Wieviel Computerlinguistik braucht der Word-Anwender?

Tobias Thelen

Institut für Semantische Informationsverarbeitung
Universität Osnabrück

E-Mail: tthelen@uos.de
www.schrifterwerb.de

Übersicht

- Textverarbeitungssysteme - Was kann Word, was kann der Anwender?
- Computerlinguistik in Word: Rechtschreib-, Grammatik und Stilprüfung
- Computerlinguistik & Orthographie
- Diskussion: Wieviel Computerlinguistik

Typologie der Funktionen eines Textverarbeitungssystems

Dokument-management	Texterfassung & -überarbeitung	Layout / Gliederung	Sprache / Inhalt
Laden/Speichern	Editorfunktionen	Zeichenformate	Rechtschreibung
Versionskontrolle	Textbausteine	Absatzformate	Grammatik
Drucken	Eingabekorrekturen	Seitenformate	Stilprüfung
Schlagworte	Objekte einfügen	Log. Gliederung.	Abstracting
	Notizen	Direkformatierg..	
		Ansischtsmodi	

Anwender„typen“ I

- „ Schreibmaschinenbenutzer“
  - Verwendet nur Direktformatierungen
  - Rechtschreibprüfung während der Eingabe
  - Verändert keine Voreinstellungen
  - Erstellt Inhaltsverzeichnis per copy/paste
  - Benutzt keine Hilfefunktionen

Anwender„typen“ II

- „ Normalbenutzer“
  - Verwendet viele Direktformatierungen
  - Gliederung per fertiger Formatvorlagen
  - Rechtschreibprüfung mit Vorschlägen
  - Beherrscht einige Standardverfahren:
    - Seitenummerierung
    - Verzeichnisse
    - Fußnoten

Anwender„typen“ III

- „ Bürobenutzer“
  - Erstellt viele gleichartige Dokumente
  - Benutzt betriebsspezifisch eingerichtete Makros/Vorlagen
  - Hauptsächlich standardisierte Verfahren
  - Inhalt besteht zu großen Teilen aus Textbausteinen
  - Rechtschreib- und evtl. Stilprüfung

Anwender„typen“ IV

- „ DTP-Benutzer“
  - Erstellt große Dokumente, evtl. in verschiedenen Fassungen/Überarbeitungen
  - Richtet Makros/Vorlagen selbst ein
  - Nutzt viele Automatisierungsmöglichkeiten
  - Interesse an umfassenden Dokumentmanagementfunktionen
  - Rechtschreibprüfung

Vorhandene CL in Word: Rechtschreibung

- Rechtschreibprüfung:
  - Einzelwort- und lexikonbasiert
  - schwache Morphologiekomponente zur Kompositazerlegung
  - Soundex-ähnlicher Mechanismus
  - Korrekturvorschläge mit einigen Heuristiken
  - „ seltsame“ Behandlung alter/neuer Rechtschreibung

Vorhandene CL in Word: Stilprüfung

- Stilprüfung:
  - Satzlänge, Konstituentenhäufung, Einbettung, Passivverwendung
  - Stilebene (gehoben, veraltet, vulgär)
  - „ Verständlichkeitsmaß“ auf Grundlage dieser Punkte
  - Gemischte Verfahren: Lexikon, einfache Zählungen und Parsing

Vorhandene CL in Word: Grammatikprüfung

- Grammatikprüfung:
  - Kongruenzchecks, Vollständigkeitsprüfung
    - Verfahren: Parsing
    - Problem: Funktioniert nur für „Standardsätze“
    - Übersieht viele Fehler und schlägt oft „Fehlalarm“
  - Unterstützung der Rechtschreibprüfung: Groß- und Kleinschreibung
    - Lexikonbasiert + Erkennung von „Substantivierung“
  - Oberflächliche Überprüfung von Interpunktion

Vorhandene CL in Word: Zusammenfassung

- Es werden eine Reihe von CL-Standardverfahren verwendet
- Das Hauptgewicht liegt auf „Effizienz“
- Es werden kaum konkrete Korrekturvorschläge gemacht
- Zielrichtung: Offensichtliche Fehler im Hintergrund erkennen und markieren.

