

Data Intensive Linguistics: Significant Events

Earliest times

c. 3000 B.C.: Egyptian hieroglyphics

Permanent language.

c. 1000 B.C.: Hebrew Old Testament texts

Why were these documents so important? Maybe their social and religious authority.

c 700-500 B.C.: Panini's phonetics

First known attempt to systematise the rules of a language

Widespread literacy and availability of text

c 700: Woodblock printing in China

c 1450: Widespread use of printing press. Now possible to have canonical versions of texts

1799-1820 : Rosetta Stone

Champollion uses Rosetta stone to decode hieroglyphics

Language interpretation can be seen as a puzzle. Easier if you have a lot of text available.

1897: Käding

11 million words of German. 5000 Prussian analysts. Used to examine spelling conventions.

Telegraphy and Telephony

1832: Morse's electrical telegraph

1875: Bell Invents telephone

Commercial motivation for scientific study of communication.

Information theory

1917: Nyquist "Certain factors influencing telegraph speed"

Theoretical limits on how much can be sent at given power

1928: Nyquist "Certain Topics in Telegraph Transmission Theory"

Tight limits on frequency band needed to transmit information

1928: Hartley "Transmission of Information"

Imagine the sender of a message equipped with a set of balls and rolling a dice.

1949: Shannon and Weaver "Mathematical Theory of Communication"

See later lecture for details.

Developments in Linguistics

1916: Saussure

How to do theoretical linguistics

1930: Bloomfield: scientific study of language

Express yourself in the same neutral and empirically verifiable way as do other sciences

1957 : Firth: "You shall know a word by the company it keeps"

Very important idea, but as pointed out then by Abercrombie, hard to do much about this in real life. No longer true now.

Statistics

1930s: RB Fisher and others

Statistical science reformulated and used to for quality control of industrial and other processes

Background technological developments

1930s: Gödel and Turing on Foundations of Computation

Formalises the notion of computation in a way is useful

1939-1945: Electronics and code breaking at Bletchley Park

Still arguably the most important piece of practical linguistics ever done.

post 1945: first practical computers

Appear to have been a direct result of code breaking efforts

Machine Translation

1949: Warren Weaver's memorandum on MT

Imagine the sentence is really in English, but has been mistakenly encoded into Russian.

1950s: Early systems: linguistically simple

Not really feasible , because of limitations on computational power and data collection. Complexity of language and size of corpus needed systematically underestimated.

Psychology

1930s: Watson's

Empirical approach, replaces Jamesian introspection.

1950s: Skinner's behaviourism

Rejected mentalism in favour of observation

1957: Syntactic Structures

A bomb under the whole empirical enterprise. You can't tell enough about the language from distributional data. Need to get inside the mind of the language user. The sentence "I live in Dayton Ohio" is much less likely than "I live in New York"

Corpus Linguistics under pressure

1960: Quirk's Survey of English Usage

Computerised by Svartvik by 1975.

1960-1980: Brown Corpus

Computational from the outset, at a time when computation was a serious matter.
Partially motivated by psycholinguistic considerations

1966 ALPAC Report

All US funding of (data intensive) work on machine translation suspended. Similar critique of perceptrons by Minsky and Paper, with similar terrible effects.

1980s: Digital text commonplace.

Also becoming possible to observe the processes of reading and writing at much smaller cost than previously.

1972 - date IBM Yorktown work on speech recognition

Training is better than guessing. Laboriously obtained linguistic encodings do not necessarily pay off.

1988: Success of statistical methods in speech recognition

Standard test suites, rigorous evaluation.

1990s : Vast computational power under many desks

Makes the analyses done by Käding almost trivial. Allows others which were previously unimaginable. Typically a theoretician's grasp is dependent on his reach, and the same is true of an experimentalist's imagination.

1992: Fillmore Corpus Linguistics or Computer aided armchair linguistics

corpus: The corpus linguist has all the primary facts he needs in the form of a corpus of approximately one zillion running words, and he sees his job as deriving secondary facts from these primary facts. At the moment he is busy determining the relative frequencies of eleven parts of speech as the first and second words of a sentence.

non-corpus: Sits in a deep soft armchair with his eyes closes and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly, says "wow, what a neat fact" and writes something down in pencil.

summary: every corpus taught me something I couldn't imagine getting any other way.

1993: Computational Linguistics Special Issue on CL Using Very Large Corpora

Defining moment:

[J93-1001](#): **Kenneth W. Church; Robert L. Mercer**

Introduction to the Special Issue on Computational Linguistics Using Large Corpora

CL 19(1):

[J93-1002](#): **Ted Briscoe; John Carroll**

Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars

[J93-1003](#): **Ted Dunning**

Accurate Methods for the Statistics of Surprise and Coincidence

[J93-1004](#): **William A. Gale; Kenneth W. Church**

A Program for Aligning Sentences in Bilingual Corpora

[J93-1005](#): **Donald Hindle; Mats Rooth**

Structural Ambiguity and Lexical Relations

[J93-1006](#): **Martin Kay; Martin Roscheisen**

Text-Translation Alignment

[J93-1007](#): **Frank Smadja**

Retrieving Collocations from Text: Xtract

CL 19(2):

[J93-2001](#): **Douglas Biber**

Using Register-Diversified Corpora for General Language Studies

[J93-2002](#): **Michael R. Brent**

From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax

[J93-2003](#): **Peter E Brown; Vincent J. Della Pietra; Stephen A. Della Pietra; Robert L. Mercer**

The Mathematics of Statistical Machine Translation: Parameter Estimation

[J93-2004](#): **Mitchell P. Marcus; Mary Ann Marcinkiewicz; Beatrice Santorini**

Building a Large Annotated Corpus of English: The Penn Treebank

[J93-2005](#): **James Pustejovsky; Peter Anick; Sabine Bergler**

Lexical Semantic Techniques for Corpus Analysis

[J93-2006](#): **Ralph Weischedel; Richard Schwartz; Jeff Palmucci; Marie Meteer; Lance Ramshaw**

Coping with Ambiguity and Unknown Words through Probabilistic Models

Assignments

Read Church and Mercer's intro <http://www.aclweb.org/anthology/J93-1001>

Read Abney's intro: <http://citeseer.nj.nec.com/abney96statistical.html>

Write a program that generates 10,000 random characters (hand-in Thursday). Python is recommended