Breaking up is easy to do: Evidence for decomposition in regular and irregular forms

Constantine Lignos
Children's Hospital of Philadelphia
University of Pennsylvania

Kyle Gorman
Oregon Health and Science University

LSA 2014
1/5/14

I. Preliminaries

Why care about decomposition?

- The objects of lexical storage should align with the grammatical representations of linguistic theory
  - Continuum between more (e.g., Goldberg 2006; Jackendoff 1975) and less (e.g., Anderson 1999, Di Sciullo and Williams 1987; Halle and Marantz 1983) storage
- Properties of these representation should be reflected in processing

Basic models

0. Whole-word lookup (e.g., Butterworth 1983)

playing

1. Full decomposition (e.g., Taft and Forster 1975)

play	ing

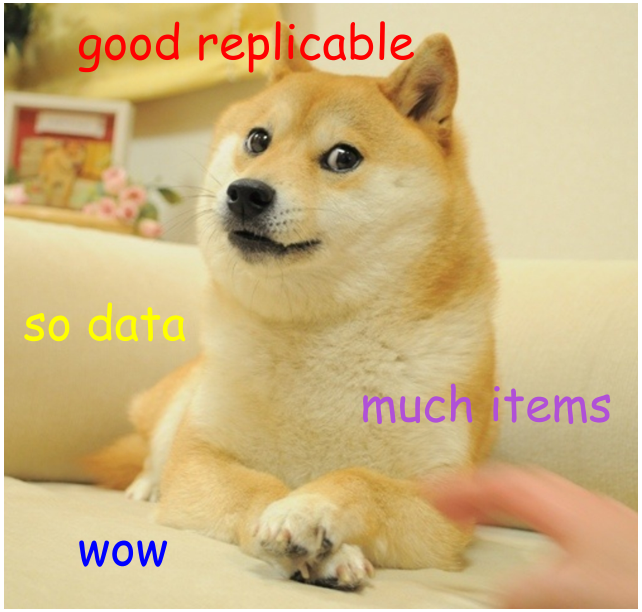
2. Dual-route model (e.g., Baayen et al. 1997)

playing

play	ing

Impasse

- After more than three decades, little consensus on role of word storage in morphological processing
  - Contradictory findings across and within experimental paradigms
  - Use of varying frequency norms, item sets, modeling strategies
- We use large scale experimental data to evaluate processing theories
  - **Massively larger** numbers of items, subjects, and trials
  - Able to evaluate many claims on same data set



Proposed model

1. Regular and irregular forms are *obligatorily* decomposed into bases and affixes
2. Access time of base is proportional to frequency *rank* of the base
3. Access time for affixed forms is sensitive to cost (to be defined later) of combining base and affix

Data

- Visual lexical decision latencies from the English Lexicon Project (ELP; Balota et al. 2007)
- Excluded incorrect responses, non-native speakers
- Analyzed regularly inflected words (*null*, *-s, -ed , -ing*)
  - 500,469 trials, 16,972 words, 8,229 lemmas, 805 subjects
- Analyzed irregulars
  - 2,355 trials, 76 words, 756 subjects
  - Items excluded: compounds, no-change, doublets

Modeling procedure

- Linear mixed effects multiple regression
- Per-subject and university random intercepts, per-subject random slopes for predictors of interest
- Baseline predictors:
    - Trial number, gender, years of education, squared word length (New et al. 2006), orthographic neighborhood density (Yarkoni et al. 2008), number of syllables

Frequency measures

- (Whole) word frequency: frequency of a complete word (*growing*) in some corpus
- Base frequency: total frequency of morphologically related words (*grow, grows, growing, grown, growth*)
- Suffix conditional probability:

  *p(***word|base**): *p(***word***) / p(***all words sharing base***)*
- Frequencies from SUBTLEX, a 51 million-word corpus of American English film subtitles
  - Best frequency estimates for behavioral measures (Brysbaert and New 2009)
  - Morphological analyses from ELP augmented by entries in CELEX2

Outline

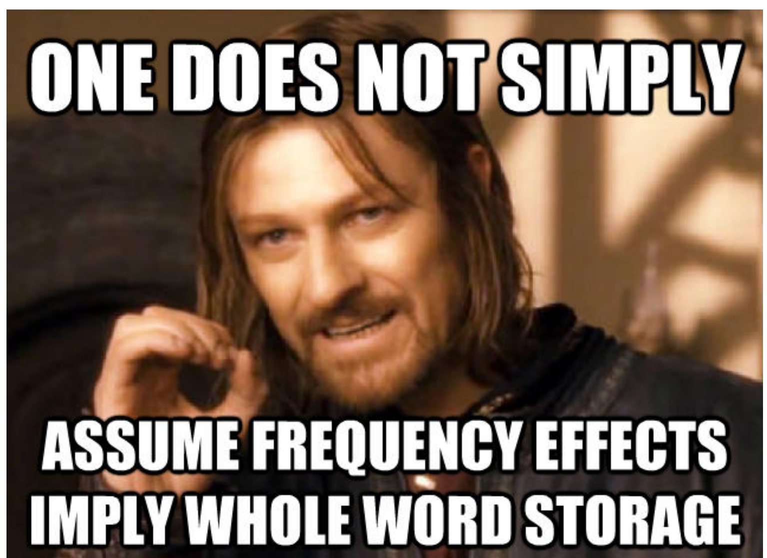
Demonstrate that:

