Computational Linguistics

Data-intensive Linguistics

Gosse Bouma

Information Science
University of Groningen

LOT Winterschool 2009

Using Corpora for Linguistic Research

Finding positive examples

linguistic intuitions of grammaticality are deeply d and seriously underes- timate the space of grammatical possibility (Bresnan et al)

Obtaining frequency information

- Psycholinguistics
  Many recent models of language comprehension have stressed the role of distributional frequencies in determining the ease of processing with a particular lexical item or sentence structure. (Roland et al)
- (Stochastic) Optimality Theory
- Computational Linguistics

Finding examples

Extraposition from Fronted Material

- Is allowed in general
- but not from comparatives (reviewer of van der Beek et al, 2001)

(1) **De vraag** is gerechtvaardigd **waarom de regering niets doet**
The question is justified why the government does not act

(2) * **Lager** was de koers nog nooit **dan bij opening**
The rates were never lower than at the opening

Corpus provides counter-examples to this claim

(3) **Nog eerder** zal de Mekong droogvallen **dan dat de premier zijn macht uit handen geeft**
It is more likely that the Mekong falls dry than that the prime-minister gives up his power (Volkskrant 1997)

Focus Particles inside PPs

(4) *Peter träumt [von nur seiner Frau] (German)
Peter dreams of only his wife

(5) *Peter droomt [van alleen zijn vrouw] (Dutch)
Peter dreams of only his wife

Highly Debated

- No agreement about data in literature (Rooth, Jacobs, Bayer, Buring and Hartman)
- General picture: Focus particles occur within PPs in English, not in German (and Dutch)

Focus Particles inside PPs

(6) ouderen [met alleen een AOW-uitkering]
elderly with only an AOW-allowance

(7) een druk programma [met ook doordeweekse wedstrijden]
A busy programme with also weekday games

(8) gevolgen variëren van depressies [tot zelfs suïcide]
consequences range from depressions to even suicide

Corpus provides many counterexamples

In Dutch, there is considerable variation as regards the preferences for Adv-P-X order versus P-Adv-X order, some having to do with pragmatic/lexical semantic factors and some with syntactic factors (possibility of relative clauses, no external particles in extraposition) (Bouma, Hendriks, and Hoeksema, 2005)

Obstacles to using Corpus Data

Corpus is not representative

Manually annotated corpora are carefully compiled but small

Automatically Annotated Corpora contain errors

- Large corpora can be annotated automatically with Part of Speech, root forms, dependency labels
- Accuracy ranges from 90% (syntax) to 97% (POS).
- Coverage of lexicon (valency information) and syntax may be limited (coordination, ellipsis, clefts, ...)

Annotation is missing

Thematic roles, word senses, focus placement, given-new distinction, coreference relations, logical form, ...

Today

Using Automatically Annotated Corpora in Linguistics

- Discuss number of studies in theoretical linguistics and psycholinguistics that make use of corpus data
- All papers make use of automatically syntactically annotated corpora (treebanks)

1. **Roland et al**: How to obtain frequency figures for syntactic constructions?
2. **Bastiaanse et al**: Should aphasiac performance be attributed to syntactic complexity or frequency?
3. **Bresnan et al**: What accounts for the dative shift?
4. **Bouma and Spenader**: Does subategorization frequency play a role in using *zichzelf* instead of *zich*?

Frequency and Language Processing

Many recent models of language comprehension have stressed the role of distributional frequencies in determining the ease of processing with a particular lexical item or sentence structure. However, there exist little relatively few comprehensive analyses of structural frequencies....

[Roland et al. (2007), Frequency of basic English grammatical structures: A corpus analysis, J of Memory and Language]

Frequency of Cleft Sentences

- Subject Cleft:
  - It was Nixon's first visit to China that set in motion...
- Object Cleft:
  - It's paper profits I am losing

Interpreting Aphasia Results

Aphasic performance of subject clefts is superior to processing of object clefts. Is this due to syntax (loss of capability to handle *traces*) or frequency?

