

# Addressees use Zipf's law a cue for semantics

Freek Van de Velde & Dirk Pijpops (University of Leuven)

(i.c.w. Mathias De Backer, Wendy Feytons, Alek Keersmaekers, Veerle Monserez, Thomas Van Hoey, and Marie Vanwingh)

# George Kingsley Zipf

- Zipf's law: Direct inverse correlation between the relative frequency of a word and its rank in a frequency list:  $F(r_i) = k \times 1/r_i$
- Zipf's law is actually a special case of Mandelbrot's law (just adding some parameters):  $F(r_i) = k \times 1/(\beta + r_i)^\alpha$
- These parameters reveal interesting aspects of the structural make-up of languages (morphological complexity), see Bentz et al. (2014, 2015)



# Zipf's law

- But there is more!

- Frequency  $\sim 1 / \text{phonetic size}$
- Frequency  $\sim 1 / \text{lexical diversity}$ :

$$a \times b^2 = k$$

- a: number of word types in frequency class
  - b: index of a frequency class
  - k: constant
- Frequency  $\sim \text{meaning}$  (difficult to operationalize – e.g. polysemy)
  - Frequency  $\sim \text{age}$  (difficult to operationalize)

Table 1  
Distributional data on German, adapted from Zipf (1965a, p. 23)

Phonetic size index (number of syllables in EU)	Total of discourse occurrences of EUs with the phonetic size index specified in column 1
1	5 426 326
2	3 156 448
3	1 410 494
4	646 971
5	187 738
6	54 436
7	16 993
8	5 038
9	1 225
10	461
11	59
12	35
13	8
14	2
15	1

Table 4  
Discourse frequency:diversity in the Latin of Plautus

b	a	k
1	5429	5429
2	1198	4792
3	492	4428
4	299	4784
5	161	4025
6	126	4536
7	87	4263
8	69	4416
9	54	4374
10	43	4300
11	44	5324
12	36	5184
13	33	5577

# Zipf's law(s)

## Zipf's basic empirical correlations

---

### Properties of individual EUs

---

Frequency

Frequent in discourse

Infrequent in discourse

Phonetic size

Small in phonetic size

Large in phonetic size

Semantic scope

Large semantic scope

Small semantic scope

Age

Old

Young

---

(Pustet 2004: 11)

# What is the explanation behind Zipf's law(s)?

- Zipf: "Principle of least effort"
- Driven by speakers:
  - "Signal simplicity naturally benefits the speaker more than the listener, as it limits the complexity of the task of physically producing an utterance by reducing the number, length, and difficulty of the units to be articulated. Obviously, though, continued erosion of the units of expression will ultimately render the listener's task more difficult." (Langacker 1977: 105, cited in Pustet 2004: 13)
- Diachronic component:
  - iron horse phenomenon
  - grammaticalization
- Question: Do speakers use short words more often, or do they shorten their long words?
- Culture > semantics > frequency > size (Pustet 2004: 19-22)

# What is the explanation behind Zipf's law(s)?

- Is Zipf's law really the speaker's sole responsibility?

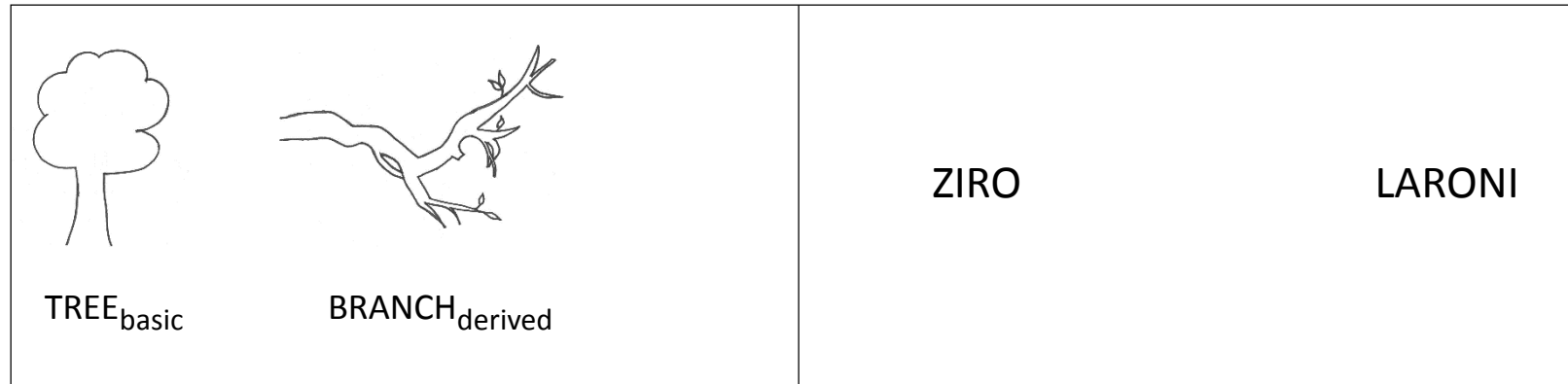
"One likely mechanism for how the lexicon comes to reflect predictability is that information content is known to influence the amount of time speakers take to pronounce a word: words and phones are given shorter pronunciation in contexts in which they are highly predictable or convey less information (...). If these production patterns are lexicalized, word length will come to depend on average informativeness." (Piantadosi et al. 2011: 3528)

