Manual annotation

Sebastian Hoffmann
University of Trier
hoffmann@uni-trier.de
Outline

❖ Why manually annotate?
❖ Brief review of existing literature on manual annotation
❖ Survey of state-of-the-art programs available for manual corpus annotation
❖ Demonstration of some tools
❖ Limitations of existing software solutions
❖ Wish-list for future corpus tools
Why manually annotate?

- Manual inspection of all texts in a corpus
- Coding of linguistic features as they are encountered
  - “bottom-up approach”
  - qualitative perspective is highlighted

Today: “top-down approach”

- focused search in corpus, followed by manual annotation of (potentially) relevant hits
Why manually annotate?

- Important concepts: Precision & Recall

RECALL measures the proportion of relevant information retrieved in response to a search procedure (the number of relevant items actually obtained divided by the total number which would have been obtained in a perfect search).

PRECISION measures the proportion of retrieved items that are in fact relevant (the number of relevant items obtained divided by the total number of retrieved items).

- optimising recall typically reduces precision
- manual work is required to clean up data
Why manually annotate?

❖ How do you find all tag questions in a corpus?

⇒ auxiliary (form of be, do, have or a modal), followed by an optional negative element and a personal pronoun, there or one

⇒ isn’t it? is there? are they? etc.

⇒ reasonable assumption: tag questions occur in utterance-final position

⇒ really???
Why manually annotate?

How do you find all tag questions in a corpus?

And right on the almost on the final whistle just before United scored in injury time, I think mid-fielder Martin Cool got in a very good volley *didn’t he* from some distance, but it really was whistling toward goal? (BNC:KS7:744)

reasonable assumption: tag questions occur in utterance-final position

really???
Why manually annotate?

- How do you find all tag questions in a corpus?

  But if immediately adjacent question mark is omitted:

  I mean *are you* talking about a hundred and fifty?  
  (BNC:F7J:360)

  The first thing he *did he* made friends amongst the young men in the college.  
  (BNC:HE3:91)

  ➞ recall is improved, precision is reduced
Why manually annotate?

❖ How do you find all tag questions in a corpus?

Automatic methods to increase precision:

• exclude if immediately preceded by $\sim wh$-word
  (e.g. Why don’t you do this?)
• exclude if verb immediately follows pronoun (e.g. Doesn’t he like cheese?)

⇒ Can this have an impact on recall?
Why manually annotate?

- Manual annotation isn’t only about removing false positives!

  Pre-existing categories of annotation in a corpus may not be adequate to answer research question

  ➞ e.g. polarity of tag questions

  ➞ e.g. pragmatic function of tag questions (cf. Tottie and Hoffmann, 2006: informational, confirmatory, facilitating, attitudinal, peremptory, and aggressive tag questions)
Reliability of manual annotation?

- Inter-rater reliability/agreement (the degree of agreement among different raters)
- Intra-rater reliability/agreement (the degree to which the same linguist is consistent in his or her own analysis)

➤ documentation of annotation process can increase reliability
➤ improves replicability of coding
Existing literature on annotation


  ➞ extensive coverage of automated annotation and manual annotation of complete texts/corpora

  ➞ manual annotation of concordance lines does not feature
Existing literature on annotation


  ➞ “problem-oriented tagging”
  ➞ example: postmodifying clauses in NPs
  ➞ useful as basis for statistical testing
  ➞ *ad hoc*, but annotation can possibly be used by “other investigators in different research projects” (p. 123)
Existing literature on annotation

  - suggests importing concordance lines into a database application
  - filtering/sorting of data in very flexible ways
Existing literature on annotation


  ➟ computer tool interacts with linguist and suggests most likely option

  ➟ automated and manual annotation may go hand in hand
Introductions to corpus linguistics


⇒ manual annotation not mentioned


⇒ manual annotation mentioned as “very important”, though need arises only “occasionally” (p.69)

⇒ no explicit guidance, as the process is considered too specific to research question
Introductions to corpus linguistics


  - discuss that automatically retrieved data may require hand-editing (pp. 71, 73)

  - mention interactive tool that aids annotation (no further specified)
Introductions to corpus linguistics


  ➞ pp. 97-8, 111 ("problem-oriented tagging")

  ➞ illustrated with study of pseudo-titles (e.g. *fugitive financier* Robert Vesco, *linguist* Noam Chomsky).

  ➞ calls for mnemonic codes
Introductions to corpus linguistics


- the only textbook that explicitly guides readers through practical steps of manual annotation

- annotation of concordance lines is referred to as “a dirty way”
On the whole...

- coverage of manual annotation is very patchy
- advantages have been clearly documented
- but treatment in introductory textbooks is not extensive
- corpus-novices may fail to be sensitized to the fundamental value of manual annotation for many types of linguistic analysis
Current corpus tools

- AntConc: no manual annotation facilities

- *WordSmith Tools* Version 5: Windows-only package; an advanced set of tools providing “an integrated suite of programs for looking at how words behave in texts” (Mike Scott)

- BNCweb: “Categorize hits” feature

- Combination of corpus tools and databases/spreadsheet applications
Current corpus tools

Demo WordSmith
WordSmith Tools

- Single letter codes can be assigned to concordance lines
- More mnemonic codes can be entered by double-clicking the “Set” field
- Multiple levels of annotation are possible (but a bit awkward)
- Annotation can be saved and re-used
- User-defined categories can also be inserted as tags into the source text. Combination of pre-existing and user-defined searches is made possible (not demonstrated in lecture).
Current corpus tools

Demo BNCweb
“In-corpus-tool” annotation

- Effective method to clean up query result
- Basic classifications of the data
- Key advantage: access to larger context of concordance lines is not lost
- User also has access to advanced functions of corpus tool (e.g. collocations, distribution analyses, etc.)
- Only one tool needs to be learnt
“In-corpus-tool” annotation

- Main limitation: lack of flexibility
  - search/filtering facilities of database tools
  - only one category can be looked at
  - No way of “annotating the annotation process”

→ exporting of data from corpus tool / importing into database/spreadsheet tool
A presenter, after all, knows that a viewer has the visual evidence to check on what is being said. (BNC: Ao4:447)

- in fiction – quoted speech or not?
- animacy of the subject?
- information status of subject – given or new?
- pragmatic aspects of usage
Current corpus tools

Demo BNCweb/Excel
Evaluating the “export-to-database” method

- Effective method to clean up query result
- Increased flexibility
- Access to larger context of query result is limited
- No access to advanced functions of corpus tool (e.g. collocations, distribution analyses, etc.)
- A second tool needs to be learnt

Wouldn’t it be nice if it was possible to combine the two approaches?
Current corpus tools

Demo BNCweb/Excel/BNCweb
References
