Wie aus Korpora ermittelte Grundformhäufigkeiten und Kookkurrenzen bei der Beschreibung des Sprachgebrauchs helfen können

Rainer Perkuhn

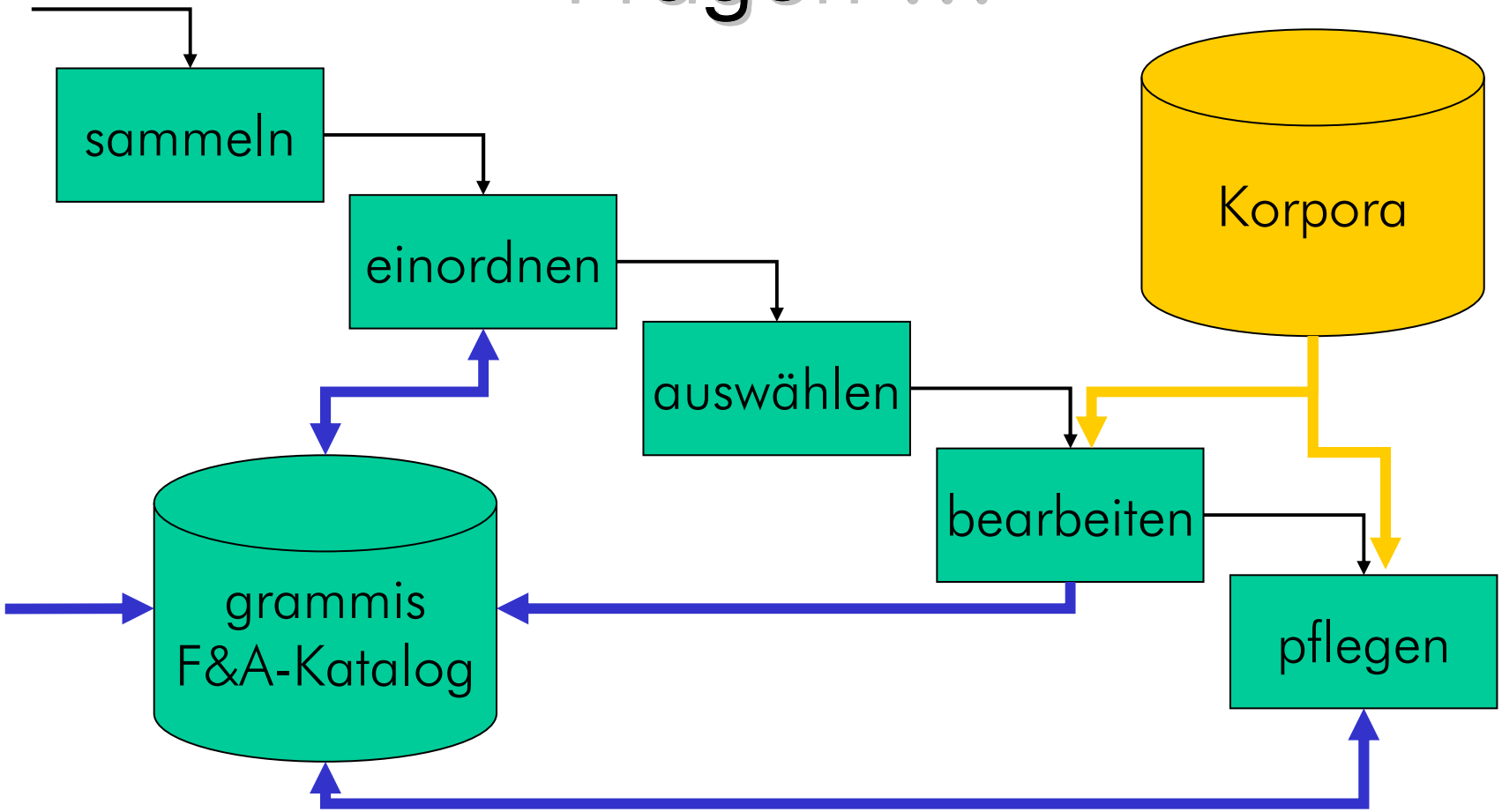
Neues aus der korpuslinguistischen Forschung
am Institut für Deutsche Sprache

Tokyo, 19.3.2008

Beispiel: grammis
Grammatik in Fragen und Antworten

- Projektgruppe:
  Bruno Strecker, Elke Donalies,
  Marek Konopka, Jacqueline Kubczak
- Erstellung von Artikeln zu grammatisch relevanten Fragestellungen
- Mischung konkret/grundsätzlich
- seit ca. einem Jahr online

