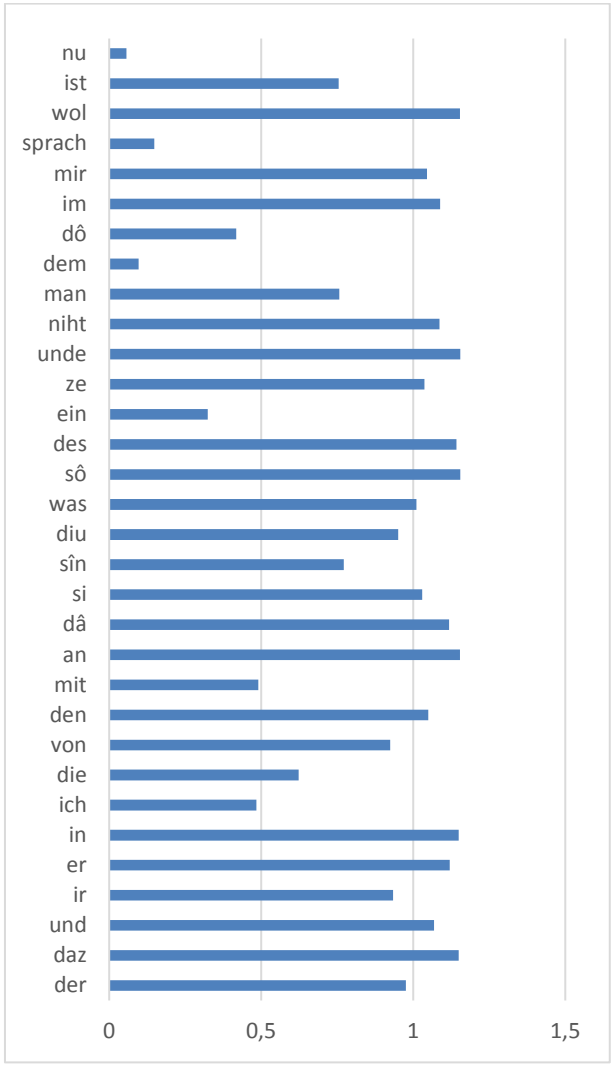
PD Dr. Friedrich Michael Dimpel: Autorschaftsattribution bei nicht-normalisiertem Mittelhochdeutsch. Bessere Erkennungsquoten durch ein Normalisierungswörterbuch

		Gesamtkorpus		Gottfrieds Tristan		Wolframs Parzival				Wolframs Willehalm	
		Ø Mess-werte	STDEV	Mess-wert	Z-Wert	Mess-wert	Z-Wert	Diff.	**Abs. Diff**	Mess-wert	Z-Wert
1	der	2,647	0,530	2,130	-0,976	2,623	-0,046	-0,930	**0,930**	3,189	1,022
2	daz	2,241	0,464	2,774	1,149	1,930	-0,671	1,820	**1,820**	2,019	-0,478
3	und	2,167	1,267	3,520	1,068	1,009	-0,914	1,982	**1,982**	1,971	-0,155
4	ir	1,760	0,408	2,141	0,933	1,810	0,122	0,811	**0,811**	1,330	-1,056
5	er	1,704	0,238	1,970	1,121	1,628	-0,319	1,439	**1,439**	1,513	-0,802
6	in	1,404	0,225	1,663	1,150	1,253	-0,669	1,818	**1,818**	1,295	-0,481
7	ich	1,343	0,079	1,304	-0,484	1,434	1,150	-1,634	**1,634**	1,290	-0,666
…									…		
200	zît	0,064	0,011	0,076	1,131	0,055	-0,768	1,899	**1,899**	0,060	-0,362

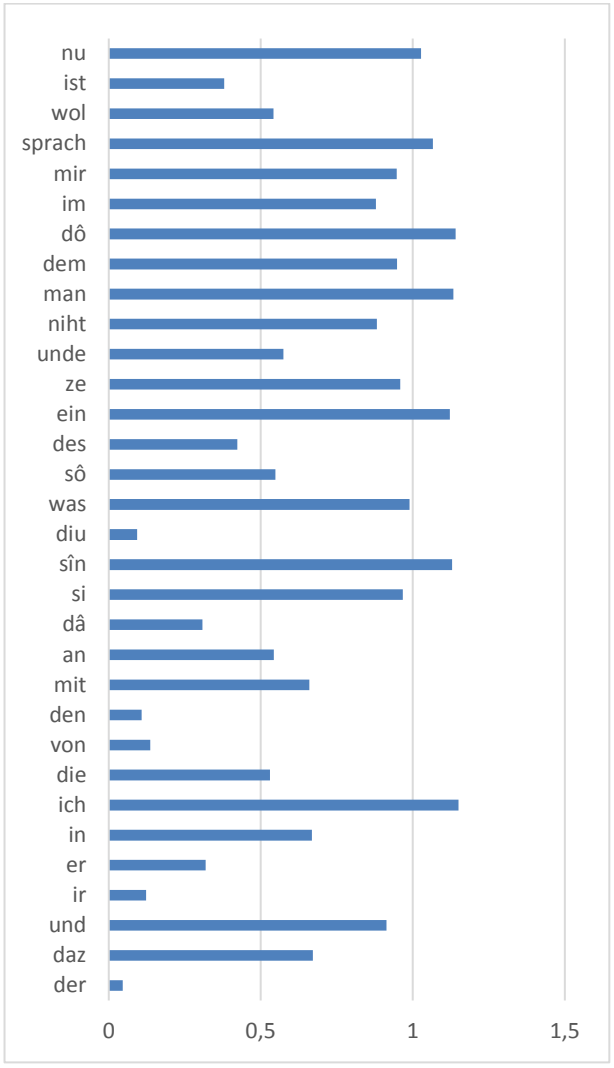
Mittelwert der absoluten Z-Wert-Diffenzen (Burrows' Delta): 1,480

