UNIB20005

Language and Computation

Corpora

Steven Bird, Department of Computing and Information Systems

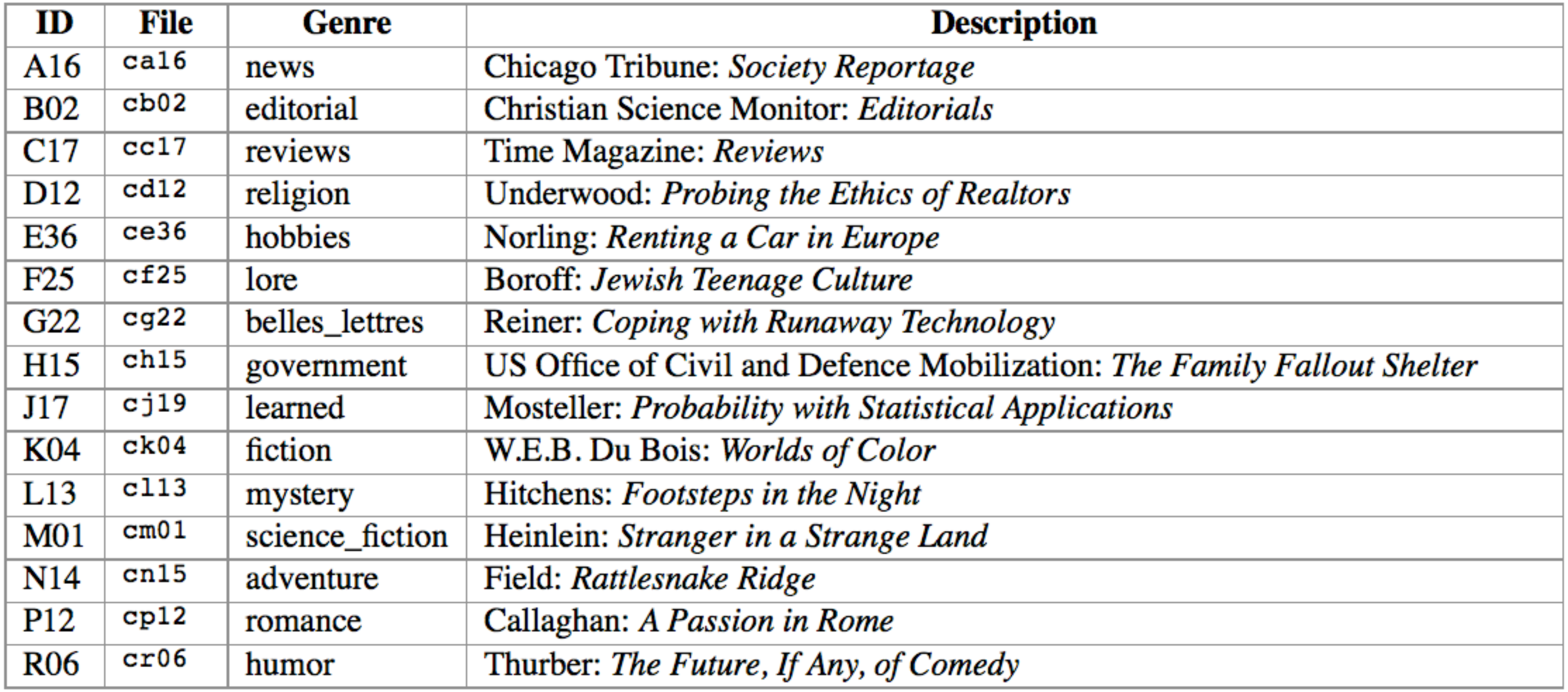
Corpus example: “must”

- “epistemic” -- in a manner that pertains to knowledge
- “deontic” -- pertaining to necessity, duty, or obligation