Fragen …


sammeln
einordnen
auswählen
bearbeiten
pflegen
Korpora
grammis
F&A-Katalog

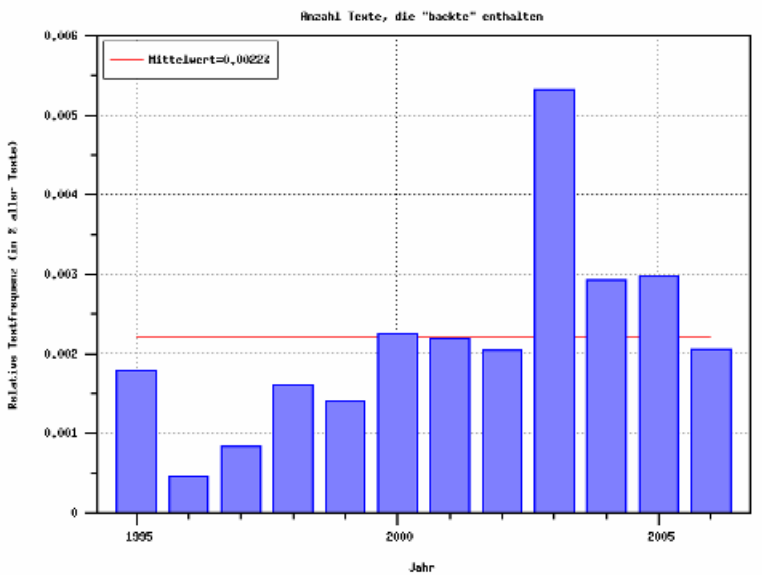
Fragen bearbeiten

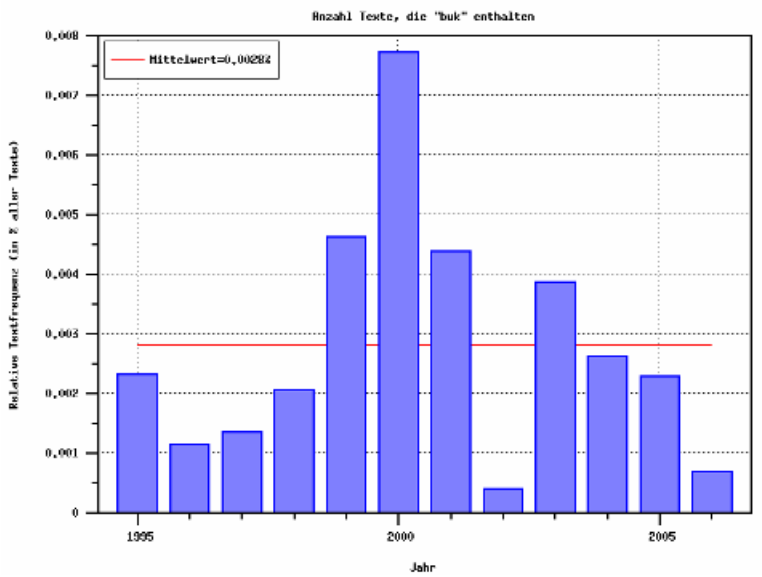
- Grammatiken heranziehen
  - ✍ Sachverhalt und geschichtlichen Hintergrund beschreiben
- Daten befragen
  - DEREKO, speziell neue Rechtschreibung
  - World Wide Web über google, speziell Chatrooms
- quantitativ bewerten, ggf. Tendenzen
  - ✍ Trend beschreiben, Empfehlung formulieren

Beispiel

- heißt es „backte“ oder „buk“?
- Sprachgebrauch weicht von Grammatiken ab
- nicht „entweder-oder“, sondern „sowohl-als-auch“
- zeitliche Entwicklung
- ggf. andere Dimensionen interessant, bisher noch nicht systematisch methodisch unterstützt

Zeitliche Entwicklung





backte vs. buk

Probleme

- Anfragen schwierig exakt zu formulieren
  - Wortformen
  - Paradigmen einer Grundform (bukst, bukest?)
  - Wortkombinationen (Stellung und Variabilität)
- Dimensionen erkennen, in Einflussfaktoren zerlegen
  - gesprochene vs. geschriebene Sprache
  - Zeit, Region, Dialekt
- im World Wide Web so gut wie unmöglich
- in geeigneten Korpora grundsätzlich machbar
  - Daten fortschreiben, Methodik verfeinern

WWW vs. DEREKO

- schelle Suche in großen Datenmengen
  - mit Hilfe von Google um Informationen (Webseiten) zu finden
  - mit Hilfe von COSMAS um sprachlichen Phänomenen nachzuspüren
- Zusammensetzung der Daten (nicht) kontrollierbar
- Bedeutung von Sonderzeichen?
- Behandlung von Funktionswörtern?
- Behandlung von Wortbildung?
- Verknüpfungen (logisch, Abstand)?

