Language Identification

A Computational Linguistics Primer

Will Fitzgerald
Powerset (Microsoft)
entish.org / powerset.com

The talk

- Introduction to Computational Linguistics using an example, Language Identification
- Review a bit of Computational Linguistics History & Current Computational Linguistics
- Look at two traditional linguistics problems using Language Identification
- Please ask questions ...

Computational linguistics

- **Linguistics** is the “scientific study of language.”

- In theory, **Computational linguistics** is the art and science of using computational means do to linguistics (cf computational chemistry, computational biology, computational material science, computational philopsophy ...

- In practice, computational linguistics has come to mean a statistical/empirical approach to linguistics.

What language is this?

What language is this?

neoda

What language is this?

neodarwini

What language is this?

hipotézis értelmezhető a neodarwini elmélet keretein belül.

What language is this?

Az evolúcióelmélet másik nagy alakja, Dawkins úgy látja, a hipotézis értelmezhető a neodarwini elmélet keretein belül.

- „Daniel C. Dennett: Darwin veszélyes ideája”
ÉRDI PÉTER
TEREMTETT VALÓSÁG
http://mek.niif.hu/05000/05015/html/index.htm

The language identification problem

- Identifying, from a sample of text or speech, the language in which the sample was produced.

Velkomstord Mine damer og herrer, det er mig en stor glæde at kunne byde velkommen til en ...	Liebe Kolleginnen und Kollegen! Im Namen unseres Hauses begrüße ich eine Delegation des ...	Καλωσόρισμα Αγαπητοί κυρίες και κύριοι συνάδελφοι, εξ ονόματος του ...
Welcome Ladies and gentlemen, on behalf of the House let me welcome a delegation...	Bienvenida Deseo dar la bienvenida a los miembros de una delegación de ...	Souhaits de bienvenue Chers collègues, je souhaite, au nom du Parlement, la ..
Hyvδt naiset ja herrat, jδlleen kerran parlamentti kokoontuu valitsemaan ...	Boas-vindas Caros colegas! Em nome do nosso Parlamento saúdo uma delegação da ...	Mina damer och herrar! Än en gång sammanträder vårt parlament för ...

Stupid language tricks

- Try this at home!

- First, get two relatively large texts (“corpora”) in different languages, and gzip them. Record their sizes.

```
[will:~/lang-id/indata] ls -la *.en
-rw-r--r--   1 will  will  16320 Feb 13  charter.en
-rw-r--r--   1 will  will  17274 Feb 13  charter.fr
[will:~/lang-id/indata] gzip charter.en
[will:~/lang-id/indata] gzip charter.fr
[will:~/lang-id/indata] ls -la *.gz
-rw-r--r--   1 will  will  5066 Feb 13 charter.en.gz
-rw-r--r--   1 will  will  5579 Feb 13 charter.fr.gz
```

Stupid language tricks (ii)

- Then, combine a text sample to be identified with each of the original corpora. The sample must come *after* each corpus.
- Gzip, and record sizes.

```
[will:/lang-id/indata] cat charter.en alouette.txt > test.en
[will:/lang-id/indata] cat charter.fr alouette.txt > test.fr
[will:/lang-id/indata] gzip test.en
[will:/lang-id/indata] gzip test.fr
[will:/lang-id/indata] ls -la *.gz
-rw-r--r--   1 will  will  5185 Feb 13 21:15 test.en.gz
-rw-r--r--   1 will  will  5691 Feb 13 21:15 test.fr.gz
```

Stupid language tricks (iii)

- Subtract the size of each original corpus from the larger corpus.
- The language causing the smaller difference will (probably) be the language of the sample text.

```
Difference between test.en.gz and charter.en.gz:

5185-5066 = 119  # English difference

Difference between test.fr.gz and charter.fr.gz:

5691-5579 = 112  # FRENCH difference

C’est français!
```

Stupid language tricks (iv)

- What about the Gettysburg Address?

```
[will:/lang-id/indata] cat charter.fr gettsyburg.txt > test.fr
[will:/lang-id/indata] cat charter.en gettsyburg.txt > test.en
[will:/lang-id/indata] cat charter.fr gettsyburg.txt > test.fr
[will:/lang-id/indata] gzip test.en
[will:/lang-id/indata] gzip test.fr
[will:/lang-id/indata] ls -la *.gz
-rw-r--r--   1 will  will  5696 Feb 13 21:20 test.en.gz
-rw-r--r--   1 will  will  6275 Feb 13 21:20 test.fr.gz

5696-5066 = 630 # ENGLISH difference
6275-5579 = 696 # FRENCH difference
```

Why does this work?

- Compression techniques look for encodings that are optimized for space.
- More redundant/more frequent codes in the original represented by smaller codes in the compressed file.
- Different languages have different code frequencies.

Some terms

- **Character bigram**: a unique two-letter long sequence (“aa”, “ab” ..)

