# Detection and Correction of OCR and Transcription Errors

*presented by Hüseyin Mergan*

*Lecturers: Caroline Sporleder & Martin Schriber*

*slides based on articles by Mihov et al (2008) and Reynaert (2005)*

**Unlocking the Secrets of the Past: Text Mining for Historical Documents (Project Seminar WS 08/09)**

**Saarland University**

**16.02.2009**

# Background

OCR (Optical Character Recogntion)

- ☑ Converts uneditable format into editable format.
  - ☑ Uneditable formats: hard copies, image files and pdf files etc.
- ☑ Not optimal even with collected English data.
- ☑ Depends on the quality of data and data itself.
- ☑ Most recent and popular projects: Project Gutenberg and Google Books

Type: different things present

Token: every individual things

Hapax legomena: word types occuring only once.

# Sofia Munich

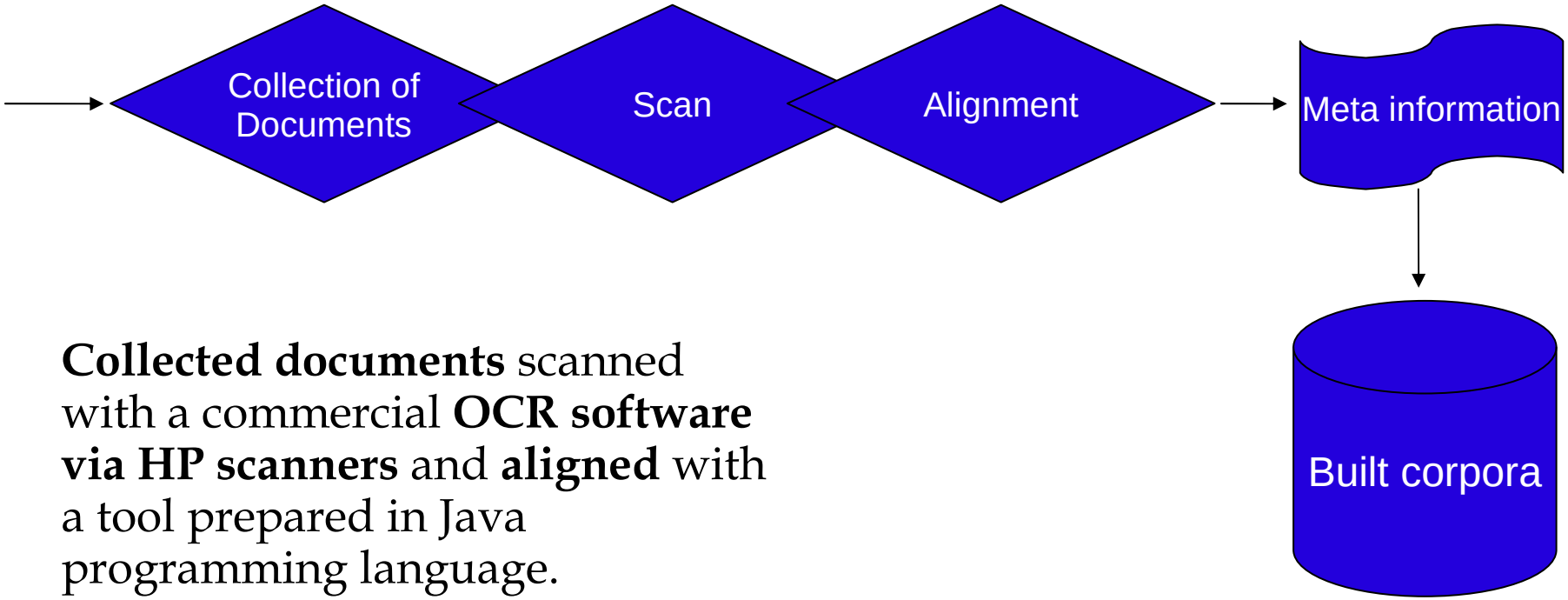
Motivation: To create corpora that can be used for researches in East European languages

- It should have distinct char., sets, genres, content and doc types

Properties:

- 2618 files (real life documents and covers wide range of document types)
- Multilingual (Bulgarian, German)
- Some docs have images, logos and strokes
- Date of documents range between 1980-2004
- Collection of fax, typewriter, laser and matrix printer
- Files are stored In PNG format scanned at 600 dpi (grey scales)*

# Sofia Munich




**Collected documents** scanned with a commercial **OCR software via HP scanners** and **aligned** with a tool prepared in Java programming language.

