ELL101: Intro to Linguistics
Computational linguistics

Tomonori Nagano <tnagano@lagcc.cuny.edu>

Education and Language Acquisition Dept.
LaGuardia Community College

August 16, 2017

Fields of linguistics

- Week 1-2: Phonetics (physical sound properties)
- Week 2-3: Phonology (speech sound rules)
- Week 4: Morphology (word parts)
- Week 5-7: Syntax (structure)
- Week 8-9: Semantics (meaning)
- Week 8-9: Pragmatics (conversation & convention)
- Week 10: First & Second language acquisition
- Week 11-12: Historical linguistics (history of language)
- Week 11-12: Socio-linguistics (language in society)
- Week 11-12: Neuro-linguistics (the brain and language)
- Week 11-12: Computational linguistics

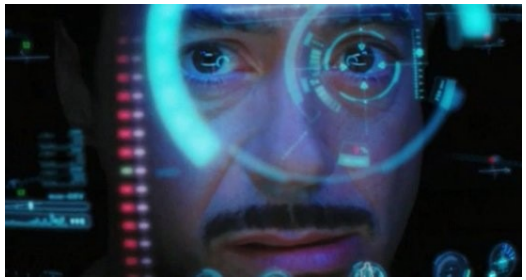
Introduction: Computational linguistics I

What is Computational Linguistics?

Simply put, computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be ”knowledge-based” (”hand-crafted”) or ”data-driven” (”statistical” or ”empirical”). (from The Association for Computational Linguistics http://www.aclweb.org)

Introduction: Computational linguistics II

- Goal: Making a computer program that fully understands human language (e.g., HAL in *2001: A space Odyssey* (1968), C3PO and R2D2 in *Star wars* (1977), KITT in *Knight Rider* (1982), Jarvis in *Iron Man* (2008))

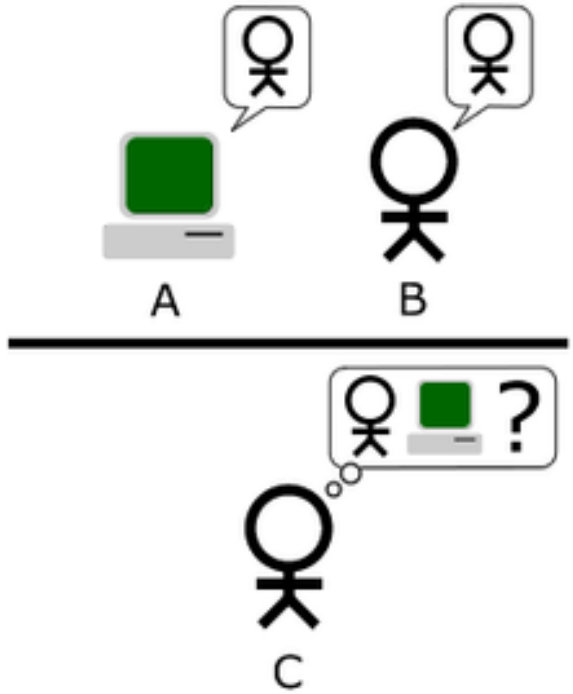


- We thought that we were close (in the 1950's)

Introduction: Computational linguistics III

Turing Test (Test for Artificial Intelligence)

In 1950 Alan Turing proposed that a machine could be termed ”intelligent” if it could respond to queries in a manner that was completely indistinguishable from a human being



