

# Korpuslinguistik

Prof. Dr. Tibor Kiss  
WS 2007/08



SPRACHWISSENSCHAFTLICHES INSTITUT

## Definitionsversuch



- "The main requirements for Statistical NLP work are computers, corpora, and software. ... knowledge of some programming language is required." (Manning/Schütze 1999:117)
  - programming language required: corpora contain too many instances of **many** phenomena
- "Corpus linguistics is perhaps best described ... as the study of language based on examples of 'real life' **language use**. ... The corpus [is] subjected to a clear, stepwise, **bottom-up** strategy of analysis." (McInery/Wilson 2001:1ff.)
- "Corpus: Endliche Menge von konkreten sprachlichen Äußerungen, die als empirische Grundlage für sprachwiss. Untersuchungen dienen. Während der Strukturalismus ... ausschließlich von beobachtbaren Corpora sprachlicher Äußerungen ausgeht, sich induktiver Aufdeckungsprozeduren bedient ... und die Gültigkeit der Aussagen auf das jeweils zugrunde liegende Corpus einschränkt, spielen Corpora in der generativen Grammatik keine wesentliche Rolle." (Bußmann 2002:143)

## Beispiel: Volk (2006)



- Kann *ob* als Präposition verwendet werden?
  - Herangehensweisen: Experiment, Selbstbefragung (Introspektion), Fremdbefragung (Grammatik, Wörterbuch), Korpusanalyse
- Introspektion
  - Rothenburg *ob* der Tauber
- Wörterbuch: (Wahrig 1996)
  - Präp. mit Dativ (veraltet): *ob* dem Wasserfall
- Fremdbefragung (google):
  - liefert Resultate, zumeist aber Straßennamen oder mittelhochd. Texte.
- Korpusanalyse:
  - 1) ... fiel schier vom Stuhl *ob* der Äußerung eines Ozeanologen ...
  - 2) Bei manchen Ölgiganten kam *ob* der Resultate gar Euphorie auf.
  - 3) ... rieben sich vergnügt die Hände *ob* des zu erwartenden Schlagabtauschs ...

## Was haben wir da gemacht?



- Fragestellung: Kann ein bestimmtes Element Z in Kontexten [... Y ... Z ...] auftreten?
  - Gibt es die Konstruktion X?
- Erste Annahme (Intuition): **ja, im Kontext [... NP<sub>dat</sub>]**
- Unterstützung (Wörterbuch): **ja im o.g. Kontext, mit Einschränkungen.**
- Korpusanalyse: **ja, im Kontext [... NP<sub>gen</sub>]**
- Fragen
  - Gibt es die Konstruktion in beiden Kontexten?
  - Hat meine Intuition gepatzt?
  - Warum finde ich im Korpus kein [... NP<sub>dat</sub>]?

## Exkurs: Geschichte



- "After [the late 1950s] the corpus as a source of data underwent a period of almost total unpopularity and neglect. ... the debate that Chomsky triggered ... is a very old one - the debate between **rationalists** and **empiricists** ... this ... division exists ... within any discipline faced with the basic decision of whether to rely on **naturally occurring observations** or to rely on **artificially induced observations**." (McInery/Wilson 2001:5)
- Bei Chomsky ist die Domäne der '**artificially induced observations**' die Kompetenz (I(nternal)-Language) und die Domäne der '**naturally occurring observations**' die Performanz (E(xternal)-Language).
  - Kompetenz: Grammatikalität ⇒ Gegenstand linguistischer Theoriebildung
  - Performanz: Akzeptabilität ⇒ nur im Verhältnis zur Kompetenz relevant.
  - "Competence is our tacit, internalized knowledge of a language. Performance is external evidence of language competence and its usage on particular occasions when factors than our linguistic competence may affect its form." (McInery/Wilson 2001:6)
- Diese Unterscheidung spiegelt die Unterscheidung zwischen *langue* und *parole* wider, die sich bei Saussure findet.

