Corpus Annotation

Gerald Penn

CSC 401
University of Toronto

`http://www.cs.toronto.edu/∼gpenn/csc401`

Corpus (*pl.* Corpora)

A corpus is a collection of text(s) or utterances

- 10^6: tiny
- 10^9: reasonable
- 10^{12}: current feasible limit for unannotated data

The most valuable corpora are those that occur naturally:

- newspaper
- collection of newspapers, e.g., 1989 *Wall Street Journal*
- complete works of Mark Twain
- all essays written last year by 8th grade students in Toronto

Notable Corpora

- Brown Corpus (1 million tokens, 61,805 types): *balanced* collection of representative genres of American English in 1960.
- Lancaster-Oslo-Bergen (LOB) Corpus: British equivalent (1970s)
- Penn Treebank (2.88 million): syntactically annotated Brown Corpus, plus others, incl. 1989 *Wall Street Journal*
- London-Lund Corpus (435K tokens): transcriptions of 87 British English conversations
- Switchboard Corpus (120 hours, ≈ 2.4M tokens) 2400 telephone conversations between US English speakers
- Hansard Corpus (ongoing): Canadian parliamentary proceedings, French-English bilingual

The Web as Corpus

- AV-ENG: all English-language web-pages indexed by Alta Vista (in 2003, $\approx$ 100 GB tokens, 350M pages).
- AV-CA (2 GB tokens): the `.ca` subset of AV-ENG
- Google n-Gram Corpus: an index of five-word sequences that appear at least 40 times in 10^{12} words of web text.
- etc.

POS Tags

POS tags label parts of speech in a corpus.

There are different *tagsets*, each with their own parochial features, e.g., the Brown Corpus tagset:

- attributes: NN-TL (title word)
- foreign word marking: FW-NN
- *elsewhere* conventions: NN vs. NNS
- implicit groupings: NN, NP, NPS, NR, ...

Category	Examples	Claws c5	Brown	Penn
Adjective	happy, bad	AJ0	JJ	JJ
Adjective, ordinal number	sixth, 72nd, last	ORD	OD	JJ
Adjective, comparative	happier, worse	AJC	JJR	JJR
Adjective, superlative	happiest, worst	AJS	JJT	JJS
Adjective, superlative, semantically	chief, top	AJ0	JJS	JJ
Adjective, cardinal number	3, fifteen	CRD	CD	CD
Adjective, cardinal number, one	one	PNI	CD	CD
Adverb	often, particularly	AV0	RB	RB
Adverb, negative	not, n't	XX0	*	RB
Adverb, comparative	faster	AV0	RBR	RBR
Adverb, superlative	fastest	AV0	RBT	RBS
Adverb, particle	up, off, out	AVP	RP	RP
Adverb, question	when, how, why	AVQ	WRB	WRB
Adverb, degree & question	how, however	AVQ	WQL	WRB
Adverb, degree	very, so, too	AV0	QL	RB
Adverb, degree, postposed	enough, indeed	AV0	QLP	RB
Adverb, nominal	here, there, now	AV0	RN	RB
Conjunction, coordination	and, or	CJC	CC	CC
Conjunction, subordinating	although, when	CJS	CS	IN
Conjunction, complementizer *that*	that	CJT	CS	IN
Determiner	this, each, another	DT0	DT	DT
Determiner, pronoun	any, some	DT0	DTI	DT
Determiner, pronoun, plural	these, those	DT0	DTS	DT
Determiner, prequalifier	quite	DT0	ABL	PDT
Determiner, prequantifier	all, half	DT0	ABN	PDT
Determiner, pronoun or double conj.	both	DT0	ABX	DT (CC)
Determiner, pronoun or double conj.	either, neither	DT0	DTX	DT (CC)
Determiner, article	the, a, an	AT0	AT	DT
Determiner, postdeterminer	many, same	DT0	AP	JJ
Determiner, possessive	their, your	DPS	PP$	PRP$
Determiner, possessive, second	mine, yours	DPS	PP$$	PRP
Determiner, question	which, whatever	DTQ	WDT	WDT
Determiner, possessive & question	whose	DTQ	WP$	WP$
Noun	aircraft, data	NN0	NN	NN
Noun, singular	woman, book	NN1	NN	NN
Noun, plural	women, books	NN2	NNS	NNS
Noun, proper, singular	London, Michael	NP0	NP	NNP
Noun, proper, plural	Australians, Methodists	NP0	NPS	NNPS
Noun, adverbial	tomorrow, home	NN0	NR	NN
Noun, adverbial, plural	Sundays, weekdays	NN2	NRS	NNS
Pronoun, nominal (indefinite)	none, everything, one	PNI	PN	NN
Pronoun, personal, subject	you, we	PNP	PPSS	PRP
Pronoun, personal, subject, 3SG	she, he, it	PNP	PPS	PRP
Pronoun, personal, object	you, them, me	PNP	PPO	PRP
Pronoun, reflexive	herself, myself	PNX	PPL	PRP
Pronoun, reflexive, plural	themselves, ourselves	PNX	PPLS	PRP
Pronoun, question, subject	who, whoever	PNQ	WPS	WP
Pronoun, question, object	who, whoever	PNQ	WPO	WP
Pronoun, existential there	there	EX0	EX	EX

Table 4.5 Comparison of different tag sets: adjective, adverb, conjunction, determiner, noun, and pronoun tags.

