APPR-NN-Sequences and their grammar

Prof. Dr. Tibor Kiss
Cogeti Heidelberg
24.11.2006

Sprachwissenschaftliches Institut

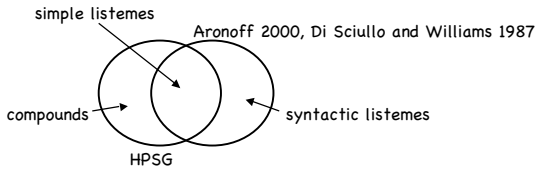
On the relation between lexicon and grammar

- HPSG's view on lexicon and grammar
  - sign → $_{\text{lexical-sign}}[] \vee _{\text{phrasal-sign}}[]$

    $_{\text{lexical-sign}}[] \wedge _{\text{phrasal-sign}}[] = \bot$ (Pollard and Sag 1987:43)
  - Partitions of sign: word, phrase
    phrase[DTRS con-struc] (Pollard and Sag 1994:396ff.)
    „sign is the ... greatest lower bound of word and phrase, but word and phrase have ... no least upper bound (i.e. they are mutually inconsistent." (Pollard and Sag 1994:31fn29)
  - cf. also the slightly more articulated characterization in Sag, Wasow, Bender (2003:473ff.)
- Clearly influenced by formal language theory
- The major distinction is simple (no DTRS) vs. complex (DTRS)
- Every complex entity is part of the grammar and thus requires a syntactic analysis

On the relation between lexicon and grammar

- Alternative view (Aronoff 2000)
  - the major divide between lexicon and grammar is not a matter of complexity, but of predictability and memorization.
  - Predictable items are described by the grammar.
  - Items to be memorized are listed in the lexicon.



Why is the alternative attractive?

- It allows to ask some question which simply do not pop up under the HPSG-view.
  - Is a syntactic construction regular in the sense that its behaviour can be predicted and extended to a possibly infinite set of instances? (Grammar should not be concerned with finite sets of instances.)
  - For which seemingly complex entities is it useful to provide grammatical descriptions? (Should all idioms receive a syntactic analysis or only decomposable idioms?)
  - Under which criteria become complex entities subject to listing?

A case study: APPR-NN

- What do we make of APPR (P) + NN (N)?
  - Die Saarbergwerke hingegen rechnen **unter-APPR Berufung-NN** auf „ernstzunehmende Energieprognosen" mit einem Exportbedarf beim Strom … [Refering to 'serious forecasts', the Saar Mining Company assumes that there is a future need to export electricity.]
  - **Unter-APPR Berücksichtigung-NN** dessen, dass das Videoband echt schlecht ist, müssen wir sagen, dass die Frisur hinkäme. [Considering that the tape was of bad quality, we would agree that it was the haircut we saw.]
  - A first guess: APPR+NN (i.e. P+Noun) = PP
- The Problem (Duden 442):
  - *Substantive mit Merkmalkombination ‚zählbar' plus Singular haben … grundsätzlich immer ein Artikelwort bei sich, und wenn es als letzte Möglichkeit der indefinite Artikel ist.* [Hence, count nouns marked singular are always combined with a determiner, and it has to be an indefinite determiner if other determiners are blocked.]

Ungrammatical sequences?

- Chafe (1968) observed anomalous idiomatic expressions like
  - by and large, no can do, trip the light fantastic, kingdom come, battle royal …" (Nunberg et al. 1994, quoting Chafe 1968)
- "[W]e do see no alternative to simply listing expressions like these." (Nunberg et al. 1994, 515)
- One solution would be to assume that APPR-NN (P+Noun) expressions are anomalous idiomatic expressions that will be listed.
- Ungrammatical sequences exist, but we do not need a grammar for them.
- But this will only work if the set of APPR-NNs is finite.

