LEMMATISIERUNG UND STEMMING IN SUCHMASCHINEN

MASTERSEMINAR SUCHMASCHINEN

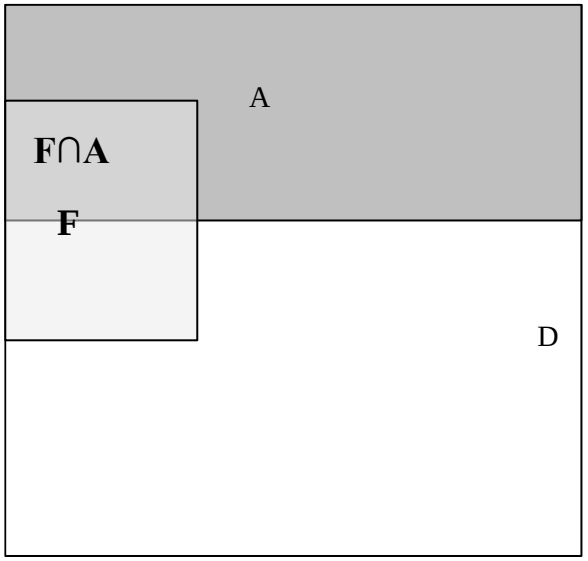
COMPUTERLINGUISTIK

SOMMERSEMESTER 2021

STEFAN LANGER

STEFAN.LANGER@CIS.UNI-MUENCHEN.DE

Trefferquote (Recall) und Genauigkeit (Precision)



Grundbegriffe Morphologie

Lexikalisches/ grammatisches Morphem

Lexikalische Morpheme haben eine semantische Funktion (*Schirm*, *sprech*-); grammatische Morpheme eine rein grammatische Funktion (z.B. Genitivsuffix).

Freies/gebundenes Morphem

Erstere kommen frei vor (*kurz*, *Schirm*), letzter nicht (-*er*, *Him*-).

Stamm (a,b), ~Basismorphem, ~Grundmorphem, ~Wurzel(b)

Morpheme oder Morphemkombinationen zu denen Flexionsendungen treten können (z.B. *Schirm* mit den Wortformen *Schirms*, *Schirme* etc.);

lexikalisches Morphem (nicht Morphemkombination), das einer Wortform oder einer Reihe von Wortformen zugrundeliegt (z.B. *Unsinnigkeit - Sinn*).

Grundbegriffe Morphologie

Affix - Suffix - Präfix - Infix - diskontinuierliche Morpheme; Flexionsaffixe, Derivationsaffixe

Affixe sind Morpheme, die zu einem Stamm treten können um entweder eine Wortform (Flexionsaffix, Z.B.: Pluralsuffix wie in *Kind* - *Kinder*) oder ein neues Wort (Derivationsaffix z.B. das denominale Suffix -*bar* in *wunder* - *wunderbar*) zu erzeugen.

Außerdem könne Affix unterschieden werden nach den Wortarten, zu denen sie treten/die sie erzeugen (nominale, verbale etc. Affixe); oder dannach, ob sie:

an den Wortanfang angehängt werden (Präfix); Bsp.: *un*- in *unmöglich*;

an das Wortende angehängt werden (Suffix), Bsp. -*bar* in *wunderbar*;

innerhalb des Wortes eingeschoben werden (Infix), Bsp. -*zu*- in *einzuschenken*;

aus mehreren Teilen bestehen (diskont. Morphem) - -*ge*-*t* in *eingeschenkt*.

Grundbegriffe Morphologie

Derivation, Konversion, Komposition, Wortbildung

Derivation (a. Ableitung) ist die Bildung neuer Wörter aus bestehenden Wörtern oder Wortstämmen. Meist geschieht dies durch Hinzufügung eines Affix. Beispiele: *Glück - Unglück*, *Glück - glücklich*.

Konversion ist liegt vor bei Übertragung in eine andere Wortart ohne explizite Veränderung (*deutsch-Deutsch*).

Komposition ist die Bildung neuer Wörter aus zwei existierenden Lexemen, z.B. *Wortbildung* aus *Wort* und *Bildung*.

Sprachtypen

Isolierende, agglutinierende, flektierende Sprachen

Unterteilung von Sprachen aufgrund der Flexionsmorphologie.

Isolierende Sprache drücken (fast) alle grammatischen Beziehungen im Satz und syntaktische Merkmale durch separate Wörter (freie Morpheme) aus (z.B. Präpositionen). Als typisches Beispiel gilt Chinesisch; aber auch Englisch hat einige Merkmale isolierender Sprachen.

Agglutinierende Sprachen realisieren gramm. Merkmale als jeweils separate Suffixe - Wörter sind leicht segmentierbar. Beispiele sind Finnisch oder Türkisch.

