Starting a sentence in Dutch

Gerlof Bouma

29 November 2013

Alfred Jodocus Kwak

1) **Ik** ben vandaag zo vrolijk.
   I am today so happy
   'I am so happy today'

2) **Vandaag** ben ik zo vrolijk.
   Today am I so happy

3) **Zo vrolijk** was ik nooit.
   So happy was I never
   'I was never this happy before.'

Word order in Dutch

- Verb second + verb last: X **V** Y Z **Vs**
- Impoverished case system
- Some word order variation
- Fronting:
  S **V** O Vs
  O **V** S Vs
  X **V** S O Vs
  ...

Nominal argument fronting in Dutch

4) **We** vieren op 5 december Sinterklaas.
we celebrate on 5 december Saint Nicolas
'We celebrate Saint Nicolas' eve on 5 December.'

5) **Sinterklaas** vieren we op 11 november
St Nic celebrate we on 11 November

Starting a sentence in Dutch

Alternation means that there are different ways of saying essentially the same thing (expressing the same relation between participants in the sentence).

Central Q: Why does a speaker choose one variant over another, wrt filling the directly preverbal position?

Type of A: general tendencies that prefer one alternate over another. These tendencies a) may conflict, b) may differ in strength c) may be of varying nature d) are ideally observed in some other language, too.

Method: formulate expectations of tendencies in terms of frequency and try to find evidence for them in a corpus.

Corpus methods

Corpus of Spoken Dutch: collected late 90s early 2000s, annotated early 2000s. 10mln words spoken data transcribed and POS-tagged, 1mln annotated with syntactic structure. About 70k main sentences.

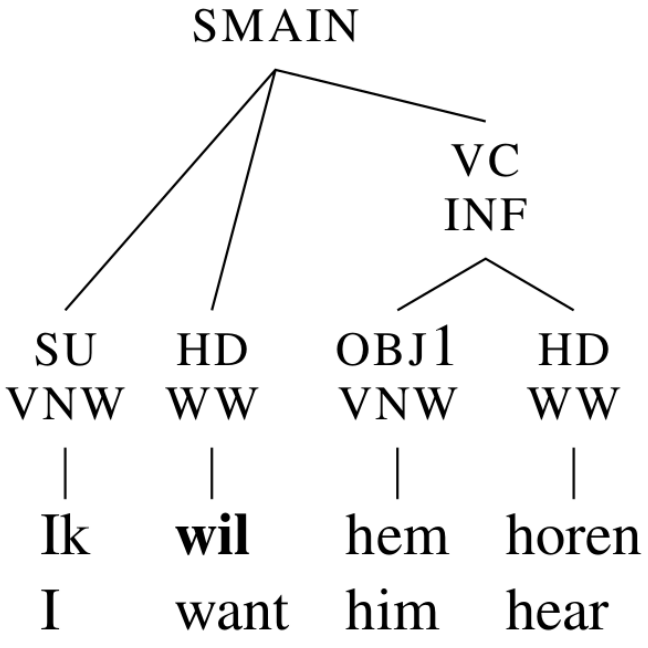
Syntactic annotations are graphs with relation labels and phrase labels.

Define linguistic high level concepts in terms of the actual CGN annotation.

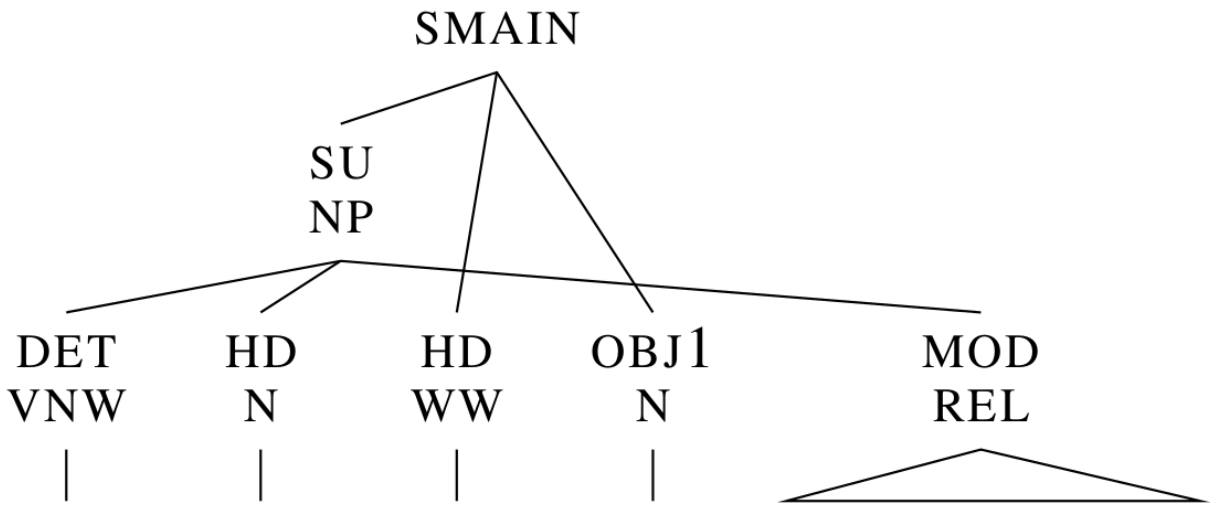
Select data.

Then: count!

CGN example tree



CGN example tree



‘Every girl that comes there is in danger.’

Statistical methods

Tables & graphs:
good to get a feel for overal relation between variables, e.g. % of fronted constituents vs grammatical function of the constituents investigated.

Regression modelling:
allows us to combine many explanatory variables (e.g., grammatical function, definiteness, length) and their effect on the predicted variable (% fronting) in the presence of each other.