Frequency of Cleft Sentences

	Wall Street Jnl	Switchboard
Subject Cleft	32	38
Object Cleft	2	0

Counts normalized per 1M words

	Wall Street Jnl	Switchboard
Subject Cleft	813	577
Object Cleft	61	0

Counts normalized per 1M sentences

- Are Subject Clefts more frequent in written than in spoken language?
- Sentence length differs between WSJ (written) and Switchboard (spoken)

Frequency of Cleft Sentences

Explanation of poor aphasia performance on Object Clefts

- Overall frequency of clefts is low (less than 1 in 1000 sentences)
- Subject clefts far more frequent than object clefts
- It is likely that Object Clefts are harder to process to begin with
- Hypothesis that processing difficulty of Object Clefts is due to inability to process with *traces* needs more evidence

Subcategorization Frequencies

1. The workers accepted salary cuts....

Subcategorization Frequencies

1. The workers accepted salary cuts....
2. The workers accepted salary cuts because of the credit crunch

Subcategorization Frequencies

1. The workers accepted salary cuts....
2. The workers accepted salary cuts because of the credit crunch
3. The workers accepted salary cuts would be necessary

Subcategorization Frequencies

1. The workers accepted salary cuts....
2. The workers accepted salary cuts because of the credit crunch
3. The workers accepted salary cuts would be necessary

Processing Issues

- Hearing *The workers accepted salary cuts....* (where continuation is unknown) is ambiguous: either a direct object or the start of a sentential complement
- Is processing difficulty influenced by frequency of **accept NP** vs **accept S**?

Subcategorization Frequencies

1. The workers accepted salary cuts were necessary
2. The workers accepted **that** salary cuts were necessary

Processing Issues

- Introduction of *that*-complementizer removes (local) ambiguity
- Does frequency of *V that S* increase if *V NP* is relatively frequent?

Methodology

Answering questions like this requires (large) syntactically annotated corpora

- Collect (per verb) frequency of various subcategorization patterns

Subcategorization Frequencies

Relative frequency of subcat frames

	BNC	BNC-Spoken	Brown	Switchbrd	WSJ
intransitive	11	14	18	32	11
transitive	30	31	32	25	29
passive	9	3	11	2	9
that S	3	3	3	2	4
bare S	4	9	1	6	7

Subcategorization Frequencies

Relative frequency of subcat frames

	BNC	BNC-Spoken	Brown	Switchbrd	WSJ
intransitive	11	14	18	32	11
transitive	30	31	32	25	29
passive	9	3	11	2	9
that S	3	3	3	2	4
bare S	4	9	1	6	7

- Frequency of subcat frames far from constant across corpora

Subcategorization Frequencies

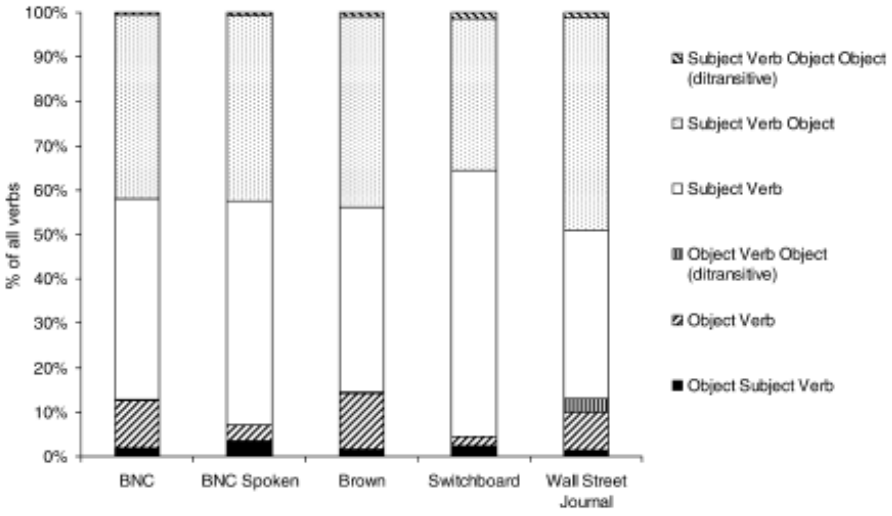


Fig. 8. Distribution of word orders across all structures in each corpus.