Flektierende Sprachen realisieren gramm. Merkmale als Affixe, allerdings oft mehrere Merkmale in einem Affix; es treten auch Stammveränderungen auf. Deutsch und Latein können als flektierende Sprachen klassifiziert werden.

Übung

Für eine oder mehrere Sprachen, die Sie kennen, begründen Sie:

- warum morphologische Analyse im Kontext von Suchmaschinen wichtig/unwichtig ist.
  - Für welche Wortarten ist morphologische Analyse wichtig?

- Welches sind in der von Ihnen genannten Sprache die besonderen Schwierigkeiten

Die Wortarten des Deutschen (CISLEX-Klassifikation)

Flektierende Wortarten

Nomen

Adjektiv

Verb

Determinator

Pronomen

Nichtflektierende Wortarten

Adverb

Partikel

Verbpartikel

Präposition

Konjunktion

Interjektion

Nominaldeklination

	stark	**schwach**	**gemischt**
sing. nom	*Tag, Kind, Nacht*	*Mensch, Hase*	*Staat*
gen	*Tages, Kindes, Nacht*	*Menschen; Hasen*	*Staats*
dat	*Tag(e); Kind(e), Nacht*	*Menschen, Hasen*	*Staat(e)*
akk	*Tag, Kind, Nacht*	*Menschen, Hasen*	*Staat*
plur nom	*Tage, Kinder, Nächte*	*Menschen, Hasen*	*Staaten*
gen	*Tage, Kinder, Nächte*	*Menschen, Hasen*	*Staaten*
dat	*Tagen, Kindern, Nächten*	*Menschen; Hasen*	*Staaten*
akk	*Tage, Kinder, Nächte*	*Menschen, Hasen*	*Staaten*

Grundformenreduzierung

Stemming	Wörterbuchbasiert	Wörterbuch + Regeln
*Dokument**en***	*Dokumenten:Dokument*	*Dokumenten:Dokument+en*
*Suchmaschin**en***	*Suchmaschinen:* *Suchmaschine*	*Suchmaschinen:* *Suchmaschine+n*
*Rahm**en***	*Rahmen:Rahmen*	*Rahmen:Rahmen+*
*Computer**s***	*Computers:Computer*	*Computers:Computer+s*
*Merkel**s***	*Merkels:?*	*Merkels:Merkel+s*

