Semantic entropy measures and the semantic transparency of noun noun compounds

Melanie J. Bell, Martin Schäfer

Anglia Ruskin University, Friedrich-Schiller-Universität Jena

Saarbrücken, 08.03.2017

Why look at semantic transparency?

- Semantic transparency plays important role in storage and processing of compound words (e.g. Libben et al. 2003)
- BUT: Semantic transparency itself is still poorly understood!
- Our questions:
  Can semantic transparency be analysed in terms of semantic entropy measures?
  What happens if, in addition to entropy measures for semantic relations, entropy measures for word senses are considered?

Semantic aspects of interest

- The semantic relation between a compound's constituents
- The senses of the constituents

Consider the N1 constituent family of *bank account*:

(1)

	relation	example
a.	IN	*bank account*
b.	FROM	*bank charge*
c.	FOR	*bank manager*
d.	...	...

(2)

	sense	example
a.	$bank_1$	*bank barn*
b.	$bank_2$	*bank clerk*
c.	$bank_3$	*bank switch*
d.	...	...

Previous studies

- Pham & Baayen (2013): entropy of semantic relations in the modifier family, relative to the lexicon as a whole, is negatively correlated with semantic transparency
- Schmidtke et al. (2015): relational entropy for individual compounds, based on the range of relations assigned to them by raters, is positively correlated with reaction time in lexical decision
- Bell & Schäfer (2016): N1 relation proportion correlates positively with semantic transparency, N2 synset proportion negatively

Which semantic entropy measures?

- Relation entropy: given a compound, how are the probabilities for specific semantic relations distributed over the constituent families
- Synset entropy: given a compound, how are the probabilities for specific readings of its constituents distributed over the corresponding constituent families

Our transparency ratings: the Reddy et al. data

- 90 compounds from the ukWaC corpus
- Transparency ratings for the whole compound, the modifier and the head collected using Amazon Turk
- 30 ratings for each task for each compound
- 2415 tokens for the whole compound

Calculating the semantic entropy measures

1. Used the annotated compound family database from Bell & Schäfer (2016)
   - Took all strings of exactly 2 nouns that follow an article in the BNC
   - Extracted constituent families for our compounds
   - Added unspaced binominal compounds from CELEX
   - Selected only those items which occur at least 5 times in the USENET corpus (Shaoul & Westbury 2010)
   - Yielding 4553 types in the N1 positional families and 9226 types in the N2 positional families
   - Coded these types for the semantic relation (after Levi 1978), and for the WordNet senses of the constituents (Princeton 2010)
2. Calculated N1 and N2 synset and relation entropies using the distribution in the corresponding families

Predictors

- Logarithmetised constituent frequencies
- Compound spelling ratio (Bell & Plag 2012)

(3) $$\text{spelling ratio} = \frac{(\text{unspaced frequency} + \text{hyphenated frequency})}{\text{spaced frequency}}$$

- N1 synset entropy, N2 synset entropy
- N1 relation entropy, N2 relation entropy

(4) $$H = -\sum_{i=1}^{n} p_i \log p_i$$

- All predictors were centered

Final model for compound transparency

Random effects:

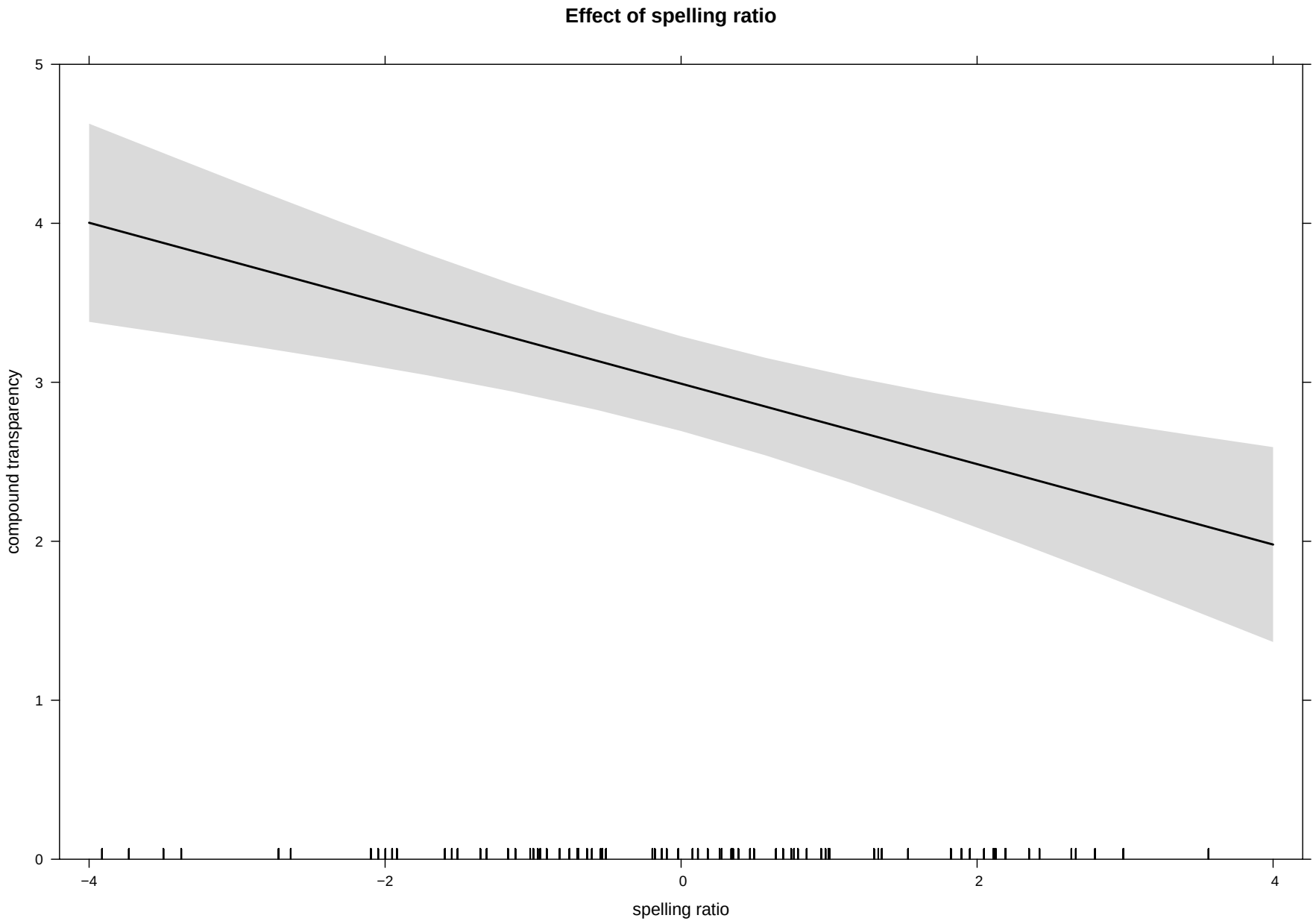
Groups	Name	Variance	Std.Dev.	Corr
rater	(Intercept)	0.151629	0.38940	
	spellingRatioCentred	0.004015	0.06336	0.91
item	(Intercept)	1.277178	1.13012	
residual		0.930239	0.96449	