$$Diff. = Z_1 - Z_2 = -0.976 - 0.046 = -0.930$$

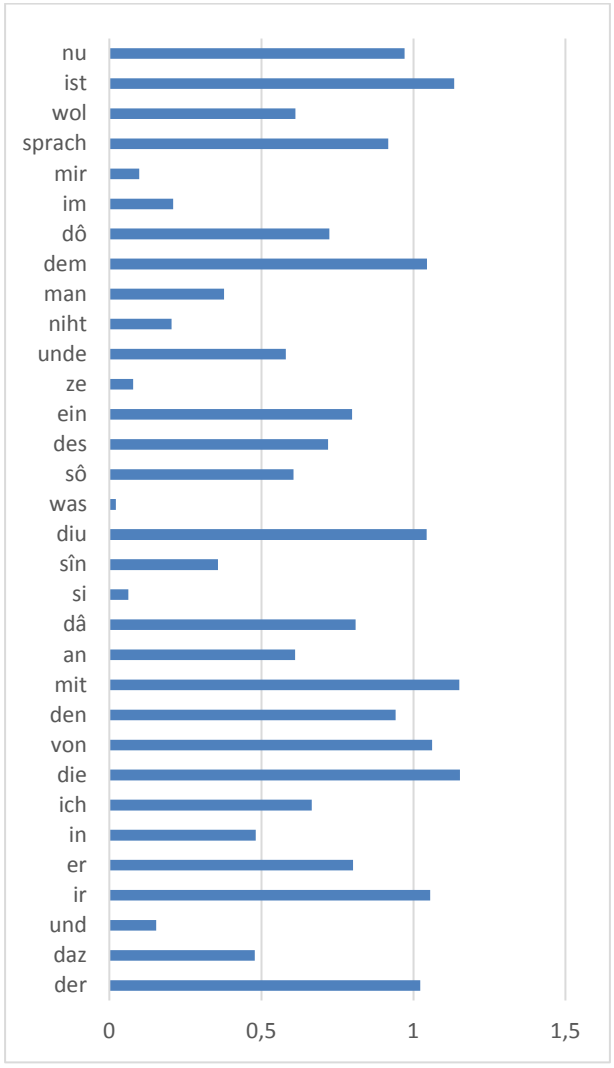
Gottfrieds ‚Tristan' – Positive Z-Werte



Wolframs ‚Parzival' – Positive Z-Werte



Wolframs ‚Willehalm' – Positive Z-Werte

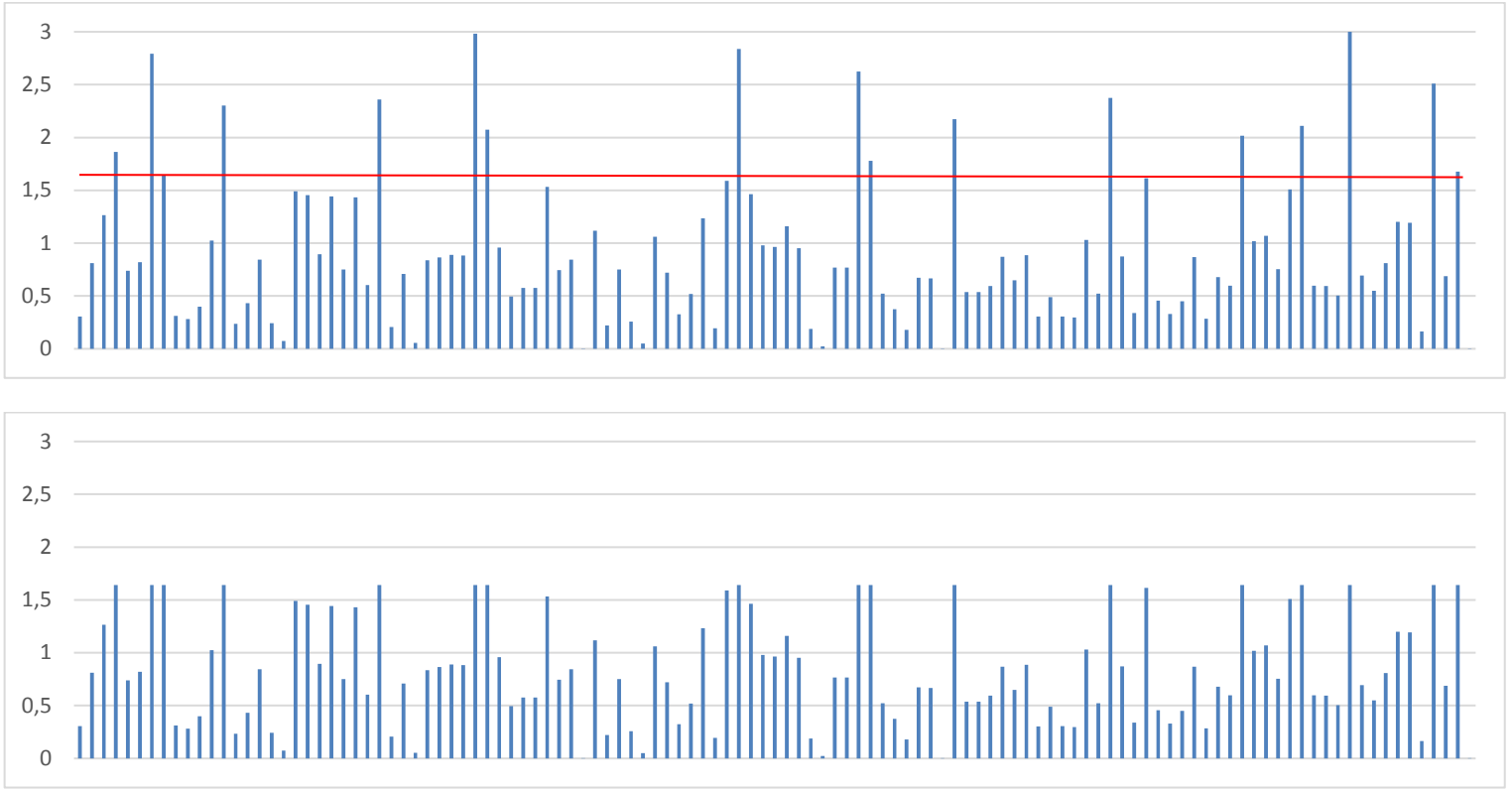


Schlüsselprofil-These versus Ausreißer-These

Beruht der Erfolg von Delta auf dem Schlüsselprofil von vielen Werten oder auf extremen Werten (‚Ausreißern‘)?

Evert / Jannidis / Pielström / Reger / Schöch / Vitt:

Das Abschneiden von extremen Z-Werten führt zu einer besseren Erkennungsquote



Probleme der Schreibung bei mittelhochdeutschen Texten

Nebeneinander von

- *und / unde / unt / vnt*
- *wîb / wîp / wib / wip*
- *kom / kam*
- *gotes / gotis*
- *sůzer / süezer*

-> Einflüsse von Dialekt, Normalisierung (durchgeführt? Art?) und Schreibergewohnheit

Stylo-R-Plot mit 37 heterogenen Texten

Rudolf von Ems:

‚Alexander', ‚Barlaam' normalisiert

‚Willehalm von Orlens', ‚Weltchronik' nicht normalisiert

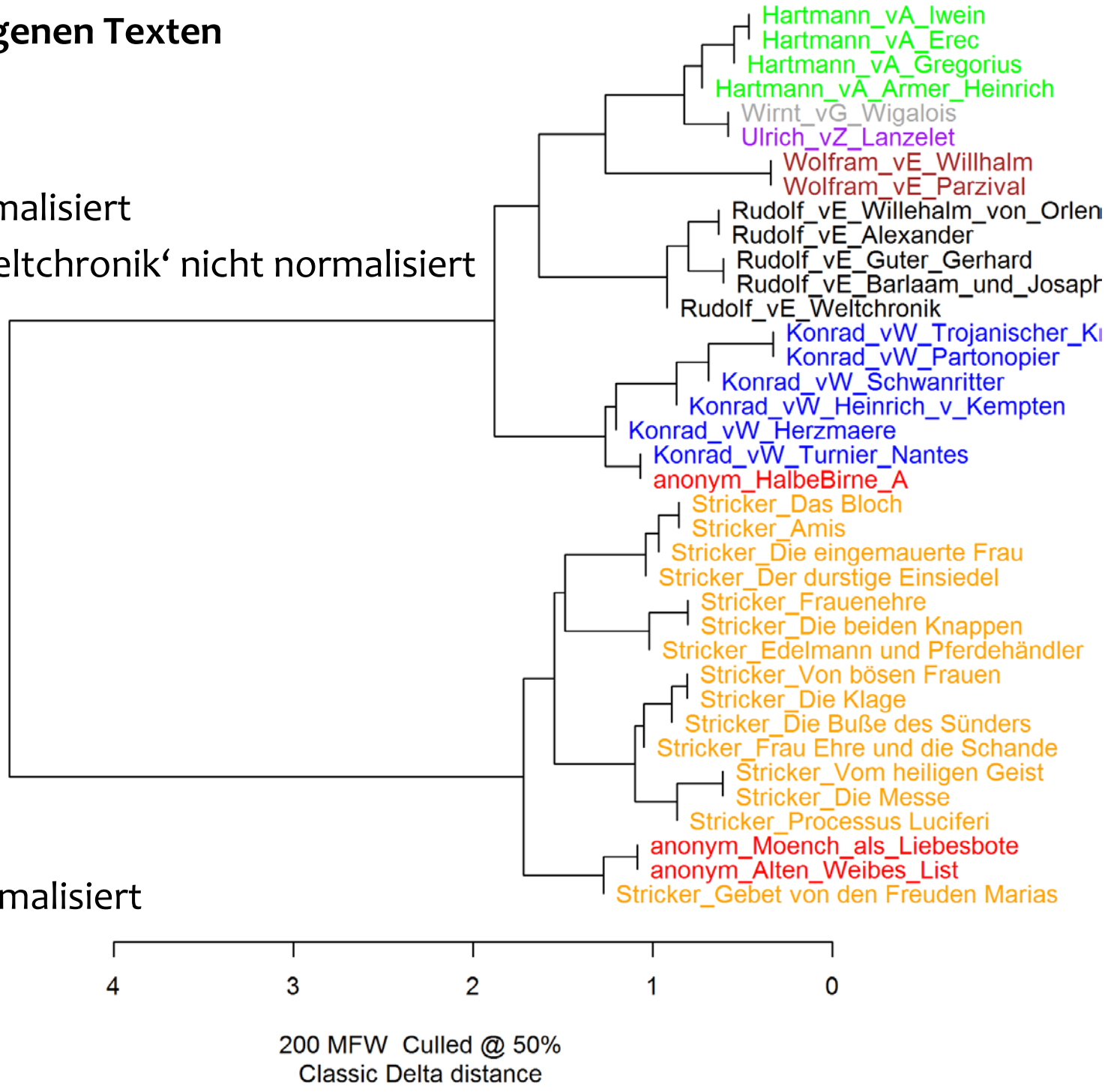
über versus *ubir*

Stricker:

‚Pfaffe Amis' normalisiert

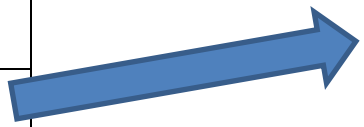
Kurzerzählungen nicht normalisiert

für versus *fur*



Validierungstest normalisierte Texte (24Rx20T): Procedere