Expectations

Expectations on the basis of results from English, German and postverbal word order alternations in Dutch:

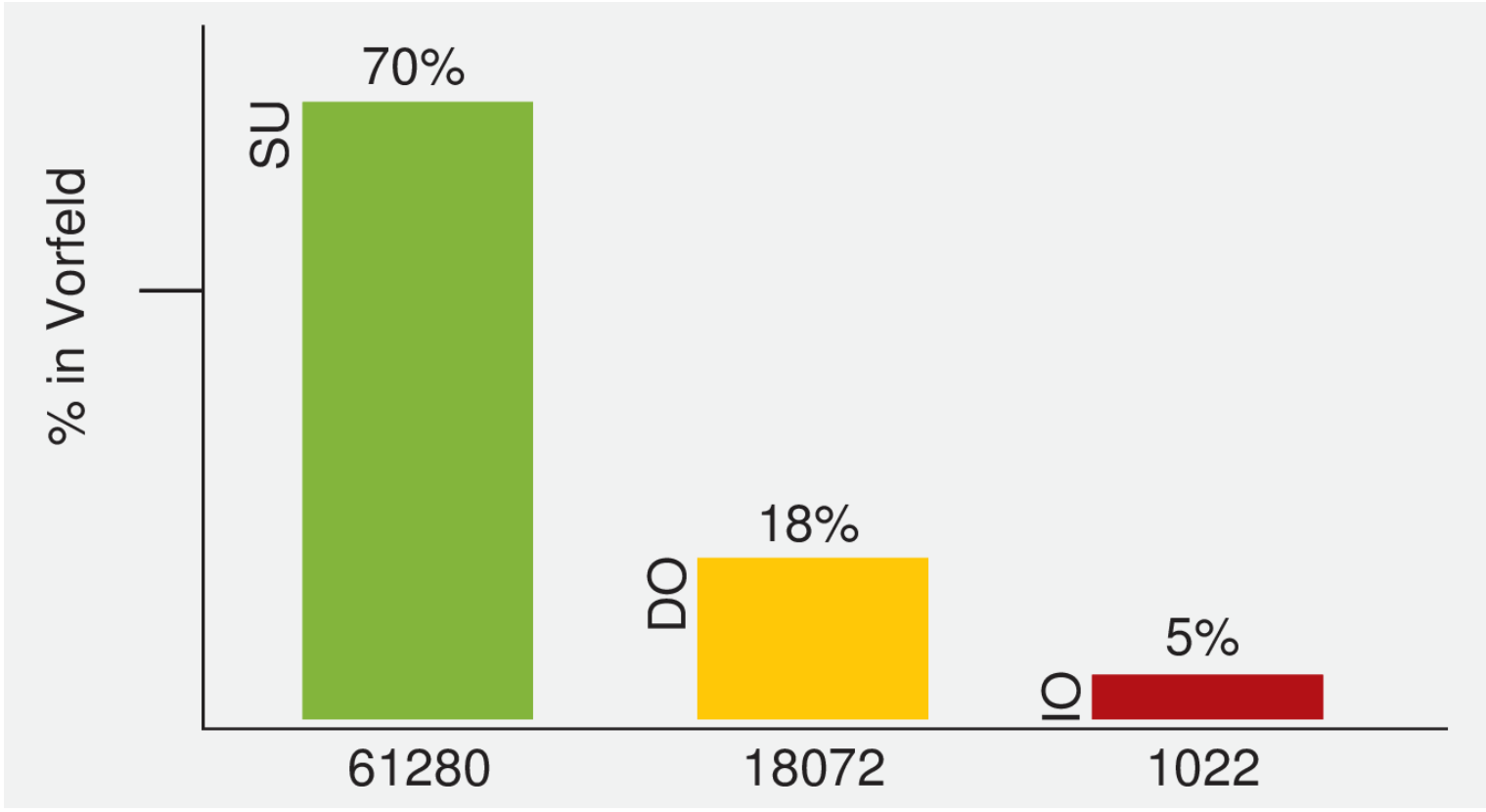
- Grammatical function: subjects front more easily than indirect objects which front more easily than direct objects
- Definiteness (NP form): pronouns $\prec$ definite NPs $\prec$ indefinite NPs
- Length/Complexity: short constituents $\prec$ long constituents (one of Behaghel's Laws)

On the basis of literature and of own findings (2nd and 3rd part of the talk):

