Lemmatisierer

Camilla Maffeis - Alba Sivianes
WiSe 21/22 - Computerlinguistik

Inhaltsverzeichnis

Lexeme und Lemmata

Wenn Fliegen hinter Fliegen fliegen, fliegen Fliegen Fliegen hinterher

→ Wie viele Wörter?

→ Was ist ein Wort?

Lexeme und Lemmata

NB: Doppelte Gliederung der Sprache - Ferdinand de Saussure

- **Langue**: (Sprach-)System.
  - Gesamtheit der Elemente einer Einzelsprache (*Phoneme, *Morpheme) und der Regeln ihrer Verknüpfung (vgl. *Morphologie, *Wortbildung, *Syntax, *Textlinguistik), das dem konkreten Sprachgebrauch, der *parole, zugrunde liegt.
- **Parole**: Sprachgebrauch.
  - Im Gegensatz zur *langue, dem abstrakten Sprachsystem, konkrete Anwendung der Sprachkenntnis durch Produktion von mündlichen oder schriftlichen Äußerungen.

Lexeme und Lemmata

Ausdrücke sind:

- auf der langue-Ebene: Bestandteile des lexikalischen Inventars einer Einzelsprache → abstrakte Einheite des Lexikons
  = **Lexeme / Lexikalische Wörter**
- auf der parole-Ebene: in konkreter Verwendung vorliegende Einheite einer Äußerung bzw. eines Textes.
  = **Wörter** (auch: **Textwörter**)

Wenn Fliegen hinter Fliegen fliegen, fliegen Fliegen Fliegen hinterher

→ 9 Textwörter, 5 Lexeme

Lexeme und Lemmata

- In konkreter Verwendung erscheinen veränderbare Lexeme immer in einer bestimmten Wortform
  (wenn auch häufig ohne grammatische Morpheme/ nur mit einem 'Null-Allomorph' verbunden)
- Im Wörterbuch erscheinen Lexeme in einer bestimmten **Zitierform/Grundform/Nennform**
  - z.B. im Deutschen:
    - Verben: im Infinitiv Präsens Aktiv (d.h. mit einem GM) → *träumen*
    - Substantive: im Nominativ Singular → *Traum*
    - Adjektive: im unflektierten Form → *traumhaft*

⇨ **Lemma** = Grundform eines Lexem; der Eintrag in einem Wörterbuch

Lemmatisierung

Lemmatisierung: Festlegung der Grundform eines Wortes → lexikalische Wörter werden auf ihre Grundforme zurückgeführt und zugeordnet

- in der maschinellen Sprachverarbeitung: die (automatische) Rückführung einer Wortform auf eine kanonische Grundform zusammen mit einer Annotierung von bestimmter linguistisch relevanter Information über die entsprechende Wortform

Lemmatisierer: morphologische Analyseprogramme, die die grammatische Form eines Wortes (→Oberflächenform) auf das Lemma zurückführen.

z.B. trefft, getroffen, traf > treffen

Lemmatisierung - Wofür braucht man das

Verarbeitung natürlicher Sprache

- Suchmaschinen
- Rechtschreibkorrektur
- Elektronische Wörterbücher
- Maschinelle Übersetzer
- Spam filtering
- Informationsextraktion
- Analyse von Big Data
- Korpuslinguistik → Analyse der Frequenz

Lemmatisierung - Anfänge

Padre Roberto Busa (1913 - 2011)

Pioner der Computerlinguistik und der DH

Index Thomisticus: Lemmatisierung der Werke von Thomas von Aquin

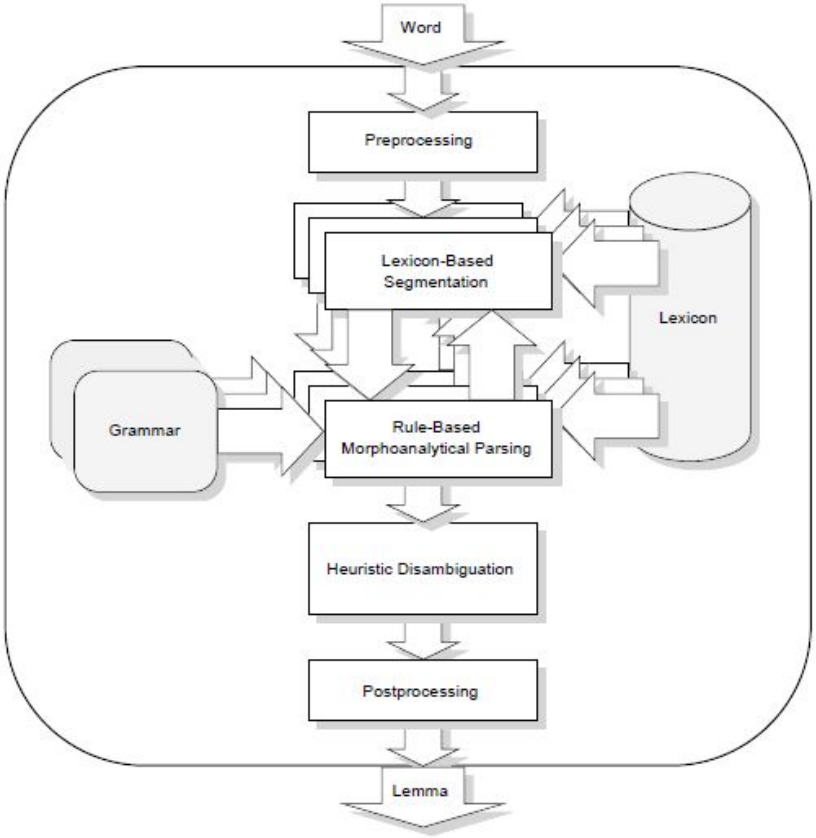
- 1946 - 1970s
- 1949: finanziert und unterstützt von IBM
- 56-bändiges Werk mit 70.000 Seiten; ca 11 Millionen Wörter

