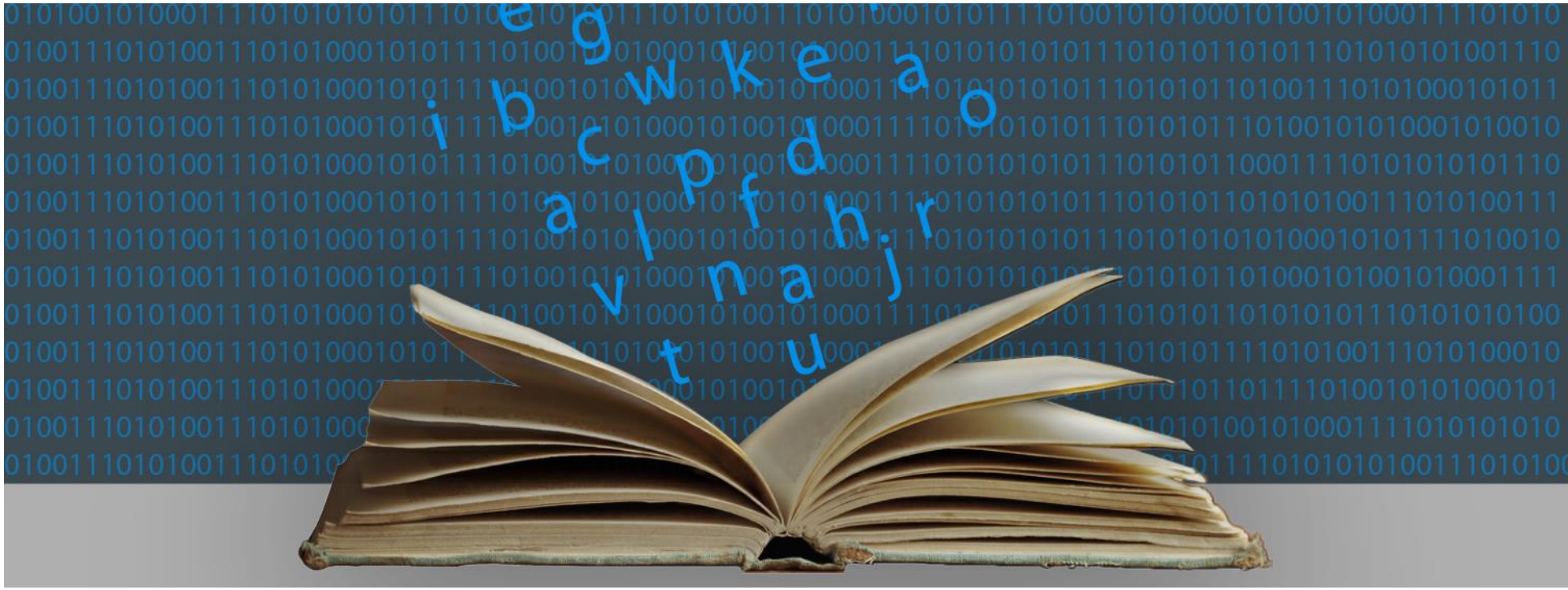
Examining a multi-layered approach for classification of OCR quality without Ground Truth



Mirjam Cuper, 17 March 2021

KB national library of the netherlands

Introduction

- Mirjam Cuper
- Data scientist
- KB - National Library of the Netherlands
- Digital heritage

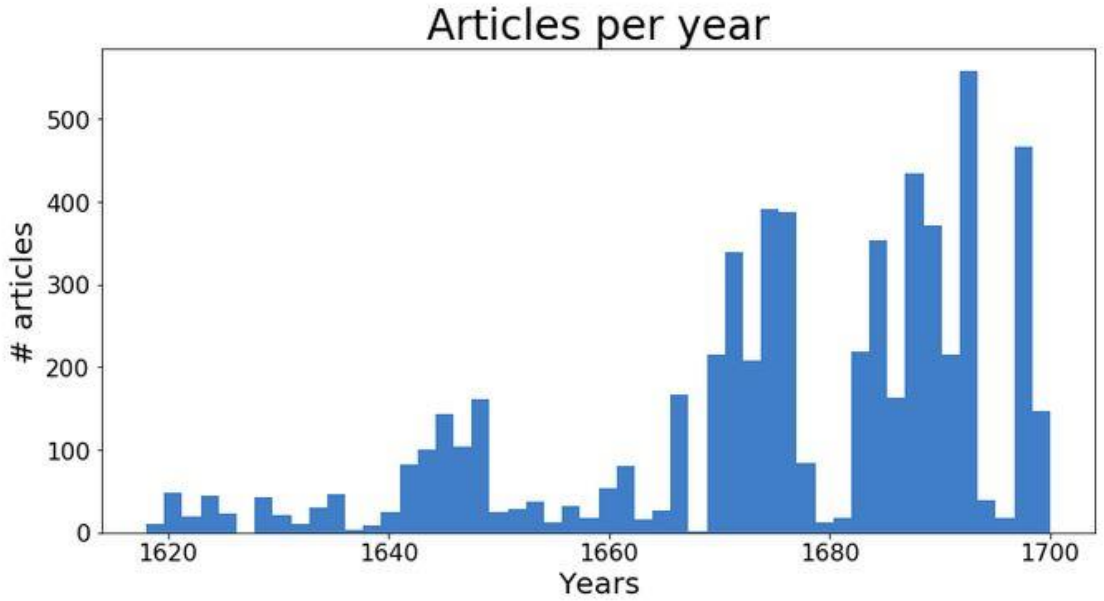
Overview

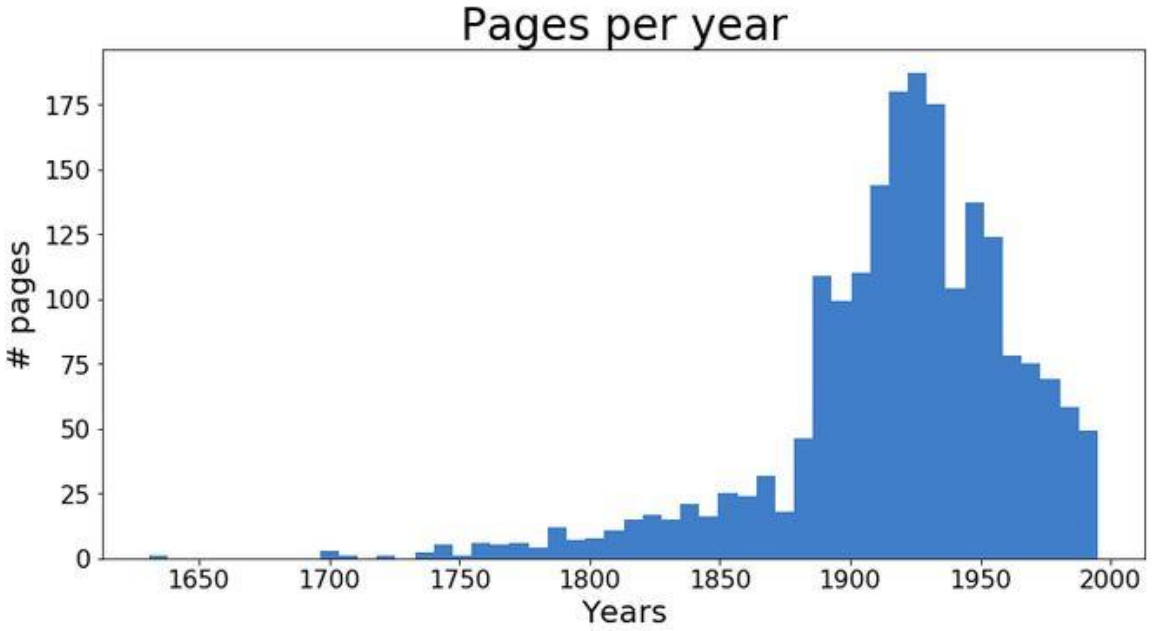
- Dataset
- Problem description
- Multi-layered approach
- First results
- Future work

Sample data

- Based on:
- +/- 30.000 news articles
- 2.000 newspaper pages

- Ground truth + OCR





WHAT IS THE PROBLEM?

OCR errors

WEST-INDIEN.

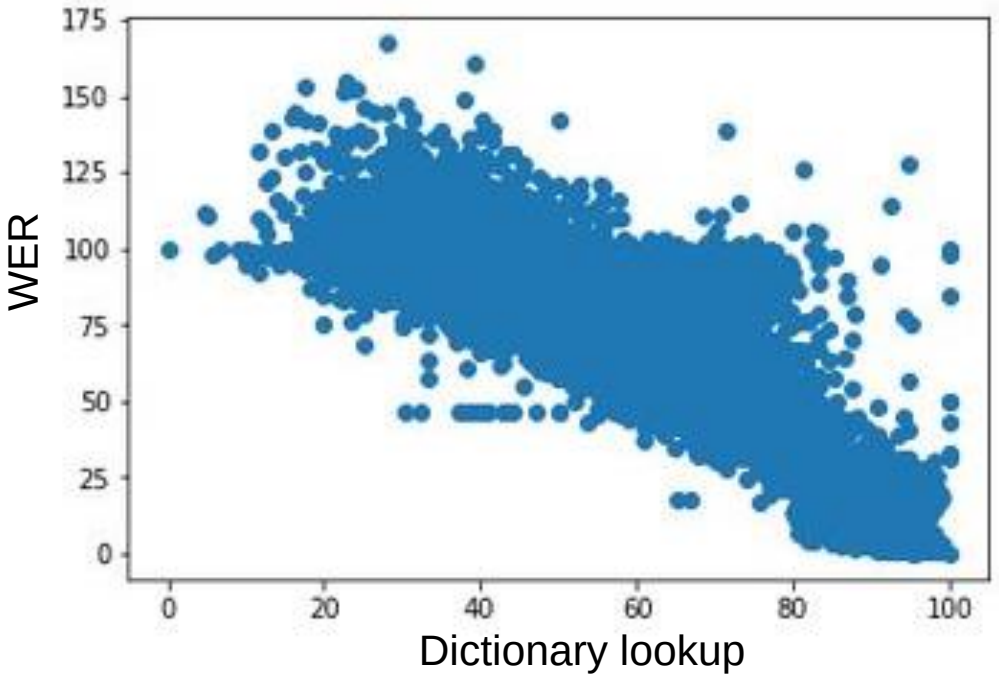
Uriname den 28 December 1682. Het Schip de Harderin, Schipper Dirck Barentſz. Kock, is alhier behouden aengekomen. Een dag a twee nae deſe arriveerde hier oock Schipper Tobias Adriaenſz., dewelcke aen St. Jago 21 Paerden ingenomen, hier levendigh overgevoert, en tot een hooge Prijs verkocht heeft. Noch een ander, zijnde een Engels Schipper, heeft mede, van Nieuw Engelandt komende, 20 Paerden hier gebracht. Met de Naturellen is het noch als vooren. Men heeft groot gebreck van Volck en Slaven. Anders ſtaet hier alles wel.

WEST-INDIEN.
SUrinimÃ¨ den 28 December 1682. Het Schip de Harderin, , Schipper Dirck Barentfz. Koek ,is alhier behouden aengekomen. Een dag a twee nae defe arriveerde hier oock Schipper Tobias Adriaenfz., dewelcke aen St. Jago2i Paerden ineenomenhier levendigh overgevoert, en tot eenhooge Prijs verkocht heeft. Noch een ander, zijnde een Engels Schipper, heeft mede, van Nieuw Engelandt komendeÂ» 20 Paerden hier gebracht. Met de Naturellen is het noch als vooren. Men heeft groot gebreck van Volck en Slaven. Anders ftaet hier alles wel.