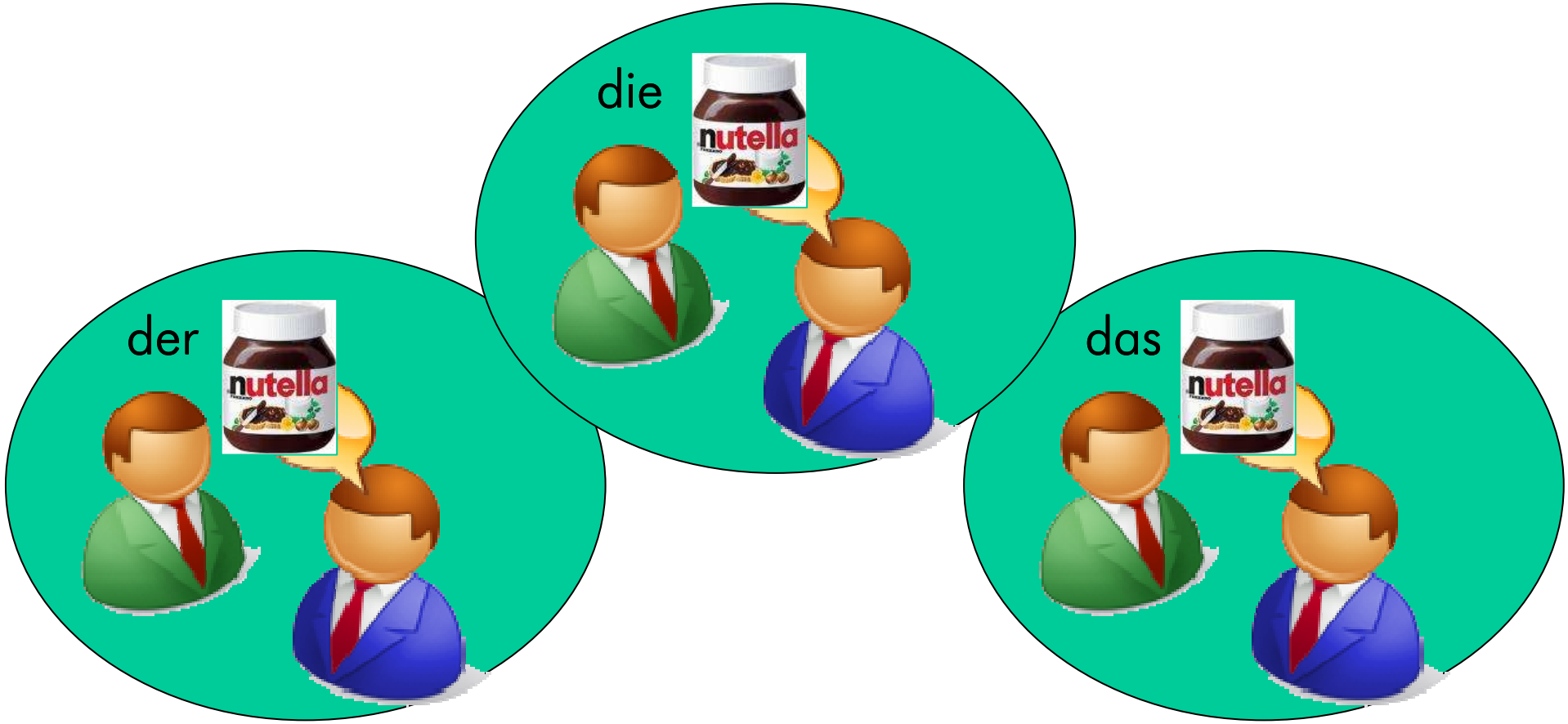
Gebrauch vs. Norm

- d.. Nutella?
- wegen dem Ändern der Rechtschreibung braucht der einzigste Papst noch lange nach Aldi gehen, um der Nutella kaufen
- wer „brauchen“ ohne „zu“ gebraucht, braucht „brauchen“ gar nicht zu gebrauchen
- der Dativ ist dem Genitiv sein Tod

