Detection and correction of OCR-errors

Souhail Bouricha

Slides based on article by Martin Reynaert (2008)

Unlocking the secrets of the past: Text Mining for Historical Documents
(WS 2009/2010)

Lecturers: Caroline Sporleder & Martin Schriber

Saarland University
22.02.2010

What is OCR?

- Optical Character Recognition
- Branch of computer sciences that involves:

  - reading text from paper

  - translating the images into a manipulated form
- OCR systems use a combination of Hardware/Software to recognize characters
- OCR technologie is said to have been born in 1951 with M. Sheppered's invention GISMO

Reasons for using OCR

- To reduce data entry errors
- To consolidate data entry
- To handle peak loads
- Human Readable
- Can be used with any printing techniques
- Scanning correction
- Eco-friendly

How does OCR work?

- Pattern Matching: compares what the OCR scanner sees as character with a library of character matrices or templates
- Feature Extraction:
  - Known as Intelligent Character Recognition (ICR)
  - This method varies by how much

  "Computer Intelligence" is applied by the manufacturer
  - The computer looks for general features such as open areas, closed shapes, diagonal lines, etc.

OCR Fonts

A font is the term given to a set of characters, for example in English language usually 0-9, A-Z and a few special characters.

Each character within a font will have a defined reproducible size and shape.

OCR's efficient?

OCR system reaches 99% word accuracy!!!

One word will have been misrecognized out of every 100 words processed

Error Sources

- Text location and format
- Print quality
- Paper quality
- Positioning a Scanner
- Writing quality

Corpora of the Cultural Heritage

1- SGD: "Staten Generaal Digitaal"
Contemporary collection comprise the published acts of Parliament (1989-95) of the Netherlands

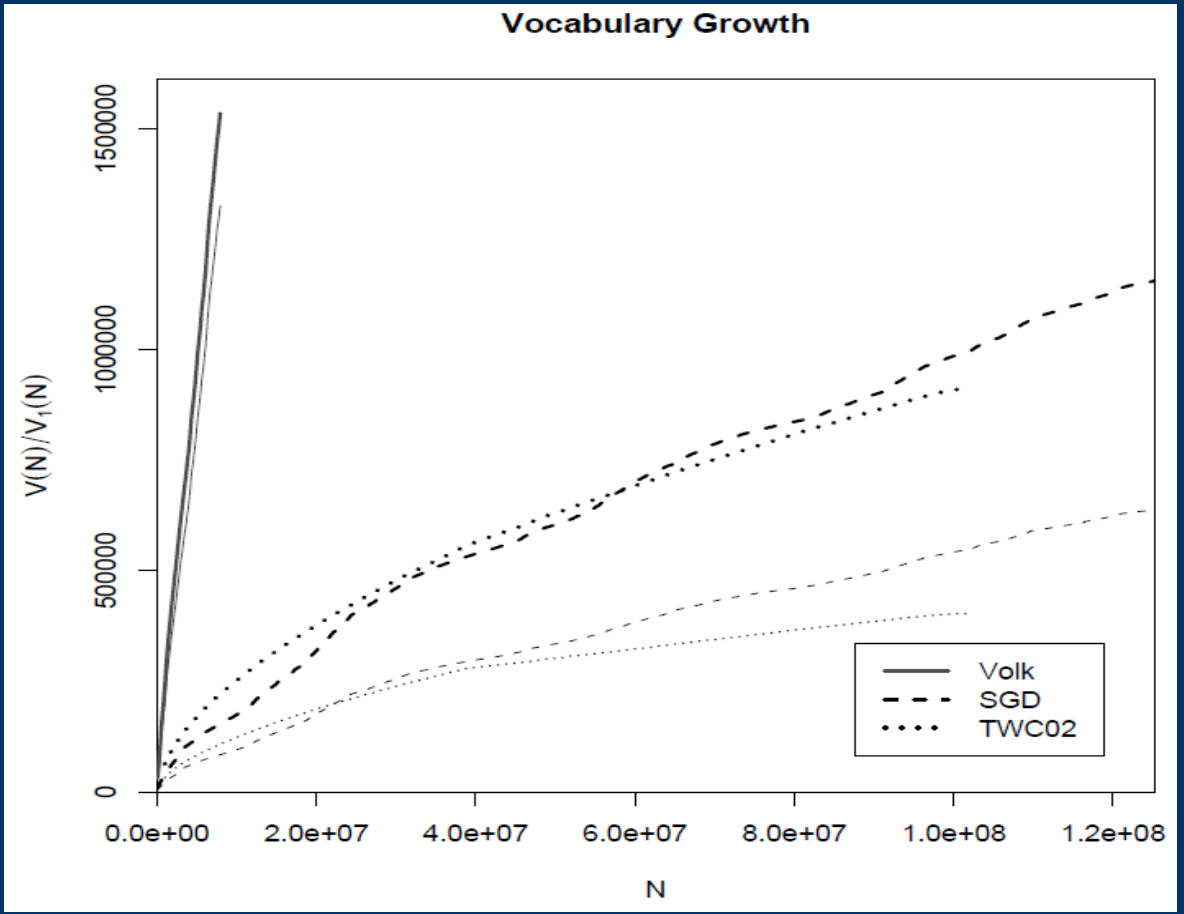
2- DDD:"Database Digital Daily newspapers"
- Historical collection
- published between 1918-46
- was written in an older Dutch spelling

