Ruprecht-Karls-Universität Heidelberg
WS07/08

Stemmingverfahren

Éva Mújdricza
Ganna Syrota

Information Retrieval, HS, WS07/08
Dozentin: PD Dr. Karin Haenelt

04.02.2008

Übersicht

- I: Stemmingverfahren
  - Grundlagen
  - Eigenschaften
  - Stemming in Suchmaschinen
  - Evaluierung
  - Typische Fehler
  - Flaches und tiefes Stemming
- II: Stemmer
  - Stemmerarten
  - Porter-Stemmer für das Deutsche
- III: Entwicklung eines Stemmers
  - für das Ukrainische
  - (Porter-Stemmer für das Ungarische)
- IV: Zusammenfassung

Übersicht

- I: Stemmingverfahren
  - Grundlagen
  - Eigenschaften
  - Stemming in Suchmaschinen
  - Evaluierung
  - Typische Fehler
  - Flaches und tiefes Stemming
- II: Stemmer
  - Stemmerarten
  - Porter-Stemmer für das Deutsche
- III: Entwicklung eines Stemmers
  - für das Ukrainische
  - (Porter-Stemmer für das Ungarische)
- IV: Zusammenfassung

Grundlagen

- **Das Ziel des IR:**

  möglichst gute **Suchergebnisse** zu liefern. Dafür werden verschiedene **Verfahren** eingesetzt.

- **Stemming** (Grundformenreduktion) ist ein Verfahren, mit dem verschiedene **morphologische Varianten** eines Wortes auf ihren **gemeinsamen Wortstamm** (stem) zurückgeführt werden

- **Die Idee:**

  die eigentliche **lexikalische Bedeutung** eines Wortes ist in seinem

  Stamm zu finden → man sucht nicht nach einer bestimmten Wortform, sondern nach möglich vielen Wortformen:

  - *Bruder – Bruders – brüderlich – Brüderlichkeiten* → *bruder*
  - *essen – aßen – essbar* → *ess*

Eigenschaften des Verfahrens

- Das Besondere an diesem Verfahren: **conflation (**Zusammenführung der Varianten eines Stammes) erfolgt möglichst **ohne morphologische Analyse**;
- leicht zu implementieren;
- Reduzieren der Filegröße bei der Indexierung (bis zur 50% durch das Speichern der Stämme anstatt der Terme);

Stemming in den Suchmaschinen

- Das Stemmingverfahren wird in folgenden internationalen Suchmaschinen verwendet:
  - Google
  - Lucene
  - Yahoo!
  - AOL-Search
  - Ask.com
  - dtSearch
  - Netscape Search

Evaluierung

- **Korrektheit** (correctness): Wie viele Stämme richtig ermittelt werden;
- **Wortanzahl-Stamm-Verhältnis** (Number of words per conflation class);

$$WSV = \frac{N}{S}$$

N : Wortformenzahl vor dem Stemming
S : Stammanzahl nach dem Stemming

- **Komprimierungsrate** (index compression): $K = \frac{N - S}{N}$

- Auswirkung auf die **Suchleistung** (durch **Precision** und **Recall** gemessen). Stemming verbessert den **Recall** fast immer und verschlechtern in der Regel die **Precision**; generell: neutrale oder positive Auswirkung (Frakes: 150)

Typische Fehler

- **Überstemmen (overstemming)**: zu viel wird entfernt → nichtverwandte Wörter werden zu einem Stamm zusammengefasst oder nicht existierende Stämme werden ermittelt:

  Politik → polit

- **Unterstemmen (understemming):** zu wenig wird entfernt → verwandte Wörter werden nicht als zusammengehörende erkannt.

  gehen → geh ↔ geht → geht

Flaches vs. tiefes Stemming

- **Flaches (nichtlinguistisches)** Stemming basiert auf statistischen Verfahren oder auf externen Datenbanken. Der Stamm wird nicht nach morphologischen Kriterien ermittelt, sondern möglichst einfach, ohne linguistisches Hintergrundwissen. → Der ermittelte Stamm ist oft **nicht grammatisch korrekt:**

  beauty (Grammatik) vs. *beauti* (Stemming)

- Eine Alternative: **Lemmatisierung**, die die Wortform auf das Lemma reduziert. Die graphemischen Änderungen (*lassen* – *ließ*) und unregelmäßige Worformen (*bringen* – *brachte*) können auch beachtet werden.

