Detection and Correction of OCR errors

By Cornelius Leidinger

TICCL

Text-Induced Corpus Clean-up - TICCL

By Martin Reynaert

http://ilk.uvt.nl/downloads/pub/papers/CICLING08.TICCL.MRE.postpublication.pdf

Text collections

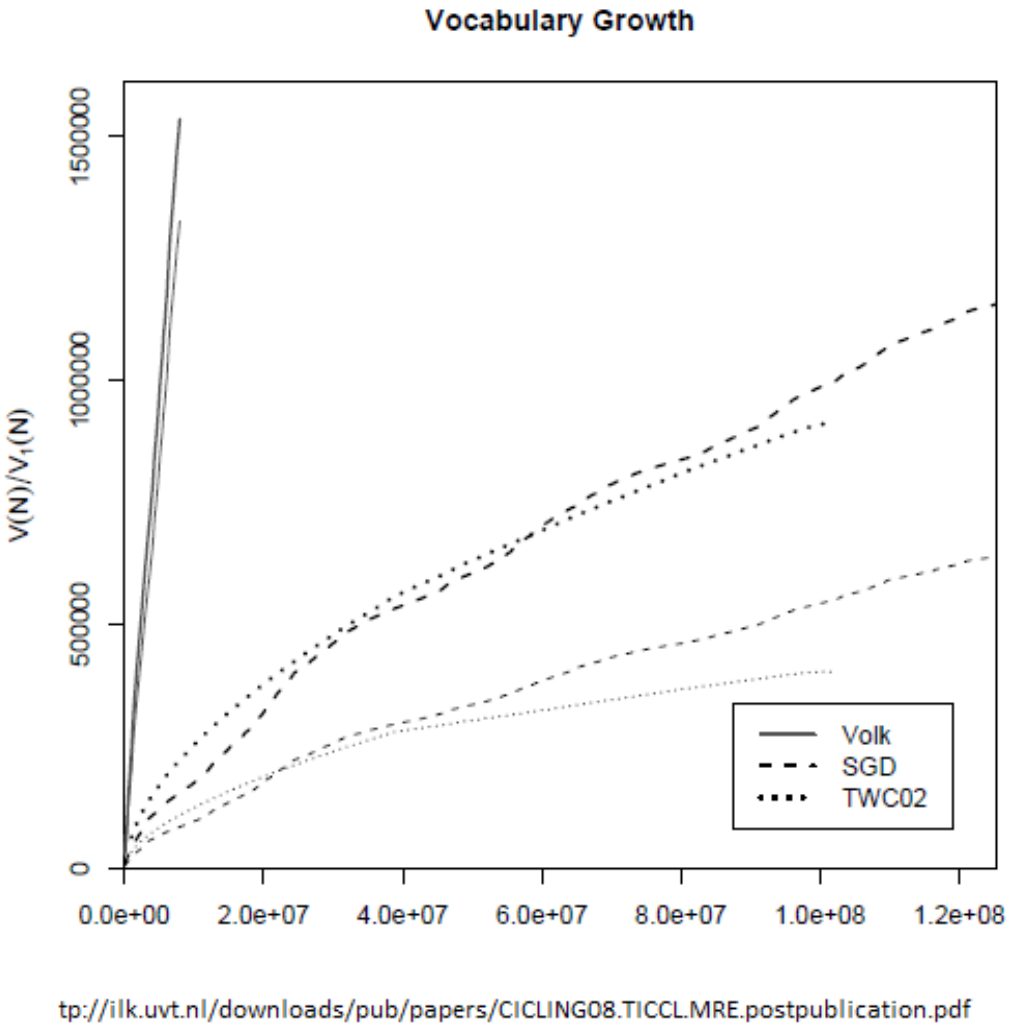
Contemporary collection: The published Acts of Parliament(1989-1995) of The Netherlands
As 'Staten-Generaal Digitaal'(SGD)

Historical collection: The 'Database Digital Daily Newspaper'(DDD) (1918-1946)
In old Dutch spelling 'De Vires-Te Winkel'