## Langue, Parole, Entdeckung, Introspektion.



- Bereits Saussure (1916) hat postuliert, dass der Gegenstand der Sprachwissenschaft nicht die *parole* sein kann, definiert als der Bereich des aktuellen Sprachgebrauchs, der allerlei extrasprachlichen Bedingungen unterworfen sein kann.
- Beispiele (nicht von Saussure, eher von Chomsky)
  - Wir zeichnen die Konversation einer Gruppe Betrunkener auf.
  - Wir zeichnen die Konversation einer Gruppe Aphasiker auf.
- Strukturalismus: Jede Analyse der *langue* hebt in der *parole* an. Durch die Analyse der *parole* entdecken (*discovery procedures*) wir die Grammatik.
- Chomsky: Wir entdecken die Grammatik, weil wir sie haben und dieses Vermögen befragen können.

## Grammatikalität - Akzeptabilität



- Komplizierter wird dies durch die Unterscheidung grammatisch vs. akzeptabel.
  - Wenn Linguisten von grammatisch (I-Language, *langue*) sprechen, meinen sie meist akzeptabel (E-Language, *parole*, aber zurückführbar auf I-Language, *langue*).
  - Ein Satz kann grammatisch sein, aber möglicherweise nicht akzeptabel (ein sehr langer Satz von Thomas Mann).
  - Ein Satz sollte eigentlich nicht akzeptabel und zugleich nicht grammatisch sein. (Dieser Satz fordert den Leser.)

## Zwei berühmte Beispiele



- "The sentence 'I live in New York' is fundamentally more likely than 'I live in Dayton, Ohio.' purely by virtue of the fact that there are more people likely to say the former than the latter." (McInery/Wilson 2001:10)
- On performing tasks:
  - **Chomsky**: The verb *perform* cannot be used with mass word objects: one can *perform a task* but one cannot *perform labor*.
  - **Hatcher**: How do you know, if you don't use a corpus and have not studied the verb *perform*?
  - **Chomsky**: How do I know? Because I am a **native speaker** of the English language.
- "The quote underlines why corpus data may be useful. Chomsky was, in fact, wrong. One can *perform magic*, ... as a check of a corpus such the BNC reveals. Native speaker intuition merely allowed Chomsky to be wrong with an air of absolute certainty." (McInery/Wilson 2001:11)
- Hier haben wir einen anderen Typ von Fragestellung, nämlich die Behauptung: X existiert nicht.

## Kritik an Korpora



- Korpora sind überflüssig, ich kann allemal meine Intuition befragen.
- Annahmen über die Frequenz einer Konstruktion sind bestenfalls irreführend (New York-Dayton)
- Korpora sind unvollständig.
  - Woher weiß ich, dass das von mir untersuchte Phänomen in der Sprache so vorkommt wie in meinem Korpus? (Randfall: Phänomen taucht nicht auf.)
  - Ist das Korpus repräsentativ?
- "Without recourse to introspective judgements, how can **ungrammatical utterances** be distinguished from ones that simply haven't occurred yet? It is only by asking a native speaker ... that we can hope to differentiate unseen but grammatical constructions from those which are simply ungrammatical and unseen ... as language is non-finite and a corpus is finite, the problem is all too real." (McInery/Wilson 2001:11)
  - "If we do not find [certain sentences or constructions in a corpus], this is an interesting and important comment on their frequency." (McInery/Wilson 2001:15)

## Repräsentativität ist eine Maxime



- Auch wenn dies nicht in allen Texten klar wird: Repräsentativität ist eine Maxime.
  - "A sample is representative if what we find for the sample also holds for the general population." (Manning/Schütze 1999:119)
  - Wenn es für einen Ausschnitt des Deutschen gilt, dass die Präposition *ob* den Genitiv regiert, dann gilt es für das Deutsche, dass die Präposition *ob* den Genitiv regiert.
  - Wenn in einem Ausschnitt des Deutschen die Konstruktion [X Y Z] nicht auftritt, gilt nicht zwingend, dass es die Konstruktion im Deutschen nicht gibt.

## Repräsentativität ist eine Maxime



- "There is no easy way of determining whether a corpus is representative, but it is an important issue to keep in mind when doing Statistical NLP work." (Manning/Schütze 1999:120)
- Reparaturversuch 1: balanced (ausgewogen) corpus
  - "put together as to give each subtype of text a share of the corpus that is proportional to some predetermined criterion of importance" (ibid.)
  - "Brown corpus: a **representative** sample of written American English as used in 1961 includes particular texts in amounts proportional to actual publications (but excluding verse)." (MS: 119)
- Reparaturversuch 2: the more principle
  - "having more training data is normally more useful than any concerns of balance, and one should simply use all the text that is available" (MS: 120)

