LSA 125 – Psycholinguistics and Syntactic Corpora

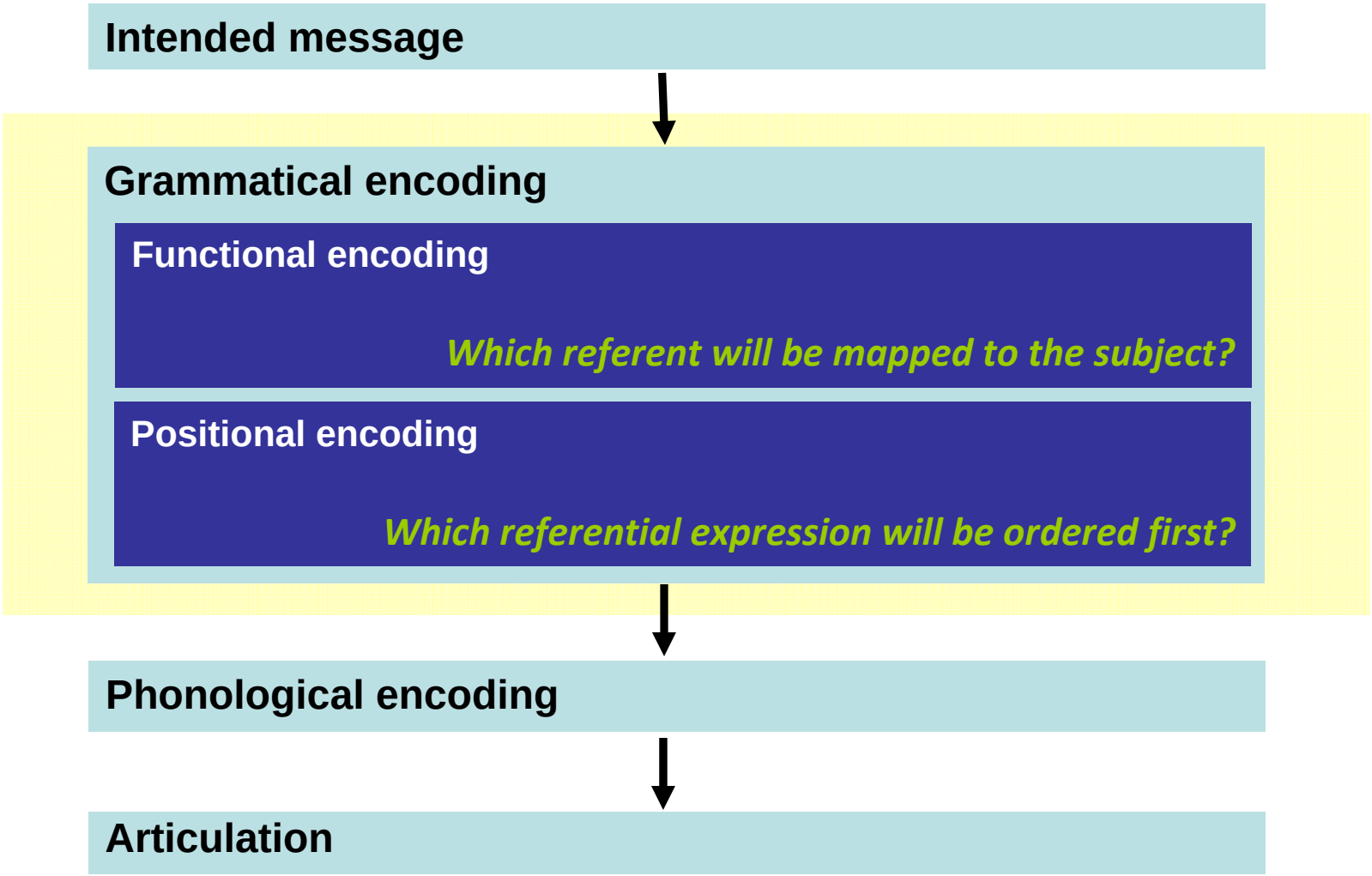
Today: *Processing accounts of syntactic variation*

LSA Summer Institute 2009, UC Berkeley

Florian Jaeger

TAs: Judith Degen, Alex Fine, and Peter Graff

Psycholinguistics



'Choice' points in production

Utterance level:	*Move the triangle to the left.*
	Select the triangle. Move it to the left.
Phrasal level:	*She gave {him the key/the key to him}*
	She already ate (dinner)
	She stabbed him (with a knife).
Word level:	*I read a book (that) she wrote.*
Morphological level:	*I've\have gone there.*
Phonological level:	t/d-deletion; final cluster reduction; vowel weakening
Phonetic level:	formant energies, F1/F2 ratio, speech rate

Methodology

- The most common approach to the study of language production are behavioral experiments
- But a corpus studies offer trade-offs that complement experiments:
  - Natural distributions
  - Heterogeneous, representative, data
  - Ecological validity (for spontaneous speech corpora)
  - But this comes with some computational and statistical challenges

Goals of this class

- Provide a background in the major psycholinguistic theories of production
- Provide an introduction to *corpus-based research on syntactic production*:
  - **Design** considerations
  - Syntactic search software **TGrep2** to extract data
  - **TGrep2 Database Tools** to combine TGrep2 output into tab-separated files
  - **R** for **statistical analysis**
- Discuss the trade-offs, limits, and some solutions to common issues in corpus-based work

Yes, we have a website

http://wiki.bcs.rochester.edu:2525/HlpLab/LSA09/

- Assignments
  - Readings and preparation for class discussion
  - Problems sets
  - Short paper
  - *Please collaborate in groups and follow instructions*
- To help, we have:
  - Tutorials on statistics (2) and **Linux/UNIX** (1 – **today!**)
  - Office hours (4h/week)
  - Tutorials on website
- So please take advantage

Today

- General background: **incremental production**
- 4/5 accounts of choice in incremental production:
  - Accessibility-based accounts
    - Availability (Levelt & Maassen, 1981; Ferreira, 1996; Ferreira & Dell, 2000)
    - Alignment (Bock & Warren, 1985; Ferreira, 1994)
  - Ambiguity avoidance (Bolinger, 1972; Hawkins, 2004)
  - Dependency-minimization (Hawkins, 1994, 2001, 2004)
  - Uniform Information Density (Jaeger, 2006; Levy & Jaeger, 2007)

- Introduce choice point environments (alternations)

Psycholinguistics theories

- Processing theories of (morpho)syntactic variation share a couple of assumptions:
  - There are often *multiple ways to convey the same message* → speakers have a '**choice**' between these different variants.
  - Speakers exhibit **preferences** at these 'choice points'
  - These preferences are attributable to some type of **processing pressure**, typically some pressure to be efficient.

→By studying speakers behavior at these choice points we can learn about production mechanisms

cnt'd

- A variety of psycholinguistic accounts of production preferences have been proposed; they differ in what is assumed to be the driving processing pressure.

- There also are non-processing-based accounts (not mutually exclusive):
  - **Gradient grammar accounts**
  - **'Paraphrase deniers'** (Wasow, p.c.): there ain't no choice
    - Semantic accounts
    - Functionalist accounts

Next Monday

- First some preliminaries
  - Choice
  - Same message
  - Incrementality

'Choice'

- Central notion to work on sentence production (Osgood & Bock, 1977; MacWhinney & Bates, 1978)
  - NB: … the late start of work on sentence production

- Choice:
  - Mostly if not always **subconscious** and **highly automatic**
  - Just refers to the presence of a choice point in the production system.
- **More choice points than immediately apparent**, not only alternations.