# Error Sources

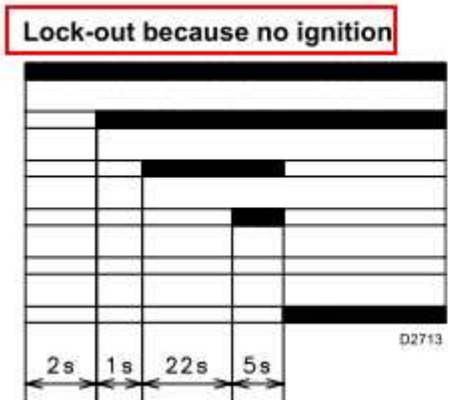
- Cyrillic letters
- Positioning on Scanner
- Paper/writing quality
- Texts location and format
- Tables
- Contrast and blurring
- Columns
- Print Quality

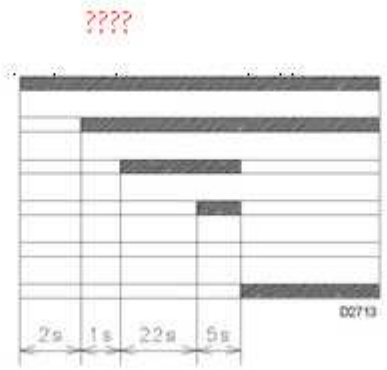
# Some Examples

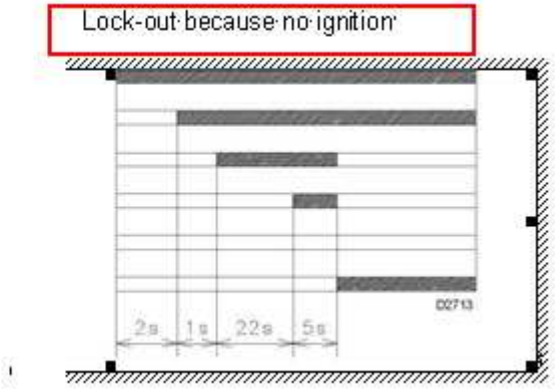


Same content one is erroneous while other is correct. Although the words are in lexicon, they were misread. This error may be result of paper quality and size.

# Some More Examples




The chart overlaps the text.

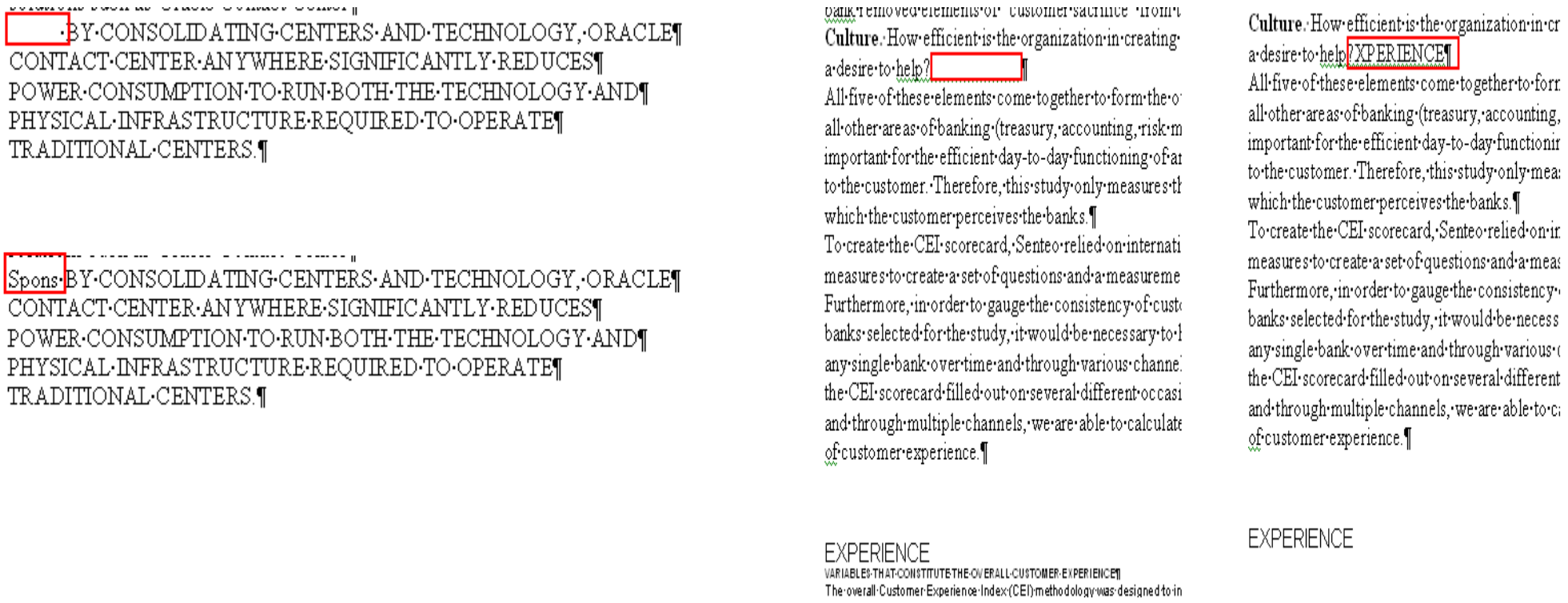
Besides, the parts in the chart is uneditable


