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

Gregory Garretson

Boston University/Mid-Sweden University

gregory@bu.edu

Part 1: The Corpus-Derived Profiles framework

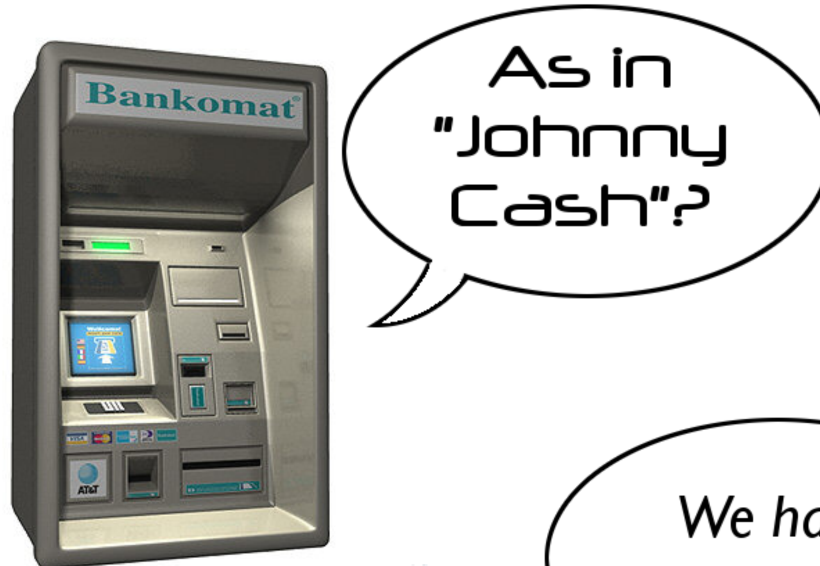
The project

- Goal 1: to create a **framework** for studying syntagmatic word relations that:
  - Comes close to instantiating Sinclair's syntagmatic model of a lexical item
  - Can be fully automatized
  - → Offers a useful *conceptual* tool for linguistic research
- Goal 2: create an **implementation** of this framework that:
  - Will work for any corpus
  - Will be cross-platform
  - Will be freely available
  - → Offers a useful *practical* tool for linguistic research

This is a work in progress—feedback is welcome!

