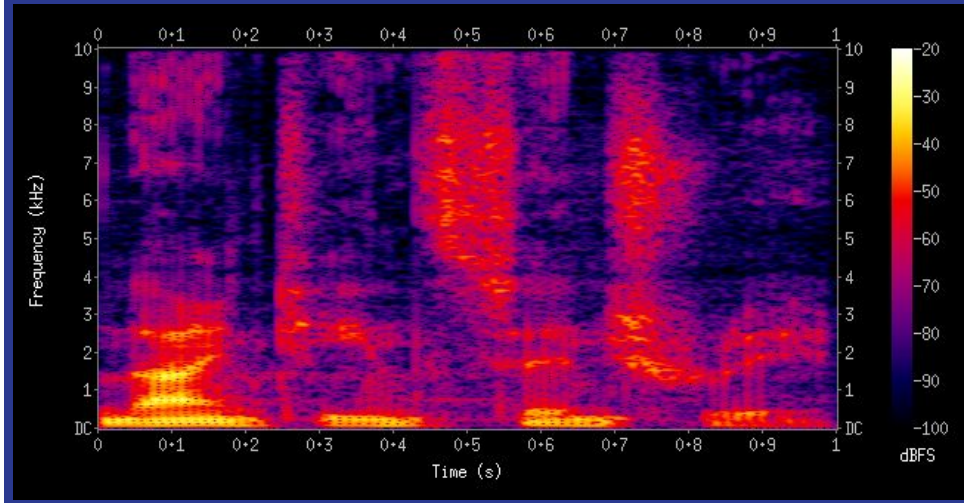
Natural Language Processing & Machine Learning for Speech

Paula Buttery, Andrew Caines, Helen Yannakoudakis;
NLIP Group, Dept. Computer Science & Technology, Cambridge.

19th century



Source: https://commons.wikimedia.org/wiki/File:Spectrogram-19thC.png

Overview

Speech vs Writing

Speech and writing share some commonalities but exhibit many differences, not just in mode of transmission, but form, construction and grammar

Speech processing

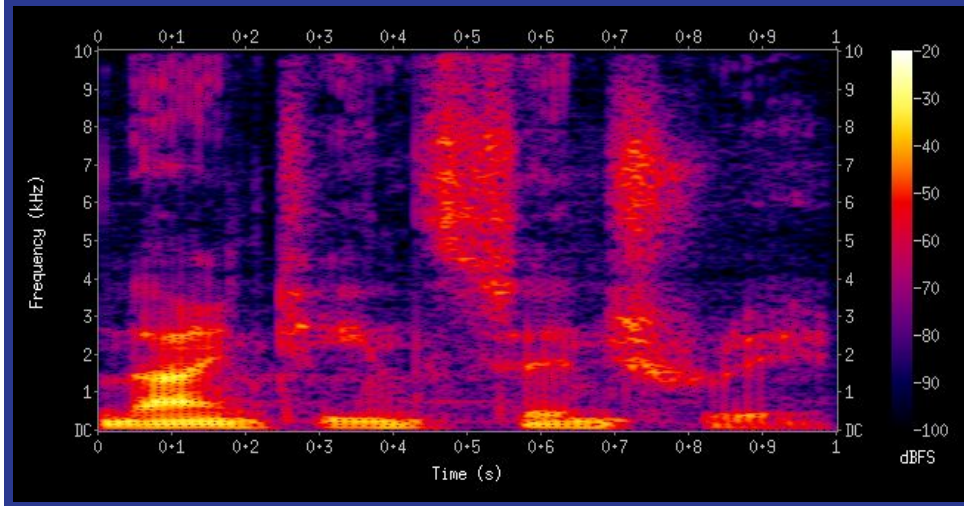
How to treat transcriptions of speech so that we can apply natural language processing techniques: more training data, test data normalisation, domain adaptation

Speech scoring

Example of NLP and machine learning for speech applications: automated assessment

Speech vs Writing

Andrew Caines



Source: https://commons.wikimedia.org/wiki/File:Spectrogram-19thC.png

Characteristics of Speech

- Speech is very different from writing
  - mode of transmission
  - phonetics, prosody, gesture (including sign language)
  - put these aside for now: consider the aspects we can examine in transcriptions
  - i.e. the lexis, morphology, syntax, semantics, pragmatics, discourse
  - note that the default speech mode involves interaction, no editing, multimodal grounding, background noise, facial expression and gesture

- Background reading:
  - Carter & McCarthy, 2017, ‘Spoken Grammar: Where are we and where are we going?’ *Applied Linguistics*.
  - Lau, Clark & Lappin, 2017, ‘Grammaticality, Acceptability, and Probability: A probabilistic view of linguistic knowledge.’ *Cognitive Science*.
  - Plank, 2016, ‘What to do about non-standard (or *non-canonical*) language in NLP.’ *KONVENS*.
  - Jurafsky & Martin, 2nd edn., Ch. 9 & 10.