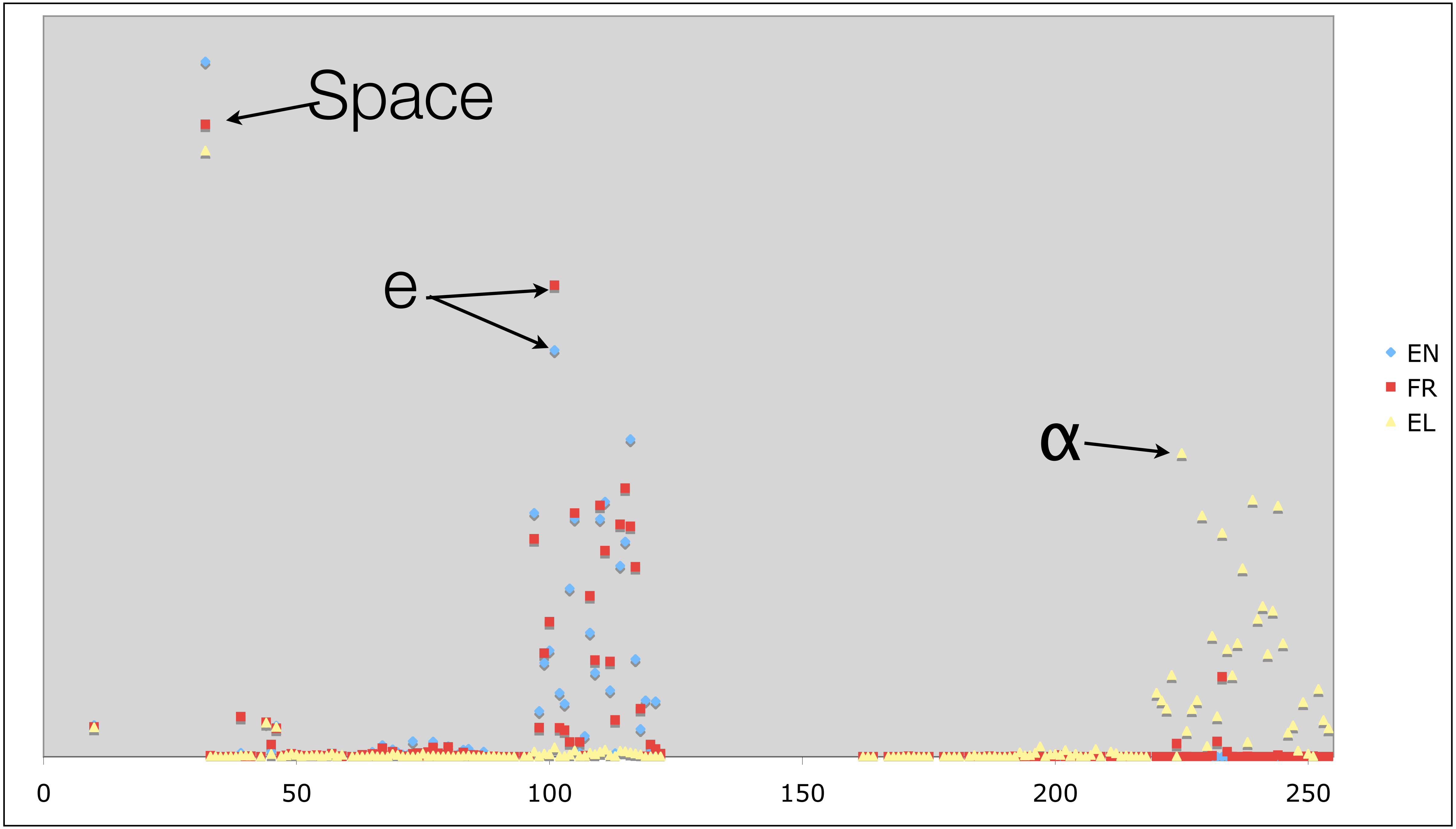
- **Character trigram**: a unique three-letter long sequence (“aaa”, “aab”, ...)