Lemmatisierung / Stemming: Implementierung

Vollformenlexikon

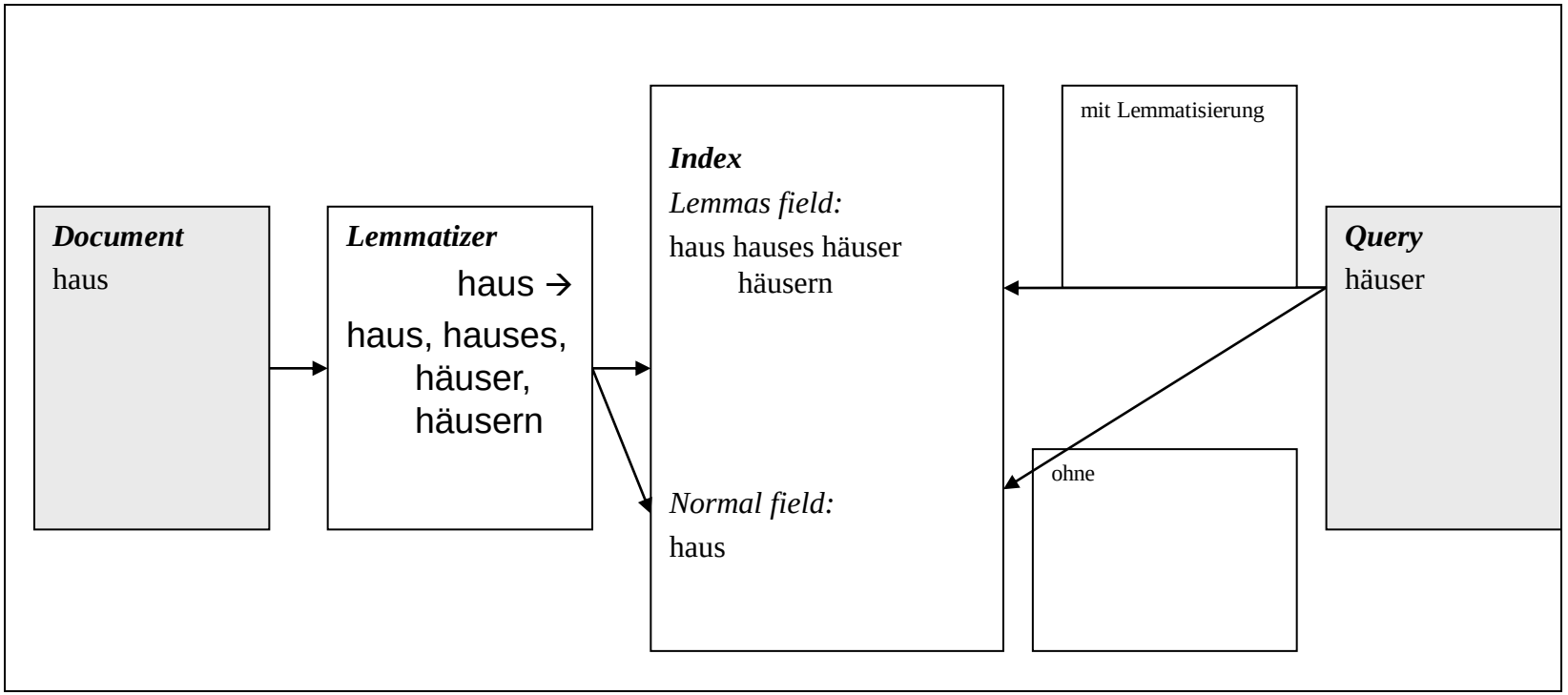
- Implementiert als Hash/Trie

Regelbasierte Morphologie

- Rechtsrekursive Grammatik / reguläre Ausdrücke
- Two-Level Morphologie

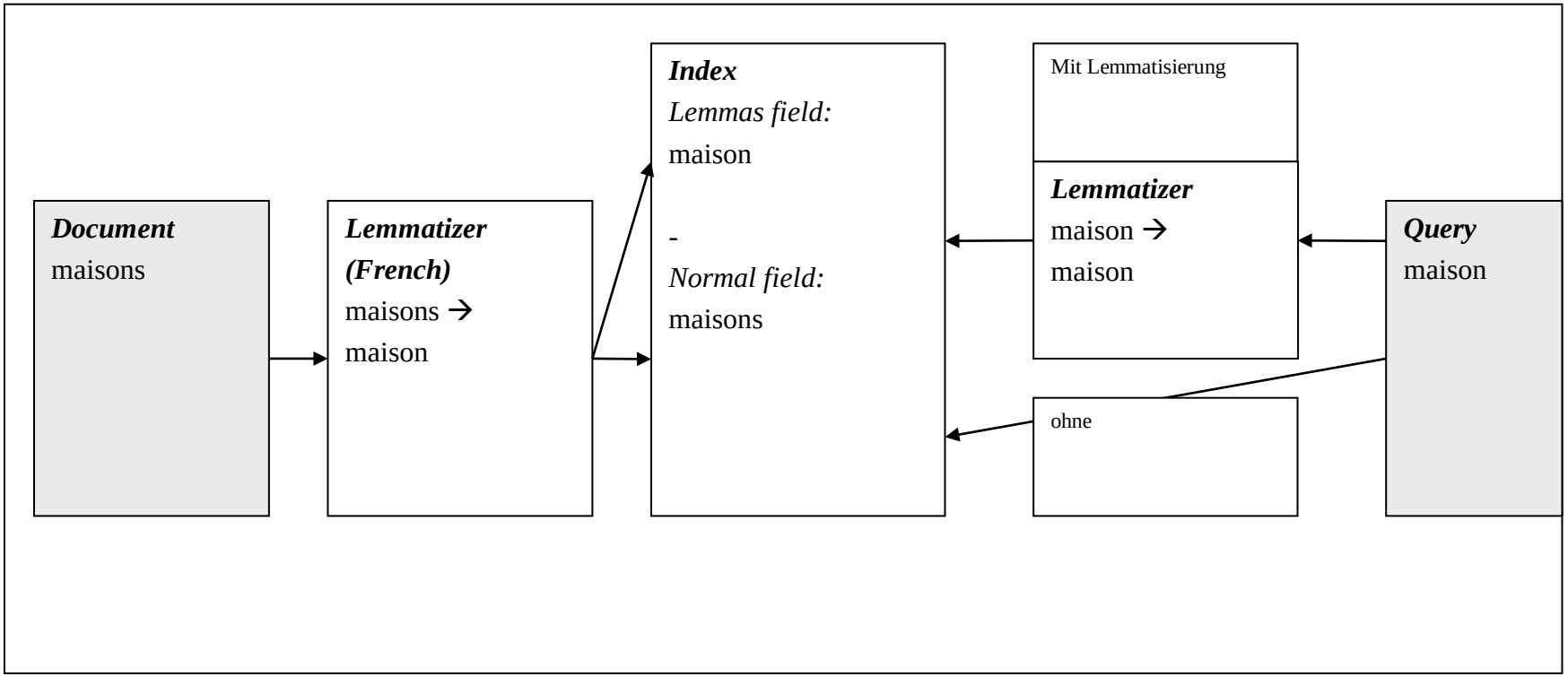
  Regelbasierte Morphologie in der Regel durch Reduktion