Fundamental questions

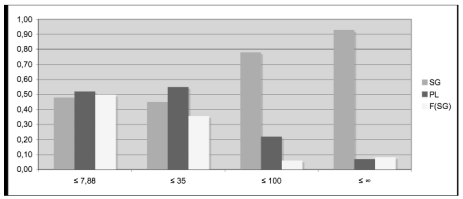
- We need a grammar to describe sequences A B if the following conditions obtain
  - There are infinitely many instances of sequences A B.
  - There is a compositional relationship between A, B and [A B] such that the meaning of [A B] can be determined on the basis of A and B.
- The big questions
  - Are there infinitely many instances of APPR-NN sequences?
  - Is there a compositional relationship between APPR, NN, and the combination of APPR+NN?
- Both questions have received negative answers.
  - Fleischer (1982, 300): „Die Bildungen sind zum größten Teil idiomatisiert …" [The combinations are mostly idiomatic …]

Accentuate the negative …

- Pretending that there are finitely many instances of APPR-NN-sequences, and that the semantics of APPR-NN-sequences is non-compositional in nature, what can be do, given an HPSG style divide between lexicon and grammar?
- Little, next to nothing.
  - APPR-NN sequences are complex, hence instances of phrasal sign and thus would require a full-fledged syntactic analysis.
- Lucky enough, negative answers should not be taken for granted.
  - (But it should be kept in mind that questions of regularity do not play a role in HPSG, and hence that the question whether a construction is finite or not will not even be raised.)

Idiomaticity

- Some interesting results of a log λ association measurement
  - Approximately 50 % of all singular occurrences of NN after 'unter' show a log l value below 7,88, i.e. are not highly enough associated to justify the idea that they are mutually dependent.
  - If P+Noun is ranked according to log likelihood, the likelihood of finding a plural noun in the top ranks is very small.
  - Combinations of P+$Noun_{pl}$ are completely regular since NP → N_{pl}.

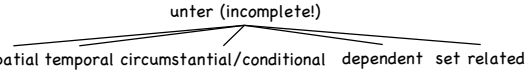


The intuition problem

- A compelling observation (brought to my attention by Joachim Jacobs) is that speakers of German
  - are unable to coin new P+Noun combinations on-the-fly and
  - usually cannot judge the grammaticality of a P+Noun combination without a given context
  - This does not hold for combinations which are built on basis of an ordinary N-N compound rule.
- This observation is in accord with the view that P+Noun are non-syntactic, or more generally, non-composed units that do not follow a rule of grammar.
  - It implies that the set of P+Noun (not built by N-N compound) may be large but that it can be listed.
  - We have to show that it is impossible to list P+Noun combinations that are not built by an N-N compound rule.

How compositional is *unter*+Noun?

- 'unter' has a complex meaning, obviously including 'below' ...
  - ... and some more.



- In compositional combinations, i.e. [$_{PP}$ P NP], all types of 'unter' can be found, while in P+Noun combinations,
  - spatial and temporal uses of 'unter' are under-represented.
  - set-related uses are very common but irrelevant, because [$_{PP}$ P [$_{NP}$ $Noun_{pl}$]] is not affected by rule 442.

Conditional/circumstantial *unter*

- Circumstantial
  - Die Gruppe von acht Schulleitern aber, die unter Anleitung des künftigen Oberschulrats Peter Daschner ... ihre Ideen zu Papier brachte, fühlt sich unverstanden.
    (The eight deans, who pinned down their ideas under the lead of PD, see themselves misrepresented.)
  - [R]und vier Milliarden Mark waren die Staubsauger und Schokoladenriegel wert, die unter Umgehung der Kassen in ihren Taschen landeten.
    (The hoovers and candy bars, who were taken by circumventing the cassiers were worthy an approximate 4 billion Marks.)
- Conditional
  - Die Arbeitsgemeinschaft Berliner Mieterberatungsgesellschaften ... hatte bereits Ende Dezember die betroffenen MieterInnen aufgefordert, die Mieterhöhung im Januar nur **unter Vorbehalt** zu zahlen.
    (The tennant advice center of Berlin had already suggested by the end of December that tennants should pay their increase of rent only with reservation.)

