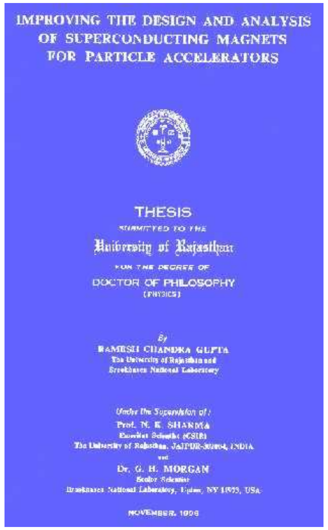
An experiment in hybrid PP attachment

1. Motivation	of the entire enterprise
2. Methods	used and rejected in the journey
3. Problems	with the approach, solved and unsolved
4. Results	in cold hard facts
5. Demo	of the integration in real time

Background

- I do constraint parsing for a living
- This is part of my PhD work
- However, this is not my PhD defense
- The thesis discusses hybrid processing of different kinds
- This talk is only about PP attachment



IMPROVING THE DESIGN AND ANALYSIS
OF SUPERCONDUCTING MAGNETS
FOR PARTICLE ACCELERATORS

THESIS
SUBMITTED TO THE
University of Rajasthan
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
(PHYSICS)

By
RAMESH CHANDRA GUPTA
The University of Rajasthan and
Brookhaven National Laboratory

Under the Supervision of :
Prof. N. K. SHARMA
Emeritus Scientist (CSIR)
The University of Rajasthan, JAIPUR-302004, INDIA
and
Dr. G. H. MORGAN
Senior Scientist
Brookhaven National Laboratory, Upton, NY 11973, USA

NOVEMBER, 1996

Motivation

- PP attachment is difficult (particularly for German)
  - 35 of 56 categories allow PP attachment
  - partially free word order
  - split auxiliary groups
  - separable verbs
  - many different factors contribute
  - as usual, full treatment is probably equivalent to full AI
- But analysing errors shows many cases that *should* be easy to get right.
- Mastering PP attachment alone would bring us 3.5% of accuracy.

More Motivation

Label	no. of errors	occurred	retrieved	percentage
PP	605	1892	1287	68.0
ADV	190	1137	947	83.3
OBJA	100	775	675	87.1
APP	94	660	566	85.8
SUBJ	88	1338	1250	93.4
S	77	1098	1021	93.0
KON	75	481	406	84.4
REL	60	167	107	64.1
CJ	57	481	424	88.1
PN	44	1885	1841	97.7
AUX	42	673	631	93.8
ATTR	32	1222	1190	97.4
KOM	32	88	56	63.6