Our hypothesis:

Zipf's size-meaning tendency is not only under evolutionary selection by speakers, in their attempt to minimize articulatory effort, but benefits addressees as well, who can use this tendency as a cue: through their life-time experience with language, they know that in general, shorter words have more unmarked meanings, and they apply this implicit knowledge when they are confronted with a new language, when other cues are absent.

# Research design

- Present test subjects with:
  1. A pair of visual stimuli, one of which is semantically more 'basic' than the other
  2. A pair of fake-language verbal stimuli, one of which is longer than the other



- Ask them to link the visual stimuli to the verbal stimuli
- We don't want the test subjects to be conscious of the differences
  1. Use filler items
  2. Use visual stimuli that differ in subtle ways
  3. Ask subjects afterwards what heuristic (if any) they applied

# Research design

- Semantic differences: difficult to operationalize anyway
- They have to be subtle, so that test subject are not immediately aware of what they are tested on.





















# Research design

- Semantically subtle differences in stimuli
  1. Urban 2011: asymmetries in lexical-semantic changes across languages
    - E.g. *bark* <- *skin*, but not: *skin* <- *bark* So: *tree-skin* vs. \**man-bark*
    - E.g. *testicles* <- *eggs*, but not: *eggs* <- *testicles*.
    - E.g. *lungs* <- *liver*, but not: *liver* <- *lungs* (no intuitions)
  - Winter et al. 2013: robust on a wide range of tests: Wikipedia-links, reaction times, word associations, dictionaries etc. (to see whether the asymmetries can best be explained by frequency, cognitive accessibility)
  2. Berlin-Kay color hierarchy (dist > 2)















# Research design

- Visual stimuli (pretested)

field	basic	derived
artefact	CAR 	TRAIN 
artefact	SHADOW 	MIRROR 
color	YELLOW 	PURPLE 
color	GREEN 	GRAY 
color	BLACK 	BLUE 
body	BREAST 	MILK 
body	HEART 	STOMACH 
nature	TREE 	BRANCH 
nature	SUN 	MOON 