Variation/Choice/Selection/Preference

- 'In the process of speaking conceptual content is expressed in linguistics form. This requires the speakers to make a variety of conceptual and linguistic **decisions**' (Levelt & Maassen, 1981:221)

- 'One way to evaluate the impact of these [production and communication] pressures is to examine the **decisions** that are made when the language production system builds a sentence …' (Ferreira & Dell, 2000:297)

- 'Substantial research … has shown that variations in animacy are associated with **variations** in syntax, such as possible case marking and voice **selection** …' (Branigan et al., 2007:173)

‘Same message’

- Consider:
  - Active vs. passive
  - *it*-clefts, left-dislocation, topicalization
  - Perspective verbs, symmetrical verbs
  - Ditransitive alternation
  - *that*-mentioning
  - auxiliary contraction

- Same meaning:
  - Same truth-conditions? Relatively fair bet
  - Same connotations? Implications? Probably not?

Semantics: just one more pressure?

- What if meaning is part of the 'pressures' that speakers try to balance? (for more detail, see Jaeger, 2006: Ch 6.2.2)
  - cf. bi-directional OT

Incrementality

- **Incrementality:** Sentences aren't planned all-in-one (Bock, 1982; Kempen & Hoenkamp, 1987; Levelt, 1989)
- Different amount of **look-ahead** at different 'levels' of planning, e.g.:
  - Subject and verb need to be planned to some extend before utterance articulation is initiated (Lindsley, 1975)
  - At least one word/syllable needs to be planned before articulation can be initiated (cf. Griffin, 2003; Smith & Wheeldon, 1997; Wheeldon & Lahiri, 1999)
  - NB: Amount of planning depends on task-pressures, etc. (cf. Griffin, 2003: 1)

→ **Accounts of speakers' choice need to be compatible with incrementality**

Accessibility & Choice in Production

- Accessibility: 'ease of representing potential referents in thought' (Bock & Warren, 1985:47)

- H: **accessibility affects speakers' choices in production**
  - cf. early observation that properties that affect word recall also seem to affect constituent order (Bock, 1980) →ordering preferences linked to lexical search (Levelt, 1979; Levelt & Maassen, 1981; Bock & Irwin, 1980) and to minimization of memory load (Levelt & Maassen, 1981)

What contributes to accessibility?

- Empathy? Perspective?
- Concreteness, imageability
- Prototypicality
- Animacy
- Givenness:
  - of referent
  - of expression
- Phonological inhibition
- Word frequency, predictability?

Alignment
(a.k.a. *indirect* effects of accessibility)

- "… conceptual accessibility is related to a hierarchy of grammatical relations" (Bock & Warren, 1985:64)

- **Grammatical functions higher on the hierarchy should be mapped onto more accessible referents**.
  - SUBJ > direct OBJ > indirect OBJ > …

- Let's call this an alignment account: **similar – though arguably more elaborate– theories have been developed in linguistics** (e.g. Aissen, 2003; Bresnan et al., 2001)
  - cf. obviation, alignment of grammatical function, discourse function, salience, and linguistic *structure*

Ditransitive Alternation

(Bresnan et al., 2007)

- Effect on (log-)odds of prepositional dative (NP PP), where theme referent is mapped to direct object.*

Prepositional dative structure:	*. . . gave [toys] [to the children]*	V NP PP
Double object structure:	*. . . gave [the children] [toys]*	V NP NP

Logistic regression model of corpus data

Model B: Relative magnitudes of significant effects.

	Coefficient	Odds ratio PP	95% C.I.
Nonpronominality of recipient	1.73	5.67	3.25–9.89
Inanimacy of recipient	1.53	5.62	2.08–10.29
Nongivenness of recipient	1.45	4.28	2.42–7.59
Indefiniteness of recipient	0.72	2.05	1.20–3.5
Plural number of theme	0.72	2.06	1.37–3.11
Structural parallelism in dialogue	–1.13	0.32	0.23–0.46
Nongivenness of theme	–1.17	0.31	0.18–0.54
Length difference (log scale)	–1.16	0.31	0.25–0.4
Indefiniteness of theme	–1.74	0.18	0.11–0.28
Nonpronominality of theme	–2.17	0.11	0.07–0.19

Availability

(a.k.a. *direct* effects of accessibility)

- But these results could be interpreted differently; what if …

  Speakers prefer to order more available material earlier in the sentence. (Bock & Irwin, 1980; Levelt & Maassen, 1981; Ferreira & Dell, 2000)

- **Principle of Immediate Mention:** 'Production proceeds more efficiently if syntactic structures are used that permit quickly selected lemmas to be mentioned as soon as possible.' (Ferreira & Dell, 2000:299)
  - **Early articulation – faster processing** (e.g. Branigan et al., 2007; Ferreira & Dell, 2000)
  - **Fluency** (Race & MacDonald, 2003)

Reduction of Complement Clauses

(Ferreira & Dell, 2000: Experiments 1 & 2)

- Higher rate of *that* if complement clause subject is less available:

Recall study:
encode →
distractor →
cued recall

The coach knew (that) I missed practice.
The coach knew (that) you missed practice.

The coach knew (that) she/he missed practice.
The coach knew (that) you missed practice.

Misenc (%)	Mem Acc (%)	Scored (%)	Full sentences produced: Encoded reduced	Encoded full	*M*
3.3	63.2	78.2	61.2	87.2	74.3
2.6	63.2	67.6	63.4	89.6	76.6
5.3	66.8	71.4	51.2	84.2	68.0
6.3	67.1	77.6	53.2	86.9	70.2

cnt'd

— a corpus study of spoken American English (Jaeger, submitted)

- Predicting full complement clauses with *that:*

Predictor	Coef. β	SE(β)	z	p
Intercept	0.119	(0.376)	0.3	> 0.7
Position(Matrix Verb)	0.948	(0.143)	6.6	< 0.0001
(1st restricted comp.)	−27.819	(5.331)	−5.2	< 0.0001
(2nd restricted comp.)	55.185	(10.794)	−5.2	< 0.0001
Length(Matrix Verb-to-CC)	0.172	(0.065)	2.7	< 0.008
Length(CC Onset)	0.180	(0.014)	12.8	< 0.0001
Length(CC Remainder)	0.026	(0.006)	4.3	< 0.0001
Log Speech Rate	−0.700	(0.129)	−5.4	< 0.0001
Sq Log Speech Rate	−0.365	(0.190)	−1.9	< 0.06
Pause	1.100	(0.108)	10.5	< 0.0001
Disfluency	0.395	(0.122)	3.2	< 0.002
CC Subject =*it vs. I*	0.037	(0.077)	0.3	> 0.6
=*other pro vs. prev. levels*	0.053	(0.033)	1.6	$= 0.11$
=*other NP vs. prev. levels*	0.111	(0.023)	4.9	< 0.0001
FQ(CC Subject Head)	−0.019	(0.028)	−0.7	> 0.4
Subject Identity	−0.317	(0.166)	−1.9	< 0.056
Word Form OCP	−0.316	(0.170)	−1.9	< 0.063
FQ(Matrix Verb)	−0.208	(0.030)	−7.0	< 0.0001
Ambiguous CC Onset	−0.116	(0.115)	−1.0	> 0.3
Persistence =*no vs. prime w/o that*	0.019	(0.067)	0.3	> 0.7
=*prime w/ that vs. prev. levels*	0.058	(0.035)	1.6	$= 0.10$
Matrix Subject =*you*	0.484	(0.152)	3.2	< 0.0015
=*other PRO*	0.616	(0.125)	4.9	< 0.0001
=*other NP*	0.862	(0.128)	6.7	< 0.0001
Male Speaker	−0.157	(0.111)	−1.4	> 0.15
Information Density	**0.639**	**(0.038)**	**16.6**	< 0.0001

Beware when reading …

- 'Accessibility' used somewhat consistently only after 1985 and onward (also be aware of the related but different use of 'accessibility' in work on referential expressions, Ariel, 1990, etc.)