## Zeitungskorpora



- Zeitungskorpora sind
  - ... nicht repräsentativ, weil eine Vielzahl von denkbaren Konstruktionen in Zeitungskorpora nicht vorkommen.
  - ... nicht ausgewogen, denn sie sind beschränkt auf eine bestimmte Variante der Sprache.
  - "Ich habe absichtlich vergessen, die Pille zu nehmen." (BILD)

## Quantitativ vs. qualitativ



- Wenn wir aus dieser Perspektive zur Entscheidung rational vs. empirisch zurück kehren, dann ergeben sich die folgenden Perspektiven:
  - Aus der Perspektive der Korpuslinguistik (empirisch, bottom-up) kann man fragen: Warum sind Deine Aussagen wahr? Wodurch lässt sich die Wahrheit Deiner Aussagen bestätigen?
    - "Corpus-based observations are intrinsically more verifiable than introspectively based judgements." (McInery/Wilson 2001:14)
    - So gibt es häufig Nichtübereinstimmung mit Grammatikalitätsurteilen in Kursen und das Beispiel \*perform mass object haben wir ja auch schon diskutiert.
    - "Yes I could say that – but I never would."
    - "introspective judgements can only become available to us when our metalinguistic awareness has developed." (McInery/Wilson 2001:13)
  - Aus der Perspektive der generativ geprägten theoretischen Linguistik (rational, top-down) kann man fragen: Inwiefern sind Deine Aussagen repräsentativ?
    - So kann es etwa ein Zufall sein, dass das untersuchte Korpus nur Belege für ob+Gen aufweist und keine Belege für ob+Dat.

## Quantitativ vs. qualitativ



- qualitative Analyse
  - detaillierte Beschreibung linguistischer Phänomene
  - seltene Phänomene genau so wichtig wie häufige (also auch keine Aussage darüber, ob es sich um Zufälle handelt oder nicht)
- quantitative Analyse
  - Klassifizierung der Phänomene
  - Frequenzbestimmung
  - statistische Modelle (Signifikanz vs. Zufall)

## Vortläufige Zusammenfassung



- Idealerweise wären Korpora repräsentativ und eine Analyse der Korpora somit auch eine Repräsentation der Analyse der Sprache.
- Bei jeder Korpusanalyse ist es notwendig, extra-sprachliche Faktoren (d.h. Performanzfaktoren) zumindest in Betracht zu ziehen.
  - Häufiges Auftreten einer bestimmten Konstruktion bei falsch (absolut nicht repräsentativ) ausgewählten Korpora (Zeitungskorpora), weil identische Agenturmeldungen in unterschiedlichen Zeitungen gemeldet werden.
- Ein Verlassen auf die Intuition allein führt ebenso wenig zum Erfolg wie ein reines Beharren auf Korpusanalysen.
  - Methodologische Monokulturen sind abzulehnen (manchmal aber historisch notwendig).
  - "The main requirements for Statistical NLP work are computers, corpora, and software. ... knowledge of some programming language is required."

## Contra Monokultur ...



- Englisch ist keine reguläre Sprache (Schweizerdeutsch und Bambara sind keine KF-Sprachen).
  - Englisch  $\cap$  reguläre Sprache L  $\Rightarrow$  nicht-reguläre L', also kann Englisch nicht regulär sein, denn reguläre Sprachen sind unter Schnitt geschlossen.
- Kontextuelle Bestimmung der Zählbarkeit von Substantiven
  - Contra Chomsky (und vielen anderen) beruht die Unterscheidung Massenomenen vs. zählbares Substantiv (Individuenomen) nicht zwingend auf lexikalischen Eigenschaften, sondern ist wohl eher eine kontextuell determinierte Eigenschaft.
  - Ist die Bestimmung dieser Kontexte qualitativ oder quantitativ?
- Produktivität: "Fähigkeit von Wortbildungselementen zur Neubildung sprachlicher Ausdrücke. P. ist ein gradienter Begriff, der aufgeteilt wird in unproduktive Elemente ..., gelegentlich produktive Elemente ... und massenhaft produktive ..." (Bußmann 2002:537)
  - P(X) als Produktivitätsmaß (Baayen 2001) bestimmt als Wahrscheinlichkeit des Auftretens eines neuen Ausdrucks vom Typ X, nachdem ich schon so-und-so-viele Ausdrücke vom Typ X gesehen habe.