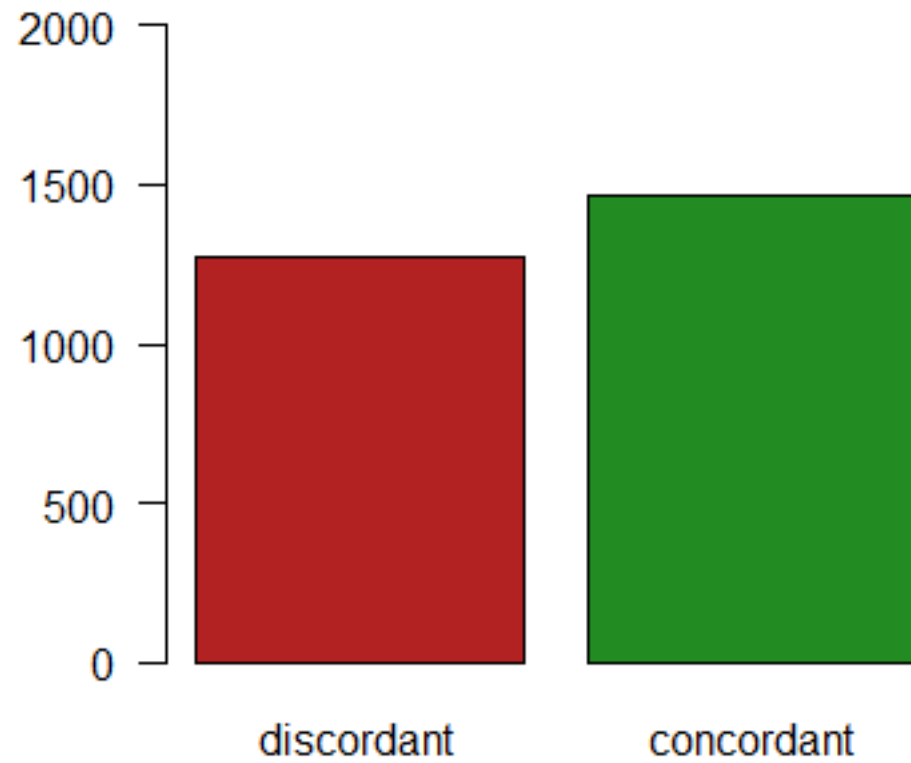
- Visual fillers

basic	derived	
NOSE 	FISH 	} non-Urban, same length (1,2,3)
ORANGE 	PINK 	
RED 	MOUNTAIN 	
HOUSE 	TABLE 	} Urban, same length (1,2)
MOUTH 	LIP 	
TOOTH 	CHEEK 	

⇒ only pairs where the basic term was **longer or equal in length** to the derived term in Dutch

# Research design

- Verbal stimuli:
  - CV syllable concatenations
  - basic vowels (i – o – a) (u excluded because it is a front round vowel grapheme in Dutch, but back round vowel in most other languages)
  - no associations with Dutch words
  - difference in syllable length (1-2, 2-3, 1-3)
  - verbal stimuli were not randomized over the visual stimuli



$p < 0.001$   
effect size for proportions ( $2 * \arcsin(\sqrt{x})$ ): 0.07  
power = 0.84