OCR systems

Commercial:
Abbyy FineReader, Nuance OmniPage

Open-source:
previously named Tesseract, now called OCRopus



tp://ilk.uvt.nl/downloads/pub/papers/CICLING08.TICCL.MRE.postpublication.pdf

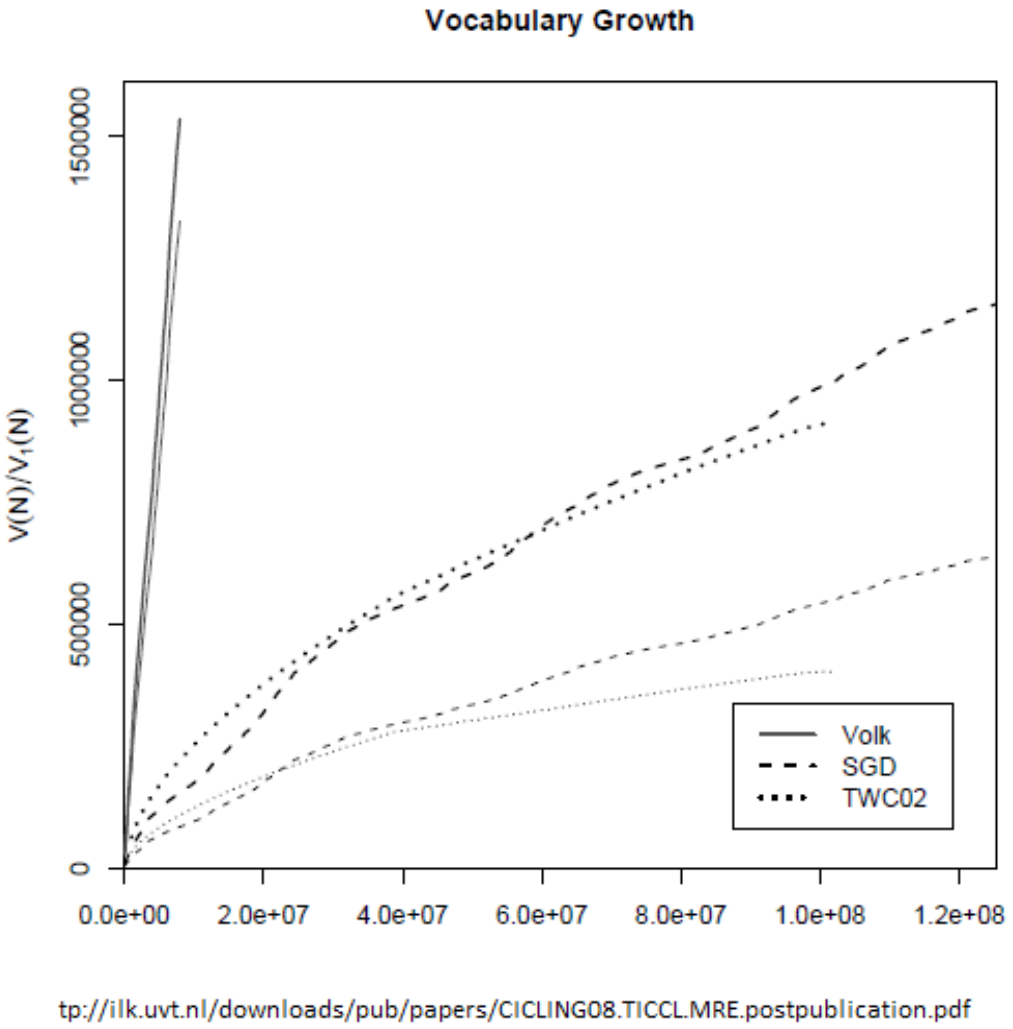
- TWC02: one year newspaper corpus, covering 2002 (born-digital)
- SGD: Staten-Generaal Digital
- Het Volk: a newspaper in the DDD

Exact values

Table 1. Corpora Statistics: Corpus, language (CD: Contemporary Dutch, HD: Historical Dutch), origin: born-digital (BD) or OCRed (OCR), number of word tokens, number of word types, type-token ratio (TTR)