## Der Rest des Kurses

- Gesprochene vs. geschriebene Daten
- Rohdaten (raw data) vs. annotierte Daten (marked-up data)
  - raw data: plain text in some electronic form
  - marked-up data: has added explicit mark-up to the text to indicate something of the structure and semantics of the document. ... raises its own questions about the kind and content of the mark-up used (Manning/Schütze 1999:119)
- Annotationsformen
  - Wort: Part-of-Speech-Tagging
  - Phrase: Chunking (ohne rekursive Einbettung)
  - Satz: Baubank mit rekursiver Einbettung
  - Semantik: Annotation von Rollen oder anderen semantischen Eigenschaften und Beziehungen



## Wie arbeitet man mit Korpora?

- Vorverarbeitung
  - Zeitungskorpora oder .wav-Dateien werden tokenisiert und in komplexe Einheiten zerlegt, wobei Wort- und Satzgrenzen identifiziert werden müssen.
  - Rohdaten-Verarbeitung (<http://www.ids-mannheim.de/cosmas2/>)
    - Frequenzinformation (bis zur Produktivitätsberechnung)
    - Sprachmodelle für Buchstaben- und Wortabfolgen (Käding 1897, Markov 1913, Spracherkennung, SMS-Systeme)
    - Identifikation von Kollokationen
    - Suche mit regulären Mustern auf der Ebene des Vokabulars
  - Korpora mit Wortartenannotation
    - Suche mit regulären Mustern auf der Basis von Kategorien (Abstraktion)
    - Sprachmodelle auf der Basis von Kategorien
    - Identifikation von basalen Subkategorisierungsmustern



## Rohdaten

- Arbeit mit Rohdaten
  - Korpus auswählen
  - Korpus tokenisieren
    - Interpunktion identifizieren und abtrennen
    - Problem der Satzgrenze vs. Abkürzung
  - Korpus ggf. auch lemmatisieren
    - 1) Ich **zolle** dieser Leistung **Respekt** .
    - 2) Man sollte hier zumindest **Respekt zollen** .
    - 3) Das ist **freie** Marktwirtschaft .
    - 4) Wir leben in der **freien** Marktwirtschaft .
  - Problem angehen ...
    - Testverfahren (selbst in Perl o.ä. implementiert oder R oder Excel oder ...)
  - COSMAS löst eine Vielzahl dieser Probleme für uns (Tokenisierung, Lemmatisierung, ...)



## Was sind Kollokationen?

- "Charakteristische, Wortverbindungen, deren gemeinsames Vorkommen auf einer Regelmäßigkeit gegenseitiger Erwartbarkeit beruht, also primär semantisch (nicht grammatisch) begründet ist: *Hund: bellen, dunkel: Nacht ...*" (Bußmann 2002:353)
- "An expression consisting of two or more words that correspond to some conventional way of saying things, ... characterized by limited *compositionality*." (Manning/Schütze 1999:151)

## Kollokationen



- (Adjazente), partiell nicht-kompositionelle Kombinationen, bei denen die freie Ersetzbarkeit *salva veritate* nicht gegeben ist.
  - häufig auftretend
  - gegenseitige Erwartbarkeit von  $w_1$  und  $w_2$  (*helllicht Tag*)
  - correspond to some *conventional way* of saying things (*Ziele setzen vs. \*Ziele planen*)
- Warum ist die Identifikation relevant?
  - wenn nicht kompositionell bzw. nicht austauschbar, dann ist die Identifikation von Kollokationen für das Sprachenlernen relevant (Wörterbücher)
  - Kollokationen sollten ggf. zusammen tokenisiert werden
  - mal sehen, wie stark konventionalisiert Sprache ist

## Reine Frequenz und ... Funktionswörter



- häufig auftretend
  - Ein einfaches Verfahren zur Identifikation von Kollokationen scheint die Ordnung von Bigrammen nach der Frequenz zu sein. Hier reagiert man auf die Annahme, dass  $w_1$ ,  $w_2$  als Kollokation gemeinhin häufiger auftritt als bei kompositioneller Verbindung.
  - aber ... vgl. Manning/Schütze (1999:154)
    - of the (80.871), in the (58.841), ... to be (15.494), ... in a (13.899), ... he said (10.007), ... has been (8753)
  - Selbstversuch (Ausschnitt NZZ 1993, nicht tokenisiert, 948.698 Wörter)
    - Zürcher Zeitung (6601), Neue Zürcher (6571), für die (4717), in der (3034), in den (1600), und die (1066), auf die (987) ... dass die (736) ... sich die (588) ... auch die (517) ... vor allem (448) ... auf der (438) ... um die (382) ... nicht mehr (355), in einer (352) ... sich in (341) ... ist die (310), nicht nur (307)
- Das Problem ist das **äußerst häufige Auftreten von Funktionswörtern**. Für die vorliegende Aufgabe bilden diese Funktionswörter die Klasse der "Stop Words".

