The American National Corpus

Everything You Always Wanted To Know . . . And Weren’t Afraid To Ask

Nancy Ide
Department of Computer Science
Vassar College

What is the ANC?

• A BNC-like corpus of American English
  • Comparable in size
  • Comparable in balance
Why bother?

- We don’t talk like the Brits!
  - Different Lexical Items
    Bobby vs. cop, underground vs. subway, lorry vs. truck, pavement vs. sidewalk, football vs. soccer…
  - Different Grammatical structures
    “She could not endure to live with him” vs. “She could not endure living with him.”
    “Have you a pen?” vs. “Do you have a pen?”
  - Different Use of Modals
    “shall” vs. “should” vs. “ought” vs. “will” vs. “would” vs. “should”
  - Different Adverbial Usage
    “Immediately I get home” vs. “As soon as I get home”
  - Different Use of Support Verbs
    “take a decision” vs. “make a decision”

How Does the ANC Differ from the BNC?

- Texts from 1990 onwards only
- A core corpus of 100 million words
  - comparable in balance to the BNC
- PLUS: a “varied” component
  - as many texts of whatever kind we can get!
  - Not necessarily balanced
    - Dictated by availability
    - Includes email, ephemera, rap lyrics, newsgroups, etc. plus historically important works from various time periods
  - Add 10% every five years
  - Layered organization
    - Dynamic component layered chronologically as added
How Does the ANC Differ from the BNC?

**Format:**
- XML markup conformant to XCES
  - XCES XML schemas
- “stand-off” document architecture
  - annotations in separate documents from main text
  - allows for multiple annotations of the same type, or with conflicting structures
  - Conducive to a distributed development model
- multiple documents vs. one big document

**Annotation:**
- Multiple part-of-speech taggings
  - “Biber” tags
  - CLAWS tags
  - Penn tags
- automatic tagging of named entities, dates, abbreviations, etc.
- sentence tagging
- syntactic bracketing
How Does the ANC Differ from the BNC?

- Produced in two stages:
  - “First release” of 10 million words
    - Rough format, no validation
  - Final release
    - 100 million words plus varied component
    - Full headers, some validation
    - Gold standard portion of 10 million words

Will there be access software?

- ANC project will provide search and access software
- Encoding via XML and stand-off architecture enables exploiting the evolving XML environment for search, access, manipulation of ANC data
  - XML Transformation Language (XSLT)
  - Resource Description Framework (RDF)
What about licensing?

- LDC
  - obtains licenses from text providers
  - issues licenses to users
- no redistribution without publisher’s permission
- “open sub-corpus” portion of the ANC
  - licensed on the model of open-source software

What is the “gold standard” component?

- NSF has funded a two-year project to hand-validate 10 million words of the corpus
  - POS tags
  - sentence boundaries
  - syntactic bracketing
- Also, create a (relatively small) ontology and identify instances in the corpus, using RDF
  - document, author, name, person, date, etc.
  - meta-data categories
How is the ANC funded?

- A consortium of publishers and software vendors
  - Pearson Education
  - Langenscheidt Publishing Group
  - HarperCollins Publishers
  - Cambridge University Press
  - Microsoft Corporation
  - Shogakukan Inc.
  - ACL Press Inc.
  - Taishukan Publishing Company
  - Oxford University Press
  - Kenkyusha Ltd.
  - IBM Corporation
  - Obunsha Publishing Co. Ltd.
  - Bloomsbury Publishing Plc
  - Benesse Corporation
  - Sanseido Co., Ltd.
  - Sony Electronics Inc.
  - Macmillan Publishers

What do they get in return?

- Immediate access to all released data
- Unavailable to other commercial users until fall, 2007
Do you have enough money?

- Operating on a shoestring and through “contributed time” of personnel at Vassar and Northern Arizona University
- Bulk of the funds pay a full-time programmer
  - Vassar:
    - N. Ide, Technical director
    - K. Suderman, Full-time programmer
  - NAU:
    - Randi Reppen, Project Manager
    - Several graduate students
- Linguistic Data Consortium contributing licensing expertise and distribution

How is the work organized?