Corpus

- representative sample of language use
  - large, balanced
- important examples
  - Gutenberg: 25k books, gutenberg.org
  - Web text, e.g. imsdb.com
  - Brown Corpus, 1M words, 500 sources
  - Reuters
  - Inaugural Address Corpus

Brown Corpus: genres



ID	File	Genre	Description
A16	`ca16`	news	Chicago Tribune: *Society Reportage*
B02	`cb02`	editorial	Christian Science Monitor: *Editorials*
C17	`cc17`	reviews	Time Magazine: *Reviews*
D12	`cd12`	religion	Underwood: *Probing the Ethics of Realtors*
E36	`ce36`	hobbies	Norling: *Renting a Car in Europe*
F25	`cf25`	lore	Boroff: *Jewish Teenage Culture*
G22	`cg22`	belles_lettres	Reiner: *Coping with Runaway Technology*
H15	`ch15`	government	US Office of Civil and Defence Mobilization: *The Family Fallout Shelter*
J17	`cj19`	learned	Mosteller: *Probability with Statistical Applications*
K04	`ck04`	fiction	W.E.B. Du Bois: *Worlds of Color*
L13	`cl13`	mystery	Hitchens: *Footsteps in the Night*
M01	`cm01`	science_fiction	Heinlein: *Stranger in a Strange Land*
N14	`cn15`	adventure	Field: *Rattlesnake Ridge*
P12	`cp12`	romance	Callaghan: *A Passion in Rome*
R06	`cr06`	humor	Thurber: *The Future, If Any, of Comedy*

Brown Corpus

She/pps was/bedz getting/vbg real/ql dramatic/jj ./.

I'd/ppss+md have/hv been/ben more/ql impressed/vbn if/cs I/ ppss hadn't/hvd* remembered/vbn that/cs she'd/pps+hvd played/vbn Hedda/np

Gabler/np in/in her/pp$ highschool/nn dramatics/nn course/ nn ./.

I/ppss didn't/dod* want/vb her/ppo back/rb on/in that/ql broken/vbn record/nn ./

Reuters Corpus: topics

acq alum barley bop carcass castor-oil cocoa coconut coconut-oil coffee copper copra-cake corn cotton cotton-oil cpi cpu crude dfl dlr dmk earn fuel gas gnp gold grain groundnut groundnut-oil heat hog housing income instal-debt interest ipi iron-steel jet jobs l-cattle lead lei lin-oil livestock lumber meal-feed money-fx money-supply naphtha nat-gas nickel nkr nzdlr oat oilseed orange palladium palm-oil palmkernel pet-chem platinum potato propane rand rape-oil rapeseed reserves retail rice rubber rye ship silver sorghum soy-meal soy-oil soybean strategic-metal sugar sun-meal sun-oil sunseed tea tin trade veg-oil wheat wpi yen zinc

Reuters Corpus

HILLSDOWN BUYS BEDDING COMPANIES FOR 23 MLN DLRS

Hillsdown Holdings Plc <HLDN.L> said its

Christie-Tyler Ltd unit would buy the European bedding making interests of Simmons Co U.S.A., Owned by Gulf and Western

Industries Inc USA <GW>, for 23 mln dlrs.

The acquisitions include <Sleepeeze Ltd> in the U.K., <Compagnie Continentale Simmons SA> in France and <Compagnia Italiana Simmons SpA> in Italy.