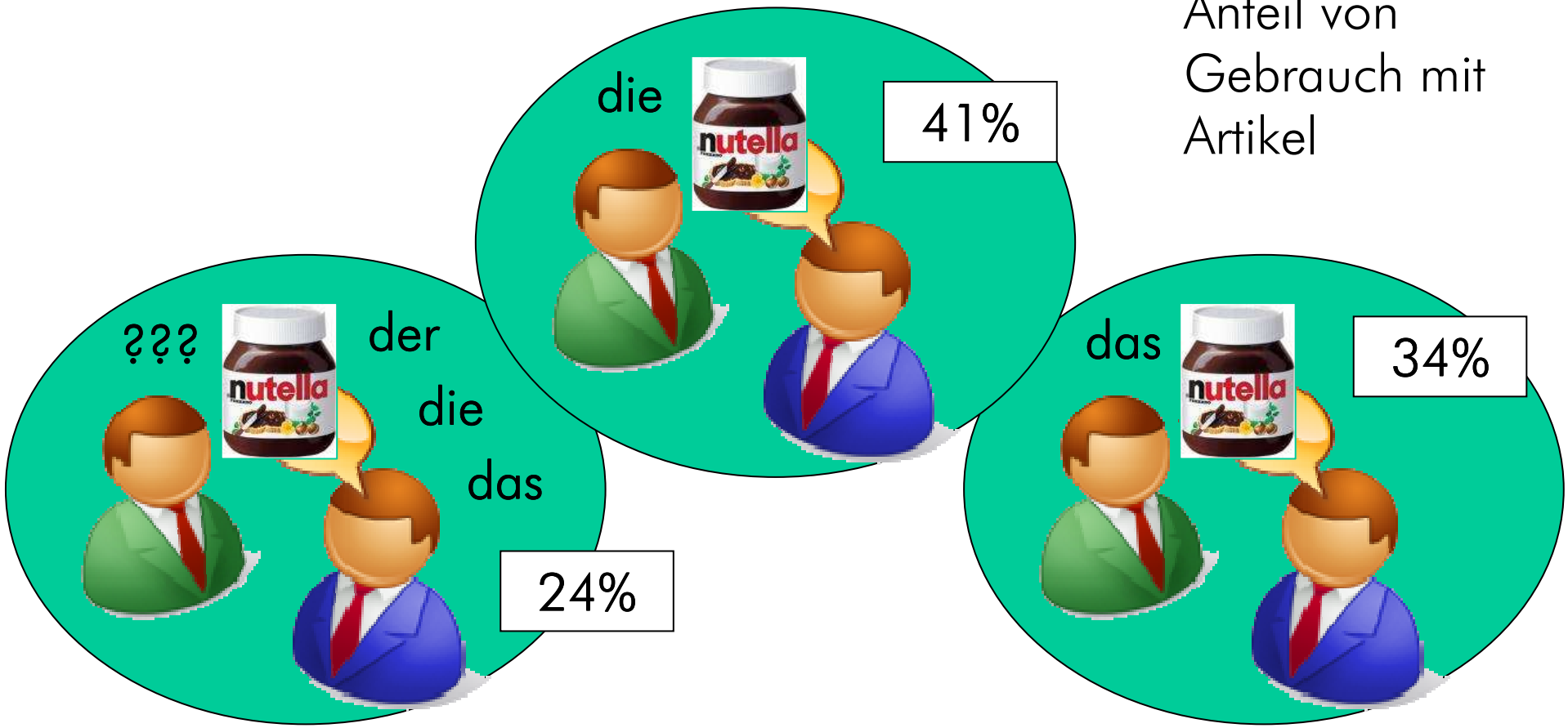
„Gebrauch“ von „ge-/brauchen“ oder „Brauch“?

- Sprachgebrauch setzt sich über Normen hinweg und hat eigene Gesetzmäßigkeiten (brauchen wir ein Wort für „nicht durstig“?)
- das, worüber wir nur im Einzelfall stolpern (weil unser Gefühl sich einmischt), steht stellvertretend für all die Schwierigkeiten, die beim Erlernen (oder Begreifen) der Sprache auftreten …