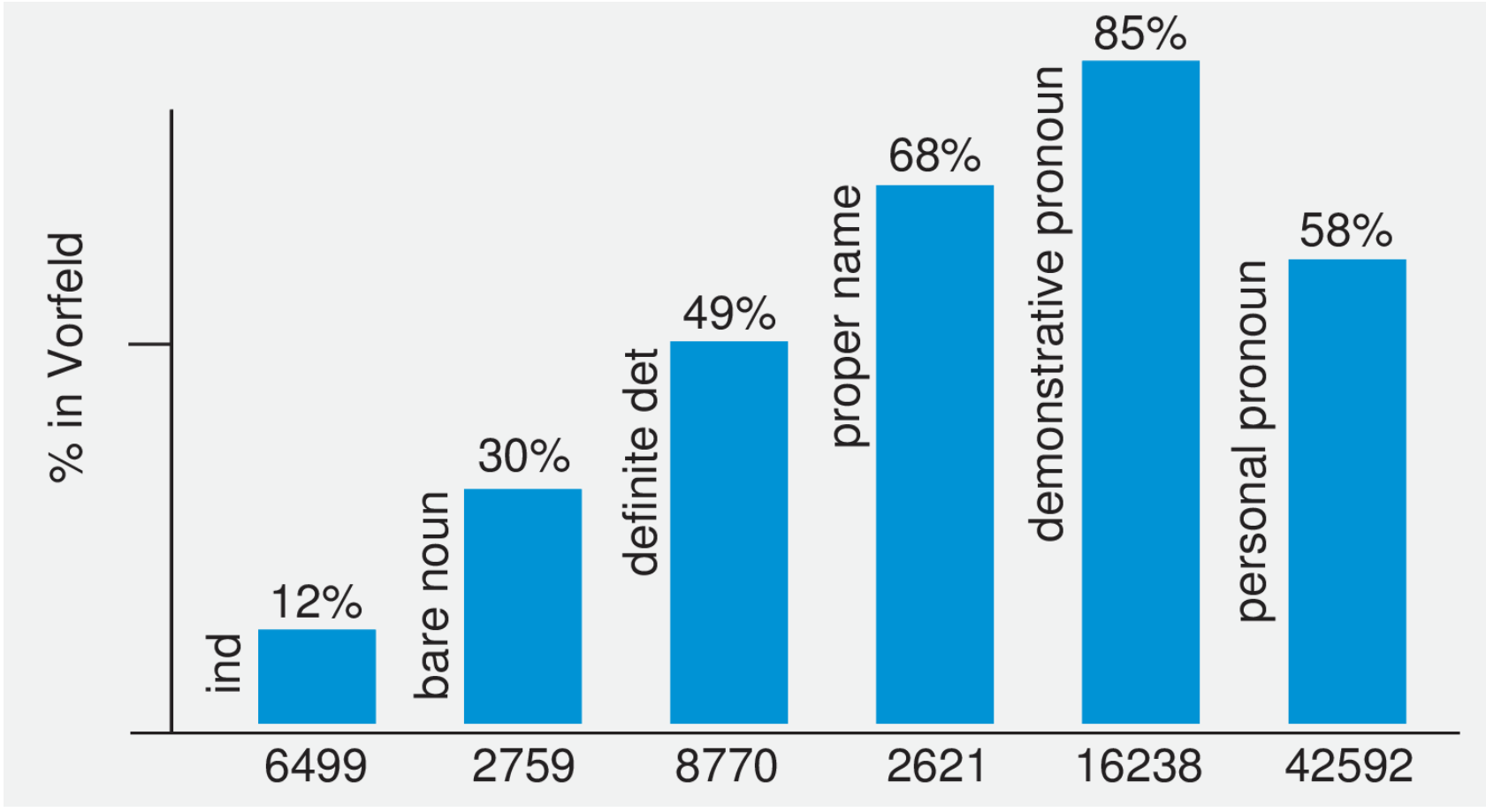
- Object does not precede subject when it hinders understanding
- 'Informative' constituents are good Vorfeld citizens

Part I
(Bouma 2008)