What this is and isn't

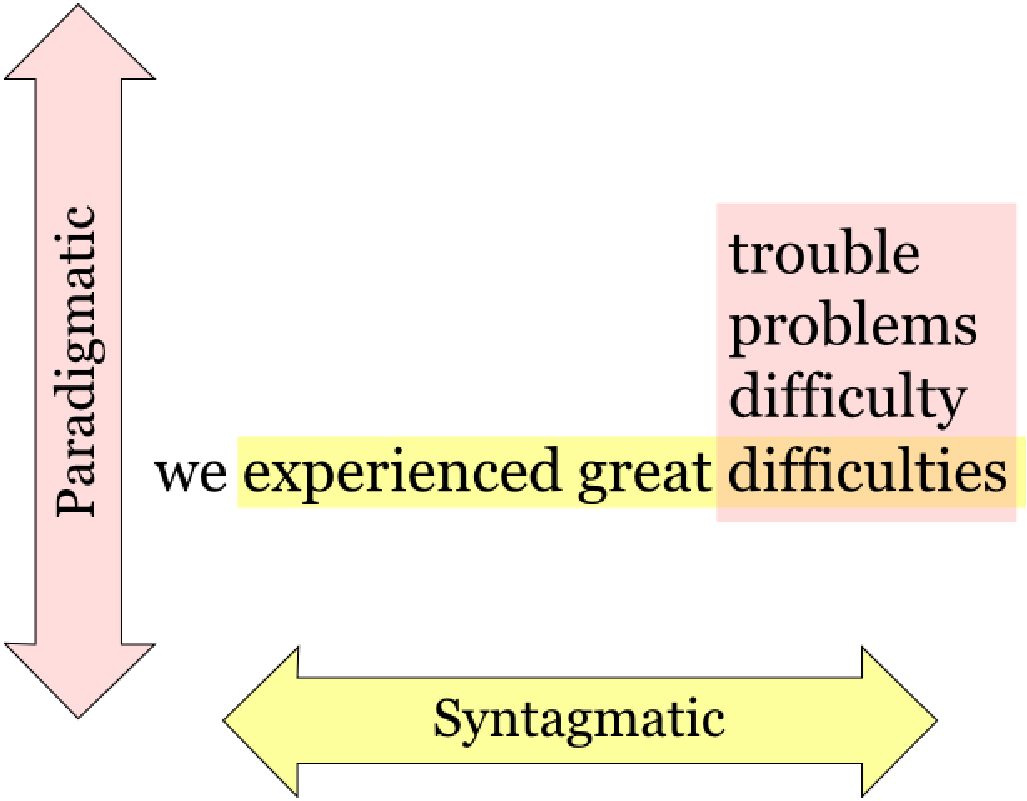
- This is not an NLP application…



- This is a *free* tool for linguistic research.

Background: syntagmatic meaning vs. paradigmatic meaning

"[...] the tradition of linguistic theory has been massively biased in favour of the *paradigmatic* rather than the *syntagmatic* dimension." (Sinclair 2004:140)



