Textual Corpora, Treebanks, and the Human Language Project

Steven Abney

Department of Linguistics
Computer Science and Engineering
School of Information

2015 Mar 30

Overview

- What kinds of (annotated) text corpora do computational linguists use?
- A research example: using existing treebanks to bootstrap treebanks in other languages
- Extrapolating into the future: building a comprehensive multilingual dataset – a Human Genome Project for language
- What a library might provide

Textual Data

Plain text

- Language samples
  - Plain text
  - Preferably large amounts (1M–1B words)
  - In as many languages as possible
- Some LDC titles:

UN Parallel Text (Complete) [1994]
European Language Newspaper Text [1995]
Japanese Business News Text [1995]
Spanish News Text [1995]
Mandarin Chinese News Text [1995]
CALLHOME Egyptian Arabic Transcripts [1997]
Portuguese Newswire Text [1999]
Korean Newswire [2000]
HUB5 German Transcripts [2003]
Chinese Gigaword [2003]
Arabic Gigaword [2003]
Czech Broadcast News Transcripts [2004]
Web 1T 5-gram Version 1 [2006]
Hungarian-English Parallel Text, Version 1.0 [2008]
Web 1T 5-gram, 10 European Languages Version 1 [2009]

- Aligned texts, multiple translations

Textual data
Lexicons

- Wordnet: lexical database
  - hypernymy, synonymy, meronymy
  - "Synset" = set of words that share a meaning
- Multi-lingual wordnets – synsets that cross languages (translation equivalents)
- Babelnet – combines Wordnets with Wikipedia and Wiktionary
- Panlex – database constructed from thousands of bilingual print dictionaries

entity
physical entity
object
whole
living thing
organism
animal
chordate
vertebrate
mammal
placental
ungulate
odd-toed ungulate
equine
horse

Textual data

Parts of speech

- the Brown corpus:
- Why?
  - Input for machine learning
  - Automatically train a system to label new text
  - First step in language-interpretation pipeline

The	at
Fulton	np-tl
County	nn-tl
Grand	jj-tl
Jury	nn-tl
said	vbd
Friday	nr
an	at
investigation	nn
of	in
Atlanta's	np$
recent	jj
primary	nn
election	nn
produced	vbd
“	“
no	at
evidence	nn
”	”
that	cs
any	dti

Textual data

Named entities

El portavoz del Consejo Político Municipal de IU en Santander, Ernesto Gómez de la Hera, explicó hoy en conferencia de prensa que este ...

