Introduction to Computational Semantics and its Applications

Ekaterina Kochmar

Computer Laboratory, University of Cambridge

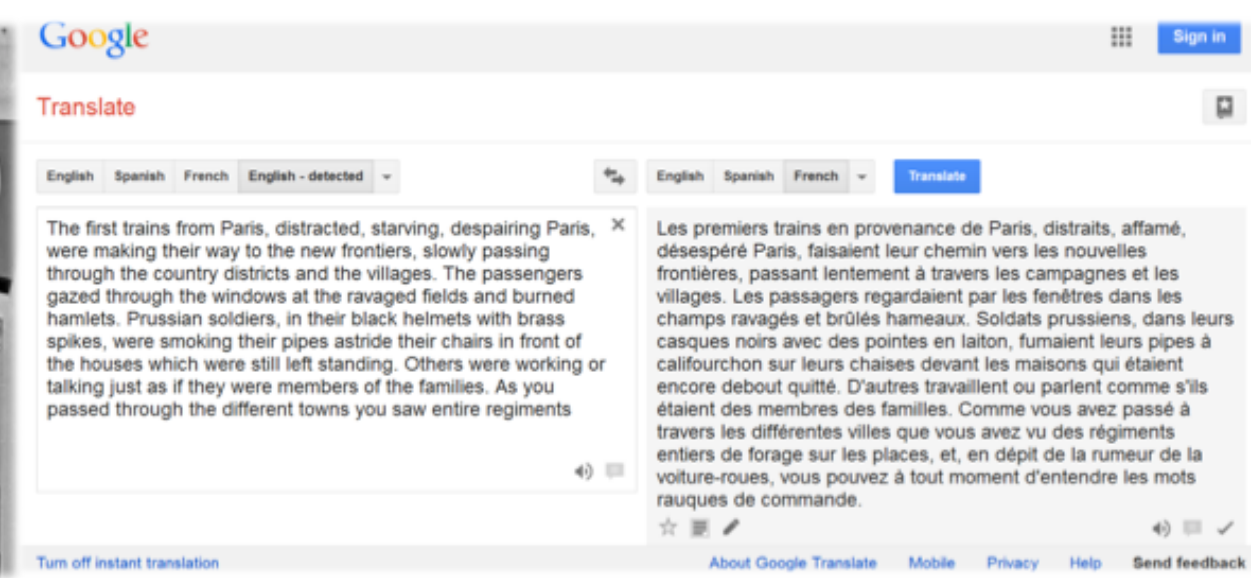
Automated Language Teaching and Assessment (ALTA) Institute

UCL, March 2018

Computational Linguistics

Introduction to Computational Linguistics:
a bit of history

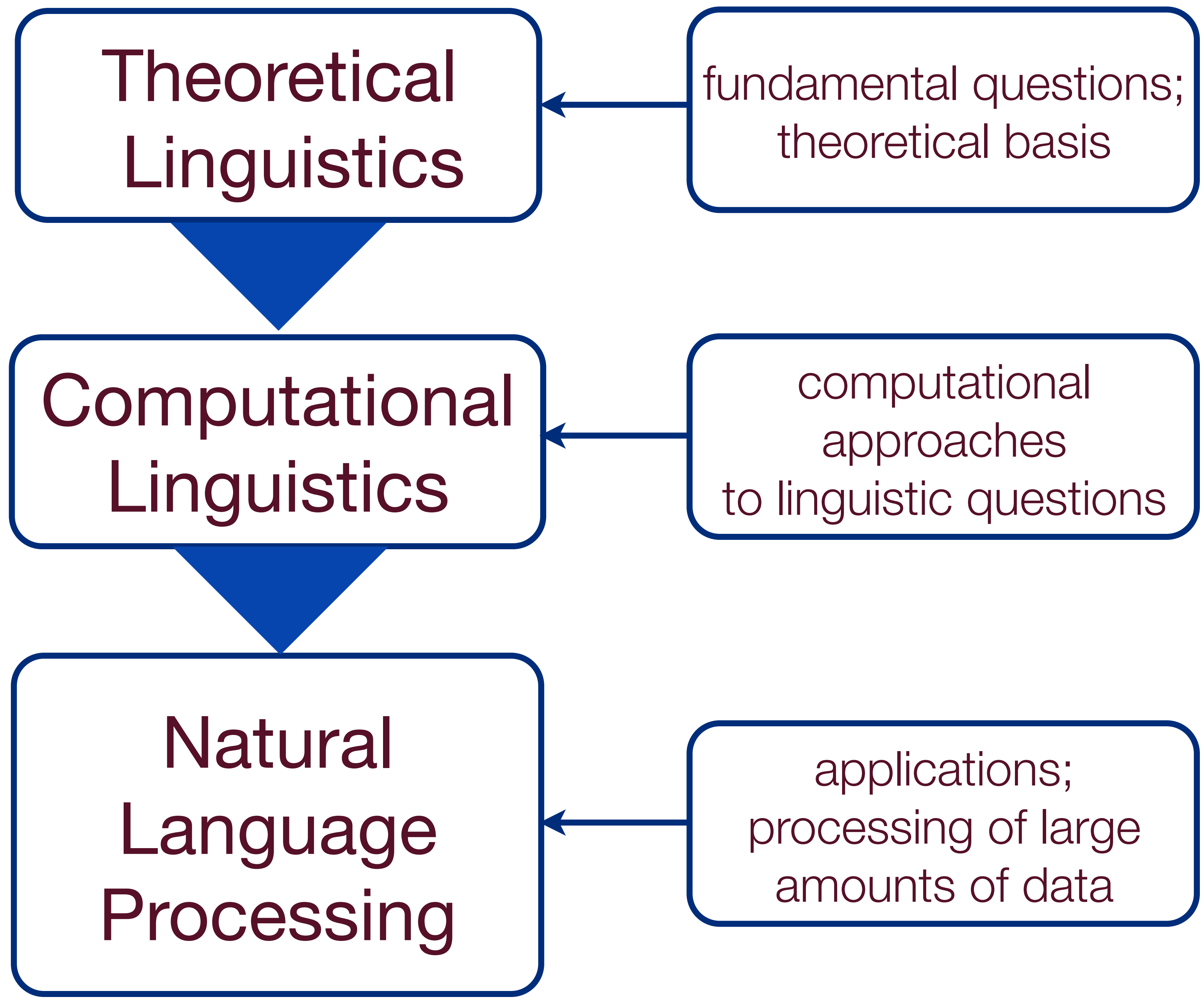
- Computational Linguistics originated in the U.S. in 1950s
- Focused on Machine Translation, particularly from Russian to English
- Deemed to be an easy computational task
- Note: this task is not perfectly solved even today...