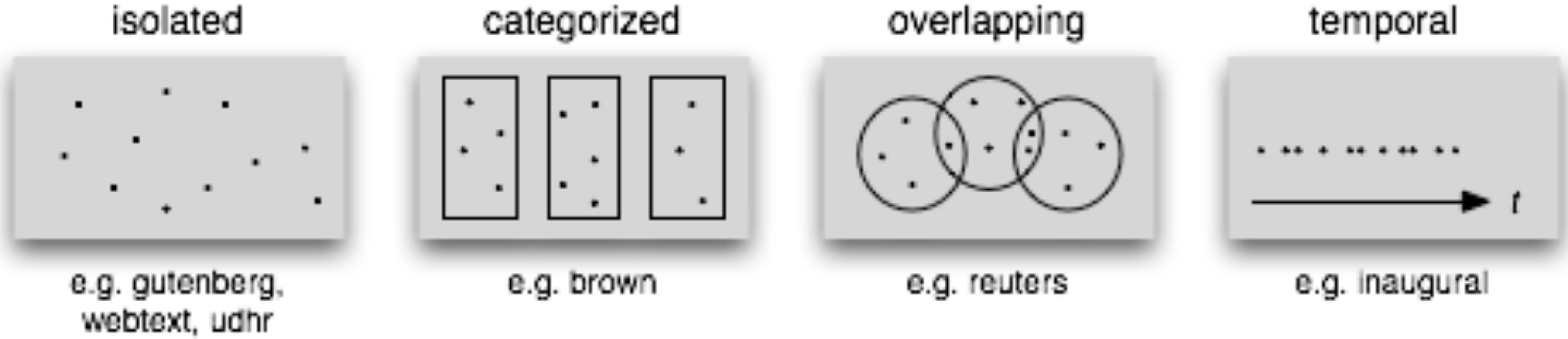
In 1986 the three businesses made pre-tax profit of around 2.5 mln stg on sales of 39 mln stg. Net assets being acquired come to around nine mln stg.

Hillsdown shares were unchanged at 266p.