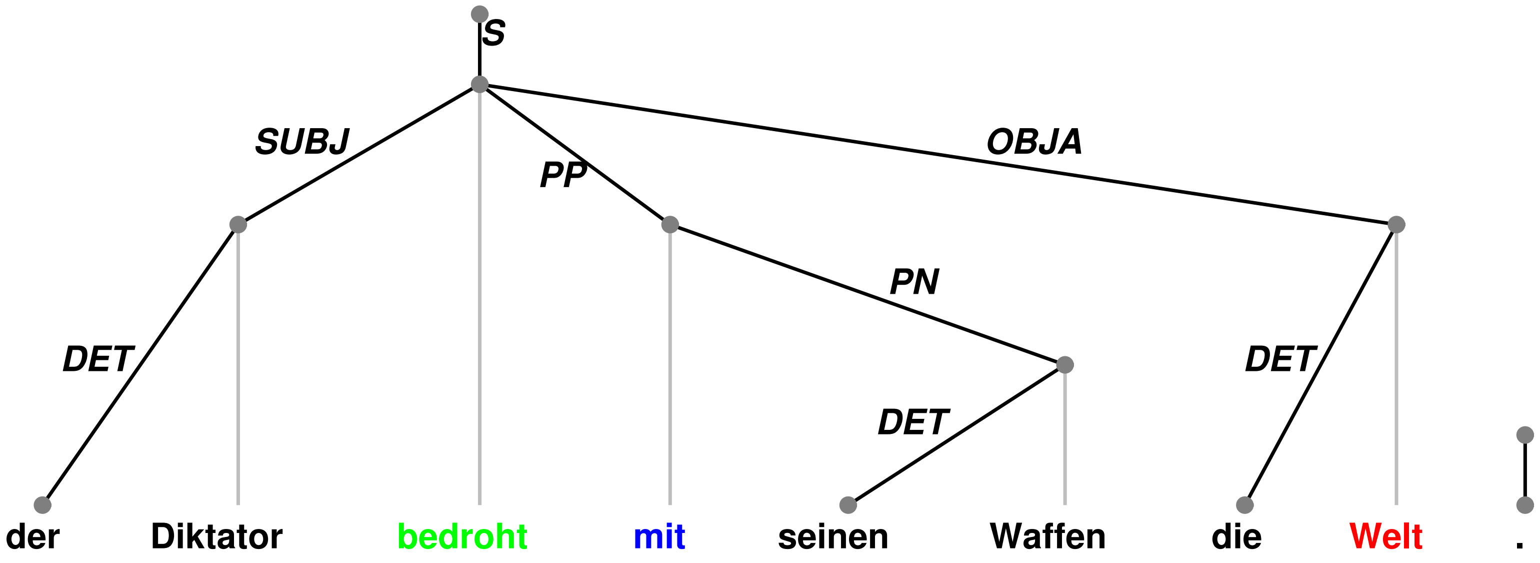
Contributing factors

- Syntax: “Ich habe *für dich* noch eine Überraschung.”
- Topology: “Die Version *für Europa* kostet 30 Euro.”
- Default assumptions: “Ich behalte die Karte *für morgen*.”
- Idioms: “Dabei werde die Regierung *auf Sofortmaßnahmen* zurückgreifen.”
- Collocation: “Man werde einen Betrag *von zwei Millionen Euro* zahlen.”
- Named Entities: “Den Großteil dieser Mittel hatte das Projekt von der Kreditanstalt *für Wiederaufbau* erhalten.”
- Style: “Time Warner hat eingewilligt, die 25,5 Prozent *von TWE von AT&T* zurückzukaufen.”
- Semantic Weakness: “Diese Bewertung kann sich nach Meinung *von Beobachtern* aber noch ändern.”

The Story So Far

- So far, we only use syntax, topology, default, and idiom constraints.
- There will now be a short example session.
- **DISCLAIMER: The constraints you are about to see have been radically simplified for expositional purposes and should not be assumed to be representative of true real-life constraints. No constraint was harmed in the making of this talk.**

