Compound stress, informativity and semantic transparency

Melanie J. Bell

Anglia Ruskin University

26th European Summer School in Logic, Language and Information
Tübingen, August 11-22, 2014
The problem: stress in NN compounds

crédit card  silk shírt
táble lamp  kitchen sínk
súmmer school  summer dréss
tóy factory  toy fáctory
Assign primary stress to a primary-stressed vowel in the context

\[ \_ \quad \ldots \, V^1 \ldots \, ]_N \]

Assign primary stress to a primary-stressed vowel in the context

\[ V^1 \quad \_ \quad \ldots \, ]_{NP} \]
The problem remains of determining under what syntactic conditions this feature [which overrides the Compound Rule] is or is not present... it can be clarified and resolved only by an investigation of the conditions, syntactic and other, under which the Compound Rule is applicable. For this reason, we will make no attempt to go more deeply into the question here... we will simply leave this question in its present unsatisfactory state.

(Chomsky and Halle 1968: 156)
Potential determinants of compound stress assignment

- structure
- semantics
- analogy based on constituents
Structural hypothesis
(e.g. Giegerich 2004)

- modifier-head structures are regularly stressed on the right constituent (*steel bridge*)

- argument-head structures are always left-stressed (*ópera singer*)

- left stress on modifier-head structures is due to lexicalisation (*ópera glasses*)
Semantic hypothesis
(e.g. Fudge 1984)

• certain semantic relations are right-stressed
  (e.g. ‘locative’ compounds, Boston hárbour)

• certain semantic classes of constituents trigger right stress
  (e.g. substance nouns as N1, silk shírt)

• lexicalised semantics goes together with left stress
  (e.g. sílk worm)
Analogical hypothesis
(e.g. Schmerling 1971, Plag 2006)

• stress assigned by analogy with similar compounds in the mental lexicon:

  Óxford Street    Oxford Road
  Régent Street    Mill Road
  Hárley Street    Upland Road
  ... Street       ... Road

  100 % left       0 % left

• 'constituent family stress bias'
Empirical evidence for determinants of stress pattern

• studies using different types of data, different varieties of English, different types of analytical tools
• most significant factor: constituent family bias
• other significant factors: semantics, lexicalisation
• no evidence for structural hypothesis
• effects are probabilistic rather than categorical
• nature of analogical effects unclear
Questions remaining

• Why do particular constituents favour one pattern or the other?

• Why are certain semantic relations associated with right stress?
What is compound stress?

(Gussenhoven 2004, Kunter & Plag 2007, Kunter 2011)

- left prominence/left stress:
  one pitch accent on left constituent

- right prominence/right stress:
  two pitch accents one on each constituent
In nineteen seventy-six, Democratic Governor Michael Dukakis fulfilled a campaign promise to de-politicize judicial appointment.

from: *Boston University Radio Speech Corpus* (BURSC, Ostendorf et al. 1995)
Two pitch accents

When a computerized call is made to a former prisoner's home phone, that person answers by plugging in the device.

from: Boston University Radio Speech Corpus (BURSC, Ostendorf et al. 1995)
Constituent effects: length hypothesis
(e.g. Jespersen 1909)

• longer compounds tend to be right-stressed

• cross-linguistic dispreference for long strings of unaccented syllables (Ladd 2008)
Constituent effects: informativity hypothesis
(e.g. Sweet 1892)

“In some compounds the uneven [left] stress seems to be the result of the second element being less logically prominent than the first, through being a word of general meaning and frequent occurrence in compounds” (Sweet 1892:288)

• uninformative elements tend to be unaccented, while more informative information is accented
Measuring informativity 1

general meaning

• semantic specificity: number of different senses (= number of synsets of N1 and N2 in WordNet)

• hypothesis: the greater the number of synsets, the less specific the constituent, the less likely to be accented
Measuring informativity 2

**frequent occurrence** in compounds

- probability of N2 (= log of **N2 family size**)
  
hypothesis: the larger N2 family size, the more probable is N2, the less likely it is for N2 to be accented

- conditional probability of N2
  (= log of **1/N1 family size**)
  
hypothesis: the more probable N2 **given N1**, the less likely it is for N2 to be accented

*cf. Shannon 1948*
Study 1: BNC (Bell & Plag 2012)

- compounds from BNC, spoken in an experiment and rated by experts (N = 3252, V = 864, 60 % left stressed)