- Arrange licensing with publishers
- Select and acquire texts
- Create header
- Tokenize, tag and lemmatize
- Transduce texts from original format to “rough” XML
- Split into sentences, add tags
- Render in XCES-compliant XML with stand-off annotation
- Distribute to users
- Provide search and access software
What software are you using?

- **XML Spy**
  - Editing, parsing, validating
  - Generate, test XSLT scripts

- **GATE**
  - All processing (sentence splitting, massaging markup, merging POS tags with files, generate stand-off annotation documents
  - Developed a protocol handler to apply XSL style sheet to file
  - Developed programming language to automate invoking bits of GATE

- **TextLightning**
  - Transduce PDF to XML

What tagger are you using?

- **Tagger developed by Doug Biber**
  - CLAWS-like tags

- **FixTag program (NAU) for validating POS tags**

- **CLAWS tags will be generated here at Lancaster**
Where do you get the texts?

- Contributions from publishers
- Contributions from software vendors
- LDC
- Donations from projects
- Web
- Scanning “ephemera”
- ACL research publications

What texts do you have so far

**WRITTEN:**
- New York Times (1.5 million)
- Oxford University Press texts (275K)
- Cambridge University Press texts
- ACL research papers (900K)
- Berlitz Travel Guides (550K+)
- Slate Magazine (5 million)
- Verbatim Magazine (?)
- Ephemera (350K)
- Government documents from the web (?)
- Kid’s corpus

**SPOKEN:**
- Switchboard (3 million)
- CallHome (50K)
- Charlotte Narrative (200K)
What texts do you expect to get?

- IBM journal
- Bloomsbury texts
- Harper Collins texts
- Pearson texts
- Macmillan texts
- Sony manuals
- ICE (spoken)
- LAWS (spoken)
- More web documents
- Email

...and anything else we can solicit!

How do you decide an author is American?

(Wish you hadn’t asked that…)

- Check spelling, lexical use
- Ask the publishers
- Check affiliation (e.g. for ACL texts)
- Use US government documents

In general, we have to assume that a text intended for an American audience is in American English

- But we can never be entirely sure
  - Hope that the size of the data allows for some noise
  - With so many non-native speakers in the US, hard to know
Why don’t you get texts from the web?

- The “American author” problem
- We have showed that web texts are skewed toward “information-dense” prose
  - Not representative of many (most) genres


When will the first release be available?

- Release scheduled for end of spring
- 10 million+ words
- “base format”
  - Markup for logical structure, sentence, word
  - Annotated for POS, lemma
  - Produced fully automatically
    - Some tags may be “rough” (e.g. cannot always identify titles, etc. reliably)
  - May not have full headers
- Not balanced for genre!
- Very basic retrieval software
  - Concordance by word, lemma, POS only
Who can get the first release?

- Consortium members
- Non-commercial users
  - No restrictions based on geographical location
- Anyone who joins the consortium at any time

Distributed by LDC

What has held up the first release?

- Delays in text acquisition
- Funky formats
  - Quark Express
  - Multi-column PDF
    - Can’t retrieve text in order
  - Even “XML tagged” text is extremely inconsistent!
    - Browsers that don’t care about consistency have led to sloppy markup habits!
    - Some so-called XML is actually HTML with < and > changed to &lt; and &gt; etc.
How does my commercial enterprise join the ANC Consortium?

• Contact me:
  Nancy Ide  ide@cs.vassar.edu

• Cost is $40,000 in two $20,000 installments

Where do I get more information?

ANC Website
http://AmericanNationalCorpus.org

Project Manager
Randi Reppen  randi_reppen@nau.edu

Technical Director
Nancy Ide  ide@cs.vassar.edu
Is this talk over?

YES!

Thank you