- Referential availability/accessibility (Bock & Irwin, 1980) = Conceptual accessibility (Bock & Warren, 1985)
  - Inherent accessibility (Prat-Sala & Branigan, 2000)
  - Derived accessibility (Prat-Sala & Branigan, 2000)

- Lexical availability/accessibility (Bock & Irwin, 1980)

Availability vs. Alignment

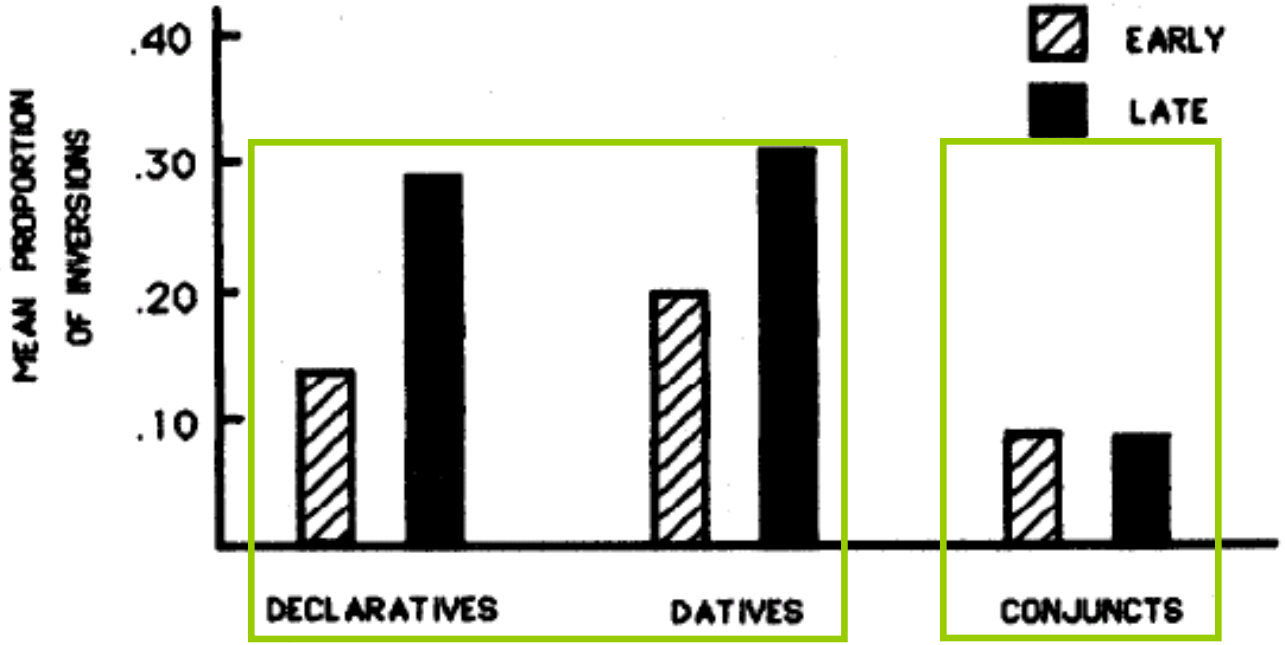
- The earliest research does not systematically try to distinguish between these two types of accounts

- But starting at least 1985, there is awareness that **grammatical function and word order are highly correlated** (in English!) (see also discussion about conceptual vs. lexical accessibility in Bock & Warren, 1985:61-64)

English: Imageability

(Bock & Warren, 1985)

Figure 1. *Mean inversion proportions (changes from a presented sentence form to an alternative form as a proportion of meaning-preserving recalls) for three sentence types. EARLY and LATE denote the locations in the presented sentences of the more imageable target noun. For declaratives, the early position was the subject, and the late position the direct object. For datives, the early position was the object immediately after the verb, and the late position was the second object. For conjuncts, the early position was the first noun in a conjunctive noun phrase, and the late position was the second noun in the same conjunctive noun phrase.*



But …

(McDonald et al., 1993)

- Animacy affects *both voice choice and conjunct order.*

TABLE 4
Percentages of Phrases Recalled in Correct and Shift Categories, Experiment 3

Original sentence type	First noun in presented phrase							
	Correct				Shift			
	Animate		Inanimate		Animate		Inanimate	
	Short	Long	Short	Long	Short	Long	Short	Long
Transitive	47	38	20	21	4	4	18	27
Conjunctive	38	38	20	26	6	6	18	20

‘Coordination is weird’

- For some time, (inconsistent) evidence from coordination was the only evidence considered when trying to distinguish between availability and alignment accounts.

- More recently, this question has been addressed cross-linguistically by studying languages where GF-assignment and order are not as systematically confounded (compared to English)

English & Spanish: Animacy, Givenness

(Prat-Sala & Branigan, 2000)

- Both left-dislocation (in Spanish) and passives (in Spanish and English) more frequent if patient more accessible.
- Interpreted as evidence for **availability**, but …?

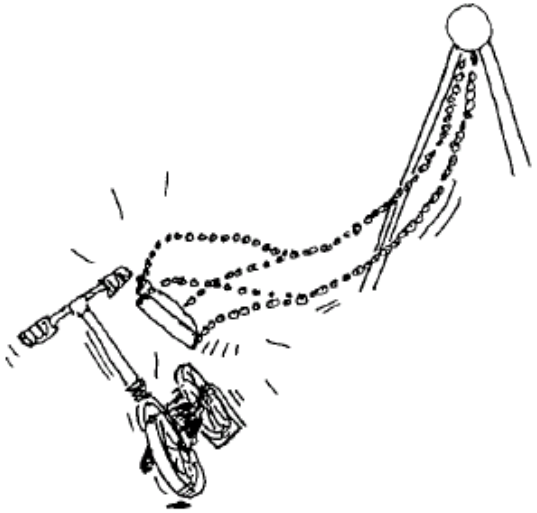


FIG. 1. Example of a picture used in Experiment 1. A swing hitting a scooter.

FIG. 2. Example of a picture used in Experiment 2. A swing hitting a man.

German: Animacy – a corpus study

(Kempen & Harbusch, 2004)

- Maybe a better example: **animacy affects *scrambling*** even when grammatical function, definiteness, and pronominality are held constant (but small numbers!)

Table 2. Linear order frequencies of (SBful,DOpro) pairs, for animate and inanimate SBful members.

	Linear order	
	SBful ≺ DOpro	DOpro ≺ SBful
SBful inanimate	11	64
SBful animate	52	56

Table 3. Linear order frequencies of (SBful,IOful) pairs, for animate and inanimate IOful members

	Linear order	
	IOful ≺ SBful	SBful ≺ IOful
IOful inanimate	3	39
IOful animate	17	20

Greek and Japanese: Animacy

(Tanaka et al., 2005; Branigan & Feleki, 1999; Branigan et al., 2007)

(a) Sta dimokratika politevmata, o politis sevete to sinda
in democratic regimes the citizen-NOM respects the law-
In democratic regimes, the citizen respects the law

(b) Sta dimokratika politevmata, to sindagma sevete o politis.
in democratic regimes the law-ACC respects the citizen-N
In democratic regimes, the citizen respects the law

(c) Sta dimokratika politevmata, to sindagma sevete ton politi.
in democratic regimes the law-NOM respects the citizen-
In democratic regimes, the law respects the citizen

(d) Sta dimokratika politevmata, ton politi sevete to sinda
in democratic regimes the citizen-ACC respects the law-N
In democratic regimes, the law respects the citizen

(a) ryokousha-ga takushii-o tukamaeta.
traveller-NOM taxi-ACC pick up-PAST
A traveler picked up a taxi.

(b) takushii-o ryokousha-ga tukamaeta.
taxi-ACC traveller-NOM pick up-PAST
A traveler picked up a taxi.