## Vorgehen: Hypothesentests



- Statistische Verfahren zur Identifikation von Kollokationen
  - Wir nehmen an, dass wir ein Verfahren haben, um Stop Words zu eliminieren.
  - Wir betrachten zunächst nur adjazente Kollokationen (keine Tests auf Mittelwert und Varianz)
  - Vier Hypothesentests
    - t-Test
    - $\chi^2$ -Test
    - log likelihood ratio ( $\log \lambda$ )
    - mutual information

## Hypothesentests



- Hypothesentests basieren auf der Annahme, dass es eine Nullhypothese gibt, die entweder verworfen (rejected) werden kann oder nicht.
- Wenn die Nullhypothese nicht verworfen werden kann, wird sie akzeptiert.
- Aus einem solchen Test folgt weder, dass eine bestimmte Hypothese mit Sicherheit bestätigt, noch, dass eine solche Hypothese mit Sicherheit verworfen werden kann, vielmehr ergibt sich eine Annahme auf einem bestimmten Signifikanzniveau.
- "The performer of a significance test seeks to determine whether the null hypothesis **is reasonable given the available data.**"

## Hypothesentests



- Wir berechnen zunächst die Wahrscheinlichkeit  $p$  des Eintretens eines Ereignisses, wenn die Nullhypothese wahr ist.
- Die Nullhypothese wird abgelehnt, wenn  $p$  zu niedrig ist bzw. unter einem bestimmten Signifikanzniveau liegt.
- $p < 0.01$  bedeutet: Der beobachtete Wert tritt unter der Nullhypothese mit einer Wahrscheinlichkeit von 1 % (oder weniger) ein.

## Was ist die Nullhypothese?

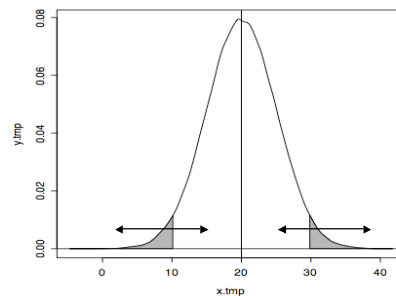


- Wir nehmen als Nullhypothese bei den Testverfahren an, dass die Wahrscheinlichkeit des Auftretens von  $w_1$   $w_2$  in einem Korpus mit der Größe  $N$  bestimmt wird als
  - $P(w_1, w_2) = P(w_1) \times P(w_2) = C(w_1)/N \times C(w_2)/N$
  - dies ist die sog. Unabhängigkeitsannahme.
- Die Alternativhypothese lautet:  $P(w_1, w_2)$  weicht für Kollokationen von dem unter der Nullhypothese bestimmten Wert signifikant ab.

## Wie ist das Signifikanzniveau zu interpretieren?



- Wie weit weicht ein beobachteter Wert von einem gegebenen oder postulierten Mittelwert ab?



## t-Test zur Identifikation von Kollokationen



- Nullhypothese: Mittelwert des Auftretens von  $w_1$   $w_2$  wird bestimmt aus dem unabhängigen Auftreten von  $w_1$  und  $w_2$ , d.h.  $P(w_1, w_2) = P(w_1) \times P(w_2)$ .
- Dieser Wert wird ins Verhältnis gesetzt zum tatsächlich beobachteten Auftreten von  $w_1$   $w_2$ :
- $t = (x' - \mu) / \sigma / \sqrt{n}$
- Hierbei sei  $\sigma = P \times (1 - P)$ . (Wir kennen die tatsächliche Standardabweichung nicht.)
- Beispiel: *new companies*
  - $x' = 8/14307668$ ;  $\mu = (15828/14307668) \times (4675/14307668)$ ;  
 $\sigma = x' \times (1 - x')$ ;  $n = 14307668$