Introduction: Computational linguistics IV

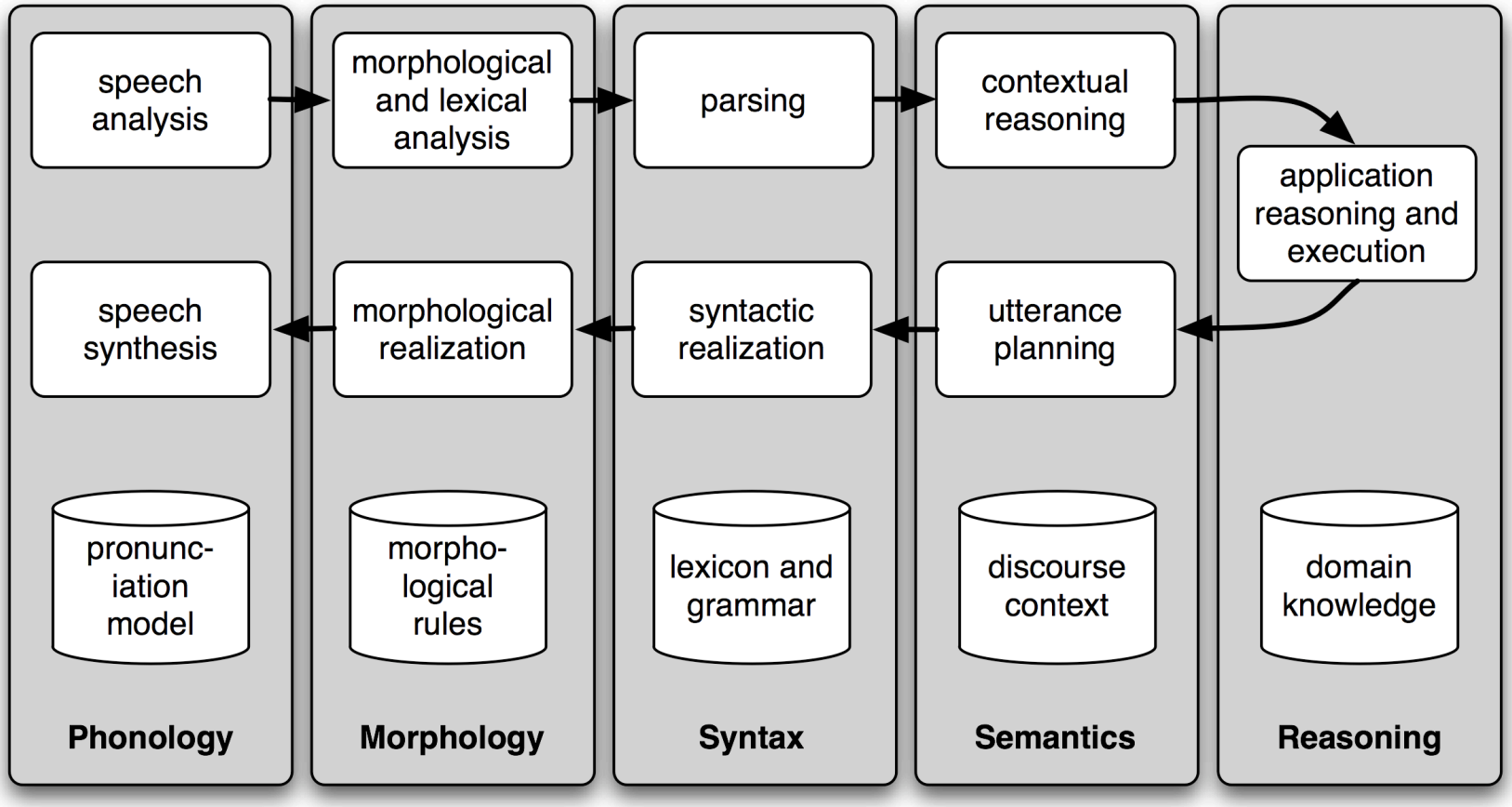
- Which ones are human utterances and which ones are machine-generated speeches?

- *Cottage cheese and chives are delicious.*

1. Human / Machine
2. Human / Machine
3. Human / Machine
4. Human / Machine
5. Human / Machine
6. Human / Machine
7. Human / Machine
8. Human / Machine
9. Human / Machine
10. Human / Machine
11. Human / Machine
12. Human / Machine

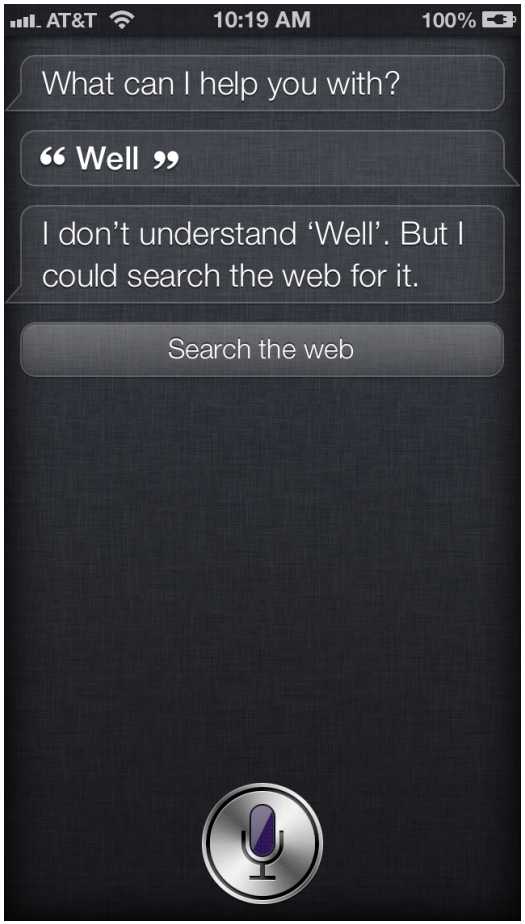
Introduction: Computational linguistics V