- coded for pertinent predictor variables
  - length in syllables
  - informativity as expectedness
  - informativity as semantic specificity
BNC: significant predictors unrelated to informativity

- Temporal
- Locative
- Made of
- Listed in dictionary
- Log spelling ratio
- Syllables after N1 main stress
BNC: informativity effects

N2 is highly specific
N1 is unspecific
Study 2: BURSC (Bell & Plag 2013)

• subset of dataset from Plag et al. (2008)
  N = 1154, V = 592, stress judged by two experts

• coded for:
  – probability and conditional probability of N2
  – synset counts of N1 and N2
  – NN frequency (lexicalisation)
  – length of NN after main stress of N1
  – no semantics

• generalised mixed effects regression analysis (speaker as random effect)
BURSC: effects of informativity 1
BURSC: effects of informativity 2
(probitilities not included in the initial model)
Predicting N1 family bias from N1 properties

Final model, adjusted R-squared=0.3007
Predicting N2 family bias from N2 properties

Final model, adjusted R-squared = 0.1071
Questions remaining

• Why do particular constituents favour one pattern or the other?

• Why are certain semantic relations associated with right stress?
Relational effects: compositionality hypothesis

in NNs with ‘even stress’ (i.e. right prominence): the logical relation [R] between the elements of a compound resembles that between the elements of a free group [a phrase], especially when the first element is felt to be equivalent to an adjective .......
even stress balances as it were the two elements against one another and puts them on a footing of equality and to some extent separates them. (Sweet 1892:288-9)
a *summer-house* is not merely a house inhabited in summer, but a house of a particular style and construction ... a *summer residence*, however, is merely (someone’s) residence in the summer, nothing more. Marchand (1969:25)
Semantic coding (Bell 2013)

Relations
• trained lexicographer
• criterion of entailment

Categories
• hypernyms in wordnet
## Predictors initially present in the semantic analysis

<table>
<thead>
<tr>
<th>N1 semantic classes</th>
<th>semantic relations entailed between N1 and N2</th>
<th>lexicalisation of NN</th>
<th>other factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>material</td>
<td>N2 is (made of) N1</td>
<td>listed in dictionary (OED)</td>
<td>N1 compound</td>
</tr>
<tr>
<td>time or place</td>
<td>N2 is at/on/in N1</td>
<td>occurs with non-spaced orthography in BNC</td>
<td>N2 compound</td>
</tr>
<tr>
<td>social group</td>
<td>N1 has N2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>adjective-like</td>
<td>NN is name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>name</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Wordnet synsets for N1 classes associated with right prominence

<table>
<thead>
<tr>
<th>N1 semantic class</th>
<th>Wordnet synset</th>
<th>Wordnet definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>material</td>
<td>substance (sense 1)</td>
<td>that which has mass and occupies space</td>
</tr>
<tr>
<td></td>
<td>fabric (sense 1)</td>
<td>cloth, material, space</td>
</tr>
<tr>
<td></td>
<td>building material</td>
<td>material used for constructing buildings</td>
</tr>
<tr>
<td>time or place</td>
<td>location (sense 1)</td>
<td>a point or extent in space</td>
</tr>
<tr>
<td></td>
<td>time period</td>
<td>period of time</td>
</tr>
<tr>
<td>social group</td>
<td>social group</td>
<td>people sharing some social relation</td>
</tr>
<tr>
<td></td>
<td>social class</td>
<td>socio-economic class</td>
</tr>
</tbody>
</table>
# Semantic predictors of right stress

