Query Spelling Correction

Verteidigung der Bachelorarbeit
von Anja Rathgeber

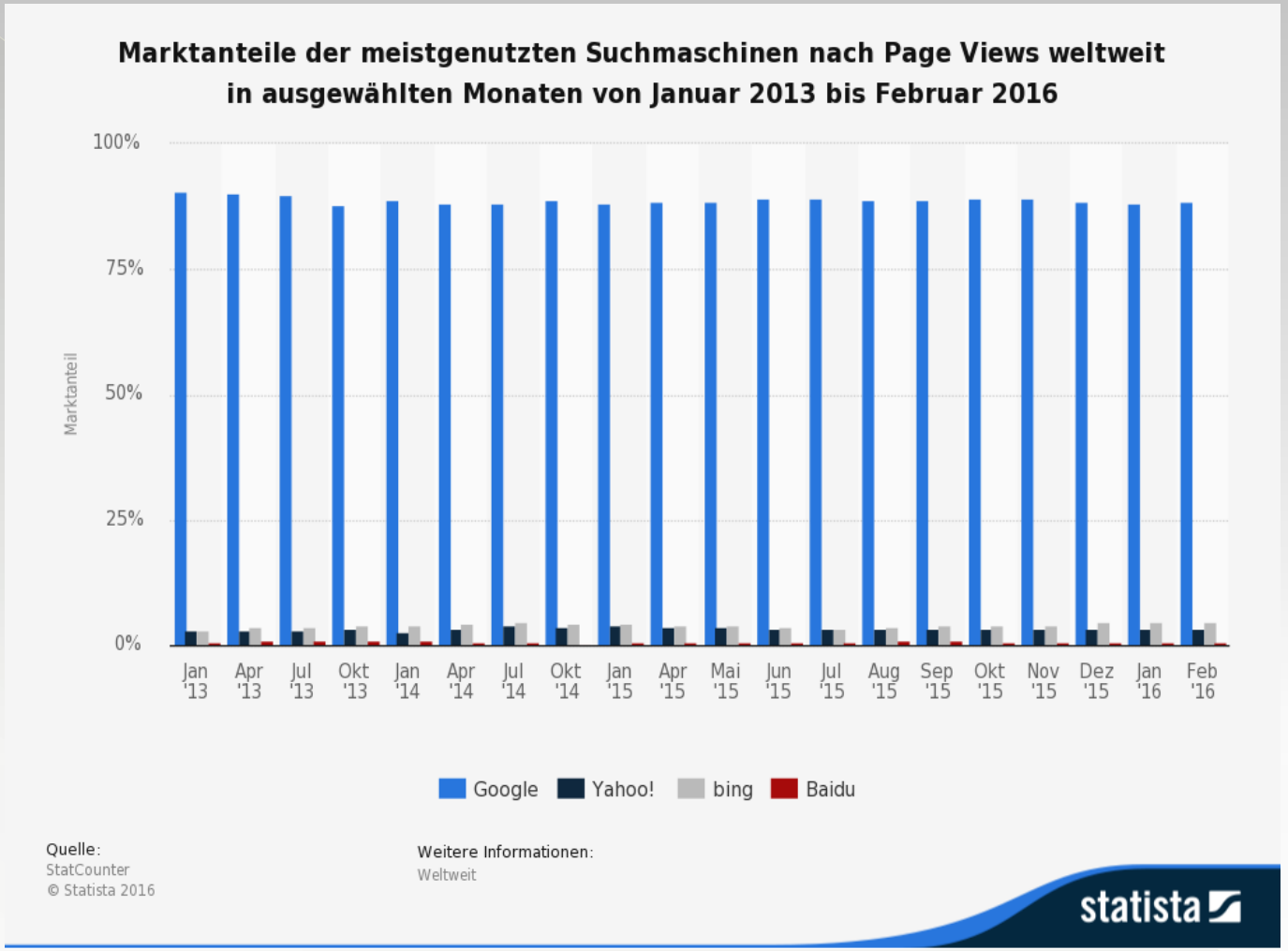
1. Gutachter: Junior-Prof. Dr. Matthias Hagen
2. Gutachter: Dr. rer. nat. Martin Potthast