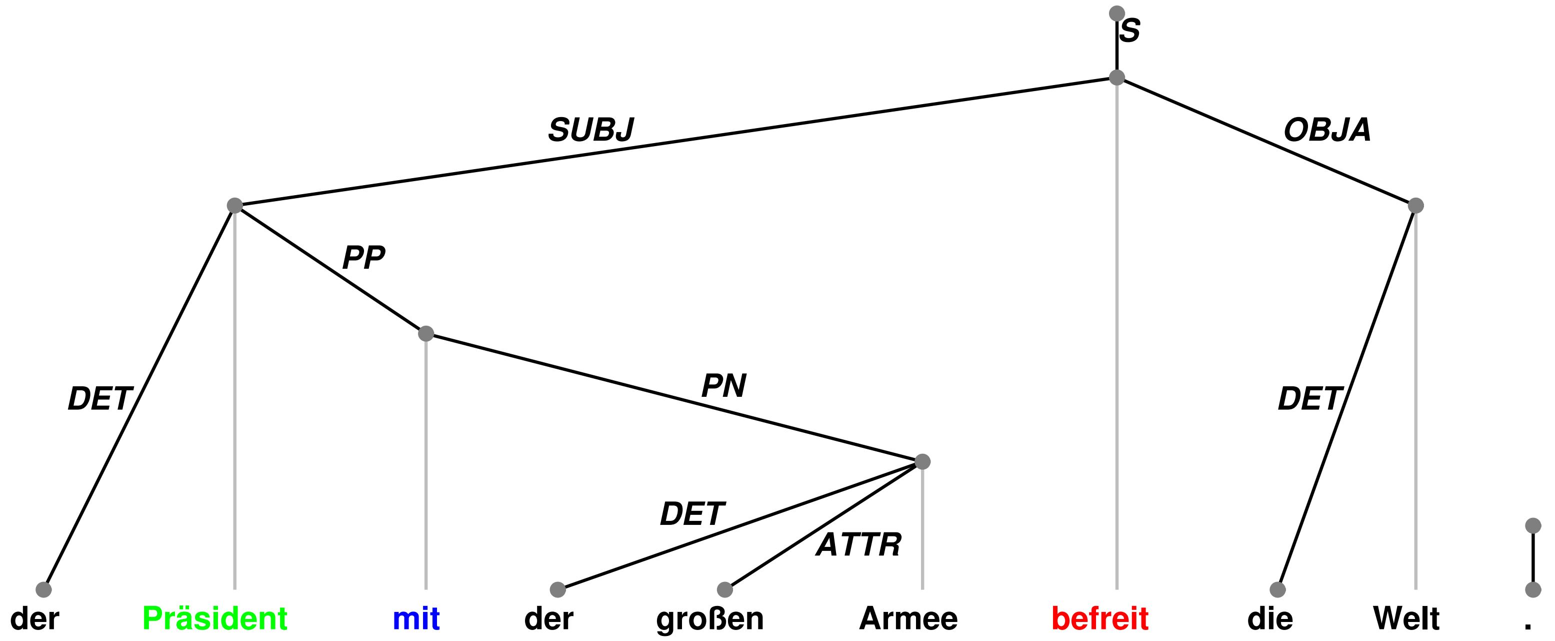
Existing constraints: syntax



```
{X/SYN} : 'Präposition vor Nomen' : order : 0.01 :
  X.label = PP
  ->
  X^cat != NN;
```

(Prepositions under nouns must follow their noun.)

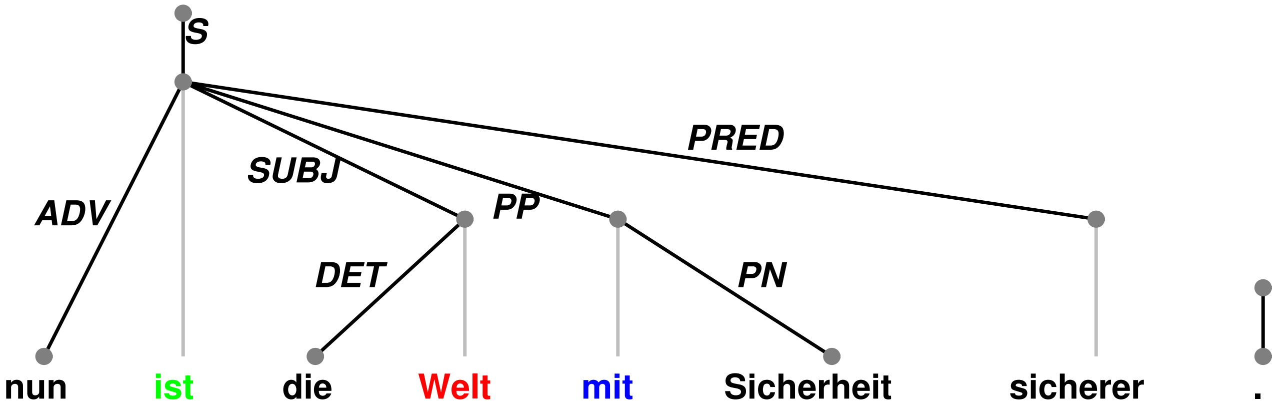
Existing constraints: topology



```
{X/SYN/\Y/SYN} : Vorfeld : top : 0.1 :
isa(X^,finit)
->
~is(X^id,S);
```

(Only one constituent precedes the verb in a main clause.)