- **Character n-gram**: a unique n-character long sequence of letters

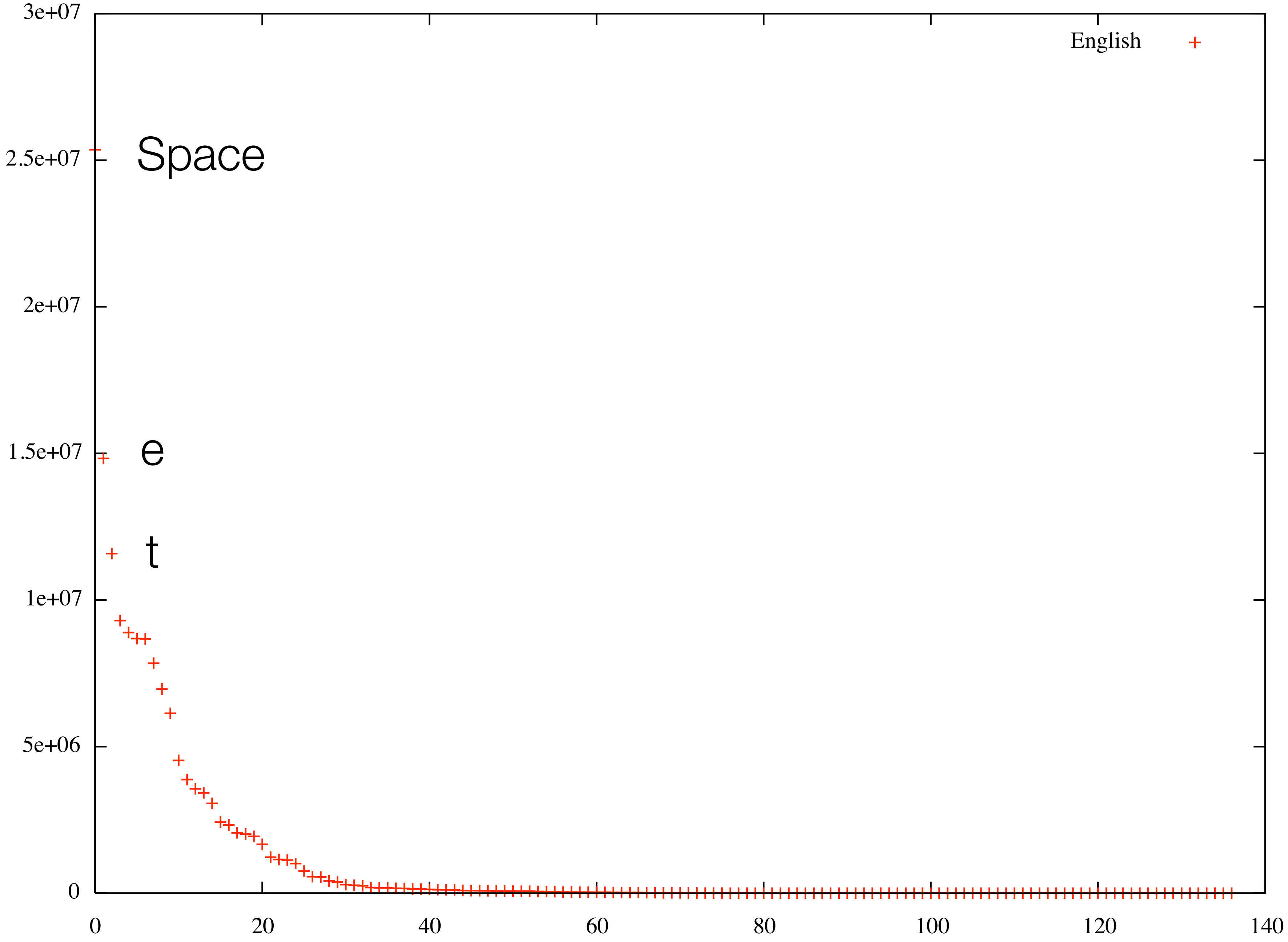
- **N-gram frequency**: how frequently an n-gram appears in (some sample) text.

- **Character encoding**: how character is represented. For example, map the integers 0-255 (one byte) to Latin characters (32 ↔ “_”, 41 ↔ “A”, 97 ↔ “a”)