ORG: Consejo Político Municipal de IU

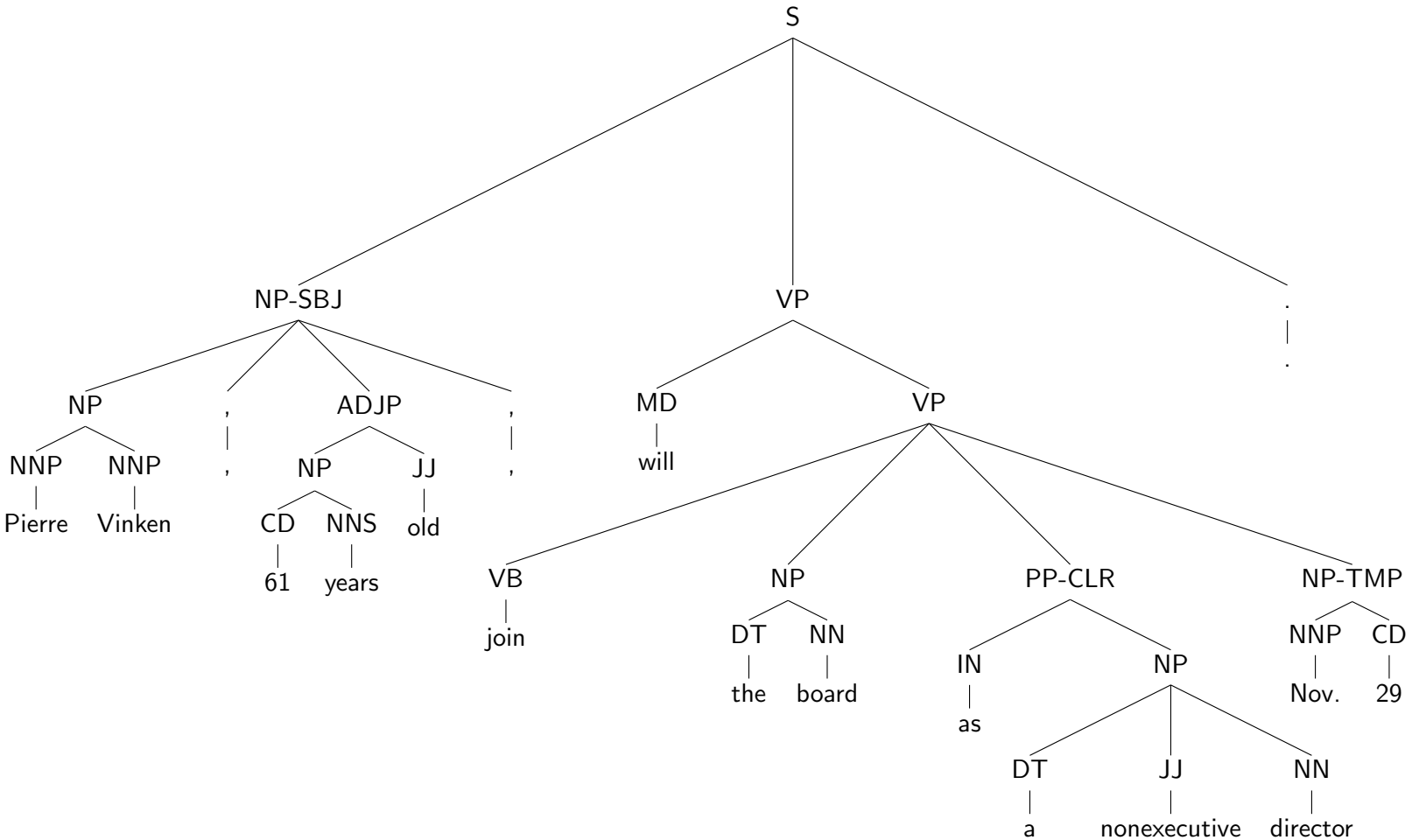
LOC: Santander

PER: Ernesto Gómez de la Hera

El	O
portavoz	O
del	O
Consejo	B-ORG
Político	I-ORG
Municipal	I-ORG
de	I-ORG
IU	I-ORG
en	O
Santander	B-LOC
,	O
Ernesto	B-PER
Gómez	I-PER
de	I-PER
la	I-PER
Hera	I-PER
,	O
explicó	O
hoy	O
en	O
conferencia	O
de	O
prensa	O
que	O
este	O

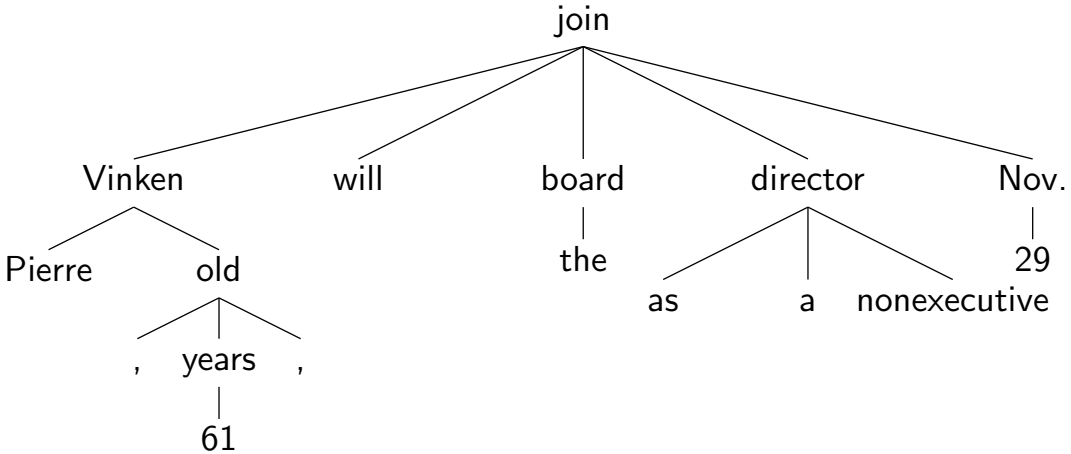
Textual data

Constituent-structure treebanks



- Penn Treebank

Textual data
Dependency treebanks



1	Pierre	NNP	2
2	Vinken	NNP	9
3	,	,	2
4	61	CD	5
5	years	NNS	6
6	old	JJ	2
7	,	,	2
8	will	MD	9
9	join	VB	0
10	the	DT	11
11	board	NN	9
12	as	IN	15
13	a	DT	15
14	nonexecutive	JJ	15
15	director	NN	9
16	Nov.	NNP	9
17	29	CD	16
18	.	.	9

- Much more compact than constituent trees, equivalent for practical purposes
- Purpose: training a parser (interpretation, translation)

Research Example

How can we learn a parser *without* a treebank?