The 'light P' hypothesis

- What about a 'light P' analysis, i.e. the semantics of *unter*+Noun is determined by Noun, and is thus falsely attributed to the semantics of *unter*?
  - Why do Nouns show up with more than one P in P+Noun combinations?
  - Why do Nouns show up with specific Ps only?
  - Why does *Vorbehalt* behave differently in different contexts?
    - *Vorbehalt:* optional PP[+gegen] complement (reservations against)
    - *unter Vorbehalt:* optional NP[+gen] complement (*PP[+gegen])
    - *mit Vorbehalt:* rarely a PP complement, no NP complement
    - *mit Vorbehalt(en):* three times more plural than singular occurrences
    - *unter Vorbehalt(en):* forty times more singular than plural occurrences

The 'light P' hypothesis

- Exchanging *mit* and *unter* ...
  - If the combination [P Vorbehalt] has a circumstantial meaning, *mit* and *unter* can be exchanged.
  - Die Saarbergwerke hingegen rechnen mit/unter Berufung auf „ernstzunehmende Energieprognosen" mit einem Exportbedarf beim Strom.
- If the phrase has a conditional (intentional) meaning, *mit* and *unter* cannot be exchanged.
  - *Die Arbeitsgemeinschaft Berliner Mieterberatungsgesellschaften hatte bereits Ende Dezember die betroffenen MieterInnen aufgefordert, die Mieterhöhung im Januar nur mit Vorbehalt zu zahlen.
  - Die Arbeitsgemeinschaft Berliner Mieterberatungsgesellschaften ... hatte bereits Ende Dezember die betroffenen MieterInnen aufgefordert, die Mieterhöhung im Januar nur unter Vorbehalt zu zahlen

P+Noun: What does the grammar say?

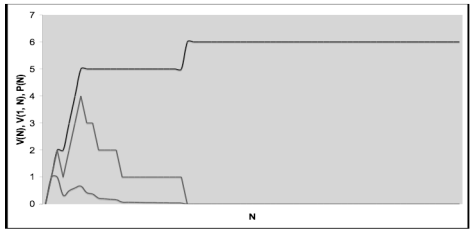
- P+Noun combinations are only rarely dealt with in grammars of German. One big expection is Helbig/Buscha (1998, 403).
  - [Präposition + Substantiv] bilden eine offene Wortklasse, die nicht vollständig aufgelistet werden kann (P+Noun make up an open word class that cannot be listed ...)
  - But, why do Helbig/Buscha (1998) call the combination a word class, a "Zusammensetzung" [combination] or "Wortgruppe" [word group] instead of using the simpler ‚phrase'?
    - [cranberry] Noun occurring in P+Noun can only appear as part of the combination.
    - [semantic weakening] The meaning of the noun is weakened in the combination.
    - [no Det] The noun cannot be used with a determiner in the combination.
    - [P+N complement] The combination of P+Noun is word like in that its complement can be substituted by many other complements.
    - [substitution] The combination can be replaced by a preposition.
    - [orthography] The noun is not capitalized in the combination or is written together with the preposition.

An empirical study

- Goal: to show that P+Noun combinations cannot consist of a finite set of elements, and hence, that P+Noun cannot be listed, despite the intuition problem.
  - restricted to *unter*+Noun
- Method: Apply Baayen's (2001) measures for productivity to syntactic combinations.
  - V(N): The number of vocabulary types in a sample of N tokens.
  - V(1, N): The number of vocabulary types in a sample of N tokens which appear only once (hapax legomena)
  - P(N) = E[V(1, N)]/N: The likelihood that a new type of a certain word class will be detected after N tokens have been sampled.
- Domain: A corpus of four consecutive editions of the Neue Zürcher Zeitung [written high-brow Swiss German], comprising a total of 106 million words; sampling occurred at subsets of 6, 12, 18, 26, 52, 78, and 106 million words.