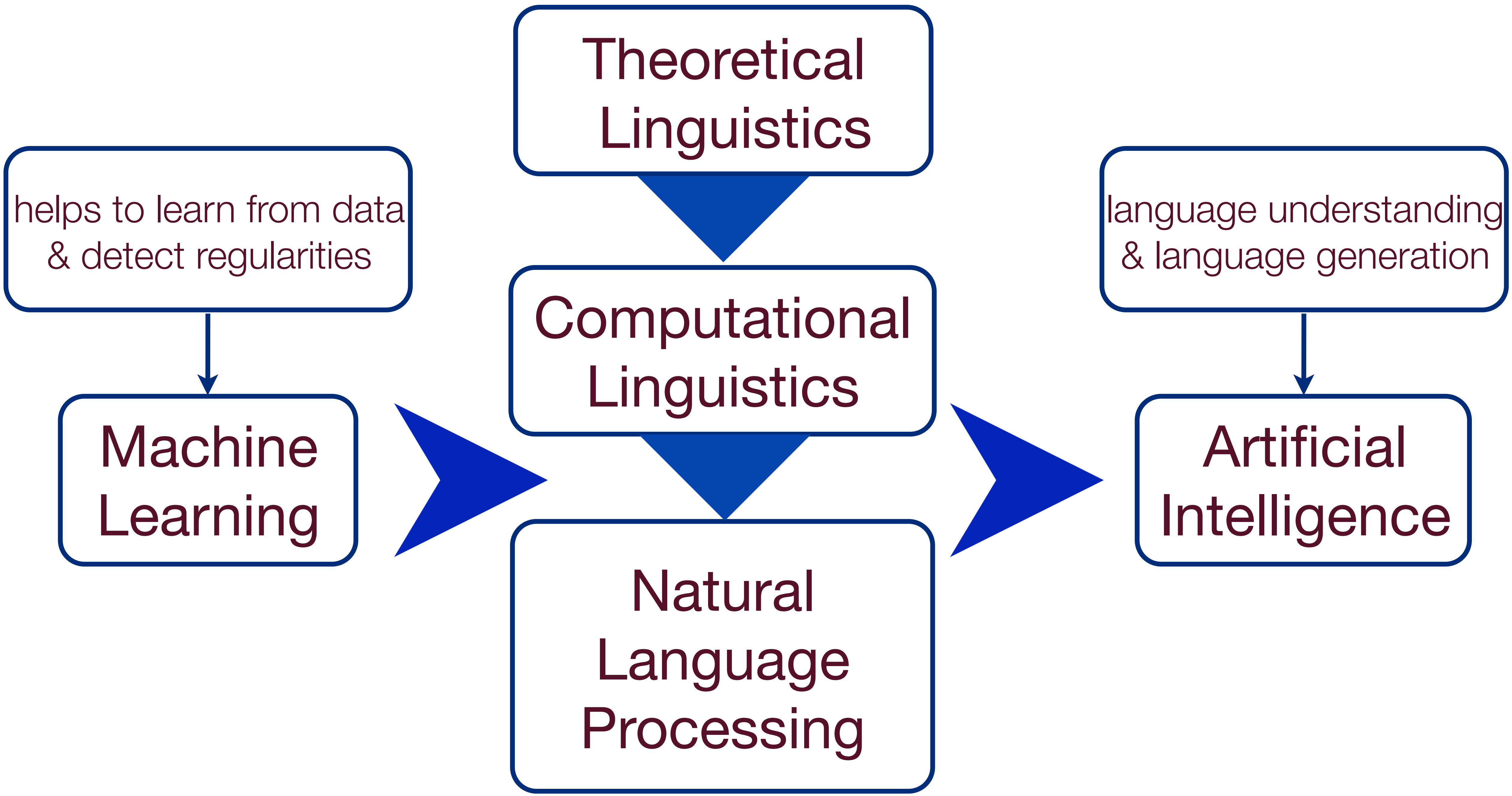
Cambridge ALTA

UNIVERSITY OF CAMBRIDGE

Computational Linguistics
and other fields



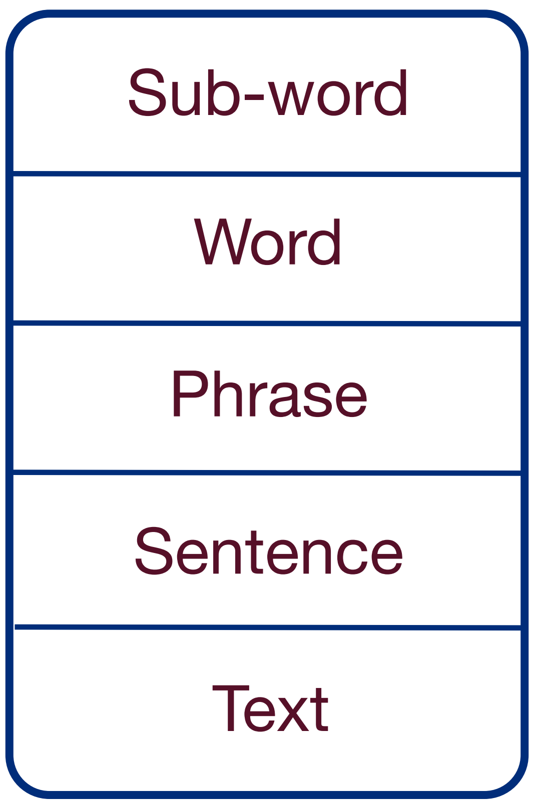
Computational Linguistics
and other fields



Computational Linguistics vs **Theoretical Linguistics**

Theoretical Linguistics	Computational Linguistics
✦ develops linguistic theory	✦ builds computational models
✦ seeks to answer fundamental questions	✦ seeks to confirm and test fundamental approaches
✦ is based on theoretical approaches	✦ rule-based or statistical, data-driven approaches
✦ theory-oriented	✦ application-oriented

Computational & Theoretical Linguistics: **Fields & Tasks**



- ✦ **Phonology/phonetics** –> speech processing, speech recognition
- ✦ **Morphology** –> morphological analysis, stemming, lemmatisation
- ✦ Word level: word segmentation, part-of-speech tagging, language modelling
- ✦ **Syntax** –> parsing
- ✦ **Semantics** –> lexical and computational
- ✦ **Discourse and pragmatics** –> discourse analysis

Fields & Tasks:
Speech processing

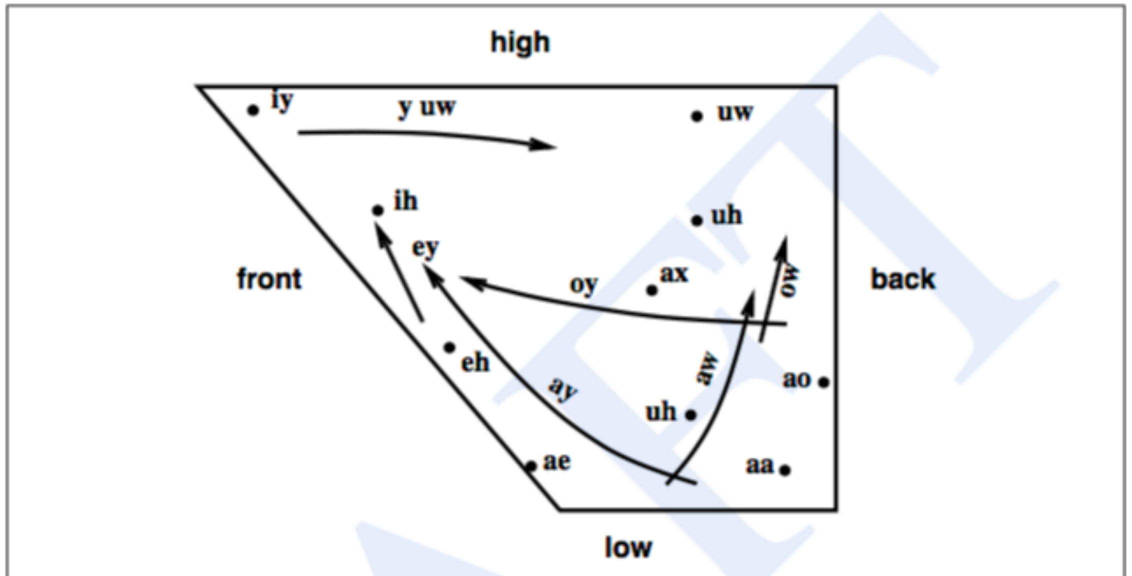


Figure 7.6 Qualities of English vowels (after Ladefoged (1993)).

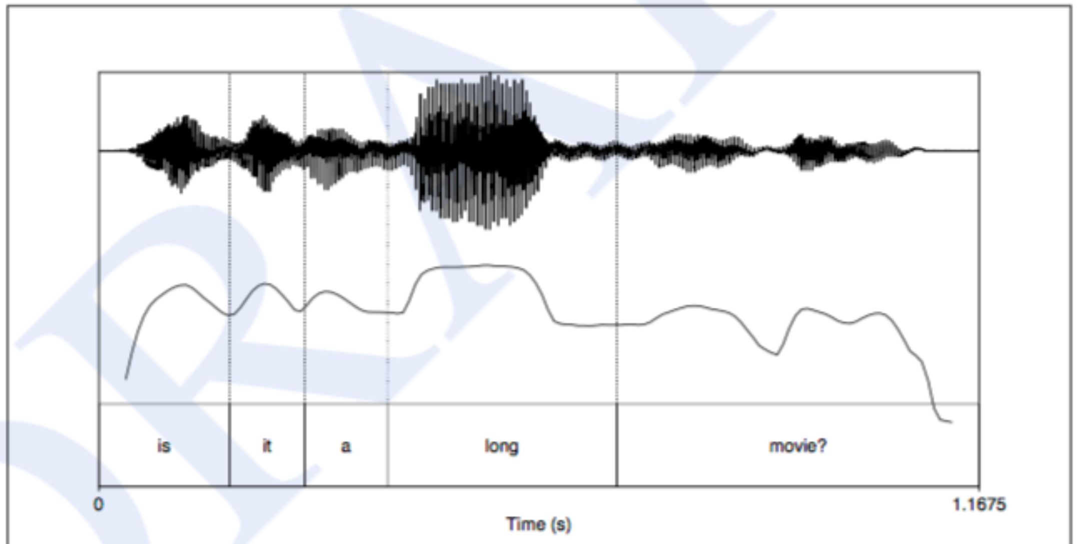


Figure 7.16 Intensity plot for the sentence "Is it a long movie?". Note the intensity peaks at each vowel, and the especially high peak for the word *long*.

- **Speech analysis**: based on what we know about phonetics and phonology, can we recognise speech, i.e. transcribe the audio signal as text?
- **Speech synthesis**: Can we generate the speech signal based on text?