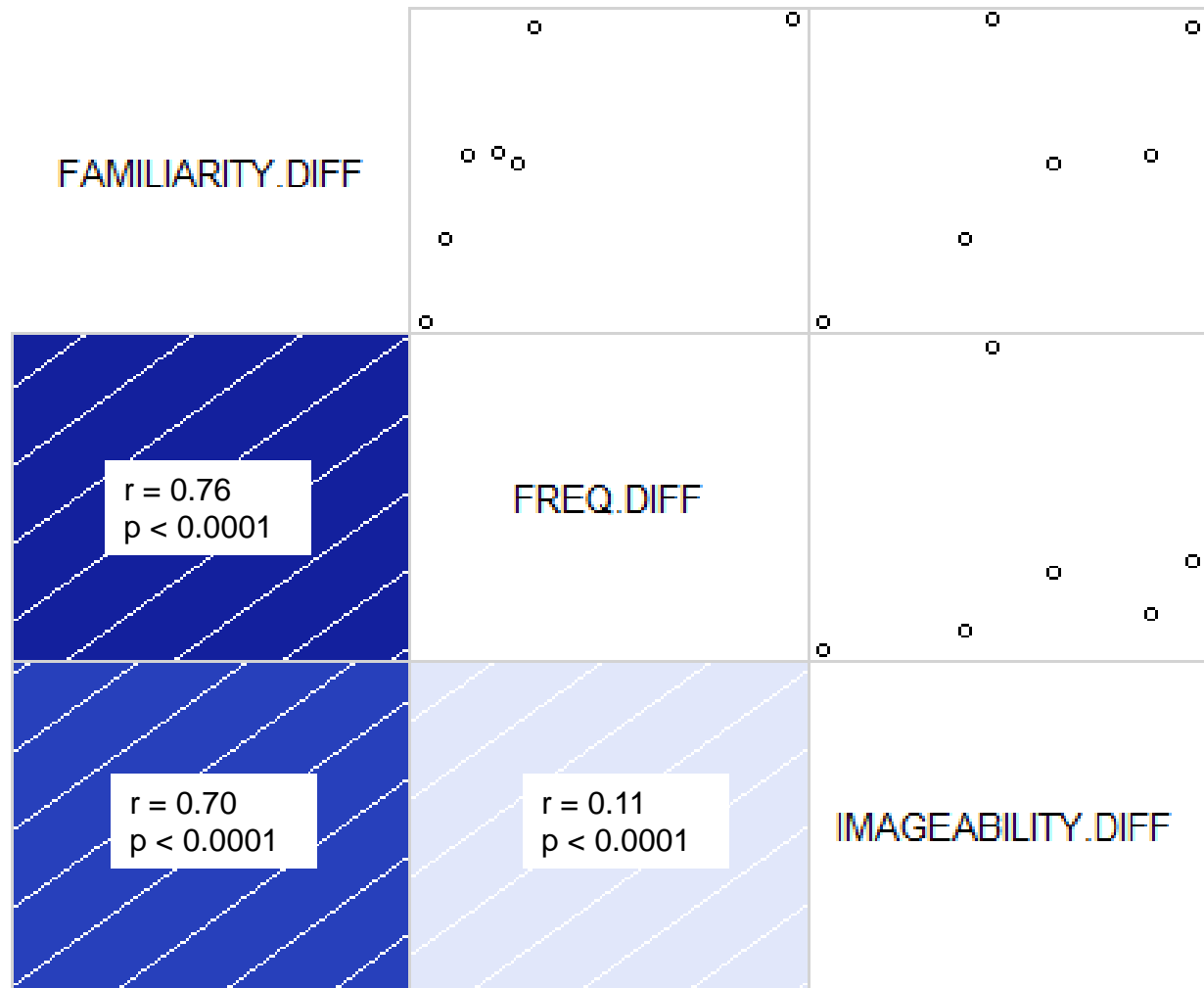
# Explanatory variables

- stimulus type: *color vs. Urban pairs*
- difference in the number of syllables of the verbal stimulus: 1, 2
- difference on psycholinguistic measures (MRC Psycholinguistic Database, Wilson 1988)
- difference between the frequency of use in spoken language (Corpus of Spoken Dutch)

	car	train	shadow	mirror	yellow	purple	green	gray	black	blue	breast	milk	heart	stomach	belly	tree	branch	sun	moon
Familiarity	634	548	536	593	555	-	583	531	603	593	555	588	578	547	486	613	529	635	585
Imageability	638	539	565	627	598	-	609	541	589	569	597	638	617	551	576	622	548	639	585
Frequency	2757	1254	166	191	669	136	756	325	979	683	227	194	650	283	848	198	767	166	

