

## Data: four domains

**nantc** : press-wire news data

**classifieds** : real estate ads from on-line newspapers

**pc110** : palmtop mailing list (e-mail like)

**rfr** : rec.food.recipes USENET messages

Corpus	nantc	ads	pc110	rfr
total # tokens	4.3m	415k	264k	209k
# NSWs	377k	180k	72k	46k
% NSW	8.8%	43.4	27.3	22.0

	EXPN	abbreviation, contractions	adv, N.Y, mph, gov't
alpha	LSEQ	letter sequence	CIA, D.C, CDs
	ASWD	read as word	CAT, proper names
	MSPPL	misspelling	geography
	NUM	number (cardinal)	12, 45, 1/2, 0.6
	NORD	number (ordinal)	May 7, 3rd, Bill Gates III
	NTEL	telephone (or part of)	212 555-4523
	NDIG	number as digits	Room 101,
N	NIDE	identifier	747, 386, 15, PC110, 3A
U	NADDR	number as street address	5000 Pennsylvania, 4523 Forbes
M	NZIP	zip code or PO Box	91020
B	NTIME	a (compound) time	3.20, 11:45
E	NDATE	a (compound) date	2/2/99, 14/03/87 (or US) 03/14/87
R	NYER	year(s)	1998 80s 1900s 2003
S	MONEY	money (US or otherwise)	\$3.45 HK\$300, Y20,000, \$200K
	BMONY	money tr/m/billions	\$3.45 billion
	PRCT	percentage	75%, 3.4%
O	SLNT	not spoken, word boundary	word boundary or emphasis character:
T			M.bath, KENT*REALTY, _really-, *** Added
H	PUNC	not spoken, phrase boundary	non-standard punctuation: "... " in
E			DECIDE... Year, "***" in \$99,9K*** Whites
R	FNSP	funny spelling	sllooooooww, sh*t
	URL	url, pathname or email	http://api.co.uk, /usr/local, phj@teleport.com
	NONE	token should be ignored	ascii art, formatting junk

## Data: NSW distributions

	Domains			
	nantc	classifieds	pc110	rfr
ASWD	83.49	28.64	64.60	72.36
LSEQ	9.10	3.00	22.60	2.11
EXPN	7.41	68.36	12.80	25.53

	Domains			
	nantc	classifieds	pc110	rfr
NUM	66.11	58.26	43.77	97.90
NYER	19.06	0.70	0.51	0.27
NORD	9.37	3.37	4.45	0.11
NIDE	2.24	5.83	37.41	0.47
NTEL	1.25	25.92	1.32	0.02

# Hand labeling

- Each NSW presented in context
  - Three words either side
- One letter choice of TAG
  - or explicit expansion
  - splits “WinNT” → “Win” “NT”
- Test of inter-labeler agreement
  - 3 labelers natic, 2268 samples,  $\kappa = 0.81$
  - 9 labelers ads, 622 samples,  $\kappa = 0.84$
- Labeling held as XML markup

```
Today I bought a Sony<w NSW="LSEQ"> NP-F530,</w><w NSW="SPLT"><ws
NSW="NUM"> 1350</ws><ws NSW="EXPN">mAh.</ws></w> Like your<w
NSW="NIDE"> 550</w> it is slightly larger than the native<w
NSW="LSEQ"> IBM</w> battery pack. It's been<w NSW="NUM"> 3</w> hours
now on it's first charge - I am charging in the <w NSW="LSEQ"> PC110.
</w>
```

## Can we find NSWs?

- Tokens not in lexicon
- Plus
  - single character tokens
  - “punctuation”
  - common abbreviations (in lexicon)
- Misses homographic abbreviations/standard words
  - “sun”, “Jan”
  - also domain specific ones, “kit” and “named”

Domain Dependent?	Detection Algorithm	Precision//Recall			
		nantc	ads	pc110	rfr
No	non-lexical	55/79	96/79	80/65	76/82
No	+ sct + abbrevs	44/93	95/91	70/90	73/96
Yes	++ abbrevs	39/93	92/92	60/91	46/97