Lemmatisierung durch Expansion von Dokumententermen



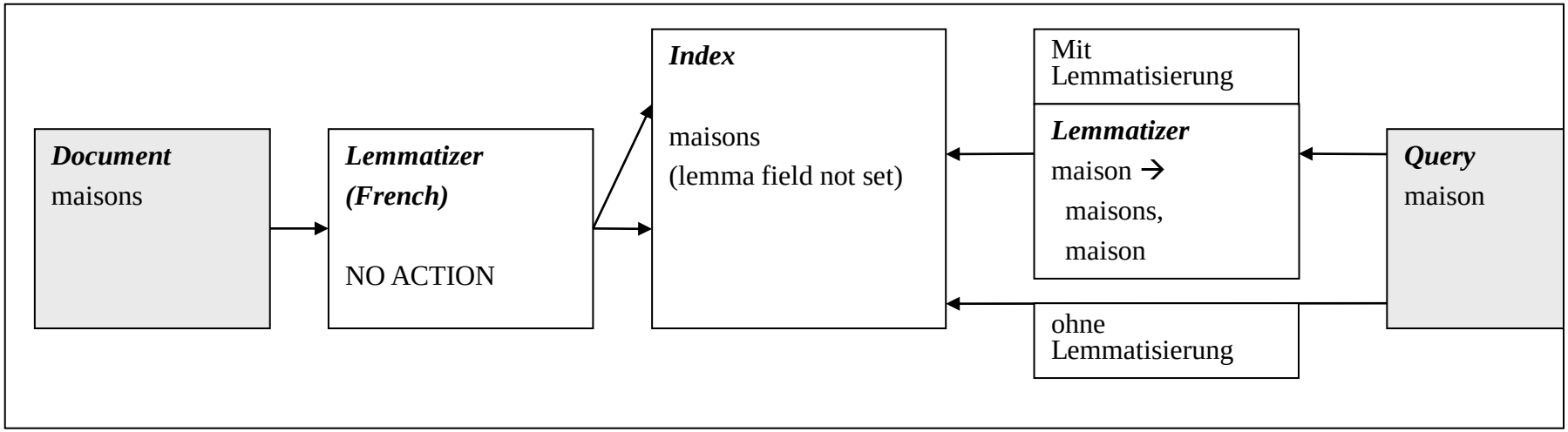
Alle Wortformen der Wörter im Dokument werden in den Index geschrieben. Die Sprache der Anfrage muss nicht bekannt sein

Lemmatisierung durch Reduktion



Wörter in Anfrage und Dokument werden auf die Grundform(en) reduziert. Dazu muss die Sprache der Anfrage bekannt sein

Lemmatisierung durch Anfrageexpansion



Elasticsearch (stemmers)

Arabic **arabic**
Armenian **armenian**
Basque **basque**
Bengali **bengali** **light_bengali**
Brazilian Portuguese **brazilian**
Bulgarian **bulgarian**
Catalan **catalan**
Czech **czech**
Danish **danish**
Dutch **dutch**, dutch_kp
English **english**, light_english, minimal_english, possessive_english, porter2, lovins
Finnish **finnish**, light_finnish
French french, **light_french**, minimal_french
Galician **galician**, minimal_galician
German german, german2, **light_german**, minimal_german
Greek **greek**
Hindi **hindi**
Hungarian **hungarian**, light_hungarian
Indonesian **indonesian**
Irish **irish**
Italian italian, **light_italian**
Kurdish (Sorani) **sorani**
Latvian **latvian**
Lithuanian **lithuanian**
Norwegian (nb)
norwegian, **light_norwegian**, minimal_norwegian
Norwegian (nn) **light_nynorsk**, minimal_nynorsk
Portuguese
portuguese, **light_portuguese**, minimal_portuguese, portuguese_rslp
Romanian **romanian**
Russian **russian**, light_russian
Spanish spanish, **light_spanish**
Swedish **swedish**, light_swedish
Turkish **turkish**

Nominalkomposita: Beispiele

Blumen|versand

Internet|such|maschine

Fuchs|schwanz

Bahn|hof

Frei|zeit

Hoch|zeit

Po|made

Tisch|fuß|ball

Wach|s|tube

Mädchen|handels|schule

Nominalkompositaanalyse

Daten:

- Nominallexikon
  - Vollformen
  - Kompositionsformen (Form mit Fugenelement)

    z.B. *Umgebungs*

  Regeln:

  Kompositum -> Kompositionsform + Vollform

  Kompositum -> Kompositionsform + Kompositum