Existing constraints: defaults



```
{X!SYN} : Präpositionalattribut : cat : 0.9 :
X.label = PP -> X^cat != NN;
```

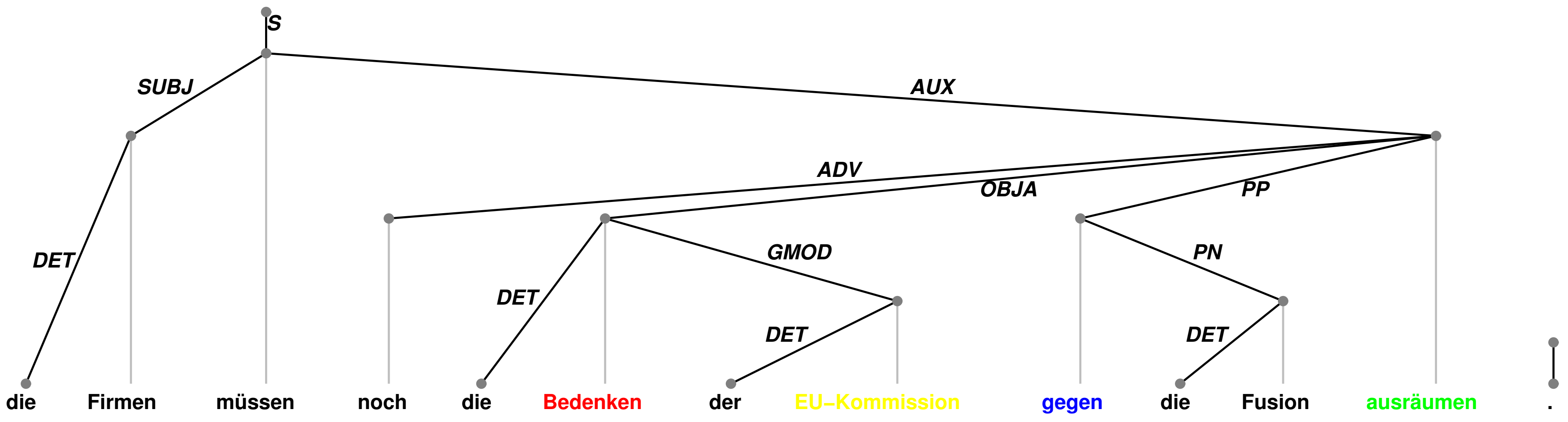
(Verbs slightly disprefer noun attachment.)

```
{X!SYN} : 'mod-Distanz' : dist : gradient(100) :
edge(X,Modifikator)
->
abs(distance(X@id,X^id) = 0;
```

(Verbs *very* slightly prefer close attachment.)

Doubt resolution

Here's a difficult attachment problem:



Die Firmen müssen die Bedenken der EU-Kommission gegen die Fusion ausräumen.

(The companies have yet to address the Commission's concerns about the merger.)

What shall we do with a doubted merger?

- There are three possible attachment points for ‘gegen’: ‘Bedenken’, ‘Kommission’, and ‘ausräumen’.
- Our rules so far predict ‘ausräumen’.
- Intuition says ‘Bedenken’.
- But that is actually the worst-scored option.
- How to formalize intuition?
- The relevant constraint would be trivial to write…
- …it’s the other 5,000,000 that get to you.

Related Work

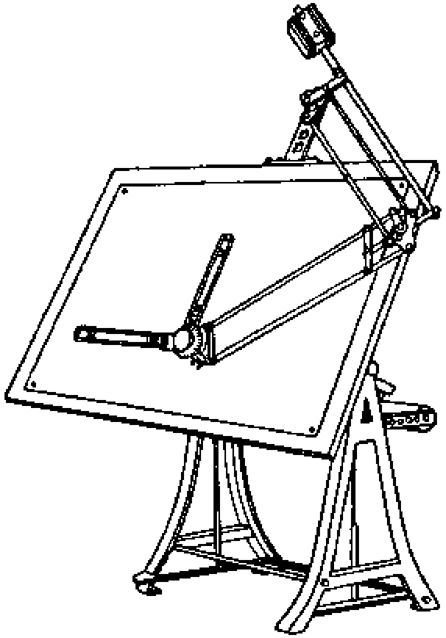
- Many people have tackled the PP attachment problem.
- But usually in a reduced version (the quadruple problem).
- This makes the problem both harder and easier.

