Corpus-based ways to introduce syntax

Stefanie Dipper

Linguistics Department
Ruhr-University Bochum

6.1.2011
Workshop HU Berlin
“Corpora in Teaching Languages and Linguistics”

Outline

1. Traditional Introductory Courses in Formal Syntax

2. BNC Queries: Distribution-based Criteria

3. BNC Queries: Constituency

4. Conclusions

Traditional syntax introductions

- (Parts of Speech)
- Relevance of word order (“precedence relation”)
- Relevance of hierarchical structure
  (structural ambiguities, “dominance relation”)
- Constituent tests
- Further topics (grammatical functions, X’-theory, “movement” phenomena, . . .)

Parts of Speech (POS)

- “Naïve” approach: semantic-based
  - e.g. nouns are words that *name* concrete objects (Latin *nomen* ‘name’)
  - works fine for “prototypical” nouns such as *child, door* but not for “problematic” nouns such as *movement, softness*

Parts of Speech (cont'd)

Hence:

- Distribution-based criteria (language-specific)
  e.g.: (English) nouns can be preceded by articles and/or adjectives
- Form-based criteria (language-specific)
  e.g.: The ending '-s' can be added to noun stems to form the plural (in English)

No one-to-one correspondance between the criteria and POS

- Certain criteria apply to different POS
- Certain words do not fulfill all criteria (e.g. irregular plural form)

Criteria are indicators of POS rather than defining criteria

Where do the criteria come from?

- Listen to the teacher
- Look them up in a grammar
- Come up with criteria by yourself

Word order and constituency

- Observation: linear order matters
  A young child saw a huge dog
  $\neq$*A huge dog saw a young child*
- "Grouping" matters
  "Can I try that dress on in the window?" — "Certainly Madam, unless you'd prefer to use the changing room."
- Recursive grouping:
  - hierarchical (constituent) structure
  - 2nd dimension

“Grouping” criteria

- “Naïve” approach: according to semantics
  E.g. words that refer to the same object form a constituent
  A young child saw a huge dog
- Classical approach: according to constituent tests
  - “Movement”: topicalization, wh-question, cleft/pseudo-cleft
  - Substitution: pronominalization
  - Interposition: adverb insertion
  - Coordination
  - . . .

Example constituent test: topicalization

Criterion: The group of words that can be topicalized (= moved to the front) is a constituent

- Ex:
  I like green beans
  Assumption: *green beans* form a constituent
  Test: *Green beans I like*
- Hence, *green beans* is a constituent (in these two sentences)
  I like [green beans] (and *[Green beans] I like*)

Problems of the traditional introduction

- Application of constituent tests: often problematic
  - e.g. topicalization of subjects
  - additional modifications (e.g. auxiliaries)
  - unclear intuitions
- Criteria are usually introduced and illustrated by made-up examples
  - → no/few connection to every-day language
- Students learn how to apply criteria/rules
  Students should (also) learn:
  - to verify such criteria
  - to develop new criteria
- Alternative approach: use of corpus data
  - connection to every-day language
  - corpus-based methods to come up with generalizations/criteria

General procedure

1.
    1. “Semantic start”: Pick some words that are prototypical representatives of a certain part of speech.
       E.g. words that denote concrete objects = prototypical nouns
       (Alternatively: “Structuralist approach”: Pick one or several arbitrary words)
    2. Create test sentences for each of these words. (Make them up or search for the words in a corpus.)
    3. By investigating your test sentences, come up with one or two hypothetical criteria.
    4. Run corpus searches, to confirm or refute the hypotheses.
2. Pick further (prototypical) words, and run the same corpus searches with them.
3. Cross-check: Once a rule has been shown to be valid for all or most of the prototypical words, run corpus searches for atypical words.

Ex: develop distribution-based criteria for nouns

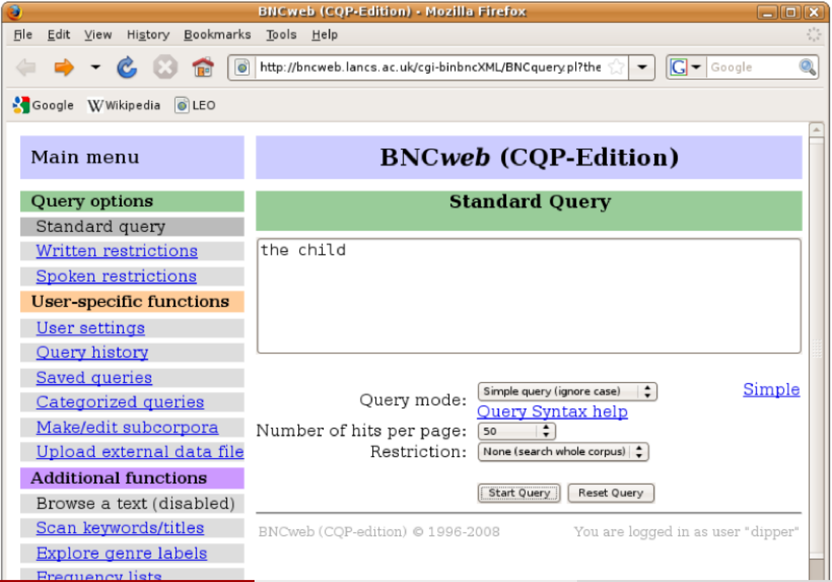
- Nouns: names of persons or things
- “Prototypical” nouns: *child, door, table, . . .*