Dictionary lookup - overview

- Correlation WER
- Relatively good result

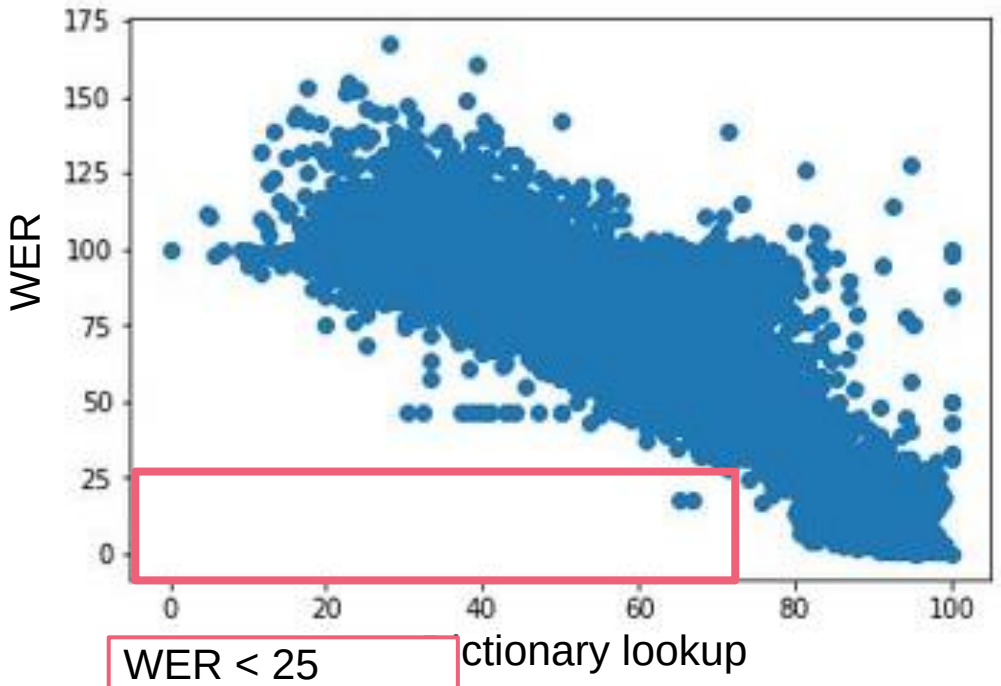
Correlation Dictionary lookup and WER



Dictionary lookup - overview

- Correlation WER
- Relatively good result

Correlation Dictionary lookup and WER

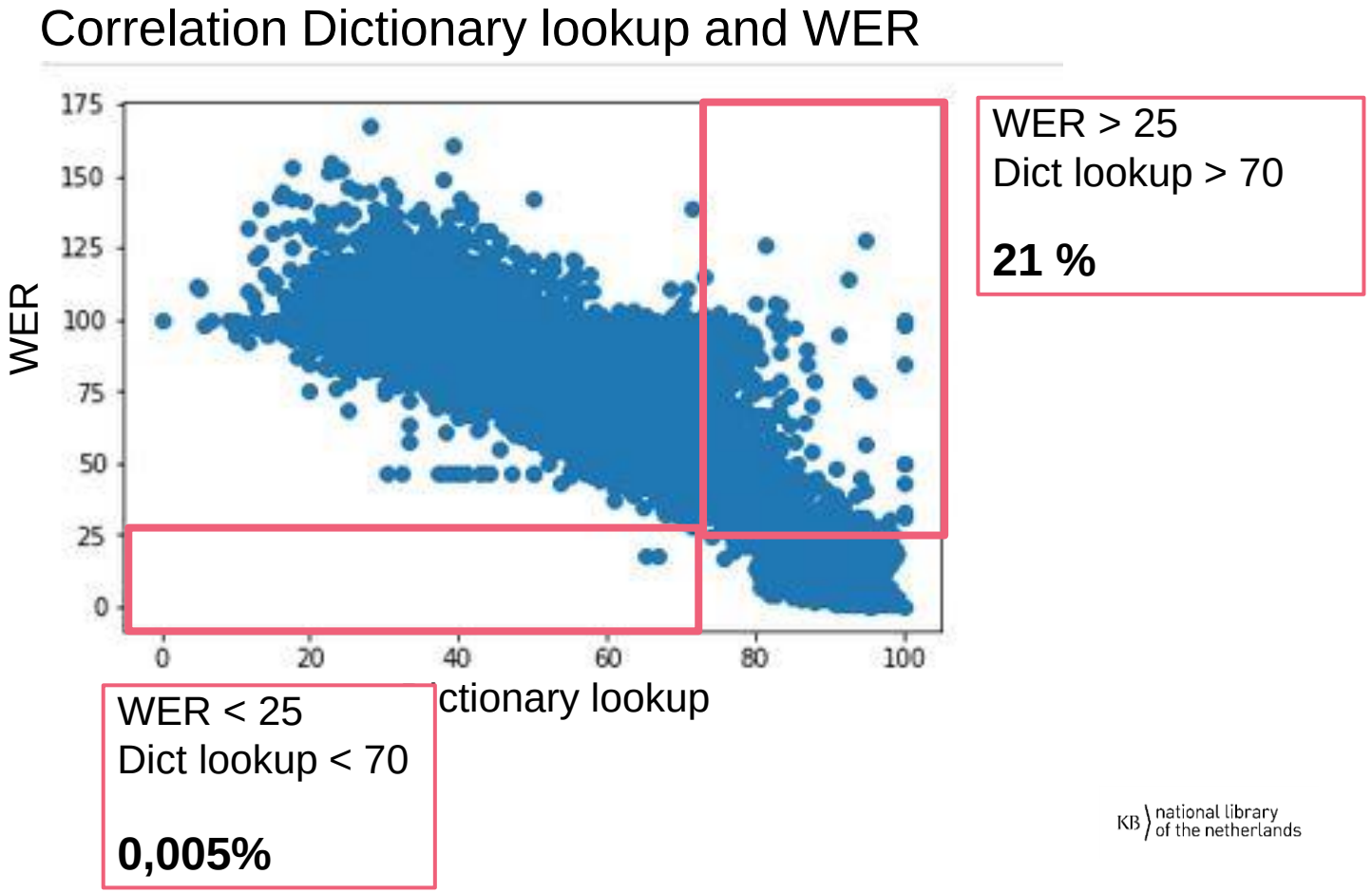


WER < 25
Dict lookup < 70

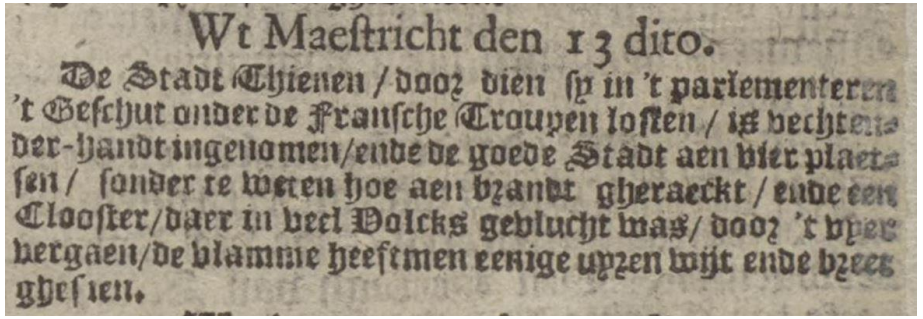
0,005%

Dictionary lookup - overview

- Correlation WER
- Relatively good result



Missing text

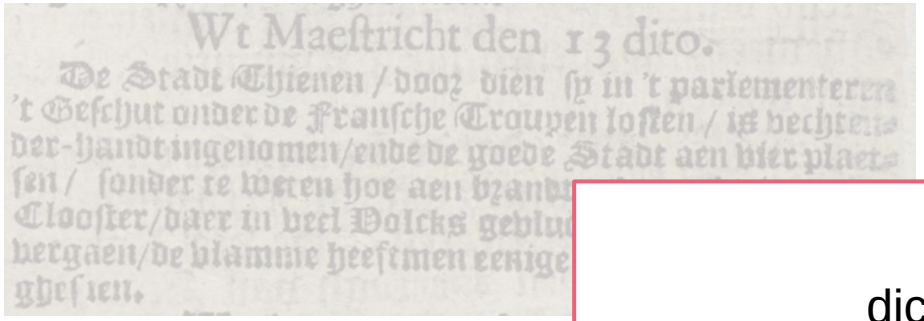


Wt Maeſtricht den 13 dito.

De Stadt Thienen / dooz dien ſy in 't parlementeren 't Geſchut onder de Franſche Trouppen loſten / is bechtender-handt ingenomen/ende de goede Stadt aen vier plaetſen / ſonder te weten hoe aen brandt geraeckt / ende een Clooſter/daer in veel Volcks geblucht was/ dooz 't vyer vergaen/de vlamme heeftmen eenige uyren wijt ende breet gheſien.

Wt Maestricht den 13 dito.
g_<Â«^taÃ¶tiâ,¬i)ieiien/boo? bien
fptn'tpartementer

Missing text



Wt Maestricht den 13 dito.
g_<Â«^taÃ¶tiâ,¬i)ieiien/boo? bien
fptn'tpartementer

dictionary:

75% of words found

Missing text

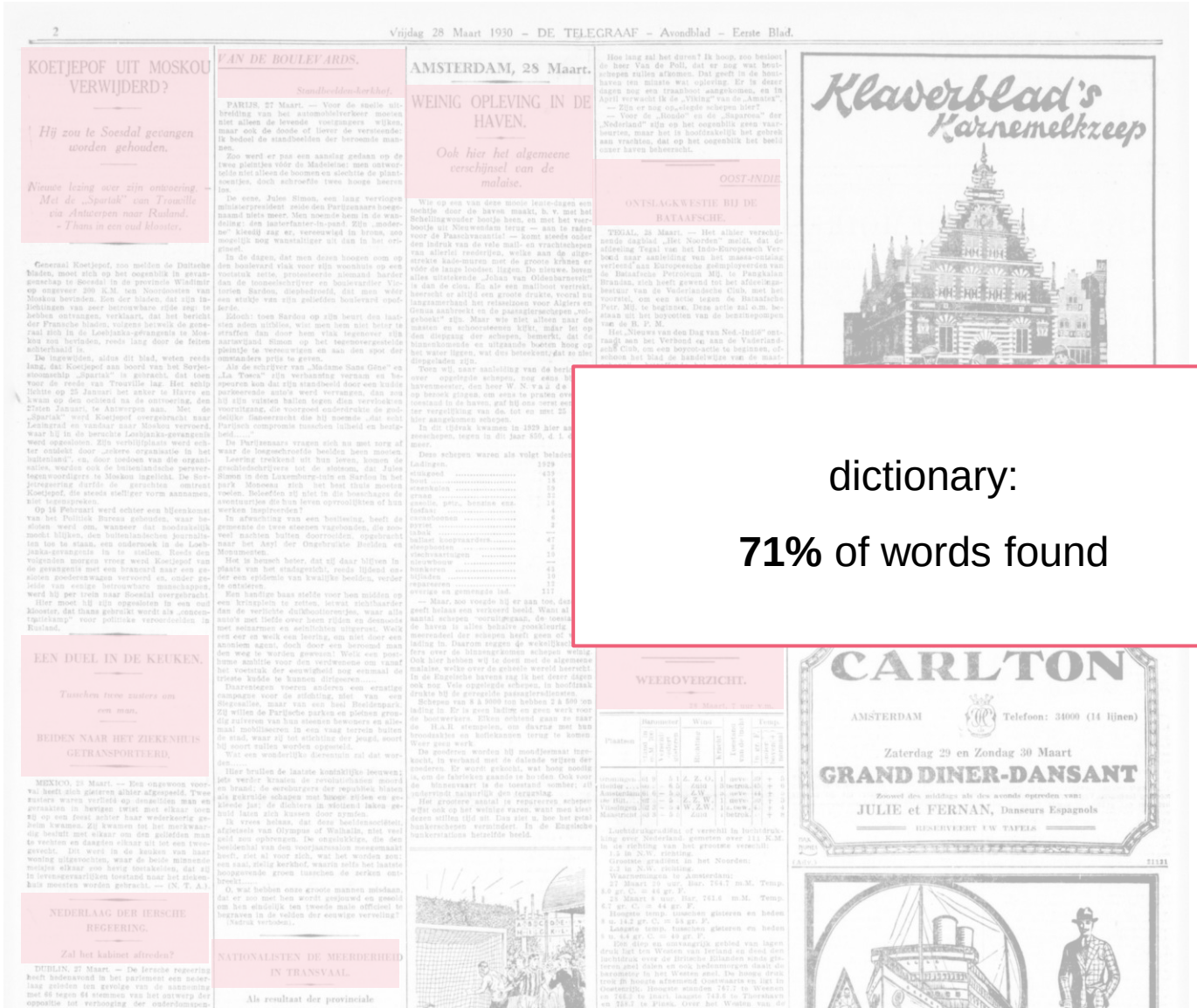
KOETJEPOF UIT MOSKOU VERWIJDERD? Hij zou te Soesdal gevangen worden gehouden. Nieuwe lezing over zijn ontvoering. — Met de „Spartak" van Trouville via Antwerpen naar Rusland. - Thans in een oud klooster. I ■ kou i lse inKe.'. - ovjet ■ schip ■ ■ ' ■ •i het rgnnl • ■ Loeb ■ ■es.lal o. » .st tiians g EEN DUEL IN DE KEUKEN. Tusschen twee zusters om een man. BEIDEN NAAR HET ZIEKENHUIS GETRANSPORTEERD. •.' i '\ ■ •ti man en - I '- - I NEDERLAAG DER IERSCHE REGEERING. Zal het kabinet aftreden?

VAN DE BOULEVARDS. Standbeelden-kerkhof. ■ ogert w loal de si. nen. ■ twor ■ los IV» • ■ In d ■ ■ .lel 1 ' naar - ■ - ■ ■ ■ campagne voor de i een tl » ■ ■ NATIONALISTEN DE MEERDERHEID IN TRANSVAAL. Als resultaat der provinciale verkiezingen. JOH • zingen voor den provincialen raad van 'I • • kaaneche partij en Z I.ab- Twee uitslagen zijn nog niet I.- i zul ■ ■

WEINIG OPLEVING IN DE HAVEN. Ook hier het algemeene verschijnsel van de malaise. maakt, b. v. met het ■ rat — aan te raden ■ hepen ida -ttge ' alles ii" een n . ral nu •rs en ■ir dB ki.'kt. maar let np i •n de -trteatjjttl U den A ' . .i de Poll er den Maan ■ ■ 19:» ïtso tti 27 II 4 8 | i S is 61 1 13 . — 1 14 3 117 141 II U's: .it al Is be; - -r_an, de toestand In «een of « , e cij - - naar ■ hun . hotu en In aan i ;. het oogenblik het beeld

OOST-INDIË. ONTSLAGKWESTIE BIJ DE BATAAFSCHE. ■ - . H. i. M. raadt -chf C en, of i ZUID-SUMATRA LANDBOUW- EN NIJVER HEIDSVEREENIGING. ■ J. 1! M. J. 1 i i ■ WEEROVERZICHT. ' . A ' ■ ■ LiUkt'i. tn 7. 55 gr. Woedende

Missing text



KOETJEPOF UIT MOSKOU VERWIJDERD?

Hij zou te Soesdal gevangen worden gehouden.

Nieuwe lezing over zijn ontvoering. — Met de „Spartak" van Trouville via Antwerpen naar Rusland. - Thans in een oud klooster.

EEN DUEL IN DE KEUKEN.

Tusschen twee zusters om een man.

BEIDEN NAAR HET ZIEKENHUIS GETRANSPORTEERD.

NEDERLAAG DER IERSCHE REGEERING.

Zal het kabinet aftreden?

VAN DE BOULEVARDS.

Standbeelden-kerkhof.

AMSTERDAM, 28 Maart.

WEINIG OPLEVING IN DE HAVEN.

Ook hier het algemeene verschijnsel van de malaise.

OOST-INDIË.

ONTSLAGKWESTIE BIJ DE BATAAFSCHE.

NATIONALISTEN DE MEERDERHEID IN TRANSVAAL.

Als resultaat der provinciale

WEERoverzicht.

Klaverblad's Karnemelkzeep

CARLTON

AMSTERDAM

Telefoon: 34000 (14 lijnen)

Zaterdag 29 en Zondag 30 Maart

GRAND DINER-DANSANT

JULIE et FERNAN, Danseurs Espagnols

RESERVEERT UW TAFELS

dictionary:

71% of words found

KOETJEPOF UIT MOSKOU VERWIJDERD? Hij zou te Soesdal gevangen worden gehouden. Nieuwe lezing over zijn ontvoering. — Met de „Spartak" van Trouville via Antwerpen naar Rusland. - Thans in een oud klooster. I ■ kou i lse inKe.'. - ovjet ■ schip ■ ■ ' ■ •i het rgnnl • ■ Loeb ■ ■es.lal o. » .st tiians g EEN DUEL IN DE KEUKEN. Tusschen twee zusters om een man. BEIDEN NAAR HET ZIEKENHUIS GETRANSPORTEERD. •.' i '\ ■ •ti man en - I '- - I NEDERLAAG DER IERSCHE REGEERING. Zal het kabinet aftreden?

VAN DE BOULEVARDS. Standbeelden-kerkhof. ■ ogert w loal de si. nen. ■ twor ■ los IV» • ■ In d ■ ■ .lel 1 ' naar - ■ - ■ ■ ■ campagne voor de i een tl » ■ ■ NATIONALISTEN DE MEERDERHEID IN TRANSVAAL. Als resultaat der provinciale verkiezingen. JOH • zingen voor den provincialen raad van 'I • • kaaneche partij en Z I.ab- Twee uitslagen zijn nog niet I.- i zul ■ ■

WEINIG OPLEVING IN DE HAVEN. Ook hier het algemeene verschijnsel van de malaise. maakt, b. v. met het ■ rat — aan te raden ■ hepen ida -ttge ' alles ii" een n . ral nu •rs en ■ir dB ki.'kt. maar let np i •n de -trteatjjttl U den A ' . .i de Poll er den Maan ■ ■ 19:» ïtso tti 27 II 4 8 | i S is 61 1 13 . — 1 14 3 117 141 II U's: .it al Is be; - -r_an, de toestand In «een of « , e cij - - naar ■ hun . hotu en In aan i ;. het oogenblik het beeld

OOST-INDIË. ONTSLAGKWESTIE BIJ DE BATAAFSCHE. ■ - . H. i. M. raadt -chf C en, of i ZUID-SUMATRA LANDBOUW- EN NIJVER HEIDSVEREENIGING. ■ J. 1! M. J. 1 i i ■ WEEROVERZICHT. ' . A ' ■ ■ LiUkt'i. tn 7. 55 gr. Woedende

Word importance

dictionary:

76% of words found

Een handleidinq of een gebriiksaanwijzing is een schriftelljke instrvctie die met een qroduct meegeleverd wordt.

De inhovd van een handlejding is:

hoe een produot te assernbleren;
hoe een product te reporeren;
hoe een pnodvct te instaileren;
hoe een product gebruikt dient te worden;
hoe een poduct te onderhovden;
hoe de instelllngen van een product aan te passen;
hoe een storimg op te lossen;
hoe een produci niet te misbruiken;
hoe comtact op te nemen voor servioe.

Word importance

dictionary:

76% of words found

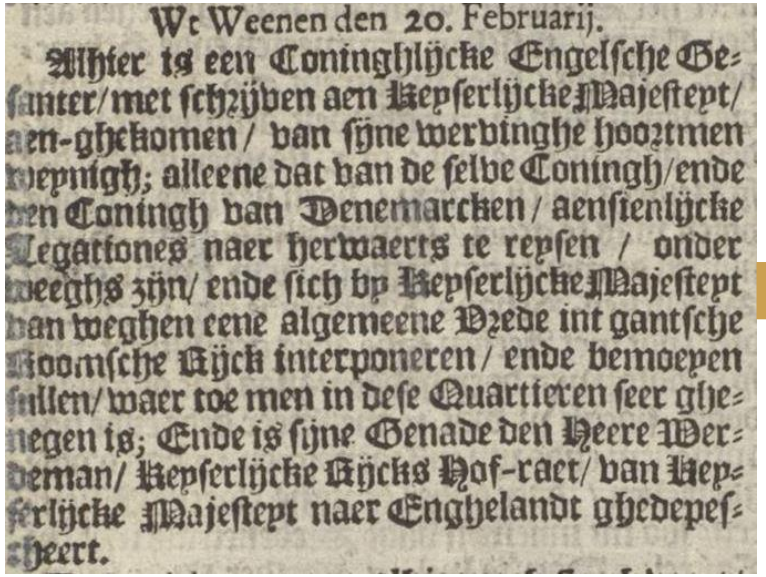
Een handleidinq of een gebriiksaanwijzing is een schriftelljke instrvctie die met een qroduct meegeleverd wordt.

De inhovd van een handlejding is:

hoe een produot te assernbleren;
hoe een product te reporeren;
hoe een pnodvct te instaileren;
hoe een product gebruikt dient te worden;
hoe een poduct te onderhovden;
hoe de instelllngen van een product aan te passen;
hoe een storimg op te lossen;
hoe een produci niet te misbruiken;
hoe comtact op te nemen voor servioe.

Time period

Article, 1636 – Ground Truth text

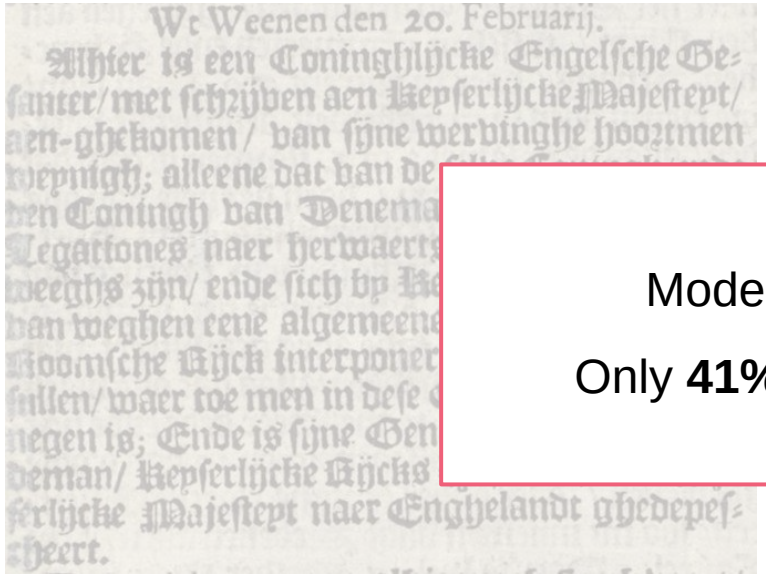


Wt Weenen den 20. Februarij.
Alhier is een Coninghlijcke Engelſche Geſanter/met ſchrijven aen Keyſerlijcke Majeſteyt/aen-ghekomen/ van ſijne werdinghe hoortmen weynigh; alleene dat van de ſelve Coningh/ende den Coningh van Denemarcken/ aenſienlijcke Legationes naer herwaerts te reyſen / onder weeghs zijn/ ende ſich by Keyſerlijcke Majeſteyt van weghen eene algemeene Vrede int gantſche Roomſche Rijck interponeren/ ende bemoeyen ſullen/waer toe men in deſe Quartieren ſeer ghenegen is; Ende is ſijne Genade den Heere Werdeman/ Keyſerlijcke Rijcks Hof-raet/ van Keyſerlijcke Majeſteyt naer Enghelandt ghedepeſcheert.