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>S.E.</th>
<th>Wald Z</th>
<th>P</th>
<th>Penalty Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td>-2.443</td>
<td>0.227</td>
<td>-10.76</td>
<td>0.000</td>
<td>0</td>
</tr>
<tr>
<td>N1isTimePlace=yes</td>
<td>1.478</td>
<td>0.468</td>
<td>3.16</td>
<td>0.0016</td>
<td>0.3162</td>
</tr>
<tr>
<td>N1isSocial=yes</td>
<td>2.225</td>
<td>0.495</td>
<td>4.5</td>
<td>0.000</td>
<td>0.3162</td>
</tr>
<tr>
<td>N1isAdjectivelike=yes</td>
<td>2.237</td>
<td>0.6312</td>
<td>3.54</td>
<td>0.0004</td>
<td>0.3162</td>
</tr>
<tr>
<td>N2isN1=yes</td>
<td>4.154</td>
<td>0.3848</td>
<td>10.79</td>
<td>0.000</td>
<td>0.3162</td>
</tr>
<tr>
<td>N2OnAtInN1=yes</td>
<td>2.498</td>
<td>0.418</td>
<td>5.98</td>
<td>0.000</td>
<td>0.3162</td>
</tr>
<tr>
<td>N1hasN2=yes</td>
<td>2.028</td>
<td>0.6236</td>
<td>3.25</td>
<td>0.0011</td>
<td>0.3162</td>
</tr>
<tr>
<td>NNisName=yes</td>
<td>2.624</td>
<td>0.9218</td>
<td>2.85</td>
<td>0.0044</td>
<td>0.3162</td>
</tr>
<tr>
<td>occursNonSpaced=yes</td>
<td>-1.018</td>
<td>0.3671</td>
<td>-2.77</td>
<td>0.0056</td>
<td>0.3162</td>
</tr>
<tr>
<td>N1compound=yes</td>
<td>1.146</td>
<td>0.5053</td>
<td>2.27</td>
<td>0.0233</td>
<td>0.3162</td>
</tr>
<tr>
<td>N2compound=yes</td>
<td>3.716</td>
<td>0.9402</td>
<td>3.95</td>
<td>0.0001</td>
<td>0.3162</td>
</tr>
</tbody>
</table>

Model C = 0.936
Hypothesis

Right stress is associated with compositional (phrase-like) semantics
Reddy et al. 2011

- aim: to predict human ratings of compound compositionality
- equated compositionality with literality
Reddy et al. data

• 90 compounds from the ukWaC corpus
• 2 definitions of each compound
• 5 randomly selected example sentences containing the compound
• first: choose most frequent definition
• second: rate compound, first or second constituent for literality
• 30 ratings for each task for each compound
Data selected
(Bell & Schäfer 2013)

• items for which the same rater had performed all three tasks
• 1337 tokens, each with chosen definitions
Modelling compositionality: predictors

• metaphorical and metonymical shifts in meaning of the first or second constituent, or the compound as a whole
• relation between constituents based on Levi (1978)
• frequencies of the constituents
• ‘spelling ratio’: proportion of tokens (in BNC) that are written unspaced
Compound relations: Levi 1978

- CAUSE tear gas
- HAVE picture book
- MAKE music box
- USE steam iron
- BE pine tree
- IN desert rat
- FOR plant food
- ABOUT oil crisis
## Compound literality

|                        | Coef | S.E.  | t    | Pr(>|t|) |
|------------------------|------|-------|------|----------|
| Intercept              | -0.5861 | 0.3207 | -1.83 | 0.0678   |
| logFreqN1              | 0.2830  | 0.0243 | 11.63| <0.0001 |
| logFreqN2              | 0.1535  | 0.0283 |  5.42| <0.0001 |
| spellingRatio          | -0.1240 | 0.0249 | -4.98| <0.0001 |
| Ametaphor=Yes          | -0.6397 | 0.0939 | -6.82| <0.0001 |
| Bmetaphor=Yes          | -0.4841 | 0.0920 | -5.26| <0.0001 |
| ABmetaphor=Yes         | -1.8411 | 0.0910 | -20.23| <0.0001 |
| In=Yes                 | 0.6041   | 0.1273 |  4.75| <0.0001 |
| For=Yes                | 0.2363   | 0.0882 |  2.68| 0.0074   |

$R^2$ adjusted = 0.459
### Literality constituent A

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>S.E.</th>
<th>t</th>
<th>Pr(&gt;t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.3791</td>
<td>0.3418</td>
<td>-1.11</td>
<td>0.2676</td>
</tr>
<tr>
<td>logFreqN1</td>
<td>0.3406</td>
<td>0.0262</td>
<td>12.99</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>logFreqN2</td>
<td>0.0953</td>
<td>0.0305</td>
<td>3.13</td>
<td>0.0018</td>
</tr>
<tr>
<td>spellingRatio</td>
<td>-0.0674</td>
<td>0.0268</td>
<td>-2.51</td>
<td>0.0122</td>
</tr>
<tr>
<td>Ametaphor=Yes</td>
<td>-1.7234</td>
<td>0.1003</td>
<td>-17.19</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Bmetaphor=Yes</td>
<td>0.8728</td>
<td>0.0987</td>
<td>8.85</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ABmetaphor=Yes</td>
<td>-1.8728</td>
<td>0.0939</td>
<td>-19.95</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>In=Yes</td>
<td>0.9275</td>
<td>0.1344</td>
<td>6.90</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