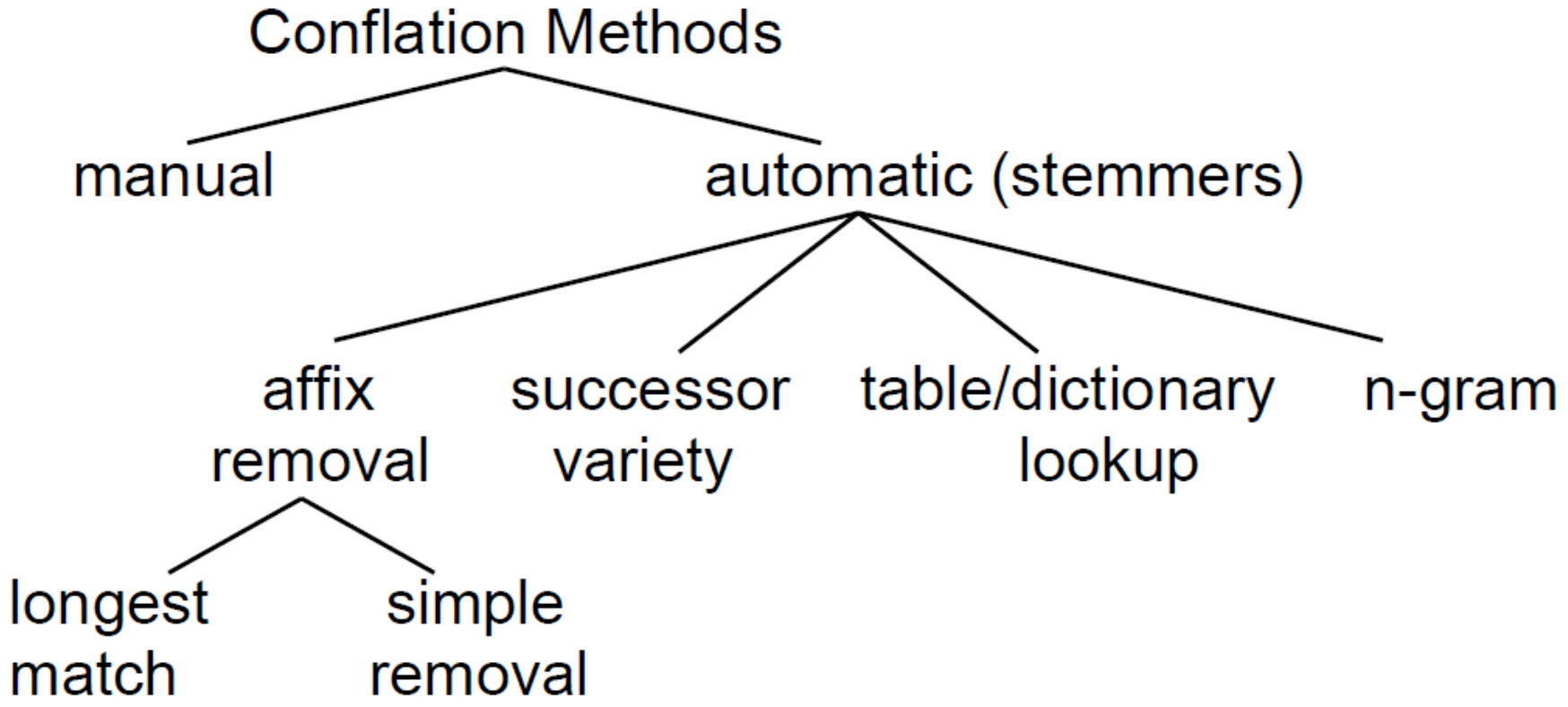
- **Tiefes (linguistisches) Stemming** bezieht auch linguistisches Wissen ein.

Übersicht

- I: Stemmingverfahren
  - Grundlagen
  - Eigenschaften
  - Stemming in Suchmaschinen
  - Evaluierung
  - Typische Fehler
  - Flaches und tiefes Stemming
- II: Stemmer
  - Stemmerarten
  - Porter-Stemmer für das Deutsche
- III: Entwicklung eines Stemmers
  - für das Ukrainische
  - (Porter-Stemmer für das Ungarische)
- IV: Zusammenfassung

Stemmerarten

- nach Frakes: 132



Stemmer: N-Gramm-Stemmer

- Zählt die **Bigramme** (**N-Grammme)**, die zwei Wortformen **gemeinsam** haben.
- Ähnlichkeitsmaß mit **Dice-Koiffizient** wird für jedes Wortformpaar im Korpus berechnet → **Ähnlichkeitsmatrix**. Die Wortformen werden geclustert (single link clustering).

$$S = \frac{2 \cdot (\text{N-Gramme}_{W1} \cap \text{N-Gramme}_{W2})}{\text{N-Gramme}_{W1} + \text{N-Gramme}_{W2}}$$

- **Beispiel** (nach Frakes, S.136):

W1: *statistics* → *st ta at ti is st ti ic cs*
Bigrammmenge: {*at cs ic is st ta ti*} (7)

W2: *statistical* → *st ta at ti is st ti ic ca al*
Bigrammmenge: {*al at ca ic is st ta ti*} (8)

Gemeinsame Bigrammmenge:
{*at, ic, is, st, ta, ti*} (6)

⇓

Ähnlichkeit: $S = \frac{2 \cdot 6}{7+8} = 0{,}8$

Stemmerarten: Lookup

- In einer Tabelle (**Table Lookup**) oder in einem Wörterbuch (**Dictionary Lookup**) wird für jede Wortform der Stamm gespeichert.
- **Beispiel** (Frakes 133):

Term	Stem
engineering	engineer
engineered	engineer
engineer	engineer

- Precision ist durch die/das gespeicherte Tabelle/Wörterbuch gewährleistet.
- Der Aufbau des Systems ist zeit- und arbeitsintensiv und die Tabelle/ das Wörterbuch braucht regelmäßig Pflege.