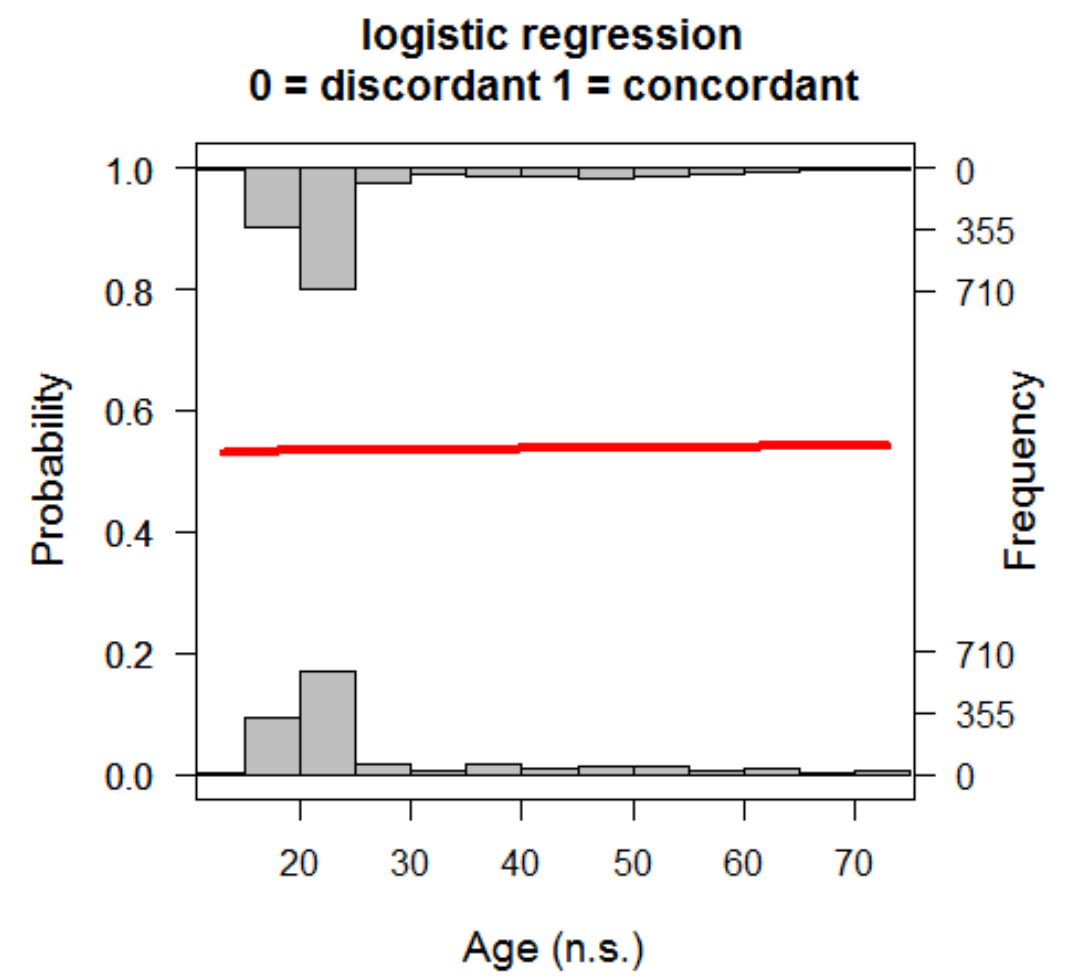
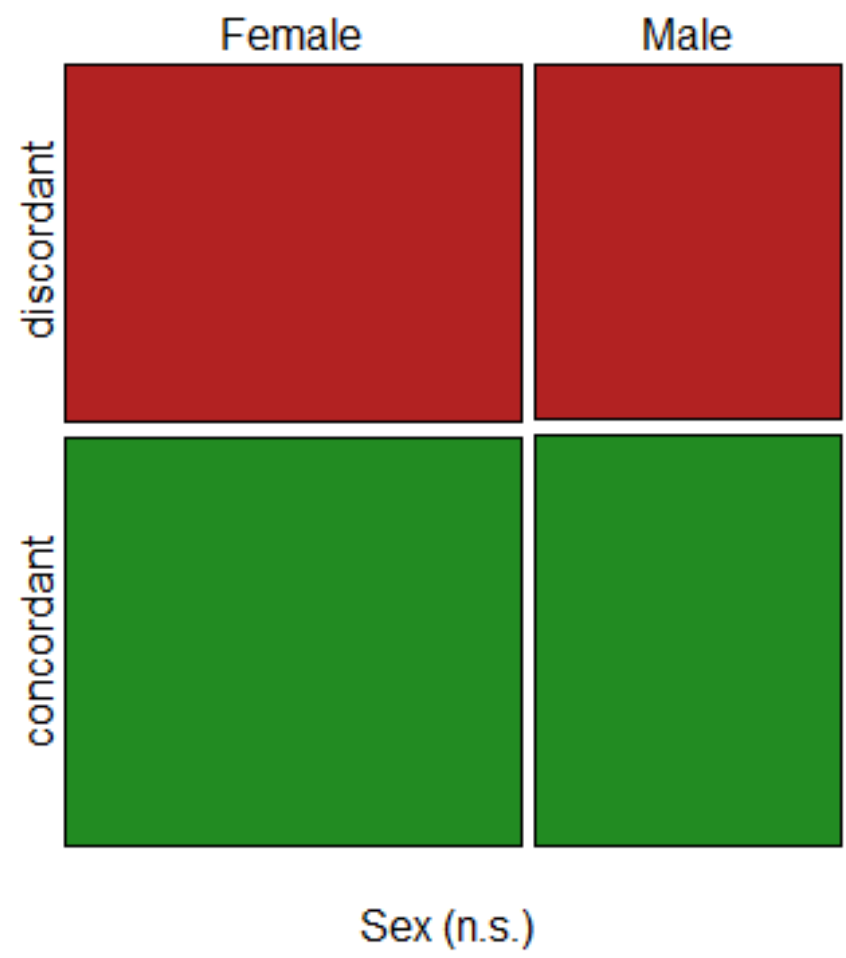
- Leave out SHADOW/MIRROR and BREAST/MILK
- Difference in familiarity is best measure