(c) takushii-ga ryokousha-o tukamaeta.
taxi-NOM traveller-ACC pick up-PAST
A taxi picked up a traveler.

(d) ryokousha-o takushii-ga tukamaeta.
traveller-ACC taxi-NOM pick up-PAST
A taxi picked up a traveler.

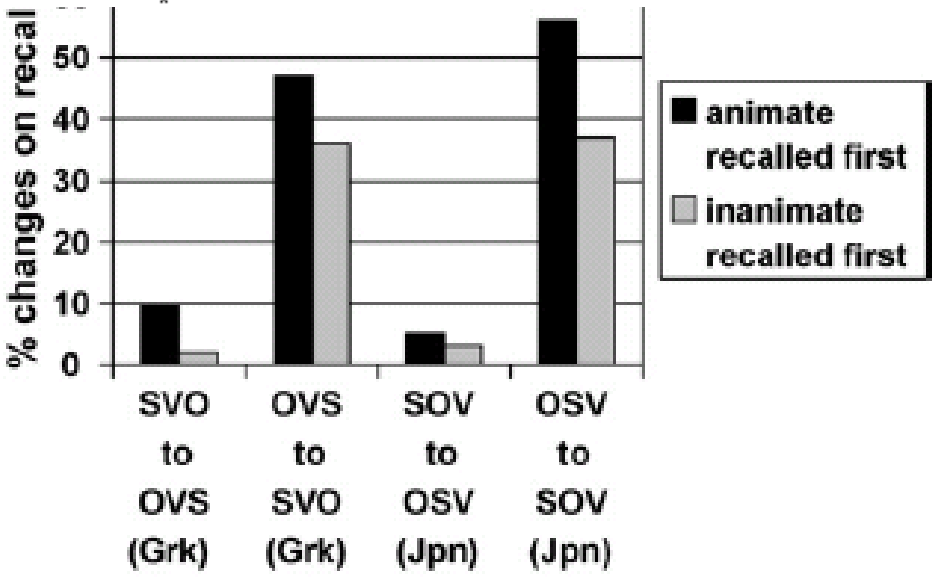


Fig. 1. Data from Branigan and Feleki (1999) and Tanaka et al. (2005): percentage of responses involving correct meaning but changes in form. For instance, *SVO to OVS* means SVO utterances recalled as OVS. *Grk*, Greek; *Jpn*, Japanese.

(a) ryokousha-ga takushii-niyotte tukamae-rare-ta.
traveller-NOM taxi-OBL pick up-PASSIVE-PAST
A traveler was picked up by a taxi.

(b) takushii-niyotte ryokousha-ga tukamae-rare-ta.
taxi-OBL traveller-NOM pick up-PASSIVE-PAST
A traveler was picked up by a taxi.

(c) takushii-ga ryokousha-niyotte tukamae-rare-ta.
taxi-NOM traveller-OBL pick up-PASSIVE-PAST
A taxi was picked up by a traveler.

(d) ryokousha-niyotte takushii-ga tukamae-rare-ta.
traveller-OBL taxi-NOM pick up-PASSIVE-PAST *118 (2008) 172 189*
A taxi was picked up by a traveler.

y, cnt'd (Tanaka et al., 2005)

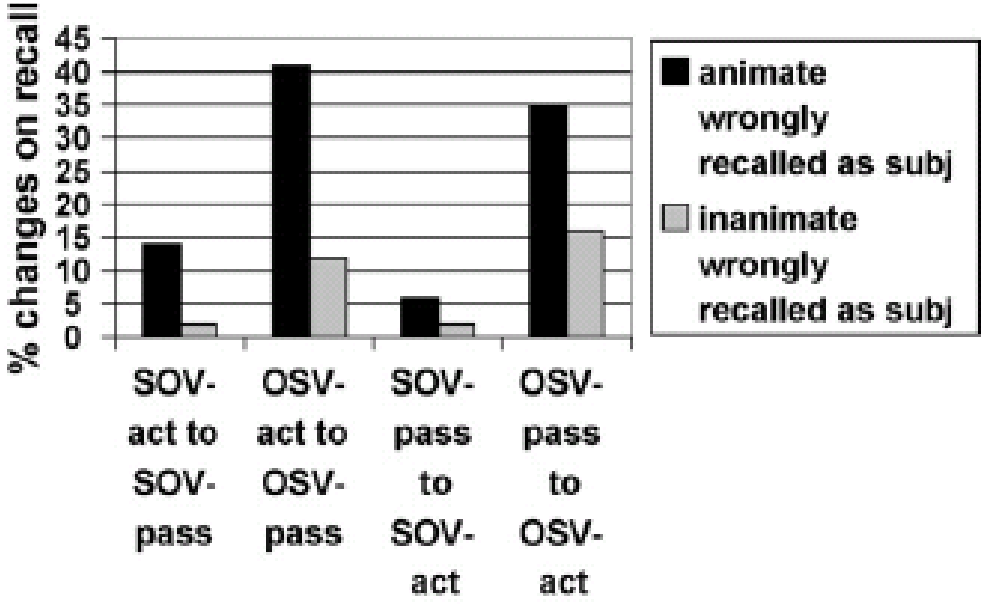


Fig. 2. Data from Tanaka et al. (2005): percentage of responses involving correct meaning but changes in voice. For instance, *SOV-act to SOV-pass* means SOV active utterances recalled as SOV passive.

What contributes to accessibility? Revisited

- How do the properties that have been shown to affect syntactic production relate to alignment vs. availability accounts? What do you think?
  - Concreteness, imageability
  - Prototypicality
  - Animacy
  - Givenness:
    - of referent
    - of expression
  - Phonological inhibition
  - Word frequency, predictability

Grammatical weight

- The **complexity** of constituents (or sometimes words) has also been hypothesized to affect speakers' choices.
- **Das Gesetz der Wachsenden Glieder:** "Von zwei Gliedern von verschiedenem Umfang steht das umfangreichere nach. " (Behaghel, 1930:85)

- Example from Wasow (1997): "… allowing us to document
  - in great detail
  - the psychology of linguistic rules
  - from infancy to old age
  - in both normal and neurologically impaired people,
  - in much the same way that biologists focus on the fruit fly *Drosophila* to study the machinery of the genes."

Ditransitive Alternation

(Bresnan et al., 2007)

- Predicting NPPP-order:

Prepositional dative structure:	*. . . gave [toys] [to the children]*	V NP PP
Double object structure:	*. . . gave [the children] [toys]*	V NP NP

length(recipient) – length(theme)

Relative magnitudes of significant effects.

	Coefficient	Odds ratio PP	95% C.I.
Nonpro inality of recipient	1.73	5.67	3.25–9.89
Inanimacy o cipient	1.53	5.62	2.08–10.29
Nongivenness o cipient	1.45	4.28	2.42–7.59
Indefiniteness of recip nt	0.72	2.05	1.20–3.5
Plural number of theme	0.72	2.06	1.37–3.11
Structural parallelism in dialogue	–1.13	0.32	0.23–0.46
Nongivenness of theme	–1.17	0.31	0.18–0.54
Length difference (log scale)	–1.16	0.31	0.25–0.4
Indefiniteness of theme	–1.74	0.18	0.11–0.28
Nonpronominality of theme	–2.17	0.11	0.07–0.19

What determines grammatical weight?

- Not clear: different proposed measures are too highly correlated!

- E.g. Szmrecsanyi (2004:1037; see also Wasow, 2002; Gomez Gallo et al., 2008):

	SL vs. NODE	ISC vs. NODE
Susanne corpus (written English)	.976**	.836**
Christine corpus (spoken English)	.989**	.916**

** significant at the .01 level

Table 2. Correlation coefficients for rankings suggested by different structural measures

cnt'd

- Both **production-oriented** and **parsing-oriented** weight accounts have been proposed.