Stemmerarten: Successor Variety

- Der **Nachfolgervielfalt**-Algorithmus (successor variety) basiert auf Untersuchungen, die für einen betrachteten Buchstaben im Wort die **möglichen Nachfolgebuchstaben** ermittelt haben (Hafer und Weiss 1974). Dabei wurde festgestellt, dass die Anzahl der möglichen Nachfolgerbuchstaben mit der Länge der Wortform oft abnimmt.
- **Nachfolgervielfalt** (NFV): wie viele und welche Buchstaben können in einem Korpus einem gegebenen Buchstaben an der i-ten Position des Wortes folgen.

Stemmerarten: Successor Variety

- **Beispiel** (Frakes, 135): Testwort: ***readable***
- KORPUS: *able, ape, beatable, fixable, read, readable, reading, reads, red, rope, ripe*

1. Ermittlung der NFV für das Testwort
2. Ermittlung der Wortsegmente.
3. Ein Segment als Stamm auswählen: Das erste Segment, wenn es in höchstens 12 Wörtern im Korpus vorkommt, sonst das zweite. (Das erste Segment könnte auch ein Präfix sein.)

- Ergebnis: *read* + *able*

Prefix	Successor Variety	Letters
r	3	e, i, o
re	2	a, d
rea	1	d
read	3	a, i, s
reada	1	b
readab	1	l
readabl	1	e
readable	1	BLANK

Stemmerarten: Affix Removal

- Stemming durch **Entfernung von Derivations- und Flexionsaffixen**.
  - oft werden **nur Suffixe** behandelt
- Diese Art ist am weitesten verbreitet.
- Überprüft die Eingabe nach definierten Affixen und entfernt sie in einer bestimmten Reihenfolge.
  - oft iterativ: Die Regeln können wiederholt angewendet werden (bis zu einem Abbruchkriterium).
- Der Stamm kann auch nach der Entfernung von Affixen nochmal geändert werden – Nachbereitung.
- Affixentfernung:
  - oft gierige Algorithmen (z.B. Porter-Algorithmus)

    Überstemming → Präzisionsverlust
- Sprachabhängigkeit: für jede Sprache verschiedene Regeln und Bedingungen.

Stemmerarten im Vergleich

Stemmerart	schnell	Implementierung	linguistisches Wissen	zusätzlicher Speicherplatz
N-Gramm	ja	leicht	nein	nein
Lookup	ja/nein	leicht	nein	ja
Successor V	ja	leicht	nein	nein
Affix Removal	?	leicht	ja	nein

Übersicht

- I: Stemmingverfahren
  - Grundlagen
  - Eigenschaften
  - Stemming in Suchmaschinen
  - Evaluierung
  - Typische Fehler
  - Flaches und tiefes Stemming
- II: Stemmer
  - Stemmerarten
  - Porter-Stemmer für das Deutsche
- III: Entwicklung eines Stemmers
  - für das Ukrainische
  - (Porter-Stemmer für das Ungarische)
- IV: Zusammenfassung

Porter Stemmer für das Deutsche

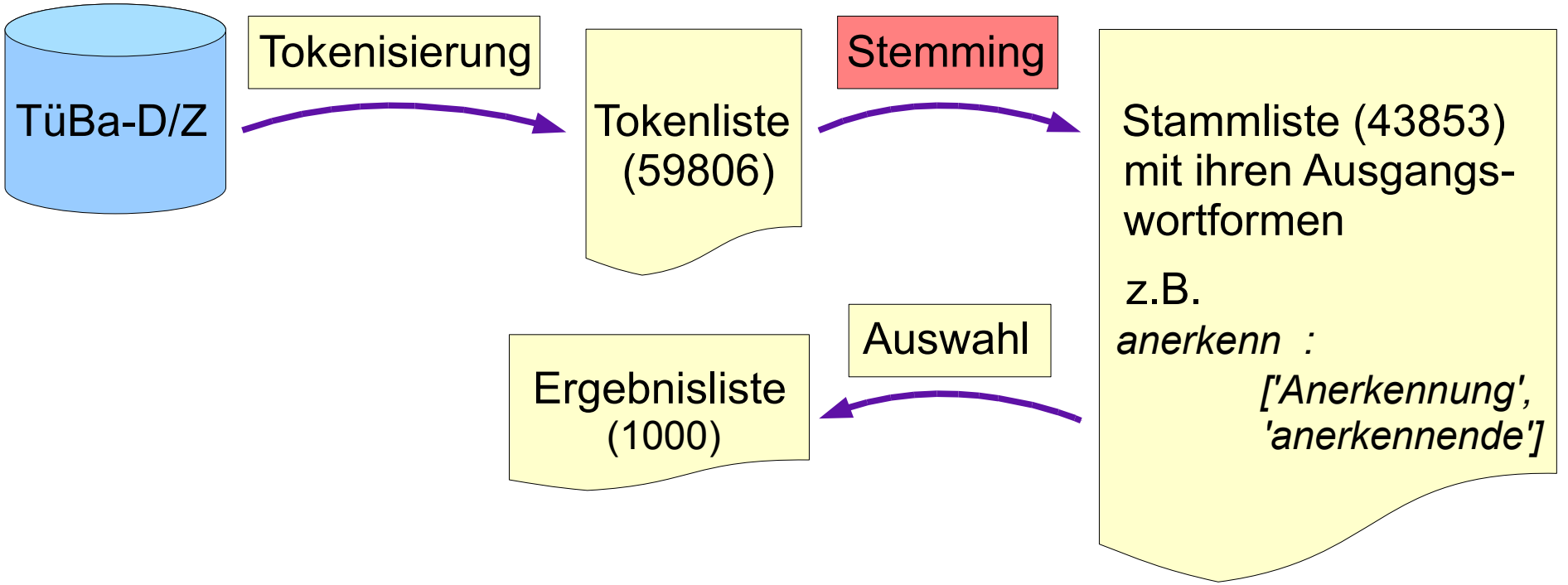
- Porter-Algorithmus: Martin Porter, 1979/80: Stemmer für das Englische
- Affix Removal Stemmer
- 3 Hauptschritte:
  1. Vorbereitungsschritt(e)
  2. Bearbeitungsschritt(e)
     - iterative Anwendung der Schritte möglich
  3. Nachbereitungsschritt(e)