Wt Weenen den 20. Februarij. Alhier is een Coninghlijcke Engelsche Gesanter, met schrijven aen Keyserlijcke Majesteyt, aen-ghekomen, van sijne wervinghe hoortmen weynigh; alleene dat van de selve Coningh, ende den Coningh van Denemarcken, aensienlijcke Legationes naer herwaerts te reysen, onder weeghs zijn, ende sich by Keyserlijcke Majesteyt van weghen eene algemeene Vrede int gantsche Roomsche Rijck interponeren, ende bemoeyen sullen, waer toe men in dese Quartieren seer ghenegen is; Ende is sijne Genade den Heere Werdeman, Keyserlijcke Rijcks Hof-raet, van Keyserlijcke Majesteyt naer Enghelandt ghedepescheert. De wervingen gaen alhier noch sterck voort, ende wert veel Meel ende Coorn, tot provianderinghe vande Keyserlijcke Armaden, in \'t Rijck ghesonden.

Time period

Article, 1636 – Ground Truth text



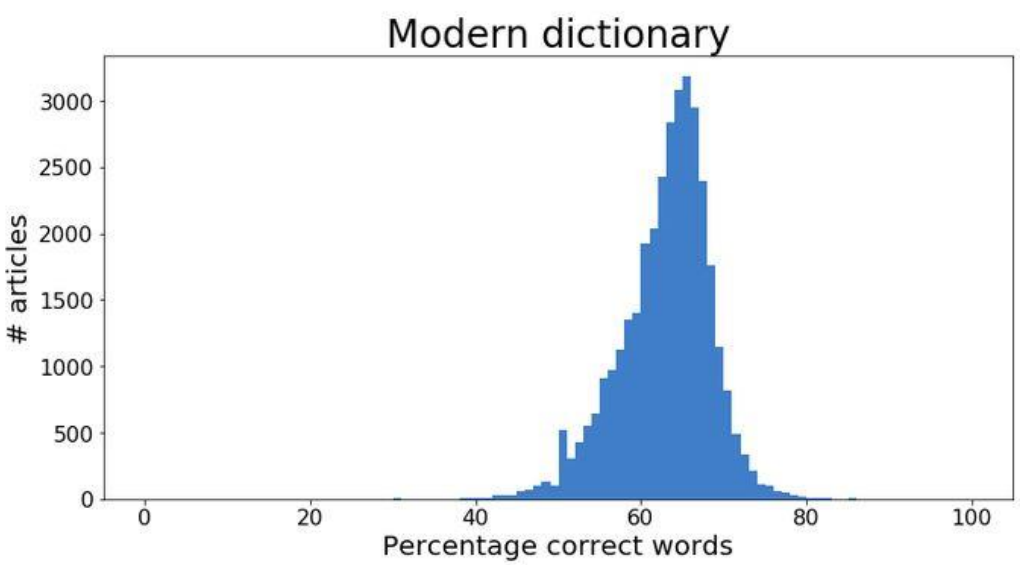
Wt Weenen den 20. Februarij. Alhier is een Coninghlijcke Engelsche Gesanter, met schrijven aen Keyserlijcke Majesteyt, aen-ghekomen, van sijne wervinghe hoortmen weynigh; alleene dat van de selve Coningh, ende den Coningh van sienlijcke Legationes naer herwaerts te ghs zijn, ende sich by Keyserlijcke Majesteyt algemeene Vrede int gantsche Roomsche ende bemoeyen sullen, waer toe men in er ghenegen is; Ende is sijne Genade den Keyserlijcke Rijcks Hof-raet, van Keyserlijcke helandt ghedepescheert. De wervingen erck voort, ende wert veel Meel ende Coorn, tot provianderinghe vande Keyserlijcke Armaden, in \'t Rijck ghesonden.

Modern dictionary:

Only **41%** of words found

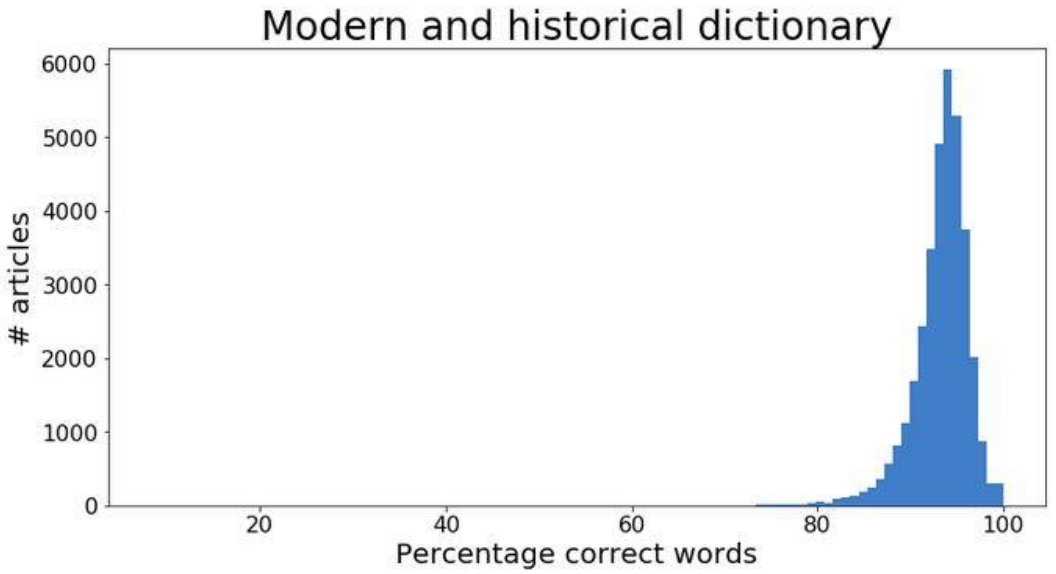
Time period

17th century articles: ground truth data



Mean: 63%

Median: 64%



Mean: 93%

Median: 94%

Language

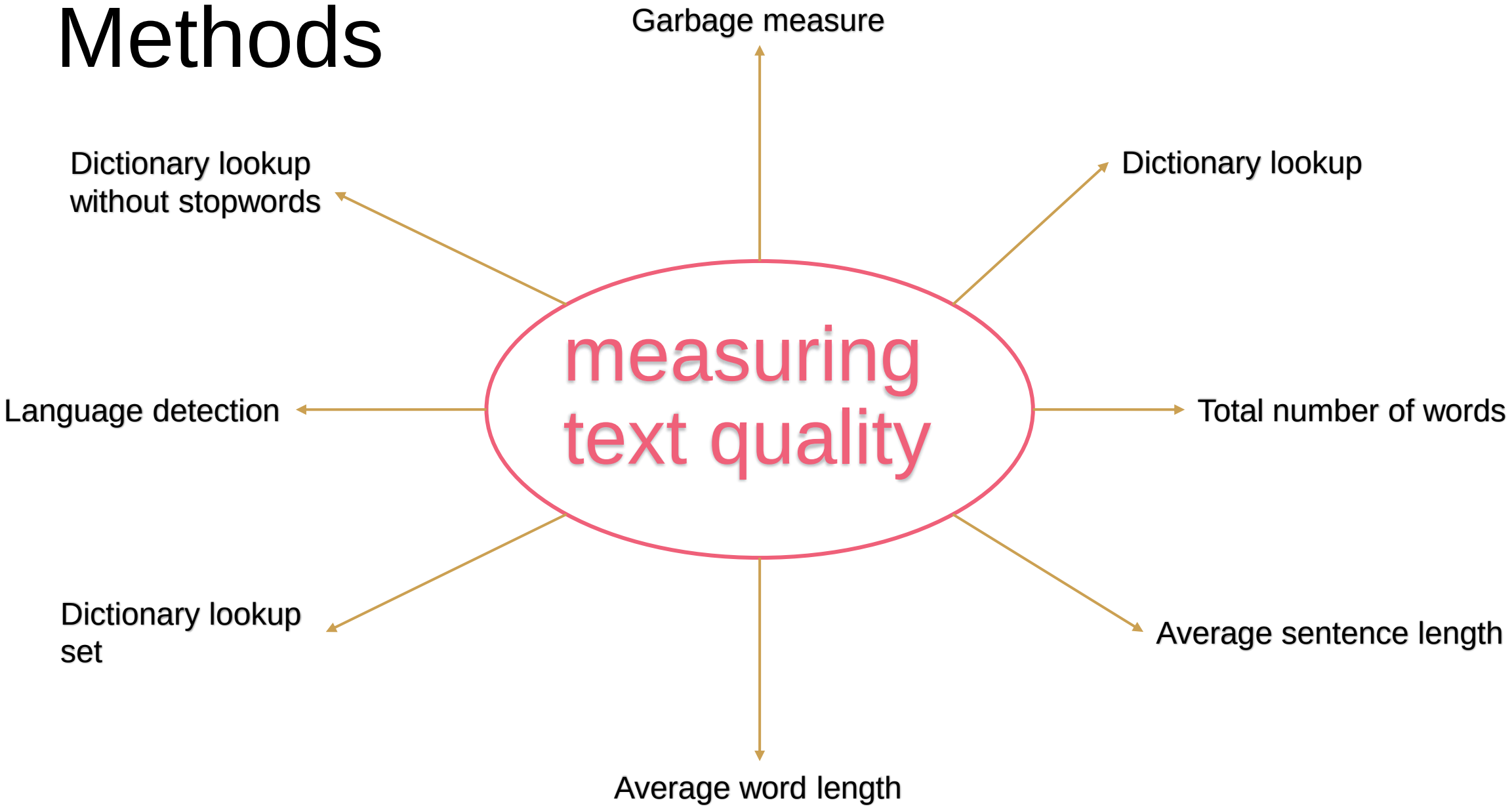
Combined dictionary:

86% of words found

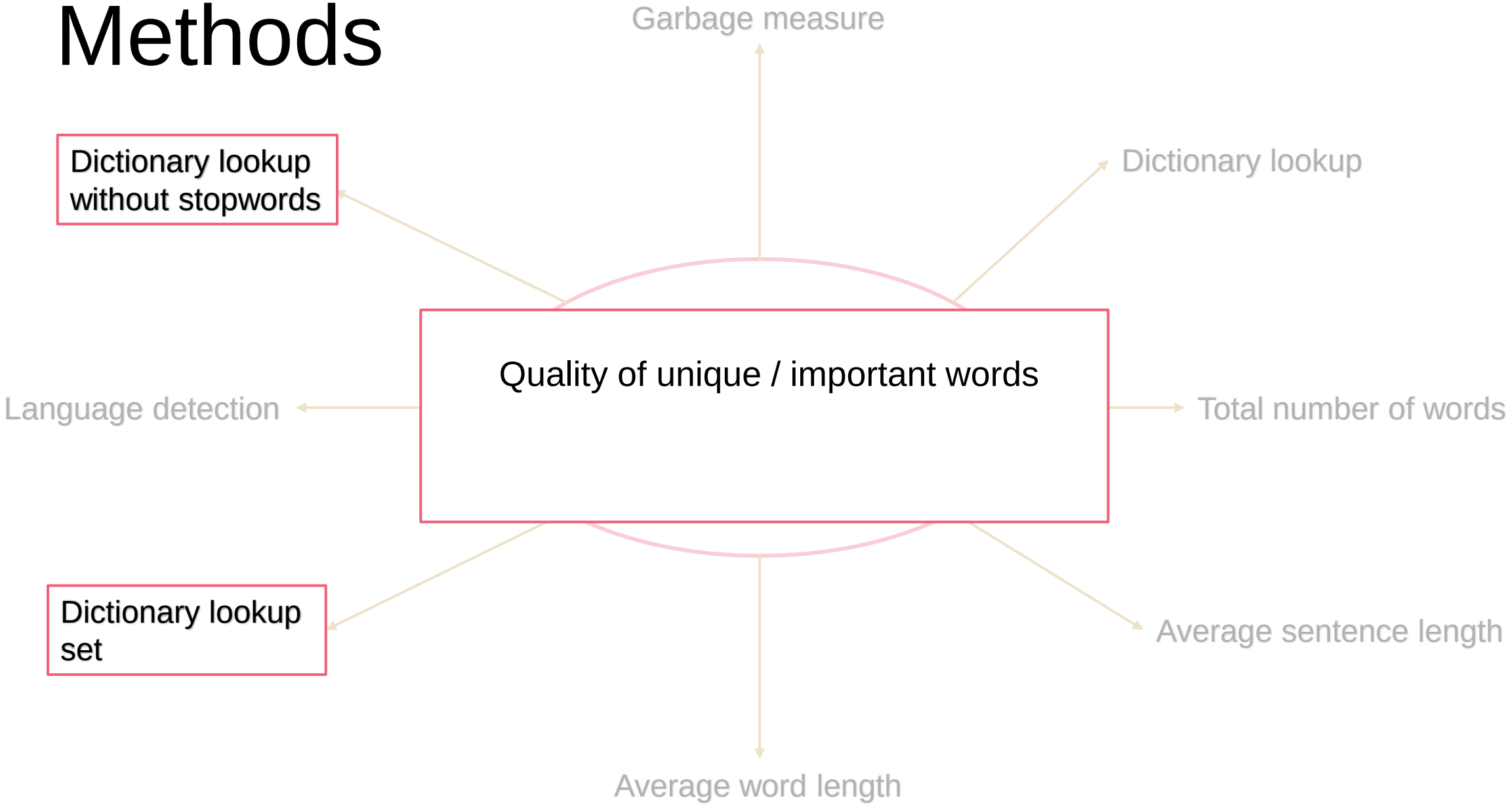
JOURNAL DE LA PROVINCE DE LIMBOURG. (N.° 5.)
Samedi, 6 Janvier 1821.
PORTUGAL Lisbonne, le 13 décembre. Les élections des députés se poursuivent avec acti vité dans cette capitale et dans les provinces; elles se font en tout conformément aux ordres qui ont été
és à ce sujet. Cependant, dans l'Alentejo on a découvert
voudrait réunie des cortès à Lamego, mais afin d'étouffer
incipe cette division, on a envoyé de cette capitale dans la
vince quelques régimens d'infanterie et de cavalerie. Nos
ec le cabinet de Madrid deviennent chaque jour plus
et plus amicales. Avant hier, nous avons vu arriver une
estafette espagnole venant de Madrid, avec des dépêches pour notre gouvernement. Depuis quelques j ours, un grand nom bre d'Espagnols affluent à Lisbonne. Le bruit court que des mouvemens sérieux ont eu lieu à Rio-Janeiro aussitôt qu'on y a appris la nou velle des événemens du Portugal. On dit que les troupes ont fait dans cette capitale la même chose qu'à Oporto; mais jusqu 'à présent, on ne connaît rien d'officiel à cet égard. ESPAGNE. Madrid, le 21 décembre. On dit que le nombre des Espagnols qui ont ordre de quitter la capitale est de 54, mais on ne connait pas encore leurs noms. On

MEASURING METHODS

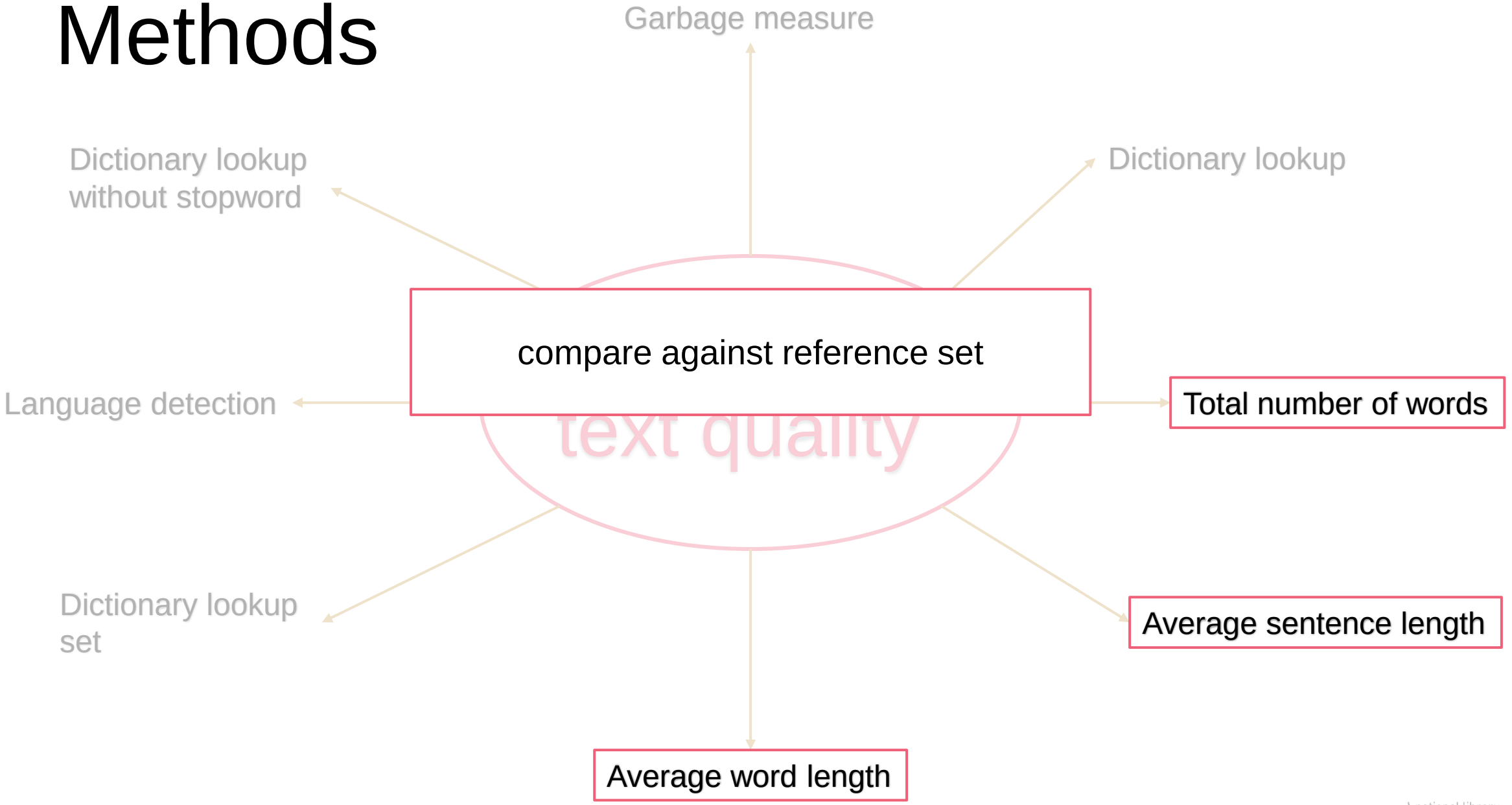
Methods


Garbage measure
Dictionary lookup without stopwords
Dictionary lookup
measuring text quality
Language detection
Total number of words
Dictionary lookup set
Average sentence length
Average word length

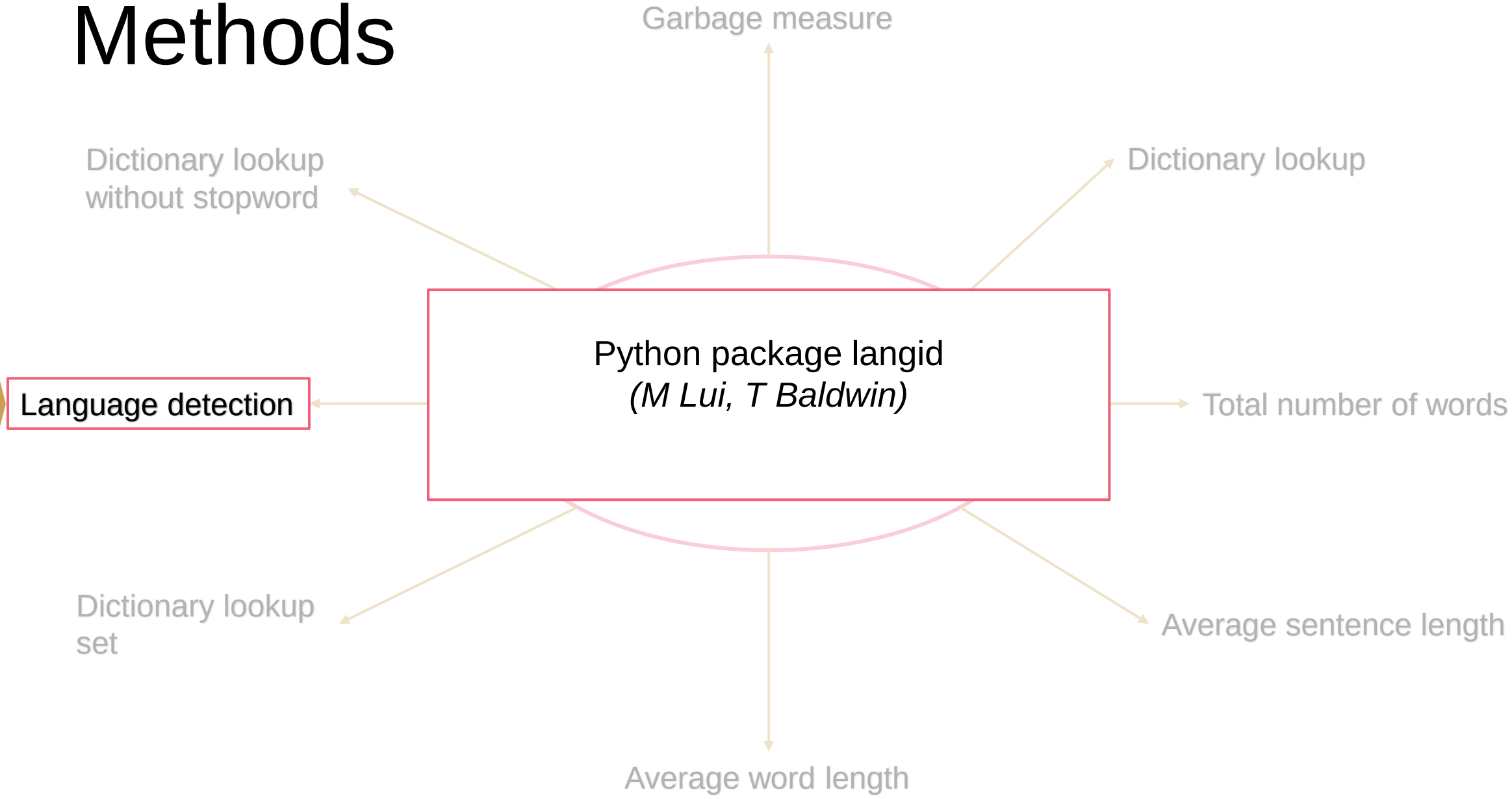
Methods


Garbage measure
Dictionary lookup without stopwords
Dictionary lookup
Quality of unique / important words
Language detection
Total number of words
Dictionary lookup set
Average sentence length
Average word length