Parsing-oriented: Minimize Domain (MiD)

(Hawkins, 2004; here taken from Hawkins, 2007:96)

- The human processor prefers to **minimize the connected sequences of linguistic forms and their conventionally associated syntactic and semantic properties in which relations of combination and/or dependency are processed**.
  - The degree of this preference is proportional to the number of relations whose domains can be minimized in competing sequences or structures, and to the extent of the minimization difference in each domain.
  - **Combination:** Two categories A and B are in a relation of combination iff they occur within the same syntactic mother phrase and maximal projections (phrasal combination), or if they occur within the same lexical co-occurrence frame (lexical combination).
  - **Dependency:** Two categories A and B are in a relation of dependency iff the parsing of B requires access to A for the assignment of syntactic or semantic properties to B with respect to which B is zero-specified or ambiguously or polysemously specified.

PP-ordering in English

(Hawkins, 1999; taken from Hawkins, 2007:97)

(19) a. The man vp[waited pp1[for his son] pp2[in the cold but not unpleasant wind]]
1 2 3 4 5
---

b. The man vp[waited pp2[in the cold but not unpleasant wind] pp1[for his son]]
1 2 3 4 5 6 7 8 9

Structures like (19) were selected from a corpus on the basis of a permutation test (Hawkins, 2000, 2001): the two PPs had to be permutable with truth-conditional equivalence (i.e. the speaker had a choice). Only 15% (58/394) of these English sequences had long before short. Among those with at least a one-word weight difference (excluding 71 with equal weight), 82% had short before long, and there was a gradual reduction in the long before short orders, the bigger the weight difference (**PPS** = shorter **PP**, **PPL** = longer **PP**):

(22)	PPL > PPS by 1 word	by 2 4	by 5 6	by 7+
[V PPS PPL]	60% (58)	86% (108)	94% (31)	99% (68)
[V PPL PPS]	40% (38)	14% (17)	6% (2)	1% (1)

Verb-particle split in English

(Lohse et al., 2004)

(7) a. Joe $_{VP}$[looked up $_{NP}$[the number of the ticket]]

1 2 3

VP PCD: IC-to-word ratio of 3/3 = 100%

b. Joe $_{VP}$[looked $_{NP}$[the number of the ticket] up]

1 2 3 4 5 6 7

VP PCD: IC-to-word ratio of 3/7 = 43%

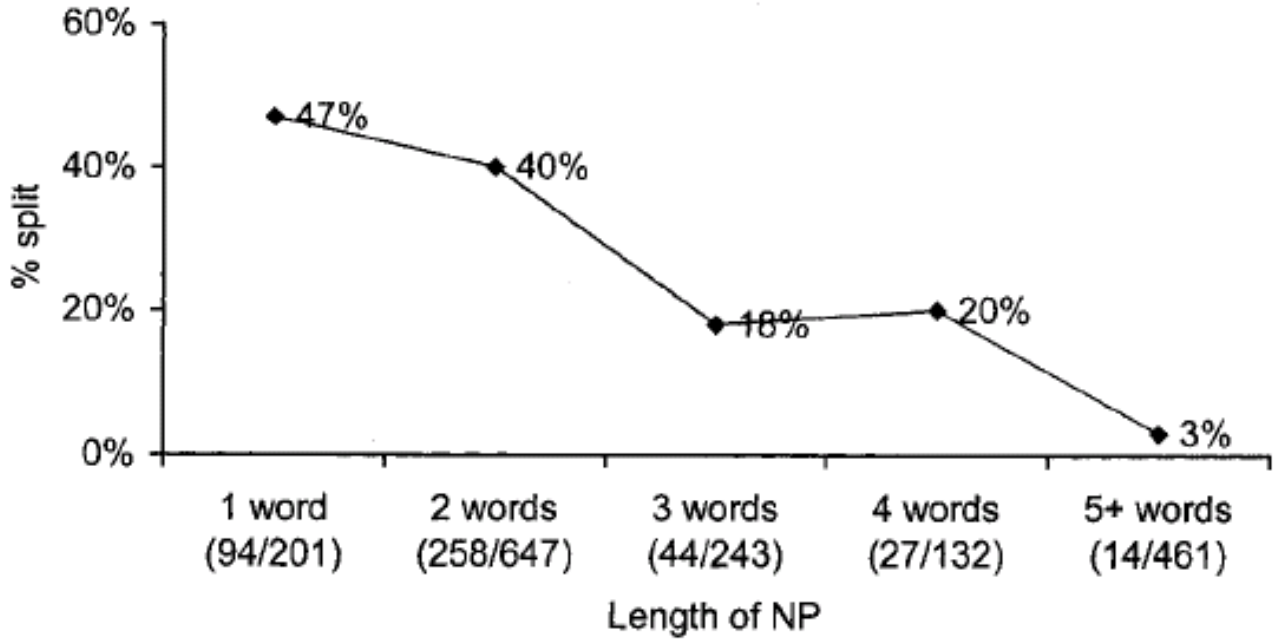


FIGURE 1. Split vs. joined by NP-length.

Hungarian NP-ordering

(Hawkins, 1994; taken from Hawkins, 2001)

(4) (a) [Döngetik $_{NP}$[facipöink] $_{NP}$[az utcakat]]. (Hungarian)
batter wooden shoes-1PL the streets-ACC
'Our wooden shoes batter the streets.'
(b) [Döngetik $_{NP}$[az utcakat] $_{NP}$[facipöink]]

$_{m}$NP = any NP constructed on its left periphery
$NP_2 > NP_1$ in word length
Actual numbers of sequences given in parentheses
An additional 21 sequences had NPs of equal length (total n = 116)

n = 95	$_{m}NP_2 > {}_{m}NP_1$ by 1 word	by 2 words	by 3+ words
$[V\ {}_{m}NP_1\ {}_{m}NP_2]$	85% (50)	96% (27)	100% (8)
$[V\ {}_{m}NP_2\ {}_{m}NP_1]$	15% (9)	4% (1)	0% (0)

Table 2
Hungarian noun phrase orderings by relative weight (Hawkins 1994: 133)

Maximize Online Processing (MaOP)

(Hawkins, 2004; here taken from Hawkins, 2007:101)

- The human processor prefers to maximize the set of properties that are assignable to each item X as X is processed, thereby **increasing O(n-line) P(roperty) to U(ltimate) P(roperty) ratios**.
  - The maximization difference between competing orders and structures will be a function of the number of properties that are unassigned or misassigned to X in a structure/sequence S, compared with the number in an alternative.
  - **On-line Property to Ultimate Property Ratio:** The OP-to-UP ratio is calculated at each word X within a connected set of words f. . .X. . .g whose property assignments differ between two competing structures S and S0. The cumulative number of properties assignable at each X is divided by the total number of properties to be assigned in the connected set in the ultimate representation of each structure, and the result is expressed as a […]. The higher these on-line percentages, the more efficient is structure S or S0, since more properties are assigned earlier

cnt'd

- Supposed to capture asymmetries not captured by MiD, whenever B depends more on A than A on B; e.g. topic-before-subject, subject-before-object.
- cf.

(13) Relative frequencies of basic word orders in Tomlin's (1986) sample (402 lgs)

SOV (168)	VSO (37)	VOS (12)	OVS (5)
SVO (180)			OSV (0)
87%	9%	3%	1%

Production-oriented weight accounts

- "… the principal reason speakers postpone the production of complex constituents is that it helps them in planning their utterances." (Wasow, 1997)

- **Arguments against parsing-oriented accounts** (cf. Wasow, 1997):
  - MiD (and to some extent MaOP) require access to entire sentence prior to decision → **incompatible with incrementality**
  - Longer phrases can precede shorter phrases, if they the shorter phrase is hard to retrieve: "That will bring to the plate … Barry Bonds"

cnt'd

- That's a very availability-oriented account of grammatical weight effects

- **Aside:** is grammatical weight the same as accessibility? Cf.:
  - "Can you hand me *some xylophones*?" vs.
  - "Can you hand me *that nice little things over there*?"