- Ein Beispiel: Eine leicht veränderte Version des Stemmers für das Deutsche, den man unter ➳ http://snowball.tartarus.org/otherlangs/german_py.txt herunterladen kann.
- Der Stemmer musste für die Demonstration angepasst werden:

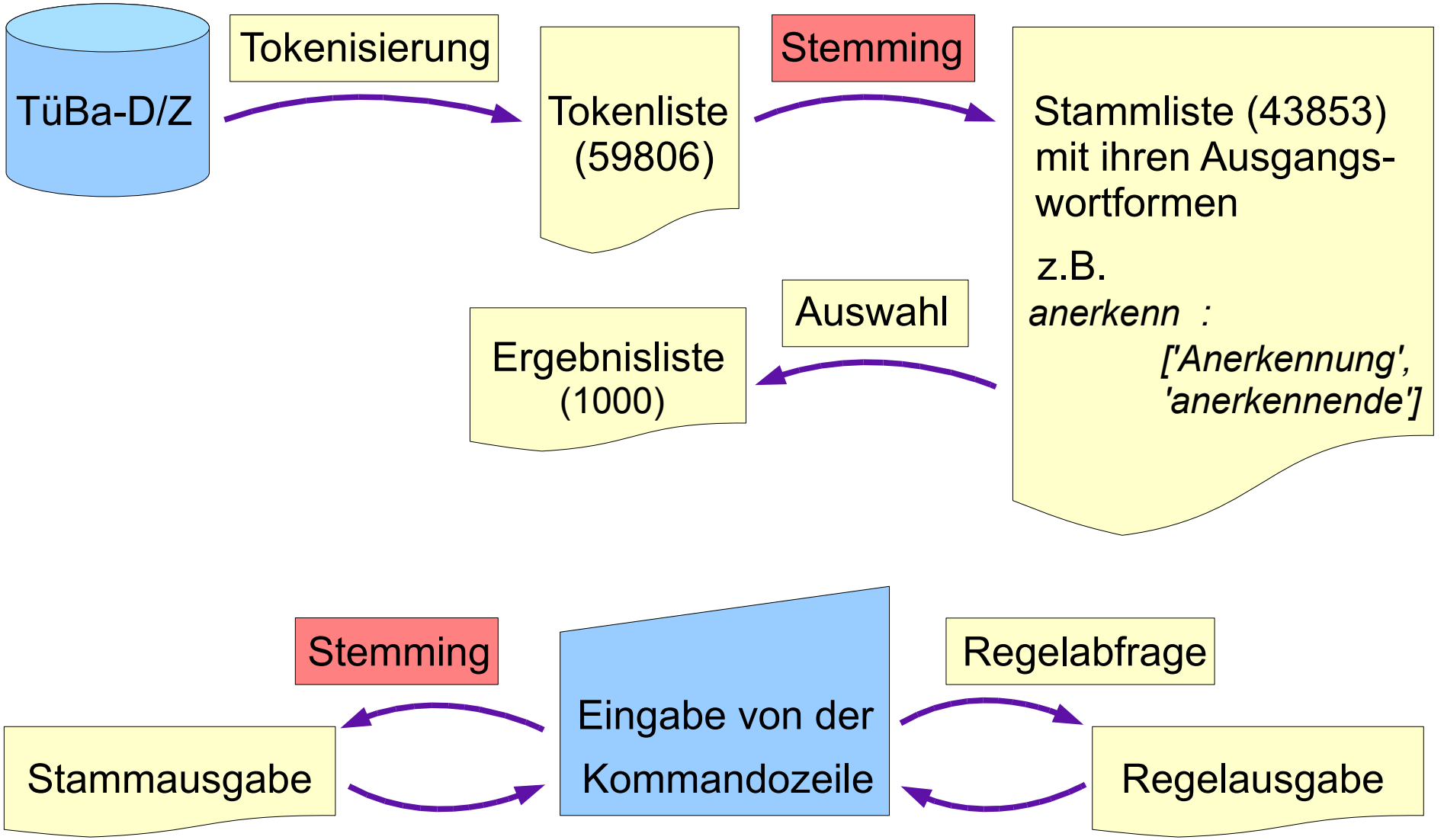
  → Stemming sowohl von der Kommandozeile aus, als auch aus Dateien möglich.