Characteristics of Speech

- BBC News ‘Brexit: May to make plea to MPs for time to change deal’ https://www.bbc.co.uk/news/uk-47187491
  - Prime Minister Theresa May will ask MPs to give her more time to secure changes to the controversial part of her Brexit deal - the Northern Irish backstop. Mrs May is due to report back to MPs this week, after trying to persuade the EU to make last-minute changes. Labour wants to hold Mrs May to her word and make sure the vote is held. The shadow Brexit secretary, Sir Keir Starmer, has said Labour has drafted an amendment which, if passed this week, would guarantee a vote by the end of the month.
- Spoken corpus examples
  - um he’s a closet yuppie is what he is (Leech 2000)
  - I played, I played against um (Leech 2000)
  - You’re happy to -- welcome to include it (Levelt 1989)
- British National Corpus conversations:
  - Oi you, he's playing with your
  - Oh let's have a, is it in there?
  - (unclear) no
  - (pause) right, we'll have another cup of tea and then we'll have that nice cake
  - https://corpus.byu.edu/bnc [KGC]

Characteristics of Speech

- BBC Radio 4 In Touch: *Navigating University*
  https://www.bbc.co.uk/sounds/play/m0001f1d (2:20)
  - Megan: When I started as an undergraduate, I'd chosen the University of Gloucestershire and when I went on the open days they were the only university who gave me a prospectus in braille. I was so made up. It was interesting because I actually applied two years in advance because I took a year out to go and teach English in Germany. And by the time I came back, all the disability staff who were clued up seemed to have gone or moved on and the disability department was completely different.
  - That was the only thing I got in braille, pretty much, the seven years I was there. So, they hooked me in and then yeah…
  - White: And didn't really follow through.
  - Megan: No and the sad thing was, as well, I'd emailed the disability department just before I got on the plane to Germany and I said – please, could you make the lecturers aware that I'm registered blind so that we can start those discussions early with two years to go. And when I started at the university I walked in to my lectures and I was met with dismay, indifference and my lecturers had no clue about me arriving at all.

Characteristics of Speech

- Speech is very different from writing
- Even when viewed in writing
- (vice versa: imagine hearing written text read aloud, as in speeches, prayers, old-school conference papers)
- Become an observer!

- Problems for NLP:
  - Disfluencies
  - Tendency for long coordinated structures / Speech-unit delimitation
  - Overlap, interruption, subject-less structures, verb-less structures, acceptability appropriateness clarity over absolute grammaticality, incomplete propositions
  - Co-construction, multimodal physical context, background inter-personal relations & common ground
  - Creativity and language play

NLP of speech

- Caines, McCarthy, Buttery, *SCNLP* 2017

Medium	Tokens	Types
speech	394,611*	11,326**
writing	394,611	27,126

Table 2: Vocabulary sizes in selected corpora of English speech and writing (* sampled from 766,560 tokens in SWB corpus; ** mean of 100 samples, st.dev=45.5).

Speech	Freq.	Rank	Writing	Freq.
you know	11,165	1	of the	4313
it's	8531	2	in the	3702
that's	6708	3	to the	2352
don't	5680	4	I have	1655
I do	4390	5	on the	1607
I think	4142	6	I am	1500
and I	3790	7	for the	1475
I'm	3716	8	I would	1427
I I	3000	9	and the	1389
in the	2972	10	and I	1361
and uh	2780	11	to be	1318
a lot	2714	12	I was	1140
of the	2655	13	don't	1125
it was	2616	14	will be	1092
I mean	2518	15	it was	1057
kind of	2448	16	at the	1044
they're	2349	17	in a	1041
I've	2165	18	like to	1036
going to	2135	19	is a	1021
lot of	2053	20	it is	998

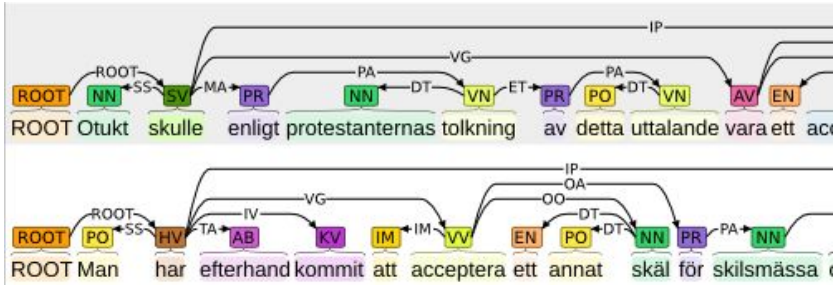
Table 3: The most frequently occurring bigrams in selected corpora of English speech (the Switchboard Corpus in Penn Treebank 3) and writing (EWT, LinES, TLE), normalised to counts per million.

NLP of speech

- Caines, McCarthy, Buttery, *SCNLP* 2017

Speech	Freq.	Rank	Writing	Freq.
VBP_PRP	51,845	1	NN_DT	48,846
NN_DT	47,469	2	NN_IN	36,274
ROOT_UH	39,067	3	NN_NN	27,490
IN_NN	26,868	4	NN_JJ	21,566
VB_PRP	24,321	5	VB_NN	19,584
ROOT_VBP	24,156	6	VB_PRP	16,320

Table 4: The most frequently occurring part-of-speech tag dependency pairs in selected corpora of English speech (the Switchboard Corpus in Penn Treebank 3) and writing (EWT, LinES, TLE), normalised to counts per million. The first tag in the pair is the head of the relation; the second is the dependent (Penn Treebank tagset).