3- TWC02: Contemporary one year newspaper corpus(2002), 5 Dutch newspapers, one called "Het Volk"

Background

Token: Number of words in a text(are repeated)
Types: abstract and unique
Ratio: Number representing a comparison between two things
Born-Digital: (Natively digital vs. Digital reformatting) Materials that originate in a digital form
Hapax legomena: A word occuring only once in a given corpus

Lexical Variation in Corpora



Corpus	Lang.	Origin	Tokens	Types	TTR
TWC2	CD	BD	92,793,519	914,026	0.985%
SGD	CD	OCR	125,209,007	1,156,998	0.924%
DDD	HD	OCR	7,950,950	1,535,529	19.31%

Categories of errors

1- Transposition
2- Insertion
3- Deletion
4- Substitution

OCR Post-correction (*TICCL*)

- Text-Induced Corpus Clean-up

- automatic

- work for most alphabetical languages

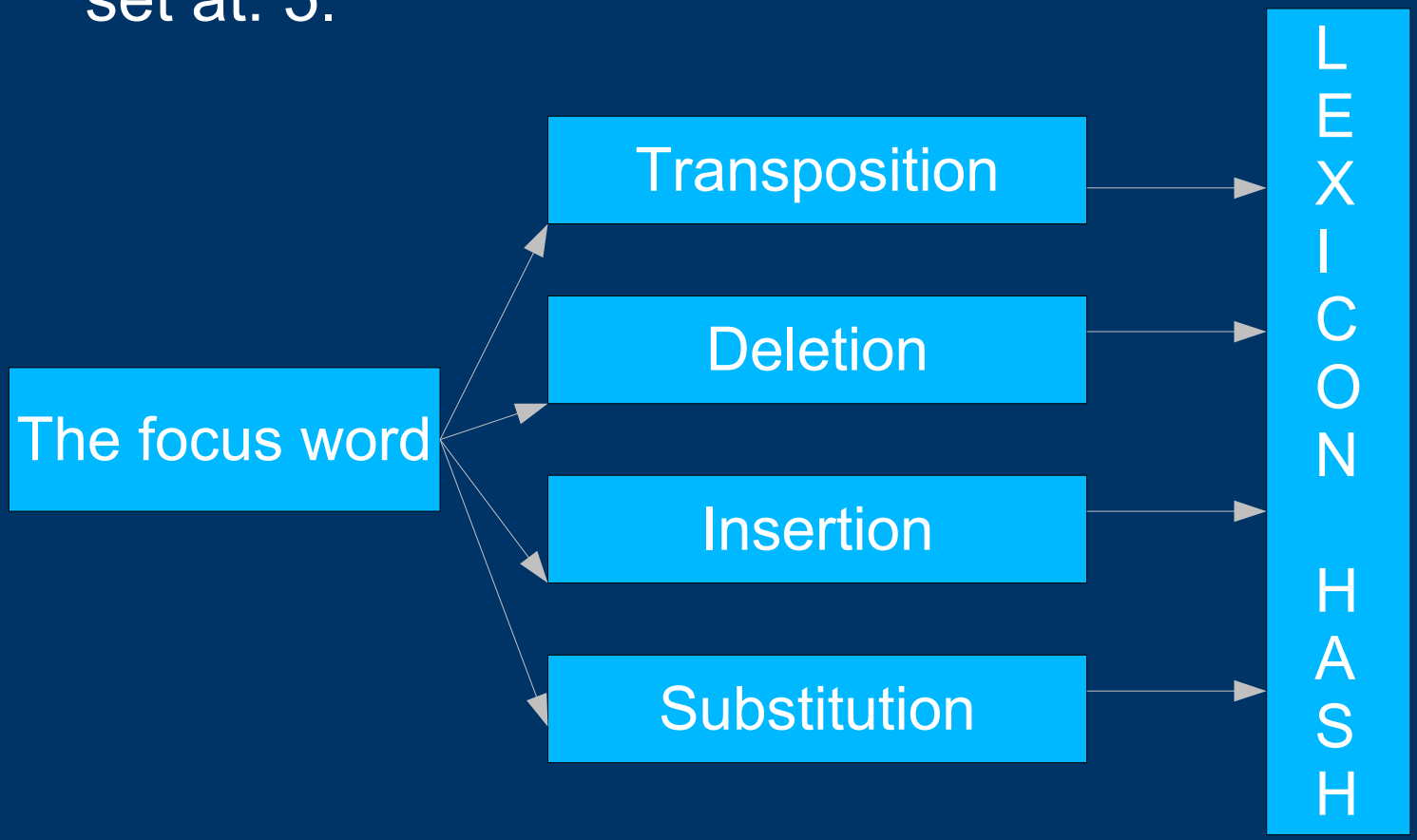
- does not try to account for unknown word types

- the system can be run with or without an extra validated word lexicon

- the system is able to drive a word type list from a backgound corpus

Anagram Hashing

The numerical value for a word string is obtained by summing the ISO Latin-1 code of each character in the string raised to a power n, where n is emperically set at: 5.



Processing Steps

1- we compare each word with the background lexicon

2- Each word in the corpus has a diffrent frequency

3- we associate the frequency of a word in the corpus with the same word in lexicon

4- TICCL reads a list of variants of the focus word (only if it's available)

5- TICCL returns: focus word and retrieved variant
(That we got through Lexicon and Morphological filter)