- **This is an open question ripe for a test …**

Problems for parsing-oriented accounts

(Jaeger, submitted)

- Predicting full complement clauses with *that:*
- Some effects support MiD
- Other weight effects argue for production-oriented accounts

Predictor	Coef. β	SE(β)	z	p
Intercept	0.119	(0.376)	0.3	> 0.7
Position(Matrix Verb)	0.948	(0.143)	6.6	< 0.0001
(1st restricted comp.)	−27.819	(5.331)	−5.2	< 0.0001
(2nd restricted comp.)	55.185	(10.794)	−5.2	< 0.0001
Length(Matrix Verb-to-CC)	0.172	(0.065)	2.7	< 0.008
Length(CC Onset)	0.180	(0.014)	12.8	< 0.0001
Length(CC Remainder)	0.026	(0.006)	4.3	< 0.0001
Log Speech Rate	−0.700	(0.129)	−5.4	< 0.0001
Sq Log Speech Rate	−0.365	(0.190)	−1.9	< 0.06
Pause	1.100	(0.108)	10.5	< 0.0001
Disfluency	0.395	(0.122)	3.2	< 0.002
CC Subject =*it vs. I*	0.037	(0.077)	0.3	> 0.6
=*other pro vs. prev. levels*	0.053	(0.033)	1.6	$= 0.11$
=*other NP vs. prev. levels*	0.111	(0.023)	4.9	< 0.0001
FQ(CC Subject Head)	−0.019	(0.028)	−0.7	> 0.4
Subject Identity	−0.317	(0.166)	−1.9	< 0.056
Word Form OCP	−0.316	(0.170)	−1.9	< 0.063
FQ(Matrix Verb)	−0.208	(0.030)	−7.0	< 0.0001
Ambiguous CC Onset	−0.116	(0.115)	−1.0	> 0.3
Persistence =*no vs. prime w/o that*	0.019	(0.067)	0.3	> 0.7
=*prime w/ that vs. prev. levels*	0.058	(0.035)	1.6	$= 0.10$
Matrix Subject =*you*	0.484	(0.152)	3.2	< 0.0015
=*other PRO*	0.616	(0.125)	4.9	< 0.0001
=*other NP*	0.862	(0.128)	6.7	< 0.0001
Male Speaker	−0.157	(0.111)	−1.4	> 0.15
Information Density	**0.639**	**(0.038)**	**16.6**	< 0.0001

Problems for production-oriented account (cf. Hawkins, 2007:106-107)

- Production-oriented accounts without reference to dependencies run into problems:
  - While English shows pretty consistent short-before-long ordering *post-verbally* (ditransitives, HNPS, etc.), …
  - left-dislocation and topicalization seem to favor long constituents → **long-before-short pre-verbally in English** (Snider, 2005:23-24)
  - **Verb-final languages seem to favor long-before-short**, too (Japanese: Hawkins, 1994; Yamashita & Chang, 2001; Korean: Choi, 2007), e.g. Hawkins, 1994 (Table taken from Hawkins, 2007:99):

(24)	ICL > ICS by 1 2 words	by 3 4	by 5 8	by 9 +
[ICS ICL V]	34% (30)	28% (8)	17% (4)	9% (1)
[ICL ICS V]	66% (59)	72% (21)	83% (20)	91% (10)

cnt'd

- This seems to point towards a dependency-based explanation in terms of MiD and MaOP (Hawkins, 2004, 2007, 2009)
  - Also supported by typological data

(16)

a.	vp[went pp[to the movies]]	b.	[[the movies to]pp went]vp
c.	vp[went [the movies to]pp]	d.	[pp[to the movies] went]vp

(17)

a. vp[V pp[P NP]] = 161 (41%)
b. [[NP P]pp V]vp = 204 (52%)
c. vp[V [NP P]pp] = 18 (5%)
d. [pp[P NP] V]vp = 6 (2%)
Preferred (17a) + (b) = 365/389 (94%)

- But preliminary evidence from German PP-ordering (Fine, 2007) suggests that the situation is (even) more complicated.

Ambiguity Avoidance

- The last account of speakers' choices to be discussed today builds on ideas from Audience Design (Brennan & Williams, 1995; Clark & Murphy, 1982; Clark, 1992; Lockridge & Brennan, 2002)

- Speaker may avoid ambiguity, or –more specifically, so called garden paths for comprehenders (Bolinger, 1972; Hawkins, 2004; Snedecker & Trueswell, 2003; Temperley, 2003)

Evidence from PP attachment

(Haywood et al., 2005)

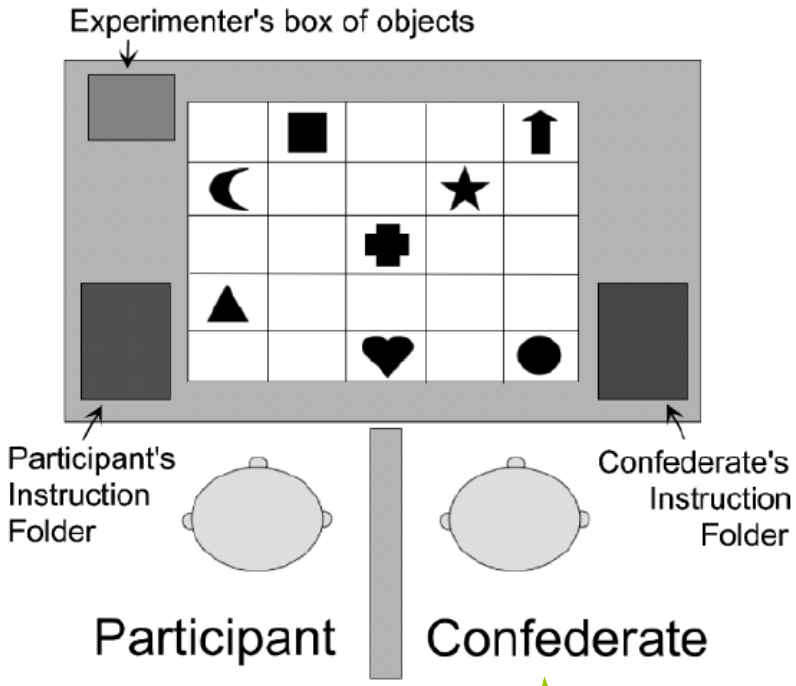


Fig. 2. Overhead view of the experimental setup.