Corpus	Medium	Units	Tokens	UAS
SWB	speech	102,900	766,560	.540
EWT	writing	14,545	218,159	.744
LinES	writing	3650	64,188	.758
TLE	writing	5124	96,180	.845

Table 5: Corpus sizes and overall unlabelled attachment scores using Stanford Core NLP; SWB=Switchboard, EWT=English Web Treebank, LinES=English section LinES, TLE=Treebank of Learner English

NLP of speech

- Caines, McCarthy, Buttery, *SCNLP* 2017

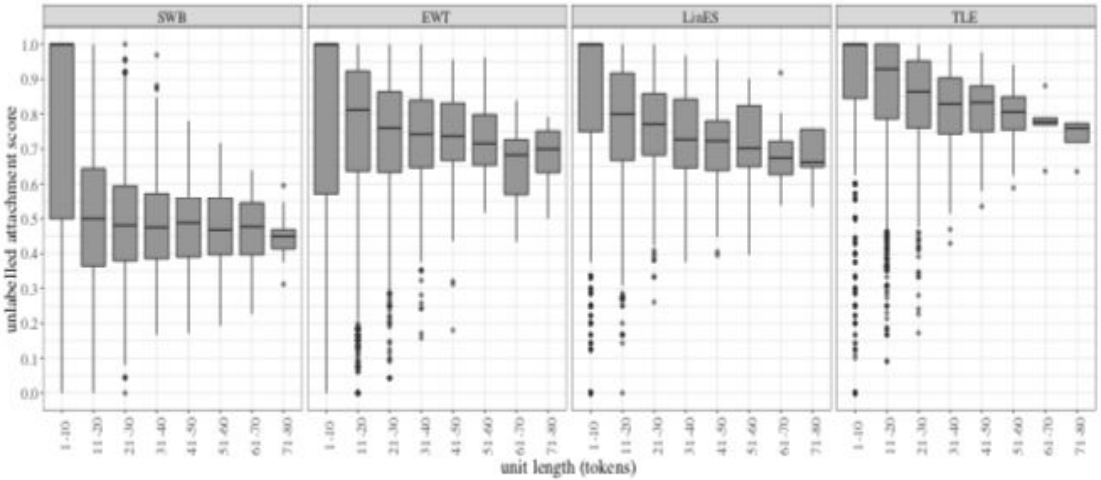


Figure 2: Unlabelled attachment scores by unit length in four English corpora.

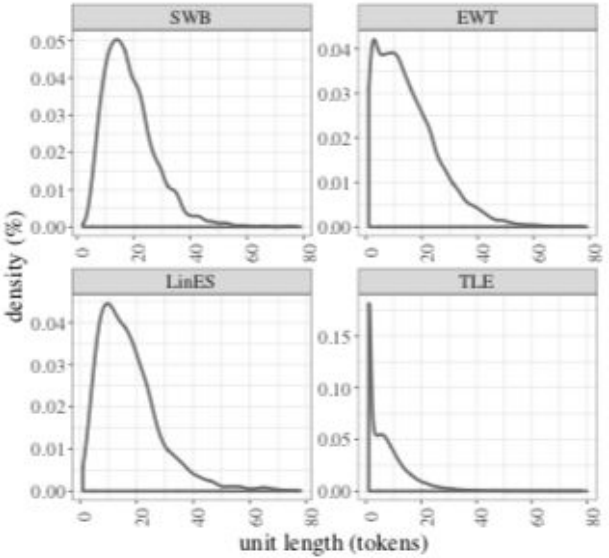


Figure 1: Density plot of unit lengths in four English corpora; SWB=Switchboard, EWT=English Web Treebank, LinES=English section LinES, TLE=Treebank of Learner English.

NLP of speech

- Caines & Buttery, *SANCL* 2014

mode	μ	Δ_{base}	¬T/frag	mode	μ	Δ_{base}	¬T/frag	mode	μ	Δ_{base}	¬T/frag
(A)	–2.599	0	.471	(A)	–2.599	0	.471	(A)	–2.599	0	.471
(B)	–2.094	+.505	.623	(BC)	–2.032	+.567	.689	(BCD)	–1.995	+.604	.715
(C)	–2.574	+.025	.484	(BD)	–2.049	+.550	.649		-	-	-
(D)	–2.563	+.036	.503	(CD)	–2.545	+.054	.523		-	-	-

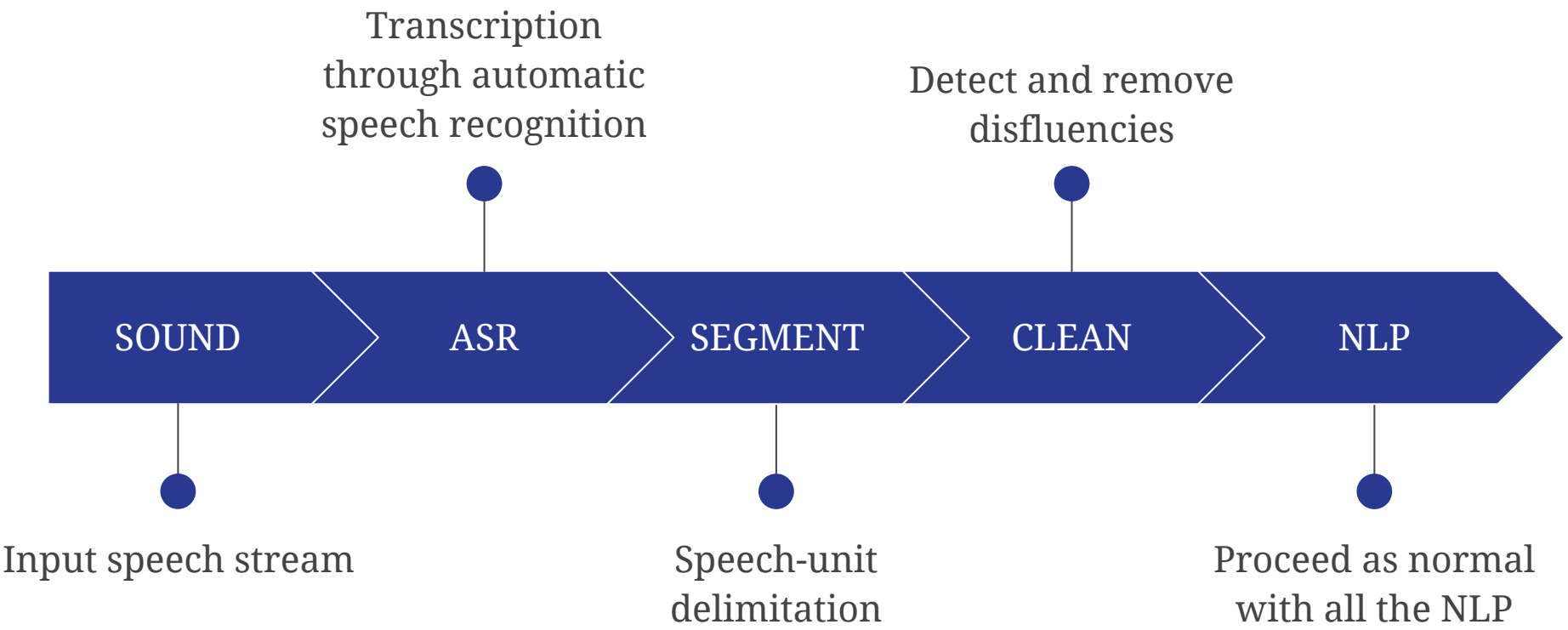
Table 3: Mean parse likelihoods, deltas to baseline and parse success rates in all transcription modes