1. Decomposition is obligatory over a wide range of frequencies
2. The *Rank Hypothesis* (Murray and Forster 2004) provides a powerful and parsimonious mechanism for explaining the observed frequency effects
3. For both regular and irregular forms, frequency effects can be best explained by this mechanism

II. The case for decomposition

Interpreting word frequency effects

- Atoms of lexical memory should show frequency effects
- But whole-word frequency effects do not entail whole-word storage:
  - Effects may be driven the frequencies of sub-components of words (i.e., base and affix frequency) which are strongly correlated with whole word frequency (Lignos and Gorman in press, New et al. 2004)



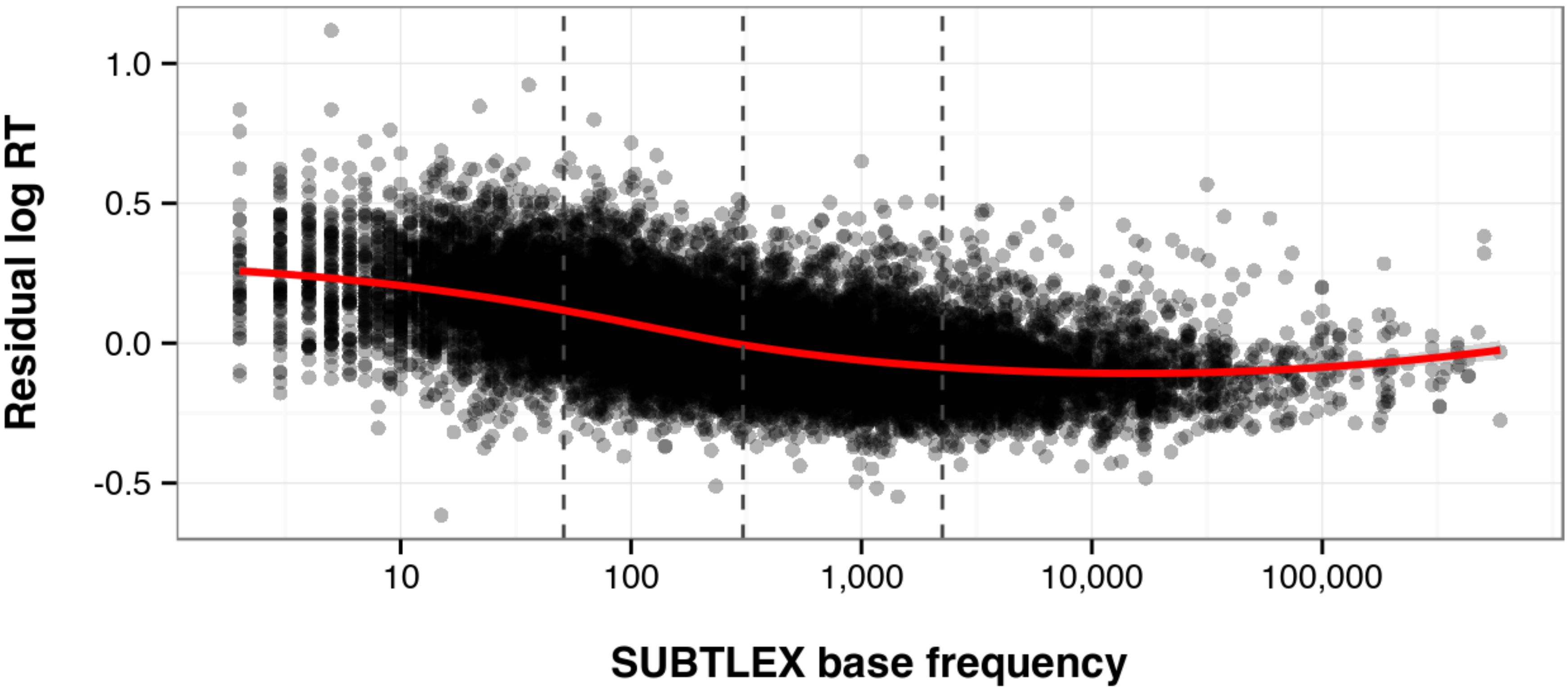
Classic findings in non-decompositionality…

- Sereno and Jongman (1997):
  - Inflected words only show whole word frequency effects, no evidence for decomposition
- Alegre and Gordon (1999):
  - Low-frequency inflected words show no whole word frequency effects, high-frequency words do
  - Interpretation: low frequency inflected forms decomposed, thus but high-frequency words not
- Baayen et al. (2007):
  - Effect of base frequency is marginal and bases are largely irrelevant in an optimal processing system

…do not withstand scrutiny

- New et. al (2004), contra Sereno and Jongman (1997):
  - French, English, and Dutch data all show same pattern of base frequency effects
- Lignos and Gorman (2012), contra Alegre and Gordon (1999):
  - Same effects of decomposition in high and low frequency regulars
- Lignos (2013), contra Baayen et al. (2007):
  - Effects of decomposition across all frequency ranges
  - Base frequency and suffix conditional probability always explain the data better than whole word frequency

The case for decomposition



- Consistent base frequency effects throughout
- Modeling all items ($p < 2.2\times10^{-16}$) and 5-quantiles (all $p < .001$) shows significant effect of decomposition
- Why this shape for frequency effect?