Number of obs: 2307, groups: rater, 119; item, 81

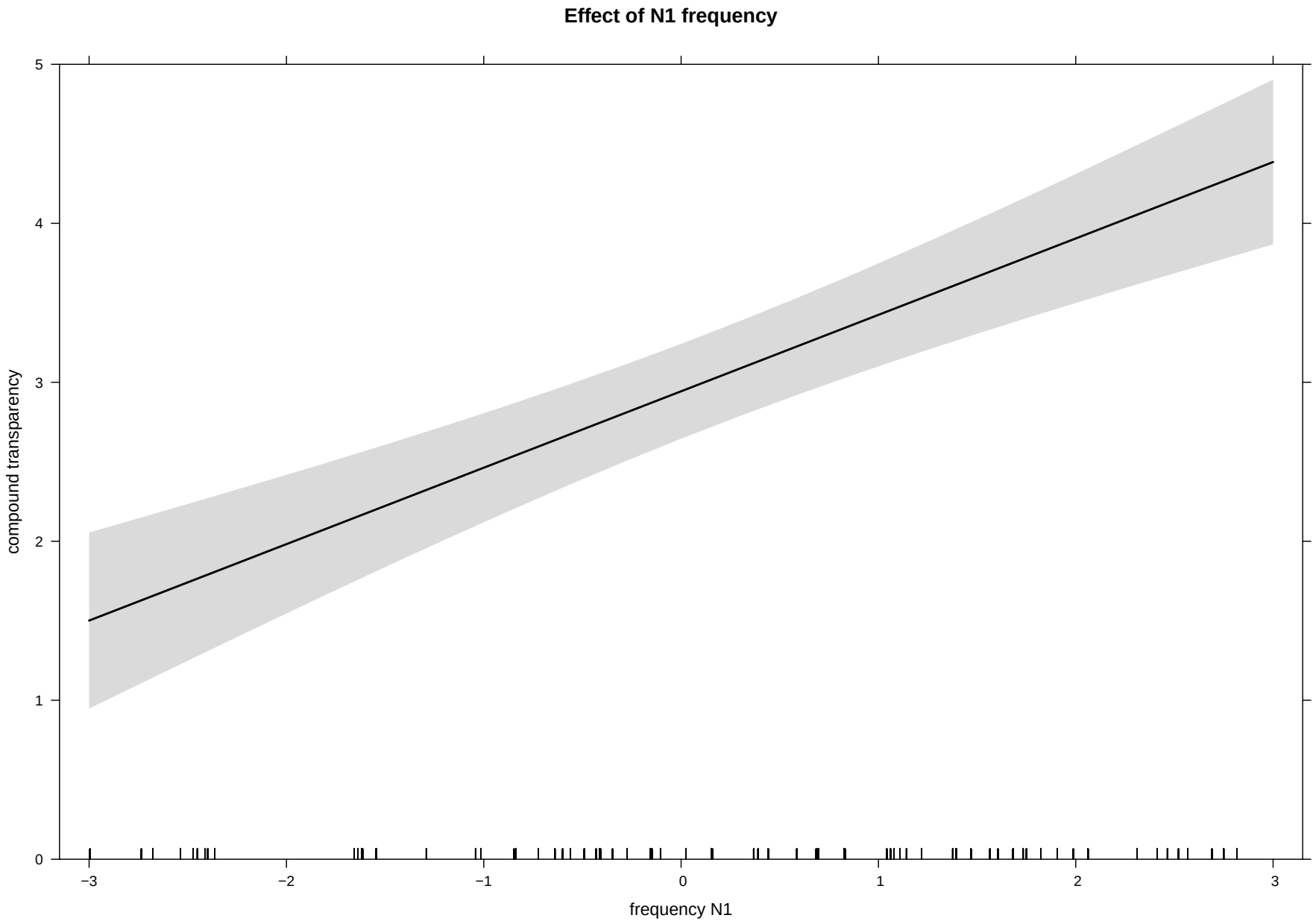
Fixed effects:

	Estimate	S.E.	df	t	Pr($>\|z\|$)
(Intercept)	2.96069	0.15224	88.64	19.448	$<$ 2e-16
spelling ratio	-0.25302	0.06904	76.69	-3.665	0.000454
N1 frequency	0.48070	0.07571	74.97	6.349	1.5e-08
N2 synset entropy	0.34824	0.18259	74.97	1.907	0.060317
N2 relation entropy	-0.15070	0.24023	75.01	-0.627	0.532353
interaction synset/relation entropy	-0.57240	0.28352	74.99	-2.019	0.047070