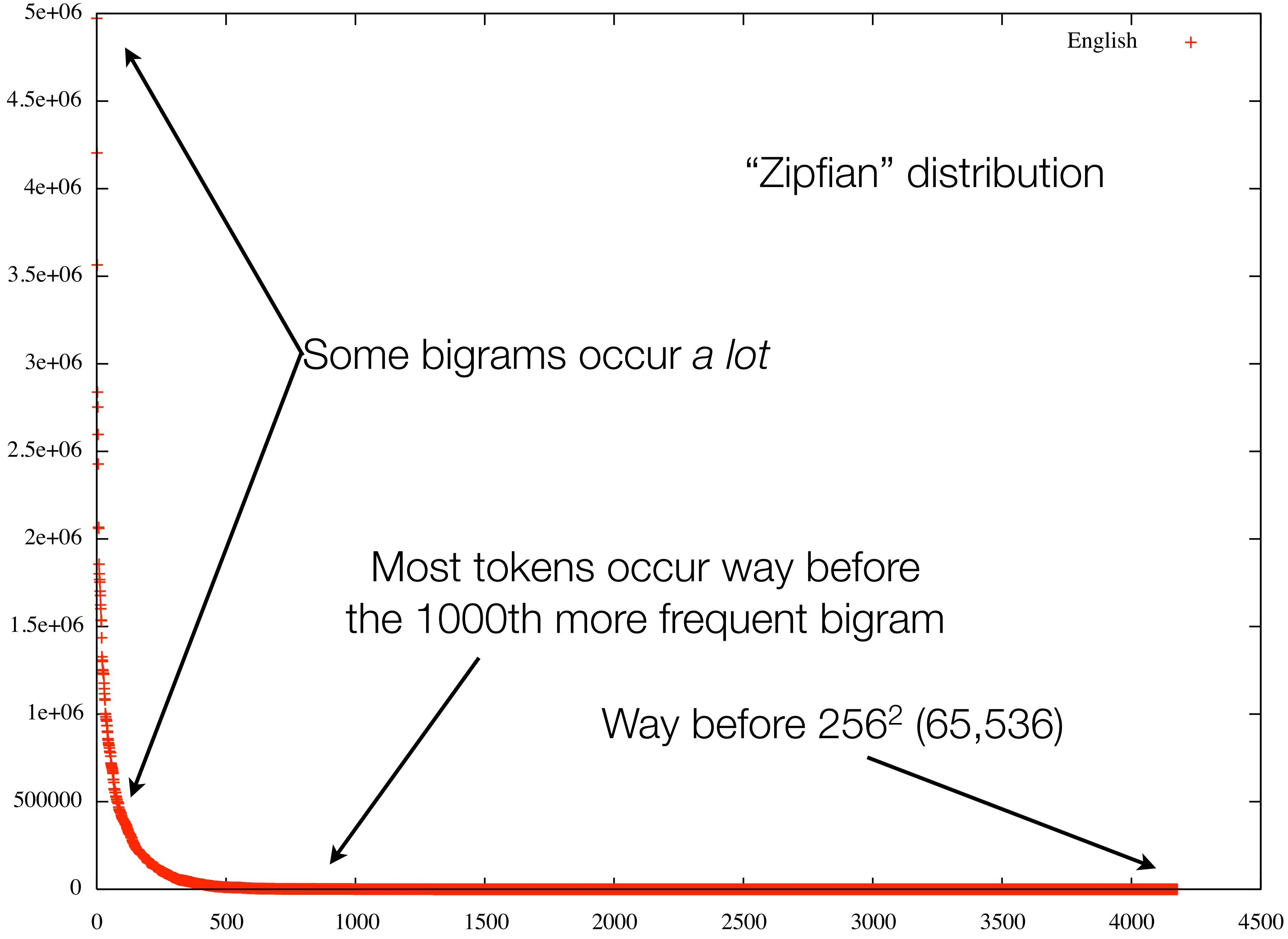
English (EN), French (FR) and Greek(EL) character frequency



English character frequency



English Character bigram frequency



“Zipfian” distribution

Some bigrams occur *a lot*

Most tokens occur way before the 1000th more frequent bigram

Way before 256^2 (65,536)

Language Identification

- The basic idea: train a language identifier on a large corpus of text from a given language. “Training” means gathering compression/frequency/information data on *n*-gram occurrence.

- Use these language identifiers to judge new texts; the fewest bits required indicate the winning identifier.

Results

- Eleven language identifiers created from a subset of the European parliamentary debate transcripts (Europarl) data, using character 2-grams

- Eleven monolingual texts, one in each language, created from John 1, then run through the language identifiers

- Bits required per bigram table (fewer bits better)

Bits/bigram		Base language of language identifier										
		DA	DE	EL	EN	ES	FI	FR	IT	NL	PT	SV
Lg of text	DA	**8.72**	9.76	14.9	10	10.7	10.9	10.6	10.8	9.70	10.9	9.30
	DE	9.91	**8.46**	14.7	9.65	10.6	10.7	10.2	10.6	9.36	10.8	9.73
	EL	25.4	26.1	**8.49**	26.3	25	26.1	25.5	24.9	25.9	25	23.1
	EN	10.2	10	15	**8.76**	10.8	11.1	10.6	10.9	9.79	11	10.3
	ES	10.4	10.5	14.3	9.91	**8.63**	11	9.75	9.68	10.4	9.22	10.5
	FI	11.1	10.5	15.4	11.3	11.8	**8.53**	11.5	11.7	11.2	12	9.82
	FR	10.5	10.6	14.1	10.5	10.2	11.3	**8.90**	10.1	10.5	10.2	11
	IT	10.2	10.2	14.1	9.86	9.46	10.5	9.75	**8.68**	10.2	9.50	10.1
	NL	10.1	9.56	14.9	10.1	11	10.8	10.7	11.1	**8.74**	11.2	10.3
	PT	10.8	10.7	14.2	10.4	9.79	11.2	10.1	10.1	10.7	**9.03**	10.7
	SV	9.52	9.67	14.9	10.4	11	10.4	10.7	11	10.3	11.2	**8.35**

Very simple algorithm

- Training:
  - For each corpus, collect frequency statistics on *n*-grams occurring in corpus *c* of length |c|.
  - Bits required for *n*-gram *i* (basically, log of relative frequency):

$$-lg\frac{f(i)}{|c|-n}$$

Simple algorithm (ii)

- Identification of a text:
  - For each language identifier,
    - Sum the number of bits required to encode the *n*-grams in the text. (Divide by number of *n*-grams).
  - The language identifier which requires the fewest bits is the best guess.

but...