# Correlogram psycholinguistic measures



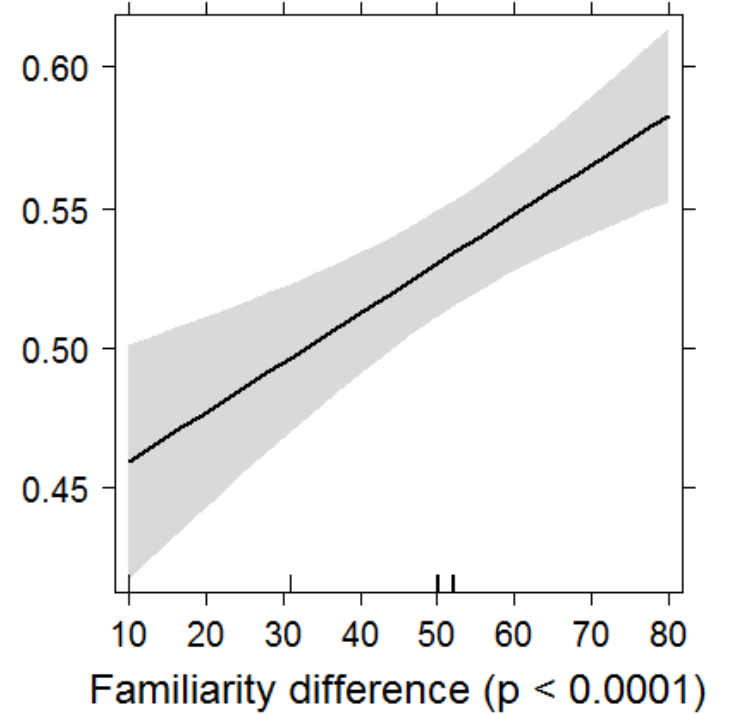
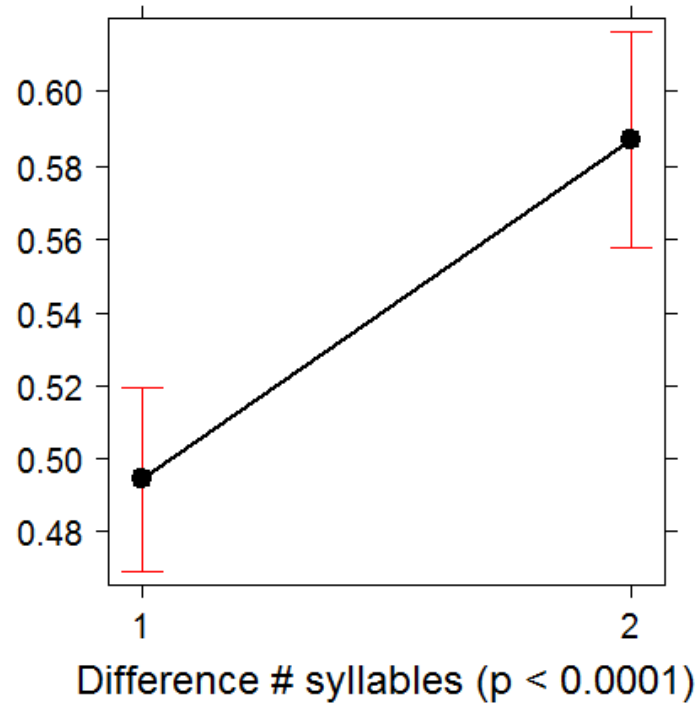
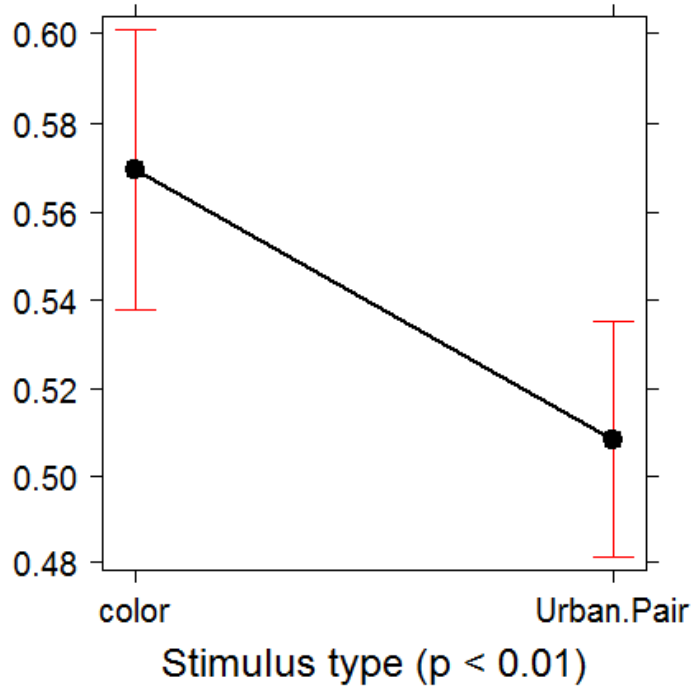
# Survey