A = ‘as is’
B = less disfluency
C = less morpho-syntactic error
D = less lexical error

What to do about speech

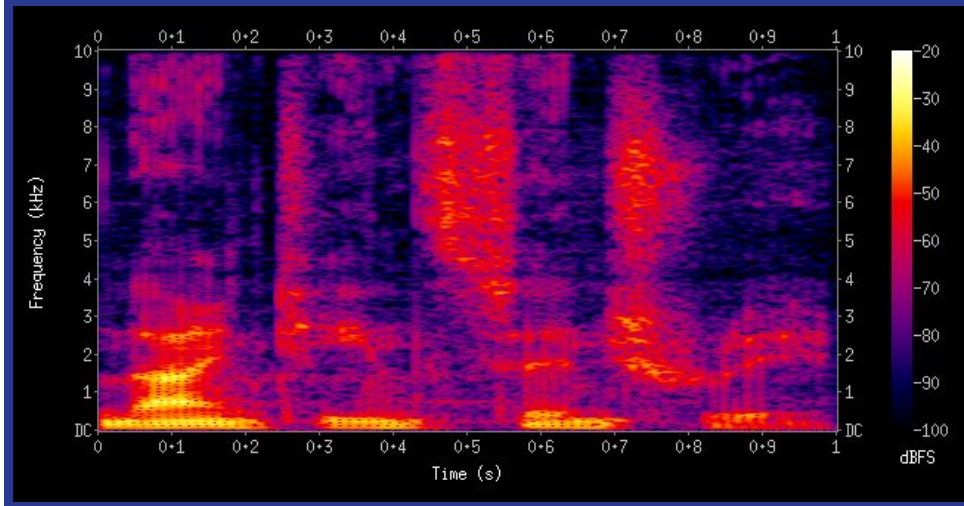
- Annotate more data
  - e.g. Switchboard, British National Corpus, CrowdED, ...
- Bring training and test data closer together: i.e. ‘normalisation’ of speech to written-like form
  - e.g. Moore et al 2015, 2016 https://aclanthology.info/papers/C16-1075/c16-1075
- Domain adaptation
  - e.g. Daumé III 2009, 2010 https://aclanthology.info/papers/W10-2608/w10-2608

The normalisation approach



Speech Processing

Paula Buttery



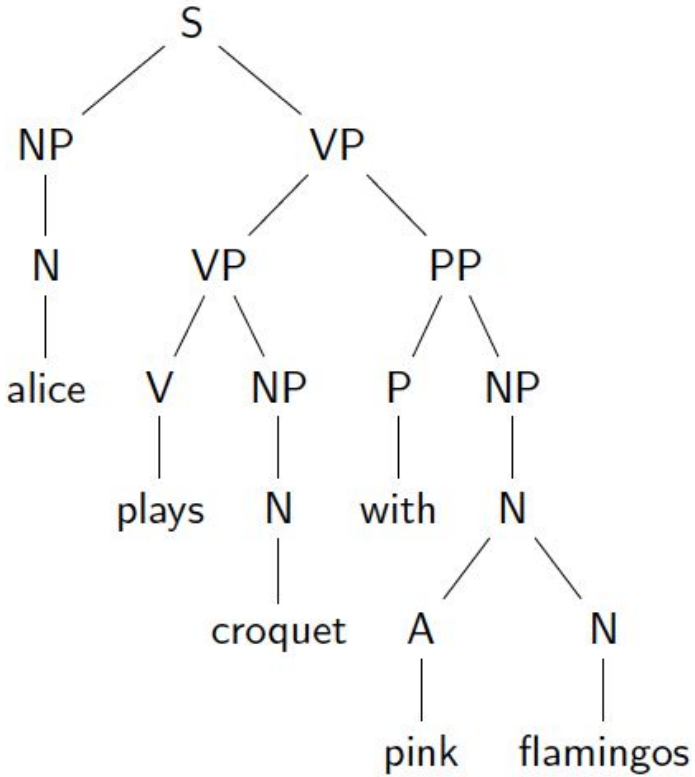
Source: https://commons.wikimedia.org/wiki/File:Spectrogram-19thC.png

Which grammar should we use?