- Research in computational linguistics is divided into different sub-tasks



Introduction: Computational linguistics VI

- Even after 60 years, we are not quite there yet (whereas the other sci-fi dreams such as space trips are getting very close)



- Famous ”I don’t understand ____. But I could search the web for it” message from a confused Siri.
- Many hacking questions
  - ”Tell me a joke”
  - ”I need a baby”
  - ”Would you marry me?”
  - ”How much wood would a woodchuck chuck if a woodchuck could chuck wood?”
  - ”Open the pod bay doors, HAL”

Brief history of computational linguistics I

- **1940-1950: Language and probability (cryptography in the WW2)**
  - Alan Turing helps break the German "Enigma" code (cipher)

    $$P(X) = \frac{r*}{N}, \text{ where } r* = (r+1) \cdot \frac{E(N_{r+1})}{E(N_r)}$$

  - Claude E. Shannon's *information theory* (e.g., entropy) at Bell Lab

    $$H(X) = \sum_{i=1}^{n} P(x_i) I(x_i) = \sum_{i=1}^{n} P(x_i) \log_b \frac{1}{P(x_i)} = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

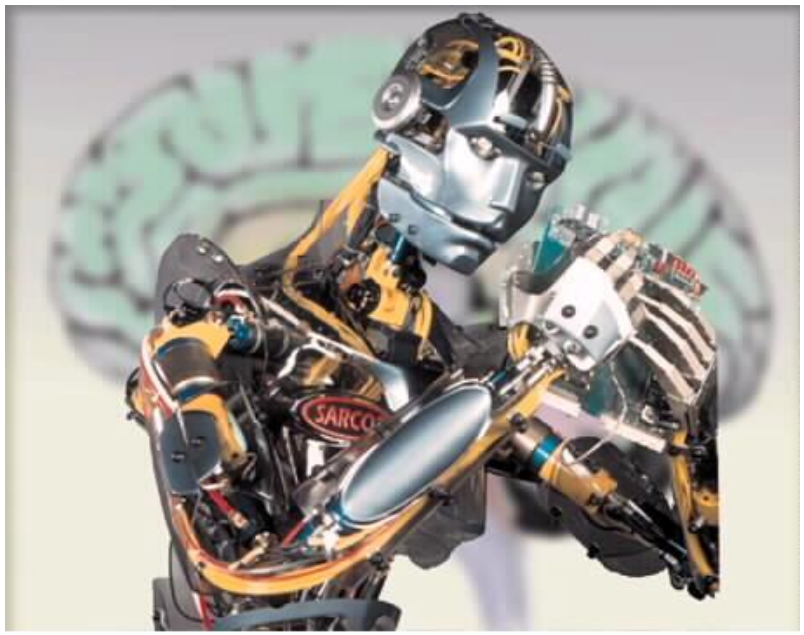
Brief history of computational linguistics II

- **1950s-1980s: Rationalist (rule-based linguists) breaks statistics**
  - Noam Chomsky's seminal book *Syntactic Structure* (1957) and his article *A Review of B. F. Skinner's Verbal Behavior*.
  - Neither sentence has ever occurred in the history of English. Any statistical model would given them the same probability (zero). However, any native speaker of English knows that one sounds far better than the other.
    - *Colorless green ideas sleep furiously.*
    - *Furiously sleep ideas green colorless.*

Brief history of computational linguistics III

- **1990s - present: The empiricists/statistical approaches strike back**
  - Advances in computational power
  - Increase in the database capacity (Google!)
  - Sophisticated algorithms

- We have yet been able to make any machine that understands and generates human language (even like a 3-year-old child).



