Need and Competition Deconstructing Quantitative Productivity

Anke Lüdeling, Berlin
Marco Baroni, Bologna/Forlì
Stefan Evert, Osnabrück

Outline

- qualitative and quantitative productivity
  - need
  - competition
- the 'too much' data
  - need
  - competition

Qualitative and quantitative productivity

- In a generative model a morphological process is either possible (grammatical) or not (ungrammatical)
→ qualitative productivity, availability
- morphologists have always wanted to express something like 'the ease with which a process can apply' (witness expressions like 'very productive', 'marginally productive' etc.)
→ quantitative productivity, profitability
- Baayen 1989, 1992, Baayen & Lieber 1991, Plag 1999, Bauer 2001, Lüdeling & Evert 2003, Meibauer, Guttropf & Scherer 2004, Nishimoto 2004, …

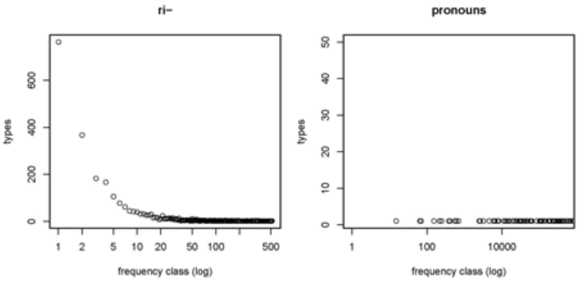
productivity measures

- a number of measures have been proposed, based on proportion of unseen types to types or on number of restrictions
(e.g., Booij 1977, Aronoff 1976)
- these have been criticized on linguistic and on mathematical grounds
- following the work of Harald Baayen (1989, 1992 etc.) productivity measures are proposed that are based on the distribution of types and tokens produced by a given word-formation process (most well-known Baayen's *P*)

productivity measures: the basic idea

- select a word-formation process
- count the types and tokens of all complex words in a given corpus
(this already implies a lot of qualitative analysis, see Lüdeling, Evert & Heid 2000)
- calculate a productivity measure
(e.g., Baayen's *P*)
- the measures rely on low frequency types
- basically: the more low frequency types are generated by the wf process, the more productive it is
– because low-frequency types indicate new formations

productivity measures: frequency spectrum



productivity measures: criticism

- mathematical: not possible to directly compare productivity measures for processes with different corpus sizes (fitting of models for extrapolation difficult)
→ discussed before, see Baayen 2001, Evert and Baroni 2005, Gaeta and Ricca (to appear)
- empirical: measures dependent on size and design of corpus
→ discussed before, used as a measure in stylometry (Tweedie and Baayen 1998) and diachronic productivity studies (Scherer 2005)
- linguistic: interpretation of the measure as purely linguistic and as inherent property of a single wf process
→ topic of this talk

interpretation of productivity measures

- *"An important property of P is that it expresses in a very real sense the probability that new types will be encountered when the item sample is increased. [...] The main interest of P is that it is the quantitative formalization of the linguistic notion of productivity."* Baayen (1992, 115)
- *"We argue that a measure of productivity based on the token frequencies of types, specifically on the number of hapax legomena for a given affix in a corpus, comes very close to according with our intuitions about productivity."* (Baayen & Lieber 1991, 801)

linguistic problems of productivity measures

- all measures of productivity rely on corpus counts and are interpreted as indices of the independent degree of linguistic productivity of a wf process
- however: the corpus counts are influenced by a number of factors (even if we assume a balanced corpus)
- the counts therefore reflect a 'mixture' of
  - need - extra-linguistic
  - competition - linguistic, sociolinguistic, psycholinguistic
  - persistence - psycholinguistic
  - 'inherent' productivity? - linguistic
  - ...