That-omission

Top 4 complementizer-less verbs in various corpora

Corpus	Verb	%Omission	% (that) S
BNC	say	69	13
	think	86	11
	know	66	5
	mean	66	4
BNC-Spoken	think	90	22
	say	81	15
	mean	94	11
	know	83	8
Brown	say	59	13
	think	86	9
	know	50	7
	suppose	76	2

That-omission

- High percentage of *that*-omission does not correlate (it seems) with high percentage of S-complements in general
- Authors suggest difference might be due to difference in meaning between **think that S** and **think S** (epistemic).
  - The government thinks that budget cuts are necessary
  - I think it is going to rain
- Other work by Roland et al: length, **(subcat) frequency**, semantic and lemma info can correctly predict 78% of presence/absence of *that* in sentential complements.

Om-omission in Dutch

(9) Het ministerie weigerde de gegevens te verstrekken
The ministry refused to deliver the data

(10) Staalbedrijven blijven weigeren om capaciteit in te leveren
Steelcompanies continue to refuse to reduce capacity

(11) Hij weigert alle medewerking
He refuses all cooperation

- Counts from CLEF-corpus (approx 80M words, newspaper)

subcat frame	count	%
weiger NP	1203	18
weiger om te	293	4
weiger te	5181	78

Verb Position in Dutch

(12) de jongen die een boek **leest**
the boy who reads a book

(13) de jongen wil een boek **lezen**
the boy wants to read a book

(14) de jongen heeft een boek **gelezen**
the boy has read a book

(15) de jongen **leest**$_i$ een boek *i* (V-2)
the boy reads a book

Verb Position in Dutch

Processing Dutch Sentences (Bastiaanse, Bouma, and Post)

Agrammatic aphasia subjects have more difficulty processing Verb-Second sentences than Verb-final Sentences

- Is this due to frequency or linguistic complexity (V2 is a derived word order)?
- Frequency counts of Verb-Second and Verb-final in Dutch
  - Which Corpus? (spoken vs written)
  - Which verbs (*grain size*)?: only lexical (or also auxiliaries), only finite (or also infinitives, participles), transitive (or also intransitive)?

Verb Position in Dutch

CGN (Spoken)

Comparison	LEX	FIN	OBJ	V-Final	V-Second
lexical trans	+	-	+	52.1	47.9
finite verbs	-	+	-	20.4	79.6
finite lexical	+	+	-	19.3	80.7
finite lexical trans	+	+	+	21.5	78.5

Algemeen Dagblad (Written)

Comparison	LEX	FIN	OBJ	V-Final	V-Second
lexical trans	+	-	+	59.7	40.3
finite verbs	-	+	-	25.9	74.1
finite lexical	+	+	-	15.3	74.7
finite lexical trans	+	+	+	27.7	72.3

Verb Position in Dutch

Interpreting Results

- Verb-second is far more frequent with finite verbs than Verb-final, in spoken and written language
- Verb-second is almost as frequent as V-final in spoken language
- (Verb-second was more frequent than V-final for verbs used in the aphasia experiments)
- **Conclusion**: It is unlikely that processing difficulty of Verb-second sentences is due to frequency

Causative Alternation

Alternation	*He melted 12 tons of lead* → *12 Tons of lead melted*
Observation	Patients with aphasia have difficulty interpreting sentences where a causative V is used intransitively
Hypothesis A	Patients have problems with Causative Alternation
Hypothesis B	Patients have problems with infrequent uses of V
Question	What is the frequency of the (in)transitive use for various verbs?

Frequency of Causative Alternation Verbs

- Requires parsed corpus
  - Subcategorization-frame used must be identified
- Ignore verbs which allow both Object Drop and Causative alternation
  - Hij kookt de aardappelen (*He cooks the potatoes*)
  - De aardappelen koken (*The potatoes are cooking*)
  - Hij kookt regelmatig (*He cooks regularly*)
- Various non-finite intransitive patterns are ambiguous
  - Het ijs is gesmolten
    - *The ice is/has melted* (passive/perfect)
  - Hij laat de suiker smelten
    - *He has someone melt the sugar*
    - *He lets the sugar melt*

Causative Alternation in TwNC (500M words)