Inhaltsverzeichnis

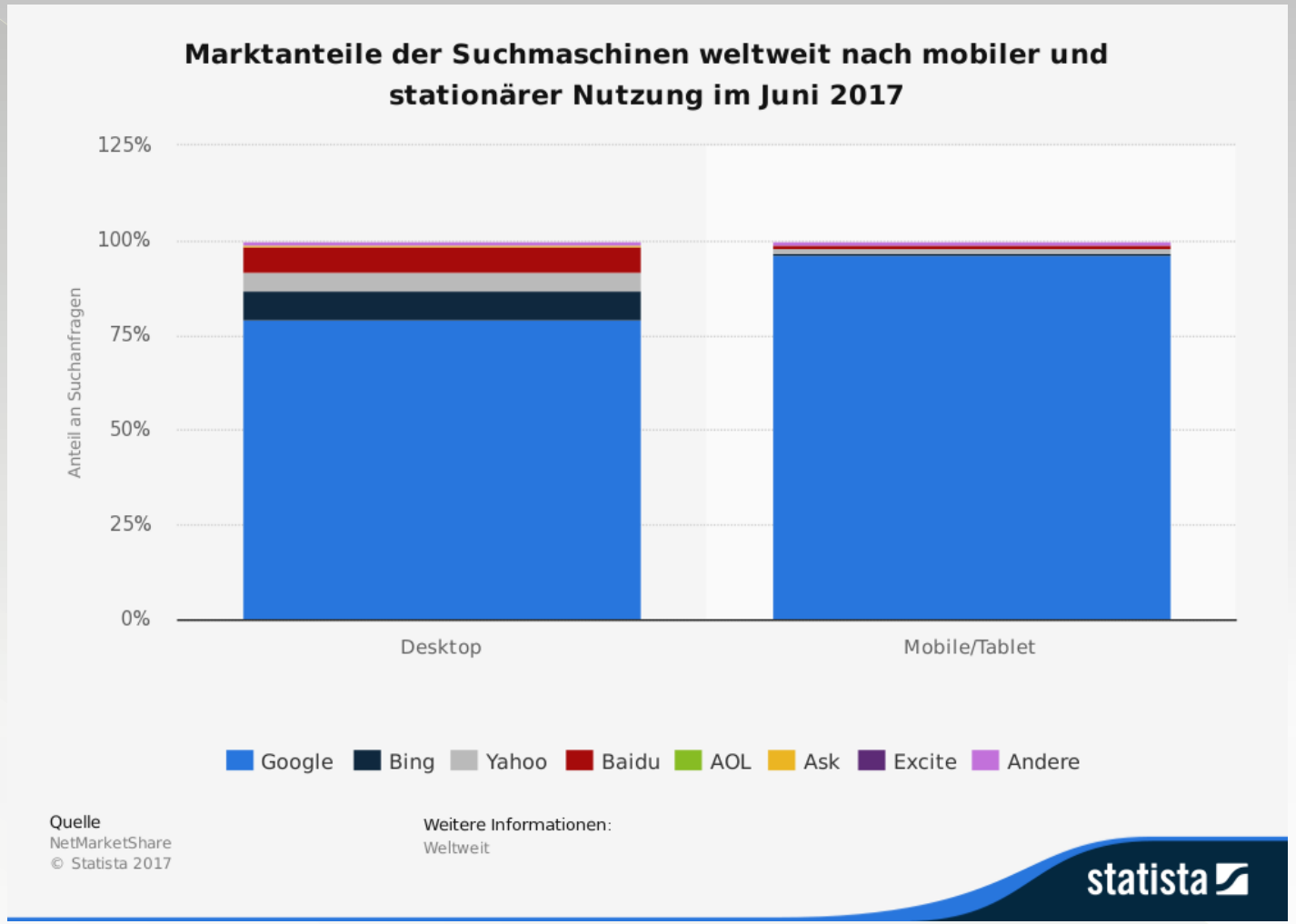
- Motivation
- Suchmaschinen
- Fehlerarten
- Vergleichbares Korpus
- Webis Query Spelling Correction 2016 Korpus
- Korpusanalyse
- Testen des neuen Korpus
- Fazit und Ausblick

Motivation

Suchmaschinen



Suchmaschinen



Fehlerarten

- **Rechtschreibfehler**

  aus Unwissen oder Phonetischer Fehler

  Bsp.: slay *[sley]* ≠ sleigh *[sley]*

- **Insertion**

  mindestens ein Buchstabe zu viel durch

  „Fat Finger“ oder „Long Press“

  Bsp.: statt *search* → *seasrch* oder *seaarch*

- **Deletion**

  mindestens ein Buchstabe fehlt, z.B. durch zu schnelles Tippen

  Bsp.: statt *search* → *serch* oder *searh*

Fehlerarten

- **Substitution**

  mindestens ein Buchstabe durch einen anderen ersetzt, z.B. durch ungewohnte Tastatur