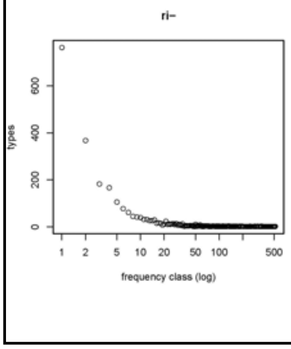
need

- corpus counts are influenced by the need to express a given thought/concept

> *Die Möglichkeit zur Bildung von Zuss. aus zwei Substantiven ist unbegrenzt. Ob solche aber wirklich gebildet werden, hängt natürlich vom Bedürfnis ab* (Paul 1920, 15)
> "The possibility to form noun-noun compounds is unlimited. Whether they are actually formed, however, depends on the need"
>
> *Words are only formed as and when there is a need for them [. . .]* (Bauer 2001, 143)

- the need to express something depends on fashion, the political situation etc. (Plag 1999)
→ **extra-linguistic** factors

need ans measures of productivity



- typical interpretation: productivity of *ri-*
- reflects the need (extralinguistic) mixed with the ‚inherent productivity' (linguistic)
- for single wf processes corpus counts do not reflect productivity

competition

- corpus counts are influenced by competition
- any need can be expressed by (in principle infinitely) many ways, morphological and syntactic
- not only competition in terms of truth-functional semantics: connotation, register, etc.
- some of the realizations are closer to each other than others
(competition cannot be modeled as random noise)
- some are more likely than others
- the likelihood of the competitors influences the likelihood of each process

aside: competition in linguistics

- Optimality Theory
  - competition between constraints
  - competition between candidates to find **the** optimal one
    – most candidates not well-formed
  - morphology: type blocking, token blocking (Plag 1999 → no genuine competition in wf)
- Minimalism
  - principles of economy

aside: competition in linguistics

- competition among well-formed objects plays a role in many linguistic fields (typically not in generative linguistics proper):
  - historical linguistics: language change, variation
  - sociolinguistics: dialects, registers, variation
- mainly descriptive, mostly no fully worked-out mathematical model of competition

„inherent“ productivity

- does it exist?
- how can we go about studying it?
  - find morphological processes that express the same need (qualitative)
  - select suitable corpus
  - find instances of the processes in the corpus
  - develop a model to account for their distribution (we are still working on this!)

the 'too much' corpus

- the 'too much' data
  - need
  - competition

find morphological processes expressing the same need

- must pertain to very specific need
- relatively ‘rare’ wf processes
- candidate instances of wf processes must be easy to spot by automated means
- the ‘too much’ data: several word formation processes that express the notion that somebody is doing too much of something and have an ‘illness’ connotation
- all instances of compounding

the ‚too much‘ heads

- non-medical *-itis*, as in *Telefonitis* ‘using the telephone too much’
- *wahn*, as in *Abbawahn* ‘playing too much music by Abba’
- *hysterie*, as in *Absicherungshysterie* ‘worrying too much about security’
- *zwang*, as in *Ausgehzwang* ‘having to go out too often’
- *sucht*, as in *Ausstattungssucht* ‘using too much equipment (in a movie)’
- *besessenheit*, as in *Besitzbesessenheit* ‘being obsessed about one’s possessions’
- *obsession*, as in *Computerobsession* ‘being obsessed about computers’
- *manie/mania,* as in *Handymanie* ‘using the mobile too much'

selecting a suitable corpus

- we need a large corpus (Lüdeling & Evert 2005)
- deWaC: more than 1.5 billion tokens of German from the Web (Baroni & Kilgarriff 2006)

collecting the data

- all potential forms in corpus extracted with regular expressions
- de-duping, clumpiness effects
- manual preprocessing necessary
  - noise
  - semantics

collecting the data: noise

- the regular expressions find words that are not built by the targeted wf processes:
  → these have to be thrown out
- typos in the data that can be clearly recognized are normalized:
  Effizienswahn / *Effizienzwahn* ’obsessing about efficiency’