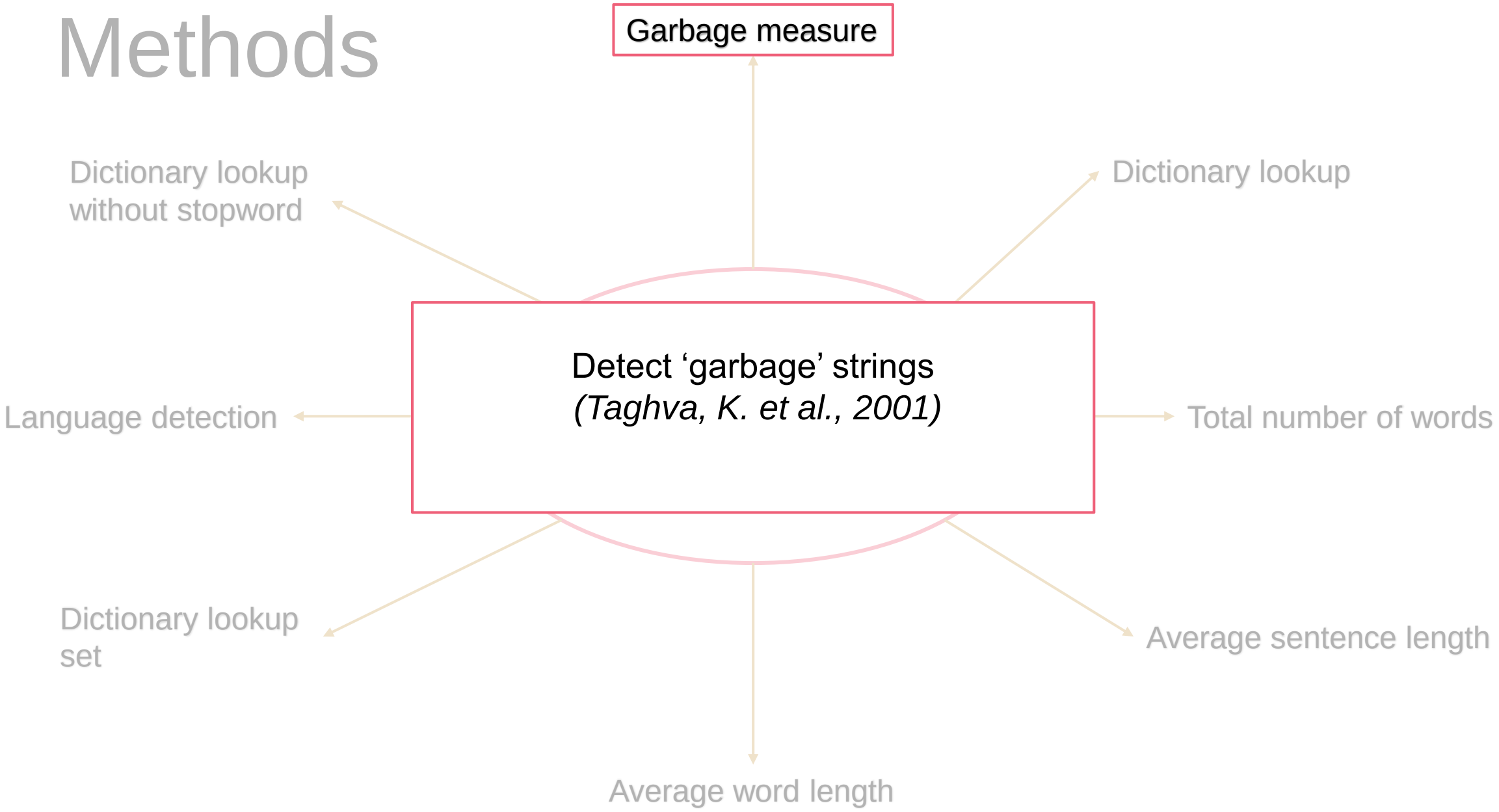
Methods


Garbage measure
Dictionary lookup without stopword
Dictionary lookup
compare against reference set
Language detection
Total number of words
text quality
Dictionary lookup set
Average sentence length
Average word length

Methods


Garbage measure
Dictionary lookup without stopword
Dictionary lookup
Language detection
Python package langid
(M Lui, T Baldwin)
Total number of words
Dictionary lookup set
Average sentence length
Average word length

Methods

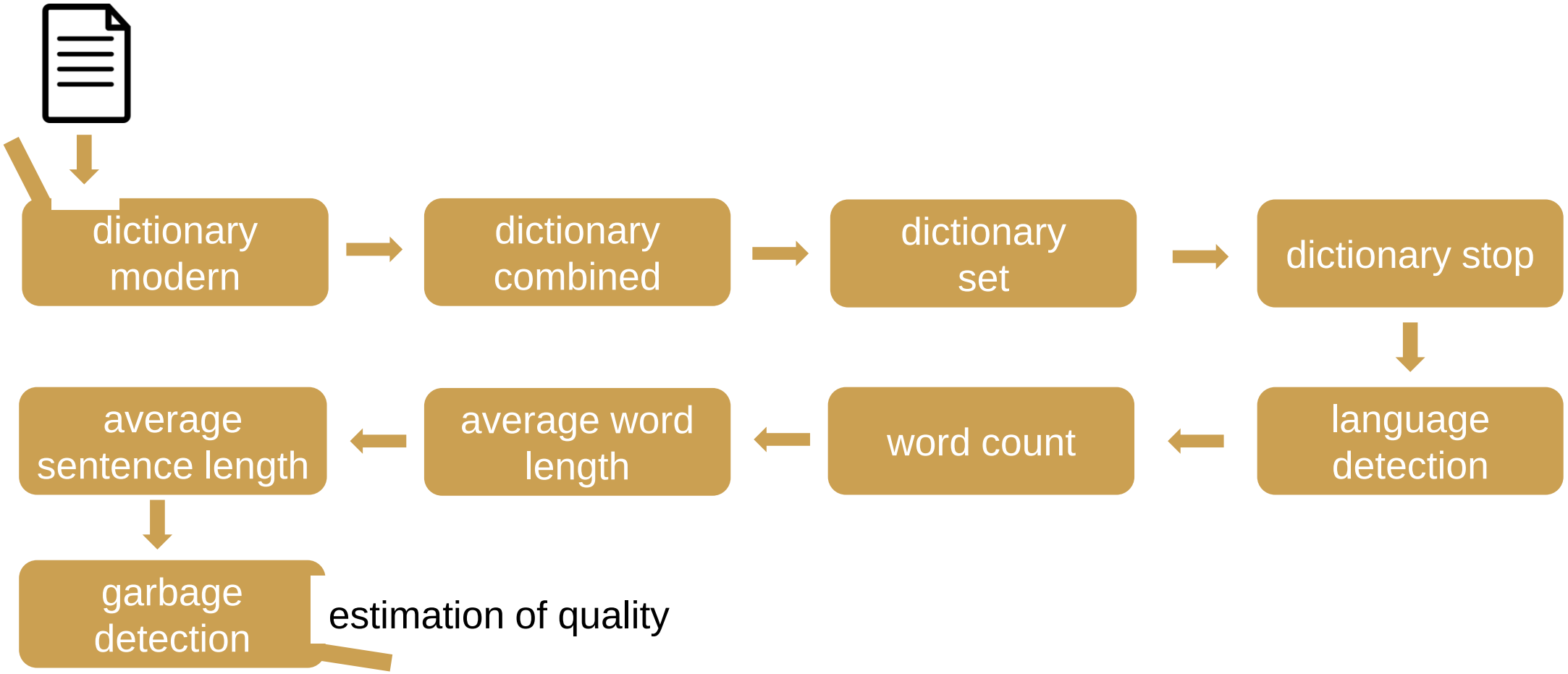

Garbage measure
Dictionary lookup without stopword
Dictionary lookup
Detect ‘garbage’ strings
(Taghva, K. et al., 2001)
Language detection
Total number of words
Dictionary lookup set
Average sentence length
Average word length

Characteristics

1. Type of text (advertisement, article, page)
2. Published year
3. Language

MULTI-LAYERED APPROACH

Pipeline

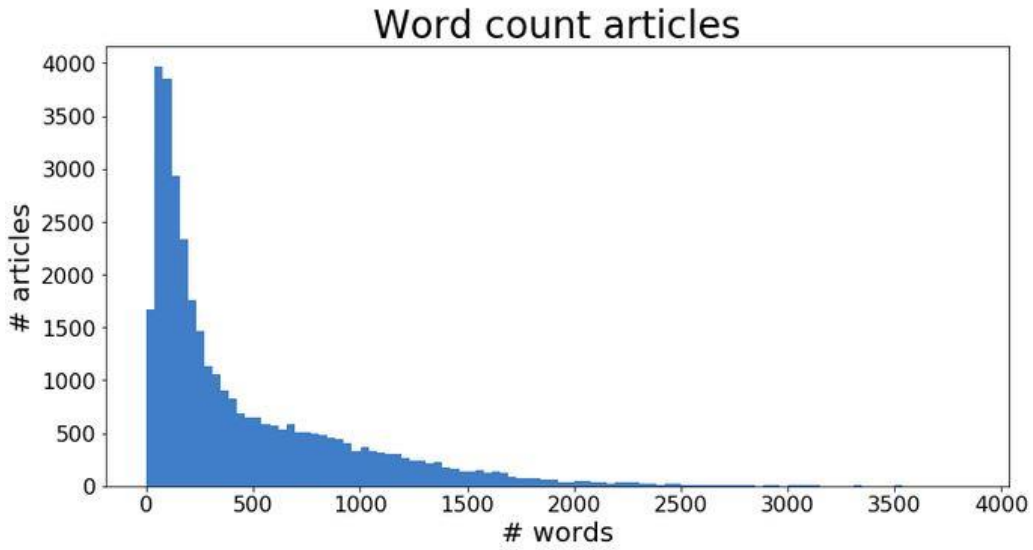

dictionary modern
dictionary combined
dictionary set
dictionary stop
language detection
word count
average word length
average sentence length
garbage detection
estimation of quality

Comparison measures

17th century news articles Ground Truth

80% of articles:

- Word count:
  - 56 – 1182 words
- Word length:
  - Mean: 4 – 5 characters
  - Median: 4 – 5 characters
- Sentence length:
  - Mean: 12-36 words
  - Median: 5-26 words