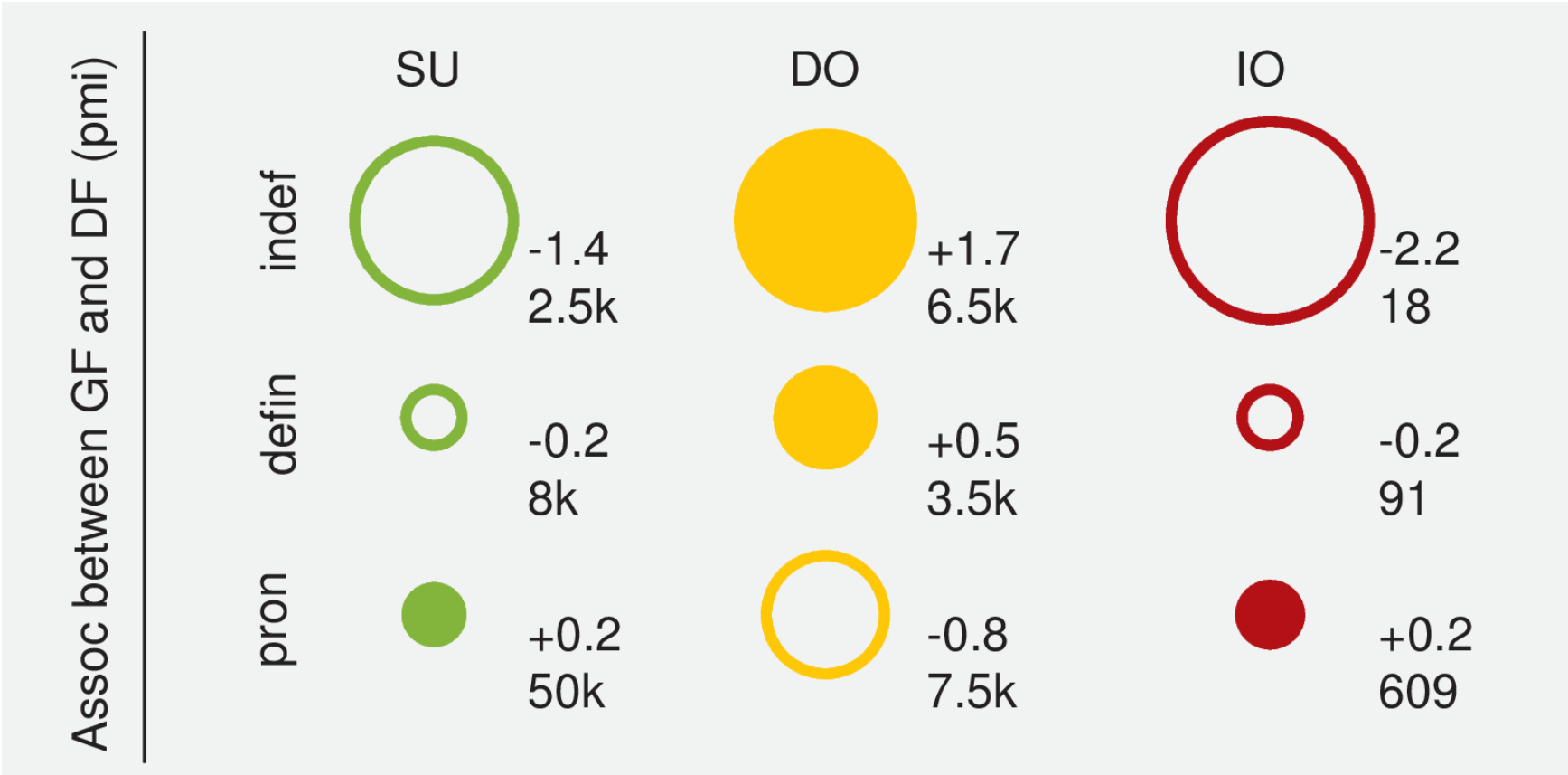
Overall grammatical function effect



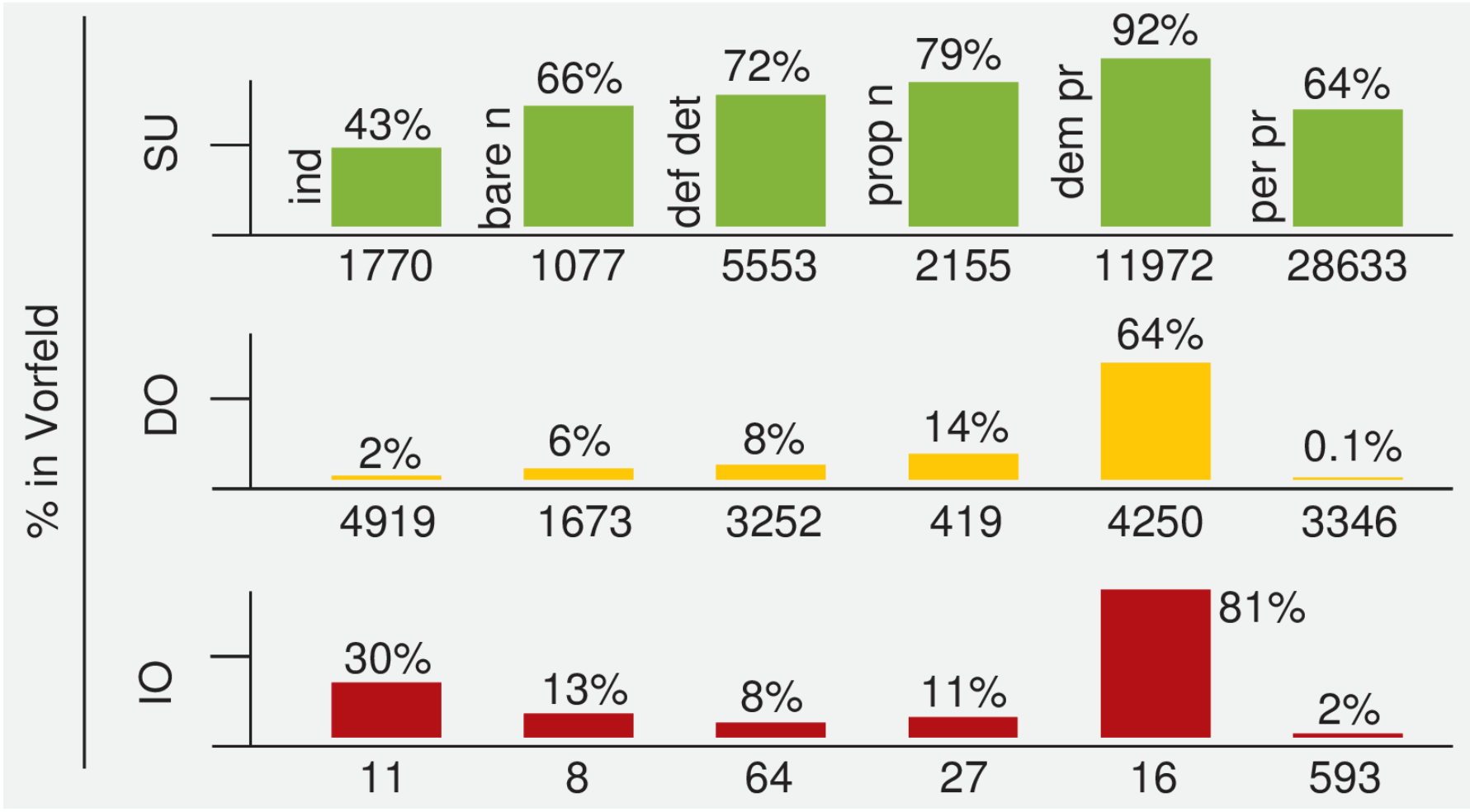
Overall definiteness effect



Association definiteness and grammatical function



Definiteness effect per function



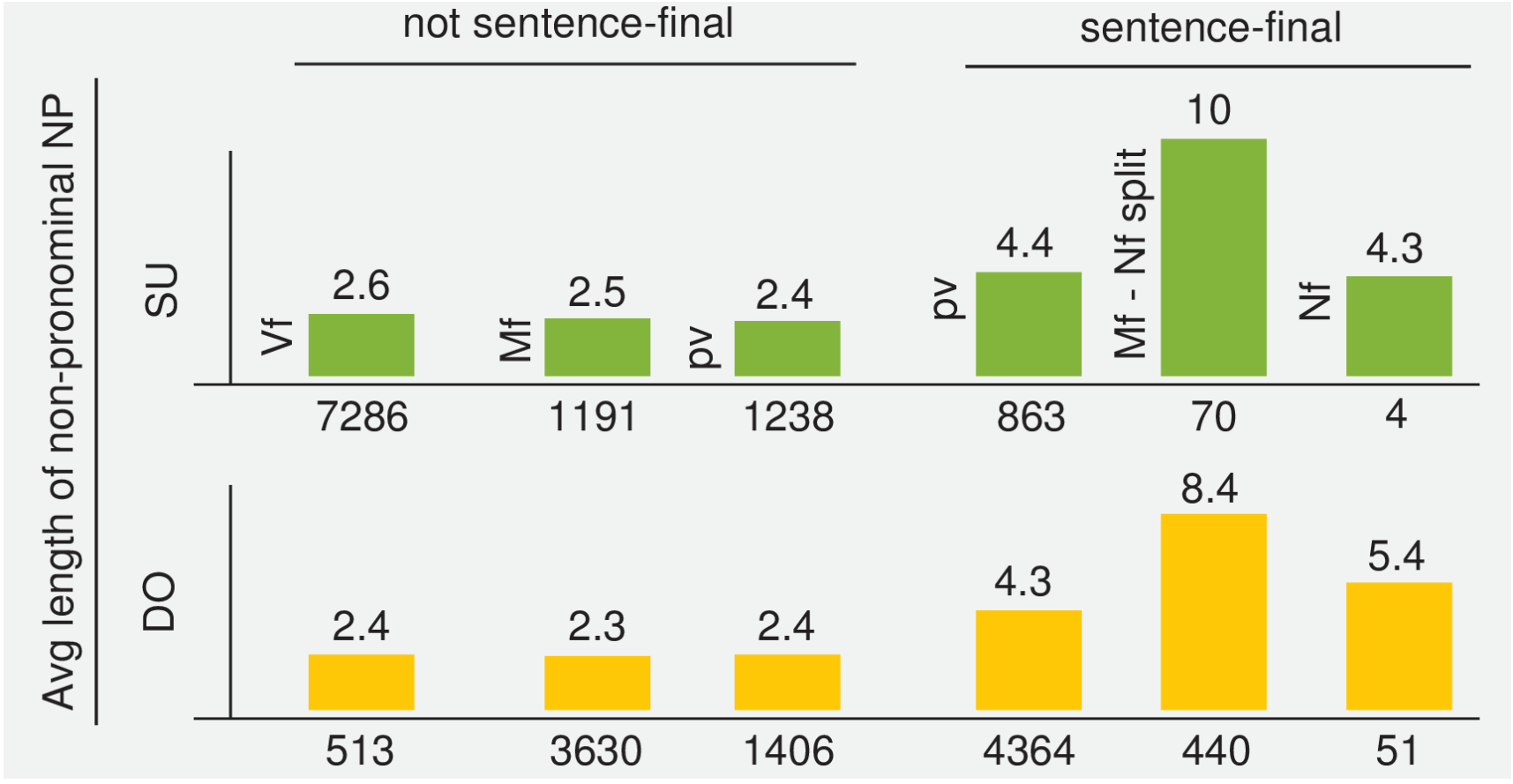
Overall length effects

Table 4.14: Summary of constituent length in words, combined data.

	Vorfeld				¬Vorfeld				
Category	#	Avg	$Q_{1,3}$	Mx	#	Avg	$Q_{1,3}$	Mx	p
nominal	46 191	**1.28**	1–1	47	33 829	**1.94**	1–2	83	<.001
n ∧ ¬pron	7 811	**2.62**	1–3		13 405	**3.37**	2–4		<.001
n ∧ ¬p ∧ ≤10	7 680	**2.42**	1–3	10	12 858	**2.83**	2–3	10	<.001
verbal	95	**2.58**	1–4	15	129	**7.37**	4–9	40	<.001
clausal	284	**6.12**	4–7	27	4 483	**9.09**	5–11	71	<.001

Note: # raw counts, *Avg* mean of lengths, $Q_{1,3}$ first and third quartile, *Mx* maximum length (minimum length is 1, except in the clausal categories, where it is 2), *p* result of 2-tailed Wilcoxon rank sum test on length of constituents in the Vorfeld vs length of constituents elsewhere.

Length per position



Interim summary

- subject $\prec$ indirect/direct objects
  no clear diff between object
- pronominal $\prec$ definite full $\prec$ indefinite full
  NB! not the personal pronouns
- short $\prec$ long
  really a fact about a different position, not a sentence wide trend or a Vorfeld fact

Part II

(Bouma 2008, 2011; Bouma & Hendriks 2012)

Word order freezing

Fronting alternations: different word orders for the same relational meaning.

Means that word order is not (always) a reliable clue for grammatical function: the object may precede the subject.

How do we know which is which?

Word order freezing: examples

(5) Mat' ljubit doč'
mother.NOM/ACC loves daughter.NOM/ACC
'Mother loves her daughter.' *Not:* 'The daughter loves her mother.'

(18) a. Patt^{h}ar t^{h}elaa todegaa.
stone.NOM cart.NOM break.FUT
'The stone will break the cart.' *Not:* 'The cart will break the stone.'

b. Jishin-ga Taroo-ga kowa-i
earthquake.NOM Taroo.NOM is afraid of
'Earthquakes are afraid of Taroo.' (SOV)