## $\chi^2$ -Test



- Eine Kritik am t-Test ist, dass hier angenommen wird, dass eine Normalverteilung vorliegt (d.h. Kollokationen sind mit Mittelwert  $\mu$  und Standardabweichung  $\sigma$  in Texten normalverteilt). Diese Annahme ist nicht korrekt.
- Der  $\chi^2$ -Test macht diese Annahme nicht. Hier wird die Diskrepanz zwischen beobachteten und erwarteten Werten betrachtet, wobei jeweils 4-Felder-Tabellen zugrunde gelegt werden. (Siehe Excel-Tabelle)

## log likelihood



- Verhältnis zweier Hypothesen
  - Nullhypothese (Hypothese 1):  $w_2$  ist nicht abhängig von  $w_1$ .
  - Alternative (Hypothese 2):  $w_2$  ist abhängig von  $w_1$ .
- Allgemeines Schema: Verhältnis zwischen dem gesamten Parameterraum und einer Teilmenge dieses Raums.
  - Hier:  $p_1$  und  $p_2$ , Wahrscheinlichkeiten für das Auftreten von  $w_2$  hinter  $w_1$  und  $\neg w_1$ ; Teilmenge:  $p_1 = p_2$ .
- Gestattet auch die Identifikation von Kollokationen, die nur sehr selten auftreten.
- Log likelihood ratio gibt Auskunft darüber wieviel wahrscheinlicher das Vorkommen eines Bigrams  $w_1 w_2$  unter der Alternativhypothese ist.
  - **powerful computers**:  $\log \lambda = 82.96$ , bedeutet: dass *powerful computers* adjazent auftritt, ist unter der Alternativhypothese  $e^{0.5 \times 82.96}$  ( $\approx 1.3 \times 10^{18}$ ) wahrscheinlicher als unter der Nullhypothese.

## Bestimmung der Wahrscheinlichkeiten



- Hypothese 1:  $P(w_2|w_1) = p = P(w_2|\neg w_1)$
- Hypothese 2:  $P(w_2|w_1) = p_1 \neq p_2 = P(w_2|\neg w_1)$
- Maximum Likelihood Estimation:
  - $p = C(w_2)/N$
  - $p_1 = C(w_1, w_2)/C(w_1)$
  - $p_2 = C(w_2) - C(w_1, w_2)/N - C(w_1)$
- Binomialdistribution: Bei  $n$  Versuchen  $k$  Erfolge mit einer Erfolgswahrscheinlichkeit von  $x$  (ohne

Zurücklegen) ist:  $\binom{n}{k} x^k \times (1-x)^{(n-k)}$

## Bestimmung der Wahrscheinlichkeiten



- Die Wahrscheinlichkeit der tatsächlich beobachteten Häufigkeiten für  $w_1, w_2$  und  $w_1, w_2$  entspricht dann:
  - $L(H_1) = b(C(w_1, w_2); C(w_1), p) \times b(C(w_2) - C(w_1, w_2); N - C(w_1), p)$
  - $L(H_2) = b(C(w_1, w_2); C(w_1), p_1) \times b(C(w_2) - C(w_1, w_2); N - C(w_1), p_2)$
- $-2 \log \lambda = -2 \log L(H_1)/L(H_2) =$   
 $-2 \log L(C(w_1, w_2), C(w_1), p) +$   
 $\log L(C(w_2) - C(w_1, w_2), N - C(w_1), p) -$   
 $\log L(C(w_1, w_2), C(w_1), p_1) -$   
 $\log L(C(w_2) - C(w_1, w_2), N - C(w_1), p_2)$
- $L(k, n, x) = x^k (1-x)^{(n-k)}$ 
  - "n über k" ist konstant und kann gekürzt werden.

## $\log \lambda$ und $\chi^2$ -Verteilung



- $-2 \log l$  ist asymptotisch  $\chi^2$ -verteilt, d.h. man kann die Grenzwerte der  $\chi^2$ -Verteilung verwenden.
  - Der Grenzwert für 99.95 %-Wahrscheinlichkeit ist 7.88.
  - In der Praxis ist aber nicht das Überschreiten des Grenzwerts alleine ausreichend, es wird auch der Wert ins Verhältnis zu dem Werten anderer Paare gesetzt.
  - most powerful ist 'kollokativer' als powerful computers.