R² adjusted = 0.499
### Literality constituent B

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>S.E.</th>
<th>t</th>
<th>Pr(&gt;t)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td>1.2383</td>
<td>0.3448</td>
<td>3.59</td>
<td>0.0003</td>
</tr>
<tr>
<td><strong>logFreqN1</strong></td>
<td>0.1224</td>
<td>0.0259</td>
<td>4.73</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>logFreqN2</strong></td>
<td>0.1443</td>
<td>0.0304</td>
<td>4.75</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>spellingRatio</strong></td>
<td>-0.1563</td>
<td>0.0264</td>
<td>-5.93</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Ametaphor=Yes</strong></td>
<td>0.8382</td>
<td>0.1009</td>
<td>8.31</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Bmetaphor=Yes</strong></td>
<td>-1.6511</td>
<td>0.0989</td>
<td>-16.70</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>ABmetaphor=Yes</strong></td>
<td>-2.0563</td>
<td>0.0978</td>
<td>-21.02</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>For=Yes</strong></td>
<td>0.2241</td>
<td>0.0929</td>
<td>2.41</td>
<td>0.0160</td>
</tr>
</tbody>
</table>

R² adjusted = 0.498
Issues

• different relations affect transparency of different constituents
• Gagné and collaborators: relational information accessed via the concepts associated with individual modifiers and heads, rather than independently of them (e.g. Gagné & Spalding 2014)
• shifts are surprisingly difficult to code
• is ‘literal’ the same as ‘not metaphorically shifted’?
New hypothesis
Bell & Schäfer (in preparation)

Transparency depends on degree of expectedness of a particular word sense and a particular relation for a given constituent
Data

- all strings of exactly two nouns that follow an article in the British National Corpus
- which also occur four times or more in the USENET corpus (Shaoul & Westbury 2010)
- from this set, extracted the positional constituent families for all noun constituents in the Reddy et al dataset
Coding

• all family members coded for
  – semantic relation between constituents
  – WordNet sense of constituent
• calculated the proportion of compound types in each constituent family with
  – each semantic relation (relation proportion)
  – each WordNet sense of the constituent in question (synset proportion)
• take these two measures to reflect the expectedness of the respective relations and WordNet senses of the constituents
## Literality constituent A

| Term                                      | Coef  | S.E.   | t     | Pr(>|t|) |
|-------------------------------------------|-------|--------|-------|----------|
| Intercept                                 | -4.6413 | 0.6593 | -7.04 | <0.0001  |
| relPropInN1Fam                           | -0.2187 | 0.6013 | -0.36 | 0.7161   |
| logFamSizeN1                              | -0.0189 | 0.0931 | -0.20 | 0.8395   |
| synsetPropInN1Fam                        | -0.2426 | 0.6152 | -0.39 | 0.6934   |
| logSynsetCountN1                          | -0.7939 | 0.2469 | -3.22 | 0.0013   |
| compoundTokenPropInN1Fam                 | 3.0130  | 0.6788 | 4.44  | <0.0001  |
| logFreqN1                                 | 0.8728  | 0.0569 | 15.34 | <0.0001  |
| relPropInN1Fam * logFamSizeN1             | 0.3311  | 0.1305 | 2.54  | 0.0113   |
| synsetPropInN1Fam * logSynsetCountN1      | 0.6855  | 0.3161 | 2.17  | 0.0303   |
| compoundTokenPropInN1Fam * logFreqN1      | -0.2804 | 0.0816 | -3.44 | 0.0006   |

$R^2$ adjusted = 0.334
Literality constituent A: effect of relation
Literality constituent A: Effect of sense
Literality constituent A: Effect of frequency

![Graph showing the effect of frequency on logFreqN1 and compoundTokenPropInN1Fam.](attachment:graph.png)
Summary

- Informative constituents tend to be stressed
- Informativity also related (inversely) to transparency of constituents
- Certain semantic relations associated with right stress
- But relationship of semantic relations to transparency is not straightforward
Thank you!
References


Bell, Melanie J. and Martin Schäfer (in preparation). Modelling semantic transparency


References, continued