Adapazarı'nı → r and ı merges.

# Error Patterns

- Cyrillic to Latin symbol substitution
  - (Ц → LJ)
- Unknown symbol substitution
  - Ğ merges with the letter just above due to the diacritic.
- Digit and case substitution
- Merging and splitting symbols (multiple contigous – multi C, multiple non contiguous multi-NC)
  - r and ı → n, r and n → m
- Merging and splitting words (1)
- False Friends
- Paragraph at the end of the each line (my own experience)

# Error Patterns (1)



·BY·CONSOLIDATING·CENTERS·AND·TECHNOLOGY,·ORACLE¶
CONTACT·CENTER·ANYWHERE·SIGNIFICANTLY·REDUCES¶
POWER·CONSUMPTION·TO·RUN·BOTH·THE·TECHNOLOGY·AND¶
PHYSICAL·INFRASTRUCTURE·REQUIRED·TO·OPERATE¶
TRADITIONAL·CENTERS.¶

Spons·BY·CONSOLIDATING·CENTERS·AND·TECHNOLOGY,·ORACLE¶
CONTACT·CENTER·ANYWHERE·SIGNIFICANTLY·REDUCES¶
POWER·CONSUMPTION·TO·RUN·BOTH·THE·TECHNOLOGY·AND¶
PHYSICAL·INFRASTRUCTURE·REQUIRED·TO·OPERATE¶
TRADITIONAL·CENTERS.¶

**Culture**.·How·efficient·is·the·organization·in·creating·
a·desire·to·help?¶
All·five·of·these·elements·come·together·to·form·the·
all·other·areas·of·banking·(treasury,·accounting,·risk·m
important·for·the·efficient·day-to-day·functioning·of·a
to·the·customer.·Therefore,·this·study·only·measures·t
which·the·customer·perceives·the·banks.¶
To·create·the·CEI·scorecard,·Senteo·relied·on·internati
measures·to·create·a·set·of·questions·and·a·measureme
Furthermore,·in·order·to·gauge·the·consistency·of·cust
banks·selected·for·the·study,·it·would·be·necessary·to·
any·single·bank·over·time·and·through·various·channe
the·CEI·scorecard·filled·out·on·several·different·occasi
and·through·multiple·channels,·we·are·able·to·calculate
of·customer·experience.¶

EXPERIENCE
VARIABLES·THAT·CONSTITUTE·THE·OVERALL·CUSTOMER·EXPERIENCE¶
The·overall·Customer·Experience·Index·(CEI)·methodology·was·designed·to·in

**Culture**.·How·efficient·is·the·organization·in·cr
a·desire·to·help?XPERIENCE¶
All·five·of·these·elements·come·together·to·forr
all·other·areas·of·banking·(treasury,·accounting,
important·for·the·efficient·day-to-day·functionir
to·the·customer.·Therefore,·this·study·only·mea
which·the·customer·perceives·the·banks.¶
To·create·the·CEI·scorecard,·Senteo·relied·on·ir
measures·to·create·a·set·of·questions·and·a·mea
Furthermore,·in·order·to·gauge·the·consistency·
banks·selected·for·the·study,·it·would·be·necess
any·single·bank·over·time·and·through·various·
the·CEI·scorecard·filled·out·on·several·different
and·through·multiple·channels,·we·are·able·to·c
of·customer·experience.¶

EXPERIENCE

Unlocking the Secrets of the Past: Text Mining for Historical Documents

(Project Seminar WS 08/09)