Not: 'Taroo is afraid of earthquakes.' (OSV)

Word order freezing in Dutch

(12) Fitz zoekt Ella op.
Fitz looks Ella up
'Fitz looks up Ella' (SVO),
not (or strongly dispreferred) 'Ella looks up Fitz' (OVS)

(13) a. Welk meisje zoekt Frank op?
which girl looks Frank up
'Which girl looks up Frank' (SVO) or 'Which girl does Frank look up' (OVS)

b. FITZ zoekt Ella op.
Fitz looks Ella up
'FITZ looks up Ella' (SVO) or 'Ella looks up FITZ' (OVS)

c. Het nummer zoekt Ella op.
The number looks Ella up
'Ella looks up the number' (OVS)
and maybe 'The number looks up Ella' (SVO)

Word order freezing in Dutch

Some evidence for freezing in Dutch, but not just related to case. Other ways of recognizing subject and object can help 'thaw' a sentence: animacy, definiteness, intonation. . .

In general: word order is taken as a good grammatical function clue when other information is lacking.

Q: do speakers actually care about this? Do they freeze sentences when other interpretation hints to the hearer are not available?

How could we investigate this in a corpus?

Object fronting per subject x object definiteness

Table 2 Object fronting by definiteness of subject and direct object in transitive clauses

	Object Definiteness			
Subject Definiteness	Indefinite full NP	Definite full NP	Pronoun	Total
Indefinite full NP	162	88	114	363
OVS (%)	*2 (1.2)*	*1 (1.1)*	*37 (32.4)*	*40 (11.0)*
Definite full NP	644	477	373	1514
OVS (%)	*13 (2.0)*	*9 (1.9)*	*113 (30.3)*	*135 (8.9)*
Pronoun	5421	2875	5972	14268
OVS (%)	*171 (3.2)*	*300 (10.4)*	*2541 (42.5)*	*3012 (21.1)*
Total	6247	3440	6459	16146
OVS (%)	*186 (3.0)*	*310 (9.0)*	*2691 (41.7)*	*3187 (19.7)*

Relative definiteness overall

Table 3 Counts and proportions of object fronting, per relative definiteness level

	Superiority	Equality	Inferiority
All word orders	8940	6611	575
OVS (%)	*484 (5.4)*	*2552 (38.6)*	*151 (26.2)*

Don't forget tendencies we learnt before

Een man heeft Ella opgezocht
a man has Ella looked up
'A man has looked Ella up',
also???? 'Ella has looked up a man.'