Firthian tradition: syntagmatic meaning

- Sinclair: model of a **lexical item**:

  "a model which reconciles the paradigmatic and syntagmatic dimensions of choice at each choice point." (Sinclair 2004:141)

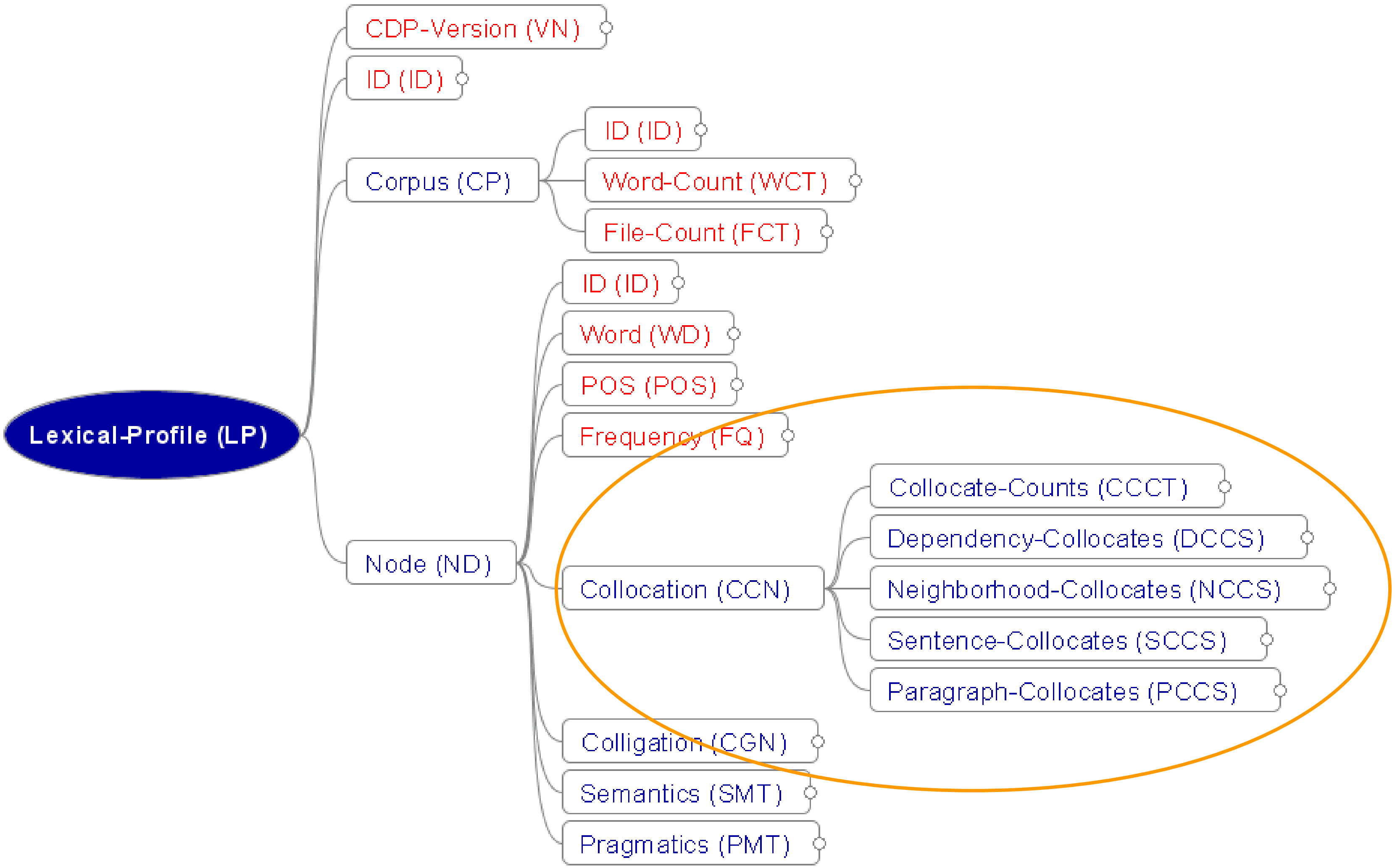
- This model contains several components:

