in newspaper text?
  - What are the collocates of *time*? What are the most common expressions it occurs in, both in academic writing and in conversation?
  - What is the difference between *sort*, *type*, and *kind*? Do they collocate with different words? Which one is most common in spoken data? How many uses does each one exhibit?

Prospective uses of the framework

- Here are some examples of ways I can envision the CDP framework being used in the future:
  - To study **a particular word** in great depth
  - To (easily) compare the differences in "behavior" of a set of **semantically related words** (e.g. synonyms, antonyms, hyponyms, etc.)
  - To compare **all the nouns** in the BNC in search of new groupings based on syntagmatic properties
  - To compare **word usage** in a learner corpus with word usage in a reference corpus, perhaps to make suggestions for more natural usage.
  - To **analyze texts** in terms of the overall strength of the semantic prosody the constituent words exhibit.

Conclusions

- The CDP framework is a **theoretical framework** for studying **syntagmatic lexical relations** in corpora.
  - Goal: expand the empirical basis for linguistic theorizing
  - Assumption: Co-occurrence relations are central to meaning in language.
- CenDiPede is an **implementation** of the framework that allows one to **create CDPs** and **query** them.
  - CenDiPede is cross-platform, free, and open-source.
- CDPs are not seen as an end result of research, but rather an **empirical basis** from which to perform research.
  - For example, they make it easy to compare different words in the same corpus, or the same word in different corpora.
- Plus: There are several potential **applied uses** of the framework

Questions I would ask you

- Does this framework appear to have any utility for your particular areas of study?
- What kind of data would you want to use it with?
- Is there other information that you would want to see included in a CDP?
- What kinds of queries would you like to be able to make?
- It is theoretically possible to list thousands of collocates. How many is enough?
- How much work are software users willing to do? How long are they willing to wait for results?

References (1)

Bartsch, S. 2004. *Structural and Functional Properties of Collocations in English.* Tübingen: Gunter Narr Verlag.

BNC. 2007. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Fellbaum, C. 1998. *Wordnet: An Electronic Lexical Database*. Cambride, MA: MIT Press.

Firth, J. R. 1968. *Selected papers of J.R. Firth, 1952-59.* Edited by F. R. Palmer. Bloomington: Indiana University Press.

Garretson, G. 2008. "Desiderata for linguistic software design." International Journal of English Studies 8:1 (special issue on "Software-aided Analysis of Language"). 67–94.

Garside, R. & Smith, N. 1997. "A hybrid grammatical tagger: Claws4". In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman. 102–121.

References (2)

Hoey, M. 2005. *Lexical priming: a new theory of words and language.* London; New York: Routledge.

Kjellmer, G. 1987. Aspects of English collocations. In Meijs, W. (ed.), *Corpus Linguistics and Beyond: Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora.* Amsterdam: Rodopi. 133-140.

Schneider, G. 2007. *Hybrid Long-distance Functional Dependency Parsing*. Unpublished PhD thesis, Institute of Computational Linguistics, University of Zurich.

Sinclair, J. M. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. M. 2004. *Trust the text: language, corpus and discourse.* London: Routledge.

Extra slides

John Sinclair: meaning in form (1)

- 4 major points Sinclair stressed:

1. Linguistic theory should be based on **empirical evidence**.
   - [...] the contrast exposed between the impressions of language detail noted by people, and the evidence compiled objectively from texts is huge and systematic. It leads one to suppose that human intuition about language is highly specific, and not at all a good guide to what actually happens when the same people actually use the language. (Sinclair 1991:4)
2. Access to **corpora** radically changes the ways we can use linguistic data in doing linguistics.
   - [...] the ability to examine large text corpora in a systematic manner allows access to a quality of evidence that has not been available before. The regularities of pattern are sometimes spectacular and, to balance, the variation seems endless. The raw frequency of differing language events has a powerful influence on evaluation. (Sinclair 1991:4)

John Sinclair: meaning in form (2)

- 4 major points Sinclair stressed:

  3. Less emphasis should be placed on **grammaticality**, and more on **naturalness**.
     - [...] the term *naturalness* is simply a cover term for the constraints that determine the precise relationship of any fragment of text with the surrounding text. (Sinclair 1991:6)
  4. Meaning derives to a greater extent than usually acknowledged from **form**, or co-text.
     - There is ultimately no distinction between form and meaning (Sinclair 1991:7).