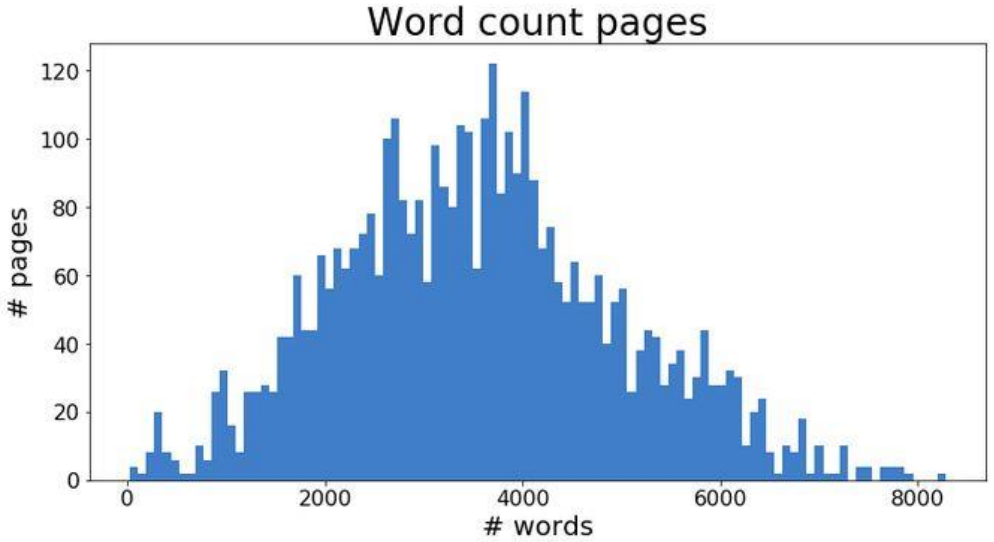


Comparison measures

Random newspaper pages Ground Truth

80% of articles:

- Word count:
  - 1760 – 5558 words
- Word length:
  - Mean: 4.9 – 5.4 characters
  - Median: 4 characters
- Sentence length:
  - Mean: 7-19 words
  - Median: 3-16 words



Pipeline – example outcomes

Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
17th century news article										
60%	80%	75%	65%	2000	nl (1.0)	4	4.5	9	15	[0.25, 0, 0, 0.25, 0.12, 0]
20th century newspaper page										
60%	80%	75%	65%	2000	nl (1.0)	4	4.5	9	15	[0.25, 0, 0, 0.25, 0.12, 0]

= good

= bad

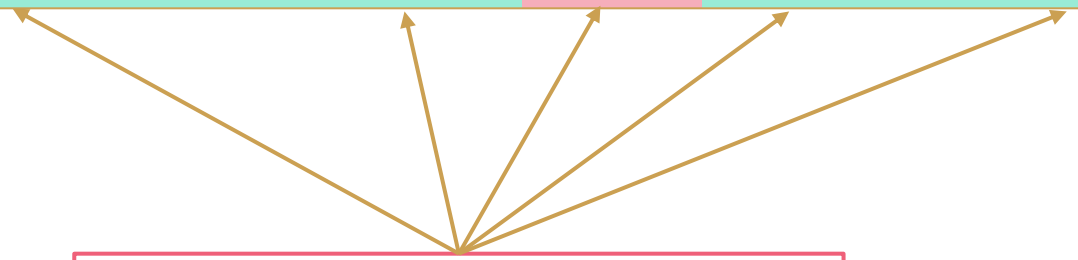
Pipeline – example outcomes

Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
17th century news article										
60%	80%	75%	65%	2000	nl (1.0)	4	4.5	9	15	[0.25, 0, 0, 0.25, 0.12, 0]
20th century newspaper page										
60%	80%	75%	65%	2000	nl (1.0)	4	4.5	9	15	[0.25, 0, 0, 0.25, 0.12, 0]

Cut-off point: 70%

Pipeline – example outcomes

Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
17th century news article										
60%	80%	75%	65%	2000	nl (1.0)	4	4.5	9	15	[0.25, 0, 0, 0.25, 0.12, 0]
20th century newspaper page										
60%	80%	75%	65%	2000	nl (1.0)	4	4.5	9	15	[0.25, 0, 0, 0.25, 0.12, 0]



Pipeline – example outcomes

Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
17th century news article										
60%	80%	75%	65%	2000	nl (1.0)	4	4.5	9	15	[0.25, 0, 0, 0.25, 0.12, 0]
20th century newspaper page										
60%	80%	75%	65%	2000	nl (1.0)	4	4.5	9	15	[0.25, 0, 0, 0.25, 0.12, 0]

Dutch

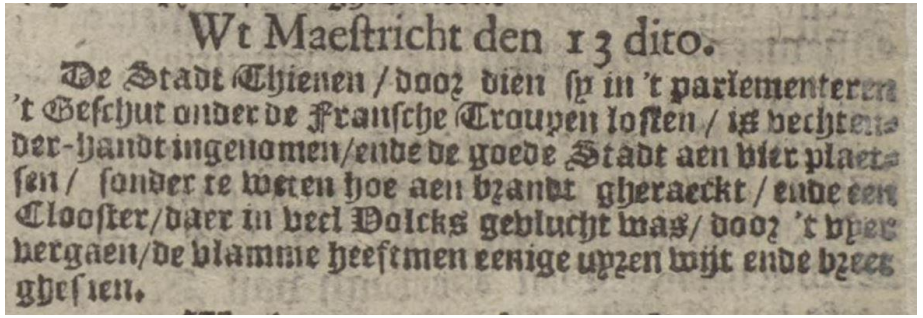
Pipeline – example outcomes

Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
17th century news article										
60%	80%	75%	65%	2000	nl (1.0)	4	4.5	9	15	[0.25, 0, 0, 0.25, 0.12, 0]
20th century newspaper page										
60%	80%	75%	65%	2000	nl (1.0)	4	4.5	9	15	[0.25, 0, 0, 0.25, 0.12, 0]

Each value < 0.25

FIRST RESULTS

Missing text


Wt Maeſtricht den 13 dito.
De Stadt Thienen / dooz dien ſy in 't parlementeren 't Geſchut onder de Franſche Troupen loſten / is bechten der-handt ingenomen/ende de goede Stadt aen vier plaetſen / ſonder te weten hoe aen vyandt geraeckt / ende een Clooſter/daer in veel Volcks gevlucht was/ dooz 't vyer vergaen/de vlamme heeftmen eenige uyzen wijt ende zyde ghe ſien.

Wt Maestricht den 13 dito.
g_<Â«^taÃ¶tiâ,¬i)ieiien/boo? bien
fptn'tpartementer

Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
50%	75%	75%	66,67%	8	Lb (0.9)	4.5	8.88	4.5	4.5	[0.25, 0, 0, 0.25, 0.12, 0]

Remarkable:
- Language Luxembourg's
- Very low word count
- Garbage detected

Missing text

2 Vrijdag 28 Maart 1930 – DE TELEGRAAF – Avondblad – Eerste Blad.

KOETJEPOF UIT MOSKOU VERWIJDERD?

Hij zou te Soesdal gevangen worden gehouden.

Nieuwe lezing over zijn ontvoering. Met de „Spartak" van Trouville via Antwerpen naar Rusland. - Thans in een oud klooster.

EEN DUEL IN DE KEUKEN.

Tusschen twee zusters om een man.

BEIDEN NAAR HET ZIEKENHUIS GETRANSPORTEERD.

NEDERLAAG DER IERSCHE REGEERING.

Zal het kabinet aftreden?

VAN DE BOULEVARDS.

Standbeelden-kerkhof.

NATIONALISTEN DE MEERDERHEID IN TRANSVAAL.

AMSTERDAM, 28 Maart.

WEINIG OPLEVING IN DE HAVEN.

Ook hier het algemeene verschijnsel van de malaise.

OOST-INDIË.

ONTSLAGKWESTIE BIJ DE BATAAFSCHE.

ZUID-SUMATRA LANDBOUW- EN NIJVERHEIDSVEREENIGING.

WEEROVERZICHT.

Klaverblad's Karnemelkzeep

HOTEL CARLTON

AMSTERDAM Telefoon: 34000 (14 lijnen)

Zaterdag 29 en Zondag 30 Maart

GRAND DINER-DANSANT

JULIE et FERNAN, Danseurs Espagnols

KOETJEPOF UIT MOSKOU VERWIJDERD? Hij zou te Soesdal gevangen worden gehouden. Nieuwe lezing over zijn ontvoering. — Met de „Spartak" van Trouville via Antwerpen naar Rusland. - Thans in een oud klooster. I ■ kou i Ise inKe.'. - ovjet ■ schip ■ ■ ' ■ •i het rgnnl • ■ Loeb ■ ■es.lal o. » .st tiians g EEN DUEL IN DE KEUKEN. Tusschen twee zusters om een man. BEIDEN NAAR HET ZIEKENHUIS GETRANSPORTEERD. •.' i '\ ■ •ti man en - I '- - I NEDERLAAG DER IERSCHE REGEERING. Zal het kabinet aftreden?

VAN DE BOULEVARDS. Standbeelden-kerkhof. ■ ogert w loal de si. nen. ■ twor ■ los IV» • ■ In d ■ ■ .lel 1 ' naar - ■ - ■ ■ ■ campagne voor de i een tl » ■ ■ NATIONALISTEN DE MEERDERHEID IN TRANSVAAL. Als resultaat der provinciale verkiezingen. JOH • zingen voor den provincialen raad van 'I • • kaaneche partij en Z I.ab- Twee uitslagen zijn nog niet I.- i zul ■ ■

WEINIG OPLEVING IN DE HAVEN. Ook hier het algemeene verschijnsel van de malaise. maakt, b. v. met het ■ rat — aan te raden ■ hepen ida -ttge ' alles ii" een n . ral nu •rs en ■ir dB ki.'kt. maar let np i •n de -trteatjjttl U den A ' . .i de Poll er den Maan ■ ■ 19:» ïtso tti 27 II 4 8 | i S is 61 1 13 . — 1 14 3 117 141 II U's: .it al Is be; - -r_an, de toestand In «een of « , e cij - - naar ■ hun . hotu en In aan i ;. het oogenblik het beeld

OOST-INDIË. ONTSLAGKWESTIE BIJ DE BATAAFSCHE. ■ - . H. i. M. raadt -chf C en, of i ZUID-SUMATRA LANDBOUW-EN NIJVER HEIDSVEREENIGING. ■ J. 1! M. J. 1 i i ■ WEEROVERZICHT. ' . A ' ■ ■ LiUkt'i. tn 7. 55 gr. Woedende

Dictionary lookup – missing text

Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
62.4%	70.8%	69%	48.8%	867	nl (1.0)	3	3.9	7	8.4	[0.33, 0, 0, 0.33, 0.02, 0.37]

Remarkable:
- Low word count for page – less than half of the minimum value
- High garbage measure

Word importance

Een handleidinq of een gebriiksaanwijzing is een schriftelljke instrvctie die met een qroduct meegeleverd wordt.

De inhovd van een handlejding is:

- hoe een produot te assernbleren;
- hoe een product te reporeren;
- hoe een pnodvct te instaileren;
- hoe een product gebruikt dient te worden;
- hoe een poduct te onderhovden;
- hoe de instelllngen van een product aan te passen;
- hoe een storimg op te lossen;
- hoe een produci niet te misbruiken;
- hoe comtact op te nemen voor servioe.