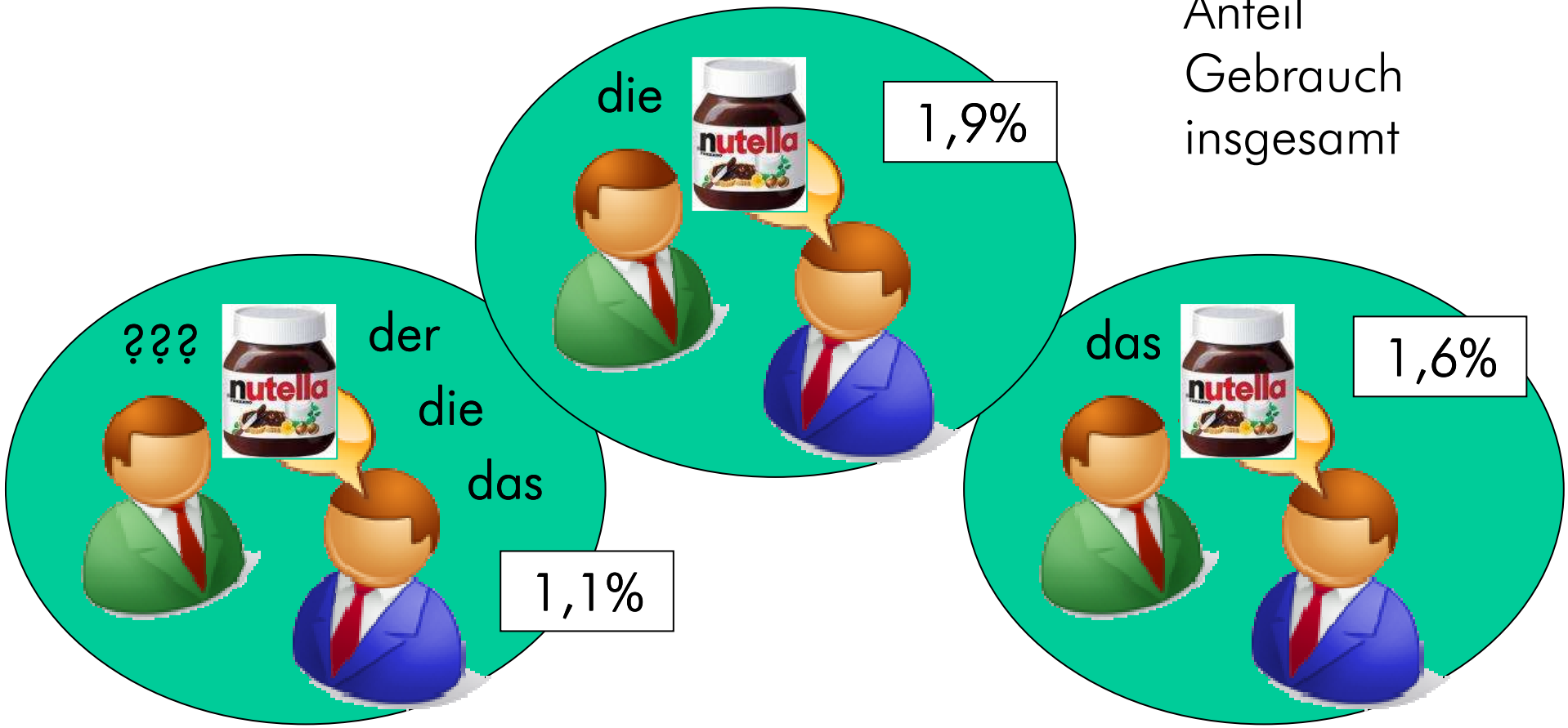
Bedarf und geteilter Brauch



Aussagekraft?



Aussagekraft?



„Deutung“

- kaum Aussagekraft
  - „der“ ist aus dem Rennen
  - ob Tendenz für „die“ abzulesen ist, bleibt fragwürdig, zumal „die“ mehr als drei Mal häufiger ist als „das“ (50 Mio. vs. 14.5 Mio.)
- wenig Gebrauch mit Artikel
  - gibt es überhaupt Bedarf sich zu einigen?
  - ist die Frage in diesem Fall relevant?

Präferenzen

- ob Notwendigkeit besteht, sich auf eine präferierte Form zu einigen, zeigt sich
  - über einen längeren Zeitraum …
  - über viele Menschen hinweg …

- gut beobachtbar:
  öffentlicher Schriftsprachgebrauch
  (Medien prägen stärker als Literatur)

Sprachgebrauch beschreiben

- korpusbasiert (DEREKO bekannt)
- Methodik bereits angesprochen
- nicht betrachtet:
  - alle Vorkommen werden beschrieben
  - nur einzelne Beispiele werden willkürlich ausgewählt
- zur Bewertung Typikalitäts- bzw. Relevanzbegriff erforderlich

Relevanz?

- Beschreibungen unterschiedlichster Art
- Nachschlagewerke und Lehrmaterial
- lehrer- oder schülerzentriert
- Sprachverstehen oder -erzeugen
- Alltagssituation oder Standardsprache

Mögliche Kriterien

- Häufigkeiten („Frequenzen“)
- Kookkurrenzen
- Grundwortschätze
- Verwendungssituationen
  (vgl. Referenzrahmen)

Worthäufigkeiten

- Wortformen vs. Grundformen
- Wortformen unnötig umfangreich
- Ungleichverteilung bei der Vielfalt im Paradigma
- vgl. Lemmabegriff in der Lexikografie
- vgl. Geläufigkeit in der Psychologie

Übersicht über Problembereiche

- Groß-/Kleinschreibung
- Trennzeichen/Bindestrich-schreibungen
- Neubildungen/Neologismen
- Fremdwörter/Anglizismen
- diskontinuierliche Konstituenten
- Eigennamen
- adjektivisch gebrauchte Partizipien
- Regionalismen
- Varianten/Varietäten (sprachlich regional, dialektal, Reform: Getrennt-/Zusammen-schreibung)
- Kurzwörter
- Akronyme, Abkürzungen z.B.
- unselbstständige Morpheme
- Verschmelzungen
- Movierung
- Häufigkeitsklassen (grafik!!)
- Grundformnennung