  → Zusätzliche Regelübersicht wurde ausgearbeitet.

Porter Stemmer (Dt.): Evaluierung



Porter Stemmer (Dt.): Evaluierung



Porter Stemmer (Dt.): 3 Schritte

- **I. Vorbereitung**
  - Definition der Vokalmenge (V) und der Konsonantenmenge (K)
  - Definition der Stoppwortliste
  - Initialisierung der Positionen P1, P2, R1, R2
  - Kleinschreibung des Eingabewortes
  - V*u*V > V*U*V, V*y*V > V*Y*V
- **II. Bearbeitung** der angegebenen Wortform: Anwendung von Reduktionsregeln mit bestimmten Bedingungen für die Entfernung von Flexions- und Derivationssuffixen.
  - Schritt 1, Schritt 2, Schritt 3
- **III. Nachbereitung**
  - Entfernung der Umlaute
  - *U* > *u*, *Y* > *y*
- ⇒ **Ausgabe** des ermittelten Stammes

Porter Stemmer (Dt.): 3 Schritte

- **R1**:
  - Entweder das Teilwort hinter der ersten VK-Folge im Wort, oder
  - das leere Wort, wenn es keine VK-Folge im Wort gibt.

  Beispiel: *Ges**undheit**, Zeit**ungen**, stum**pf**, bar_*
- **P1**: Startposition von R1

- **R2**:
  - Entweder das Teilwort im R1 hinter der ersten VK-Folge, oder
  - das leere Wort, wenn es im R1 nicht mehr als eine VK-Folge gibt.

  Beispiel: *Gesun**dheit**, Zeitun**gen**, stumpf_, bar_*
- **P2**: Startposition von R2

Porter Stemmer (Dt.): Schritt 2

- Die Bearbeitungsschritte: Stemmen von Nomen, Adjektive und Verben (eingeschränkt).

1	2	3a	3b	3c	3d
e	*est*	*igend*	*ig*	*erlich*	*lichkeit*
em	*er*	*igung*	*ik*	*erheit*	*igkeit*
en	*en*	-------	*isch*	*enlich*	-------
ern	-------	*end*		*erheit*	*keit*
er	*st*	*ung*		-------	
es				*lich*	
-------				*heit*	
s					

- Reihenfolge: zuerst Flexions-, danach Derivationsendungen (nur Suffixe)
- Gierige Methode: möglichst viel entfernen – auch mehrere Endungen in einem Schritt.
- Nicht iterativ, aber die gleichen Endungen kommen in mehreren Schritten vor.

Porter Stemmer (Dt.): Schritt 2

- Die Bearbeitungsschritte: Stemmen von Nomen, Adjektive und Verben (eingeschränkt)

1	2	3a	3b	3c	3d
e	*est*	*igend*	*ig*	*erlich*	*lichkeit*
em	*er*	*igung*	*ik*	*erheit*	*igkeit*
en	*en*	-------	*isch*	*enlich*	-------
ern	-------	*end*		*erheit*	*keit*
er	*st*	*ung*		-------	
es				*lich*	
-------				*heit*	
s					

- Beispiele:
  - *armes > arm*
  - *lieben > lieb*
  - *meins > mein*
  - *Henkels > henkel*

Porter Stemmer (Dt.): Schritt 2

- Die Bearbeitungsschritte: Stemmen von Nomen, Adjektive und Verben (eingeschränkt)

1	2	3a	3b	3c	3d
e	*est*	*igend*	*ig*	*erlich*	*lichkeit*
em	*er*	*igung*	*ik*	*erheit*	*igkeit*
en	*en*	-------	*isch*	*enlich*	-------
ern	-------	*end*		*erheit*	*keit*
er	*st*	*ung*		-------	
es				*lich*	
-------				*heit*	
s					

- Beispiele:
  - *bearbeitest > bearbeit*
  - *[$_1$einfacheren >] einfacher > einfach (↔ [$_1$schöneren >] schöner > schoner)*
  - *[$_1$derbsten >] derbst > derb*

Porter Stemmer (Dt.): Schritt 2

- Die Bearbeitungsschritte: Stemmen von Nomen, Adjektive und Verben (eingeschränkt)

1	2	3a	3b	3c	3d
e	*est*	*igend*	*ig*	*erlich*	*lichkeit*
em	*er*	*igung*	*ik*	*erheit*	*igkeit*
en	*en*	-------	*isch*	*enlich*	-------
ern	-------	*end*		*erheit*	*keit*
er	*st*	*ung*		-------	
es				*lich*	
-------				*heit*	
s					