Verb		Trans	%	Intrans	%
verkleinen	*to diminish*	1.067	93	81	7
vergroten	*to increase*	3.692	93	273	7
oplossen	*to solve*	3.878	81	884	19
verminderen	*to decrease*	8.442	69	3.844	31
verbeteren	*to improve*	2.852	64	1.613	36
breken	*to break*	6.246	61	4.044	39
opwarmen	*to heat up*	215	60	142	40
verbranden	*burn*	660	57	506	43
smelten	*to melt*	381	34	734	66
stabiliseren	*to stabilize*	71	30	177	70
ontdooien	*to defrost*	66	29	163	71
veranderen	*to change*	4.219	27	11.411	73
afkoelen	*to cool down*	96	19	402	81
verslechteren	*to deteriorate*	422	14	2.688	86
verdrink	*to drown*	171	11	1.373	89

Dative Shift (Bresnan et al)

(16) Susan gave toys to the children

(17) Susan gave the children toys

What governs dative shift?

- Difference in Meaning?
  - change of state: NP NP
  - change of place: NP to NP
- Various Variables
  - discourse accessibility, length, animacy, definiteness, pronominality)

Dative Shift and Meaning

Theoretical Literature

Idioms and 'verbs of imparting of force' suggest restrictions on meaning correspond with restrictions on dative shift

(18) That movie gave me the creeps

(19) * That movie gave the creeps to me

(20) I pushed the box to John

(21) * I pushed John the box

Dative Shift and Meaning

Searching the Web

The web provides natural examples of patterns claimed to be impossible

(22) Orson Welles used to give the creeps to countless child listeners

(23) This story will give the creeps to people who hate spiders

(24) As player A pushed him the chips, all hell broke loose

(25) He pulled himself a piece of pie

- Note that longer arguments tend to be placed at the end

Dative Shift and Meaning

Conclusions from Bresnan et al

- Linguistic intuitions of ungrammaticality are a poor guide to the space of grammatical possibility
- Usage data reveals generalizations we are sometimes blind to

Predicting Dative Shift from multiple variables

Statistical Model

- Predict 1 (NP PP) or 0 (NP NP)
- Given variables
  - semantic class
  - recepient pronominal?
  - theme pronominal?
  - recepient given?
  - ...
- Each example sentence from the corpus provides values for the variables and an outcome (1 or 0).
- Assign a weight to each variable using logistic regression and maximum likelihood estimation, which maximizes the number of cases where the model predicts the correct outcome.

Predicting Dative Shift from multiple variables

- Data from Switchboard corpus
- NP NP = 0, NP PP = 1
- Baseline (always predict 0) = 79

		Predicted		% Correct
		0	1	
Observed	0	1796	63	97
	1	115	386	77
			Overall:	92

Predicting Dative Shift from multiple variables

Are all variables necessary?

- Variables predicting NP PP (1) outcome:
  - verb type = (future) transfer of possession (give, owe, promise)
  - recepient non-given, non-pronoun, indefinite, inanimate
- Variables predicting NP NP (0) outcome:
  - verb type = communication (tell), prevention deny
  - theme non-given, non-pronoun, indefinite, non-concrete

Is the model OK?

- Model generalizes to unseen data, other corpora (WSJ), across speakers, taking lexical bias (verb) into account

Conclusions

We have found that linguistic data are more probabilistic than has been widely rec- ognized in theoretical linguistics. We have examined a body of ecologically valid data-spontaneous language use in natural settings-using statistical techniques for 28 analyzing multiple variables. And we have constructed a model that can predict the choice of dative structures with 94% accuracy, and can resolve persistent questions about usage data.
(Bresnan et al.)

Reflexives preceding the Subject

Which verbs allow reflexive before the subject?

- In Dutch, subject normally precedes the object (also if this is a reflexive pronoun).
- Sometimes, reflexive pronoun precedes the subject
- Which verbs do allow this word order?
  - Inherent Reflexives (i.e. occur only with reflexive object)
  - Other restrictions?