https://www.corpusthomisticum.org/it/index.age

Wie funktioniert es

2 Typen von Lemmatisierung → kann abhängig bzw. unabhängig vom Text geführt werden:

- **morphologische Lemmatisierung**: die Wortformen werden isoliert analysiert → unabhängig vom syntaktischen Kontext
  - wichtig für DH: kann leichter automatisiert werden
- **morphosyntaktische Lemmatisierung**: die Wortformen werden innerhalb des syntaktischen Kontext analysiert
  - immer eindeutig: Der Zusammenhang der Wortform zum Satz gibt ihren Wert an.



Aus: https://www.ids-mannheim.de/fileadmin/kl/dokumente/glemmrep.pdf

Preprocessing

Durchführung von Aufgaben, die nicht direkt mit dem Lemmatisierungsproblem zusammenhängen.

Der Präprozessor ist als Filter implementiert, um verschiedene Eingabeformate zu berücksichtigen.

Lexikon

Den Textwörtern müssen Informationen zugeordnet werden → LEXIKON

LEXIKON: konstruiert durch die Extrahierung von Informationen aus Wörterbücher und durch die Insertion von neuen lexikalischen Einheiten

- Enthält morphologische und morpho-syntaktische Beschreibungen der Lexeme
- Für jedes Lexem ein odere mehrere Einträge mit alternativen morpho-syntaktischen Paradigmen
    - z.B: *er* →als Pronom, als Verbpräfix, als Nomensuffix...

Lexikon-basierte Segmentierung

Lexikon-basierte bottom-up-algorithmus zur Morphemerkennung → Segmentierung der Input-Wörter

Der Algorithmus erzeugt für jede eingegebene Verbform einen erschöpfenden Segmentierungsbaum.

Die Segmentierungen zeigen die verschiedenen Möglichkeiten der Zerlegung der Eingabezeichenfolge.

Parsing

Segmentierungsbaum → Morpho-analytischer Parser → fehlerhafte Segmentierungslesungen werden verworfen

Parser

- arbeitet nach einem Regelwerk für Flexion, Derivation und Wortkompositionsmorphologie des Deutschen
- additional constraints: to validate the gender, number and case dependencies, the consistency of affixes and linking morphemes, etc
- The parser also assigns weights to every accepted segmentation reading based on the parse tree used and on the affinity information provided by the segmentation algorithm
- Der Parser weist auch jedem akzeptierten Segmentierungs-Lesen ein Gewicht zu, basierend auf dem verwendeten Parse-Baum und den vom Segmentierungsalgorithmus bereitgestellten Affinitätsinformationen

Context Switching

- Context switching driver: steuert die Segmentierung- und Parsingsprozesse
- Der Controller unterbricht den Segmentierungsalgorithmus, wenn dem Segmentierungsbaum ein neuer Zweig hinzugefügt wird, und es wird ein Prozess zwischen dem Parser und dem Controller eingeleitet, in dem beide die neuen Segmentierungen überprüfen.
- Erfolgreich geparste Segmentierungsbäume werden mit Gewichten versehen und zur weiteren Disambiguierung auf einen Stapel gelegt.
- Schlecht geformte Reads werden verworfen.

Disambiguation

→ eine der schwierigsten Aspekten bei der Entwicklung eines deutschen Lemmatisierers

→ Bestimmung der korrekten *readings* unter denen, die das Parsingprozess “bestanden” haben; der Entscheidungsprozess zur Ermittlung der letzten Bestandteile eines zusammengesetzten Wortes im Deutschen.

- zB Gehalt → geh- (gehen) + Alt
  - Fehler, aber folgt einem im Deutschen sehr produktiven Wortbildungsmuster: <Verb-Präsens-stem>+<Subst>, zB Gehweg

Dabei werden die akzeptierten Lesarten auf dem Baumstapel in der Reihenfolge abnehmender Wahrscheinlichkeiten geordnet, um die Lemma-Namen auszuwählen.

Verwirft der Disambiguator alle Lesungen, die eine deutlich schlechtere Bewertung als die beste haben.