collecting the data: other readings

- all heads have medical readings
  → have to be thrown out
  -itis ’inflammation’, as in *Arthritis* ’inflammation of the joints’
  sucht ’addiction’, as in *Drogensucht* ’drug addiction’
- with all heads we find compounds that have readings other than the “too much” reading
  → have to be thrown out
  Behördenzwang ’force by the authorities’
  Medienhysterie ’hysteria caused by the media’

competition 1: categorical

	besessen heit	hysterie	-itis	manie	obsession	sucht	wahn	zwang
simplex N	✓	✓	✓	✓	✓	✓	✓	✓
complex N	✓	✓	✓	✓	✓	✓	✓	✓
deverbal N	✓	✓	✓	✓	✓	✓	✓	✓
V	✓	✓	✓	✓	✓	✓	✓	✓
Adj	✓	✓	✓	✓	✓	✓	✓	✓
neocl	✓	✓	✓	✓	✓	✓	✓	✓
Engl	✓	✓	✓	✓	✓	✓	✓	✓

competition 2: in context

- is there competition in a given context?
- speaker‘s perspective: is there a choice between several options to express the same concept?
- comparable contexts in the data (our analysis)
- very small Web-experiment (10 participants)
  with 'too-much' contexts and specific contexts, ratings from 1 (very good) to 6 (unacceptable)

competition 2: in context

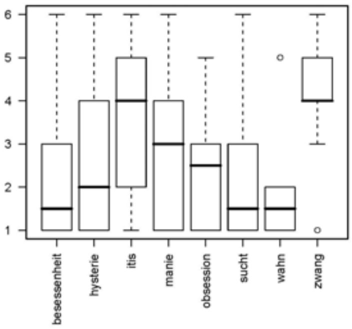
Das hieße, dass ich auch auf 20 Prozent verzichten müsste. Dann kann ich mir die Dauerkarte aber nicht mehr leisten. Egal, _______ macht doch eh nur Probleme. Streit mit der Freundin. Zeitverschwendung.

> „This would mean that I would have to make do with 20 percent less. Then I can no longer afford a season's ticket. Anyway, _______ only causes problems. Fights with the girl-friend. Waste of time."

diese blöde Fußballsucht
dieser blöde Fußballzwang
dieser blöde Fußballwahn
diese blöde Fußballmanie

diese blöde Fußballbesessenheit
diese blöde Fußballobsession
diese blöde Fußballhysterie
diese blöde Fußballitis

competition 2: in context



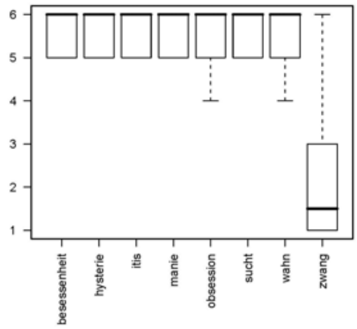
competition 2: in context

Im Jahr 1571 übernahm Philipp II. die Regierungsgeschäfte, nachdem er an der Universität Ingolstadt eine sorgfältige und natürlich streng katholisch ausgerichtete Ausbildung erfahren hatte. Es muss angenommen werden, dass ein zwölfjähriger Junge, dem die Machtfülle \n eines Regenten übertragen worden ist, vielerlei Einflüssen aus seinem Beraterkreis ausgesetzt war. Wohl unter dem Einfluss der \n mütterlichen Familie verfügte er als erstes ______ zum katholischen Gottesdienst.

> In 1571, Philipp II took over government, after a thorough and strictly catholic education at the University of Ingolstadt. It has to be assumed that a 12-year old boy who is invested with the power of a regent is exposed to manifold influences from the circle of his advisors. Apparently under the influence of his maternal family he first of all decreed _______ of the catholic service.