  — you don't have the entire context to make the decision.

  + you only have two choices, not n.
- Still, they can usually predict the majority of attachments correctly.

Sounds great. Do it.

- OK, can we extract the collocation of ‘Bedenken’/‘gegen’ from a treebank?
- Alas, no: our 55,913 trees provide not even one instance.
- (And we wouldn’t want to learn anything from one instance, anyway.)
- Back to the drawing board?



Red Alert

- We'll just have to switch to *unsupervised*, then.
-
- This poses a higher risk, but allows more resources.
- We can easily get many millions of raw sentences.
- We just have to bite the bullet and assume that co-occurrence is an approximation of subordination.

Method

- Input: TAZ corpus (18,000,000 sentences of newspaper copy)
- tokenize and assign POS tags automatically
- stem content words via our existing CDG lexicon
- find co-occurrence of content words with prepositions
- find verb/preposition pairs within 10 words
- find adjacent noun/preposition pairs
- Quiz: which word/preposition pair do you think will occur most often?

And the winner is:

- The most common pair is ‘geben’/‘in’ (270460 occurrences).
- Hmmm. Not very illuminating.
- Perhaps we should have normalized against prior probability?
- Adaptation: calculate $f_{w+p}/(f_w \cdot f_p / \sum f_i)$
- What will now be the strongest collocation?

Size does matter

- The strongest collocation from 9 years of TAZ stories is ‘II:Hattig’/‘contra’ (1/1/1132)
- What the $†ϟ%✠??
- Well, *every* occurrence of ‘II:Hattig’ was accompanied by ‘contra’…
- …and you can’t get a much better divisor than 1.
- Perhaps we should ignore such rare combinations altogether.
- What will be the strongest collocation that occurs at least 100 times?

Rare, or medium rare?

- The strongest collocation with more than 100 appearances is…
- …‘peu’/‘à’ (177/536/4977)
- Aahhhh, that sounds a lot better.
- In fact, only two among the first fifty items appear doubtful.
- Spot the mis-extracted attachments!

Reading between the lines

Score	pair	*f*
1171.56	peu/à	177/536/4977
111.46	verbergen/hinter	3992/8205/77074
93.02	zurückbleiben/hinter	1268/3123/77074
87.37	verschanzen/hinter	450/1180/77074
84.01	zurückstehen/hinter	216/589/77074
63.73	identifizieren/anhand	104/5889/4893
61.94	Milligramm/pro	279/947/83981
61.46	zurückfallen/hinter	463/1726/77074
51.45	verstecken/hinter	2170/9663/77074
50.28	auskommen/ohne	3229/5488/206645
44.11	herziehen/hinter	119/618/77074
43.72	pendeln/zwischen	941/1554/244584
42.71	verantworten/wegen	2632/6658/163447
40.95	befördern/jenseits	140/5029/12004
40.58	stecken/hinter	4654/26274/77074
39.34	Diskrepanz/zwischen	515/945/244584
38.39	nachweisen/anhand	106/9963/4893
38.06	subsumieren/unter	208/245/393885
38.03	zerreiben/zwischen	197/374/244584
36.96	anklagen/wegen	2339/6837/163447
36.90	verbüßen/wegen	303/887/163447
36.30	verurteilen/wegen	11233/33432/163447
35.12	geistern/durch	873/937/468478
35.00	lauern/hinter	356/2330/77074
34.66	Kluft/zwischen	1207/2514/244584
33.95	Cent/pro	400/2477/83981
33.09	hindurchgehen/durch	151/172/468478
32.51	verlaufen/entlang	147/9997/7987
32.22	überprüfen/anhand	105/11759/4893
32.00	ankämpfen/gegen	718/700/566060
31.85	einstellen/mangels	148/26377/3111
31.82	verstoßen/gegen	8070/7910/566060
31.23	belangen/wegen	255/882/163447
31.19	Pfennig/pro	1380/9303/83981
29.59	schwanken/zwischen	1017/2481/244584
29.04	rangieren/hinter	225/1775/77074
28.06	auflehnen/gegen	340/378/566060
28.04	tingeln/durch	520/699/468478
27.98	streunen/durch	121/163/468478
27.17	Misstrauen/gegenüber	163/1354/78238
27.16	anrennen/gegen	202/232/566060
27.00	rühren/her	461/4642/64942
26.99	ermitteln/wegen	3957/15841/163447
26.70	kosten/inklusive	294/37095/5242
26.06	verklagen/wegen	677/2807/163447
25.81	polemisieren/gegen	547/661/566060
25.57	Skepsis/gegenüber	428/3778/78238
25.55	protestieren/gegen	15204/18563/566060
25.53	Spagat/zwischen	495/1400/244584
25.48	scharen/hinter	116/1043/77074