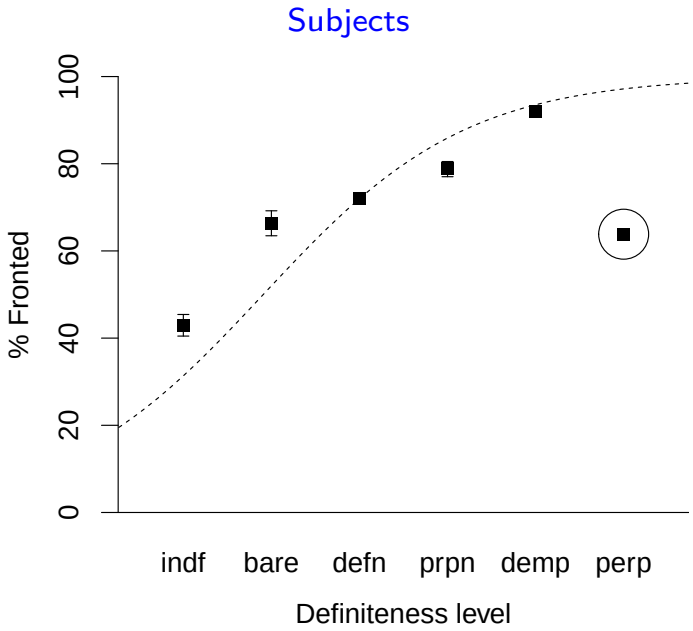
Parameter	Estimate
Intercept	−4.729
Subject complexity	0.083
Object complexity	**−0.721**
Subject form	
Indefinite full NP (base)	
Bare nominal	−0.722
Definite full NP	−0.220
Proper name	−0.450
Demonstrative pronoun	**−2.498**
Personal pronoun	−0.295
Object form	
Indefinite full NP (base)	
Bare nominal	**1.063**
Definite full NP	**1.742**
Proper name	**1.949**
Demonstrative pronoun	**5.459**
Personal pronoun	**−2.228**
Relative definiteness	
Superiority	**1.214**
Equality (base)	
Inferiority	−0.437

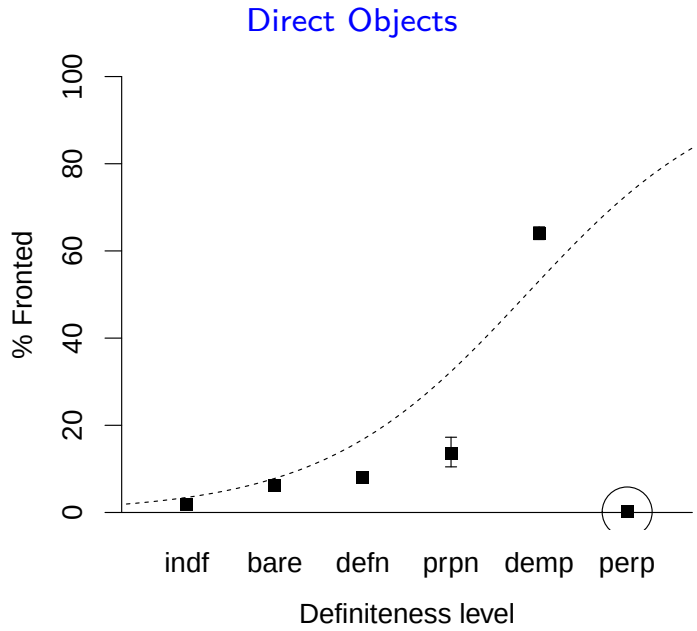
Interim summary

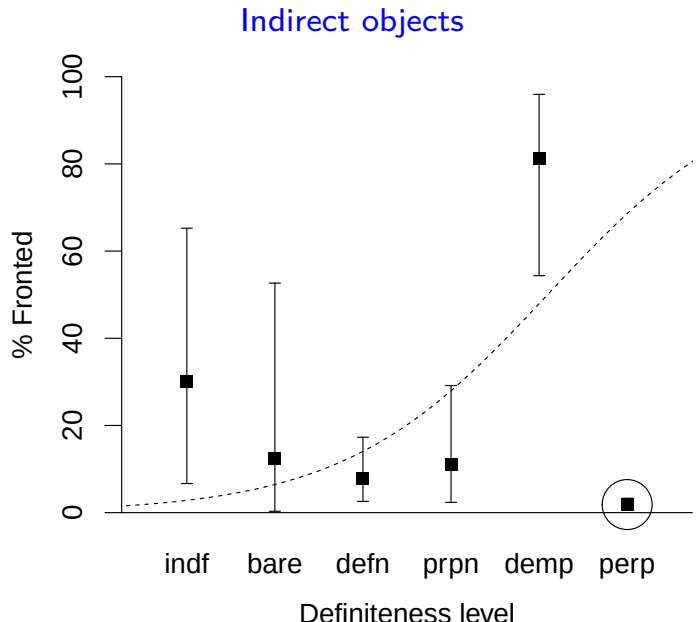
- Yes, speakers care about freezing! Object-before-subject word order used more when relative definiteness can act as a clue to interpretation.
- "Communicative success" influences choice for alternate.
- Partial word order freezing: a trend rather than a on-off effect.
- Answers to linguistic questions that are almost impossible to answer by intuition, because of all the correlations.

Part III
(ongoing)

A puzzle in the definiteness effect







Demonstratives in Dutch

6) **Dat** vieren **we** op 5 december.
that celebrate we on 5 December
'We celebrate that (=St Nic's eve) on 5 December.'

Demonstrative pronouns in Dutch (*dat, die*): frequent in discourse, may pick up salient, human referents, topic shift device (Van Kampen 2010).

