Regularity and Variability in English Compound Stress Assignment:
a Case for Analogy

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
Structure

• The empirical basis
  – English compound stress

• The theoretical foundation
  – an analogical approach to word-formation

• Modelling compound stress with AM
  – (featuring an introduction to AM, as we go along)

• Summary & conclusion
the empirical basis: English compound stress
English compound stress

• variation between two possibilities: left stress, right stress

• left-stressed examples
  apple juice  window washer  Óxford street
  téabag  cátfood  chéeese man
  méal time  scíence group

• right-stressed examples
  apple píe  glass dóor
  car rádio  gold éarring
  easter hóliday  kitchen dóor

(cf. e.g. Plag 2006, Plag, Kunter, Lappe 2007; Plag, Kunter, Lappe, Braun 2008, Arndt-Lappe 2011a, Bell 2011)
A number of factors have been identified:

- constituent family: N1, N2
- semantics:
  - semantic relations between N1 and N2
  - semantic categories of N1 and N2
- some others (not in focus today)

Semantic effects

• Certain semantic relations are right-stressed (e.g. ‘locative’ compounds, *Boston hârbour*).

• Certain semantic classes of constituents trigger right stress (e.g. substance nouns as N1, *silk shîrt*).

• 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.

| Óxford Street | Oxford Róad | state administration |
| Régent Street | Mill Róad | state budget |
| Hárley Street | Upland Róad | state benefits |
| ... Street | ... Róad | state house |
| ... | ... | state funds |
| ... | ... | state ...

100 % left 0 % left 10 % left

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

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

- leakage of rules
- interaction of local and general effects
leakage of rules

Not all compounds that have a particular N1, N2, and / or semantics have left or right stress

example

semantic rule: ‘Assign right stress to compounds where a material relation holds between N1 and N2.’

apple pie cherry pie fruit sundae
metal bridge metal rim metal strip

exception:
apple juice (orange juice, strawberry juice, etc.)
interaction of local and general effects

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

example

semantic rule: ‘Assign right stress to compounds where a material relation holds between N1 and N2.’

apple pie cherry pie fruit sundae
metal bridge metal rim metal strip
= more general effect

exception:
ápple juice (órange juice, stráwberry juice, etc.)
constituent family effect = more local effect
introducing an alternative to the rule-based view: analogy
Proposal

• an analogical view of word-formation solves some of the problems of approaches involving deterministic rules
• especially:
  – leakage of rules
  – interaction of local and general effects
• 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).
Analogy in synchronic word-formation theory

Definition
A new complex word is formed on the basis of an existing base-derivative pair in the lexicon, on the basis of the perceived similarity between the new word and its analogue.

Status in much of the rule-based literature
• Analogies are
  • exceptional,
  • occasional (i.e. nonproductive) formations that are based on
  • high degrees of formal and semantic similarities.

Typical examples
_E. sea-scape_, formed on the basis of _land-scape_  
_G. zwei-sam_, formed on the basis of _ein-sam_  
(from: Bauer 1983: 96)
Analogy in synchronic morphology

formal expression: proportional analogy

```
base of analogue

land

landscape

analogue

= 

base of new word

sea

seascape

new word
```
Analogy in synchronic morphology

an often-observed fact: analogies can give rise to productive patterns

\[
\begin{array}{c}
\text{base of analogue} \\
\text{land} \\
\text{landscape} \\
\text{analogue} \\
\hline
\text{base of new word} \\
\text{sea} \\
\text{seascape} \\
\text{new word} \\
\end{array}
\]

... giving rise to new words: cloud-scape sky-scape water-scape

dream-scape winter-scape wire-scape

(Bauer 1988: 96)
Word formation as analogy: describing stress assignment in olive pie as an analogical process

the hypothesis: compound stress is assigned by analogy

lexemes in the Mental Lexicon

```
base of analogue

pie

apple pie, cherry pie, apricot pie, pork pie

analogues

base of new word

pie

?olive pie

new word

necessary assumption: analogues can be sets of words
```
Problems with the analogical hypothesis

• unpredictable: It is not clear where and on which basis an analogy will occur
  – 'similarity': which kind? (e.g. formal? semantic? which properties make potential base-analogue pairs 'similar' to a potential new word?)
  – similarity is a gradient concept
  – analogue selection: For a given new word, there are always many potential analogues

• no clear distinction between productive and unproductive processes, regular and irregular processes
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

• the theory becomes testable

• algorithm:
  – AM(L) (Skousen 1989, 1992 et seq.; an alternative: TiMBL, Daelemans et al. 1999 et seq.)
Modelling compound stress with AM