** Here and on the other slides: the mages are adopted from* ***Jurafsky and Martin. Speech and Language Processing.*** *Second edition. 2009*

Fields & Tasks:
Text segmentation & normalisation

- What is a basic linguistic unit? –> **Word?**
  - Is ‘***U.S.’*** one word?
  - Is ‘***theory-based’*** one word?
  - Is ***‘.’*** part of the word as in ***‘Mr.’***?
  - What about ***‘;)’***?
- The notion of a word depends on language:
  - *FR* ***“l’ensemble”***
  - *GER* ***“Lebensversicherungsgesellschaftsangestellter”*** =
    Lebens-versicherungs-gesellschafts-angestellter =
    ‘ life insurance company employee ’

** Here and on the other slides: the mages are adopted from* ***Jurafsky and Martin. Speech and Language Processing.*** *Second edition. 2009*

Fields & Tasks:
Text segmentation & normalisation

- Chinese – no spaces to separate words
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Sharapova now lives in US southeastern Florida

- Japanese – many alphabets mixed

フォーチュン500社は情報不足のため時間あた$500K(約6,000万円)

Katakana Hiragana Kanji Romaji

* Here and on the other slides: the mages are adopted from ***Jurafsky and Martin. Speech and Language Processing.*** *Second edition. 2009*

Fields & Tasks: **Morphology**

- Words are built of smaller units – **morphemes**
- **Morphology**: *inflectional* (to express grammatical category) and *derivational* (to change the lexical category in related words)
- Richness of plural form morphology in English:
  - word -> word s, book –> book s
  - fox –> fox es, hero –> hero es
  - ax –> ax es and axe s <– axe
  - city –> citi es, morphology –> morphologi es
  - leaf –> leav es, shelf –> shelv es
  - foot –> feet, man –> men, mouse –> mice
  - corpus –> corpora, phenomenon –> phenomena

Fields & Tasks: **Morphology**

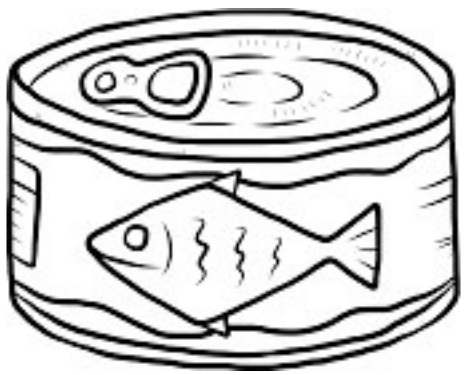
- Richness of morphological forms in many other languages is higher:
  - cf. Turkish: ***Uygarlastiramadiklarimizdanmissinizcasina*** –
    '(behaving) as if you are among those whom we could not civilise' =
    Uygar - las - tir - ama - dik - lar - imiz - dan - mis -siniz-casina =
    'civilised'-'become'-'cause'-'not able'-'past'-'plural'-'p1pl'-'abl'-'past'-'2pl'-'as if'

- With the computational models we want to recognise:
  - *book* and *books* – {*book*}; *is*, *are*, *was*, *been* – {*be*} –> **lemmatisation**
  - *automate*, *automation*, *automated*, *automatic* – {automat} –> **stemming**

** Here and on the other slides: the mages are adopted from* ***Jurafsky and Martin. Speech and Language Processing.*** *Second edition. 2009*

Fields & Tasks:
Sequence labelling and modelling

- **Part-of-speech tagging**
  - We can fish
PRON AUX VB   vs   We can fish
PRON VBP NOUN



- **Language modelling**:
  - lectu__
  - Today's lecture will take ___

Fields & Tasks: **Syntax**

I saw a man in the park [with a telescope]

I saw a man in the park [with a telescope]

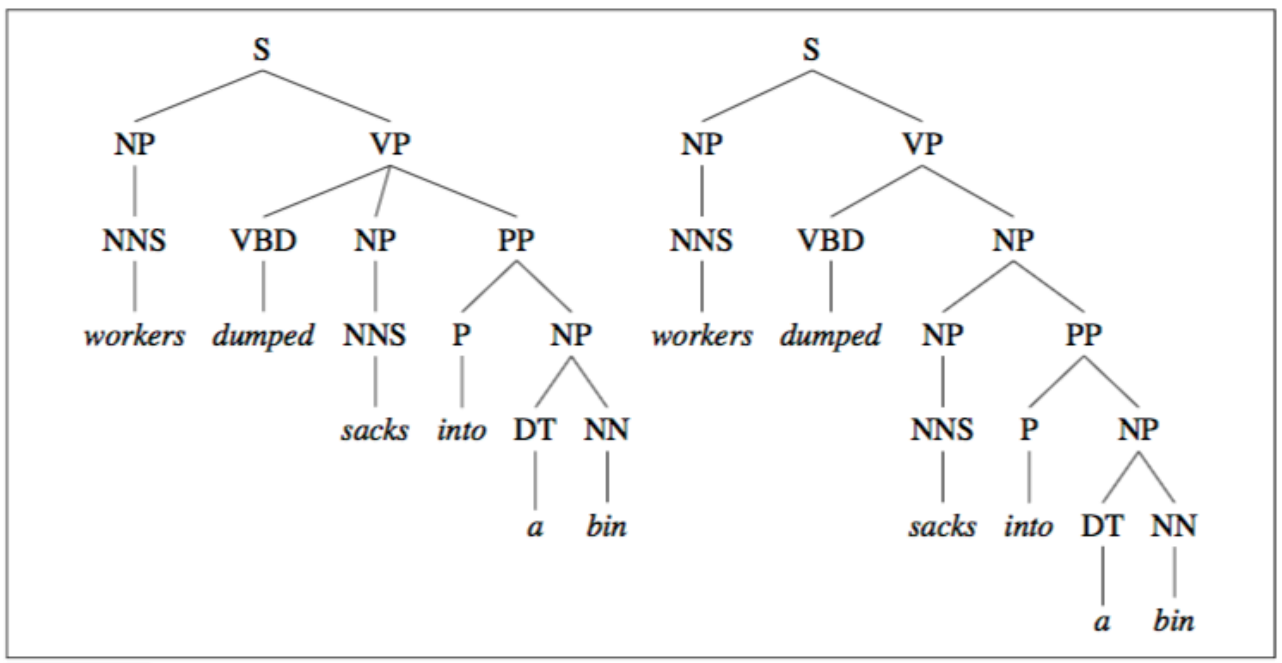


Figure 13.5 Two possible parse trees for a **prepositional phrase attachment ambiguity**. The left parse is the sensible one, in which "into a bin" describes the resulting location of the sacks. In the right incorrect parse, the sacks to be dumped are the ones which are already "into a bin", whatever that might mean.

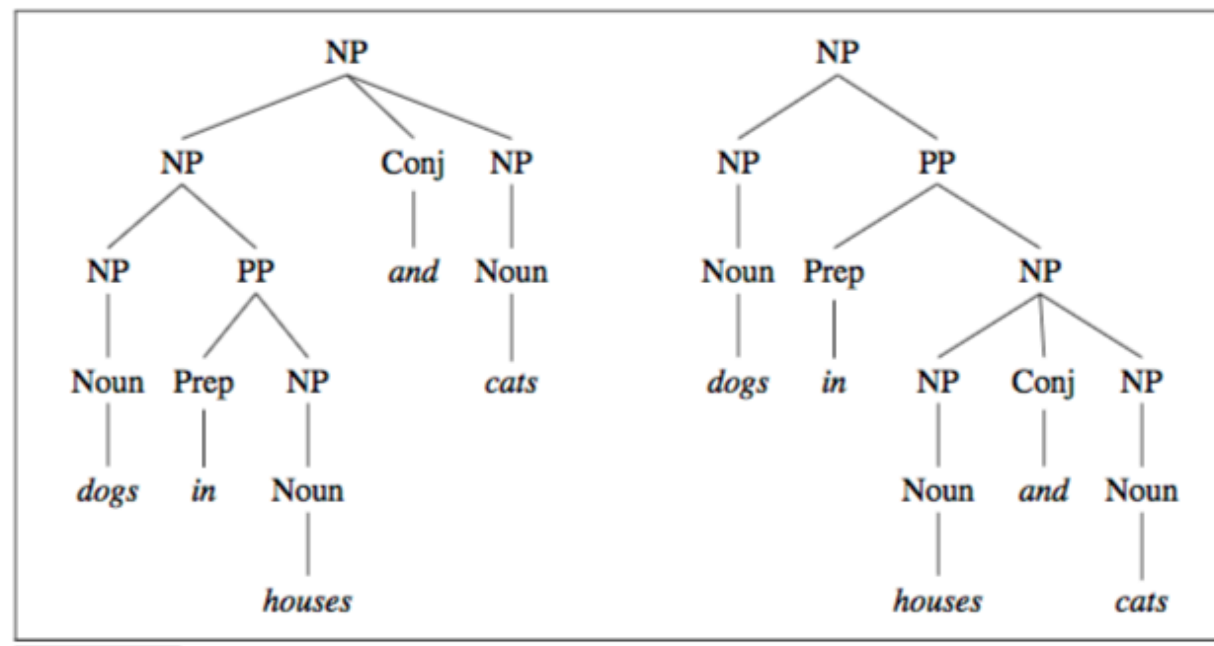


Figure 13.7 An instance of coordination ambiguity. Although the left structure is intuitively the correct one, a PCFG will assign them identical probabilities since both structures use exactly the same set of rules. After Collins (1999).