Ratekorpus 24 Texte, Autoren wie im Trainingskorpus	
Text A	Autor 1
Text B	Autor 1
Text C	Autor 2
Text D	Autor 3
…	…



Trainingskorpus 20 Texte, ein Text pro Autor	
Autor 1	Text E
Autor 2	Text F
Autor 3	Text G
…	…

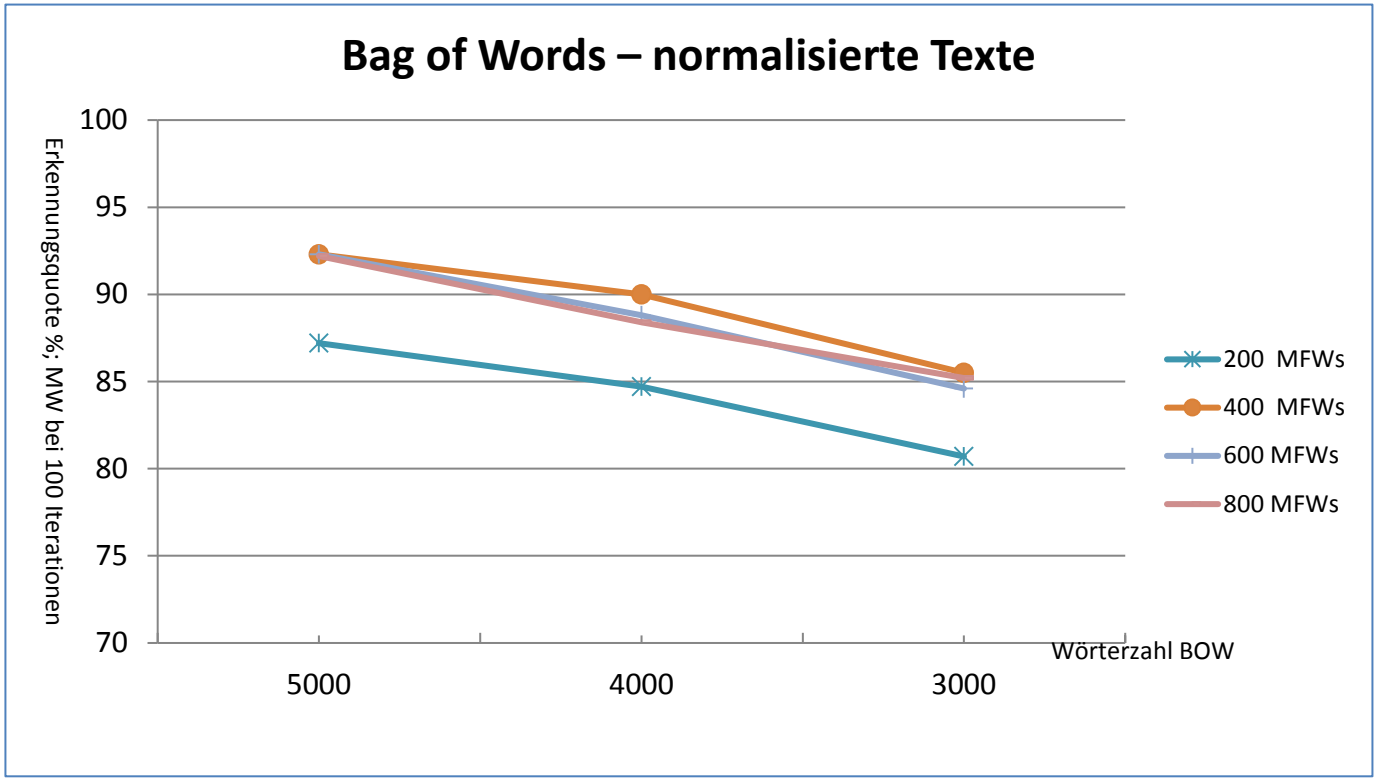
Erkennungsquote %: Wieviele Texte aus dem Ratekorpus haben den niedrigsten Deltawert zum richtigen Autor aus dem Trainingskorpus?