Let’s consider grammars you’ve encountered:

- Phrase Structure Grammars
- Dependency Grammars
- Categorial Grammars
- Feature Structure Grammars

Phrase Structure Grammars



$G = (\mathcal{N}, \Sigma, S, \mathcal{P})$ where
$\mathcal{P} = \{A \to \alpha \mid$
$A \in \mathcal{N}, \alpha \in (\mathcal{N} \cup \Sigma)^*\}$

```
;;; Disable rules which deal with elliptical dialogue-like text as
;;; they tend to overapply elsewhere

(defparameter +disabled-rules+
    '(|V1/do_gap-r| |V1/have_gap-r| |V1/be_gap-r| |V1/mod_gap-r|
      |P1/prt-of| |P1/prt-r|
      ))

pjb48$ echo "I don't want to lecture now. You'll have to" | ./scripts/rasp.sh
(|I:0_PPIS1| |do:1_VD0| |not+:2_XX| |want:3_VV0| |to:4_TO| |lecture:5_VV0| |now:6_RT| |.:7_.|) 1 ; (-9.447)
gr-list: 1
(|ncsubj| |want:3_VV0| |I:0_PPIS1| _)
(|aux| |want:3_VV0| |do:1_VD0|)
(|ncmod| _ |want:3_VV0| |not+:2_XX|)
(|xcomp| |to| |want:3_VV0| |lecture:5_VV0|)
(|ncmod| |prt| |lecture:5_VV0| |now:6_RT|)



(|you:0_PPY| |will+:1_VM| |have:2_VH0| |to:3_TO|) 0 ; ()
gr-list: 1
```

With speech rules deactivated:

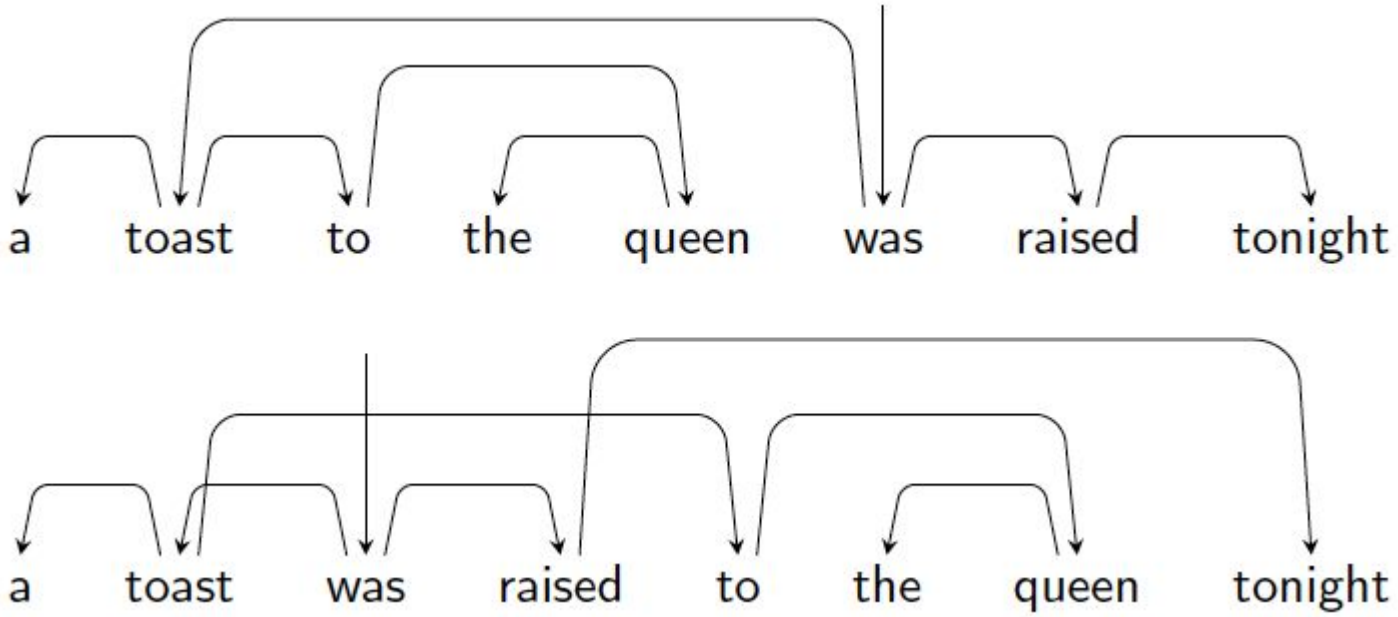
(|you:0_PPY| |will+:1_VM| |have:2_VH0| |to:3_TO|) 0 ; ()
gr-list: 1

With speech rules activated:

(|you:0_PPY| |will+:1_VM| |have:2_VH0| |to:3_TO|) 0 ; ()
gr-list: 1
(|ncsubj| |have:2_VH0| |you:0_PPY| _)
(|aux| |have:2_VH0| |will+:1_VM|)

(|T/frag|
 (|S/np_vp| |you:0_PPY|
  (|V1/modal_bse/--| |will+:1_VM| (|V1/have_gap-r| |have:2_VH0|)))
 |to:3_TO|)