Word importance

Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
73.3%	76%	58.1%	40%	76	nl (1.0)	3	5.43	15	25.67	[0.01, 0, 0, 0, 0, 0.01]

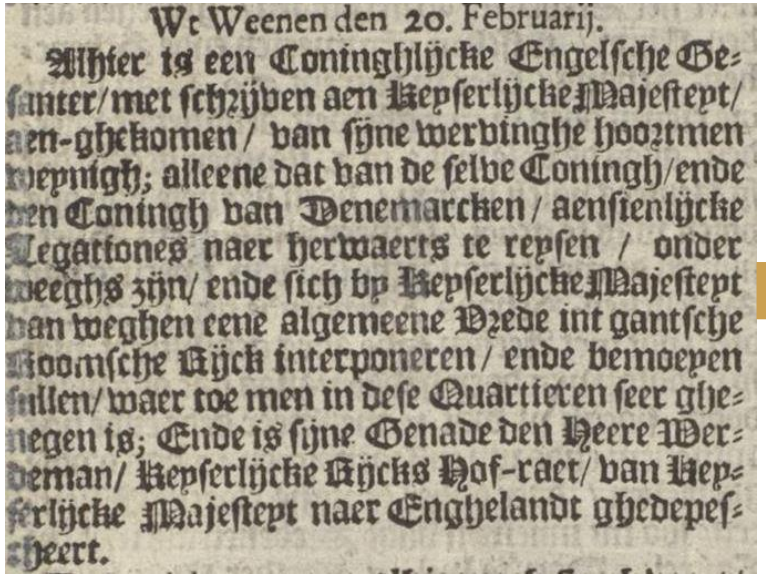
Een handleidinq of een gebriiksaanwijzing is een schriftelljke instrvctie
die met een qroduct meegeleverd wordt
hoe een phodvct te installeren;
hoe een product gebruikt dient te worden;
hoe een poduct te onderhovden;
hoe de instelllngen van een product aan te passen;
hoe een storimg op te lossen;

Not every measure can be evaluated

Remarkable:
- Dict set and dict stop considerably lower

Time period

Article, 1636

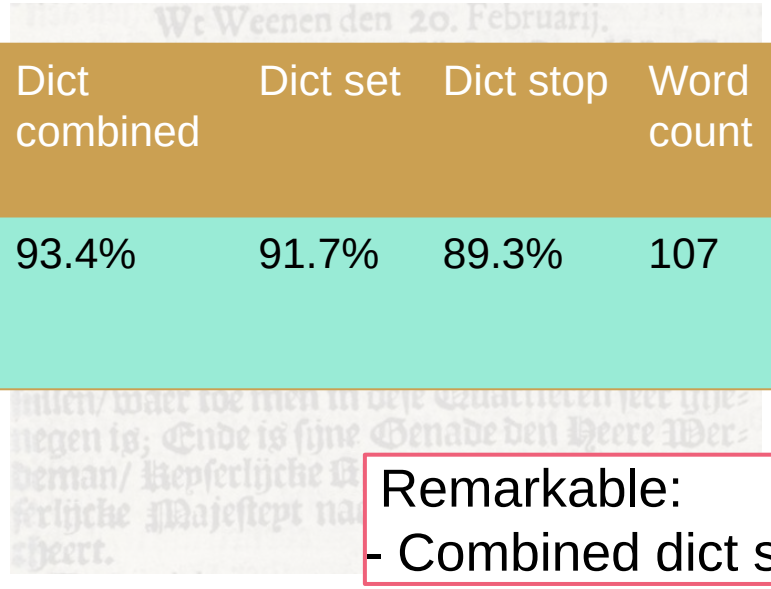


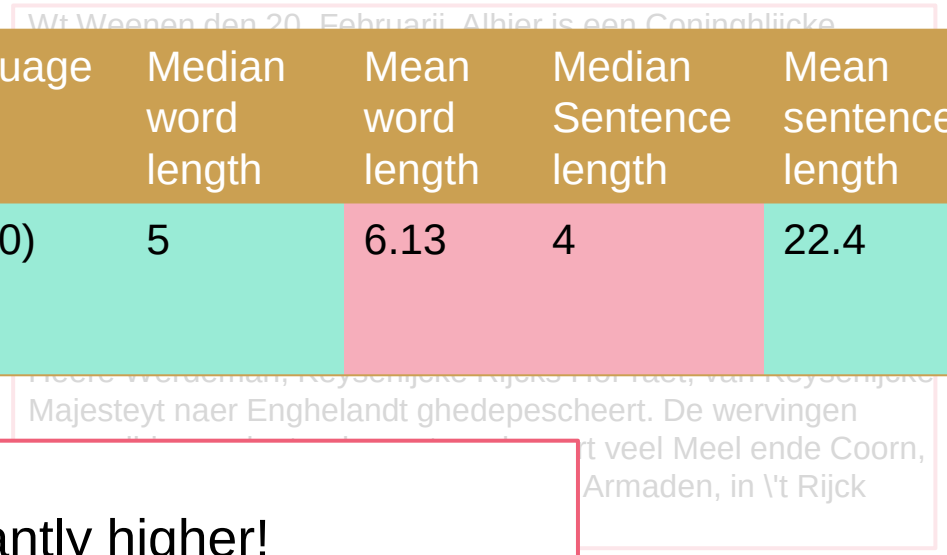
Wt Weenen den 20. Februarij.
Alhier is een Coninghlijcke Engelſche Geſanter/met ſchrijven aen Keyſerlijcke Majeſteyt/aen-ghekomen/ van ſijne werdinghe hoortmen weynigh; alleene dat van de ſelve Coningh/ende den Coningh van Denemarcken/ aenſienlijcke Legationes naer herwaerts te reyſen / onder weeghs zijn/ ende ſich by Keyſerlijcke Majeſteyt van weghen eene algemeene Vrede int gantſche Roomſche Rijck interponeren/ ende bemoeyen ſullen/waer toe men in deſe Quartieren ſeer ghenegen is; Ende is ſijne Genade den Heere Werdeman/ Keyſerlijcke Rijcks Hof-raet/ van Keyſerlijcke Majeſteyt naer Enghelandt ghedepeſcheert.

Wt Weenen den 20. Februarij. Alhier is een Coninghlijcke Engelsche Gesanter, met schrijven aen Keyserlijcke Majesteyt, aen-ghekomen, van sijne wervinghe hoortmen weynigh; alleene dat van de selve Coningh, ende den Coningh van Denemarcken, aensienlijcke Legationes naer herwaerts te reysen, onder weeghs zijn, ende sich by Keyserlijcke Majesteyt van weghen eene algemeene Vrede int gantsche Roomsche Rijck interponeren, ende bemoeyen sullen, waer toe men in dese Quartieren seer ghenegen is; Ende is sijne Genade den Heere Werdeman, Keyserlijcke Rijcks Hof-raet, van Keyserlijcke Majesteyt naer Enghelandt ghedepescheert. De wervingen gaen alhier noch sterck voort, ende wert veel Meel ende Coorn, tot provianderinghe vande Keyserlijcke Armaden, in \'t Rijck ghesonden.

Dictionary lookup – timeperiod

Article, 1636





Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
41%	93.4%	91.7%	89.3%	107	nl (1.0)	5	6.13	4	22.4	[0.04, 0, 0, 0.03, 0, 0.09]

Remarkable:
- Combined dict significantly higher!

Dictionary lookup – languages

Combined dictionary:

86% of words found

JOURNAL DE LA PROVINCE DE LIMBOURG. (N.° 5.)
Samedi, 6 Janvier 1821.
PORTUGAL Lisbonne, le 13 décembre. Les élections des députés se poursuivent avec acti vité dans cette capitale et dans les provinces; elles se font en tout conformément aux ordres qui ont été
és à ce sujet. Cependant, dans l'Alentejo on a découvert
voudrait réunie des cortès à Lamego, mais afin d'étouffer
incipe cette division, on a envoyé de cette capitale dans la
vince quelques régimens d'infanterie et de cavalerie. Nos
ec le cabinet de Madrid deviennent chaque jour plus
et plus amicales. Avant hier, nous avons vu arriver une
estafette espagnole venant de Madrid, avec des dépêches pour notre gouvernement. Depuis quelques j ours, un grand nom bre d'Espagnols affluent à Lisbonne. Le bruit court que des mouvemens sérieux ont eu lieu à Rio-Janeiro aussitôt qu'on y a appris la nou velle des événemens du Portugal. On dit que les troupes ont fait dans cette capitale la même chose qu'à Oporto; mais jusqu 'à présent, on ne connaît rien d'officiel à cet égard. ESPAGNE. Madrid, le 21 décembre. On dit que le nombre des Espagnols qui ont ordre de quitter la capitale est de 54, mais on ne connait pas encore leurs noms. On

Dictionary lookup – languages

JOURNAL DE LA PROVINCE DE LIMBOURG. (N.° 5.)
Samedi, 6 Janvier 1821.
PORTUGAL Lisbonne, le 13 décembre. Les élections des députés se poursuivent avec acti vité dans cette capitale et dans les provinces; elles se font en tout conformément aux ordres qui ont été
és à ce sujet. Cependant, dans l'Alentejo on a découvert
gouvernement. Depuis quelques j ours, un grand nom bre
d'Espagnols affluent à Lisbonne. Le bruit court que des mouvemens
l'on y a appris la nou velle
troupes ont fait dans cette
capitale la même chose qu'à Oporto; mais jusqu 'à présent, on ne connaît rien d'officiel à cet égard. ESPAGNE. Madrid, le 21 décembre. On dit que le nombre des Espagnols qui ont ordre de quitter la capitale est de 54, mais on ne connait pas encore leurs noms. On

Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
42.7%	82.8%	71,8%	76.4%	803	fr (1.0)	4	4.74	16	18.06	[0.11, 0, 0, 0.1, 0, 0.09]

Remarkable:
- Predicted language

FUTURE WORK

More testing

- More tests and more data
- Prediction of correct texts
- Comparison measured quality with ground truth

Quantifying results

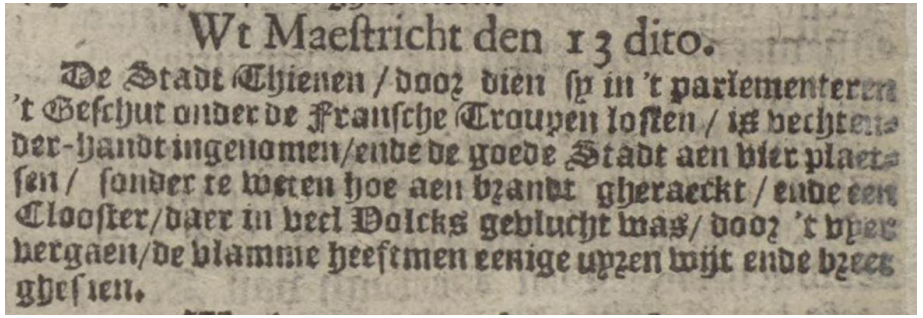
From table to one final score:

- ‘good’, ‘average’, ‘bad’

Weights

- Importance
- Characteristics

Weights - example


Wt Maeſtricht den 13 dito.
De Stadt Thienen / dooz dien ſy in 't parlementeren 't Geſchut onder de Franſche Troupen loſten / is bechten der-handt ingenomen/ende de goede Stadt aen vier plaetſen / ſonder te weten hoe aen vyandt geraeckt / ende een Clooſter/daer in veel Volcks geblucht was/ dooz 't vyer hergaen/de vlamme heeftmen eenige uyzen wijt ende vzeet gheſien.

Wt Maestricht den 13 dito.
g_<Â«^taÃ¶tiâ,¬i)ieiien/boo? bien
fptn'tpartementer

Dict modern	Dict combined	Dict set	Dict stop	Word count	Language	Median word length	Mean word length	Median Sentence length	Mean sentence length	Garbage
50%	75%	75%	66,67%	8	Lb (0.9)	4.5	8.88	4.5	4.5	[0.25, 0, 0, 0.25, 0.12, 0]

Word count and language more important?

Adding more measures

- N-grams
- Part of speech
- Letter ratio
- WER (without ground truth)

Ideas?

mirjam.cuper@kb.nl

@CuperMirjam

Questions?

KB national library of the netherlands