Validierungstests normalisierte Texte (24Rx20T): Ergebnisse

Methode: Bag-of-Words

Zufällige Auswahl von 3.000 / 4.000 / 5.000 Wortformen je Text

100 Iterationen wg. Zufallseinfluss bei der Wortselektion; die Erfolgsquote als Mittelwert dieser 100 Iterationen



Durchschnittliche Erkennungsquote für 200–800 MFWs bei 5.000 Wortformen: **91,0%**

Abhängigkeit von Edition und Handschrift?

Erkennungsquote für:	
‚Titurel' (Lachmann)	100%
‚Parzival' (Lachmann)	100%
‚Parzival' (Leitzmann)	99,7%
‚Parzival' Handschrift D	100%
‚Parzival' Handschrift G	99,1%
‚Parzival' Handschrift M	73,3%

Trainingskorpus: 20 Texte, darunter Lachmanns normalisierte ‚Willehalm'-Ausgabe

Jeweils nur ein Ratetext von Wolfram getestet bei 100 Iterationen und eine Bag-of-Words mit 5.000 Wortformen

Lachmann-‚Parzival'	Handschrift G	Handschrift M
man bôt ein badelachen dar: des nam er vil kleine war. sus kunder sich bî frouwen schemn, vor in wolt erz niht umbe nemn. die juncfrouwen muosen gên: sine torsten dâ niht langer stên. ich wæn si gerne heten gesehn, ob im dort unde iht wære geschehn. wîpheit vert mit triuwen: si kan friwendes kumber riuwen.	man bot ein badelachen dar. des nam er vil chleine war. sus chunder si bi frouwen schemen. vor in wolt erz niht vmbe nemen. die ivnchfrouwen mousen gen. si getorsten da niht lenger sten. die wane ich gerne heten gesehen. obe im dort vnden iht waere geschehen. wipheit vert mit triwen. si chan frivndes chumber riwen.	man bot eyn badelachen dar desz nam her vil cleine war susz konde her sich by vrouwen scemen vor yn woldirz nicht vmme nemen dy juncfrouwen musten gen sine gistorsten da nicht lenger sten die wene ich gerne hetten gisen ob yme dort vndir icht were giscen wipheit vert mit truwen sie kan frundesz kummer ruwen
56 Token	22 Abweichungen (von 56)	34 Abweichungen (von 56)

Teilnormalisierung und Normalisierungswörterbuch

Vorstufe: Sonderzeichen, Umlaute bereinigen

Bau eines Normalisierungswörterbuchs:

A) Perl-Skript:

- Suche jeweils 400 MFWs in nicht-normalisierten Texten
- Vorschlagsliste generieren aus Lachmann-Hartmann-Korpus via Levenshtein-Distanz
- User bestätigt oder ersetzt die zugeordnete normalisierte Wortform

B) Datengeschenk Sonja Glauch: Normalisierungsdaten aus „Lyrik des deutschen Mittelalters“

A)-B) => Mapping zu über 1.100 Wortformen

C) Datengeschenk Thomas Klein: Normalisierungsdaten aus dem Referenzkorpus Mittelhochdeutsch

A)-C) => Mapping zu über 120.000 Wortformen

Bereinigung des Normalisierungswörterbuchs

Probleme im Normalisierungswörterbuch:

- Mhd. *sluc* wird zu *sluoc* (nhd. „er schlug“) normalisiert. Mhd. *sluc* kann als stf. jedoch auch als normalisierte Form verwemdet werden („ein Schluck“)
- Abweichende Mappings: Wortformen zu *chunich* oder *küninc* normalisiert statt zu *künec*

Bereinigungsworkflow

- Frequenzbasierte Sortierung
- Bereinigung von Sonderzeichen (Nasalstrich, Superskripte, ...)
- Abgleich der Mappings zwischen Skript, Lyrik-Projekt und Referenzkorpus-Daten
- Beseitigung von Doppelmappings (zur gleichen diplomatischen Formen nur die häufigste Normalform behalten)
- Potentiell normalisierte Wortformen als diplomatische Form enthalten? Abgleich mit Vollformenwörterbuch
- Manuelle Korrektur von „ungültigen“ Normalformen anhand des Vollformenwörterbuchs

Vollformenwörterbuch (unvollständiger Prototyp)