1. general: How does AM perform, overall?
2. How does AM account for leakage and the interaction of local and general effects?
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>yes</td>
<td>no</td>
<td>right</td>
</tr>
<tr>
<td>cat</td>
<td>food</td>
<td>no</td>
<td>yes</td>
<td>left</td>
</tr>
<tr>
<td>chocolate</td>
<td>raisin</td>
<td>yes</td>
<td>yes</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>yes</td>
<td>left</td>
</tr>
</tbody>
</table>

exemplars in the lexicon

<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>yes</td>
<td>no</td>
<td>right</td>
</tr>
<tr>
<td>cherry</td>
<td>pie</td>
<td>no</td>
<td>yes</td>
<td>right</td>
</tr>
<tr>
<td>pork</td>
<td>pie</td>
<td>yes</td>
<td>yes</td>
<td>right</td>
</tr>
<tr>
<td>chicken</td>
<td>burger</td>
<td>yes</td>
<td>no</td>
<td>left</td>
</tr>
<tr>
<td>olive</td>
<td>oil</td>
<td>yes</td>
<td>no</td>
<td>right</td>
</tr>
</tbody>
</table>

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>olive</td>
<td>pie</td>
<td>yes</td>
<td>no</td>
<td>???</td>
</tr>
</tbody>
</table>

new word

4x right, 1x left

stress: right (majority choice)
Which exemplars end up in the set of analogues?

Problem 1: What are the features on the basis of which similarity is computed?
Which exemplars end up in the set of analogues?

Problem 1: What are the features on the basis of which similarity is computed?

- N1, N2
- semantic features

The similarity space is computed on the basis of the features given by the researcher coded features in our experiment: olive pie, pork pie, apple pie, olive oil, walnut oil, zucchini pie, vegetable stirfry, beef steak, apricot crumble, fruit cake, cheese cake.
Which exemplars end up in the set of analogues?

Problem 2: How much similarity does it take to end up in the analogical set?

the similarity space
Which exemplars end up in the set of analogues?

Problem 2: How much similarity does it take to end up in the analogical set?

AM determines, for each individual data item, which combination of shared features behaves homogeneously w.r.t. stress assignment. 'homogeneity': a measure of how informative a given feature combination is w.r.t. the task.
the experiment
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, Arndt-Lappe 2011a 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!
AM experiment – overall performance

How does that work?

- Let us look at
  - the coded features that are helpful in the classification task
  - the analogical sets that the stress decisions are based on

- crucial observation: they vary in size, resulting in more local and less local generalisations
The interaction of local and less local effects in the simulation
features that AM found useful

- constituent family
  - N1: Christmas catalogue, Christmas trimmings
  - N2: mice problem, circulation problem

- semantic relations favouring rightward stress:
  - material: silk shirt
  - temporal: Monday morning
  - locative: London school
  - copulative: singer songwriter

- semantic categories favouring rightward stress:
  - N1 is a time: easter holiday
  - N1 is a location: town church
  - N1 is adjective-like: end wall, glass bowl
  - N1 is a material: iron leg, glass bowl
  - N1 is a social group: family evening

(categories & codings were taken from Bell 2011)
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:**
  - 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 members of the mid-size Analogical Set share a semantic feature
  – ‘N1 has adjective-like qualities’

BUT

• Not all compounds that have the feature ‘N1 has adjective-like qualities’ are in the mid-size Analogical Set

• Not for all compounds that have the feature ‘N1 has adjective-like qualities’ does the mid-size Analogical Set apply

(N = 61)
Large analogical sets

• 178 exemplars that reoccur together in 29 analogous sets
• stress is mostly left

**examples**

<table>
<thead>
<tr>
<th>alarm business</th>
<th>antiques day</th>
<th>arm bands</th>
<th>art centre</th>
</tr>
</thead>
<tbody>
<tr>
<td>assessment piece</td>
<td>attache case</td>
<td>baby stuff</td>
<td>banking job</td>
</tr>
<tr>
<td>begging bowl</td>
<td>bike things</td>
<td>bin day</td>
<td>bingo money</td>
</tr>
</tbody>
</table>

**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

• compound stress: a phenomenon that provides a challenge for accounts based on deterministic rules

  • leakage: generalisations are non-deterministic
  • local vs. less local generalisations:
    – semantic factors: 'less local'
    – constituent family: 'more local'
Summary & conclusion – methodological 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!*  

*And thanks to:  
Melanie Bell for working with me on this, and Gero Kunter and Ingo Plag for fruitful discussion and input.