N-Gram Based Natural Language Classification for Single Novel Words

December 12, 2006

Mike Buchanan

The Problem

I have an undated, post-WWII photograph of the gates to a Jewish cemetery in Skopje (formerly Yugoslavia, now Republic of Macedonia).

Underneath the Hebrew text are the words in block letters: "IZRAELITSḰO POKOPALIŠČE" (the diacritics being my best guess). My questions:

* What language is this?
* What does the text mean?

-- Many thanks, Deborahjay 07:34, 11 December 2006 (UTC)

Is Letter Frequency a Solution?

- Letter frequency analysis of that suggests it is Bosnian, but possibly Czech, Croatian, Serbocroatian, Lithuanian, Slovak, and Slovenian.
- “Diacritic on "Ḱ" should not be there — maybe a damage on the inscription or photo?” Duja 10:31, 12 December 2006 (UTC)

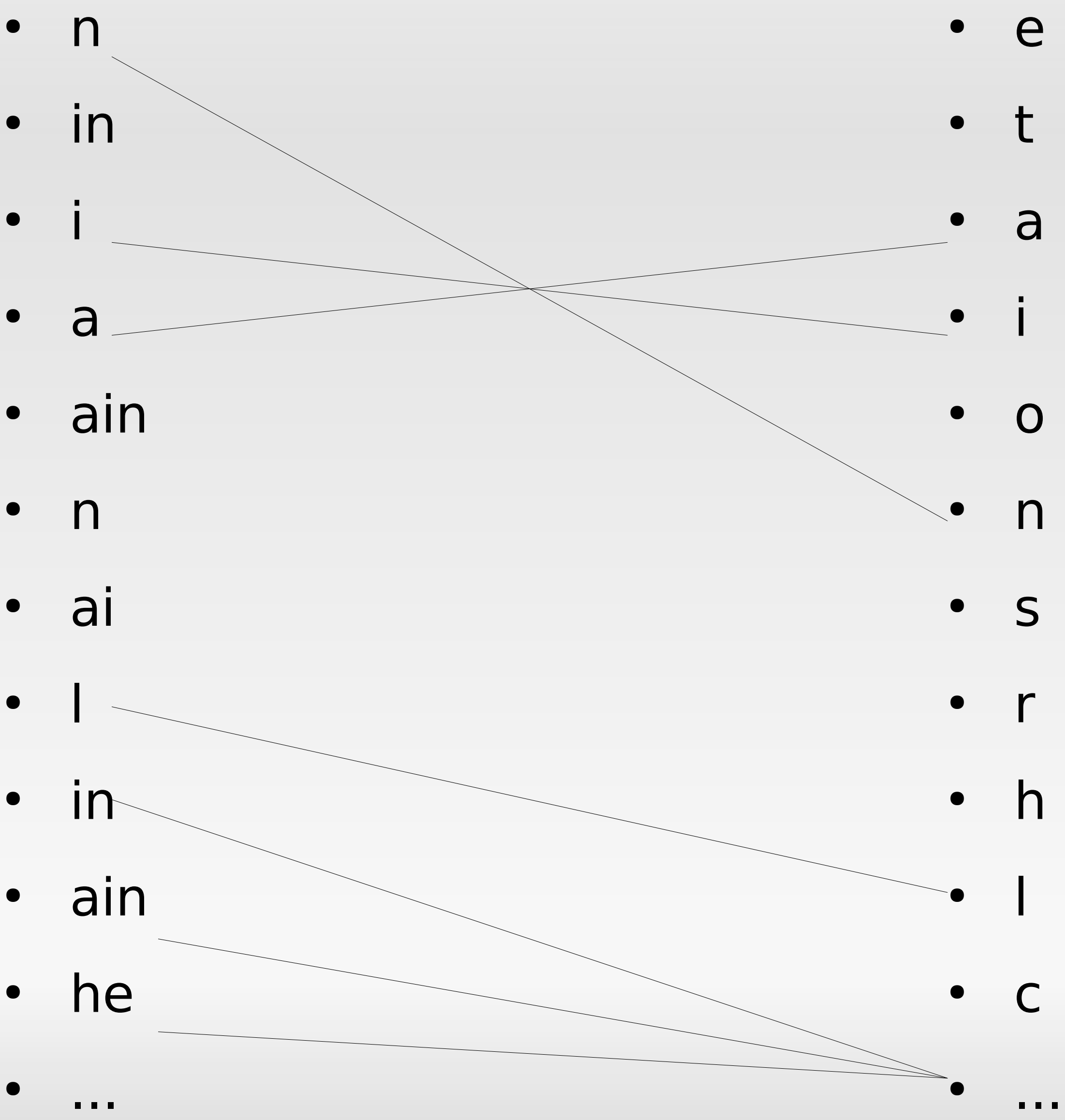
N-Grams

- Look at more than one letter at a time.
- “abcdef” becomes:
  - a b c d e f
  - ab bc cd de ef
  - abc bcd cde def
  - abcd bcde cdef

N-Gram Frequency

- “The rain in Spain falls mainly on the plains.”
- n, in, i, a, ain, n , ai, l, in , ain , he, s, p, e, h
- By letter frequency, that could be quite a few languages. Many fewer languages have the “ai” dipthong.

Distance Metric



- n
- in
- i
- a
- ain
- n
- ai
- l
- in
- ain
- he
- …

- e
- t
- a
- i
- o
- n
- s
- r
- h
- l
- c
- …

Misses

- What do you do when an n-gram is in one sample but not another?
- How far is “Википедија” from English?