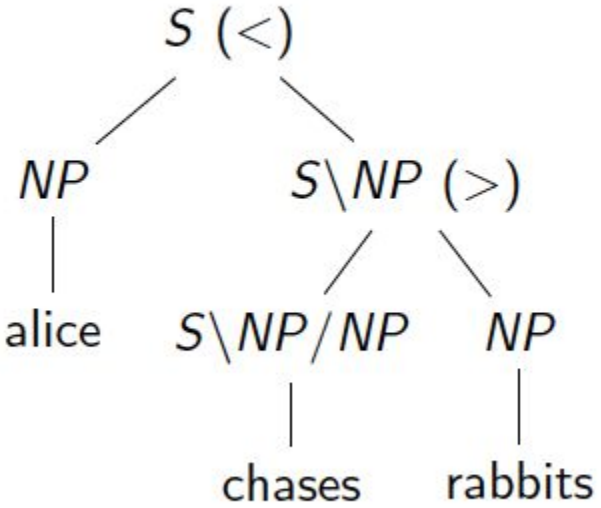
Dependency Grammars



Categorial Grammar

$$
\begin{aligned}
Pr &= \{S, NP\} \\
\Sigma &= \{alice, chases, rabbits\} \\
S &= S \\
\mathcal{R} &= \{(alice, NP), (chases, S\backslash NP/NP), \\
&\quad (rabbits, NP)\}
\end{aligned}
$$

$$
\dfrac{\dfrac{alice}{NP}\,\mathcal{R} \qquad \dfrac{\dfrac{chases}{S\backslash NP/NP}\,\mathcal{R} \qquad \dfrac{rabbits}{NP}\,\mathcal{R}}{S\backslash NP}\,>}{S}\,<
$$



Feature Structure Grammars

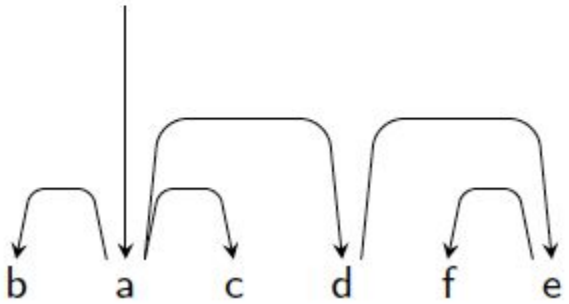
$$\left\langle \hat{}s, \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \mathbf{N} \\ \text{AGREEMENT} & pl \end{bmatrix} \end{bmatrix} \right\rangle$$

$$\left\langle \text{fox}, \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \mathbf{N} \\ \text{AGREEMENT} & [] \end{bmatrix} \end{bmatrix} \right\rangle$$

$$\begin{bmatrix} \text{HEAD} & \begin{bmatrix} \mathbf{N} \\ \text{AGREEMENT} & pl \end{bmatrix} \end{bmatrix} \sqcup \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \mathbf{N} \\ \text{AGREEMENT} & [] \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \mathbf{N} \\ \text{AGREEMENT} & pl \end{bmatrix} \end{bmatrix}$$

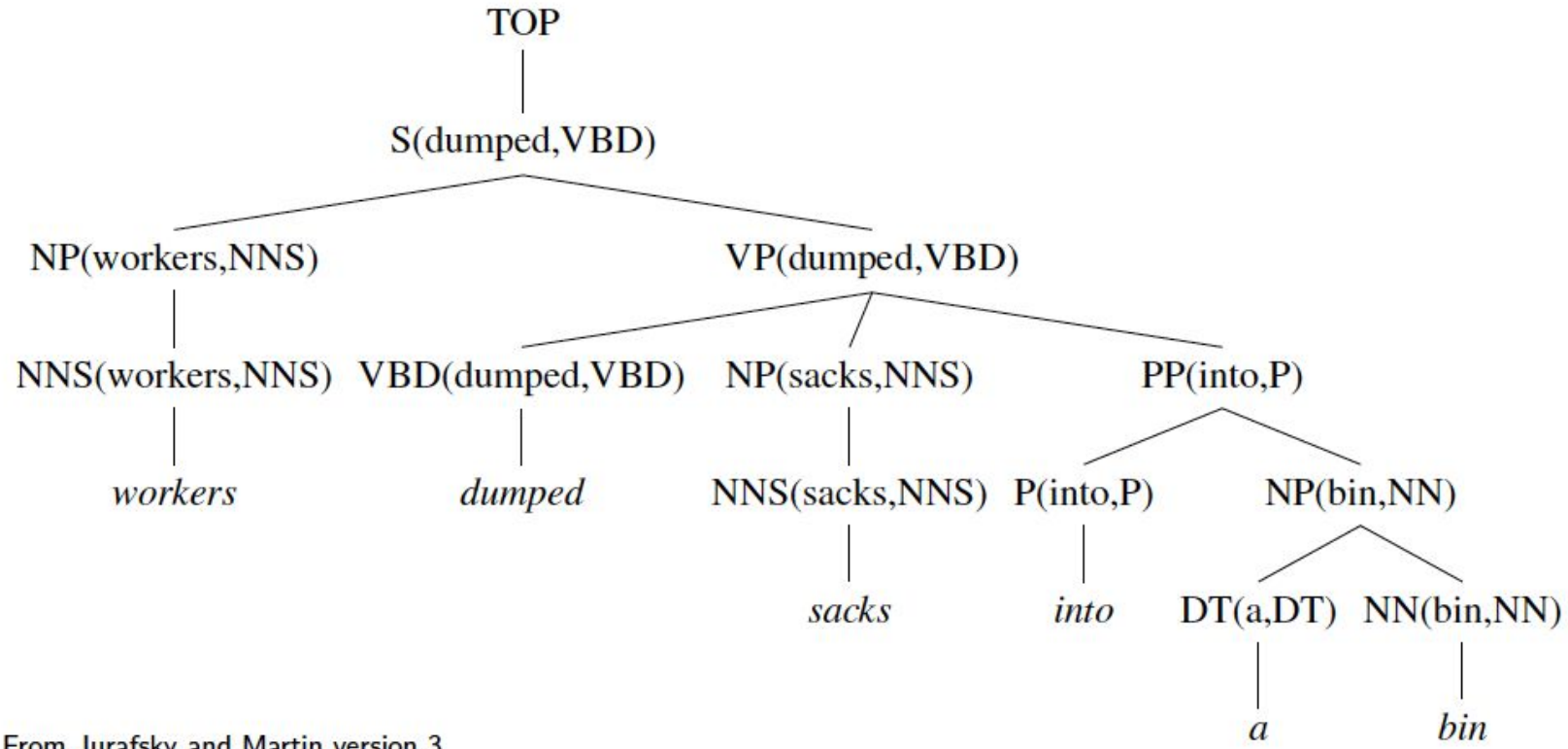
Parsing can be informed by extra linguistic info