„Schieflage“ in den Daten

- „Unausgewogenheit“ in der Zusammensetzung der Daten („zuviel Sport oder Polizeiberichte“)
- virtuelles Korpus mit entsprechend gedämpften Anteilen

Regionalismen

- regional Überrepräsentiertes (aus dem Großraum Mannheim)
- Ausreißer statistisch abschätzen und menschlich bewerten
- ggf. aufgrund Vergleichswerte anderer Daten Frequenz anpassen

Präverbfügungen

- aka: abtrennbare Präfixe
- diskontinuierliche Konstituenten nicht zuverlässig automatisch zu erkennen
- Verteilung, welche Formen wie oft mit/ohne Präfix vorkommen, anhand von Testdaten abschätzen
- Werte für betroffene Grundformen neu abschätzen (Fügung↑, Präposition↓, Grundverb↓)

Eigennamen

- für welche Werke überhaupt relevant?
- Überlappungen verfälschen Frequenzen: „Kohl“ als Bundeskanzler oder als Gemüse
- stichprobenartig Vorkommen einordnen, Verteilung abschätzen und auf Gesamtheit hochrechnen

Lemmatisieren

- prinzipiell theoretisch verstanden
- Flektion, Derivation, Komposition
- aber:
  - Mehrdeutigkeiten („Floh“ von „fliehen“ oder „Floh“)
  - sprachliche Kreativität überschreitet Grenzen („einzig“ doch steigerbar?)
  - Sprachwandel kann nicht vorausgeahnt werden

Adjektivisch gebrauchte Partizipien

– „spannend“ vs. „entspannend“:
  eigene Grundform (da Adjektiv) oder nur
  Teil des Paradigmas des Verbs?

Neubildungen/Neologismen

- unbekannte Paradigmen bzw. Erweiterung
- Abweichen von bisher nicht durchbrochenen Konventionen (z.B. „politischste“)
- neue Paradigmen empirisch erschließen

Nennung der Grundform

- immer dieselbe Form aus dem Paradigma?
- selbst wenn diese gar nicht belegt ist?
- selbst wenn diese nicht alle Formen des Paradigmas suggeriert?

Wörterbuchabgleich

- Probleme: Lücken und Leichen
- Abgleich kaum aussagekräftig
  - kein Sortierkriterium für Relevanz
  - Vergleichswörterbücher stehen nicht in beliebigem Umfang zur Verfügung

Angabe der Häufigkeit

- Häufigkeiten, egal ob absolut oder relativ, sind keine aussagekräftigen Größen
  - **nicht vergleichbar** mit anderen Korpora gleicher oder unterschiedlicher Größe
- bewährt haben sich Häufigkeitsklassen, die das Verhältnis zur Frequenz des häufigsten Eintrags ausdrücken („der/die/das“)
  - **vergleichbar** mit anderen Korpora gleicher oder unterschiedlicher Größe

Häufigkeitsklassen

– Verhältnis logarithmisch (wie oft zu verdoppeln?)

$N = hk(\textit{lemma}) := \lfloor \log2(f(\textit{d-})/f(\textit{lemma})) \rfloor$

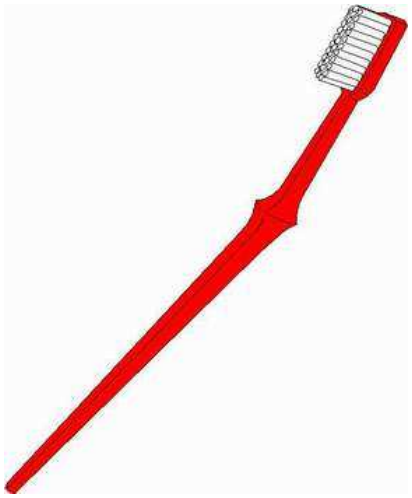
also: $f(\textit{lemma}) \approx f(\textit{d-})/2N$

$N =$	0	1	2	3	4	5		10		17
$2^N =$	2^0	2^1	2^2	2^3	2^4	2^5	…	2^{10}	…	2^{17}
$2^N =$	1	2	4	8	16	32		1024		131072
Bsp.	d-	…	und	von	als	Jahr		aktuell		Lachmuskel

– Zusammenspiel von Einheiten in Welt und Sprache und innerhalb der Sprache in vielen Aspekten nicht formal definierbar oder regelgeleitet …