- Beispiele:
  - *Vervollständigung > vervollstand, Einigung > einig*
  - *Ermittlung > ermittl (↔ Endung > endung)*
  - *zitierend > zitier (↔ sitzend > sitzend)*

Porter Stemmer (Dt.): Schritt 2

- Die Bearbeitungsschritte: Stemmen von Nomen, Adjektive und Verben (eingeschränkt)

1	2	3a	3b	3c	3d
e	*est*	*igend*	*ig*	*erlich*	*lichkeit*
em	*er*	*igung*	*ik*	*erheit*	*igkeit*
en	*en*	-------	*isch*	*enlich*	-------
ern	-------	*end*		*erheit*	*keit*
er	*st*	*ung*		-------	
es				*lich*	
-------				*heit*	
s					

- Beispiele:
  - *lebendig > lebend (↔ fleißig > fleissig)*
  - *Politik > polit*
  - *Portugiesisch > portugies*

Porter Stemmer (Dt.): Schritt 2

- Die Bearbeitungsschritte: Stemmen von Nomen, Adjektive und Verben (eingeschränkt)

1	2	3a	3b	3c	3d
e	*est*	*igend*	*ig*	*erlich*	*lichkeit*
em	*er*	*igung*	*ik*	*erheit*	*igkeit*
en	*en*	-------	*isch*	*enlich*	-------
ern	-------	*end*		*erheit*	*keit*
er	*st*	*ung*		-------	
es				*lich*	
-------				*heit*	
s					

- Beispiele:
  - *Besonderheit > besond*
  - *unehelich > unehe (↔ ehelich > ehelich)*

Porter Stemmer (Dt.): Schritt 2

- Die Bearbeitungsschritte: Stemmen von Nomen, Adjektive und Verben (eingeschränkt)

1	2	3a	3b	3c	3d
e	*est*	*igend*	*ig*	*erlich*	*lichkeit*
em	*er*	*igung*	*ik*	*erheit*	*igkeit*
en	*en*	*-------*	*isch*	*enlich*	*-------*
ern	*-------*	*end*		*erheit*	*keit*
er	*st*	*ung*		*-------*	
es				*lich*	
-------				*heit*	
s					

- Beispiele:
  - *Wahlmöglichkeit > wahlmoglich (↔ Möglichkeit > moglichkeit, obwohl Moglichkeit > moglich !!)*
  - *Geschwindigkeit >* $_{3d/1}$*geschwind*
  - *Sauberkeit > sauber*

Porter Stemmer (Dt.): Beispiele

- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*
- *leiten* *leit*
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit*
- *Geschwindigkeit* *geschwind*
  Geschwindigkeiten *geschwind*
- *geworfen* *geworf*
- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt*
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Beispiele

- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*

- *leiten* *leit*
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit*
- *Geschwindigkeit* *geschwind*
  Geschwindigkeiten *geschwind*
- *geworfen* *geworf*
- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt*
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Beispiele

- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*

  ✓
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*

- *leiten* *leit*
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit*
- *Geschwindigkeit* *geschwind*
  Geschwindigkeiten *geschwind*
- *geworfen* *geworf*
- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt*
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Beispiele

- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*

- *leiten* *leit*
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit*

- *Geschwindigkeit* *geschwind*
  Geschwindigkeiten *geschwind*

- *geworfen* *geworf*

- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt*
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Beispiele

- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*

- *leiten* *leit* ✓
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit* ✗

- *Geschwindigkeit* *geschwind*
  Geschwindigkeiten *geschwind*

- *geworfen* *geworf*

- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt*
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Beispiele

- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*

- *leiten* *leit*
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit*

- *Geschwindigkeit* *geschwind*
  Geschwindigkeiten *geschwind*

- *geworfen* *geworf*

- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt*
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Beispiele

- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*

- *leiten* *leit*
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit*

- *Geschwindigkeit* *geschwind* ✘
  Geschwindigkeiten *geschwind*

- *geworfen* *geworf*

- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt*
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Beispiele

- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*
- *leiten* *leit*
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit*
- *Geschwindigkeit* *geschwind*
  Geschwindigkeiten *geschwind*
- *geworfen* *geworf*
- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt*
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Beispiele

- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*
- *leiten* *leit*
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit*
- *Geschwindigkeit* *geschwind*
  Geschwindigkeiten *geschwind*
- *geworfen* *geworf* ?
- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt*
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Beispiele

- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*
- *leiten* *leit*
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit*
- *Geschwindigkeit* *geschwind*
  Geschwindigkeiten *geschwind*
- *geworfen* *geworf*
- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt*
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Beispiele