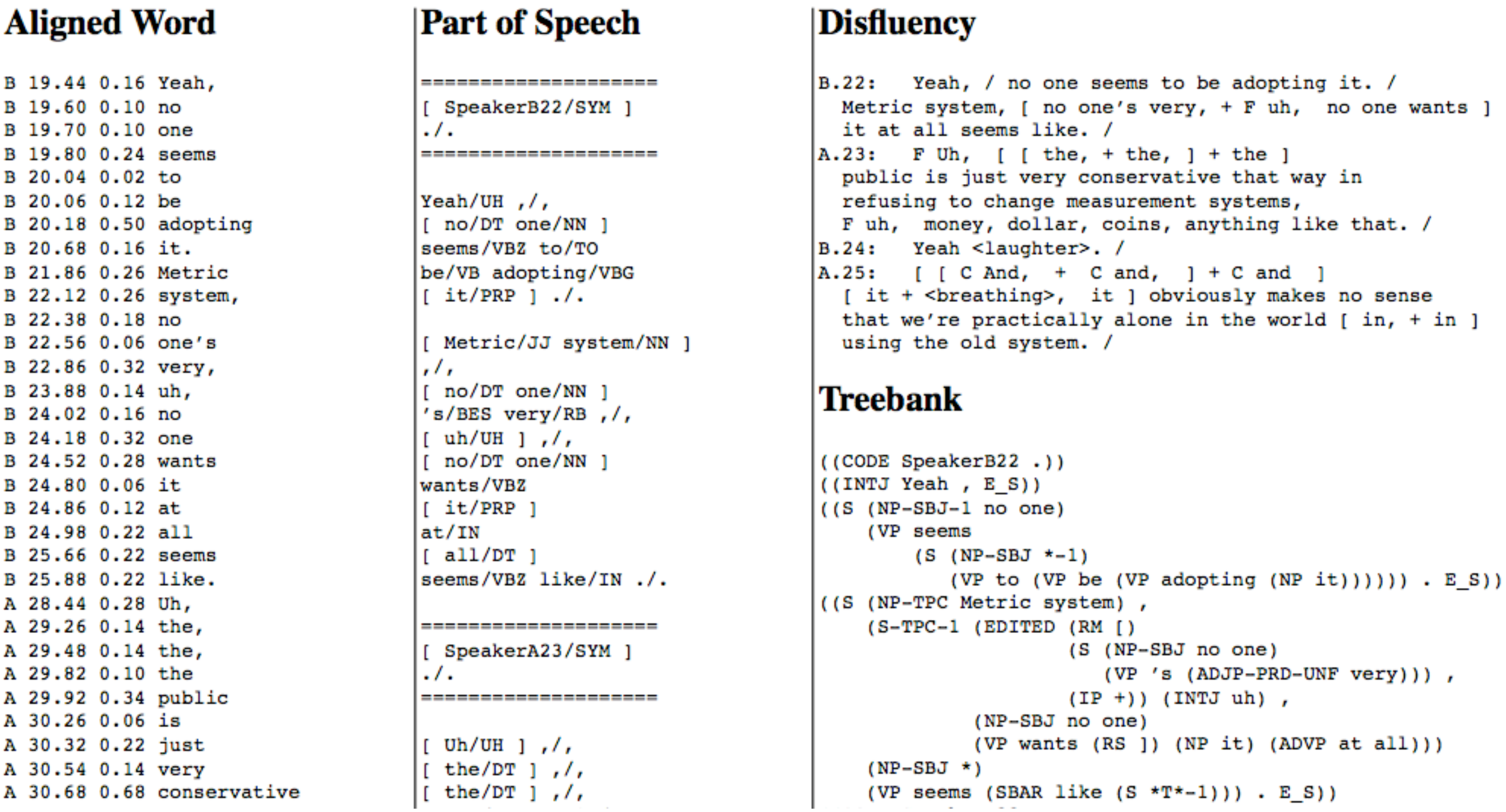
Inaugural Address Corpus

In unfolding to my countrymen the principles by which I shall be governed in the fulfillment of those duties my first resort will be to that Constitution which I shall swear to the best of my ability to preserve, protect, and defend. That revered instrument enumerates the powers and prescribes the duties of the Executive Magistrate, and in its first words declares the purposes to which these and the whole action of the Government instituted by it should be invariably and sacredly devoted -- to form a more perfect union, establish justice, insure domestic tranquillity, provide for the common defense, promote the general welfare, and secure the blessings of liberty to the people of this Union in their successive generations. Since the adoption of this social compact one of

Text Corpus Structure



Annotated Corpora



Aligned Word

```
B 19.44 0.16 Yeah,
B 19.60 0.10 no
B 19.70 0.10 one
B 19.80 0.24 seems
B 20.04 0.02 to
B 20.06 0.12 be
B 20.18 0.50 adopting
B 20.68 0.16 it.
B 21.86 0.26 Metric
B 22.12 0.26 system,
B 22.38 0.18 no
B 22.56 0.06 one's
B 22.86 0.32 very,
B 23.88 0.14 uh,
B 24.02 0.16 no
B 24.18 0.32 one
B 24.52 0.28 wants
B 24.80 0.06 it
B 24.86 0.12 at
B 24.98 0.22 all
B 25.66 0.22 seems
B 25.88 0.22 like.
A 28.44 0.28 Uh,
A 29.26 0.14 the,
A 29.48 0.14 the,
A 29.82 0.10 the
A 29.92 0.34 public
A 30.26 0.06 is
A 30.32 0.22 just
A 30.54 0.14 very
A 30.68 0.68 conservative
```

Part of Speech

```
====================
[ SpeakerB22/SYM ]
./.
====================

Yeah/UH ,/,
[ no/DT one/NN ]
seems/VBZ to/TO
be/VB adopting/VBG
[ it/PRP ] ./.

[ Metric/JJ system/NN ]
,/,
[ no/DT one/NN ]
's/BES very/RB ,/,
[ uh/UH ] ,/,
[ no/DT one/NN ]
wants/VBZ
[ it/PRP ]
at/IN
[ all/DT ]
seems/VBZ like/IN ./.

====================
[ SpeakerA23/SYM ]
./.
====================

[ Uh/UH ] ,/,
[ the/DT ] ,/,
[ the/DT ] ,/,
```