** Here and on the other slides: the mages are adopted from* ***Jurafsky and Martin. Speech and Language Processing.*** *Second edition. 2009*

Fields & Tasks:
Semantics

- **Lexical Semantics**: word senses and relations between word senses
    - *I went to the <u>bank</u> and withdrew money from my account*
    - *I went to the <u>bank</u> and had a walk along the river*

- **Computational Semantics** (Vector Semantics): representation of word (and larger linguistic units) meaning in a shared semantic space

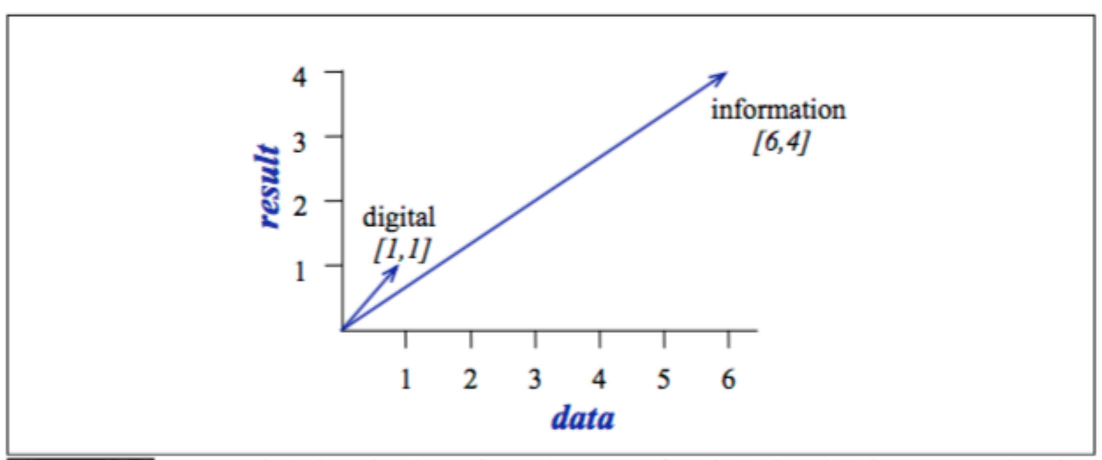


Figure 15.5 A spatial visualization of word vectors for *digital* and *information*, showing just two of the dimensions, corresponding to the words *data* and *result*.

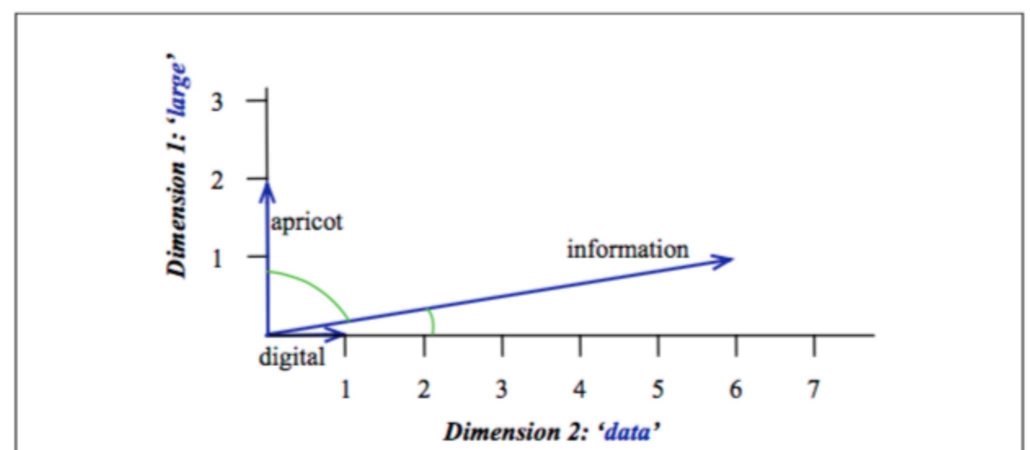


Figure 15.10 A graphical demonstration of the cosine measure of similarity, showing vectors for three words (*apricot*, *digital*, and *information*) in the two dimensional space defined

* *Here and on the other slides: the mages are adopted from **Jurafsky and Martin. Speech and Language Processing.** Second edition. 2009*

Computational Semantics

Computational Semantics

- ✦ Our goal is to build a **computational model of word meaning** so that a machine can understand the words, derive the meaning of phrases and detect the anomalies

- ✦ Luckily, there are **compositional distributional (as well as distributed) semantic models** that can help us:

  - **distributional/distributed models** helps capturing individual words' meaning

  - **compositional semantic models** help successfully (or unsuccessfully) combine the individual meanings into the meaning of a longer phrase

- *Mikolov et al. (2013)* showed that computers can reason about word meaning similarly to humans using an example of word analogy:

Man is to ***woman*** as

king is to ____ ?

- *Mikolov et al. (2013)* showed that computers can reason about word meaning similarly to humans using an example of word analogy:

Man is to ***woman*** as

king is to ***queen*** ?

- What the solution boils down to is:

MEANING(***WORD***) = MEANING(***king***) - MEANING(***man***) + MEANING(***woman***)

- How do we know what words mean?

Who is a queen?

Computational Semantics:
Learning through experience

Learning through experience

- We **read** about *kings* and *queens*
- We **hear** about them on the news
- We **see** them on the TV or, perhaps, even in person
- => We build our semantic model of what the words *king* and *queen* mean based on our experience
- How can a machine learn the meaning of a word?

Key assumptions of distributional semantics

- **Key assumption**: word meaning can be approximated by a word's distribution

"You shall know a word by the company it keeps" (Firth)

- **Method**: represent words with distributional vectors, dimensions = co-occurrence with a predefined set of context words

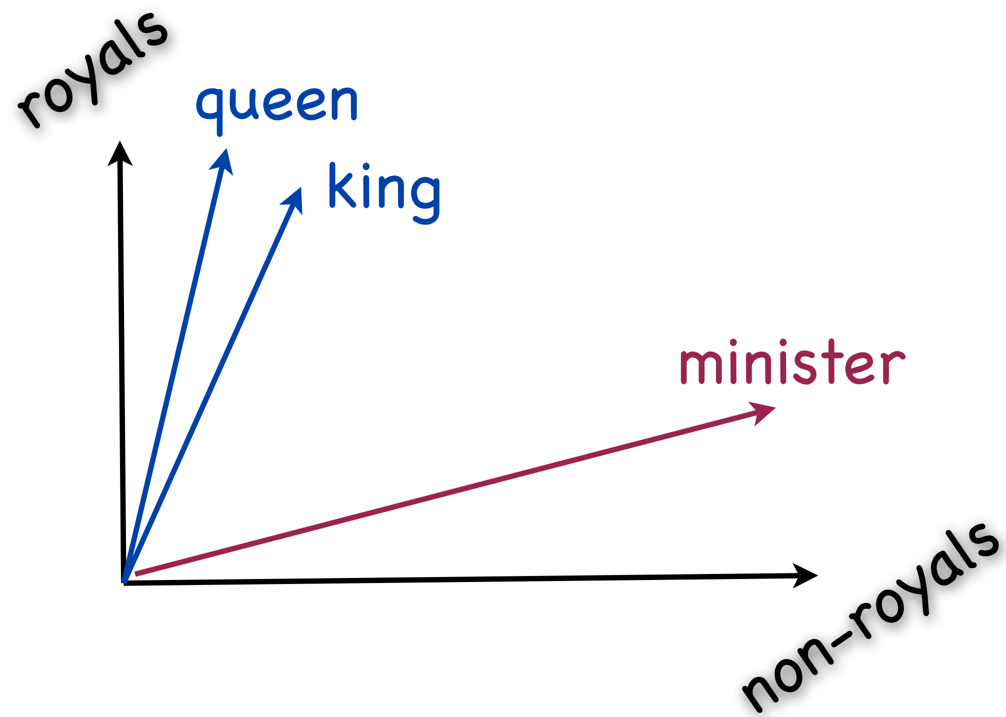
- **Hypothesis**: semantically similar words occur in similar contexts and, therefore, will be represented with similar vectors in the semantic space

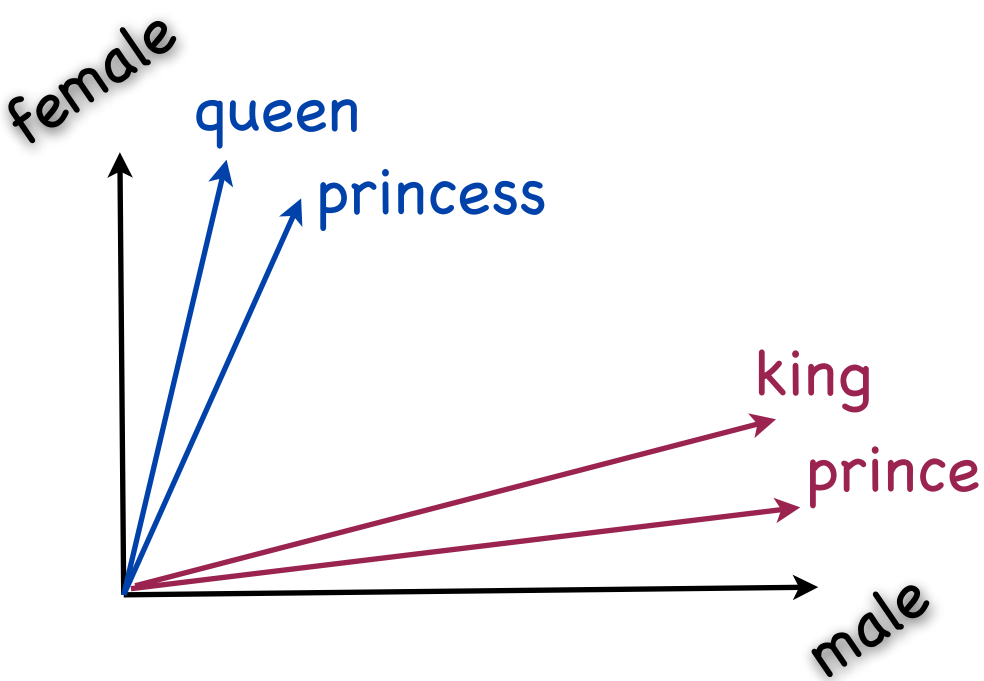
- A nice property of a direct interpretation of word meaning through vectors in space