Orthographie und Linguistik

Phonographie

- Basis-GPK
- Schärfung
- Dehnung
- Silbentrennendes h
- /s/-Schreibung
- /ks/-Schreibung

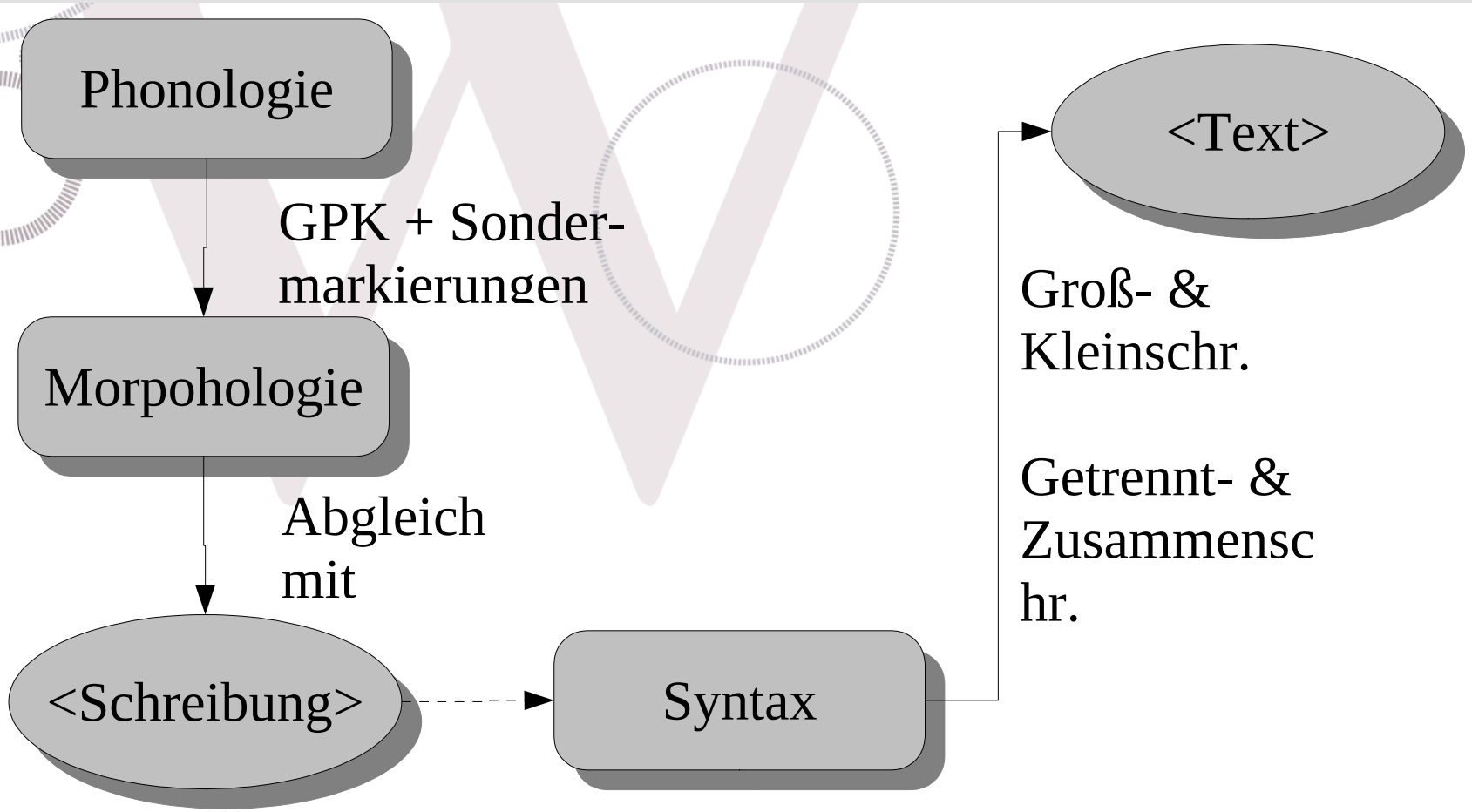
Logographie

- Groß- & Kleinschreibung
- Getrennt- & Zusammenschr.

Interpunktion

- Satzgrenzen-markierung
- Kommasetzung

Herleitung der Schreibung



Zusammenfassung Linguistik

- Die deutsche Rechtschreibung ist zu einem sehr großen Teil regelhaft aus anderen linguistischen Größen herleitbar.
- Diese Größen sind aber in einem Textverarbeitungssystem i.d.R. nicht bekannt.
- Retrieval aus einem Lexikon Bootstrapping-Problem

Computerlinguistische Verbesserungen

- Mehrstufige Verfahren je nach vorhandenem Wissen
- Linguistisch fundierte Soundex-Verbesserungen
- Stringabstandsmaße mit Heuristiken
- NP-Parsing für Groß-/Kleinschreibung

☐ Wie wirkt sich das auf die Effizienz aus?

Diskussion

- Metapher: „Weißes Blatt und Schreibmaschine“
- Metapher: „Autor, Lektor, Setzer“
- Metapher: „Diktat und Sekretärin“
- Was brauchen/wollen Anwender?
- Kann Computerlinguistik das Erforderliche leisten?