helpful or not

Put the penguin in the cup …

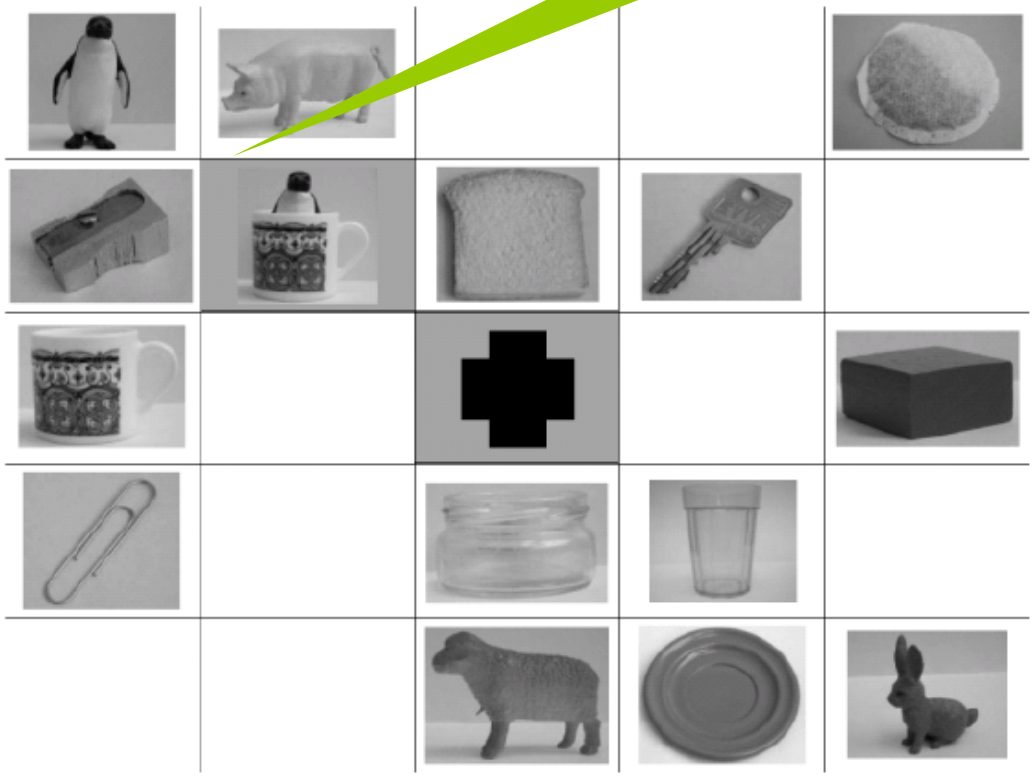


Fig. 1. Sample target card. (Color photographs were used in the experiment.) This card comes from the end of a round, when there was only one geometric shape left to be covered up with an object.

cnt'd

TABLE 1

Proportion of Target Responses Including That's *or Any Disambiguating Word*

Condition	Helpful confederate	Unhelpful confederate
Responses with *that's*		
Prime without *that's*		
Unambiguous (one referent) context	.13 (.20)	.17 (.26)
Ambiguous (two referent) context	.25 (.30)	.17 (.26)
Prime with *that's*		
Unambiguous (one referent) context	.50 (.38)	.33 (.32)
Ambiguous (two referent) context	.53 (.40)	.49 (.38)
Responses with any disambiguating word or words		
Prime without *that's*		
Unambiguous (one referent) context	.15 (.19)	.17 (.26)
Ambiguous (two referent) context	.29 (.30)	.18 (.26)
Prime with *that's*		
Unambiguous (one referent) context	.50 (.38)	.33 (.32)
Ambiguous (two referent) context	.60 (.38)	.50 (.38)

Proportions analyzed with ANOVA → has to be interpreted with caution

But …

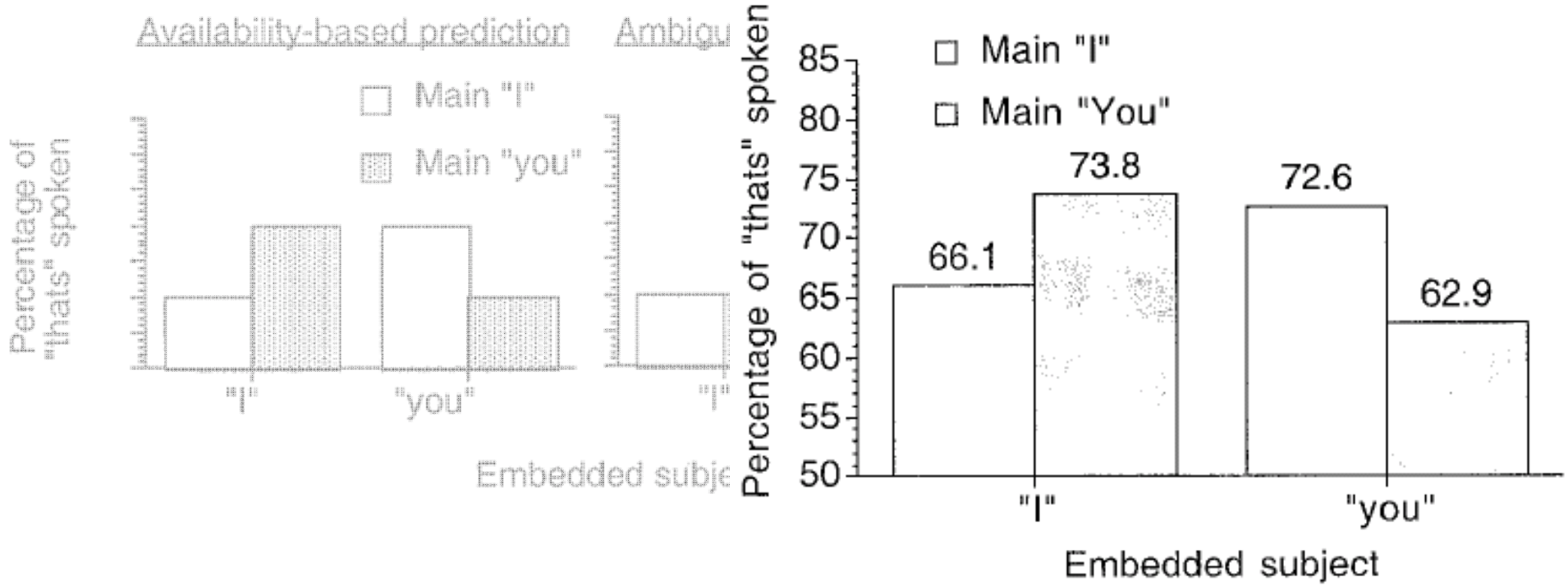
- … other studies –many of which contained better controls than previous studies– did *not* find ambiguity avoidance effects (Arnold et al., 2004; Ferreira & Dell, 2000; Jaeger, 2006; Kraljic & Brennan, 2005; Roland et al., 2005; Schafer et al., 2000)

- … and others found only very weak evidence (Jaeger, submitted, in prep)

No sign of ambiguity avoidance

(Ferreira & Dell, 2000:316-318)

Repeated?		Ambiguous?
Yes	I knew (that) I missed practice.	No
No	You knew (that) I missed practice.	No
No	I knew (that) you missed practice.	Yes
Yes	You knew (that) you missed practice.	No



Replicated in corpus study

(Roland et al., 2005:25)

- *that*-mentioning in complement clauses in British English: **no sign of ambiguity avoidance**

Table 7
Frequencies of *that* ellipsis by type of post-verbal NP

Type of post-verbal NP	SC-0	SC-that	% *that* ellipsis
Full NP	49,543	91,691	35.1
Unambiguous pronoun	49,754	19,855	71.5
Ambiguous pronoun	37,474	12,710	74.7

No PP-attachment avoidance

(Arnold et al., 2004:Experiments 2 & 3)

Table 1
Sample stimuli and results for Experiment 1

Stimuli
(1) Potential ambiguity
(a) The foundation gave a museum in Philadelphia Grant's letters to Lincoln.
(b) The foundation gave Grant's letters to Lincoln to a museum in Philadelphia.
(2) No potential ambiguity
(a) The foundation gave a museum in Philadelphia Grant's letters praising Lincoln.
(b) The foundation gave Grant's letters praising Lincoln to a museum in Philadelphia.