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%

http://ilk.uvt.nl/downloads/pub/papers/CICLING08.TICCL.MRE.postpublication.pdf



tp://ilk.uvt.nl/downloads/pub/papers/CICLING08.TICCL.MRE.postpublication.pdf

- TWC02: one year newspaper corpus, covering 2002 (born-digital)
- SGD: Staten-Generaal Digital
- Het Volk: a newspaper in the DDD

Example for word 'regeering'

Table 2. Twenty variants (multiple non-contiguous errors) for the focus word 'regeering' produced by apparent random substitutions of the focus word's last character(s), besides the recurring substitution of an 'e' by 'c'

regecrin	regecrinc	regecring'	regecrinj	regecrino	regecrins	regecrinz
regecrin-	regecrincr	regecrini	regecrink	regecrinp	regecrint	regecrinü
regecrina	regecrinf	regecrinic	regecrinn	regecrinr	regecrinx	

http://ilk.uvt.nl/downloads/pub/papers/CICLING08.TICCL.MRE.postpublication.pdf

Insertion, Deletion, Substitution

Insertion: 'regeering' → 'regeeriing'

Deletion: 'regeering' → 'regeerng'

Substitution: 'regeering' → 'regecring'

Transposition, Multi-C, Multi-NC

Transposition: 'regeering' → 'regeeirng'

Multi-C: multiple contiguous error
'regeering' → 'regeermg'

Multi-NC: multiple non-contiguous error
'regeering' → 'rcgecring'

Statistics

Table 3. SGD 1989-1995: overview and statistics per LD of error-types encountered in a sample of 5,047 non-word variants

Category	LD 1	LD 2	LD 3	LD 4	LD 5	LD 6	LD 7	Total	%
deletion	221	10	3	1				235	4.66
insertion	1,980	27	6	11				2,024	**40.10**
substition	1,065	49	37	3		1		1,155	**22.89**
transposition		26						26	0.52
multi-C		722	30	10	1	1		779	**15.46**
multi-NC		303	271	101	22	5	2	710	**14.09**
run-on words	67							67	1.33
split word	32							32	0.63
TOTAL	3,380	1,138	347	126	23	7	2	5,047	
%	66.98	22.55	6.88	2.50	0.46	0.14	0.04		100.00

http://ilk.uvt.nl/downloads/pub/papers/CICLING08.TICCL.MRE.postpublication.pdf

Statistics

Table 4. DDD 'Het Volk' 1918: overview and statistics per LD of error-types encountered in a sample of 3,799 non-word variants

Category	LD 1	LD 2	LD 3	LD 4	LD 5	LD 6	Total	%
deletion	31	27	1	12			71	1.87
insertion	133	25	3	4			165	4.34
substition	575	276	109	2			962	**25.32**
transposition		3					3	0.08
multi-C		203	193	9	2	1	412	**10.85**
multi-NC		810	1,277	77	15	3	2,182	**57.44**
run-on words	2						2	0.05
split word	2						2	0.05
TOTAL	743	1,344	1,583	104	17	4	3,799	
%	19.56	35.38	41.67	2.74	0.45	0.11		100.0

http://ilk.uvt.nl/downloads/pub/papers/CICLING08.TICCL.MRE.postpublication.pdf

TICCL

Unsupervised, scalable, fully automatic
– no training, largely language-independent.

Anagram Hashing

Use a bad hashing function to get all word strings in the corpus, that have the same subset of characters.

Assign them a large number as index

Nummerical value for a word string

For characters use ISO Latin-1 code value

A → 41 → 65

Z → 5A → 90

a → 61 → 97

z → 7a → 122

Example

'regeering' =
114^5
+ 101^5
+ 103^5
+ 101^5
+ 101^5
+ 114^5
+ 105^5
+ 110^5
+ 103^5
= large number

Anagrams

Anagrams will be identified through their common numerical value produced by the bad hash function. These are called 'angram hash'.