Brief history of computational linguistics IV

- Two different approaches in computational linguistics (actually with totally opposite philosophy about the language)

Data-driven approach

- Heavy use of statistical analysis of data
- Mostly based on the machine learning techniques
- Little use of linguistic knowledge
- Robust in the noisy (actual) language data

Rule-based approach

- Heavy use of handcrafted rules
- Considerable manual work
- Reflecting evidence from linguistic research
- No sufficient robustness in the noisy (actual) language data

Challenges in computational linguistics I

Challenges

- **Machine is not fast enough**: Average people speak 10-15 phonemes per second and artificially sped-up speech (like on the radio) produces 40-50 phonemes per second. We can still understand those speech with fairly high accuracy.
- **Machine is not creative enough**: A human can generate an utterance that he or she has never heard before.
- **Machine is not flexible enough**: A human can understand a wide range of people who speak the same language (in spite of the regional accents)
- **Machine is not rational enough**: A human can make an immediate/online judgment as to which possible interpretation makes sense the most.

Challenges in computational linguistics II

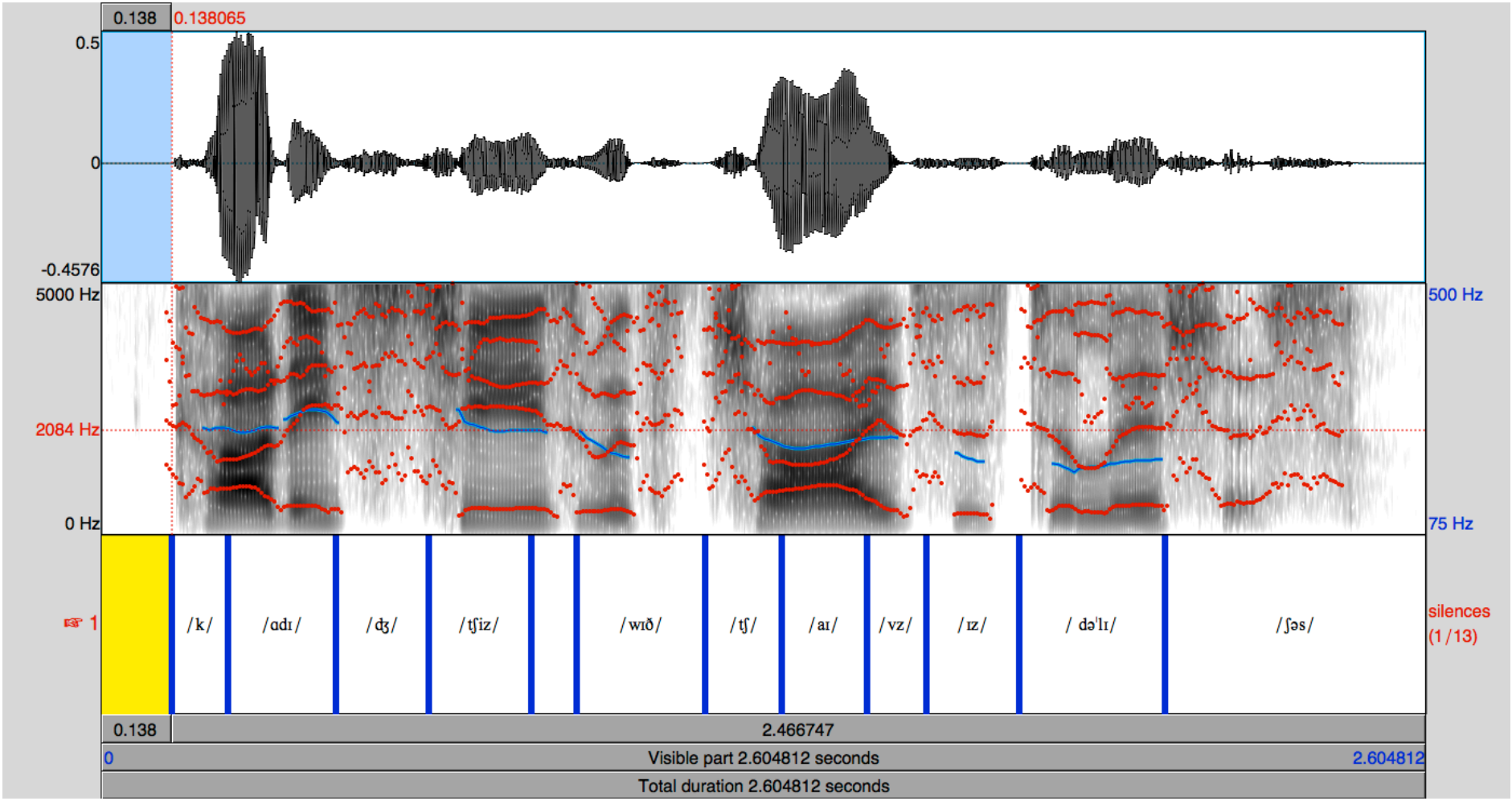
- Fast-talk commercials by John Moschitta (FedEx and Jetblue)