  "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)

The Corpus-Derived Profiles framework

- Allows the creation of **lexical profiles** (CDPs) of words in corpora.
  - A given word will have 1 profile for a given corpus.
  - Each profile contains information on:
    - The corpus and the **node**
    - The node's collocational relations
    - The node's colligational relations
    - The node's semantic relations (semantic preference)
    - The node's pragmatic relations (semantic prosody)
- Defines a **query language** for executing queries of
  - A single profile
  - A combination of profiles: **either**
    - ***different words in the same corpus*, or**
    - ***the same word in different corpora***

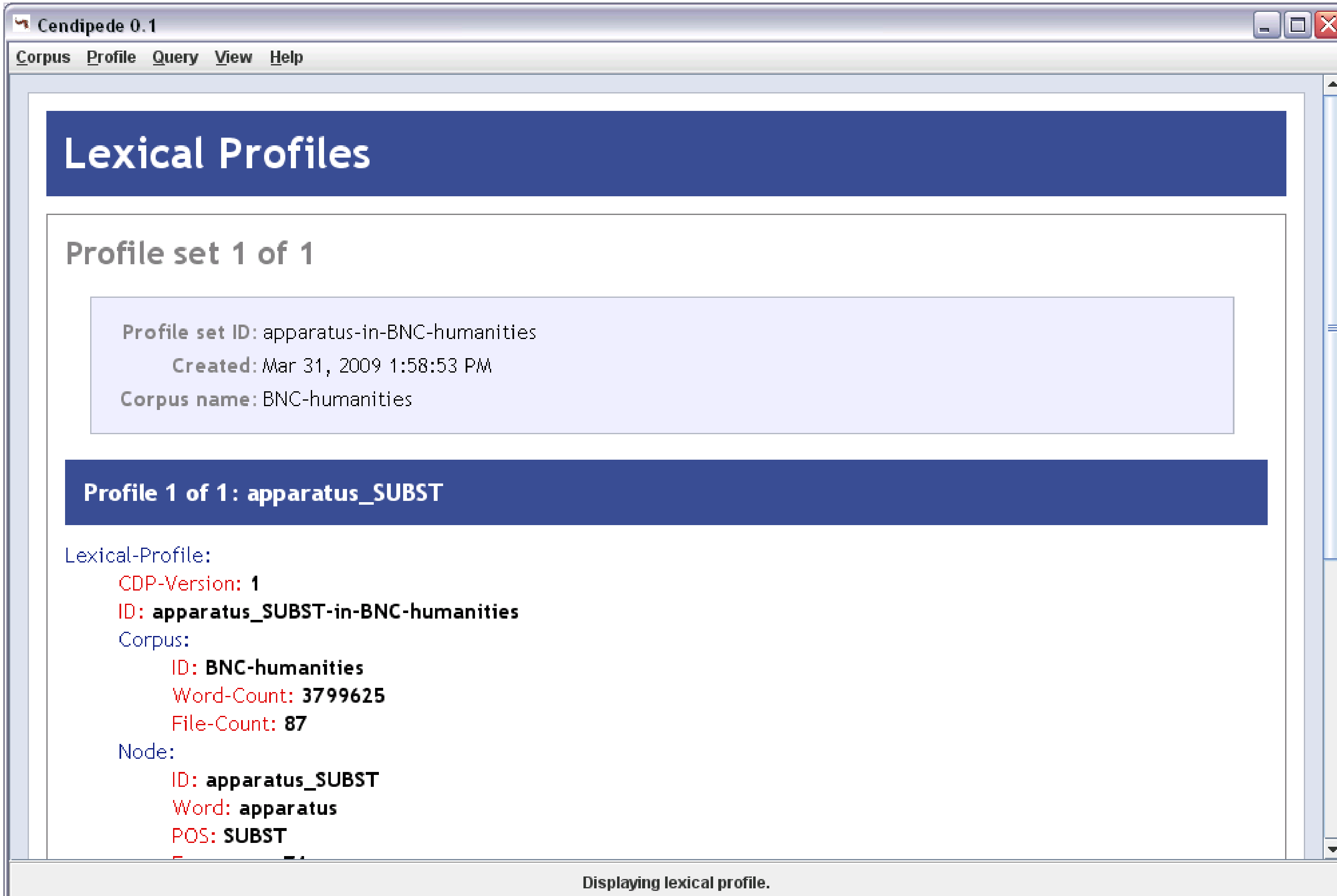
The structure of a lexical profile



CenDiPede: an implementation of the framework

- The framework is a **conceptual model** only. But it is designed to be **implemented computationally**.
- I have created an implementation of the framework called **CenDiPede**:
  - ✓ Stupid name
  - ✓ Creates CDPs (lexical profiles) for any parseable corpus
    - Includes Pro3Gres dependency parser (Schneider 2007)
  - ✓ Written in Java; cross-platform
  - ✓ Easy to use; pleasing GUI
  - ✓ Interactive view of each profile
  - ✓ KWIC and Paragraph views of the corpus tokens
  - ✓ Fully implements CDP Query Language
  - ✓ Outputs data in different formats (E.g. to a spreadsheet)
  - ✓ Operates at a relaxing pace (but can create batches of profiles)

Viewing a profile in CenDiPede



KWIC view of tokens in Cendipede

KWIC View

Sort first by: choose | Then by: choose | Then by: choose | Then by: choose | Sort | Show PARA View | Send To Browser

Row	Collocate	Pre-context	Node	Post-context	File
1	state	loyalist statelet , with its appropriate state	apparatus	of coercion and alternative paramilitary units . For	A07.xml
2	state	the founding of the state . The Irish state	apparatus	is certainly secular : priests have no role in it .	A07.xml
3	state	opposed any attempt to incorporate the church into the	apparatus	of the state and in this was , as already	A07.xml
4	state	endemic in the growth of the modern state	apparatus	, including its welfare institutions . It is also important	A07.xml
5	state	quoted Whyte 1980 : 29-30) ' □ The Irish state	apparatus	assisted public and family morality by the fairly heavy	A07.xml
6	state	shelter of an increasingly powerful state	apparatus	, at the cost of both social cohesion and personal	A66.xml
7	state	procedures , of the interests of the state	apparatus	and of national policies . Yet in relation to our local	BMP.xml
8	state	state , and how far it is part of the state	apparatus	. Nor would this position necessarily be fixed : a	BN8.xml
9	state	established unions working against the state	apparatus	could not hope to obtain for their members (labour	G1R.xml