- LimeSurvey
- Via social media
- 464 respondents
- Retaining only
  - native speakers of Dutch
  - respondents who didn't say they applied Zipf's law as a heuristic
- N = 395
- color stimuli: controlled for colorblindness (n = 24)

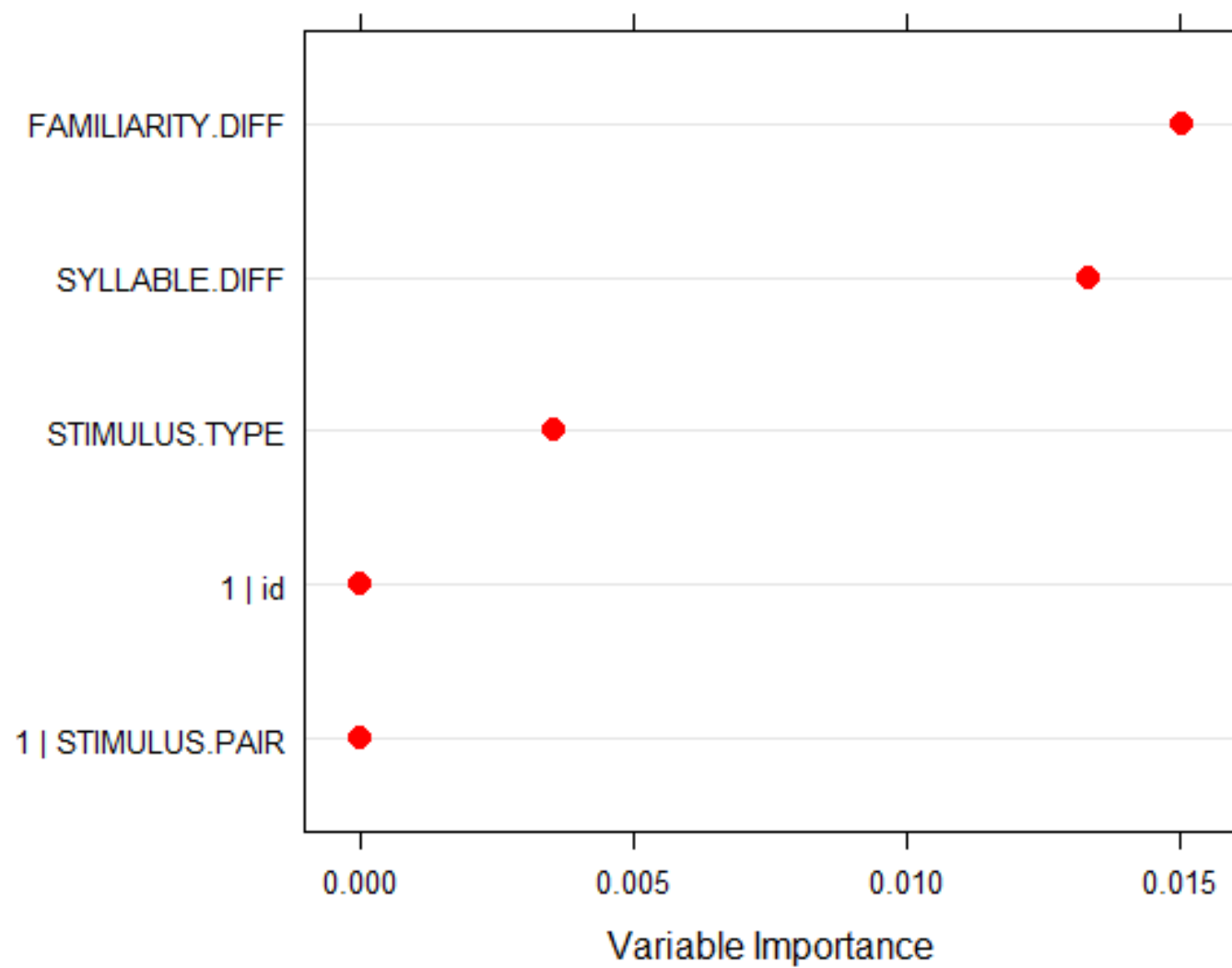




Mixed-effect logistic regression (crossed random intercepts for stimuli and test subjects)  
Effect plots: fitted probability of a concordant response



## Random forest



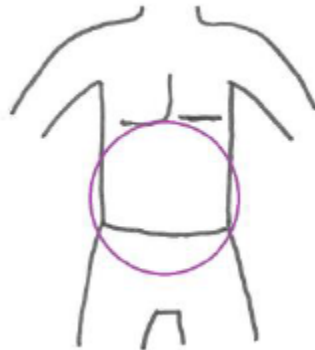
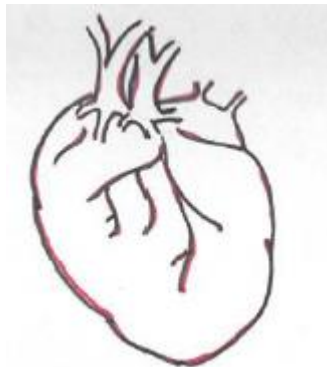
# Discussion

- Zipf's size-meaning correlation is generally considered as the result of a shortening of words
- This shortening is beneficial to the speaker, who avoids unnecessary energy expenditure in expressing.
- The traditional view is that this inclination of the speaker is detrimental for the hearer, who strives for clarity of expression by maximal distinctiveness.
- However, Zipf's size-meaning correlation also holds a benefit for the hearer, as the predictable nature of this tendency can be capitalized on in decoding the signal.
- Our results show that this is indeed what hearers do, even when the differences between the words are subtle enough to stay below the level of consciousness.
- With more obvious differences between the visual stimuli, the effect size becomes bigger, as the results in Lewis & Frank (subm.) show.



# Shortcomings and future work

- Are the test subjects really addressees? You might say that they act as speakers, picking a word for a concept. It might be a good idea to see whether we get the same results with for example eyetracking.
- What other factors are at play?
  - Formality: formal words tend to be shorter. You could try to manipulate the context of use