Disfluency

```
B.22:   Yeah, / no one seems to be adopting it. /
  Metric system, [ no one's very, + F uh,  no one wants ]
  it at all seems like. /
A.23:   F Uh,  [ [ the, + the, ] + the ]
  public is just very conservative that way in
  refusing to change measurement systems,
  F uh,  money, dollar, coins, anything like that. /
B.24:   Yeah <laughter>. /
A.25:   [ [ C And,  +  C and,  ] + C and  ]
  [ it + <breathing>,  it ] obviously makes no sense
  that we're practically alone in the world [ in, + in ]
  using the old system. /
```

Treebank

```
((CODE SpeakerB22 .))
((INTJ Yeah , E_S))
((S (NP-SBJ-1 no one)
    (VP seems
        (S (NP-SBJ *-1)
           (VP to (VP be (VP adopting (NP it)))))) . E_S))
((S (NP-TPC Metric system) ,
    (S-TPC-1 (EDITED (RM [)
                      (S (NP-SBJ no one)
                         (VP 's (ADJP-PRD-UNF very))) ,
                      (IP +)) (INTJ uh) ,
             (NP-SBJ no one)
             (VP wants (RS ]) (NP it) (ADVP at all)))
    (NP-SBJ *)
    (VP seems (SBAR like (S *T*-1))) . E_S))
```