The CDP query language

- Query language features a long form:
  1. "find every Dependency-Collocate where Joint-Frequency is greater than 20 and sort descending by Joint-Frequency"
  2. "find all Neighborhood-Collocate where MI-score rank is less than 101 and show LL-Score, T-Score and sort by MI-score"
  3. "find all non-overlap between Paragraph-Collocate and Neighborhood-Collocate where POS is SUBST and display max 50"
- And a short form:
  1. "find all DCC where JFQ > 20 and sort by JFQ"
  2. "find all NCC where MI < 101 and show LL, TS and sort by MI"
  3. "find all non-overlap between PCC and NCC where POS = SUBST and display max 50"

Example of a profile query

Choose Profile Terms

Type a query using the CDP query language:

```
find all non-overlap between PCC, NCC
pos = SUBST
show JFQ, CSC, MI
sort by CSC
```

OK

Query Results

Profile Query Results

Overview

Your query:

```
find all non-overlap between PCC, NCC
pos = SUBST
show JFQ, CSC, MI
sort by CSC
```

Canonicalized query:

```
FIND ALL NON-OVERLAP BETWEEN PCC, NCC WHERE POS = SUBST
AND DISPLAY JFQ, CSC, MI
AND SORT DESCENDING BY CSC
```

No. of profiles: 1

Total hits: 69

Profile 1 of 1

Profile ID: apparatus_SUBST-in-BNC-broadsheet

Node: apparatus_SUBST

Corpus: BNC-broadsheet

Hits: 69

PCC:ID: **struggle_SUBST**	CSC: **90**	JFQ: **6**	MI: **12.2137**	POS: **SUBST**
PCC:ID: **Doctrine_SUBST**	CSC: **89**	JFQ: **4**	MI: **15.9049**	POS: **SUBST**
PCC:ID: **massacre_SUBST**	CSC: **89**	JFQ: **5**	MI: **13.3199**	POS: **SUBST**
PCC:ID: **Adamec_SUBST**	CSC: **86**	JFQ: **4**	MI: **13.8174**	POS: **SUBST**
PCC:ID: **schools_SUBST**	CSC: **86**	JFQ: **6**	MI: **11.0181**	POS: **SUBST**
PCC:ID: **army_SUBST**	CSC: **84**	JFQ: **6**	MI: **10.8496**	POS: **SUBST**

Part 2: Collocation

Two approaches to the study of collocation

- Most agree: **collocation = repeated co-occurrence of words**
- But that's where the agreement ends.
- Two broad traditions in collocation research:

Nesselhauf's (2004) terms:

Frequency-based approach	Phraseological approach

My suggested terms:

Syntagmatic approach	**Paradigmatic approach**
Firth	Benson
Halliday	Cowie
Hoey	Hausmann
Renouf	Howarth
Sinclair	Mel'čuk
Etc.	Etc.

Aspects of a theory of collocation

- I have identified eight ways in which definitions of collocation tend to vary. I call these **aspects** of collocation:

Ontology
Extent
Symmetry
Frequency
Scope
Relation
Variation
Opacity

Ontology

- Question: **What kind of phenomenon is collocation?**
- One view: Collocation is a *textual* relation between words in a corpus.
  - E.g. Sinclair 1991; Partington 1998
- Another view: Collocation is a *psychological* relation between words in the mental lexicon.
  - E.g. Leech 1974; Hoey 2005
- Yet another view: Collocation is an *abstract* semantic relation between words that may be instantiated with different lexical items.
  - E.g. Cowie 1998; Nesselhauf 2005

(Note: red dot means "seems to me to be the most popular view")

Extent

- Question: **How many words may a collocation consist of?**
- One view: A collocation includes a single node and a single collocate.
  - E.g. Jones & Sinclair 1974
- Another view: A collocation may include multiple words.
  - E.g. Firth 1957; Renouf & Sinclair 1991; Bartsch 2004