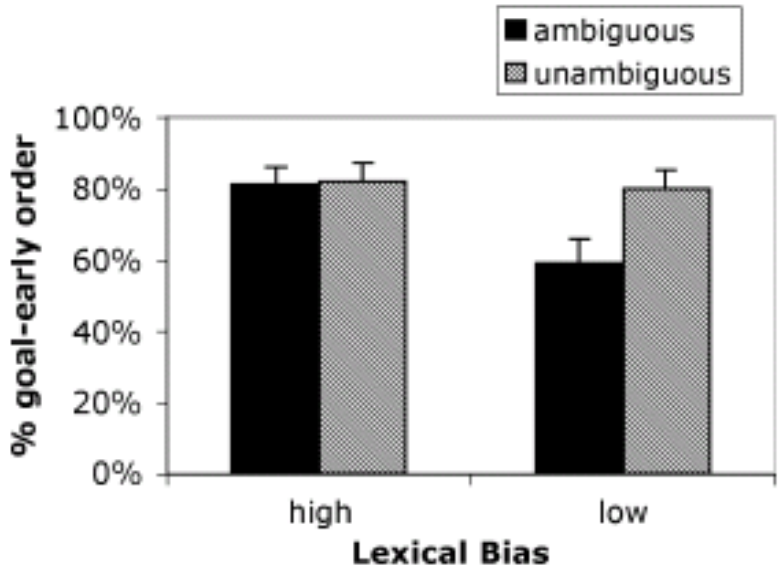


Fig. 1. Experiment 2 results (Lexical Bias × Ambiguity).

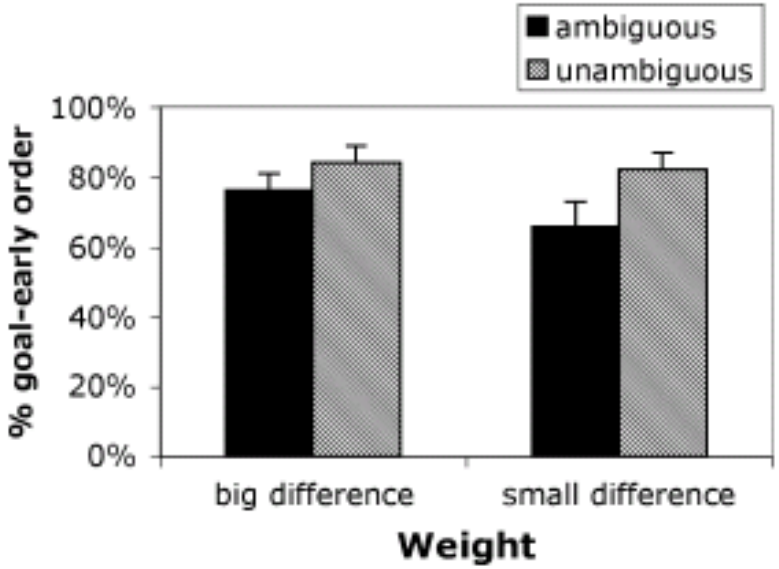


Fig. 2. Experiment 2 results (Weight × Ambiguity).

Garden-path avoidance vs. Ambiguity Avoidance

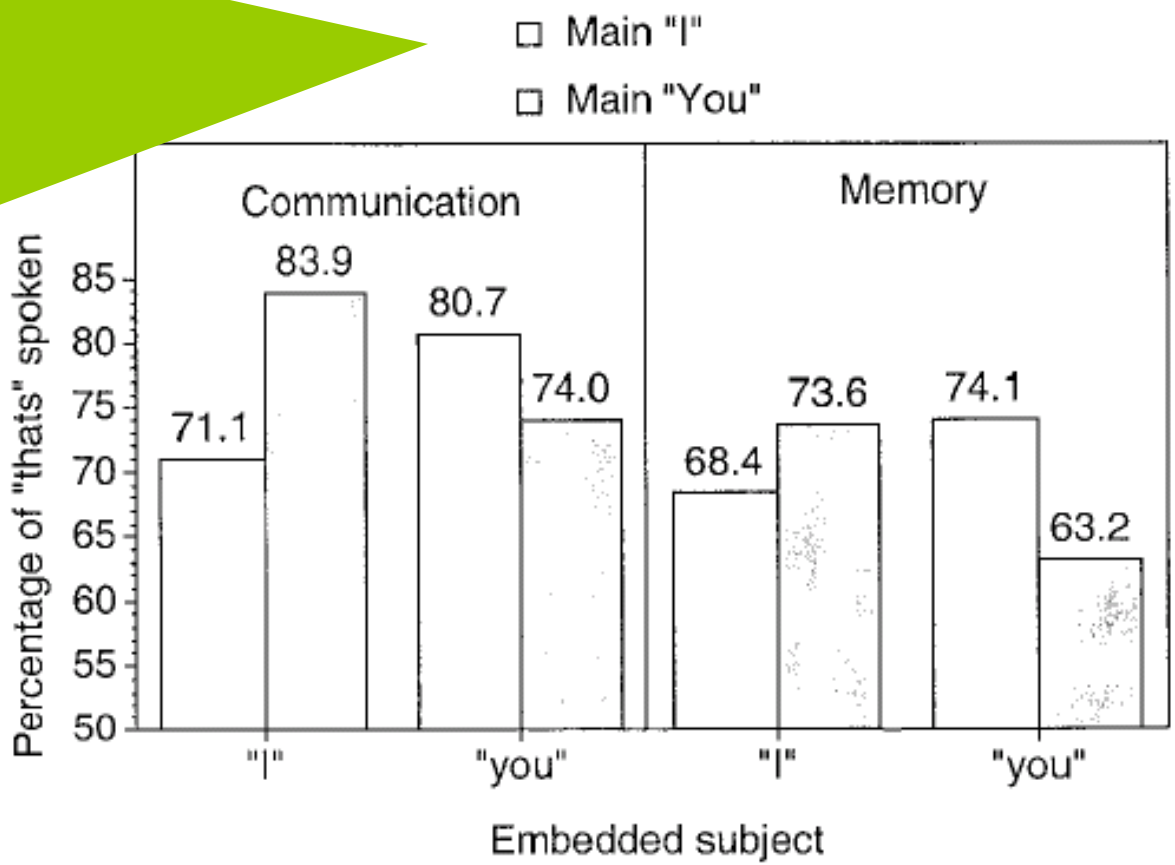
- There is some evidence that speakers **avoid long-lasting ambiguities where contextual biases would mislead comprehenders into garden-paths** (Jaeger, 2006, submitted, in prep)

A different type of Audience Design?

(Ferreira & Dell, 2000:324)

[…] speakers (the *communication* group) looked at and spoke to their addressees when they produced the sentence. Addressees rated the clarity of speakers' productions with pen and paper on a 7-point scale. […] The group of speakers not paired with addressees (the *memory* group) performed the identical task, except that they did not speak to an addressee and no instructions about maximizing clarity were given.

- More complementizers when audience present.



Different types of Audience Design

- Maybe speakers do design their utterances to their audience, but they do not bother to avoid ambiguity, or at least not most:
  - Real ambiguities, as in *cases that can create serious garden paths,* are rare (cf. Jaeger, 2006, submitted; see also **pragmatic ambiguity**, Wasow, 2002)
- Collateral signals (Jaeger, 2005): *that* in complement and relative clauses could *signal* production difficulty (Clark & Fox-Tree, 2002)

Table 6: Model improvement for each of the disfluency measures

	Fillers		Suspension/Restart	
	In NP	In NSRC	In NP	In NSRC
Coefficient in model	-0.02	0.89	-0.2	0.55
Change in -2log-LH	0	19.5	0.4	11.8
Significance level of χ^2	n.s.	$p < 0.001$	n.s.	$p < 0.001$

Different types of Audience Design

- Maybe speakers do design their utterances to their audience, but they do not bother to avoid ambiguity, or at least not most:
  - Real ambiguities, as in *cases that can create serious garden paths,* are rare (cf. Jaeger, 2006, submitted; see also **pragmatic ambiguity**, Wasow, 2002)
- Collateral signals (Jaeger, 2005)

- Also (next Monday): Uniform Information Density (Frank & Jaeger, 2008; Jaeger, 2006, submitted; Levy & Jaeger, 2007)