What would Cavnar and Trenkle do?

- Previous N-gram based approaches have limited their frequency profiles to some small constant length (~300), so a miss obviously cost 300.
- My frequency profiles are not limited to finite length: the more training data, the longer the profile.
- No obvious answer.

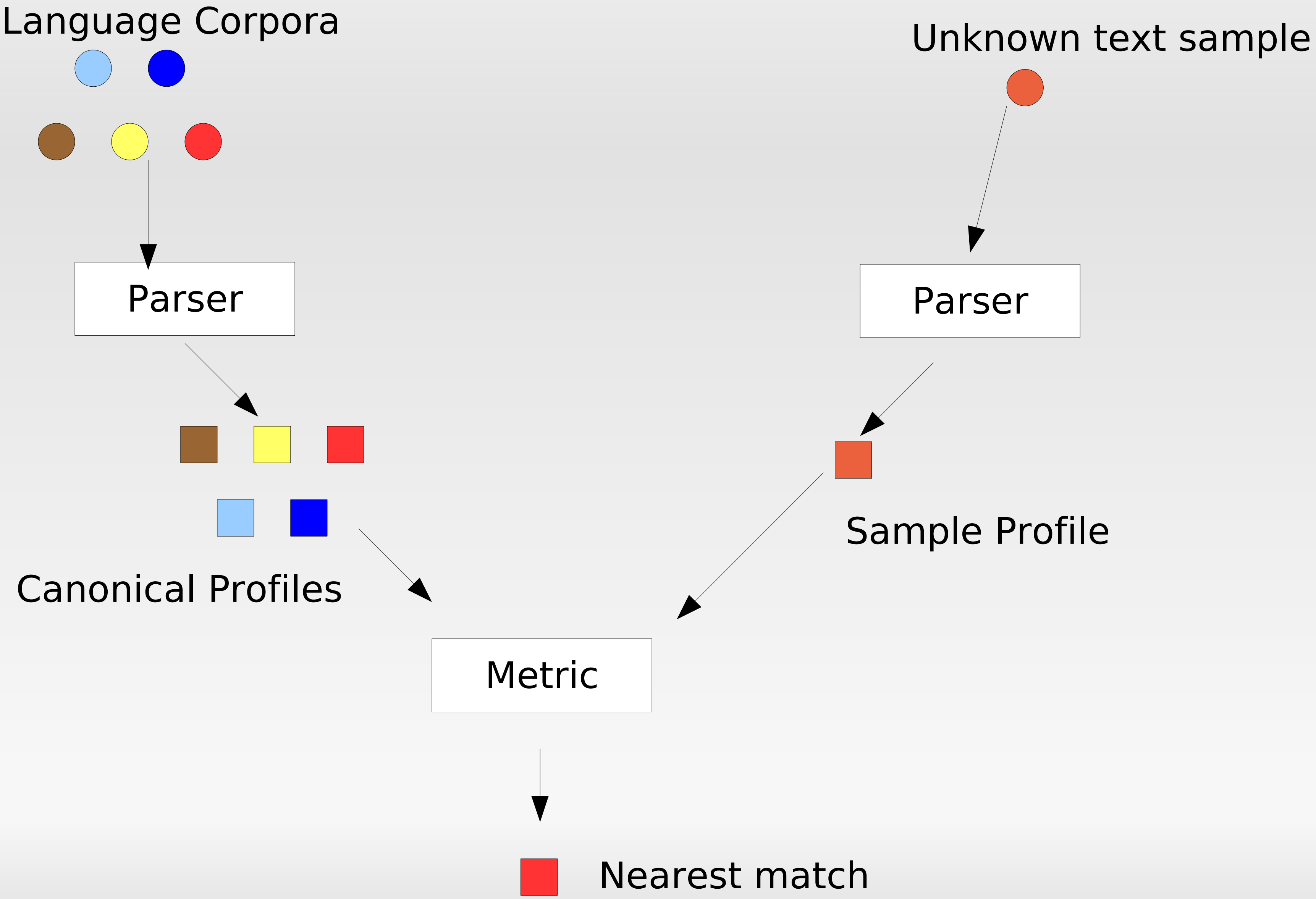
Idea 1: Individual Size

- Misses for a given language depend on that language's frequency profile length.
- Misses for English cost lots, misses for Zulu almost nothing.
- Result: Zulu does very well.
- Observation: The null language always wins.
- This might be desirable if some possible languages have scarce training data.

Idea 2: Uniform Miss Cost

- We want to be fairer to languages with large profiles.
- Let's agree on one miss cost for all languages.
  - Maximum: Unfair to short profiles.
  - Minimum: Unfair to long profiles.
  - Mean: Just right?

General organization



Language Corpora

Unknown text sample

Parser

Parser

Canonical Profiles

Sample Profile

Metric

Nearest match

Where to get corpora?

- It's pretty easy to get corpora for English, most European languages, etc. What about our example from Macedonia?
- Wikipedia is in 250 languages, from Afrikaans to Zulu, including Lojban, Navajo, Manx, Assyrian Neo-Aramaic, and Klingon.
- Markup is a bit painful, but I'm now the master of sed.

Does it work?

- My network says that “IZRAELITSḰO POKOPALIŠČE” is Slovenian.
- “It's definitely Slovenian” -Duja 10:31, 12 December 2006 (UTC)
- No quantitative results yet, because the network is too much fun to play with.
- Demo!