Examples

1. The child over there laughed.
2. I don’t know this child.
3. He shut the door.
4. It was a heavy door.

- Hypothetical criterion: *Nouns often follow the word ‘the’*
- Run corpus searches: investigate the left context of *child, door*

Corpus BNC, accessed via the interface BNCweb



BNCweb (CQP-Edition) - Mozilla Firefox

http://bncweb.lancs.ac.uk/cgi-binbncXML/BNCquery.pl?the

Main menu

BNCweb (CQP-Edition)

Query options

Standard query

Written restrictions

Spoken restrictions

User-specific functions

User settings

Query history

Saved queries

Categorized queries

Make/edit subcorpora

Upload external data file

Additional functions

Browse a text (disabled)

Scan keywords/titles

Explore genre labels

Frequency lists

Standard Query

```
the child
```

Query mode: Simple query (ignore case)

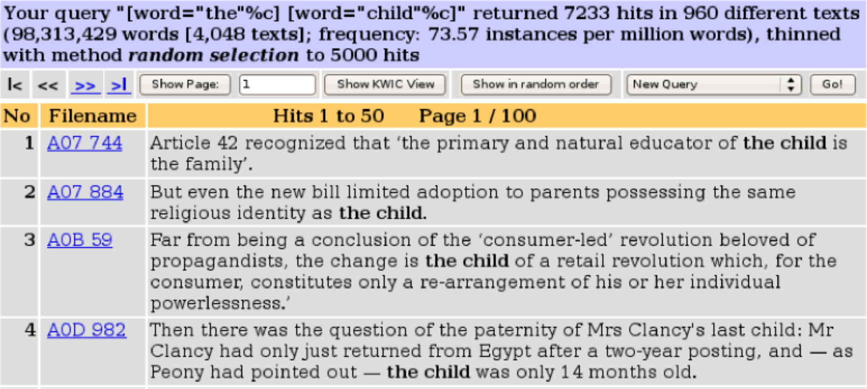
Number of hits per page: 50

Restriction: None (search whole corpus)

BNCweb (CQP-edition) © 1996-2008

You are logged in as user "dipper"

BNCweb: result window



Your query "[word="the"%c] [word="child"%c]" returned 7233 hits in 960 different texts (98,313,429 words [4,048 texts]; frequency: 73.57 instances per million words), thinned with method ***random selection*** to 5000 hits

|< << >> >| Show Page: 1 Show KWIC View Show in random order New Query Go!

No	Filename	Hits 1 to 50 Page 1 / 100
1	A07 744	Article 42 recognized that 'the primary and natural educator of **the child** is the family'.
2	A07 884	But even the new bill limited adoption to parents possessing the same religious identity as **the child**.
3	A0B 59	Far from being a conclusion of the 'consumer-led' revolution beloved of propagandists, the change is **the child** of a retail revolution which, for the consumer, constitutes only a re-arrangement of his or her individual powerlessness.'
4	A0D 982	Then there was the question of the paternity of Mrs Clancy's last child: Mr Clancy had only just returned from Egypt after a two-year posting, and — as Peony had pointed out — **the child** was only 14 months old.

Corpus searches

- Query expression: `+ child`
- Result sorted according to absolute frequencies

No.	Lexical items	No. of occurrences	Percent
1	the child	1510	30.2%
2	a child	1066	21.32%
3	and Child	152	3.04%
1	the door	2552	51.04%
2	front door	376	7.52%
3	next door	293	5.86%

→ Up to now: hypothesis confirmed

Next steps

- Test further prototypical nouns, e.g. *table, chair, man*
  . . . let us assume: done . . .
- Cross-check: test atypical words
  - e.g. words that describe events, e.g. *laughed, broke, moved* (past tense occurs more frequently in the BNC — and it is less ambiguous . . .)