  Bsp.: statt *search* → *sesrch* oder *seqrch*

- **Transposition**

  Vertauschung von zwei Buchstaben, z.B. durch blindes Schreiben

  Bsp.: statt *search* → *saerch* oder *serach*

Vergleichbares Korpus

- *Speller Challenge TREC Data* im Januar 2011 von Microsoft im Rahmen der Microsoft Speller Challenge veröffentlicht (entstanden aus „2008 Million Query Track“ Daten-Set)
- 5892 Anfragen, wovon 311 als falsch geschrieben bewertet wurden
- 1122 Anfragen, mit mindestens einem zum Original unterschiedlichen Vorschlag zur Schreibweise
- Korpus besteht nur aus Kleinbuchstaben und enthält keine Sonderzeichen
- Jede einzelne Anfrage geprüft durch bis zu drei unabhängige Experten

Webis Query Spelling Correction 2016 Korpus

- **Ziel:** *Webis Query Spelling Correction 2016 Korpus*, welches für jede fehlerhafte Anfrage mindestens eine Korrektur enthält
- 2010 entstandene *Webis Query Segmentation Corpus* mit 54.944 Anfragen diente als Grundlage
- Enthielt für 8.364 Anfragen genau eine Korrektur durch damalige Fehleranalyse

Webis Query Spelling Correction 2016 Korpus

- **1. Entfernung von Duplikaten**

  →19 Anfragen entfernt

- **2. Semiautomatische Rechtschreibkontrolle**

  → Nutzung von umfangreichen englischen Wörterbuch und zusätzlich Liste von Marken und Firmen

  → Mit Python-Script jede Zeile Wort für Wort durchlaufen und überprüft

  → Berechnung der Editier-Distanz zwischen einem falschen Wort und einer möglichen Korrektur

  → Korrekturvorschläge für 20.123 Wörter, welche anschließend manuell geprüft wurden mittels Google

  → Annotation von Sonderzeichen und Akzenten (853 Anfragen)