(26) Het was reeds bekend dat *een deel van hen* **zich** in Jeruzalem bevond .
It was known already that some of them were located (SELF) in Jeruzalem

(27) In het grijze gebouw bevindt **zich** *het Rijksarchief*
In the grey building, the National Archive is located (SELF)

(28) Bij deze beslissing legt **zich** *Ajax* neer
Ajax accepts (SELF) this decision

Zich-Subj vs Subj-Zich

82.4 (563)	17.6 (120)	ontspin#refl
70.5 (117)	29.5 (49)	wreek#sbar_subj_refl_no_het
59.4 (1559)	40.6 (1064)	dien_aan#part_refl(aan)
52.9 (925)	47.1 (822)	vorm#refl
49.1 (368)	50.9 (381)	ontvouw#refl
47.4 (1130)	52.6 (1252)	teken_af#part_refl(af)
43.5 (54)	56.5 (70)	teken_af#part_refl_ld_pp(af)
37.9 (36)	62.1 (59)	formeer#refl
36.3 (8479)	63.7 (14909)	bevind#refl_ld_pp
36.2 (21)	63.8 (37)	strek#refl
33.2 (269)	66.8 (541)	verzamel#refl
32.7 (738)	67.3 (1516)	bevind#refl_ld_adv
32.2 (39)	67.8 (82)	sluit_aan#part_refl(aan)
31.0 (303)	69.0 (675)	wreek#refl
29.5 (4083)	70.5 (9757)	doe_voor#part_refl(voor)
29.3 (34)	70.7 (82)	bouw_op#part_refl(op)
29.3 (176)	70.7 (424)	open#refl
28.7 (45)	71.3 (112)	verhef#refl
27.4 (414)	72.6 (1098)	openbaar#refl

Algemene Nederlandse Spraakkunst

Zich-su word order is possible for verbs that have a somewhat 'bleeched' semantics, and express that something exists or comes into existence

ontspinnen, aandienen, vormen, ontvouwen, aftekenen, formeren, bevinden, verzamelen, voordoen, opbouwen, openen, verheffen, openbaren, ...

Two reflexive pronoun forms (Bouma and Spenader)

(29) Brouwers schaamt **zich**/*__zichzelf__ voor zijn schrijverschap.
Brouwers is ashamed of his writing

(30) Duitsland volgt **zichzelf** niet op als Europees kampioen.
Germany does not succeed itself as European champion

(31) Wie **zich**/**zichzelf** niet juist introduceert, valt af.
Everyone who does not introduce himself properly, is out.

- Are there differences between *zich* and *zichzelf*?
- What determines the choice between *zich* and *zichzelf*?

Properties of strong and weak reflexive pronouns

- *Zichzelf* is the strong, marked, less frequent, form
- Only *zichzelf* can be fronted (approx. 100 ex. in 470M word corpus)

(32) **Zichzelf** vereeuwigde Erdmann in de figuur van Thomas
Erdmann immortalized himself in the character of Thomas

(33) **Zichzelf** nam hij daarbij niet als voorbeeld
He did not take himself as example with this

- Only *zich* can appear between finite verb and subject

(34) Ruim 50 jaar geleden voltrok$_{vfin}$ **zich** [de watersnoodramp]$_{su}$
The flouding-disaster happened over 50 years ago

(35) Al vroeg bevinden$_{vfin}$ **zich** [duizenden supporters]$_{su}$ in het stadion
Already early, thousands of fans resided in the stadion

What governs the choice between two forms?

- Inherent reflexive verbs take only weak *zich*

(36) Brouwers vergist zich/*zichzelf
Brouwers mistakes himself

(37) Bush bemoeit zich/*zichzelf met Big Three
Bush occupies himself with Big Three

- Corpus does contain counterexamples:

(38) Hij verbeeldt zichzelf oogcontact te hebben
He imagines himself to have eye-contact

What governs the choice between two forms?