Something old, something new

Gundel (1988):

- *Given-before-new principle*
  State what is given before what is new in relation to it.2
- *First-things-first principle*
  Provide the most important information first.

(also Mithun 1987, Givon 1988, a.o.)

Something old, something new

A good candidate for fronting is one that is definite and informative.

A demonstrative pronoun is a supercandidate: highly definite (pronoun) but informative (topic-shift/signalling function)

Does this carry over to non-pronominal NPs? How do we determine informativity?

Proposal:
informativity as unexpectedness/untypicality of a constituent wrt its clausal context, calculated from lexical statistics.

Informativity: Surprisal

$$\text{Surprisal}(\textit{verb}, \textit{arg}) = -\log p(\textit{arg}|\textit{verb})$$

Verb-argument Surprisal: the surprisal of the subject/object given the verb. Statistics from automatically parsed X00MIn words written Dutch (Van Noord 2010).

Between 0 (fully predicted, no info) and inf (completely unpredicted, max info)

Data overview

Subset of data from Part I, subject and object data where the subject/object is non-pronominal.

- Direct objects: 10k observations (5% fronted)
- Intransitive subjects: 7k5 observations (65% fronted)
- Transitive subjects: 2k5 observations (75% fronted)

Data overview

Informative:

[$_{obj}$ overleg] bedoel je
'Deliberation, you mean?'

we zoeken graag [$_{obj}$ alternatieve reisvormen]
'We like to find alternative ways to travel'

nu wil ik [$_{obj}$ chips]
'now I'd like some crisps.'

[$_{obj}$sokken] hebben we nog niet opgeschreven
'Socks, we haven't written down yet.'

Data overview

Uninformative:

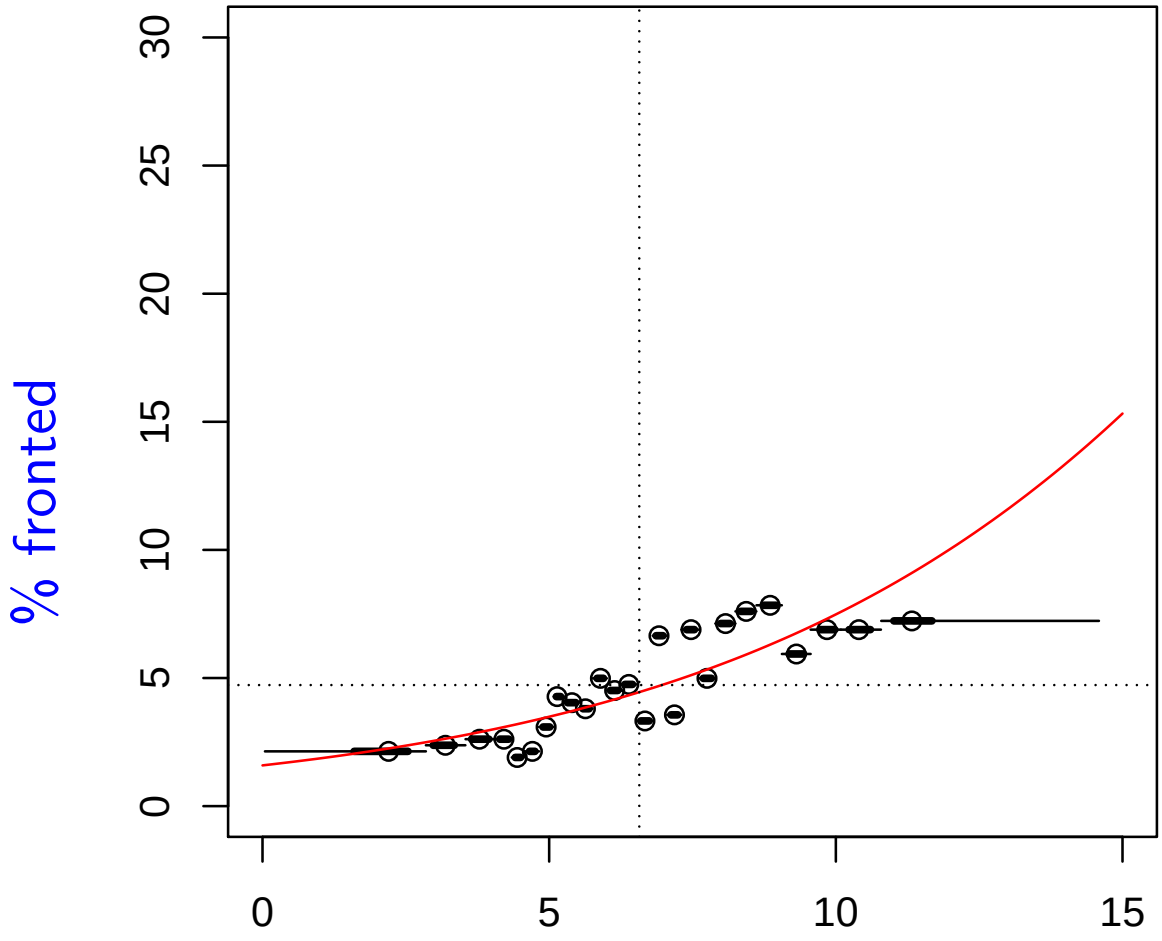
Nederland viert vandaag [$_{obj}$ Koninginnedag de verjaardag van prinses Juliana]
'Today, the Dutch celebrate queensday, Juliana's birthday'

Cocu speelt [$_{obj}$ de bal] naar Bergkamp
'Cocu plays the ball to Bergkamp'

de ME voerde [$_{obj}$ charges] uit
'The riot police performed charges'