Saarland University

16.02.2009

# Corpora of Cultural Heritage (CCH)

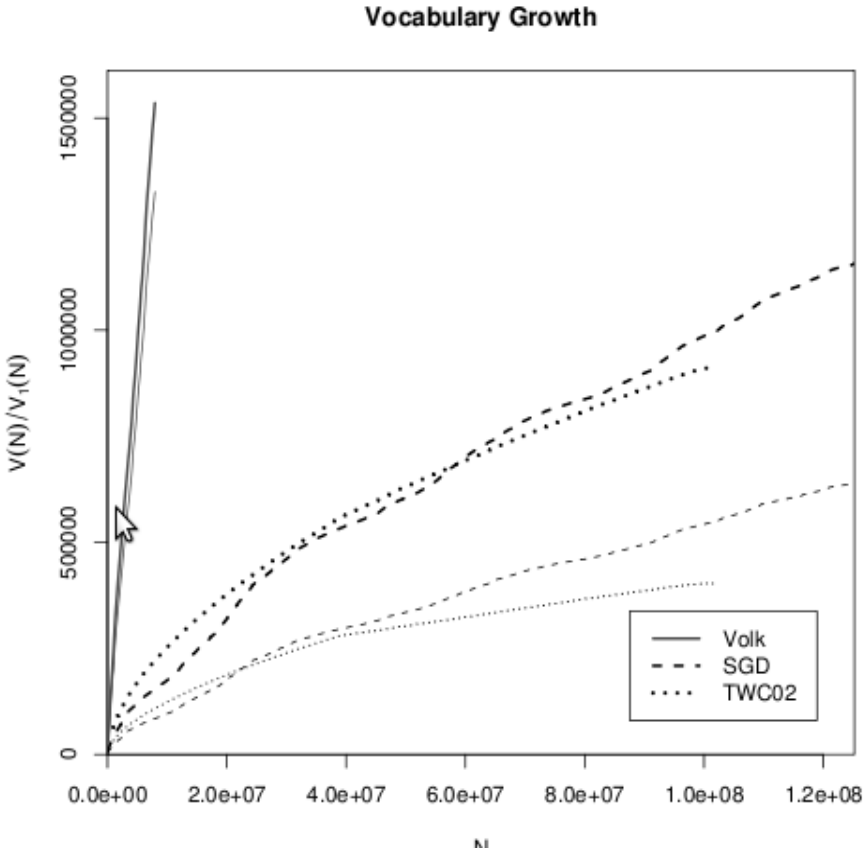
Content:

- Comprises contemporary and historical texts
- Contemporary texts: Acts of Parliement (SGD) (1989-95)
- History texts: selection of daily news papers (DDD) (1918-46)

Properties:

- Spell checkers: DeVries-Te Winkel (for Historical Texts) and the version updated in 1954 (for Contemporary texts)
- Monolingual (ignoring spell checker versions)
- Pilot project for the newspaper archive in National Library
- Collection of fax, typewriter, laser and matrix printer
- Files are stored In PNG format scanned at 600 dpi (grey scales)*
- TICCL project

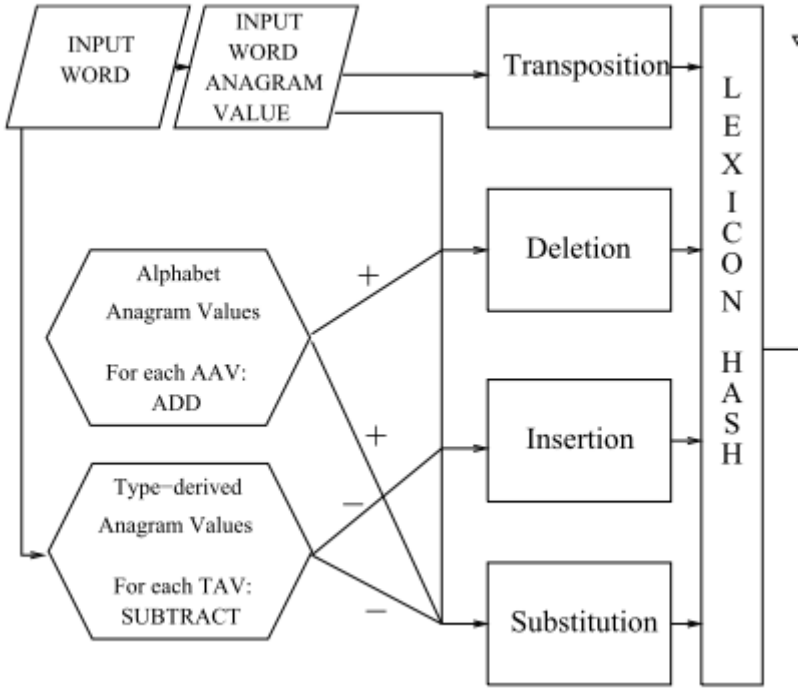
# CCH - Lexical Variation




| 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% |

Word frequency and change of spelling over time play important role in terms of. post correction and spell checking w.r.t. Historical documents.