$$
\begin{aligned}
\Sigma &= \{a...z\} \\
D &= \{\mathcal{L}, \mathcal{R}\} \\
s &= \text{s} \\
\mathcal{P} &= \{(a \rightarrow b, \mathcal{L} \mid c, \mathcal{R} \mid d, \mathcal{R}) \\
&\quad (d \rightarrow e, \mathcal{R}) \\
&\quad (e \rightarrow f, \mathcal{L})\}
\end{aligned}
$$



STACK	BUFFER	ACTION	RECORD
	bacdfe	SHIFT	
b	acdfe	SHIFT	
ba	cdfe	LEFT-ARC	$a \rightarrow b$
a	cdfe	SHIFT	
ac	dfe	RIGHT-ARC	$a \rightarrow c$
a	dfe	SHIFT	
ad	fe	SHIFT	
adf	e	SHIFT	
adfe		LEFT-ARC	$e \rightarrow f$
ade		RIGHT-ARC	$d \rightarrow e$
ad		RIGHT-ARC	$a \rightarrow d$
a		TERMINATE	$root \rightarrow a$

Parsing can be informed by features



From Jurafsky and Martin version 3

Parsing can be informed by extra linguistic info

1. produce a parse forest using simple version of the grammar
   i.e. find possible parses using coarse-grained non-terminals, e.g. *VP*
2. refine most promising of coarse-grained parses using complex grammar
   i.e with feature-based, lexicalised non-terminals, e.g. *VP*[*buys*/*VBZ*]

- **Coarse-grained step** can be **efficiently parsed** using e.g. CKY
- But the simple grammar **ignores contextual features** so best parse might not be accurate
- **Output a pruned packed parse** forest for the parses generated by the simple grammar (using a beam threshold)
- **Evaluate remaining parses with complex grammar** (i.e. each coarse-grained state is split into several fine-grained states)

Excerpt from the Spoken section of the British National Corpus

Set your sights realistically haven't you? And there's a lot of people unemployed. And what are you going to do when you eventually leave college? If you get there. You're not gonna step straight into television. Mm right then, let's see now what we're doing... Where's that recipe book for that chocolate and banana cake? Chocolate and banana cake which book was it? Oh right. Oh, some of these chocolate cakes are absolutely mm mm mm. Right, what's the topping? what's that? Icing sugar, cocoa powder and vanilla essence. Oh luckily I've got all those I think, yes!

Speech-unit delimitation

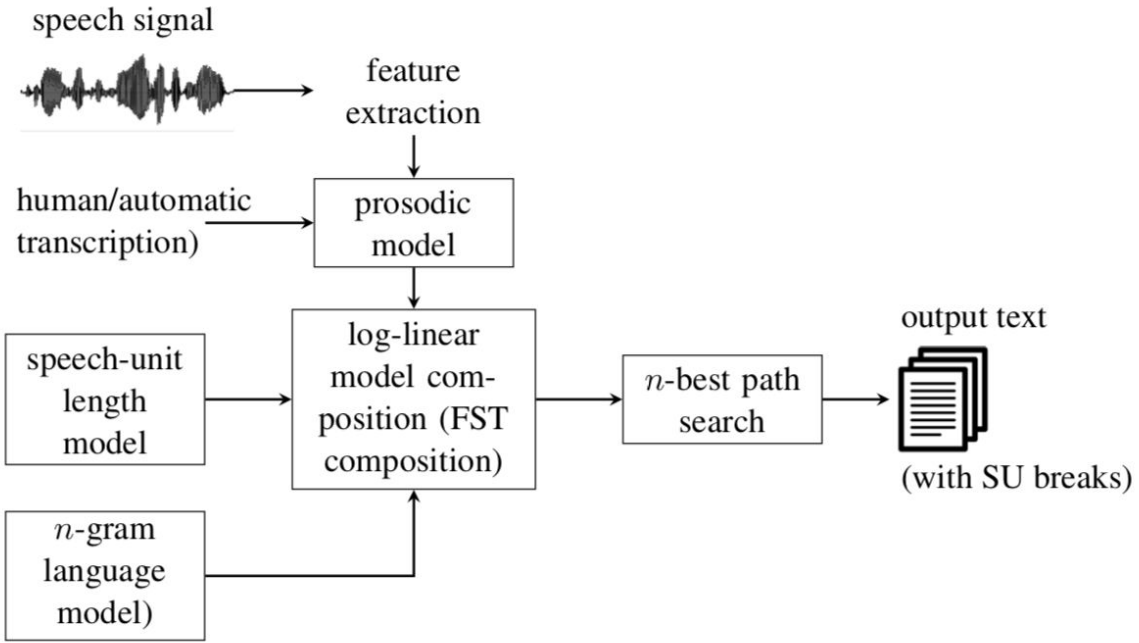


Figure 1: System diagram.

