Explaining English Compound Stress Analogically

Sabine Arndt-Lappe, Universität Siegen,
Melanie J. Bell, Anglia Ruskin University
Objectives

• empirical
  – corpus study
  – test how compound stress assignment in the corpus can be modelled with the help of a computational analogical model (AM, Skousen 1989, 1992, Skousen et al. 2002 et seq.)

• theoretical
  – English compound stress functions analogically
  – AM is able to model and predict the interplay of different types of factor influencing compound stress
  – An analogical approach to compound stress assignment is superior to a categorical, rule-based approach
English compound stress

• variation between two possibilities: left stress, right stress

• left-stressed examples
  apple juice    window washer    Oxford street
  tea bag       cat food        cheese man
  meal time     science group

• right-stressed examples
  apple pie     glass door      
  car radio     gold earring
  easter holiday kitchen door

• predictive factors include constituent family and semantic properties

(cf. e.g. Plag 2006, Plag, Kunter, Lappe 2007; Plag, Kunter, Lappe, Braun 2008, Arndt-Lappe 2011a, Bell 2011)
Semantic effects

• 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 (*sílk worm*).
Constituent family effects

Stress is assigned by analogy with compounds in the mental lexicon that share either N1 or N2.

<table>
<thead>
<tr>
<th>100 % left</th>
<th>0 % left</th>
<th>10 % left</th>
</tr>
</thead>
<tbody>
<tr>
<td>Óxford Street</td>
<td>Oxford Road</td>
<td>state administration</td>
</tr>
<tr>
<td>Régent Street</td>
<td>Mill Road</td>
<td>state budget</td>
</tr>
<tr>
<td>Hárley Street</td>
<td>Upland Road</td>
<td>state benefits</td>
</tr>
<tr>
<td>... Street</td>
<td>... Road</td>
<td>state house</td>
</tr>
<tr>
<td></td>
<td></td>
<td>state funds</td>
</tr>
<tr>
<td></td>
<td></td>
<td>state ...</td>
</tr>
</tbody>
</table>

'constituent family stress bias'
Problems for accounts of the variation

• Variation is systematic and productive, but cannot be captured in terms of deterministic rules

• Not all compounds that have a particular N1 or N2, and not all compounds with a certain semantics have the predicted stress pattern (e.g. 'made of'-relation: apple pie vs. apple juice)

• Interaction between rather local effects (based on constituent family, affecting few compounds) and more general effects (based on semantics) is unclear.
Proposal

• an analogical view of word-formation solves many of the problems of approaches involving deterministic rules

• especially:
  – variability / leakage of rules
  – interaction of local and general effects (an aspect of productivity)

• The criticism against analogical views of word-formation that they are
  – vague
  – non-predictive, and
  – not testable

is not true for some computational analogical models (in particular: AM, Skousen 1989, 1992, Skousen et al. 2002).
The hypothesis:
Compound stress is assigned by analogy

necessary assumption: analogues can be sets of words
computational analogical models

- can operationalise the notions of 'similarity' and 'sets of analogues' (alternative term: 'analogical sets')
- can model variation, categorial behavior and leakage
- analogy becomes predictive and predictable
- the theory becomes testable
- algorithm:
  - AM(L) (Skousen 1989, 1992 et seq.; an alternative: TiMBL, Daelemans et al. 1999 et seq.)
The basic architecture of an analogical model

<table>
<thead>
<tr>
<th>feature 1</th>
<th>feature 2</th>
<th>feature 3</th>
<th>feature 4</th>
<th>stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>afternoon</td>
<td>break</td>
<td>no</td>
<td>temporal</td>
<td>right</td>
</tr>
<tr>
<td>cat</td>
<td>food</td>
<td>no</td>
<td>no</td>
<td>left</td>
</tr>
<tr>
<td>chocolate</td>
<td>raisin</td>
<td>no</td>
<td>no</td>
<td>right</td>
</tr>
<tr>
<td>coffee</td>
<td>jar</td>
<td>no</td>
<td>no</td>
<td>left</td>
</tr>
<tr>
<td>sports</td>
<td>center</td>
<td>no</td>
<td>no</td>
<td>left</td>
</tr>
</tbody>
</table>

**exemplars in the lexicon**

**set of analogues / analogical set**

<table>
<thead>
<tr>
<th>feature 1</th>
<th>feature 2</th>
<th>feature 3</th>
<th>feature 4</th>
<th>stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>pie</td>
<td>material</td>
<td>no</td>
<td>right</td>
</tr>
<tr>
<td>cherry</td>
<td>pie</td>
<td>material</td>
<td>no</td>
<td>right</td>
</tr>
<tr>
<td>pork</td>
<td>pie</td>
<td>material</td>
<td>no</td>
<td>right</td>
</tr>
<tr>
<td>chicken</td>
<td>burger</td>
<td>material</td>
<td>no</td>
<td>left</td>
</tr>
<tr>
<td>olive</td>
<td>oil</td>
<td>material</td>
<td>no</td>
<td>right</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>feature 1</th>
<th>feature 2</th>
<th>feature 3</th>
<th>feature 4</th>
<th>stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>olive</td>
<td>pie</td>
<td>material</td>
<td>no</td>
<td>???</td>
</tr>
</tbody>
</table>

4x right, 1x left

stress: right (majority choice)
Computing Analogical Sets

the similarity space

different degrees
different dimensions

olive stone
olive pie
olive bread
apple pie
cottage pie
fruit cake
Computing Analogical Sets

the similarity space

different degrees
AM starts with the most similar exemplars.

different dimensions
Along all dimensions, AM tries to include more distant exemplars.