  → 140 Anfragen aus anderen Sprachen entfernt

Webis Query Spelling Correction 2016 Korpus

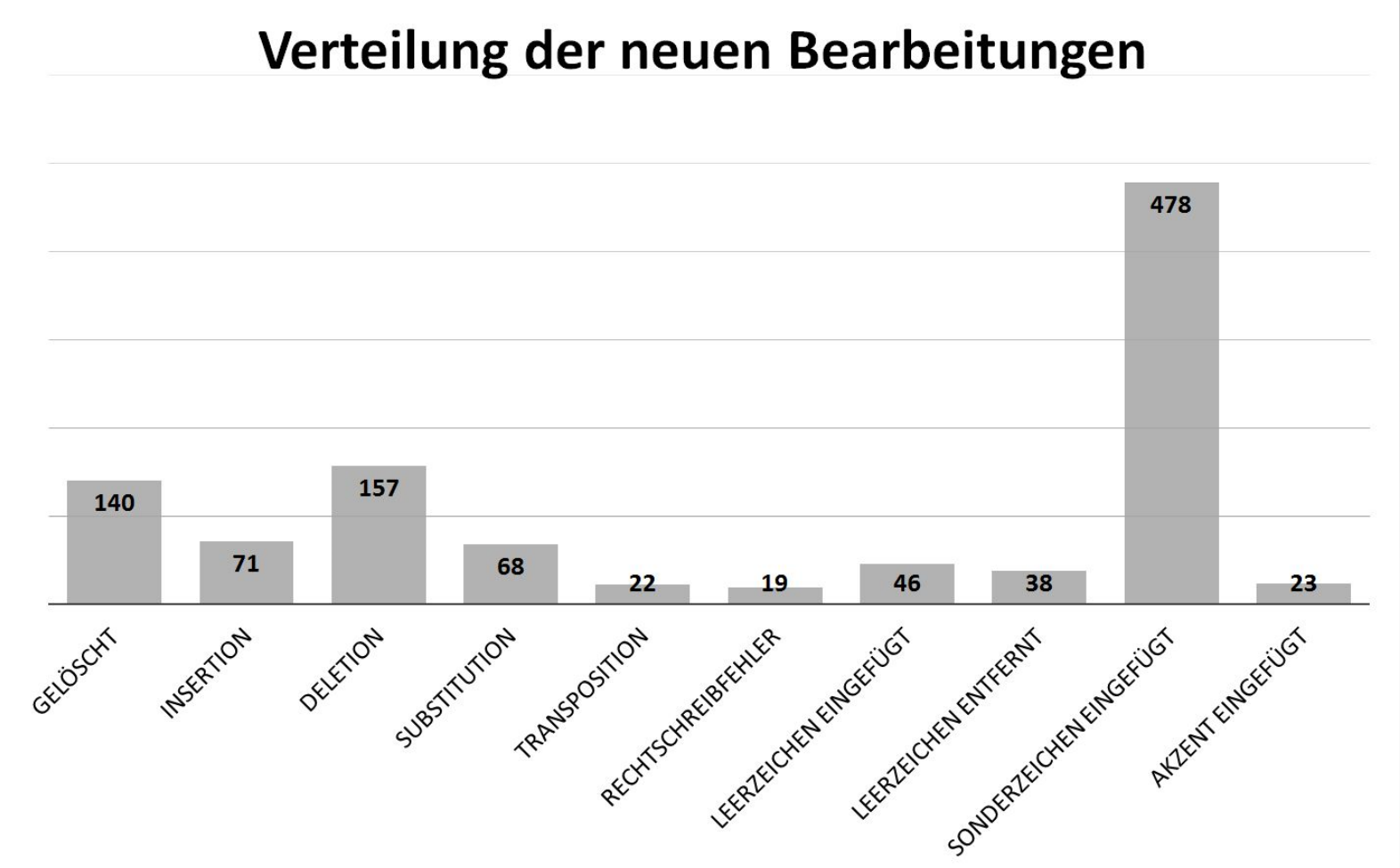


Abbildung 1: Fehlerverteilung der bearbeiteten Anfragen im zweiten Schritt

Webis Query Spelling Correction 2016 Korpus

- **3. Review der Rechtschreibkontrolle von 2010**

  → Erneute manuelle Kontrolle der damaligen 8.364 Korrekturen

  → Für rund 350 Anfragen neue Entscheidung über Korrektur

  - **Schwierigkeiten:** Entscheidung wann Korrektur sinnvoll, weil der Intent der Anfrage betroffen ist, Ausschreiben von Abkürzungen, Varianten bei verschiedenen Möglichkeiten

Webis Query Spelling Correction 2016 Korpus

- **Spalte 1:** Originale Anfrage des Nutzers
- **Spalte 2:** Bei fehlerhaften Anfrage die korrigierte Variante (inklusive fehlender Sonderzeichen) Wenn keine Korrektur => originale Anfrage aus Spalte 1
- **Spalte 3:** Erste mögliche weitere Variante, ausgeschriebene Variante für Buchstabenauslassungen, bei Akzentzeichen wurden diese hier von den Vokalen entfernt, Et-Zeichen „&“ durch das Wort „and“ ersetzt, sonst leer
- **Spalte 4 und 5:** Weitere mögliche Varianten für die Anfragen, sonst leer
- **Spalte 6 und 7:** Leer für optische Trennung

Webis Query Spelling Correction 2016 Korpus

- **Spalte 8:** Entfernung aller Sonderzeichen nach vorher festgelegten Regeln ausgehend von Spalte 2, Akzentzeichen von Vokalen wurden hier ebenfalls entfernt, somit für jede Anfrage eine Variante ohne Sonderzeichen
- **Spalte 9, 10 und 11:** Ansatz von Spalte 8 angewandt auf die Varianten aus Spalte 3,4 und 5
- **Spalte 12, 13 und 14:** Leer für optische Trennung
- **Spalte 15:** Kommentarspalte ausschließlich zur Information über Fehlerart oder andere Bemerkungen

Webis Query Spelling Correction 2016 Korpus

- Insgesamt 54.772 Anfragen
- Bei 9.033 Anfragen wurden Fehler gefunden und Bearbeitungen vorgenommen
- Für 643 Anfragen wurde eine weitere Variante der Schreibweise hinzugefügt

  → Davon für 13 Anfragen eine dritte Variante

  → 4 Anfragen zusätzlich noch eine vierte Variante
- Kommentare wurden für 9.044 Anfragen eingefügt