- Accidental reflexive verbs can occur both with zich and zichzelf
- If a verb is rarely used reflexively, it has a stronger preference for the strong form (Haspelmath, 2004, Smits, Hendriks, Spenader, 2007, Hendriks, Smits, Spenader, 2008)

Corpus Research

For all transitive, accidental reflexive, verbs

1. Count number of non-reflexive object arguments
2. Count number of weak reflexive arguments
3. Count number of strong reflexive arguments

Prediction: 1/(1+2+3) correlates with 3/(2+3)

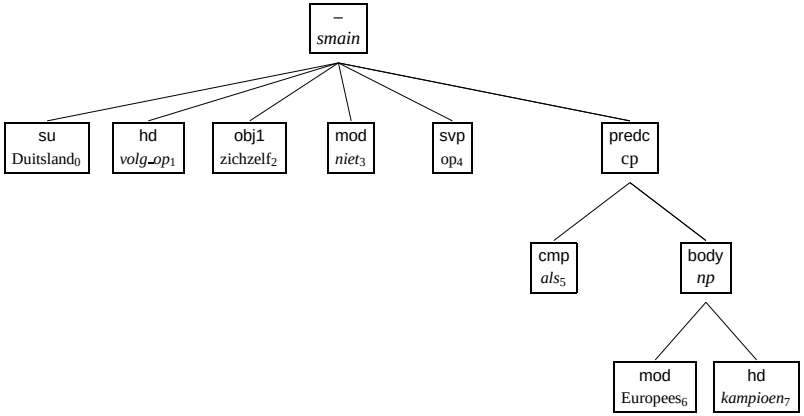
Treebank

- Counting verbs and their object arguments requires syntactic annotation
- Obtaining sufficient data for specific verbs (especially for reliable weak/strong reflexive counts) requires large amounts of data
- Only automatically constructed treebanks are large enough

Twente-News Corpus

- 470 M words of Dutch newspaper text (1994-2005)
- Automatically annotated with root-forms, POS-tags, and dependency relations using the Alpino-parser (van Noord, 2007)

Syntactic Annotation



Germany does not succeed itself as European champion

Previous work

Smits et al. 2006

- 80M word corpus (CLEF corpus, part of TwNC),
- 45 transitive verbs, manual selection of relevant cases,

Hendriks et al. 2007

- 300M word corpus (parts of TwNC)
- 32 selected transitive verbs, manual selection of relevant cases
- included 1st & 2nd person cases, non-reflexive cases = pronouns

This paper

- 470M word corpus (TwNC)
- all relevant transitive verbs,
- only 3rd person subjects, only object pronouns

Counting verbs or counting verb senses?

(39) De bedrijven maakten foute rekeningen op
*The companies **produced** wrong bills*

(40) De schelpdieren maken al het voedsel op
*The shellfish **take** all the food*

(41) Als ik 240 rijd, kan mijn assistente zich rustig opmaken
*If I drive 240, my assistent can still **put make-up on***

(42) De showbizz maakt zich op voor het huwelijk van het jaar
*The showbizz **prepares** itself for the marriage of the year*

- Better to count verb senses

Counting verbs or counting verb senses?

- Subcategorization-frames disambiguate between some senses

(43) De bedrijven maakten$_{part_trans(op)}$ foute rekeningen op
*The companies **produced** wrong bills*

(44) De schelpdieren maken$_{part_trans(op)}$ al het voedsel op
*The shellfish **take** all the food*

(45) Als ik 240 rijd, kan mijn assistente zich rustig opmaken$_{part_trans(op)}$
*If I drive 240, my assistent can still **put make-up on***

(46) De showbizz maakt$_{part_refl_pc_pp(op,voor)}$ zich op voor het huwelijk van het jaar
*The showbizz **prepares** itself for the marriage of the year*

- We counted occurrences of ⟨verb,subcategorization-frame⟩ pairs

Preliminary Corpus Observations

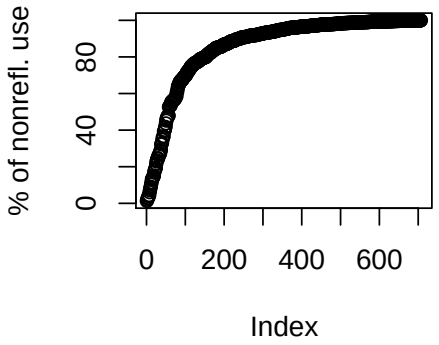
- 736 ⟨verb,subcat-frame⟩ pairs occur ≥50 times, and ≥ 10 times with a reflexive