Her Majesty the Queen
The Queen's speech during the State Visit to...
Buckingham Palace is the Queen's official London residence...
The Crown of Queen Elizabeth
The Queen Mother

	he	*she*	*royal*
queen	20	581	389
king	599	18	344

Computational Semantics:
Distributional Semantic Models





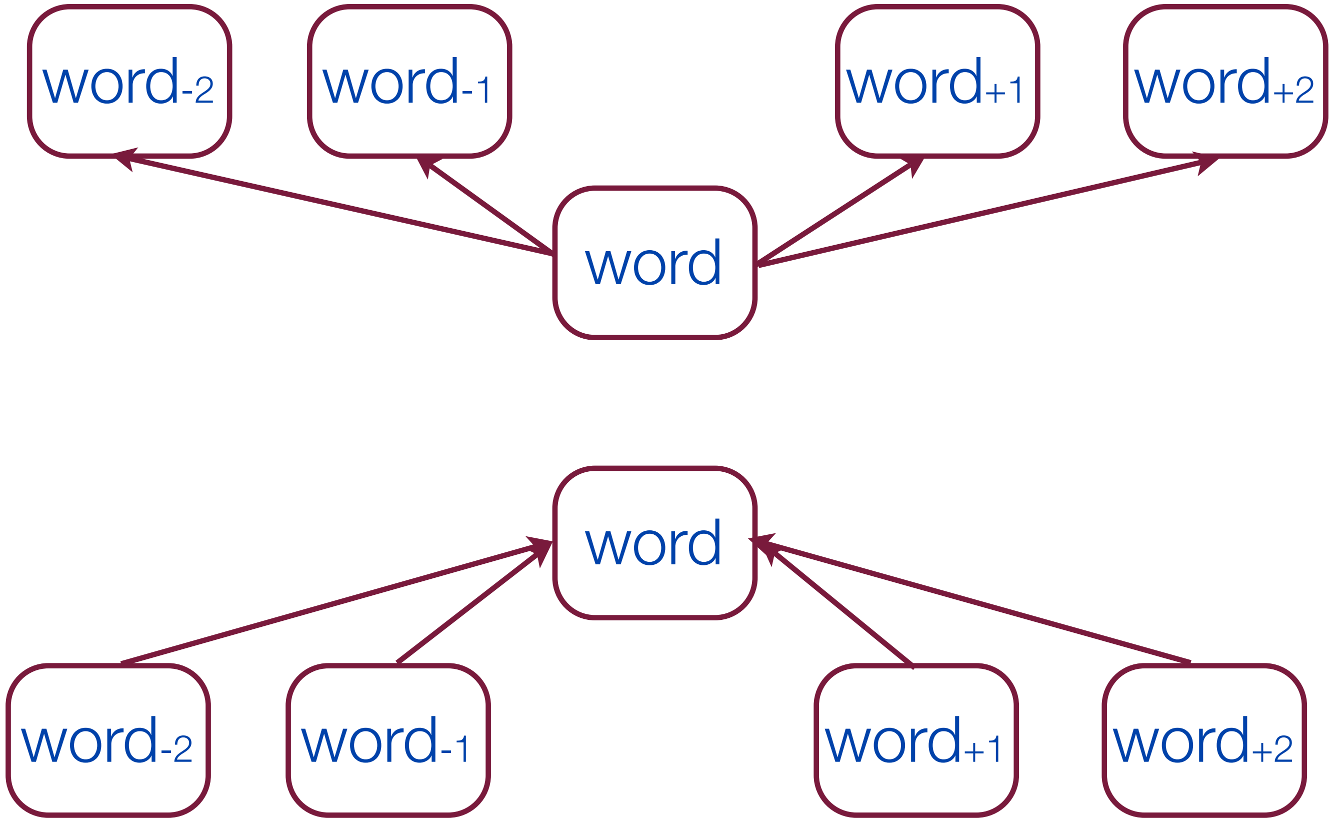
Represent words as vectors
How should we build them?
What are the dimensions?

Learn from the data
Build vectors using the surrounding words
-> Distributional models of word meaning

Word meaning representations

- **Distributional models**: build word vectors using contexts

- **Distributed models** (word embeddings): dense low-dimensional (300) representations where each dimension encodes some distinct property



Word meaning representations

- **Distributional models**: build word vectors using contexts

- **Distributed models** (word embeddings): dense low-dimensional (300) representations where each dimension encodes some distinct property

- Essentially: different ways to build **word vectors**

- A bit of math:
  - How to measure semantic similarity? Use cosine (distance) measure

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Word meaning interpretation

- *Mikolov et al. (2013)* showed that computers can reason about word meaning similarly to humans using an example of word analogy:

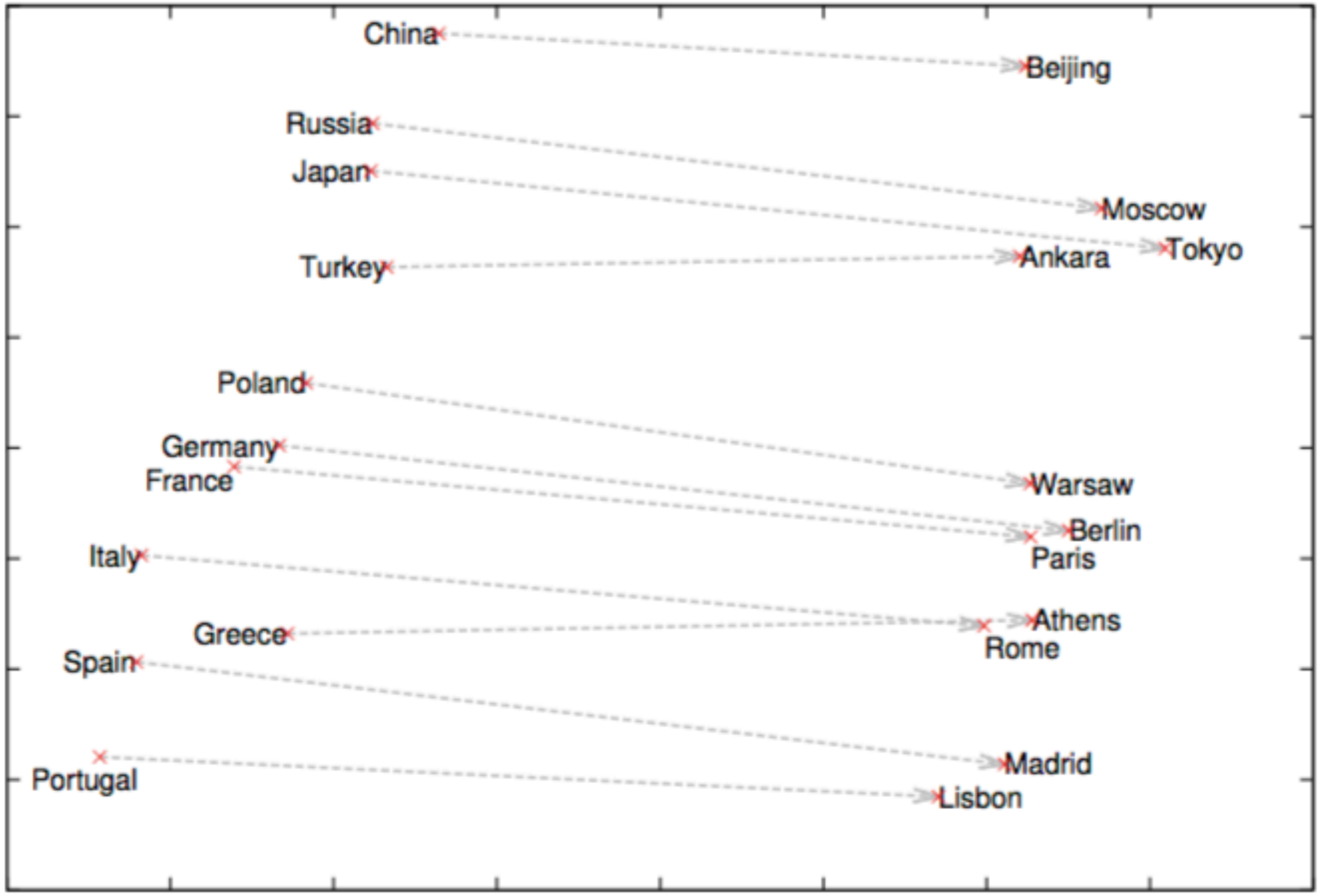
Man is to ***woman*** as

king is to ***queen*** ?