Postprocessing

Nachdem die Disambiguierung abgeschlossen ist, werden die vom Parser gesammelten internen morphoanalytischen Informationen über die eingegebene Wortform auf die folgende Oberflächenstruktur reduziert:

<word_form>

<composition_tag><derivation_tag><lemma_name>

Der Rest der internen morpho-analytischen Informationen, die der Parser während der Analyse über die eingegebene Wortform gesammelt hat, wird verworfen.

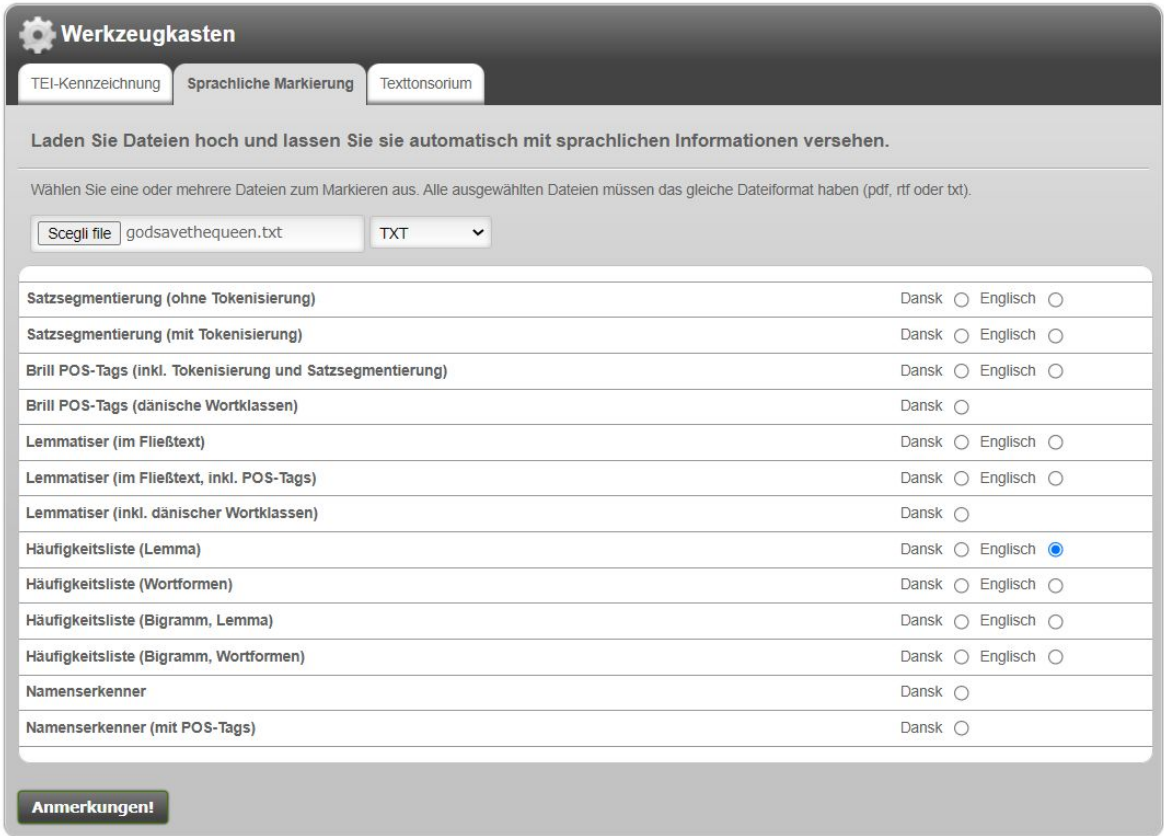
Beispiel

https://clarin.dk/clarindk/toolchains-wizard.jsp

Beispiel:

God save the queen - Lemmatisiert inkl. POS-Tags

God Save The Queen - Häufigkeitsliste



Quellen

- Busch, Albert, and Oliver Stenschke. *Germanistische Linguistik: eine Einführung*. Narr Francke Attempto Verlag, 2018.
- Stadler, Heike. "Die Erstellung der Basislemmaliste der neuhochdeutschen Standardsprache aus mehrfach linguistisch annotierten Korpora." (2014).
- Glück, Helmut, Rödel, Michael (Hrsg.): Metzler Lexikon Sprache. 5.Aufl. Stuttgart 2016.
- https://www.digitale-edition.at/o:konde.115
- Pirelli, Alessandro. *Classificazione di documenti pre-elaborati con tecniche di stemming*. Diss.
- Explora! Science Now, 2. Folge
- Sánchez, Evelín Perdomo, et al. "ANÁLISIS DE LOS PROCESOS DE LEMATIZACIÓN Y ESTEMIZADO EN LINGÜÍSTICA COMPUTACIONAL." (2017).
- Cyril, Belica: **WP2-Lemmatizer**, Final Report. MLAP93-21 MECOLB. Luxembourg 1994.
- SALEM : e. Verfahren zur automat. Lemmatisierung dt. Texte / Hrsg.: Sonderforschungsbereich 100 „Elektron. Sprachforschung", Projektbereich A. Aus Beitr. von Hans Eggers... Zsgest. von Hans Eggers ... Tübingen : Niemeyer, 1980.
- https://www.clarin.eu/blog/clarin-dk-presents-cst-lemmatizer