Category	Examples	Claws c5	Brown	Penn
Verb, base present form (not infinitive)	take, live	VVB	VB	VBP
Verb, infinitive	take, live	VVI	VB	VB
Verb, past tense	took, lived	VVD	VBD	VBD
Verb, present participle	taking, living	VVG	VBG	VBG
Verb, past/passive participle	taken, lived	VVN	VBN	VBN
Verb, present 3SG *-s* form	takes, lives	VVZ	VBZ	VBZ
Verb, auxiliary *do*, base	do	VDB	DO	VBP
Verb, auxiliary *do*, infinitive	do	VDB	DO	VB
Verb, auxiliary *do*, past	did	VDD	DOD	VBD
Verb, auxiliary *do*, present part.	doing	VDG	VBG	VBG
Verb, auxiliary *do*, past part.	done	VDN	VBN	VBN
Verb, auxiliary *do*, present 3SG	does	VDZ	DOZ	VBZ
Verb, auxiliary *have*, base	have	VHB	HV	VBP
Verb, auxiliary *have*, infinitive	have	VHI	HV	VB
Verb, auxiliary *have*, past	had	VHD	HVD	VBD
Verb, auxiliary *have*, present part.	having	VHG	HVG	VBG
Verb, auxiliary *have*, past part.	had	VHN	HVN	VBN
Verb, auxiliary *have*, present 3SG	has	VHZ	HVZ	VBZ
Verb, auxiliary *be*, infinitive	be	VBI	BE	VB
Verb, auxiliary *be*, past	were	VBD	BED	VBD
Verb, auxiliary *be*, past, 3SG	was	VBD	BEDZ	VBD
Verb, auxiliary *be*, present part.	being	VBG	BEG	VBG
Verb, auxiliary *be*, past part.	been	VBN	BEN	VBN
Verb, auxiliary *be*, present, 3SG	is, 's	VBZ	BEZ	VBZ
Verb, auxiliary *be*, present, 1SG	am, 'm	VBB	BEM	VBP
Verb, auxiliary *be*, present	are, 're	VBB	BER	VBP
Verb, modal	can, could, 'll	VM0	MD	MD
Infinitive marker	to	TO0	TO	TO
Preposition, to	to	PRP	IN	TO
Preposition	for, above	PRP	IN	IN
Preposition, of	of	PRF	IN	IN
Possessive	's, '	POS	$	POS
Interjection (or other isolate)	oh, yes, mmm	ITJ	UH	UH
Punctuation, sentence ender	. ! ?	PUN	.	.
Punctuation, semicolon	;	PUN	.	:
Punctuation, colon or ellipsis	: . . .	PUN	:	:
Punctuation, comma	,	PUN	,	,
Punctuation, dash	–	PUN	–	–
Punctuation, dollar sign	$	PUN	not	$
Punctuation, left bracket	([{	PUL	(	(
Punctuation, right bracket	)] }	PUR	)	)
Punctuation, quotation mark, left	' "	PUQ	not	"
Punctuation, quotation mark, right	' "	PUQ	not	"
Foreign words (not in English lexicon)		UNC	(FW-)	FW
Symbol	[fj] *		not	SYM
Symbol, alphabetical	A, B, c, d	ZZ0		
Symbol, list item	A A. First			LS

Table 4.6 Comparison of different tag sets: Verb, preposition, punctuation and symbol tags. An entry of 'not' means an item was ignored in tagging, or was not separated off as a separate token.

POS Tags

In many respects, they are internally inconsistent:

- clitics:
  - PPS + MD (she'll)
  - BEZ* (isn't)
  - MD + RB (ca n't)
  - NN + $ (cat 's)
  - NP$ (children's)
- shape vs. form:
  - VBG (both gerund and participle)

. . . and inconsistent with each other:

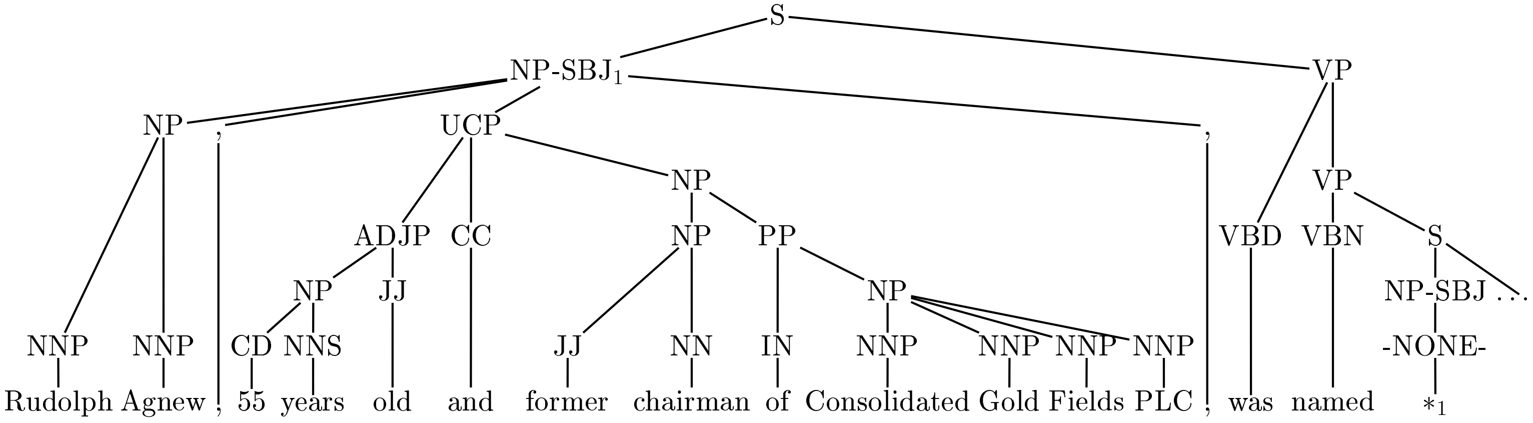
- punctuation classes
- subordinating conjunctions: *because*, *while*, . . .
  - with coordinating conjunctions
  - with prepositions

Kinds of Annotation

- part-of-speech (POS)
- lemmatization
  - find the true word-type, e.g., *listen* vs. *listener*
  - sometimes requires finding true word sense, e.g., *lie/lay* vs. *lie/lied* vs. *lay/laid*
- stemming (e.g. Porter or Lovins stemmer)
  - (mostly) used to approximate lemmatization
  - strips off affixes (derivational and inflectional):
    - *listen*, *listener* → *listen*
    - *arbitrariness* → *arbitrary*
    - but also *business* → *busy*
  - result may not even have dictionary entry or POS:
    - *geographer*, *geography*, *geographic* → *geograph*

Kinds of Annotation

- part-of-speech (POS)
- lemmatization
- stemming
- syntactic structure:
  - usually glorified CF parse trees, e.g., Penn Treebank



  - typed feature structures, e.g., LiNGO Redwoods

Genres/Subgenres used by the Brown Corpus

Table 1: Genre and subgenre codes for the Brown Corpus

I. Press

- A. Reportage
- B. Editorial
- C. Reviews

II. Miscellaneous

- D. Religion
- E. Skills and hobbies
- F. Popular lore
- G. Belles-lettres

III. Formal documents

- H. Government and institutional
- J. Learned

IV. Fiction

- K. General
- L. Mystery
- M. Science fiction
- N. Adventure
- P. Romance
- R. Humour

Genres are not the same as topics or sources, . . .
but they are somewhat confused here.

Mark-Up Languages

This is the only general solution

- SGML: ancestor of all most common mark-ups, dates back to 1960s
- XML: simplified SGML

These use *document type definitions (DTDs)*

- HTML: most common DTD (web-pages)

In NLP, one finds mark-up in:

- annotated resources
- textual sources, e.g., using web pages to train statistical tools
  - DTD (esp. HTML) often used to hack layout
  - DTD usually doesn't annotate the right structures for NLP
  - Document often deviates from DTD
  - So we generally ignore the DTD
  - Style sheets may one day be of use

End-of-Sentence Marking

Even this is not easy
Heuristic algorithm:

1. Provisionally accept all '.', '?', and '!' as EOS boundaries
2. If boundary is followed by quotation mark, move beyond mark, e.g., *Wayne said, "Me like hockey."*
3. Disqualify full-stop boundaries that are:
   - preceded by known non-final abbreviations, e.g., *Prof.*, *vs.*
   - preceded by known abbreviations, but not followed by capitalized word, e.g., *Sammy Davis Jr. died this morning in his Los Angeles home.*
4. Disqualify '?' and '!' boundaries followed by lower case letter or known name, e.g., *"What did you see?" John said.*

End-of-Sentence Marking

Other approaches:

- *Decision tree* [Riley, 1989], trained on:
  - (type) case
  - length of preceding and following words
  - prior probability of every word to occur before/after EOS boundary
- *Neural network* [Palmer & Hearst, 1997], trained on:
  - window of POS probability vectors
  - 3 tokens on each side

  98.5% accuracy
- *Maximum Entropy modelling* [Mikheev, 1998], 99.25 % accuracy