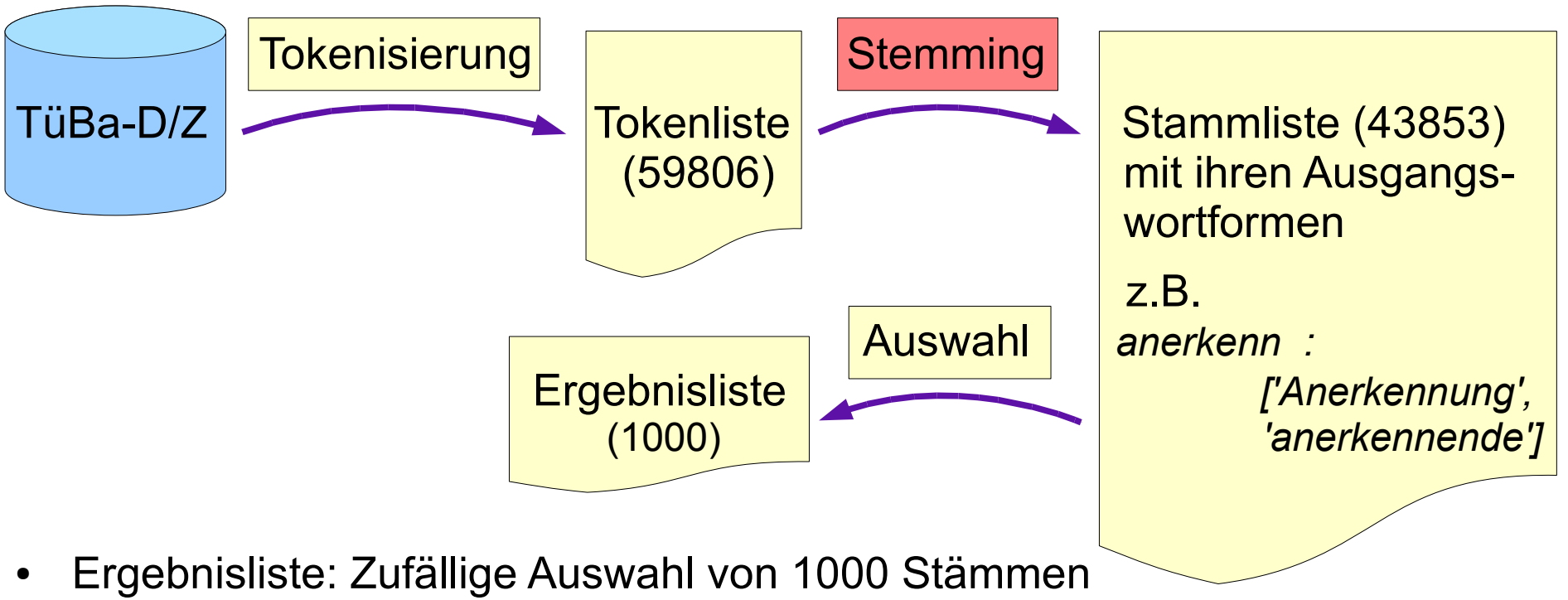
- *Sterns* *stern*
  Stern *stern*
  Sternen *stern*
  Sterne *stern*
- *beeindrucken* *beeindruck*
  beeindruckend *beeindruck*
  beeindruckender *beeindruck*
  beeindruckendsten *beeindruck*
- *Vollzeitstellen* *vollzeitstell*
- *Wasserversorger* *wasserversorg*
  Wasserversorgung *wasserversorg*
- *leiten* *leit*
  Leiter *leit*
  Leiters *leit*
- *Bundeswahlleiter* *bundeswahlleit*
- *Geschwindigkeit* *geschwind*
  Geschwindigkeiten *geschwind*
- *geworfen* *geworf*

- *geliebt* *geliebt*
  geliebtes *geliebt*
  Geliebten *geliebt* ✘
- *weinte* *weint*
  weint *weint*

Porter Stemmer (Dt.): Evaluierung

- Korpus: **TüBa-D/Z**: Tübinger Baumbank des Deutschen / Schriftsprache, 3. Version (14.07.2006) [auf den PCPool-Rechnern zugänglich]
  - syntaktisch manuell annotiertes Korpus der "die tageszeitung" (taz)
  - ca. 27000 Sätze, 470000 Wörter (Tokens) (27.09.2007)
- Nach der **Ausfilterung** von Zahlen und Stoppwörtern blieben
  - 396734 Wortformen
  - 59806 Tokens
- Nach dem **Stemming** blieben
  - 43853 Stämme = 73,32 % der ursprünglichen Tokenanzahl
  - Komprimierungsrate: $\frac{59806-43853}{43853}=0{,}3637$

Porter Stemmer (Dt.): Evaluierung



- Ergebnisliste: Zufällige Auswahl von 1000 Stämmen
- Annotierung der Ergebnisliste (manuell):
  - Korrektheit: korrekt – überstemmt – unterstemmt
  - Wortarten: Verb, Nomen, Adjektiv, andere Wortart, fremdsprachiges oder unbekanntes Wort; Eigenname

Porter Stemmer (Dt.): Evaluierung

- **Ergebnisse**:

		576	424
		Entfernung	**keine Entf**
534	**korrekt**	172	362
183	**Unterstemming**	122	61
283	**Überstemming**	282	1 (!)