„Fürwahr“

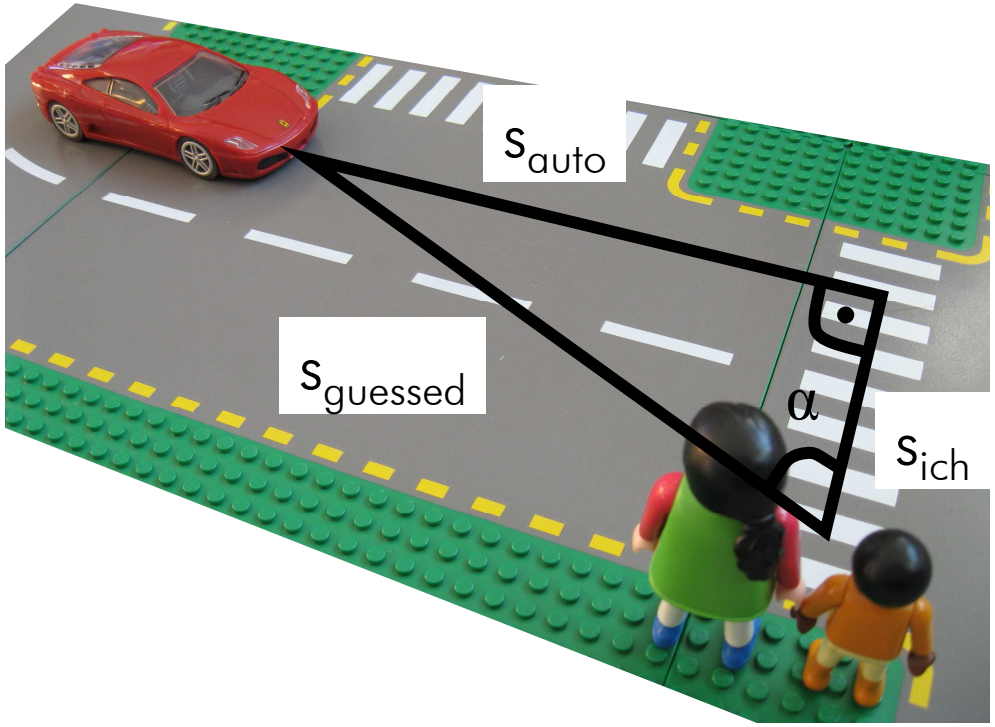
– Wiebke, 9 Jahre:
[als Reaktion auf ein scherzhaftes „Wohl bekomm‘s“]
„Fürwahr … ich weiß zwar nicht, was es bedeutet, aber ich finde, es passt irgendwie …“




HEUREKA !
DIE SCHLÜMPFE !
Da seid ihr ja !
Hatte Omnibus euch hergeschlumpft?
Wir haben die Schlumpfe mitgeschlumpft!
Jetzt könnt ihr Bruno Böse schlumpfen !
Schlumpft euch vor !
Die beiden Schlümpfe sind noch da !
Da haben wir aber Schlumpf gehabt !

Annahme

- die fundamentalen Mechanismen der Sprache
  (sicher: ihres Erwerbs,
  vermutlich auch: ihrer kognitiven Verarbeitung)
  - sind nicht direkt oder per Introspektion zu erheben
  - sind partiell funktional äquivalent zu einer statistischen Bewertung des Kontextverhaltens der Wörter im ‚Sprachinput'
- vgl. Mathematik beim Straßenüberqueren



$$s_{auto} = v_{auto} \cdot t_{auto}$$

$$s_{ich} = v_{ich} \cdot t_{ich}$$

$$\tan(\alpha) = \frac{s_{auto}}{s_{ich}}$$

$$\tan(\alpha) = \frac{v_{auto} \cdot t_{auto}}{v_{ich} \cdot t_{ich}}$$

$$\tan(\alpha) \cdot \frac{v_{ich}}{v_{auto}} = \frac{t_{auto}}{t_{ich}}$$

$$t_{ich} < t_{auto} \Leftrightarrow 1 < \frac{t_{auto}}{t_{ich}}$$

$$1 < \tan(\alpha) \cdot \frac{v_{ich}}{v_{auto}}$$

Kookkurrenz wichtig, weil …

- Annäherung an Bündelung von Ereignissen, die Musterhaftes enthalten als Input für unseren „kognitiven Apparat"
- kompakte Präsentation kann den langwierigen Prozeß des Wartens auf „vergleichbare Ereignisse" verkürzen