- Lots of small details to consider:
    - Encoding of original corpus
    - Number of possible *n*-grams (*e.g.*, 256^n)
    - Training vs. testing corpora

One big detail

- What to do about missing n-grams?
  - Most n-grams will be missing, especially in the “other” languages
  - -lg(0) is undefined
  - Requires ‘smoothing.’ For character n-grams, probably ok to use -lg(1/count), but not for word n-grams – why?

    See Dunning Statistical identification of Language, 1994

Related detail

- Most n-grams have very (or even very, very, very) low frequencies. Consequence of:
  - Large encoding space (Consider word n-grams)
  - Zipfian distribution
- Often use log probabilities instead (Of course, this is almost the same as information value)

A History of Computational Linguistics in Four Slides

Computational Linguistics empiricism

- Claude Shannon formalizes the maths of information (late 40s)

- Warren Weaver’s memo on machine translation (1949):

  If one examines the words in a book, one at a time through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of words. “Fast” may mean “rapid”; or it may mean “motionless”; and there is no way of telling which. But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning . . . It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?

Reaction

- Chomskyan linguistics and descendants emphasizes “discrete” models over “analog” ones; syntax, semantics, ... (50s on)

- Schankian and other “good old fashioned AI” approaches focus on semantics and complex models (70s on)

New Empiricism

- Cheap, fast computers and memory; vast amounts of data; intelligent researchers resurrect empirical approaches, e.g.:
  - Speech recognition, Natural Language Processing at Bell, IBM (1990s)
  - Special issue of *Computational Linguistics* "Using Large Corpora" (1993):

    From the introduction: "When the idea first arose to publish a special issue of CL on using large corpora, the topic was not generally considered to be part of mainstream CL, in spite of an active community working in this field."

Some papers from a recent ACL conference

- Scaling Phrase-Based Statistical Machine Translation to Larger Corpora and Longer Phrases
- A Hierarchical Phrase-Based Model for Statistical Machine Translation
- Dependency Treelet Translation: Syntactically Informed Phrasal SMT
- A Probabilistic Framework for the Evaluation of Text Summarization Systems
- Supervised and Unsupervised Learning for Sentence Compression
- Word Sense Disambiguation vs. Statistical Machine Translation
- Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning

What Makes a Model of a Modern CL Paper?

A good paper...

- Solves a real problem, using real data over large domains
- Is mathematically sophisticated, empirically based
- Has a clear evaluation metric

Evaluating language identification algorithms

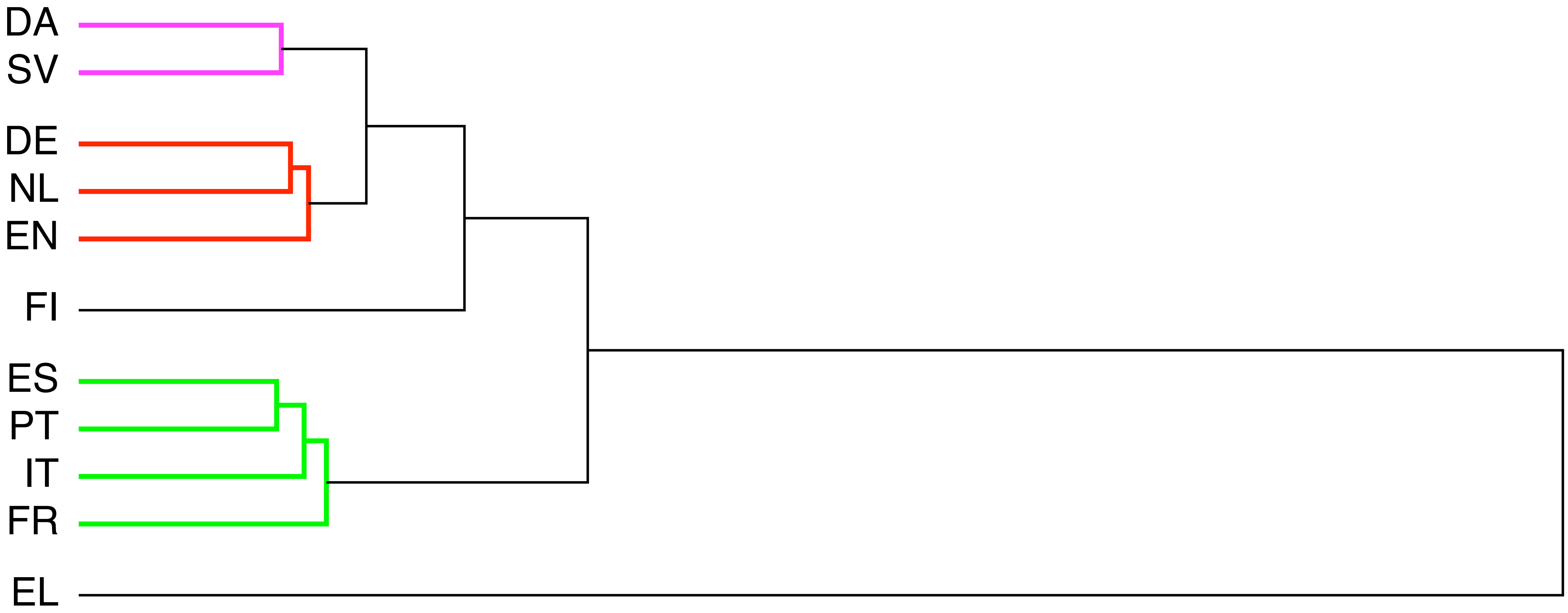
- Examples for this talk are anecdotal
- One typical evaluation metric: divide corpora into ten parts, train on nine, test on one; repeat ten times
- Another is to use standard evaluation corpora