- Extraktion von Lemmata und grammatikalischen Angaben aus der Trierer-Wörterbuch-CD (2004)
- Generierung von normalisierten Flexionsformen
- Starke Verben: Angaben zur Stammform erlauben Paradigma-Generierung mit grammatischem Wechsel und Auslautverhärtung
- Schwache Verben mit sog. Rückumlaut werden generiert, wenn im Sg. Präs. ein umgelauteter Vokal sowie Positionslänge oder Naturlänge vorliegen
- Nomina: Bislang nur bei Grundformen (ohne Präfix) werden vorhandene Angaben zu Genitiv/Plural/Endung verwendet, um Umlaut und Auslautverhärtung zu berücksichtigen

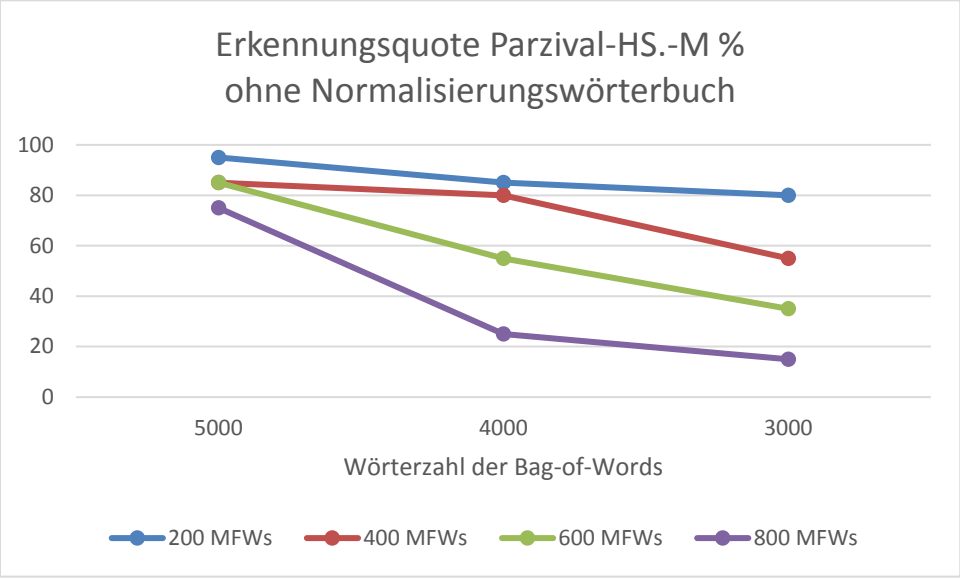
Unvollständig (etwa: keine Varianten) und teils unkorrekt => Das Ziel im Delta-Kontext ist lediglich, bei hochfrequenten Wortformen die automatische Teilnormalisierung zu verbessern.

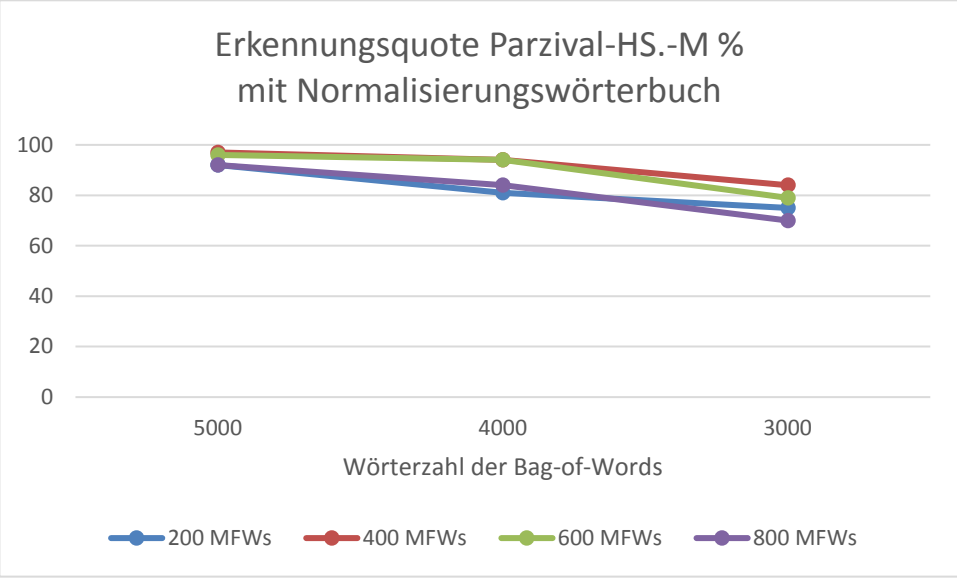
‚Parzival'-Handschrift M

Steigerung der durchschnittlichen Erkennungsquote für Vektoren von 200-800 MFWs bei einer Bag-of-Words mit 5.000 Wörtern:

Ohne Normalisierungswörterbuch: Ø 73,3%

Mit Normalisierungswörterbuch: Ø 97,1%





Neuer Test: nicht-normalisiertes Ratekorpus, gemischtes Trainingskorpus

Probleme Korpusbau:

- Digitale Verfügbarkeit
- Mindestens zwei Texte des gleichen Autors
- Texte länger als 5.000 Wörter (viele Predigten und Kurzerzählungen sind kürzer)
- Eder / Rybicki / Jannidis / Schöch meiden Gattungsmischung und Mischung von Vers-Prosa

Probleme beim Test von nicht-normalisierten Texten gegen normalisierte des gleichen Autors: Streuung bzw. fehlende Streuung.