Preliminary Evaluation

- The adjusted formula seems to produce credible results.
- Of the first 100 items, only 4 seem unintuitive:
  'befördern'/'jenseits', 'rühren'/'her', 'geschlossen'/'hinter', 'Direkt'/'neben'
- These appear to be mis-taggings rather than mis-collocations.
- What about our original problem?

Rank	score	pair	*f*
1372	4.96	Bedenken/gegen	1529/9618/566060
4216	2.47	ausräumen/gegen	185/2336/566060
51422	0.13	Kommission/gegen	223/52415/566060

- Wow! Our intuition was confirmed by the facts!

All is not well

- We have found an instance where statistical information works well.
- However, there many reasons why things can go wrong.
- The following is just a subset of problems with the method I am proposing:
  - sparse data
  - extraction errors
  - lexical ambiguity
  - independence assumptions
- Solving one problem often creates new problems.

Problem: Sparse data

- There is no such thing as ‘enough data’.
- Particularly not in German:
  - nouns are constantly formed anew
  - words can vary in their forms because of inflection
- Therefore, we will often have to attach unknown words.
- Solution: *backing off*
  - Map ‘sagen’, ‘sagte’, ‘gesagt’ etc. to ‘sagen’
  - Map ‘Bundesverteidigungsministerium’ to ‘Verteidigungsministerium’
  - . . . or even to ‘Ministerium’ or ‘NN’
- This assumes that the compound noun behaves like the base noun.

Problem: Sparse data

- But often the attraction radiates from the extra noun instead:

  “Man muß einen *Versorgungs*grad von 25% der Bevölkerung
  mit UMTS erreichen.”

  “Sie mußten Unsummen für die *Eintritts*karte *in den UMTS-Markt* zahlen.”

- or from an attribute:

  “Nun werden auch *wiederbeschreibbare* DVDs *mittels blauer Laser* verkauft.”

  “Der Rücktritt wird spätestens auf der *letzten* Sitzung *vor den Ferien* erwartet.”

- or even a morpheme:

  “Ellison wurde *um 300 Millionen* reich*er*.”

Problem: Sparse data

- Some word classes are intrinsically rare: NE, CARD
- We will never be able to learn the lexical attraction for ‘436’…
- …and probably not for ‘Richard’, either.
- Solution: Back off to ‘CARD’ and ‘NE’ immediately
- Problem: This is not always the right thing to do.
  - *Gates am Montag
  - Frankfurt am Main
  - 2000 bis 2002
  - *2002 bis 2000

Problem: Extraction errors

- Base form mapping is a good idea because it counters data sparseness, but it introduces new problems.
- Why do these pairs emerge from the algorithm:
  - raten/unter
  - zeihen/in
- Problem: base form computation is ambiguous.
- Idea: always assume the more common infinitive
- POS tagging, spelling, etc. are likewise uncertain.
- The noise problem is by definition unsolvable.