- What the solution boils down to is:

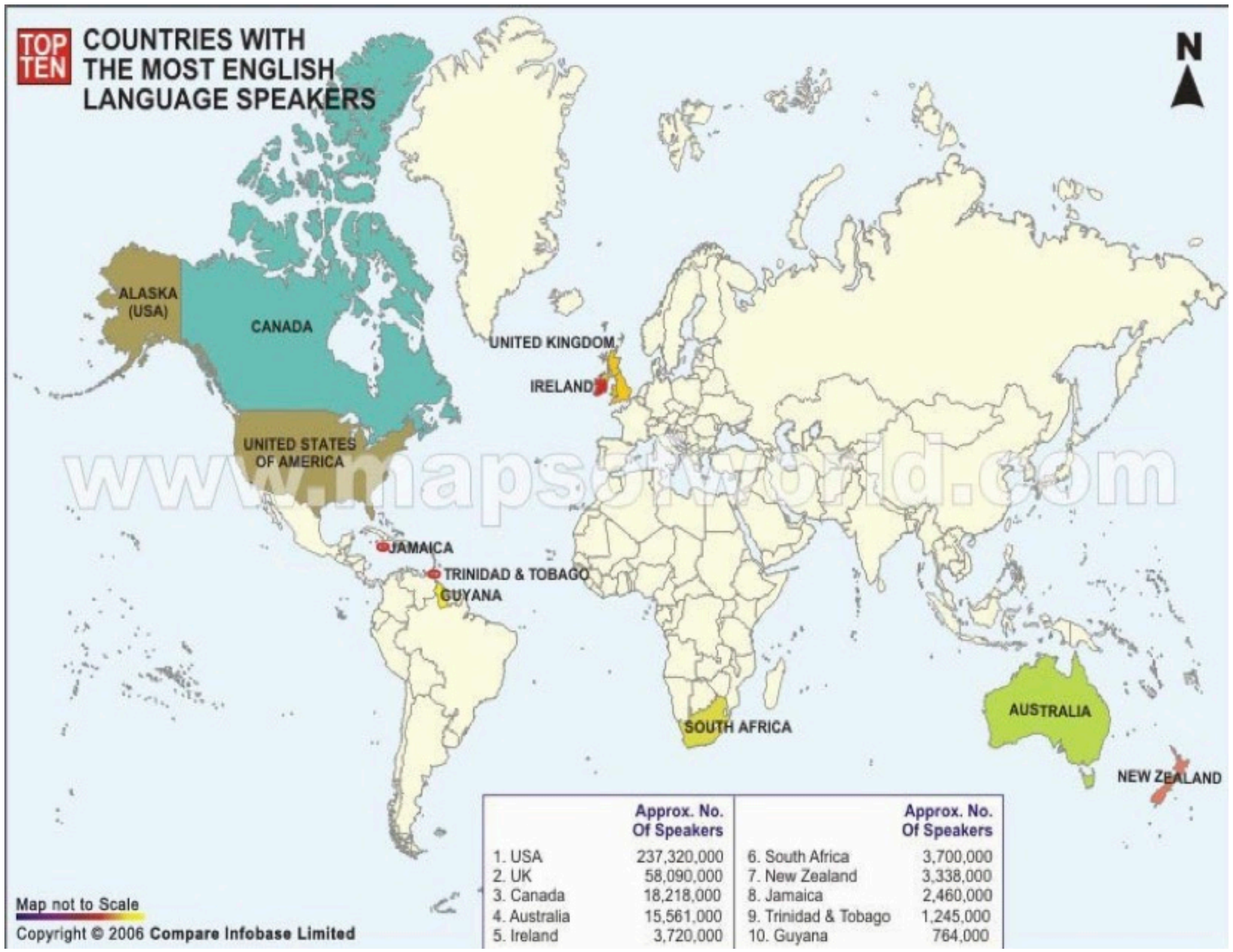
REP(***WORD***) = REP(***king***) - REP(***man***) + REP(***woman***)

- Check your intuitions
- Input: *Russia* is to *Moscow* as *China* is to ___ ?
  - *France*
  - *Germany*
  - *Greece*
  - *Italy*
  - *Japan*
  - *Poland*
  - *Portugal*
  - *Spain*
  - *Turkey*



Computational Semantics
&
Second Language Learning

Learner Errors
English Today



- About **7,000** known living languages
- Native speakers of English – about **5.52%**
- The rest – non-native speakers (language learners)
- The University of Cambridge: 18,000 students, of which **3,500** are international students from **>120** different countries

Learner Errors
Why this matters

Keywords: Text classification, hierarchical classification, feature selection, feature weighting

Abstract. In recent years, there have been extensive studies and rapid progresses in automatic text classification, which is one of the hotspots and key techniques in the information retrieval and data mining field. Feature extraction and classification algorithm are the crucial technologies for this problem. This paper firstly proposed feature extraction algorithm based on key words, the algorithm selected key words set from special part of scientific papers, and employed mutual information to extract features. And then, proposed an improved hierarchical classification method, and realized hierarchical classification of Chinese scientific papers.

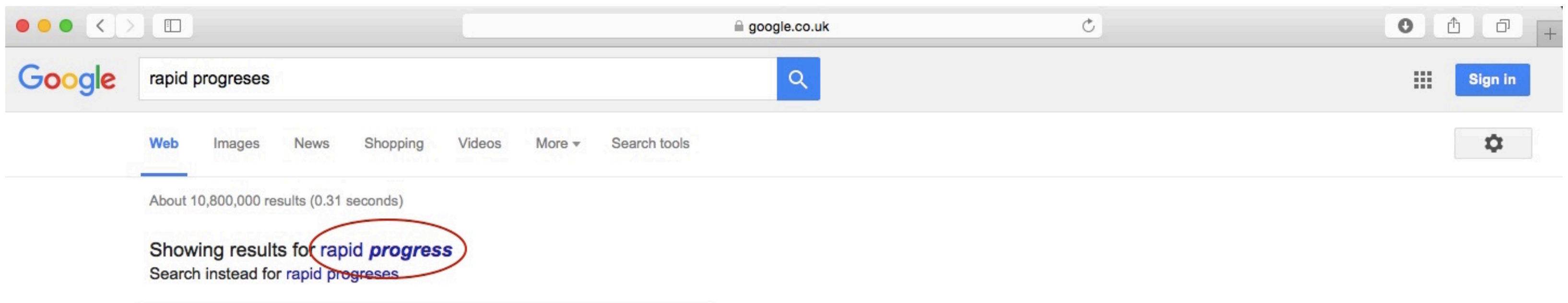
Introduction

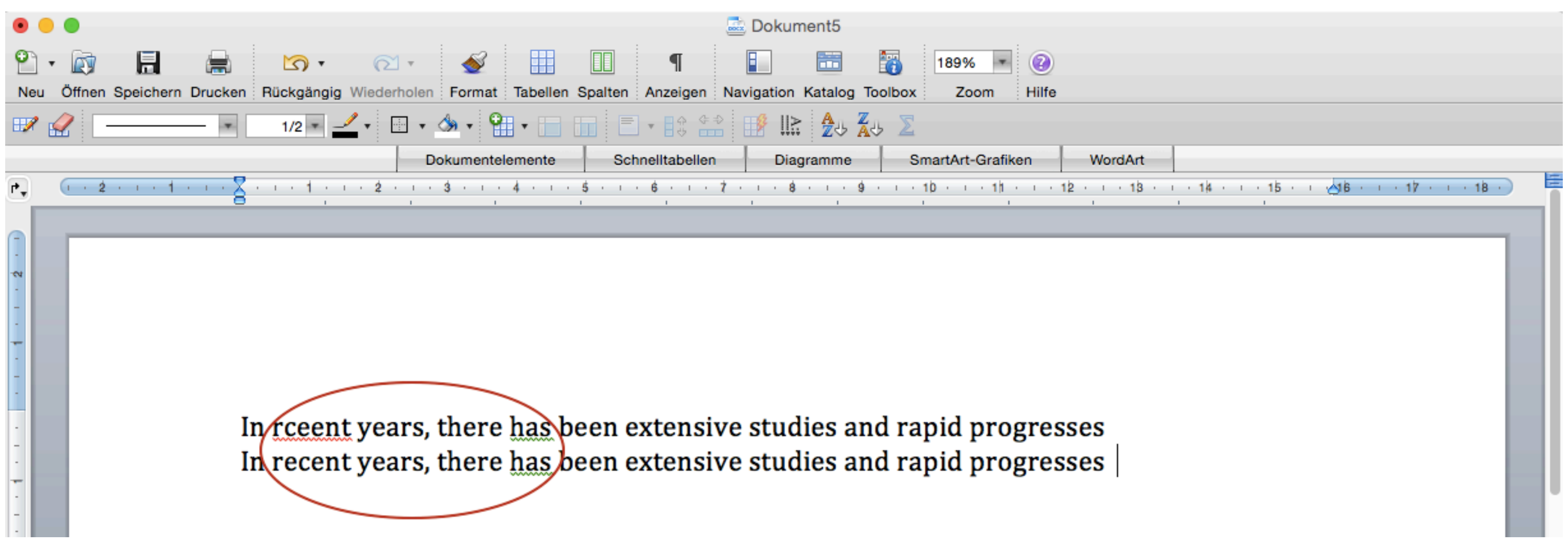
Goal of automatic text classification system is an orderly organization of the text sets, to organize the similar and related texts together. As a tool of knowledge organization, it provides more effective search strategies and more accurate query results for information retrieval.[1]

- ✦ In scientific text, it is particularly important that the ideas are clearly expressed
- ✦ What we aim to do:
  - analyse the text
  - detect the problematic areas
  - suggest corrections
  - ideally, do all of the above automatically

Learner Errors

State-of-the-art





- Currently, widely used spell-checkers and grammar-checkers can only detect and correct a **limited set** of errors (e.g., spelling, typos, some grammar)
- However, if you've picked a completely incorrect word they are unlikely to ask you if you have "***meant powerful computer instead of strong computer?***"

Does incorrect word choice impede understanding?