Challenges in computational linguistics III

- Let's assume a speaker of English with a vocabulary size of 20,000.
  - Assuming that words occur in any order (without any rule), the possible number of 5-word sentences is:
    $20,000 \times 20,000 \times 20,000 \times 20,000 \times 20,000 = 3.2 \times 10^{21}$
    (3200 trillion trillion possible sentences)
  - If it takes about one second to produce one sentence, we need 1.01×10^{14} years (one hundred trillion years) to say all the possible 5-word sentences.
- Almost all 5-word sentences you utter have never been used in the history of English. Let's use Google (undisputedly the largest language database to this date) to confirm.
  - "a pancake with fresh strawberries" (8 results)
  - "a pancake with fresh meatballs" (0 result)
  - "a stew with fresh meatballs" (0 result)
  - "a stew and fresh bananas" (0 result)

Challenges in computational linguistics IV

- Spectrogram is a visualization of a voice (a spectral density of a sound).



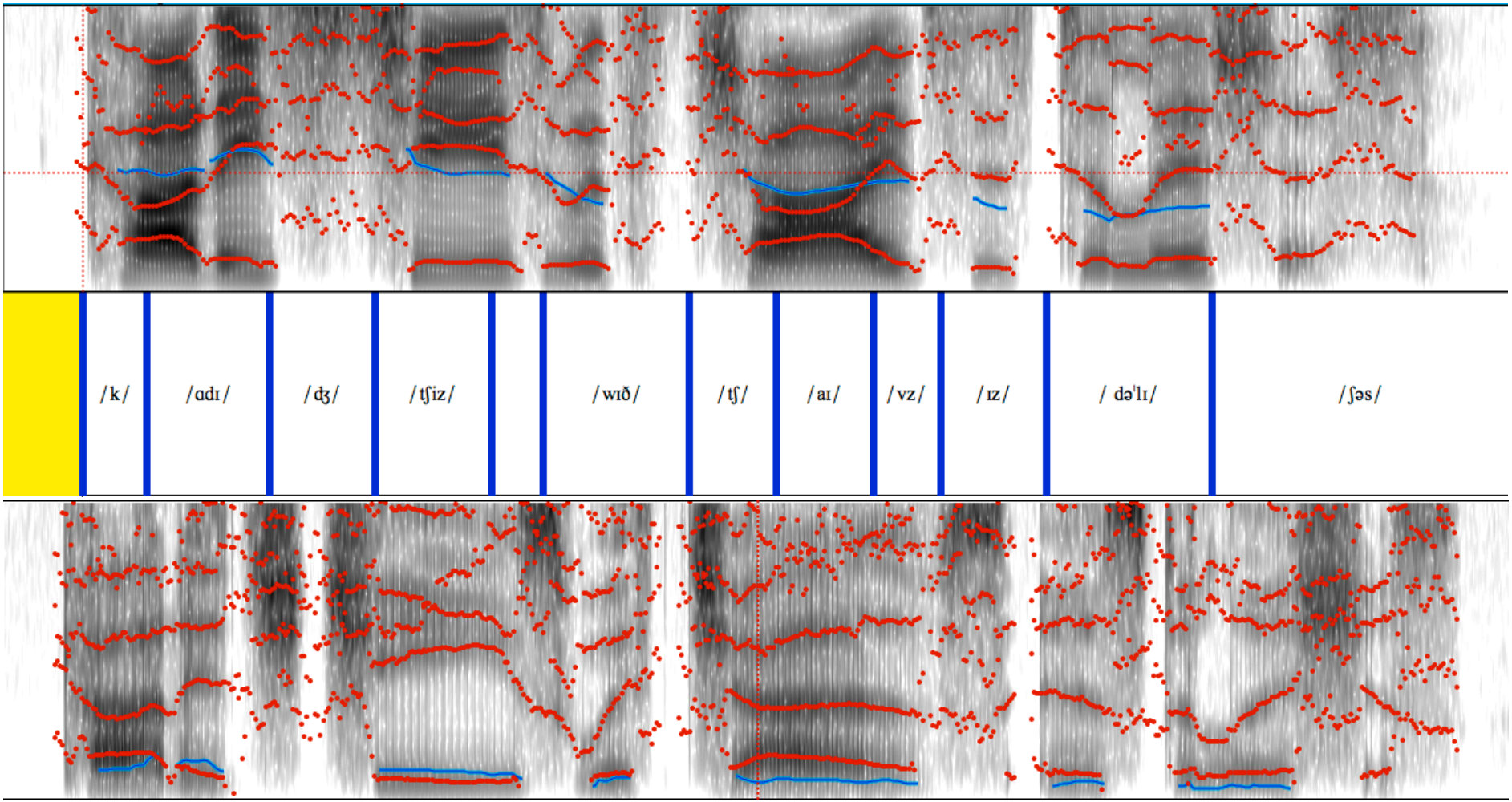
Challenges in computational linguistics V