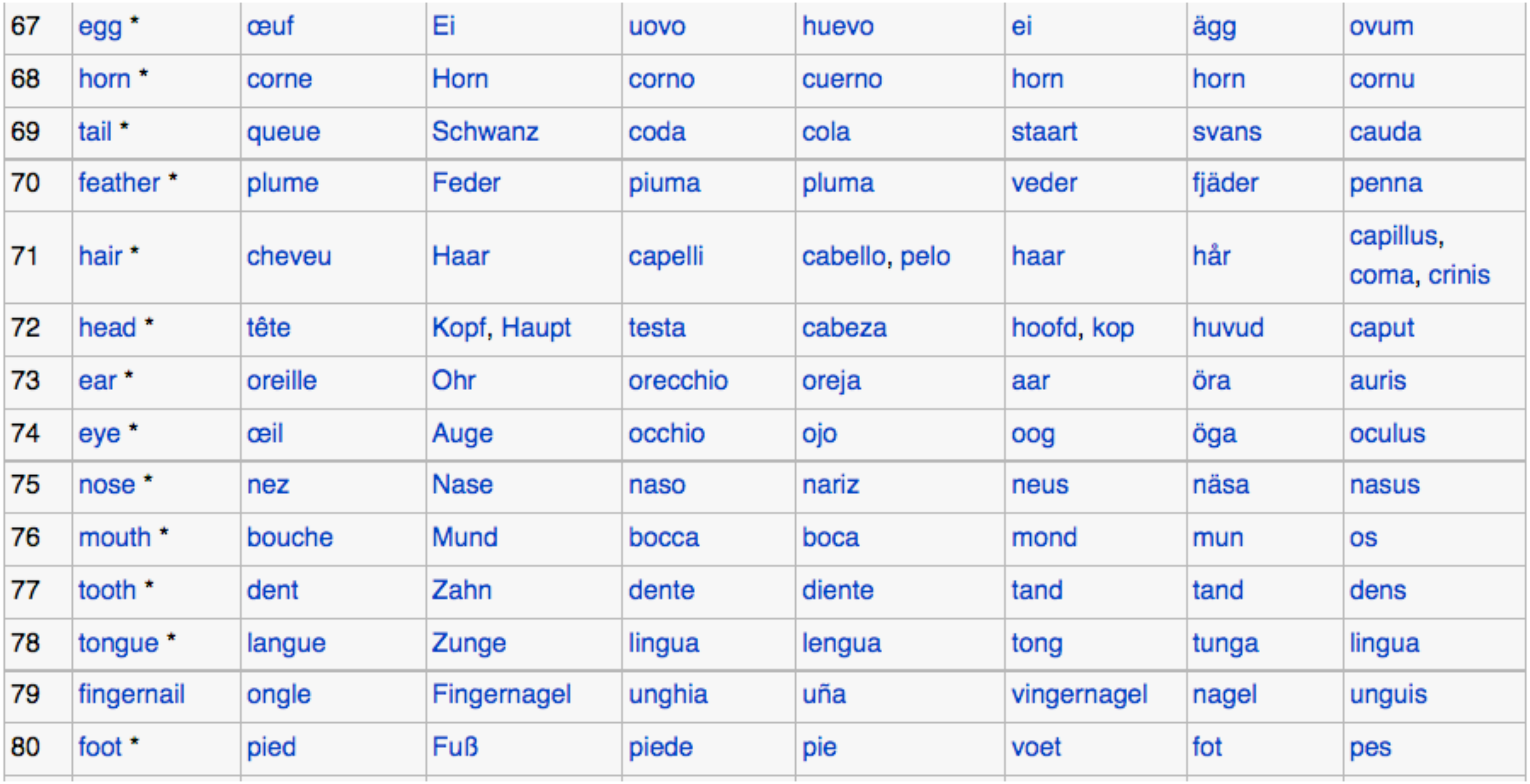
Another resource type: lexicons

- record vs temporal structure
- CMU Pronunciation Lexicon
- Swadesh Wordlists
- Princeton Wordnet

CMU Pronlex

- UNUSABLE 1 AH0 N Y UW1 Z AH0 B AH0 L
- UNUSED 1 AH0 N Y UW1 Z D
- UNUSUAL 1 AH0 N Y UW1 ZH AH0 W AH0 L
- UNUSUAL 2 AH0 N Y UW1 ZH UW0 AH0 L
- UNUSUAL 3 AH0 N Y UW1 ZH W AH0 L
- UNUSUALLY 1 AH0 N Y UW1 ZH AH0 W AH0 L IY0
- UNUSUALLY 2 AH0 N Y UW1 ZH UW0 AH0 L IY0
- UNUSUALLY 3 AH0 N Y UW1 ZH W AH0 L IY0

Swadesh Wordlist



67	egg *	œuf	Ei	uovo	huevo	ei	ägg	ovum
68	horn *	corne	Horn	corno	cuerno	horn	horn	cornu
69	tail *	queue	Schwanz	coda	cola	staart	svans	cauda
70	feather *	plume	Feder	piuma	pluma	veder	fjäder	penna
71	hair *	cheveu	Haar	capelli	cabello, pelo	haar	hår	capillus, coma, crinis
72	head *	tête	Kopf, Haupt	testa	cabeza	hoofd, kop	huvud	caput
73	ear *	oreille	Ohr	orecchio	oreja	aar	öra	auris
74	eye *	œil	Auge	occhio	ojo	oog	öga	oculus
75	nose *	nez	Nase	naso	nariz	neus	näsa	nasus
76	mouth *	bouche	Mund	bocca	boca	mond	mun	os
77	tooth *	dent	Zahn	dente	diente	tand	tand	dens
78	tongue *	langue	Zunge	lingua	lengua	tong	tunga	lingua
79	fingernail	ongle	Fingernagel	unghia	uña	vingernagel	nagel	unguis
80	foot *	pied	Fuß	piede	pie	voet	fot	pes

Princeton Wordnet

- online demo
- senses, lemmas, synsets
- lexical relations
  - hypernym / hyponym
  - holonym / meronym
    - part/substance/member
- entailment

Lexicon

Abstraction: fielded records

key	field	field	field	field
key	field	field	field	field

Eg: dictionary

wake: weɪk, [v], *cease to sleep...*
walk: wɔːk, [v], *progress by lifting and setting down each foot...*