Fiktives Beispiel:

- Wort „über“: 100 x in normalisierten Texten
- Nicht-normalisiert: 40 x „ubir“, 40 x „uber“, 20 x „vbir“

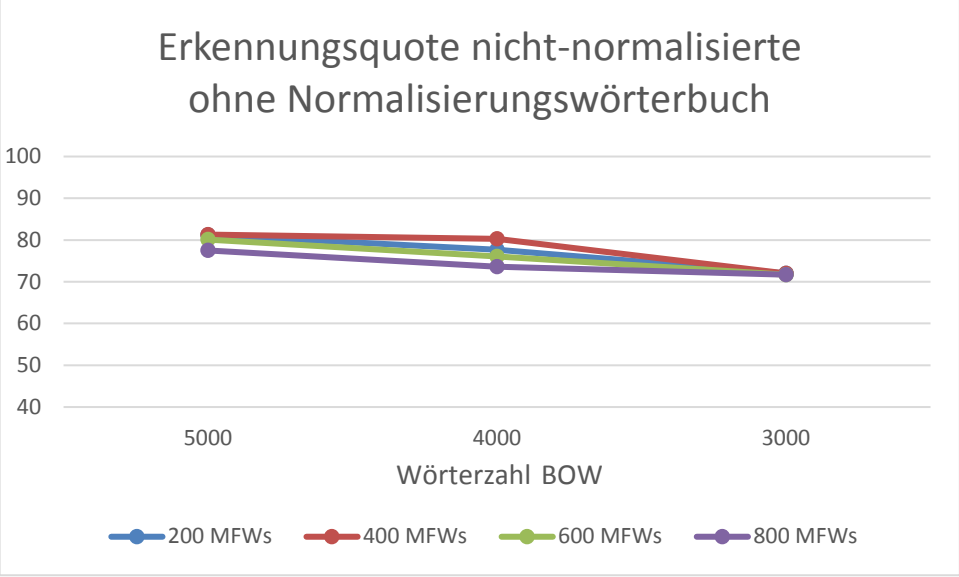
Höhere Erkennungsquote durch Normalisierungswörterbuch und Z-Wert-Begrenzung:

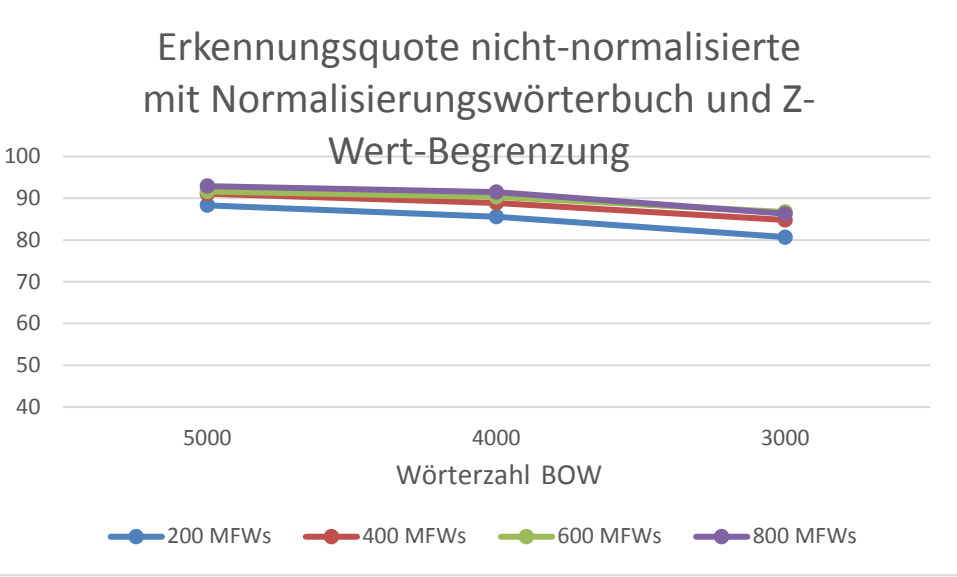
Steigerung von 80% auf 91% (bei 5.000 Wortformen in den Bag-of-Words)

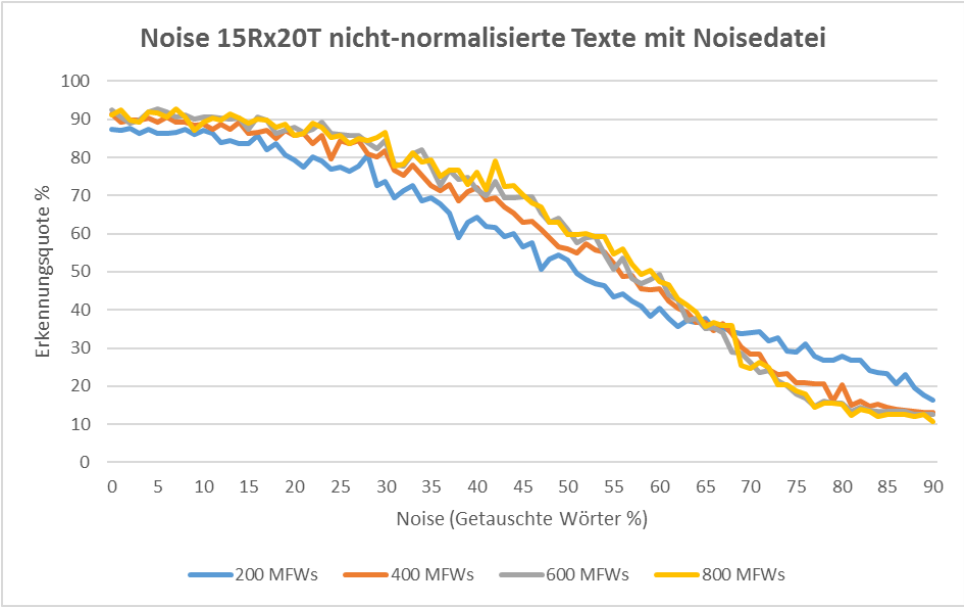
Nicht-normalisiertes Ratekorpus: 15 Texte von 10 Autoren.