Korpusanalyse

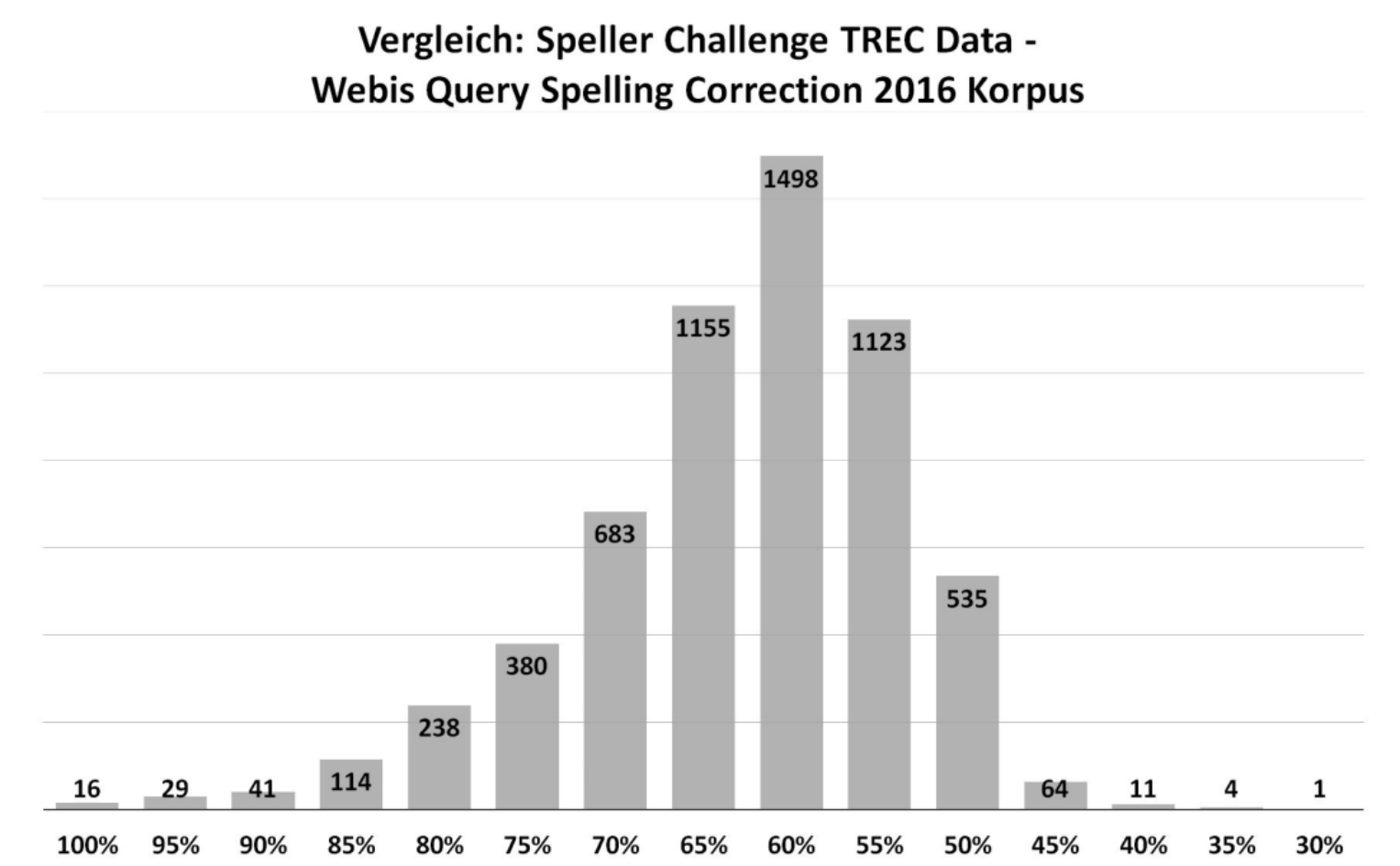


Abbildung 2: Vergleich von Speller Challenge TREC Data und Webis Query Spelling Correction 2016 Korpus auf Ähnlichkeit

Korpusanalyse

	Webis Query Spelling Correction 2016		Speller Challenge TREC Data	
	Absolut	Relativ	Absolut	Relativ
1 Wort	0	0,000%	178	3,021%
2 Wörter	1	0,002%	629	10,675%
3 Wörter	24477	44,689%	988	16,768%
4 Wörter	14933	27,264%	817	13,866%
5 Wörter	7841	14,316%	534	9,063%
6 Wörter	3887	7,097%	1381	23,439%
7 Wörter	1884	3,440%	762	12,933%
8 Wörter	969	1,769%	390	6,619%
9 Wörter	507	0,926%	195	3,310%
10 Wörter	273	0,498%	18	0,305%