Eg: comparative wordlist

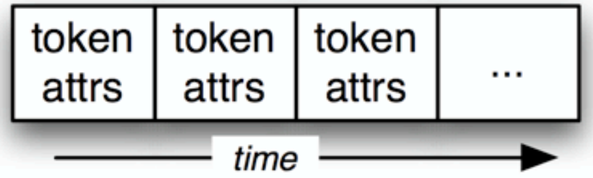
wake; aufwecken; acordar
walk; gehen; andar
write; schreiben; enscrever

Eg: verb paradigm

wake	woke	woken
write	wrote	written
wring	wrung	wrung

Text

Abstraction: time series



Eg: written text

A long time ago, Sun and Moon lived together. They were good brothers. ...

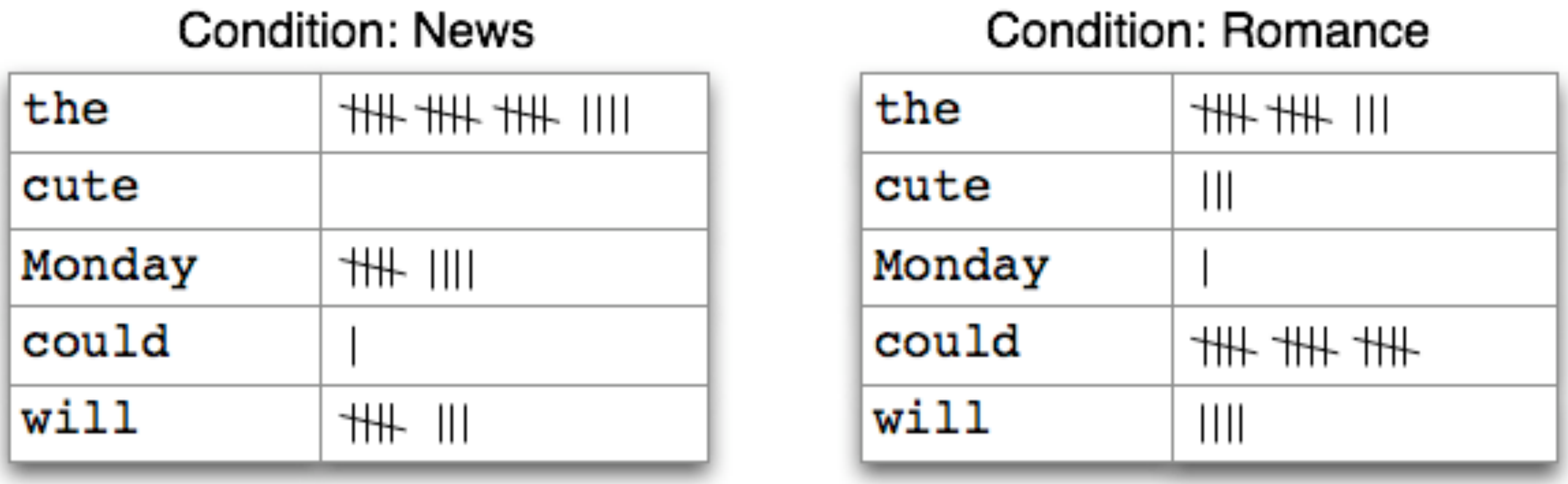
Eg: POS-tagged text

A/DT long/JJ time/NN ago/RB ,/, Sun/NNP and/CC Moon/NNP lived/VBD together/RB ./.

Eg: interlinear text

Ragaipa		irai		vateri	
ragai	-pa	ira	-i	vate	-ri
PP.1.SG	-BEN	RP.3.SG.M	-ABS	give	-2.SG

Conditional Frequency Distributions



Condition: News

the	𝍸 𝍸 𝍸 ||||
cute	
Monday	𝍸 ||||
could	|
will	𝍸 |||

Condition: Romance

the	𝍸 𝍸 |||
cute	|||
Monday	|
could	𝍸 𝍸 𝍸
will	||||