Measuring non-productivity

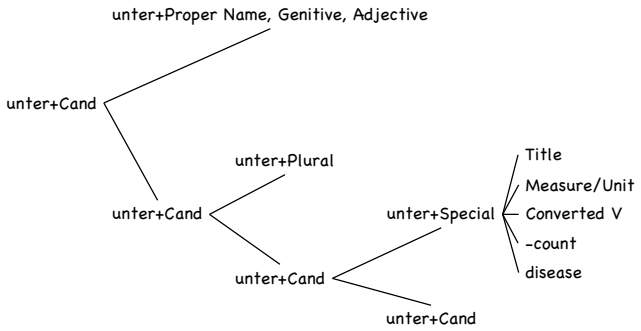
- Assume a language consisting of the six numbers of a six-sided die.
  - How does E[V(N)] develop as N gets larger?
  - How does E[V(1, N)] develop as N gets larger?



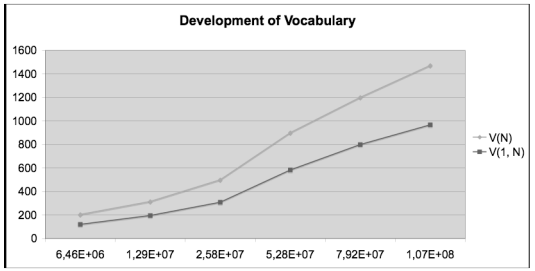
Measuring rule applications

- Instead of measuring productivity – P(N) – as the likelihood of a new word type occurring after N tokens have been sampled, we measure the likelihood of a new $word_2$ occurring after a fixed $word_1$ after N tokens have been sampled.
  - We measure the likelihood that a new instance of a rule of grammar which combines w_1 and w_2 occurs after N tokens have been sampled.
  - P(N) is measured as E[V(1, N)]/N.
  - While there are methods to estimate E[V(1, N)] (Evert 2004), they are problematic when it comes to measuring syntactic units.
  - Currently we set E[V(1, N)] as empirical V(1, N) and measure P(N) accordingly as V(1, N)/N.

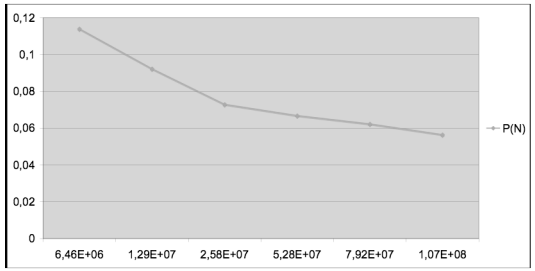
Which elements are considered?



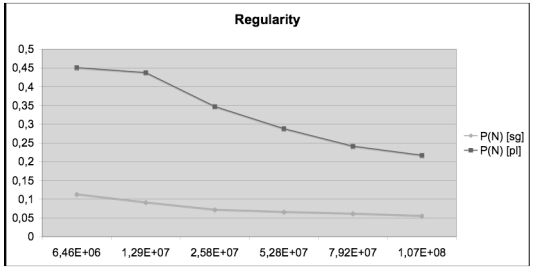
Measuring rule applications



Measuring P(N)



Controlling P(N)



Perspectives

- The combination unter+Noun$_{sg}$ cannot be listed.
- P+N combinations are more often compositional than non-compositional.
  - Claim: P+N combinations are just as often non-compositional as any other combination of X and Y in the language under observation.
- The rule for combining P+N does not seem to be a rule which the speaker has tacit knowledge of ...
  - speakers cannot easily produce new P+N
  - speakers cannot easily judge P+N
- We have to explore the rule type which allows a combination of P+N despite a speaker's unability to creatively produce the combination.
- We need a grammar of P+N combinations which takes the aforementioned considerations into account.