No.	Lexical items	No. of occurrences	Percent
1	he laughed	732	16.5%
2	She laughed	565	12.73%
3	and laughed	344	7.75%
1	he broke	387	7.74%
2	She broke	252	5.04%
3	and broke	242	4.84%

→ Criterion is, again, confirmed, and seems useful

From POS to constituents

- POS criteria: word-based investigations
- Constituents: are based on POS rather than words
- BNC: provides POS annotations
  - automatic annotations, based on criteria such as the ones that we have developed
  - hence, we can expect or even predict erroneous annotations!
  - e.g. *He's England's answer to Tom Selleck and I think he should be the next James Bond, **man**'*
    → *man* tagged as V-N

BNC classes and their frequencies

No.	Part of speech	BNC Tag	No. of occurrences	Percent
1	noun	SUBST	25,491,812	22.74%
2	verb	VERB	17,861,343	15.93%
3	punctuation	STOP	13,606,160	12.14%
4	preposition	PREP	12,842,940	11.46%
5	adjective	ADJ	11,818,917	10.54%
6	article	ART	8,690,652	7.75%
7	pronoun	PRON	7,906,511	7.05%
8	adverb	ADV	6,505,396	5.80%
9	conjunction	CONJ	5,656,592	5.05%
10	other	UNC	1,343,981	1.20%
11	interjection	INTERJ	378,021	0.03%

Plus: finer-grained POS tags: NN1, NN2, NN0, NP0 for SUBST, etc.

Chance co-occurrence

- A problem for our account:
  Frequently-occurring POS = frequent neighbors
- Solution:
  Compare observed vs. expected frequencies of POS co-occurrences
  → If the observed (actual) frequency is considerably higher than the expected frequency, the POS neighbors are characteristic neighbors
- I.e. use collocation measures rather than raw frequency counts to come up with criteria

Collocation measures

General idea: we compare 4 frequencies

1. A and B co-occur (adjacent to each other)
2. A occurs but not B
3. B occurs but not A
4. Neither A nor B occurs (within the sentence)

Representation by a contingency table:

	AT0	not-AT0	Sum
NN1	4.5 mio	10 mio	14.5 mio
not-NN1	4 mio	93.5 mio	97.5 mio
Sum	8.5 mio	103.5 mio	112 mio

NN1: common noun, singular
AT0: article

Collocation measures

	AT0	not-AT0	Sum
NN1	4.5 mio	10 mio	14.5 mio
not-NN1	4 mio	93.5 mio	97.5 mio
Sum	8.5 mio	103.5 mio	112 mio

- Observed frequencies: $p(A, B)$ (A-B occurring together) $p(AT0, NN1) = 4.5/112 = .04$
- Expected frequencies: $p(A) * p(B) =$ $p(AT0) * p(NN1) = 8.5/112 * 14.5/112 = .13 * .08 = .01$
- Mutual Information (MI — one version):

$$I(A, B) = log_2(\frac{p(A, B)}{p(A) * p(B)})$$

$$= log_2(.04/.01) = 1.4$$

Collocations

Note:

- Collocation scores cannot be compared to each other, in general (the score depends on the absolute number of matches)
- But we can compare scores of all left neighbors of the same POS, or left neighbors with right neighbors of the same POS

Collocations for SUBST: Comparison of all left neighbors

No.	POS tags	Exp. freq.	Obs. freq.	Log-likelihood
1	AT0 SUBST	374.157	1135	1135.5794
2	AJ0 SUBST	274.882	788	693.7394
3	DPS SUBST	60.271	209	227.0251
...				
44	PUN SUBST	477.225	133	-375.1645
45	PNP SUBST	213.940	5	-389.6526

- BNCweb: Log-likelihood = default collocation measure
- AJ0: adjectives, DPS: possessive pronouns, PUN: general punctuation mark, PNP: personal pronoun

→ Scores confirm our noun criterion

Collocations for SUBST, ART: Comparison of left with right neighbors

No.	POS tags	Exp.freq.	Obs.freq.	Log-likelihood
1	AT0 **SUBST**	374.157	1135	1135.5794
2	AJ0 **SUBST**	274.882	788	693.7394
1	**SUBST** PUN	488.498	1193	839.3770
2	**SUBST** PRF	134.008	490	585.6784
1	PRP **ART/DET**	321.883	1445	2421.4327
2	PRF **ART/DET**	123.836	478	611.9658
1	**ART/DET** NN1	642.810	2246	3093.7606
2	**ART/DET** AJ0	284.565	863	832.2167

→ Scores can be used as evidence for constituent boundaries

Goals and challenges

- Basic concepts of syntax
- Scientific argumentation

- Lexical ambiguities: *table*

Goals and challenges

Application of the POS criteria: manually tagging texts

- Procedure: for each word:
  (i) determine its semantics and protopyical POS
  (ii) check distribution- and form-based criteria
  (and pick another POS, if necessary)
- “I know already what nouns are”
  - *(make a good) impression*: prototypical noun?
- Uni-directional criteria
  - “Nouns often occur after adjectives”
- Currently available information
  - “Prepositions occur in front of nouns”

Summary

- Fundamental syntactic concepts: based on corpus evidence
- Parts of speech: boot-strapping approach:
  Come up with criteria for prototypical words, and successively add more words and more criteria
- Constituency: collocation strength used as an indicator of constituent boundaries
- Similar to Structuralist approach
  - corpus evidence rather than introspective tests
- But:
  - starts with prototypical words (semantically defined)
  - makes use of collocation measures