Tabelle 1: Vergleich von *Speller Challenge TREC Data* und *Webis Query Spelling Correction 2016 Korpus* auf Anfragenlänge

Korpusanalyse

- **Fehleranalyse**
  - Damerau-Levenshtein-Distanz (1965)

$$D_{0,0} = 0$$
$$D_{i,0} = i \qquad 1 \leq i \leq m$$
$$D_{0,j} = j \qquad 1 \leq j \leq n$$

$$D_{i,j} = min \begin{cases} D_{i-1,j-1} & + 0 \; falls \; u_i = v_j \\ D_{i-1,j-1} & + 1 \; (Substitution) \\ D_{i,j-1} & + 1 \; (Insertion) \\ D_{i-1,j} & + 1 \; (Deletion) \end{cases}$$
$$(i = 1, 1 \leq j \leq n) \vee (1 \leq i \leq m, j = 1)$$

$$D_{i,j} = min \begin{cases} D_{i-1,j-1} & + 0 \; falls \; u_i = v_j \\ D_{i-1,j-1} & + 1 \; (Substitution) \\ D_{i,j-1} & + 1 \; (Insertion) \\ D_{i-1,j} & + 1 \; (Deletion) \\ D_{i-2,j-2} & + c \; (Transposition), falls \; u_i = v_{j-1} \wedge u_{i-1} = v_j \end{cases}$$
$$2 \leq i \leq m, 2 \leq j \leq n$$

Korpusanalyse

`spelimgcorrectoin` → `spelimgcorrectoin`

`spelling correction` → `spellingcorrection`

1. Prüfe Leerzeichen und Sonderzeichen
2. Damerau-Levenshtein-Matrix wird erstellt
3. Backtrace der Matrix zur Fehlerermittlung

```
spel imgcorrectoin
spellingcorr ection
eeeedeseeeeeieeetee
```

e = equal, i = Insertion, d = Deletion, s = Substitution, t = Transposition

4. Zählen der Fehler

→ 1 Leerzeichen-Fehler, 0 Sonderzeichen-Fehler, 1 Insertion-Fehler, 1 Deletion-Fehler, 1 Substitutions-Fehler und 1 Transformations-Fehler vor