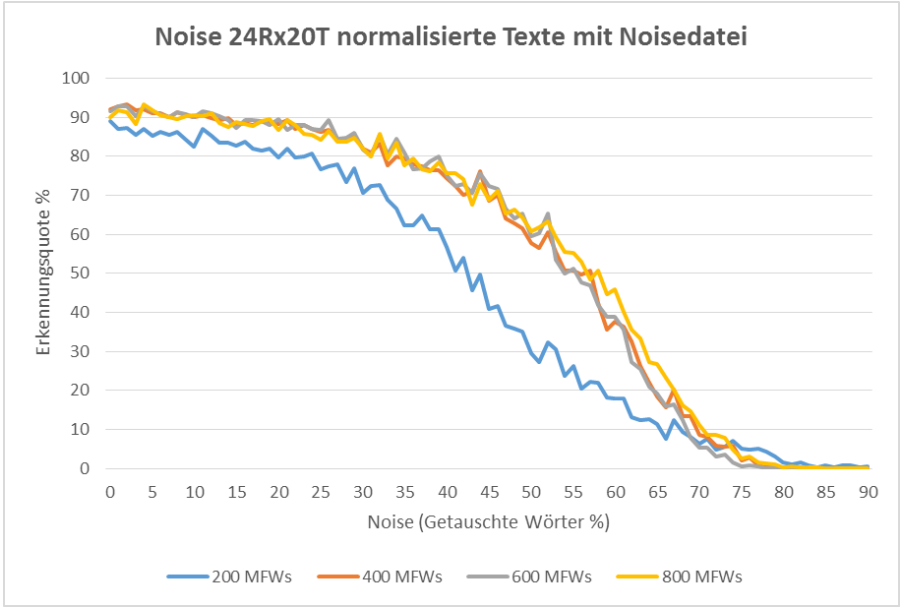
Gemischtes Trainingskorpus: 20 Texte. Wenn verfügbar: nicht-normalisierter Trainingstext zum Ratetextautor, sonst normalisierter Trainingsstext zum Ratetextautor.

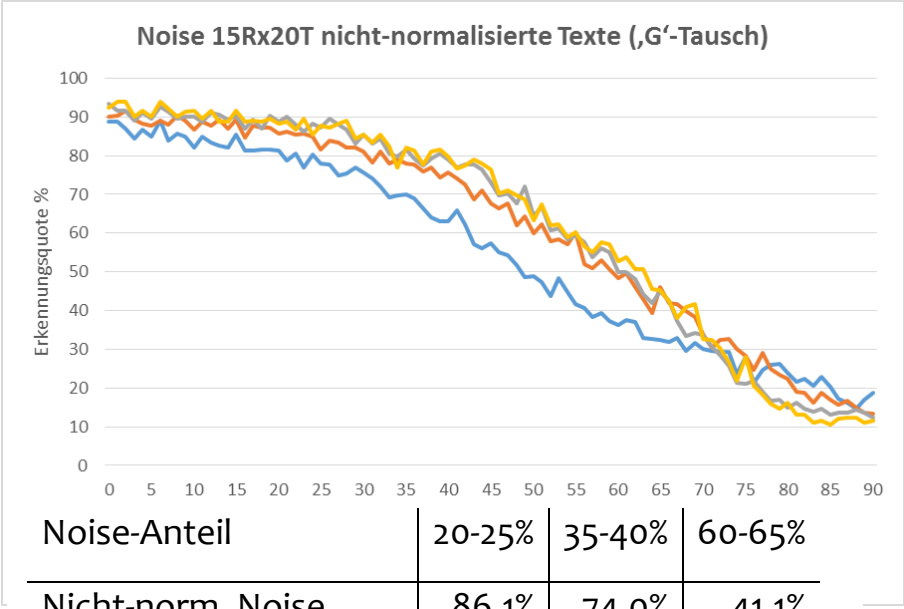
Zwert-Begrenzung (Evert): Z-Werte ab |1,64| werden auf 1,70 gesetzt.











Noise-Anteil	20-25%	35-40%	60-65%
Nicht-norm. Noise	86,1%	74,0%	41,1%
Normalisierte Noise	87,3%	77,5%	30,6%
Nicht-norm. ‚G'-Noise	86,9%	78,9%	47,2%
Normalisierte ‚G'-Noise	86,8%	74,4%	16,8%