It does so if the more distant exemplars behave like the more similar group w.r.t. stress assignment. => minimised uncertainty
Methodology

Data: NN constructs

- 406 compounds extracted from the British National Corpus, for which stress was produced and rated consistently across all 4 recordings done for Bell (2011), which have a constituent family for N1 or N2 or both of these,

- 241 are left-stressed, 165 are right-stressed

Coded features

- N1, N2 (in spelling)
- semantic properties and relations found relevant in Bell (2011), Plag et al. (2007, 2008)

Setup

'leave-one-out', corpus is tested on itself
AM experiment – overall performance

<table>
<thead>
<tr>
<th>features given as information source</th>
<th>F-score(_{\text{left}}^*) (% correct predictions)</th>
<th>F-score(_{\text{right}}) (% correct predictions)</th>
<th>F-score(_{\text{average}}) (% correct predictions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>constituent family &amp; semantic features</td>
<td>0.94 (95%)</td>
<td>0.92 (90%)</td>
<td>0.93 (93%)</td>
</tr>
</tbody>
</table>

*cf. Daelemans & Bosch 2005 for discussion of performance measures

⇒ AM predicts stress assignment correctly for 93% of the data!
⇒ Predictions are almost equally good for left and right stress!

How does AM do this?
three types of analogical set

small sets
(1-15 exemplars)
75% of the data

mid-size sets
(53-55 exemplars)
11% of the data

large sets
(179-183 exemplars)
7% of the data
Small analogical sets

- Clear effect of constituent family in all small sets (all sets are comprised of members of the constituent family)
- Stress is sometimes left, sometimes right

Examples, sets of two:

- cat food
- banana sandwich
- football quiz

- cat muck
- lamb sandwiches
- football nights

- convenience food
- salmon sandwiches
- football party

⇒ 'local analogies'
Mid-size analogical sets

• 53 exemplars that reoccur together in 45 analogical sets
• stress is right

examples:
baby boy plastic wallet bastard teacher bitch teacher
cotton sheets glass bowl glass dish gold jewellery
gold band leather bags toy cups

⇒ wider similarity space: For these 45 new words, exemplars that share certain semantic features appear together in the analogical set. The fact that they reappear more often gives the impression of a 'rule'. But there is no rule.

⇒ 'less local analogies': give the impression of a productive rule
What do compounds in the ‘mid-size sets‘ have in common?

- **semantic relations:** esp.
  - material: yes (34), no (19)
  - temporal: no
  - locative: no
  - copulative: no

- **semantic categories of N1:**
  - N1 is a time: no
  - N1 is a location: no
  - N1 is adjective-like: yes
  - N1 is a material: yes (48), no (5)
  - N1 is a social group: no

(categories & coding were taken from Bell 2011)
leakage in the mid-size set

- All 53 members of the mid-size Analogical Set share a semantic feature
  - 'N1 has adjective-like qualities'

BUT

- There are 61 'adjective-like' compounds in the dataset!
Large analogical sets

- 178 exemplars that reoccur together in 29 analogical sets
- stress is mostly left

**Examples**
- alarm business
- antiques day
- arm bands
- art centre
- assessment piece
- attache case
- baby stuff
- banking job
- begging bowl
- bike things
- bin day
- bingo money

**Wide similarity space**: analogical sets consist of exemplars NOT having a semantics that favours right stress.

The fact that they share 'negative' values for the relevant semantics gives the impression of a 'default situation'.
(cf., e.g., Derwing & Skousen 1989, Eddington 2000 for inflection)

The most **non-local type of analogy** conceivable
Compound stress assignment by analogy

• AM is highly successful in modelling compound stress on the basis of constituent family and semantics

• one single mechanism produces effects for which three different mechanisms are invoked in other frameworks:
  – different degrees of productivity
  – local, exceptional analogies
  – rules and default rules

• key: interplay of local and less local analogies
Summary & conclusion – empirical level

• AM was used to test how the challenges could be solved within an analogical theory of word-formation
  
  – very good overall predictive power

  – non-deterministic behaviour is expected in an analogical model

  – interaction of local and less local generalisations as well as default situations are epiphenomena of 'gang behaviour' among analogical sets
summary and conclusion – theoretical level

• AM constitutes a testable version of a theory of word-formation that is based on analogy.
• rules vs. analogy? Compound stress assignment provides evidence
  – against rules that are deterministic and independent of the lexicon
  – in favour of an approach that assumes no strict distinction between the lexicon and rules, and that allows for systematic variability. An analogical approach of the type implemented in AM is one plausible possibility.
Thank you very much for your attention!
References


Bell, Melanie & Ingo Plag. 2010. Informativeness is a determinant of compound stress in English: Ms, submitted for publication.


Skousen, Royal & Theron Stanford. 2007. AM::Parallel. available from http://humanities.byu.edu/am