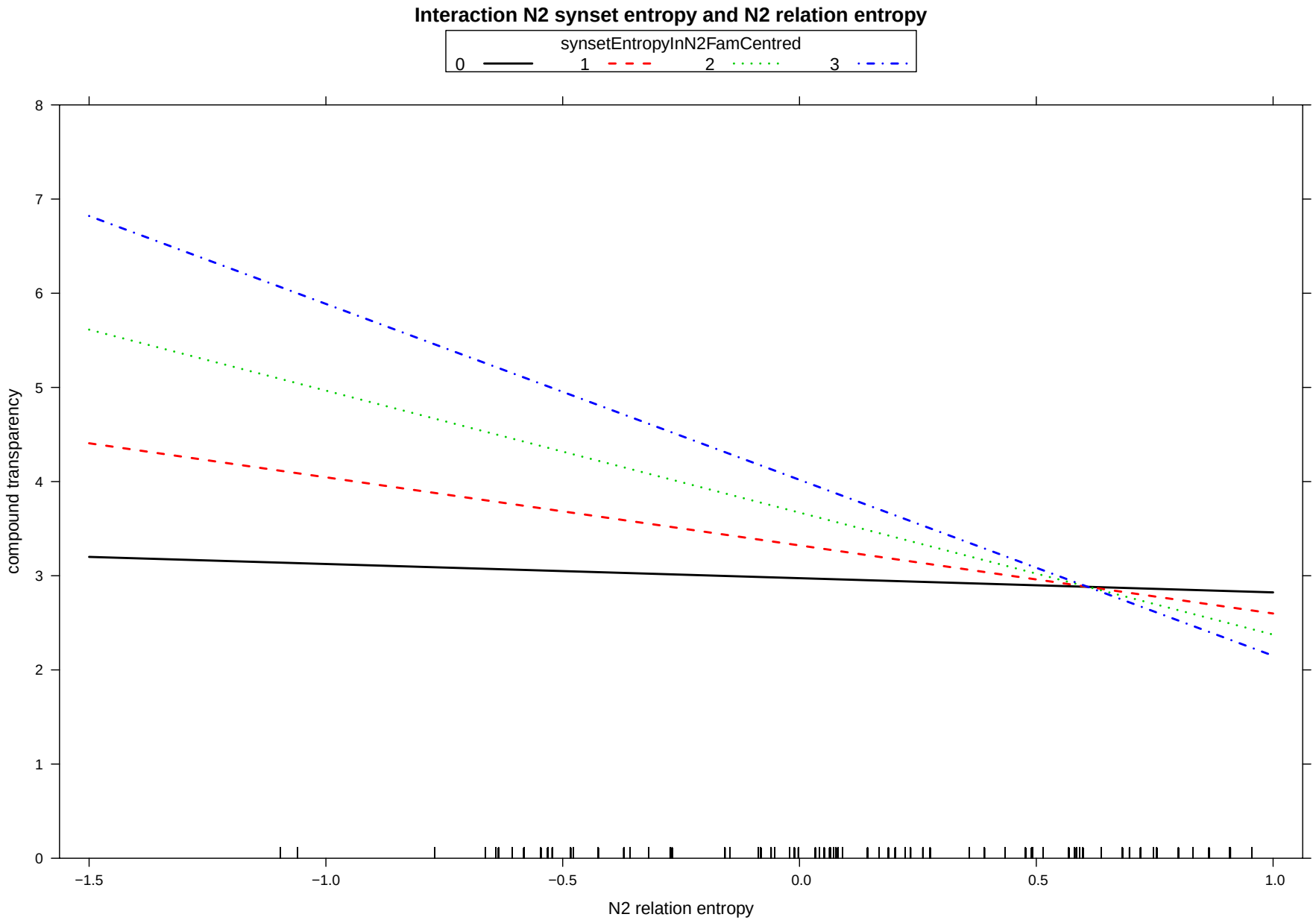
The effect of spelling ratio



The effect of N1 frequency



The Interaction N2 synset entropy and N2 relation entropy



Interpreting the effects: the main effects

- The two main effects of spelling ratio and N1 frequency mirror the effects found in Bell & Schäfer 2016
  - Positive correlation with N1 frequency: reflex of expectedness
  - Negative correlation with spelling ratio: operationalising lexicalisation, reflex of non-compositional interpretations

Interpreting the effects: the interaction

Relation entropy not negatively correlated with semantic transparency across the board

- Low N2 synset entropy:
  N2 relational entropy does not make much of a difference
- High N2 synset entropy:
  N2 relation entropy correlates negatively with compound transparency (mirroring Pham & Baayen's 2013 finding for modifier families)
- The most transparent compounds have high synset entropy but low relation entropy

Conclusions

- First evidence that the interaction of entropy measures based on the head families plays a role for perceived transparency
- Overall further evidence for differentiated roles of modifier and head in compound processing
- Target compounds are all high frequent compounds, exploration of these measures on less frequent compounds is a task yet to be done.

Thank you!