  - Iconicity: length / sound symbolism
  - Visual stimuli: our 'worst' pair was HEART/STOMACH, but the visual stimuli differed in anatomical detail



## References

- Bentz, C., D. Kiela, F. Hill & P. Buttery. 2014. Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory* 10(2): 175-211.
- Bentz, C., A. Verkerk, D. Kiela, F. Hill & P. Buttery. 2015. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE*.
- Langacker, R.W. 1977. Syntactic reanalysis. In Ch.N. Li (Ed.), *Mechanisms of syntactic change*. Austin: University of Texas Press. 57-139.
- Lewis, M.L. & M.C. Frank. Submitted. The length of words reflects their conceptual complexity.
- Piantadosi, S.T., H. Tily & E. Gibson. 2011. Word lengths are optimized for efficient communication. *PNAS* 108(9): 3526-3529.
- Pustet, R. 2004. Zipf and his heirs. *Language Sciences* 26: 1-25.
- Urban, M. 2011. Asymmetries in overt marking and directionality in semantic change. *Journal of Historical Linguistics* 1: 3-47.
- Wilson, M. 1988. MRC psycholinguistic database: Machine-usable dictionary. Version 2.00. *Behavior Research Methods, Instruments, & Computers* 20(1): 6-10.
- Winter, B., G. Thompson & M. Urban. 2013. Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure. In: E. A. Cartmill, S. Roberts, H. Lyn, & H. Cornish (eds.), *10th International Conference on the Evolution of Language*. New Jersey: World Scientific. 353-360.
- Zipf, G.K. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard: Harvard University Press.
- Zipf, G.K. 1935. *The Psycho-biology of Language. An Introduction to Dynamic Philology*. Boston: Houghton Mifflin Company.
- Zipf, G.K. 1949. *Human behavior and the principle of least effort*. Cambridge (Mass.): Addison-Wesley.