Problem: Extraction errors

- Lexical attraction may be falsely suggested by correlation between more than two items.
- Why does Sprecher/gegenüber score higher than Sprecher/von?
- Because it is indirectly correlated:
  "Das sagte ein Sprecher gegenüber der FAZ."
- As a result, similar sentences will likely be mispredicted.
- Idea: detect high-scored correlations that also have indirect correlations, and treat them with suspicion.

Problem: Lexical ambiguity

- Verbs can have different senses in different contexts.
- Meaning obviously affects attraction behaviour:

  "Da kostete der Ritter von dem Met."

  "Jetzt kostet ein Anteilsschein von Freeserve wieder unter 400 Pence."

  "Wir vertreiben Produkte vom Anwendungsserver bis zur Festplattenschraube."

  "Wir vertreiben Ratten von der Wiese in den Kanal."
- Pipe dream: This could only be handled with sense-tagged corpora.

Problem: Lexical ambiguity

- German verbs have another trick up their sleeve: *prefixing*
- Prefixed verbs are essentially different verbs, even if the prefix is syntactically distant.
- Sometimes they behave identically: fahren/nach, weiterfahren/nach
- Sometimes they don't: reagieren/auf, abreagieren/an
- Solution: pair each PTKVZ with the closest VVFIN, and add it to the infinitive.
- This must be done both during extraction and prediction.

Problem: Lexical ambiguity

- Many prepositions have two variants differing only in case.
- But case cannot be accurately measured from raw text!
- The difference is important for lexical attraction:
  "Man konnte die Führung *auf drei Punkte ausbauen.*"
  "*Das Bild auf der Titelseite* schockiert die Welt."
  "Man wolle die Position auf dem Markt für Mobiltelefonie ausbauen."
- Even with the same case, prepositions can have multiple senses:
  "Melden Sie dem General den Weg über den Paß!"
  "Melden Sie dem General den Weg über das Internet!"
  "Melden Sie dem General den Weg über sein Handy!"

Problem: Lexical ambiguity

- Auxiliary verbs are transparent with respect to PP attachment…
- …but they sometimes function as full verbs and then behave very differently.
- Example: “Ich bin gegen Atomkriege” vs. “Ich bin gegen die Wand gerannt”
- (Actually this is not much of a problem, since we normalize PP attachment to the full verb anyway, but it still disturbs the penalty calculation.)
- Solution: detect pseudo-full auxiliary verbs by the absence of a full verb…
- …and then treat them like full verbs.

Problem: Lexical ambiguity

- ‘von’ is a special case for attachment:
- it is both a preposition and a passive morpheme
- Therefore, it reacts to category rather than lexeme.
- Verbs that occur mainly in the passive would get far too high ‘von’ counts: beschuldigen/von
- Solution: ignore ‘von’ with passives when extracting…
- …and allow it with all passives when attaching.
- (Actually, it is also a genitive morpheme…)

Problem: Independence assumptions

- The assumption that only base form matters is wrong.
- ‘außer’ is strongly attracted to ‘geraten’…
- …but otherwise, more to nouns than verbs.
- “Danach soll der Minister außerdem nach Warschau fahren.”
- Hmmm. ‘außerdem’ behaves differently from ‘außer’.
- Idea: we could give up normalizing of prepositions.
- But the same case could be made for any word category really.