Using Language Identification to do Two Traditional Linguistics Studies

Results of running language identifiers on English text (5131 characters)

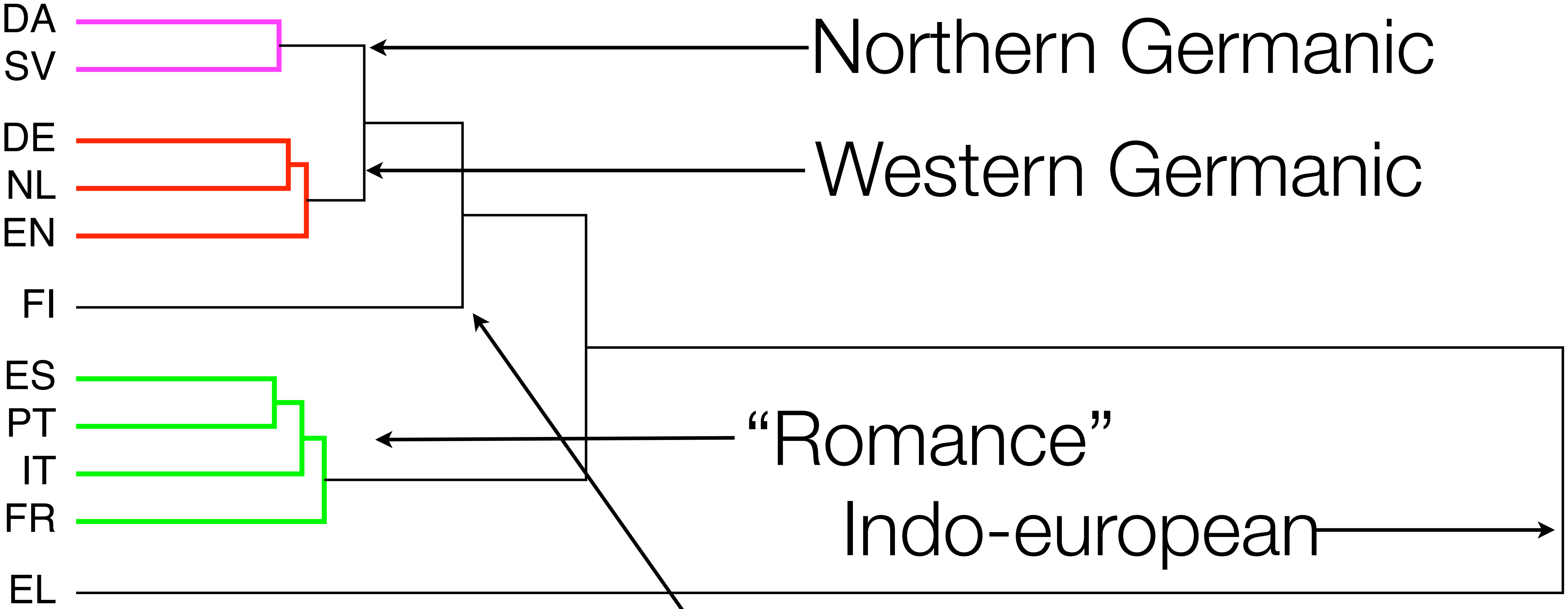
Language of identifier		Bits per bigram
EN	English	8.76 (vs. 16)
NL	Dutch	9.79
DE	German	10.02
DA	Danish	10.23
SV	Swedish	10.34
FR	French	10.62
ES	Spanish	10.75
IT	Italian	10.88
PT	Portuguese	11.01
FI	Finnish	11.10
EL	Greek	14.97

Cheap historical linguistics



Language relatedness dendrogram created using bigram results

Cheap historical linguistics (II)



Northern Germanic

Western Germanic

“Romance”

Indo-european

Finnish misplaced within Indo-european

Code-switching

- **Code-switching** is the act of changing from one language to another in mid-discourse.