Korpusanalyse

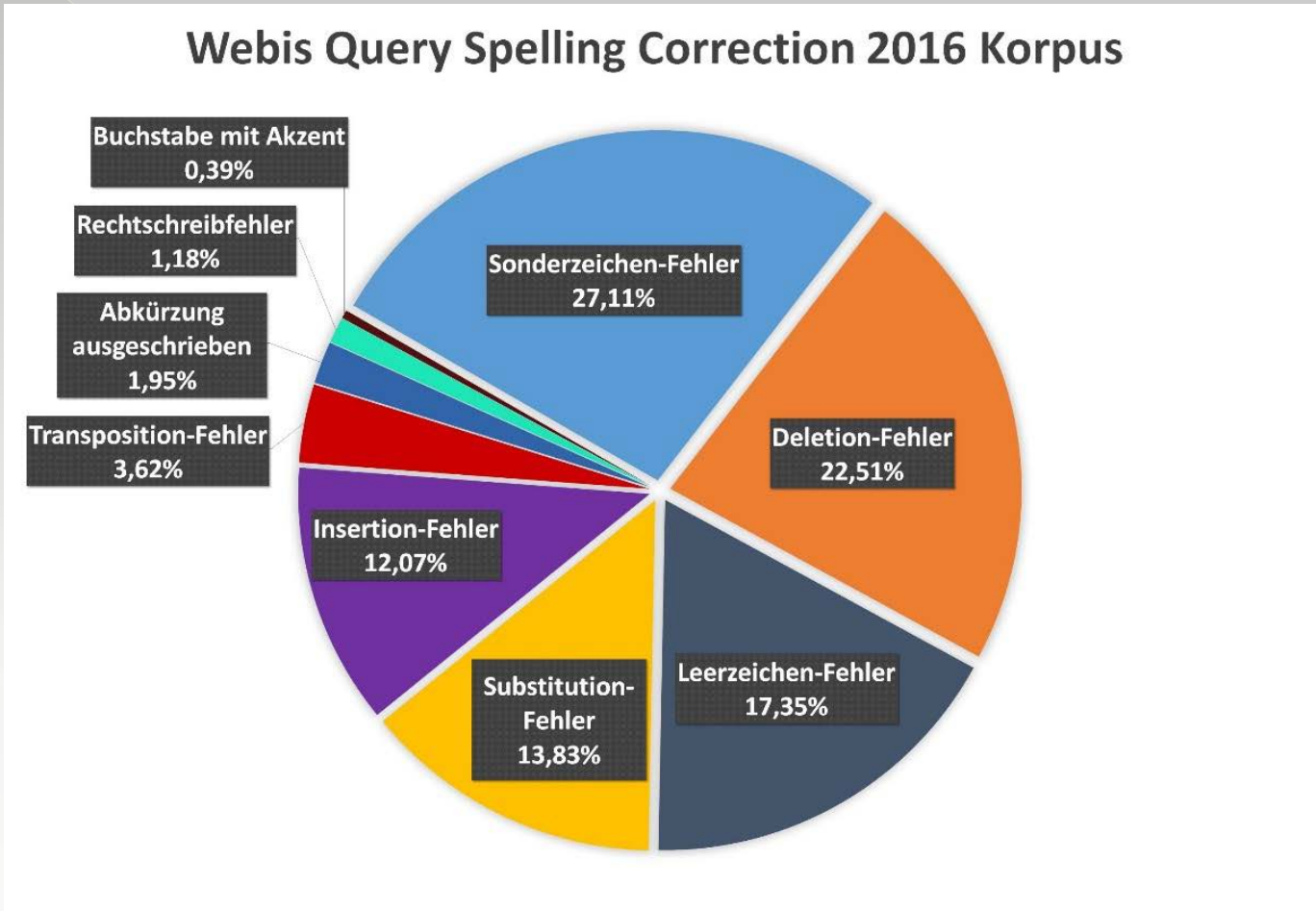


Abbildung 3: Fehlerverteilung für Webis Query Spelling Correction 2016 Korpus

Korpusanalyse

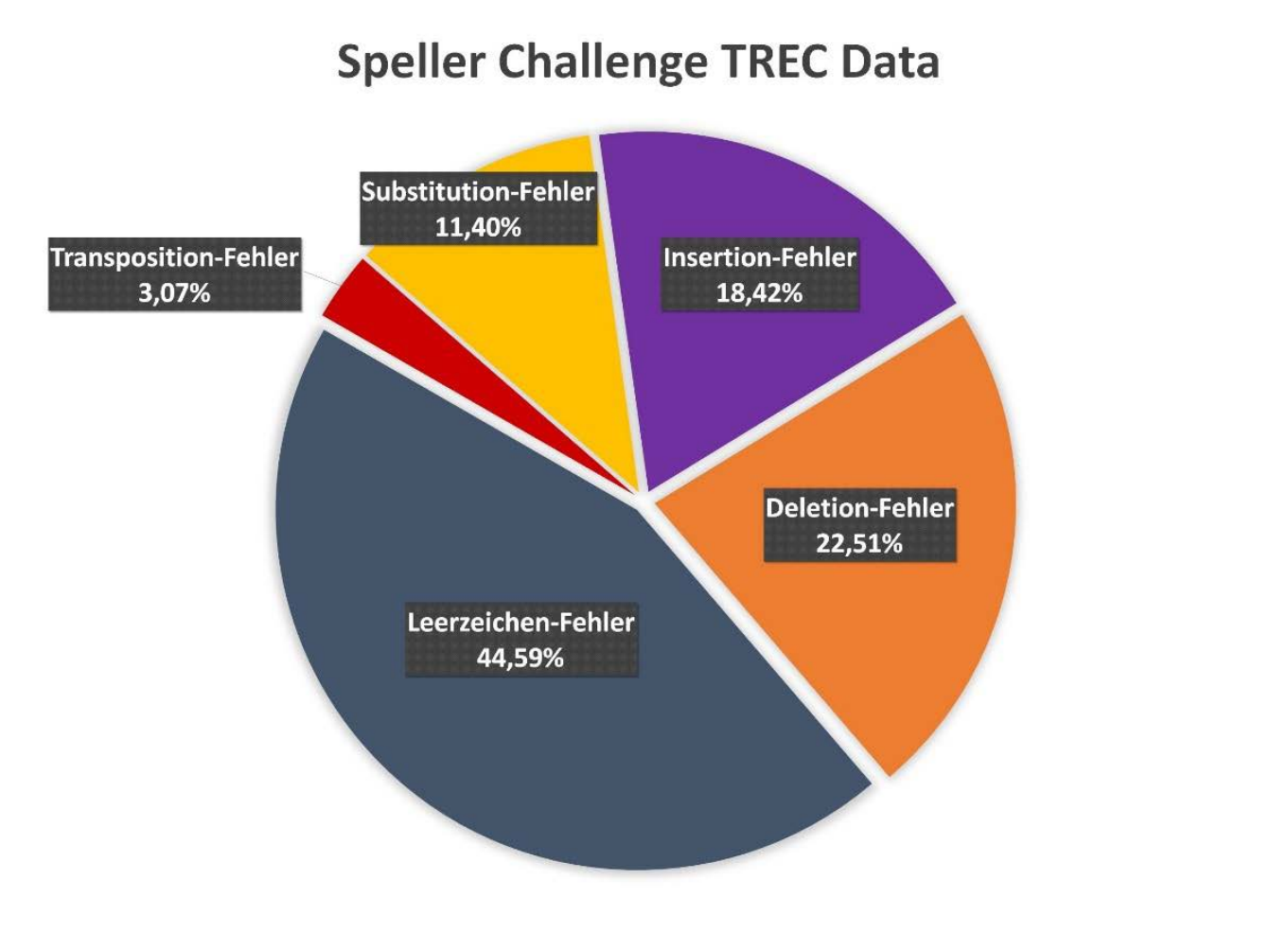


Abbildung 4: Fehlerverteilung für Speller Challenge TREC Data Korpus

Testen des neuen Korpus

„Pythia“

- Beitrag von Peter Nalyvayko zur Microsoft Speller Challenge
  - Liste von möglichen Korrekturen wird mit einem trainierten Noisy Channel Model erstellt
  - Modifizierte Levenshtein-Distanz, um die Similarität der korrigierten und fehlerhaften Wörter zu berechnen

    (verschiedene Wichtung: Deletion und Insertion: 1,2 , Substitutionen werden: 2,0 und Transpositionen: 1,0)
  - Ersten Wertung über die bedingte Wahrscheinlichkeit
- Bewertung der Wahrscheinlichkeiten und finale Sortierung mittels Microsoft Web N-gram Service

Testen des neuen Korpus

- F-Maß (harmonischen Mittel zwischen Genauigkeit und Trefferquote) ermöglicht einen direkten Vergleich des *Webis Query Spelling Correction 2016 Korpus* mit dem *Speller Challenge TREC Data* Korpus

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

$$Genauigkeit\ (precision) = \frac{Anzahl\ der\ \text{übereinstimmenden}\ Korrekturen}{Anzahl\ an\ vorgeschlagen\ Korrekturen\ von\ "Pythia"}$$

$$Trefferquote\ (recall) = \frac{Anzahl\ der\ \text{übereinstimmenden}\ Korrekturen}{Anzahl\ der\ Korrekturen\ im\ Korpus}$$