# TICCL



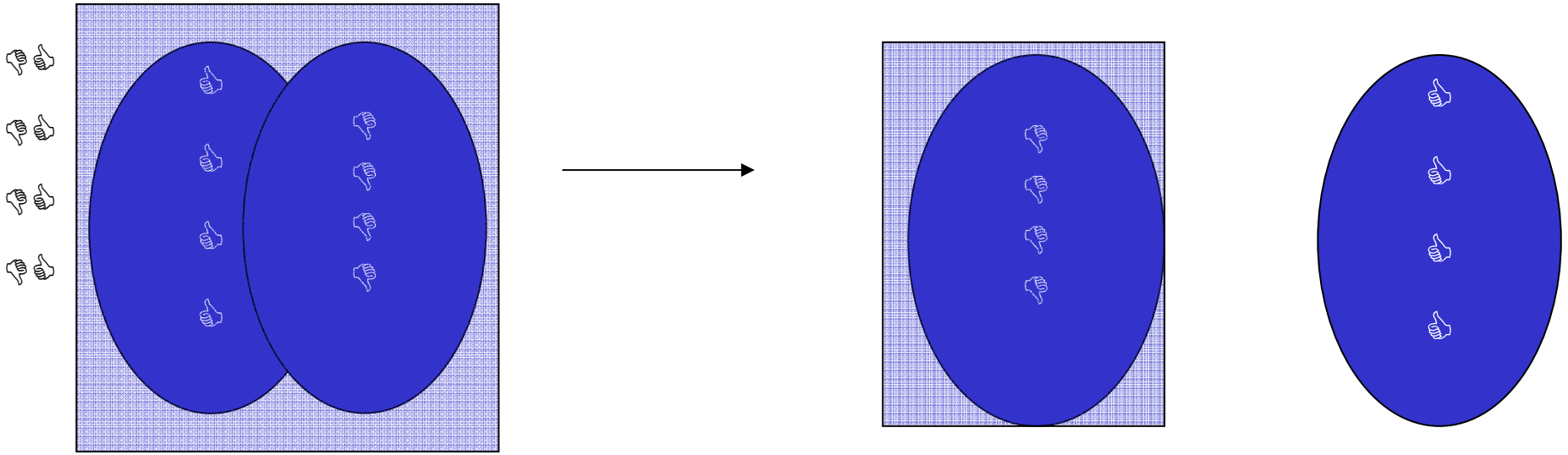

Raynaer, “Corpus Induced Corpus Cleanup”, 2006.

- Typographical variants of words
- Bring the words within bounds (Levensthein distance*: minimum number of operations needed to transform one string into the other )
- Frequency comparison
- Focus word: variants of word strings focused on

# TICCL

- Anagram hashing: the numerical value for a word string is obtained by summing the ISO Latin-1 code value of each character in the string raised to a power n, where n is empirically set at: 5 (e.g. CAT: C + A + T = 675 + 655 + 845 = 6,692,535,156 / TAC: T + A + C = 845 + 655 + 675 = 6,692,535,156)
- For all the variants retrieved, the task we address is determining whether the variant is in fact a perfectly acceptable word in the language in its own right, whether or not this is a perfectly acceptable morphological variant, a perfectly acceptable orthographical variant perhaps to another portion of the language community, viz. English versus American usage – or whether the word variant retrieved constitutes a word form unacceptable to any sizeable portion of the language community. If the latter is the case, we will call the word variant a non-word in that particular language, or typo for short.

# Process



Compare each word with background lexicon

Process the rejected

Focus word

# References


- Reynaert, M. “Non-Interactive OCR Post-Correction for Giga-Scale Digitization Projects”, 2008
- Reynaert M. “Corpus Induced Corpus Cleanup”, 2006
- Mihov, S et a. A Corpus for Comparative Evaluation of OCR Software and Postcorrection Techniques. Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05), pp. 162-166, 2005.
- Manning C. *Foundations of Statistical Natural Language Processing*, Masachuttes and London: MIT Press, 2000