- Motivations
  - Linguistics: ultimate subject matter is **human language capacity** = ability to learn language.
  - Google: access to data in all languages
  - DARPA: decision support in crisis management; information extraction from news media and social media
- Treebanks
  - I know of treebanks for 43 languages:

Arabic	English, Middle	Hungarian	Romanian
Armenian, Ancient	English, Old	Icelandic	Russian
Basque	Estonian	Indonesian	Slavonic, Old Church
Bulgarian	Finnish	Italian	Slovene
Catalan	French	Japanese	Spanish
Chinese	German	Karuk	Swedish
Czech	Gothic	Korean	Thai
Danish	Greek	Latin	Turkish
Dutch	Greek, Ancient	Polish	Ugaritic
English	Hebrew	Portuguese	Vietnamese
English, Early Modern	Hindi-Urdu	Portuguese, Medieval	

  - But there are 6800 languages (Ethnologue)

Research Example

Learning a dependency parser for a new language

- Accuracy measure: percentage of governors correctly identified

Monolingual grammatical inference	47%
Delexicalized transfer	52%
Multi-source delexicalized transfer	55%
Adaptation using bitexts	59%
Adaptation using bitexts + language relationships	62%
Supervised training	84%

McDonald et al 2012

- Existing methods neglect bilingual dictionaries

Research Example

A challenge: low-resource languages

- Making the methods practical for **low-resource languages**
  - Well-resourced languages: $\sim 50 = 0.7\%$
  - E.g., Google translates 57 languages
  - All but 18 are Indo-european, none are endangered.
- Example: machine translation
  - Current methods require 2–10 million words of **bitext** for training
  - Largest source of bitext is the Bible: 0.8 M words, 459 languages (7%).
  - New Testament: 0.1 M words, 1213 languages (18%)

The Human Language Project

Bootstrapping resources for low-resource languages

- Where we would really like to go
  - Comprehensive language resources
  - The Human Genome Project for languages
- Made urgent by language endangerment
  - "Low resource" = digitally endangered
  - 33% endangered, another 10% vulnerable
  - Half of the world's languages have fewer than 6,000 speakers.
  - 4% have gone extinct since 1950
  - Current rate of extinction: 2 languages/month
  - Projections: 50–90% loss by end of century

The Human Language Project

Resources

- What do we mean by resources?
  - Target-language plain text
  - Monolingual dictionaries with parts of speech
  - Bilingual dictionaries
  - Morphological paradigms
  - Bitext
  - Treebank
- All expressible in a simple data format

1	Pierre	NNP	2
2	Vinken	NNP	9
3	,	,	2
4	61	CD	5
5	years	NNS	6
6	old	JJ	2
7	,	,	2
8	will	MD	9
9	join	VB	0
10	the	DT	11
11	board	NN	9
12	as	IN	15
13	a	DT	15
14	nonexecutive	JJ	15
15	director	NN	9
16	Nov.	NNP	9
17	29	CD	16
18	.	.	9

What Libraries Might Provide

Scanned books

- Traditional language description
  - A grammar and a lexicon, maybe a text collection
  - Printed books, for human consumption
  - All that is available for most languages

Dat.	*Ako-ṭhen, ṭheċ̀, sen, seċ̀*, to themselves.
Acc.	*Ako*, themselves.
Abl.	*Ako-khǫn, khǫnaḱ*, from themselves.
Loc.	*Ako-re, talare*, in, on themselves.

Lars Skrefsrud, A grammar of the Santhal language (1873)

घोषण ghosh-aṇa, *a.* sounding; *n.*, **ā**, *f.* proclamation.
घोषवत् ghosha-vat, *a.* sounding, roaring; sonant (*gr.*): **-i**, *f. kind of* lute; **-vriddha**, *m.* elder of a herdsmen's station.
घोषि ghósh-i, **घोषिन्** ghosh-in, *a.* sounding; noisy: (*n*)**-i**, *f. pl. kind of* demon.

Arthur Macdonell, A Sanskrit-English dictionary (1893)

What Libraries Might Provide

Scanned books

- Difficulties
  - OCR is a huge problem – poor or nonexistent for non-roman scripts, poor for text with diacritics
  - Possible alternative: crowd-sourcing
  - Transcription is one thing, conversion to dataset is another

What Libraries Might Provide

Libraries as digital archives?

- Biggest current provider: Linguistic Data Consortium
  - Heavy emphasis on speech, languages with overwhelming commercial and intelligence value (English, Chinese, Arabic, western Europe)
  - Expensive
- Language documentation archives
  - Archiving of traditional field notes, recordings
  - Small scale, little support for or awareness of computational methods
  - Access is often highly restricted

What Libraries Might Provide

Libraries as digital archives?

- What is lacking: archives that provide
  - Free public access
  - Comprehensiveness
  - Machine consumption
- Broader issues: incentives/hindrances to producing resources
  - Recognition of dataset production as publication
  - Production of machine-oriented datasets from copyrighted print works
  - Ability to publish annotation of others' datasets