Mixed categories and gradient grammatical constraints
Gabriel Doyle and Roger Levy
University of California - San Diego

gdoyle@ling.ucsd.edu
rlevy@ling.ucsd.edu

How is speaker choice in a lexic syntactic-alternation influenced by the mixed-category status of one of its alternants?

Hypothesis

Environment Prototypicality: Alternants are preferred in environments prototypical of their syntactic category/categoricals.

Alternation

We investigate

needs to be done ~ needs doing

in British English.

Both alternants are acceptable in many contexts:

The couch needs to be cleaned [the to be form]
The couch needs cleaning [the ing form]

This is not completely free variation:

You need to show the way
You need showing the way

We all need to be cooled down after the debate
We all needed cooling down after the debate

But the awkward variants do occur:

Kewell needs showing the exit door.

This is a bug and the ubuntu/etheral package needs teaching to use ‘sudo’ (user-password) instead of ‘su’.

Mixed Categories in the Alternation

• English gerunds are a mixed category (Malaud 2000:31) with both verbal nominal properties:
  • Verbal: gerunds can govern other NPs
  • Verbal: gerunds are modified by adverbs, not adjectives

Nominal: gerund phrase have NP-type external distributions

Mixed: optional subject can be accusative or genitive

• But the past participle has no nominal properties

• Needs to be done should be favored in more prototypically nominal environments

Semantic (Near) Equivalence

The ideal stochastic lexic-syntactic variable is not subject to categorical meaning-based constraints (see, e.g., Weiner and Labov 1983)

We investigated 4 possible categorical semantic constraints:

• Propositions 1-3, by Murphy (referenced in Murray, Frazer, & Simon 1996)
• Proposition 4, our own

Counter-examples to all these proposals occur in the BNC

Categorical semantics do not drive the alternation

Semantic-determination Proposals

Proposal 1: to be implies agent as possessor

John’s car needs to be washed (⇒ John will wash my car)

John’s car needs washing (⇒ John will wash my car)

But:

He said Prost would almost certainly be granted his super licence, but said his behaviour in using “insulting terms” in his criticism of FISA would still need to be considered by the world council next month. (BNC)

Agent = the world council = possessor

Proposal 2: ing implies benefit to subject

My books need to be sold (⇒ Being sold benefits the books)

My books need selling (⇒ Being sold benefits the books)

But:

If the reservoir is holed then the remote brake servo will not work and swapping.

The remote brake servo does not benefit from being removed

Proposal 3: ing implies pre-existing subject

My paper needs to be written (***)

My paper needs writing

But:

I do not believe that the current management at British Rail is capable of building the project, although I believe that if needs building, [winc]

Subject (the project) does not exist yet

Proposal 4: achievement and state verbs (Vendler 1957) require to be

Achievement (instantaneous event) and state verbs lack a progressive tense, so can you use the -ing alternant for these?

(*) The form, the past tense (**) The facts need facing.

But:

39 ing achievements, 47 to be achievements

5 ing states, 28 to be states

Needs ~ as a stochastic variable

• Alternation not driven by identifiable semantic constraints
• A gradient/stochastic approach is thus reasonable

We use mixed-effects logistic regression (Agresti 2002, Renor & Levy 2006, Bresnan et al., 2007, Jaeger 2008) as a modeling framework:

• Categorical and continuous constraints both influence a probabilistic outcome
• Constraint strength identifiable with coefficient magnitude
• Constraints can “dominate” one another by having successively larger coefficients

• “Ganging up” and other softer effects also possible

Dataset

1004 automatically extracted, hand-filtered and hand-annotated examples from the BNC.

Retrospectively sampled (Agresti 2002) for balance: 502 examples of each alternant.

Potential sources of variation (model predictors)

CATEGORICAL PREDICTORS

• animacy (animate v. inanimate)***
• concreteness (concrete v. abstract)***
• definiteness (definite v. indefinite v. unclear)***
• pronoun (pronominal v. non-pronominal)***
• relativization (relative clause v. non-relativized)***
• construction’s tense (future v. past v. present v. other)***
• inflection of need v. (v other)
• modality (spoken v. written)
• aspect (state v. process v. extended event v. instant event v. unclear)***

QUANTITATIVE PREDICTORS

subject length (smoothed log; length in words)

verb length (in syllables)***

verb frequency (smoothed log: from CELEX)***

POST VERBAL MATERIAL

• post-verbal dependent length (smoothed log; length in words)***

• ambiguously attached phrase length (smoothed log; length in words)***

Sites needed to be monitored continuously

in order to pick up fluctuations.

RANDOM EFFECT OF VERB

Each verb can have its own idiosyncratic preference for an alternant

These preferences are assumed to be normally distributed

Model Results: verb-specific preferences

• Verbs vary considerably in their idiosyncratic preferences
• Standard deviation of estimated random effects = 0.546

Some common verbs, such as do, remind, and replace, have strong preferences

Structural Bias & environment prototypicality

We quantify environment prototypicality probabilistically:

• Theoretically, we want the structural bias to be large and positive
• In practice, we crudely approximate this with the first constituent C, following the need-selected verb

High structural bias should favor the (nominal) need-ing form

RESULT: When added to the model, structural bias is of marginal significance (likelihood-ratio test: p=0.07)

Mutate it highly correlated (p<0.91) with post-verbal dependent length

Significance of the latter drops dramatically (to p<0.03)

Future work: disentangle from dependent length with finer-grained measures of structural bias

Conclusions

• Speakers are aware of the gerund’s mixed category status when composing a sentence and are influenced by environment prototypicality when choosing between mixed and single category forms