- Sonorant (vowels, glides, liquids) show 4-5 formants (density peaks)
- Stop sounds modify the placement of formants in the surrounding vowels.
- Bilabial sounds cause a lowering of the formants
- Velar sounds show f_2 and f_3 coming together
- Alveolar sounds cause less systematic changes in formants

Vowel	Formant f1	Formant f2
u	320 Hz	800 Hz
o	500 Hz	1000 Hz
ɑ	700 Hz	1150 Hz
a	1000 Hz	1400 Hz
y	320 Hz	1650 Hz
ɛ	700 Hz	1800 Hz
e	500 Hz	2300 Hz
i	320 Hz	2500 Hz

Challenges in computational linguistics VI

- These two utterances are identical, but produced by different speakers.



Challenges in computational linguistics VII

- Ambiguity is an enemy of computational linguistics
  - *The chicken is ready to eat.*

Challenges in computational linguistics VIII

- Computers simply cannot appreciate ambiguity in the way human being does.
  - John enjoys painting his models nude.
  - Visiting relatives can be boring.
  - The men attacked the rioters with knives.
  - The girl hit the boy with a book.
  - He's a Tibetan history teacher.
  - We invited short men and women.
  - The chicken is ready to eat.
  - The killing of tyrants is justified.
  - Flying planes can be dangerous.
  - Painting mice can be entertaining.

Epilogue

- Computational linguists also kill time searching for online trivia.
  - Google: Google Gravity
  - Google: Epic Google
  - Google: Annoying Google
  - Google Translation: `ttchtt kkkkk, bsch. ttchtt kkkkk, bsch. ttchtt kkkkk, ttchtt kkkkk, ttchtt kkkkk, bsch. pv zk pv pv zk pv zk kz zk pv pv pv zk pv zk zk pzk pzk pvzkpkzvpvzk kkkkkk bsch`
  - http://www.funny-google.com

Bibliography I

Clark, A., Fox, C., and Lappin, S. (2010). *The Handbook of Computational Linguistics and Natural Langauage Processing*. Wiley-Blackwell.

Jurafsky, D. and Martin, J. (2006). *Speech and Language Processing*. Prentice Hall, xxx, 2nd edition.

Kilgarriff, A. (2005). Language is never, ever, ever random. *Corpus linguistics and linguistic theory*, 1:263–276.

Lee, L. (2004). "I'm sorry dave, I'm afraid i Can't do that": Linguistics, statistics, and natural language processing circa 2001. *Computer Science*, pages 111–118.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Mass.

Meyer, C. (2004). *English Corpus Linguistics An Introduction*. Cambridge University Press, Cambridge, Mass.

Mitkov, R. (2004). *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford, UK.

Sparck Jones, K. (2007). Computational linguistics: What about the linguistics. *Computational Linguistics*, 33(3):437–441.

Sproat, R., Samuelsson, C., Chu-Carroll, J., and Carpenter, B. (2002). Computational linguistics. In Aronoff, M. and Rees-Miller, J., editors, *The Handbook of Linguistics*, chapter 25, pages 609–636. Blackwell Publishers, Malden, Mass.

This presentation slide was created with LaTeX and *beamer* LaTeX style.