Speech-unit delimitation

category	features
pause duration	· pause before w_i · pause after w_i
phone duration	· final phone in w_i · sum of vowel phones in w_i · longest phone in w_i
f0	· max.f0 in w_i · min.f0 in w_i · max.f0 in w_{i+1} · min.f0 in w_{i+1} · end of w_i − start of w_{i+1} · end of w_i − recording min.f0 · mean of w_i − recording min.f0 · start of w_{i+1} − recording min.f0 · mean of w_{i+1} − recording min.f0
energy	· per f0

Table 1: List of prosodic features for the current word token (w_i).

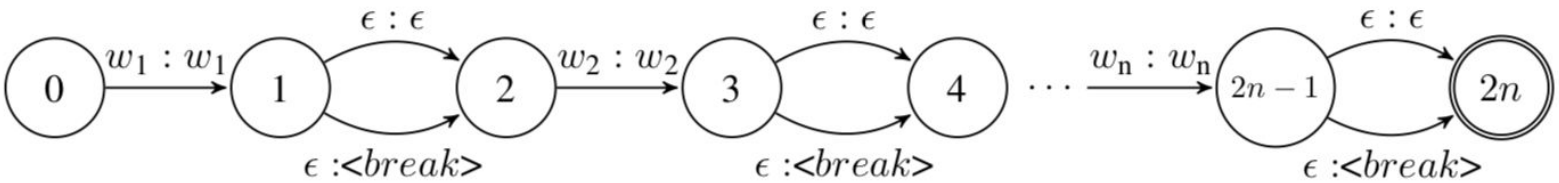


Figure 2: Prosodic model of an input string of length n.

Speech-unit delimitation

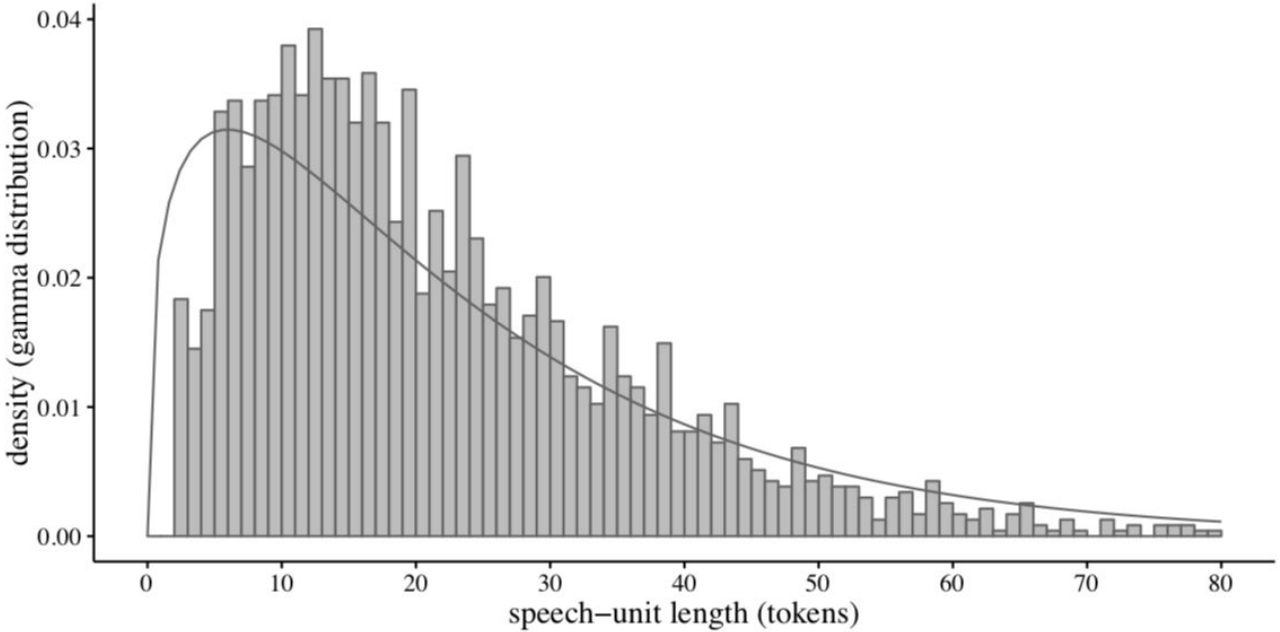


Figure 4: Density plot of speech-unit lengths in the BULATS learner corpus.

Speech-unit delimitation

BULATS PM	LM	SLM	BLEU-like	*p*	*r*	*F*
gold	.	.	.	.	.	.
(a) PM_{SVM}	BULATS	BULATS	0.51	0.409	0.582	0.48
(b) $PM_{r.6LR}$	BULATS	BULATS	0.75	0.64	0.5	0.56
(c) $2.5*PM_{r.6LR}$	BULATS	BULATS	0.75	0.639	0.617	0.628
(d) $5*PM_{r.6LR}$	BULATS	BULATS	0.74	0.653	0.686	0.669
(e) $5*PM_{r.6LR}$	CLC	BULATS	0.74	0.653	0.693	0.673
(f) $5*PM_{r.6LR}$	SWB	BULATS	0.74	0.656	0.693	0.674

Table 3: Speech-unit delimiter output evaluation (PM: prosodic model;