- Korrektheit: 53,4 %
- viel mehr Überstemming als Unterstemming – gieriger Algorithmus
- Eigennamenanteil: 16,7 %
- Nomina insgesamt: 68,7 %
- Leider wurde ein Stamm falsch annotiert – ohne Entfernung von Endungen „zuviel Entfernung”.

Porter Stemmer (Dt.): Evaluierung

- **Fehleranalyse**:
  - Die Umlaute wurden nicht abgetrennt, *ß* wurde nicht durch *ss* ersetzt, *Ü, Ö* wurden nicht in Kleinschreibung umgesetzt.
    - Grund: zwei verschiedene Kodierungen: z.B. „ß" =
      - „\xe1" – Kommandozeile: ?
      - „\xdf" – Einlesen der Datei: nach der Ascii-Tabelle
  - von *-ie* wurde „e" abgetrennt, z.B. *Strategie > strategi*
  - Stammteil als Endungen erkannt: *servieren > servi, Fallobst > fallob, Trinkflasche > trinkflasch, Luftkampagne > luftkampagn*
  - Endung nicht als Ganze erkannt: *Schäfchen > Schäfch, Kindermädchen > kindermädch, AnwohnerInnen > anwohnerinn*
  - Überstemming, z.B.
    - *Geschwindigkeit > geschwind*
    - *Luftverschmutzung > luftverschmutz*

Porter Stemmer (Dt.): Evaluierung

- **Verbesserungsvorschläge**:
  - für -*chen* vor dem ersten Schritt einen zusätzlichen Schritt einfügen
  - Von -*ie* „e" nicht abtrennen *(Energie > energi)*
  - Behandlung von *-innen/-Innen > in (AnwohnerInnen > anwohnerinn)*
  - ? Entfernung der Endung *-in*
  - Behandlung von *-ieren > ier*, bzw. *-ier > -ier (servieren > servi)*
- **Frage**: bekommt man bessere Ergebnisse, wenn man die Komposita mit Bindestrich grundsätzlich auseinandernimmt? *(Lenau-Grundschule, Assistenz-Programm, Do-it-yourself-Verfahren)*
- *Hunde halten – Hundehalter, die Luft verschmutzen – Luftverschmutzung*
  - Dieses Phänomen lässt die Endung des zusammengesetzen Nomens nicht abtrennen.
- Nicht aufgeklärtes Problem:
  - *Müller > muller, aber Muller > mull*, *Möglichkeit > moglichkeit, obwohl Moglichkeit > moglich*

Übersicht

- I: Stemmingverfahren
  - Grundlagen
  - Eigenschaften
  - Stemming in Suchmaschinen
  - Evaluierung
  - Typische Fehler
  - Flaches und tiefes Stemming
- II: Stemmer
  - Stemmerarten
  - Porter-Stemmer für das Deutsche

- III: Entwicklung eines Stemmers
  - für das Ukrainische
  - (Porter-Stemmer für das Ungarische)
- IV: Zusammenfassung

Stemmer für das Ukrainische

Link:

STEMMER FÜR DAS UKRAINISCHE

Stemmer für das Ungarische

Link:

EIN PORTER-STEMMER FÜR DAS UNGARISCHE

Zusammenfassung

- Stemming ist ein schnelles, leicht implementierbares Verfahren;
- die am meisten verbreitete Methode – Affix Removal – ist stark sprachabhängig;
- ungelöste Probleme:
  - Eigennamen werden auch gestemmt (ca. 15 % aller Wörter!):

    Neubauer – neubau (Kurze Namen haben „Glück”, weil viele Stemmer die kurzen Wortformen nicht oder nicht immer stemmen. Beispiel: *Müller*)
  - unregelmäßige Formbildung:

    go – went

    matrix – matrices
  - Homonymie (Stamm- und Affixebene)
  - Komposita

Quellen

- **William Frakes (1992) Stemming Algorithms**. In: Frakes, William; Baeza-Yates, Ricardo (eds.): Information Retrieval. Data Structures and Algorithms. Prentice Hall: New Jersey, Kap. 8 (S.131-160).
- http://tartarus.org/~martin/index.html (Stand: 12.12.2007)
- http://snowball.tartarus.org/otherlangs/german_py.txt (Stand: 12.12.2007)
- http://snowball.tartarus.org/algorithms/german/stemmer.html (Stand: 26.12.2007)
- http://www.sfs.uni-tuebingen.de/de_nf_asc_resources.shtml (Stand: 18.01.2007)
- http://www.sfs.uni-tuebingen.de/de_tuebadz.shtml (Stand: .12.2007)
- http://www.comp.lancs.ac.uk/computing/research/stemming/Links/error.htm (Stand: 01.02.2008)
- http://scholar.google.de/scholar?hl=de&lr=&cluster=6157911103063237267 (Stand: 01.02.2008)
- http://www.google.com/support/bin/static.py?page=searchguides.html&ctx=basics (Stand: 26.01.2008)

Quellen

- http://www.dtsearch.com/CS_DeveloperTools.html#languages (Stand: 26.01.2008)
- http://www.stn-international.de/help/srchhelp.htm (Stand: 26.01.2008)