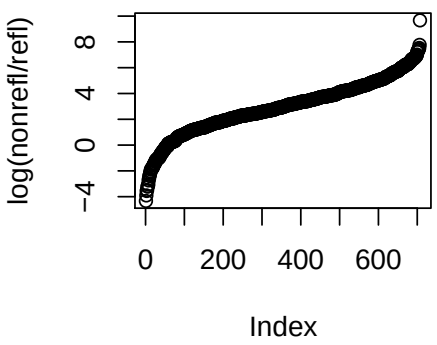
verb	nonrefl		refl		*zich*		*zichzelf*	
	#	%	#	%	#	%	#	%
straf (*to punish*)	1060	95.7	47	4.3	2	4.2	45	95.8
bescherm (*to protect*)	4921	96.4	186	3.6	95	51.1	91	48.9
vastketenen (*to chain*)	24	34.8	45	65.2	43	95.6	2	4.4

Strong Refl	≥ 95	≥ 50	≤ 8
Non-Refl Use	97.1%	95.1%	72.0%
# Verbs	44 (6%)	247 (34%)	187 (25%)

Percentages vs log of the ratio

- Distribution of non-reflexive vs reflexive use and strong reflexive vs weak reflexive use is not normal
- Taking the log of the ratio of non-reflexive over reflexive use (and strong reflexive over weak reflexive use) gives a more normal curve

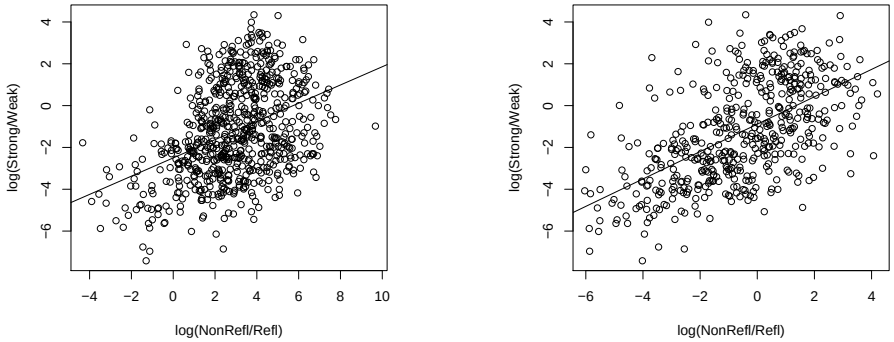




Counting all NPs or only pronouns?

- What counts as a relevant instance of non-reflexive use?
  - All non-reflexive object NPs?
  - Only non-reflexive object pronouns? (Haspelmath)
  - Only 3rd person non-reflexive pronouns? (Hendriks et al, 2008)

All nonreflexive NPs vs Pronouns



	# verbs	r^2	std err
all NPs	736	0.162	2.07
pronouns	594	0.293	1.98
3rd pers pro's	500	0.332	1.97

Comparison with Hendriks et al 2008

- Hendriks et al: $r^2 = 0.45$ for 32 selected verbs
- 24 of these verbs occur ≥ 50 times, and ≥ 10 with a reflexive
- for these 24 verbs, $r^2 = 0.547$
- Fully automatic data collection is as reliable as manually controlled selection...

Discussion

- Why do 32 (24) selected verbs score better?
  - Less ambiguous? More frequent?
- Why does contrasting reflexive use with non-reflexive pronoun use give better scores?
  - More coherent verb senses?
  - Restricts relevant cases to animate objects (as is the case for reflexives)?

Discussion

- What other factors might predict strong vs weak reflexive use
  - sentence position
  - stress
  - focus

	zichzelf	*zich*		*zichzelf*	*zich*
alleen (only)	109	1	*nu (now)*	16	1
ook (also)	214	9	*wel (certainly)*	14	0
niet (not)	30	9	*min of meer (more or less)*	21	0
slechts (only)	2	0	*alleen maar (only)*	13	1
zelfs (even)	7	0	*zo (that way)*	12	0

Conclusions

- Correlation between non-reflexive use and preference for strong reflexive pronouns can be demonstrated on fully automatic annotated and collected data
- Using more data for more verbs did not show higher correlation than in previous work
- Other factors that might explain choice between strong and weak reflexive pronoun (stress, focus) are hard to obtain automatically from corpora.