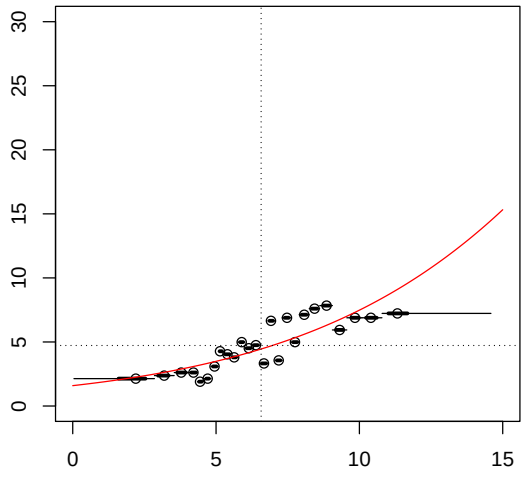
lemmingen plegen [$_{obj}$ geen massaal zelfmoord]
'Lemmings do not commit suicide *en masse*'

Correlations in non-pronominal direct object data

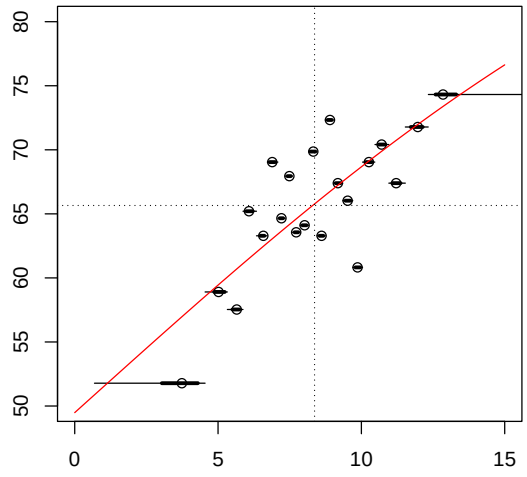


Correlations in all three datasets

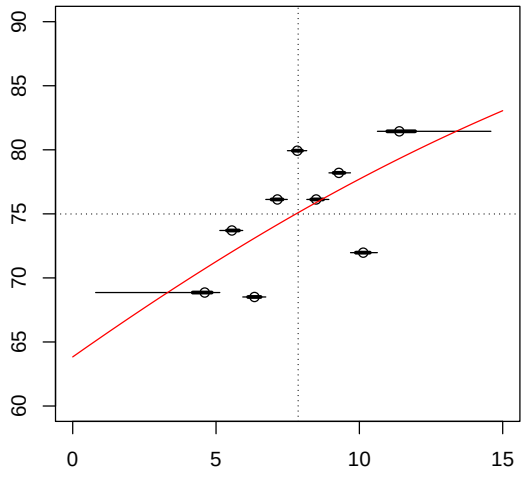
Direct objects



Intransitive Subjects



Transitive subjects



Modelling expectations

Something old, something new:

Regression modeling should show a possitive effect of informativity on fronting in each data set (first-things-first) and maintain the definiteness effects found before (given-before-new)

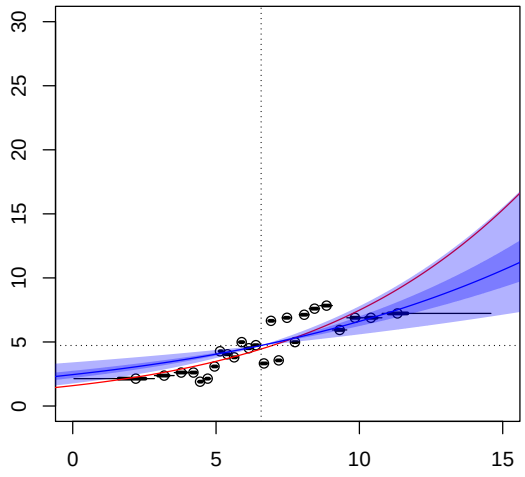
Modelling

Fixed effects:

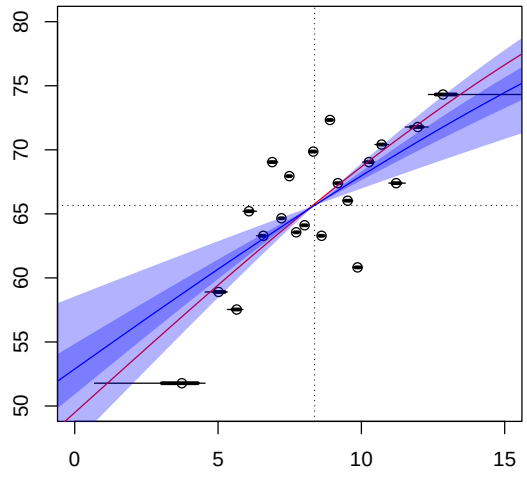
- Definiteness (indefinite, bare NP, definite, proper name)
- NP subclass (universal pro, existential pro, demonstrative det)
- Constituent length
- Sentence length
- Definiteness of other argument (not intrans subj)
- **Informativity**

Modelling results for informativity effect

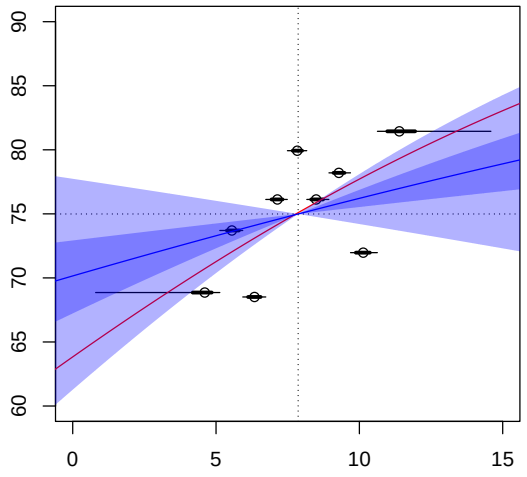
Direct objects



Intransitive Subjects



Transitive subjects



Modelling

Direct objects:

- Informativity effect
- Known definiteness effects reconfirmed.

Intransitive subjects:

- Informativity effect
- Known definiteness effects reconfirmed.

Transitive subjects:

- No clear informativity effects found
- No reconfirmation of known definiteness effects

Summary part III

Support for the SO SN hypothesis of fronting in Spoken Dutch in intransitive subject and direct object data: effects of both definiteness and informativity.

We can use auxiliary statistics from automatically annotated data in our corpus experiments.