Kookkurrenz

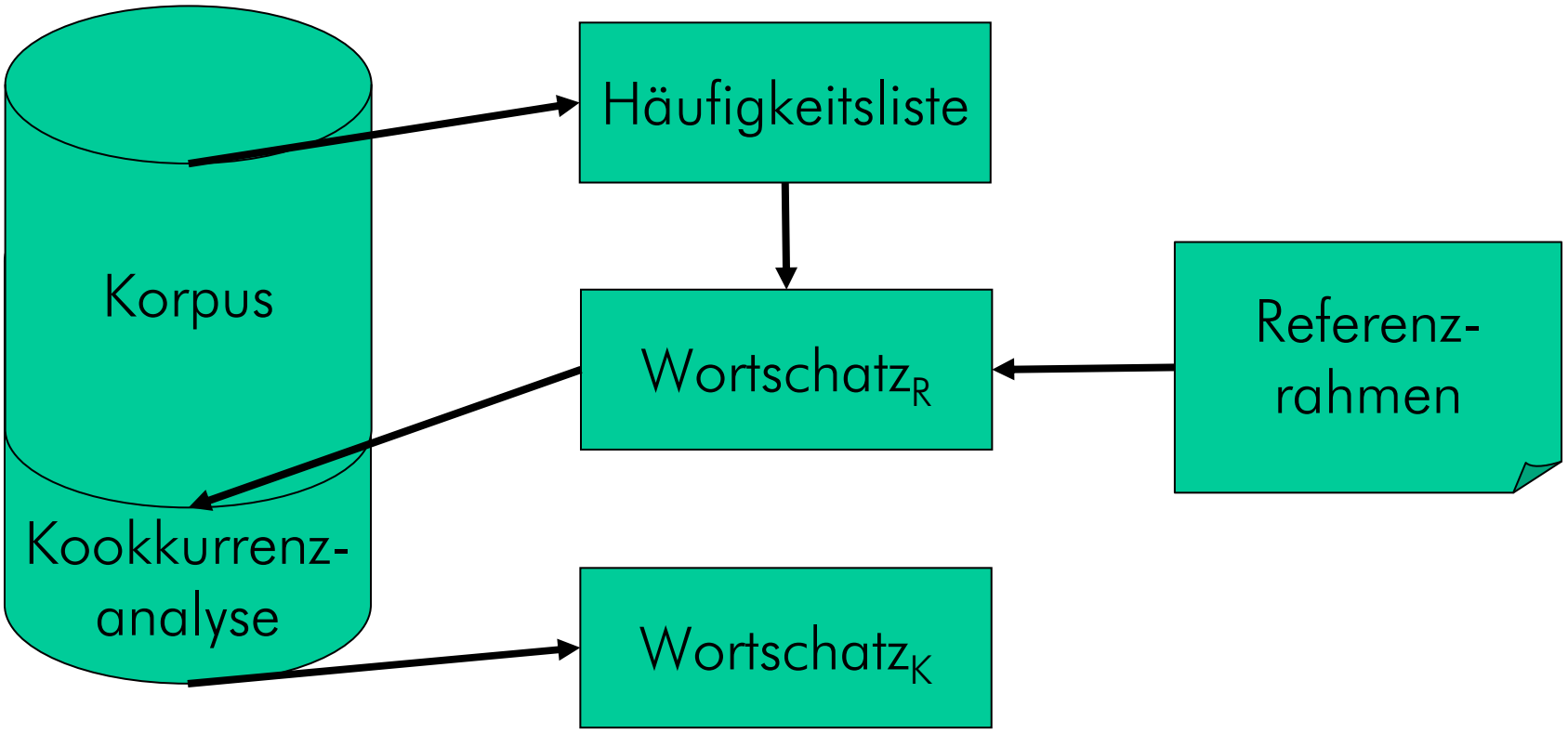
- bezogen auf eine bestimmte Datengrundlage
- hier: ideal für Beschreibung mit Ziel
  - Verstehen geschriebene Standardsprache
- anzupassen für andere Ziele
- Bezugswörter oder Kookkurrenzpartner auf bestimmte Wortschätze (z.B. Referenzrahmen A1) filtern

Wortschatz/Referenzrahmen

- Wortschatz „veraltet“
- Referenzrahmen nicht extensional
- Reflektion anhand Häufigkeiten
- welche Wörter werden häufig in den vom Referenzrahmen beschriebenen Situationen eingesetzt? (hier fehlt die Vorarbeit!!!)

Entscheidung?

- Entscheidung aus Zusammenspiel aller Mitspieler
- Häufigkeit in Wechselwirkung mit Referenzwortschätzen
- Kookkurrenzen


Korpus
Kookkurrenz-
analyse
Häufigkeitsliste
Wortschatz$_R$
Referenz-
rahmen
Wortschatz$_K$

Nicht gelöst …

- Rezeption vs. Produktion
- (mündliche) private Kommunikation vs. (schriftlicher) öffentlicher Diskurs
- regional vs. überregional
- literarische Sprache lesen können vs. bei (oder auch nach ☺) Geschäftsverhandlungen sicher auftreten können

Nicht gelöst …

- „Helferlein-Wörter“, nicht unbedingt standard-nahe Wörter, die Verstehens- oder Formulierungsprobleme überspielen helfen
  - vgl. „truc“ = „Ding“
  - „machen“, „tun“

Vielen Dank …

… für Ihre Aufmerksamkeit.

Informationen zur Methodik:
http://www.ids-mannheim.de/kl/
korpuslinguistik@ids-mannheim.de