Symmetry

- Question: **Is collocation a symmetrical or asymmetrical relation?**
- One view: “A is a collocate of B” = “B is a collocate of A”
  - E.g. Stubbs 2001
- Another view: There is a difference between collocation with a more frequent word and collocation with a less frequent word.
  - E.g. Sinclair 1991
- Yet another view: There is a difference in the semantic contribution of node and the collocate.
  - E.g. Hausmann1989; Mel’čuk 1998

Frequency

- Question: **How often must words co-occur to be a collocation?**
- One view: No requirement of repeated co-occurrence.
  - E.g. Sinclair 1991 (in mention of “casual collocations”)
- Another view: They must co-occur 2 or more times in the corpus.
  - E.g. Kjellmer 1987; Kennedy 1990; Altenberg 1998
- Yet another view: The words must co-occur often enough to pass a statistical test.
  - E.g. Church & Hanks 1990; Kilgarriff 1992; Biber 1993; Stubbs 1995

Scope

- Question: **What counts as co-occurrence?**
- Two sub-aspects:
  - **Distance:** How near each other must the words be?
  - **Structure:** Must they co-occur within some structure?
- One view: The collocate must be between L5 and R5 (i.e., with max 4 words intervening) or similar.
  - E.g. Stubbs 1996; Bartsch 2004
- Another view: The collocate must be within the same phrase/ sentence/paragraph/text.
  - E.g. Hoey 2005 (in discussion of textual collocates)
- Combined requirement also possible: e.g. collocate must be within 5 words *and* in the same sentence as node.

Relation

- Question: **Must the words be in a grammatical relation?**
- One view: No direct relation is necessary.
  - E.g. Sinclair 2004
- Another view: The words must stand in a direct grammatical/ syntactic relation.
  - E.g. Kjellmer 1994; Bartsch 2004; Nesselhauf 2005

Variation

- Question: **How much may the forms vary?**
- Two sub-aspects:
  - **Inflectional variation:** May word inflections vary?
  - **Lexical variation:** May other words be substituted?
- One view: Each form (*find, finds, found*) has its own collocations.
  - E.g. Sinclair 1991; Bartsch 2004; Hoey 2005
- Another view: Each lemma (*FIND*) has its own collocations.
  - E.g. Halliday 1966; Benson et al. 1997
- Yet another view: A collocation is abstract and may be instantiated by different words (*find, discover,* etc.) ***but only to a limited extent***.
  - **Commutability**: (restricted) ability to substitute other words without changing the meaning
  - E.g. Hausmann 1989, Cowie 1998; Nesselhauf 2005

Opacity

- Question: **Must the meaning of the words be opaque?**
- One view: There is no requirement of opacity in collocation.
  - E.g. Moon 1998; Sinclair 2004
- Another view: The meaning of the collocation must not be determinable from the meanings of the component words.
  - E.g. Cowie 1998; Bartsch 2004

Gradient variation between approaches

Mel’čuk 1998

Hausmann 1989

Cowie 1998

Benson et al. 1997

Nesselhauf 2005

Bartsch 2004

Moon 1998

Hoey 2005

Kjellmer 1994

Stubbs 1996

Biber 1993

Sinclair 2004

Renouf & Sinclair 1991

Note: This figure is meant to be suggestive, not strictly accurate.

Collocation in the CDP framework (1)

- Collocation is seen in the CDP framework as a relation between words in text (not a psychological relation).
  - This has several advantages:
    - Does not assume an understanding of mental processes
    - Statements about data may be with confidence and precision
    - Allows for precise comparison of genres, text types, etc
    - Ameliorates problem of the “statistical individual”
- Collocation may involve multple words, but is generally discussed as a relation between 1 node and 1 collocate.
- Collocation is seen as inherently symmetrical, though information on asymmetry is recorded in a profile.

Counts of what?

"Counts on the page are not the same as counts in the head."