→ *Besuchszwang* ‚obligatory attendance'

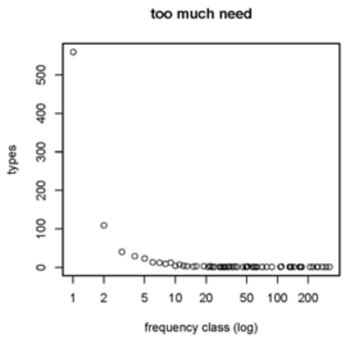
competition 2: in context



need and productivity

- for each non-head, compute frequency with whole set of need-satisfying heads
- e.g.: *Anglizismus* occurs 7 times with *-itis*, 4 times with *wahn*
- the 'too much' need frequency of *Anglizismus* is 11
- we measure productivity of extra-linguistic need to create a compound noun expressing notion that something is done too much. . .
- not a word-formation process!

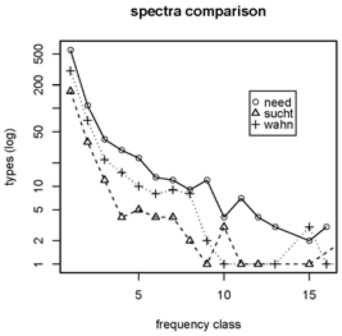
need and productivity



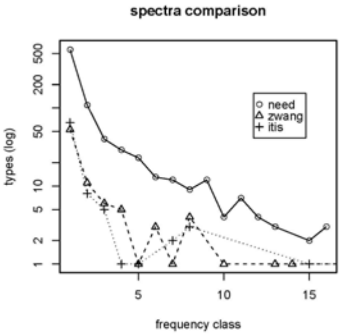
need and productivity

- the “too much” need has the typical shape of a productive process
- Germans feel stressed out about many different, constantly new things being done too much, and want to talk about it
- we cannot conclude that *sucht* or other word formation process is productive based on frequency spectrum
- productive shape is (at least to some extent) reflection of productivity of need
- what is the relation between the frequency spectrum of the overall need and the frequency spectrum of the specific processes?

need and productivity



need and productivity



need and productivity

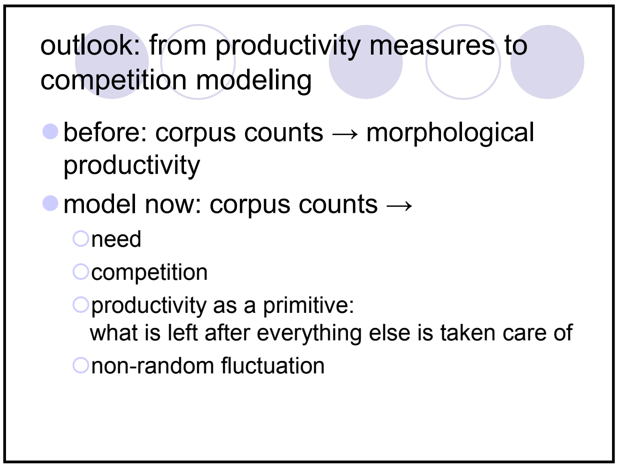
- frequency spectra of *sucht* and *wahn* are close reflections of need spectrum
- *zwang* and *-itis* patterns suggest some autonomy from need
(or data sparseness issues?)

conclusion

- quantitative approaches to productivity do not measure what everybody thinks they should measure
(i.e., an inherently linguistic property of a wf process)
- instead, all measures based on corpus data are influenced by a mixture of extra-linguistic factors, as well as irrelevant linguistic factors
- we must take need and competition into account
- productivity linguistically interesting only in the presence of competition

outlook: from productivity measures to competition modeling

- data
  - more data, better data
- factors that influence the choice
  - constant factors
    - non-categorical phonological, morphological, syntactic, semantic effects pertaining to heads, non-heads and their combination
e.g.: *obsession* prefers simplex nouns
  - contingent factors
    - stylistic, discourse-related, persistence-driven, collocational effects



outlook: from productivity measures to competition modeling

- before: corpus counts → morphological productivity
- model now: corpus counts →
  - need
  - competition
  - productivity as a primitive:
    what is left after everything else is taken care of
  - non-random fluctuation

Thank you.
Grazie.

Danke.