The unique numerical values are called 'anagram values' (AV) and 'anagram keys'

AnagramValueAlphabet

This Alphabet contains singel values that refer to a single, a combination of two or three characters (more are possible)

a-zA-Z
aa, ab,ba, ...
aaa, aab, aba, baa, ...

FocusWordAlphabet

Contains all AnagramValues present in the focus word

How it works

For substitutions:

Substract value from FocusWordAlphabet
Add value from AnagramValueAlphabet

Example

Focus word 'regeering'
Minus AV 'e'
Plus AV 'c'

OCR-errors: 'rcgeering', 'regcering' and 'regecring'

Insertions

Also substitution:

Subtract zero
Add a value from AnagramValueAlphabet

Deletions

Also substitution:

Subtract vlaue from FocusWordAlphabet
Add zero

Transposition

The value doesn't change

Execution

The system do all substitutions for all values of AnagramValueAlphabet and all values of FocusWordAlphabet for a FocusWord and so it retrieves all focus word variants up to LD 3

Normalization

Up to now the SGD had 187 different characters

All text is lowercased

All punctuation marks, except hyphens and apostrophes, are rewritten as a '2'

All numbers are rewritten as a '3'

Uppercased diacritic characters are rewritten as '4' (Ö,Ü,Ä)

Lowercased diacritic characters are rewritten as '5' (ö,ü,ä)

After normalization there are 32 characters left

Result

It returns the variants in pairs:

(focusword, retrieved variant)

Table 5. Overview of the SGD and DDD focus words and their observed numbers of variants which constitute the evaluation sets. Capitalized words are proper names

Focus SGD	#	Focus 'Het Volk'	#
Achtienribbe-Buijs	23	Amsterdam	307
Amsterdam	43	Annexionisten	20
Bolkestein	18	België (Belgium)	104
Jorritsma-Lebbink	33	Bismarck	10
Nieuwenhoven	22	Compiègne	3
Rotterdam	47	Hindenburg	32
Wolffensperger	25	Nederlandsche (Dutch)	572
belasting (tax)	36	Posthuma	264
belastingen (taxes)	56	Richthofen	7
belastingplichtige (taxable person)	41	Trotzky	45
belastingplichtigen (taxable persons)	37	Wilhelmina	42
doelstelling (aim)	82	Zeeuwsch-Vlaanderen	19
doelstellingen (aims)	58	belasting (tax)	102
evaluatie (evaluation)	44	belastingen (taxes)	34
faciliteiten (facilities)	27	distribueeren (to distribute)	52
goedkeuring (approval)	36	eenheidsworst (unity sausage)	21
inkomstenbelasting (income tax)	81	regeering (government)	1468
motorrijtuigenbelasting (motor vehicle tax)	70		
studiefinanciering (study financing)	93		
vennootschapsbelasting (corporate tax)	52		

Evaluation

True Positives, False Positives, False Negatives

Recall, Precision

F-score

Table 6. Overview of the SGD performance scores

Measured at	Items retrieved			At LD			Cumul. to LD		
LD	TP	FN	FP	R	P	F	CR	CP	CF
LD 1	466	4	7	0.991	0.985	0.988	0.991	0.985	0.988
LD 2	284		129	1.000	0.688	0.815	0.995	0.847	**0.915**
LD 3	106	1	525	0.991	0.168	0.287	0.994	0.564	0.720
LD 4	11	11	133	0.500	0.076	0.133	0.982	0.522	0.682
LD 5	1	6	22	0.143	0.043	0.067	0.975	0.515	0.674

Table 7. Overview of the DDD performance scores

Measured at	Items retrieved			At LD			Cumul. to LD		
LD	TP	FN	FP	R	P	F	CR	CP	CF
LD 1	380	6	4	0.984	0.990	0.987	0.984	0.990	0.987
LD 2	1112	9	114	0.992	0.907	0.948	0.990	0.927	**0.957**
LD 3	1558	3	613	0.998	0.718	0.835	0.994	0.807	0.891
LD 4	25	9	46	0.735	0.352	0.476	0.991	0.798	0.884