Testen des neuen Korpus

	Speller Challenge TREC Data	Webis Query Spelling Correction 2016 Korpus
F-Maß	0.9185730419	0,6579752269
Genauigkeit	0.9352770727	0,5331053872
Trefferquote	0.9024552090	0,8592346752

Tabelle 2: Evaluierung von Speller Challenge TREC Data und Webis Query Spelling Correction 2016 Korpus

Probleme:

- Keine Korrektur von Sonderzeichen durch „Pythia“
- Für einen Teil der Anfragen liefert „Pythia“ eine hohe Anzahl an möglichen Korrekturen, welche allerdings zu einem großen Teil als sehr unwahrscheinlich bewertet werden

Fazit und Ausblick

- **Fazit:**
  - *Webis Query Spelling Correction 2016 Korpus* aus fast 55.000 Suchanfragen mit Korrektur aller Fehler
  - Schwierigkeit: kontextbasierte Korrektur von Anfragen → individuell für jede Anfrage entschieden
  - Hinzufügen von Varianten bei verschiedenen möglichen Schreibweisen
  - Fertiges *Webis Query Spelling Correction 2016 Korpus* kann nun für ein umfangreiches Training von Algorithmen zur *query spelling correction* genutzt werden

- **Ausblick:**
  - Korrekte Annotation von Großschreibungen zur weiteren Optimierung
  - Mögliche Ergänzung von Varianten mit kontextuellen Inhalt der Anfrage

Vielen Dank für Ihre Aufmerksamkeit.

Für Fragen stehe ich Ihnen jetzt zur Verfügung.