- It’s been a topic of sociolinguistics research for some time.

Web examples

French and English weblog entry with comments, http://martinepage.com/blog/

Pause lunch. Je décide de regarder une entrevue à la télé avec un jeune romancier québécois plutôt populaire dans notre coin de la blogosphère francophone. Sympa. Puis tout à coup, une petite révélation de sa part: il avoue ne pas lire de livres, ou très peu. Il aime bien lire mais la vie lui offre d'autres stimulations ailleurs. Mais quand il lit, c'est bien, ça lui plait. Comme une inhabituelle visite au musée qui nous fait penser qu'on devrait y aller plus souvent. Mais on n'y va jamais plus souvent.

...

five blue said... ben... MOI je pense que c'est mal! m'enfin... les gens qui ne lisent pas, de façon générale, je les trouve peu intéressants - pas qu'ils ne puissent pas l'être, mai je ne peux pas connecter, c'est comme s'ils venaient d'une autre planète...

AJ said...I think you can draw a distinction between keeping-up-on-trends in one's field, and being literate in general. And I understand his comment completely.

It's like when you cross that boundary from being a music fan, and listening to lots of new albums a year, to becoming a musician, and barely listening to any new music at all. (I admit to this.)

Dutch, German and English business weblog, http://www.interdependent.biz/main/index2.html

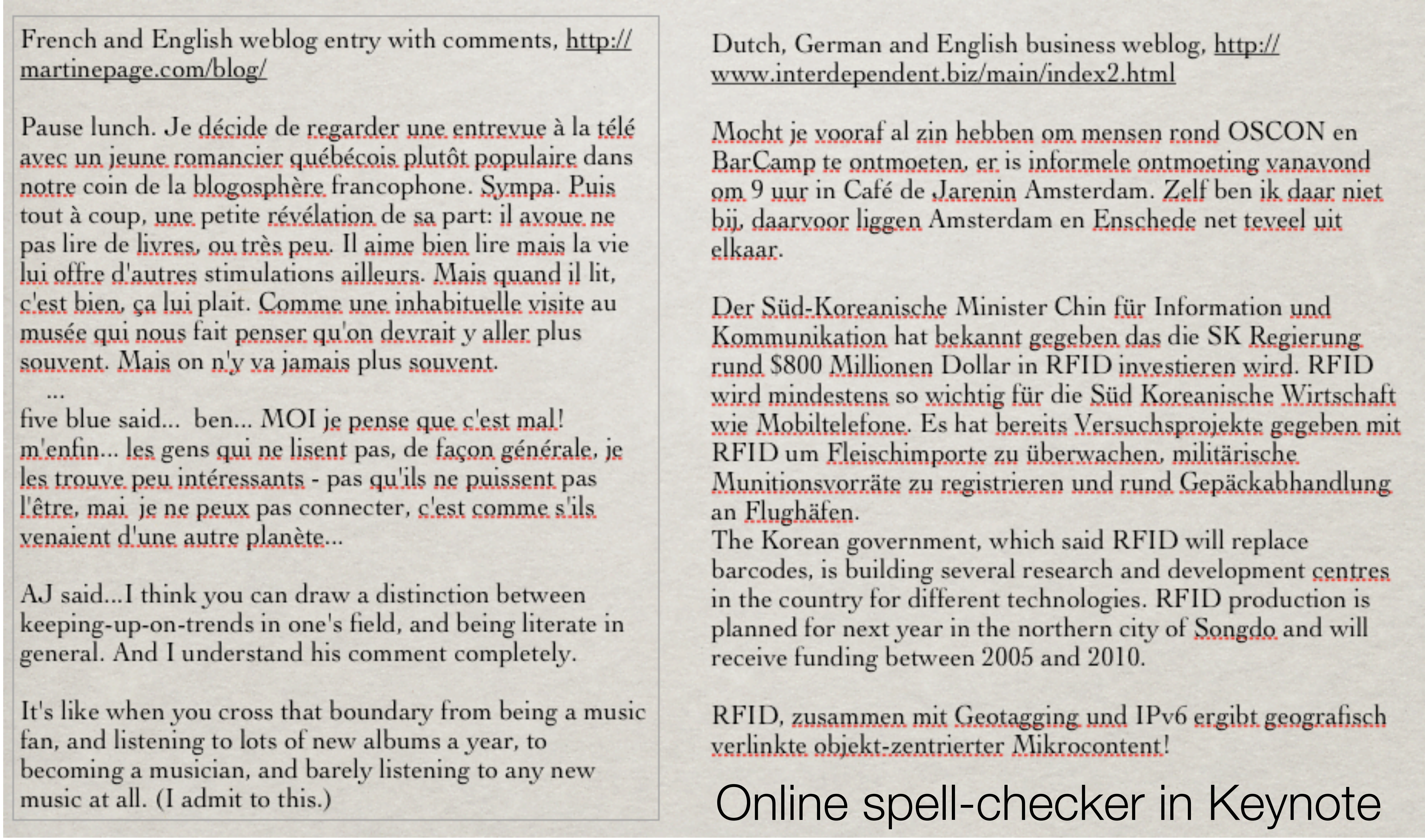
Mocht je vooraf al zin hebben om mensen rond OSCON en BarCamp te ontmoeten, er is informele ontmoeting vanavond om 9 uur in Café de Jarenin Amsterdam. Zelf ben ik daar niet bij, daarvoor liggen Amsterdam en Enschede net teveel uit elkaar.