Problem: Independence assumptions

- The assumption that the kernel noun is irrelevant is wrong.
- Both erwarten/von and Nutzen/von are strong collocations.
- “Man erwarte einen Nutzen von 100 Millionen Euro.”
- But obviously not with every kernel.
- Idea: learn triples instead of pairs
- Meta-Problem: Please buy me 100,000,000,000 words of German…

Problem: Independence assumptions

- Some PP kernels behave like adverbials.
- The strongest collocation in the following sentence is: Pauschalpreis/ohne
- “Die ISPs könnten solch einen Pauschalpreis ohne weiteres finanzieren.”
- But apparently, ‘ohne weiteres’ is not your average ‘ohne’;
  it attaches as freely as an adverb
- Idea: detect common kernels and count which ones defy the normal distribution
- These PPs might be metaphorical and should then be exempted from attraction penalties.
- Contrariwise, some PP/kernel pairs are strongly attracted to certain verbs:
  ‘zur Anwendung kommen’, ‘unter Verdacht geraten’

Problem: Independence assumptions

- Two PP in the same sentence are not independent from each other.
- “Wir müssen im Zeitraum von Januar bis März nach Berlin fahren.”
- $\Longrightarrow$ ‘von’ and ‘bis’ are positively correlated.
- “Unser Reporter spricht mit dem Regisseur von mehr als 30 Filmen.”
- $\Longrightarrow$ ‘mit’ and ‘von’ are negatively correlated for ‘sprechen’,
  although both are good collocations on their own.
- Spot the independence assumption:
  “Jetzt hat sich auch der letzte Vertreter der Spielebranche *von* einem Auftritt auf der CeBIT verabschiedet.”
- Solution: Live with it.

Putting the data to work

- Enough of these gloomy thoughts!
- We need a formula to translate the information to CDG penalties.
- The extracted preferences should be reflected by the penalties.
- But the penalties should not be too strong altogether.
- What about this formula:

$$p = \max(1, \min(1 - (2 - \log_3(s))/50, 0.8))$$

- For instance, this will map 5 → 0.989 and 0.5 → 0.95.
- Finally we normalize all alternative scores so that the highest one is 1.[*]

[*] Basic CDG writing technique, read ‘Writing and Using CDG’ when it is finished.

Results

- And now, the moment we have been waiting for:

Method	uses which PP data?	Syntactical accuracy
baseline	none	90.7%
supervised	treebank only	91.9%
unsupervised	taz data only	91.9%
added	both tables added	92.0%
successive	use taz as backoff only	91.9%

- Yes, part of the theoretical benefit can be reaped!
- PP attachment itself rises from 68% to 79%.
- Note that the supervised/unsupervised figures are actually the same (although the results are not the same!).
- This is probably because the supervised data were mostly in-corpus, while the unsupervised data were from a different corpus.

Real-Time demonstration

- The main lesson here is that the predominance of common cases makes up for the multitude of special cases.
- This is, after all, what statistical methods are about.
- I shall now perform some annotation work with no safety net whatsoever:
  - Watch as the intrepid annotator reviews automatically parsed newsticker sentences.
  - The PP module was not available when these trees were computed, but it is now active.
  - Will we manage to observe dependencies snap into the right place at a single mouse click? Stay tuned…

Future Work

- Statistical methods can make significant improvement to PP attachment.
- Although PPs are probably the subordination type that needs it most, similar reasoning could be applied to others.
- Take coordination:

  "Nach Meinung vieler Experten steht hinter diesen Geboten weniger die Sorge um nicht ausreichende Netzkapazitäten, sondern viel mehr der Wille, den Einstieg von Newcomern in das deutsche Mobilfunkgeschäft zu verhindern."

  "Sorge" is a much better attachment for "Wille" than "Netzkapazitäten", simply because the two are closely related semantically.
- Exploiting this will definitely require a semantic model of German.
- But hey, don't we own GERMANET?

Future Work

- Apposition is very similar to coordination: “Der deutsche Geschäftsführer von PSINet, Helmut Blank, sagte...”

  “Geschäftsführer” is a much better attachment for the name “Helmut” than “PSINet” because it denotes a human.

- This is arguably even easier; we need only to mark nouns that are subtypes of persons (Präsident, Mittelstürmer, Ausländer)

- Object/subject disambiguation relies heavily on semantics:
  - “Die Mutter sieht die Tochter.” (no support)
  - “Konkursgerüchte drücken Kurs der Amazon-Aktie” (supports normal syntax)
  - “Dies Vorhaben will die Regierung noch dieses Jahr verabschieden.” (supports inverted syntax)