Error	Correction	Error type	Problematic to understand?
I am * student	I am **a** student	*Missing article*	
Last year I went ***in** London on a business trip	Last year I went **to** London on a business trip	*Wrong preposition chosen*	
***big** history ***large** knowledge …	**long** history **broad** knowledge …	*Wrong adjective chosen*	

Does incorrect word choice impede understanding?

Error	Correction	Error type	Problematic to understand?
I am * student	I am **a** student	*Missing article*	?
Last year I went ***in** London on a business trip	Last year I went **to** London on a business trip	*Wrong preposition chosen*	?
***big** history ***large** knowledge …	**long** history **broad** knowledge …	*Wrong adjective chosen*	✓

Learner Errors

Example

Depending on the word type, the change in the original meaning can be **significant**:

When somebody uses an expression ***big history*** do they mean "*academic discipline which examines history from the Big Bang to the present*"?

Content Words

Content words vs. Function words

Back to linguistics...

Function words	Content words
✦ link and relate the words to each other	✦ express the meaning of the expression
✦ are very frequent in language	✦ are conceptual units
✦ examples – articles and prepositions: *I am a student at the University of Cambridge*	✦ examples – nouns, verbs and adjectives: *I study Computer Science at the University of Cambridge. The course is very intensive*

Content Words

How to solve the task of ED in content words?

- Errors in content words (**nouns**, **verbs**, **adjectives**) are diverse → difficult to generalise and learn regularities from the data
- The contexts are also more diverse → we might never see exactly the same context around content words again and learn anything about the features
- Corrections cannot be represented as a finite set applicable to all nouns, all verbs or all adjectives in language, and they always depend on the original incorrect word
- Content words are not just linking other words, they express **meaning** → we should take semantics into account

Content Words

Types of errors in content words

- Words are confused because they are **similar in meaning**:

  He gave a small speech (short speech)

- Words are confused because they have **similar form**:

  It includes articles over ancient Greek sightseeings as the Alcropolis or other famous places (ancient sites)

- There are some other, **less obvious** reasons:

  Deep regards, John Smith (kind regards)

- Interpretation depends on the **context**, and the chosen words simply don't fit:

  The company had great turnover, which was noticable in this market (high turnover)

Semantic Approach
Semantic Space construction

	give	*last (v)*	*build*	*topic*	*big*	...
speech	85	18	0	33	1	...
talk	84	23	0	38	0	...
house	0	2	67	0	56	...

Semantic Approach

Can any language expression be modeled this way?

What happens when we try applying same models to longer expressions?

- We might find **100** examples with the word ***speech***, **50** of which will be about ***long speech***, **2** about ***45-minutes speech*** and **none** about ***7-minutes speech*** (or ***small speech***)

- That means, longer expressions (***1-hour speech***, ***1-hour long speech***) will necessarily have sparser and less reliable vectors

- Also, we won't be able to say anything about either ***7-minutes speech*** or ***small speech*** – if we don't see it in the data, does it means both are implausible / nonsensical? Have we just not looked carefully enough?

Semantic Approach
Compositional Semantics methods

Instead of relying on **distributional** information for longer phrases, let's use distributions of words within phrases and build vectors for longer phrases in a **compositional** way

- **Component-wise additive** model:

$$c_i = a_i + b_i$$

$$(\text{small\_speech})_i = \text{small}_i + \text{speech}_i$$

- **Component-wise multiplicative** model:

$$c_i = a_i \times b_i$$

$$(\text{small\_speech})_i = \text{small}_i \times \text{speech}_i$$

Semantic Approach

Measures of semantic anomaly

- Earlier, we have assumed that the computational semantic representation of words will tell us something about correctness of our examples

- Once we have modeled the phrases computationally, how can we distinguish between the representations for the correct and for the incorrect phrases?

- Since there is a direct geometric interpretation for the semantic vectors, we assume that **certain properties of the vectors** will highlight the differences

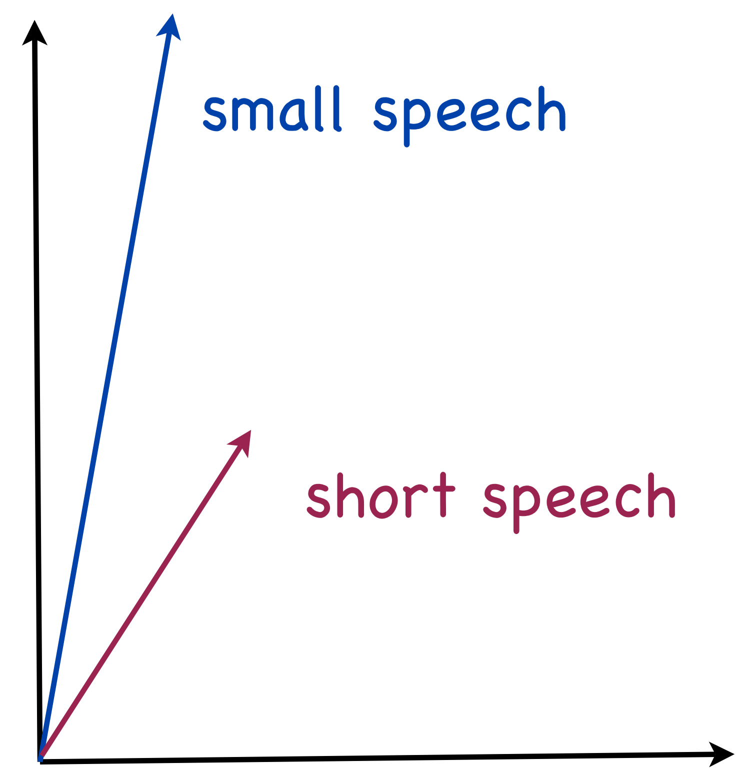
Semantic Approach

Vector length as a measure of semantic anomaly

In anomalous combinations, the counts in the input vectors are distributed differently → some “incompatible dimensions” would receive low counts → anomalous phrase vectors are expected to be **shorter** than vectors of the acceptable phrases

$$\|\boldsymbol{X}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}.$$

short	88	5
speech	92	2
small	0	30
short+speech	**180**	**7**
short×speech	**8096**	**10**
small+speech	**92**	**32**
small×speech	**0**	**60**

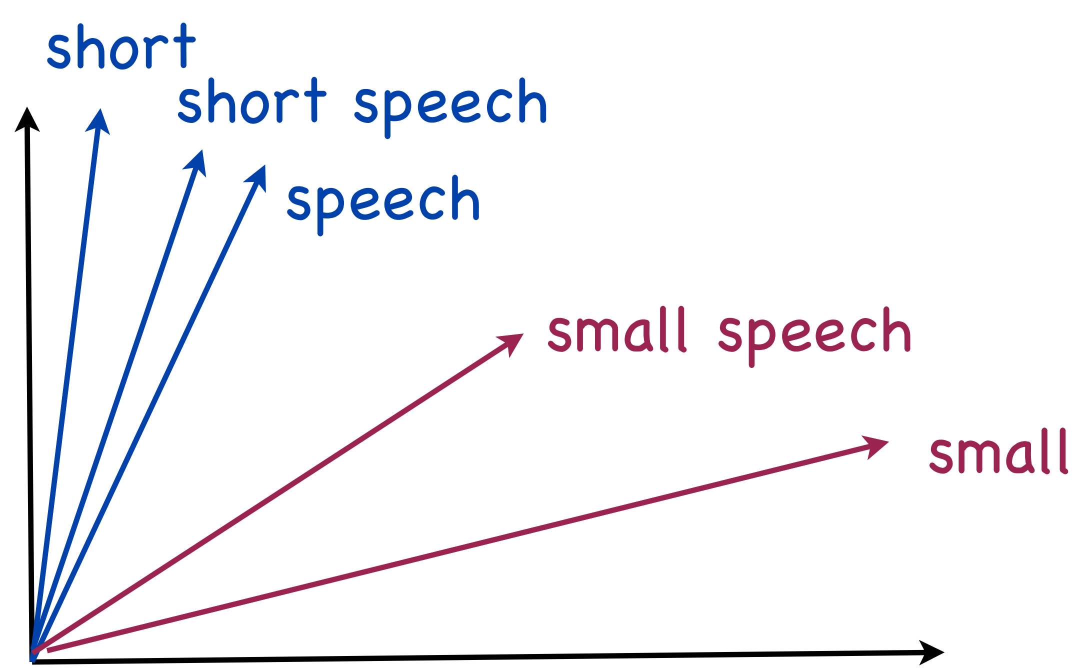


len(short + speech) = 180 len(short × speech) = 8096
len(*small + speech) = 97 len(*small × speech) = 60