Der Süd-Koreanische Minister Chin für Information und Kommunikation hat bekannt gegeben das die SK Regierung rund $800 Millionen Dollar in RFID investieren wird. RFID wird mindestens so wichtig für die Süd Koreanische Wirtschaft wie Mobiltelefone. Es hat bereits Versuchsprojekte gegeben mit RFID um Fleischimporte zu überwachen, militärische Munitionsvorräte zu registrieren und rund Gepäckabhandlung an Flughäfen.

The Korean government, which said RFID will replace barcodes, is building several research and development centres in the country for different technologies. RFID production is planned for next year in the northern city of Songdo and will receive funding between 2005 and 2010.

RFID, zusammen mit Geotagging und IPv6 ergibt geografisch verlinkte objekt-zentrierter Mikrocontent!

Multilingual spell-checking



French and English weblog entry with comments, http://martinepage.com/blog/

Pause lunch. Je décide de regarder une entrevue à la télé avec un jeune romancier québécois plutôt populaire dans notre coin de la blogosphère francophone. Sympa. Puis tout à coup, une petite révélation de sa part: il avoue ne pas lire de livres, ou très peu. Il aime bien lire mais la vie lui offre d'autres stimulations ailleurs. Mais quand il lit, c'est bien, ça lui plait. Comme une inhabituelle visite au musée qui nous fait penser qu'on devrait y aller plus souvent. Mais on n'y va jamais plus souvent.

...

five blue said... ben... MOI je pense que c'est mal! m'enfin... les gens qui ne lisent pas, de façon générale, je les trouve peu intéressants - pas qu'ils ne puissent pas l'être, mai je ne peux pas connecter, c'est comme s'ils venaient d'une autre planète...

AJ said...I think you can draw a distinction between keeping-up-on-trends in one's field, and being literate in general. And I understand his comment completely.

It's like when you cross that boundary from being a music fan, and listening to lots of new albums a year, to becoming a musician, and barely listening to any new music at all. (I admit to this.)

Dutch, German and English business weblog, http://www.interdependent.biz/main/index2.html

Mocht je vooraf al zin hebben om mensen rond OSCON en BarCamp te ontmoeten, er is informele ontmoeting vanavond om 9 uur in Café de Jarenin Amsterdam. Zelf ben ik daar niet bij, daarvoor liggen Amsterdam en Enschede net teveel uit elkaar.

Der Süd-Koreanische Minister Chin für Information und Kommunikation hat bekannt gegeben das die SK Regierung rund $800 Millionen Dollar in RFID investieren wird. RFID wird mindestens so wichtig für die Süd Koreanische Wirtschaft wie Mobiltelefone. Es hat bereits Versuchsprojekte gegeben mit RFID um Fleischimporte zu überwachen, militärische Munitionsvorräte zu registrieren und rund Gepäckabhandlung an Flughäfen.
The Korean government, which said RFID will replace barcodes, is building several research and development centres in the country for different technologies. RFID production is planned for next year in the northern city of Songdo and will receive funding between 2005 and 2010.

RFID, zusammen mit Geotagging und IPv6 ergibt geografisch verlinkte objekt-zentrierter Mikrocontent!

Online spell-checker in Keynote

Code-switching identification

- Algorithm 1: Define a window size, *s*, and run language identification on each window.

- Algorithm 2: Do language identification by logical or syntactic unit (paragraph, sentence, phrase).

Examples

- Two English/French weblogs
- Dutch/English/German business weblog
- Exploratory results

Is the focus on numbers good for the field?

- Despite this talk, there’s lots to do without a lot of mathematics
- Focus on quantitative (vs. qualitative) evaluation is good (working on large scales).
- Still room for exploratory research...

In conclusion...

- Recommendations for aspiring Computational Linguists of the statistical kind:
    - Take **computer science** (**machine learning**, NLP)
    - Take **linguistics**
    - Take **discrete mathematics**, statistics and combinatorics, Bayesian statistics
- New old/paradigm: Combine linguistic (syntax/semantics/pragmatics) with stats

Thank you