- If we agree with these assertions, we should see the **syntagmatic relations** observed in corpora as central to meaning in language.

The idiom principle

- Sinclair identified two tendencies in language that operate in tension with one another:
  - The **open-choice principle**:
    - At a choice point, a speaker has a number of options for filling a particular “slot” (e.g., choice of verb).
    - Also referred to as the “terminological tendency”.
  - The **idiom principle**:
    - “[…] a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments” (Sinclair 1991:110). (e.g., *of course*)
    - Also referred to as the “phraseological tendency”.

A new focus on “naturalness”

- John Sinclair:
  - “If we accept that the requirements of coherence and communicative effectiveness shape a text in many subtle ways, the term *naturalness* is simply a cover term for the constraints that determine the precise relationship of any fragment of text with the surrounding text.” (1991:6)
- Michael Hoey:
  - The problem with [previous linguistic theories] is that they account only for what is possible in a language and not for what is natural. This book is concerned, in part, with how naturalness is achieved and how an explanation of what is natural might impinge on explanations of what is possible.” (2005:2)
  - […] accoding to the theories of the lexicon that have dominated linguistic thought for the past 200 years there is no reason to regard the naturalness or clumsiness of […] sentences as being of any importance. (2005:6)

Choosing statistical tests

- Why MI score, *t*-score, and Log-Likelihood?
  - The **MI score** (pointwise mutual information) is a measure of the strength of the association between two words. It works well with sparse data.
    - BUT tends to prioritize low-frequency words
  - The ***t*-test** is a widely used test that compares the probability of two words co-occurring to chance.
    - BUT assumes normally distributed probabilities of co-occurrence.
  - The **Chi-squared test** doesn't assume a normal distribution, and works well with large probabilities.
    - BUT it doesn't work well with sparse data.
  - The **log-likelihood score** works well with sparse data and copious data.
    - BUT it is less widely used
    - BUT it is very complex to calculate

The CLAWS tagset

- The CLAWS simplified tagset and frequency in the BNC:
  - SUBST 25,504,973
  - VERB 17,869,010
  - PREP 12,848,079
  - ADJ 11,824,837
  - ART 8,694,665
  - PRON 7,909,809
  - ADV 6,508,896
  - CONJ 5,659,060
  - UNC 1,166,271
  - INTERJ 378,111

The CDP dependency categories

- The CDP dependency relation categories (which derive from the Pro3Gres dependency types) are these (and the inverse of each):

- PrepNounPhrase
- VerbSubject
- VerbDirectObject
- NounModifier
- NounDeterminer
- VerbPrepPhrase
- NounPrepPhrase
- ConjunctionConjoined
- VerbAuxiliary
- ModifiedAdverb
- VerbSubordClause
- ModifiedPrepObject
- VerbComplementizer
- VerbPredAdj
- NounRelClauseVerb
- ModifiedRelClauseObj
- ModifiedAppositive
- NounParticiple
- PossessorPossessee
- ModifiedRelClauseSubj
- ModifiedControlledSubj
- VerbAdjunct
- NounAdjective
- VerbPassiveSubject
- VerbSecondObj
- NounStrandedPrep
- ComparatorAdjective
- NounIngForm
- SubjectRelClauseVerb
- NounQuantifier
- ModifiedControlledObj
- AdjectiveDeterminer

The Pro3Gres parser

- Pro3Gres is “a robust, hybrid, deep-syntatic dependency-based parsing architecture” (Schneider 2007)
- Pro3Gres = PRObabilistic, PROlog-implemented, Parser-based Robust Grammatical Relations Extraction System (*ibid*)