Semantic Approach

Cosine to the input words as a measure of semantic anomaly

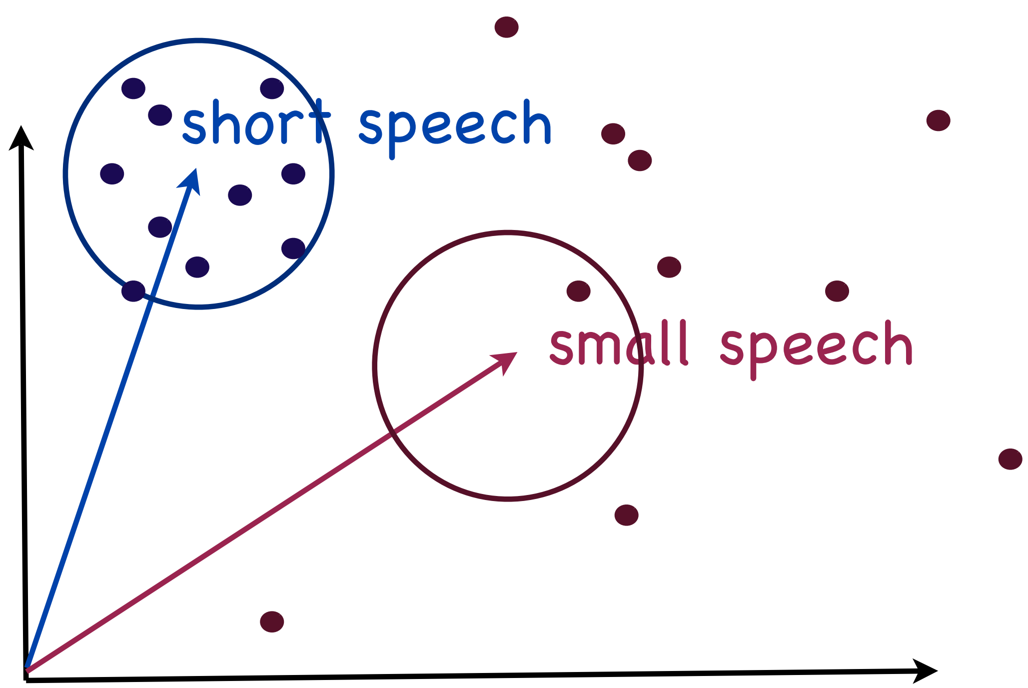
Anomalous phrases are less similar to the input nouns (verbs, adjectives), and the semantic space provides a direct interpretation of the similarity of two words via their distance in the space → vectors of the anomalous word combinations are expected to have **lower cosine (similarity)** to the input noun/verb/adjective vectors



Semantic Approach

Neighbourhood density as a measure of semantic anomaly

Anomalous phrase vectors are expected to not have any specific meaning → they are expected to not be closely surrounded by other words with similar meaning → have sparser neighbourhoods in the semantic space. We measure this as an **average cosine** (= distance) to the 10 nearest neighbours



Semantic Approach

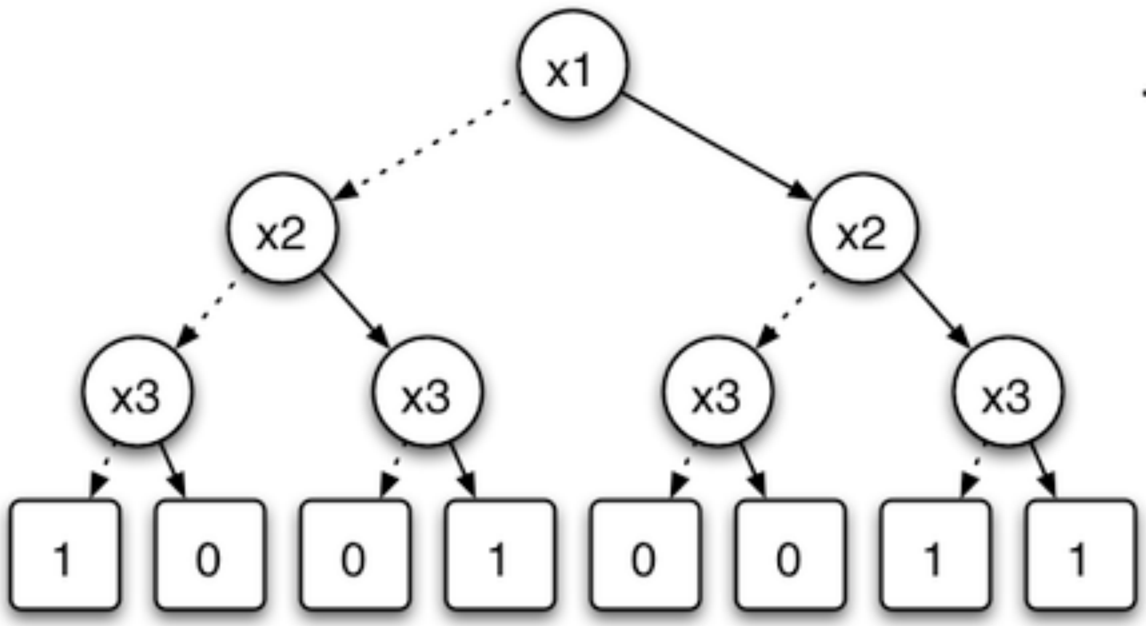
Component overlap as a measure of semantic anomaly

We assume semantically acceptable phrases to be placed in the neighbourhoods populated by **similar words and combinations**, and calculate the proportion of neighbours containing the same words as the input phrases. We expect this **proportion** to be lower for the anomalous phrases (**lower overlap**)

short speech	*small speech*
• [x] **speech**	• quantity
• **short** [x]	• **small** amount
• talk	• person
• ...	• ...

Semantic approach:

Machine Learning classifier for ED



x1	x2	x3	*f*
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	1

- We apply ***Decision Tree Classifier*** to our task
- Two classes – ***correct*** (0) and ***incorrect*** (1)
- At each node, the classifier checks whether the value of the feature falls within a certain value interval (e.g., whether ***len<0.5*** or ***len>=0.5***) and follows the relevant path
- The algorithm makes sure the most discriminative rules are applied first

Semantic approach:
Results

Content word combinations	**Accuracy** (averaged over 5 folds)	**Lower bound** (=majority class distribution)	**Upper bound** (=annotator agreement)
adjective-noun	**0.6535** ± 0.0189	0.5084	**0.7467** ± 0.0221
verb-noun	**0.6491** ± 0.0188	0.6086	**0.8467** ± 0.0377

ED System

Further evaluation of the ED system

- ***Precision*** = #(instances that belong to class ***n*** & are identified by the system as belonging to class ***n***) / #(all instances identified by the system as belonging to class ***n***)

$$\text{Precision} = \frac{tp}{tp + fp}$$

- ***Recall*** = #(instances that belong to class ***n*** & are identified by the system as belonging to class ***n***) / #(instances in the data that actually belong to class ***n***)

$$\text{Recall} = \frac{tp}{tp + fn}$$

- ***F-measure*** – harmonic mean of the two

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

	Predicted (+)	Predicted (-)
Actual (+)	***tp***	***fn***
Actual (-)	***fp***	***tn***

ED System

Class-specific performance of the ED system

Content word combinations	**Precision**	**Recall**	**F1**
adjective-noun, ***correct***	0.6173	0.7226	0.6558
adjective-noun, ***incorrect***	0.7071	0.5898	0.6409

ED System

Class-specific performance of the ED system

Content word combinations	Precision	Recall	F1
verb-noun, ***correct***	0.6027	0.3192	0.4174
verb-noun, ***incorrect***	0.6637	0.8630	0.7503

ED System

Summary on the ED system

- We have showed that our algorithm detects errors with high accuracy
- There is still some room for improvement – it is close to, but does not yet reach human performance on this task
- The features derived using semantics and trying to capture the meaning of the words are useful
- The algorithm shows high precision → it is reliable → learners can use it to detect errors in their writing

Thank you!

- Further information:
  - http://www.cl.cam.ac.uk/~ek358/
  - Ekaterina.Kochmar@cl.cam.ac.uk
- Datasets:
  - http://www.cambridgeenglish.org
  - http://www.cl.cam.ac.uk/~ek358/an-dataset.xml
  - http://ilexir.co.uk/applications/adjective-noun-dataset/
- Useful resources:
  - Jurafsky and Martin. Speech and Language Processing. Second Edition, 2009 (https://web.stanford.edu/~jurafsky/slp3)