– N. C. Ellis, May 28, 2009, 9:15 AM

Collocation in the CDP framework (2)

- Collocates must pass a statistical test.
  - Composite score based on ***t*-score**, **MI score**, and **log-likelihood score**
- Scope may vary; the framework defines four types of collocates:
  - **Paragraph collocates** (in same paragraph)
  - **Sentence collocates** (in same sentence)
  - **Neighborhood collocates** (within 5 words)
  - **Dependency collocates** (in direct grammatical relation)
- Only dependency collocates must exhibit a grammatical relation.
- The default assumption is that every form has its own collocations.
  - But forms may be combined, to profile a lemma.
  - Even lemmas may be combined, to profile an abstract concept.
- There is no requirement of opacity.

Advantages of this definition of collocation

- Can be implemented computationally (i.e. automatized)
- Does not rely too heavily on any one statistical test
- Provides different types of information via different types of collocates
- Allows for great leeway in the types of things investigated:
  - forms, lemmas, phrases, etc.
- Provides a good basis for studying colligation, semantic preference, and semantic prosody
- Framework constitutes a well-defined standard that makes studies:
  - *easy to describe*
  - *easy to replicate*
- (Hopefully) provides a useful tool for linguistic research

Thank you!

References (1)

Altenberg, B. 1998. On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. In Cowie, A. P. (ed.) *Phraseology: Theory Analysis and Applications*. Clarendon Press, 101-122.

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

Benson, M, E. Benson & R. Ilson (eds.). 1997. *The BBI dictionary of English word combinations*. Amsterdam: John Benjamins.

Biber, D. 1993. Co-Occurrence Patterns among Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition. *Computational Linguistics* 19, 3, 531-538.

Church, K. W. and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16, 1, 22-29.

Cowie, A. (ed). 1998. *Phraseology: theory, analysis, and applications.* New York : Oxford University Press.

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.

References (2)

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.

Halliday, M. 1966. Lexis as a linguistic level. In C. E. Bazell et al. (Eds.), *In Memory of J. R. Firth*. 148-162.

Hausmann, 1989. Le dictionnaire de collocations. In F. J. Hausmann, H. E. Wiegand, & L. Zgusta (eds.), *Wörterbücher, Dictionaries, Dictionnaires.* Ein internationales Handbuch zur Lexikographie. Berlin: de Gruyter.

Kennedy, G. 1990. Collocations: Where grammar and vocabulary teaching meet. In S. Anivan (Ed.), *Language Teaching Methodology for the Nineties.* Singapore: SEAMEO Regional Language Centre.

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

References (3)

Jones, S. and J. M. Sinclair. 1974. English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie* 24, 1, 15-61.

Kilgarriff, A. 1992. *Polysemy*. Unpublished PhD thesis, University of Sussex.

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.

Kjellmer, G. 1994. *A Dictionary of English Collocations: Based on the Brown Corpus.* Oxford: Clarendon Press.

Leech, G. 1974. *Semantics*. London: Penguin.

Mel'čuk, I. 1998. Collocations and Lexical Functions. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications*. Oxford University Press.

Moon, R. 1998. *Fixed expressions and idioms in English: a corpus-based approach.* Oxford University Press.

Nesselhauf, N. 2004. What are collocations? In D. Allerton, N. Nesselhauf, and P. Skandera (Eds.), *Phraseological Units: Basic Concepts and Their Applications*. Basel: Schwabe.

References (4)

Nesselhauf, N. 2005. *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Renouf, A. and J. M. Sinclair. 1991. Collocational frameworks in English. In K. Aijmer and B. Altenberg (Eds.), English Corpus Linguistics. Studies in Honor of Jan Svartvik. London: Longman. 128-143.

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.

Stubbs, M. 1995. "Corpus evidence for norms of lexical collocation". *Principle and practice in applied linguistics*. Oxford University Press.

Stubbs, M. 1996. *Text and corpus analysis: computer-assisted studies of language and culture.* London: Blackwell.

Stubbs, M. 2001. *Words and phrases: corpus studies of lexical semantics.* London: Blackwell.