The Rank Hypothesis

- Murray and Forster (2004)
  - Rank (position in frequency-ordered list) better predictor of behavioral data than log frequency

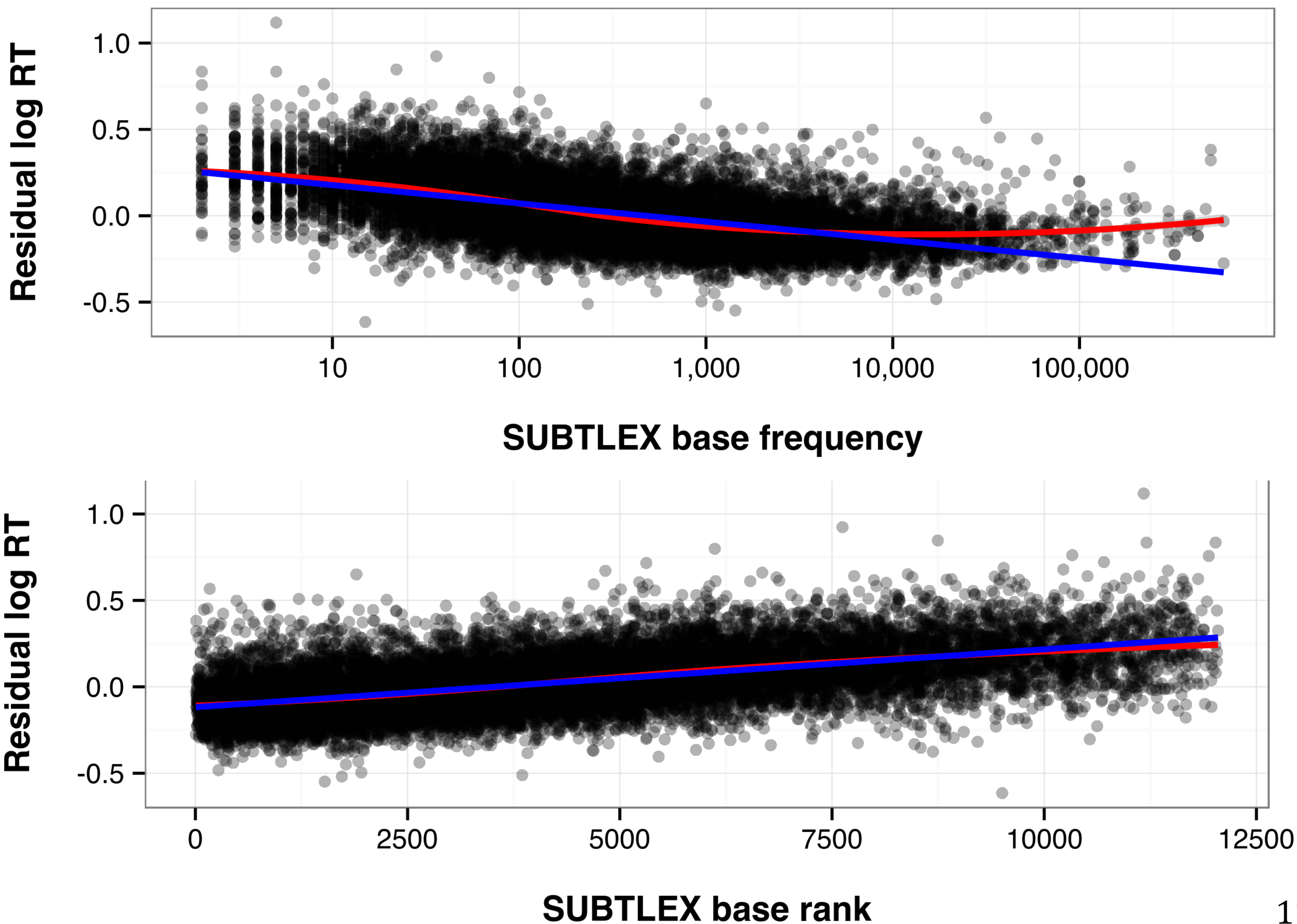
    ☐ Root 1

    ☐ Root 2

    …

    ☑ Root N
- Specifics
  - Words binned into separate lists; statistically equivalent to one list even if not implemented that way in the brain
- Predicts rank better predictor of latency than frequency

Comparing rank and base frequency



Summary for regulars

- No apparent discontinuity in base frequency effects; strong effect of decomposition throughout
- Rank provides better and more parsimonious predictor of access time than frequency

III. Modeling irregulars

Evidence for irregular decomposition

- Irregular past tense verbs exhibit base frequency effects (Kelliher & Henderson 1990)
- Irregulars fully prime their bases (e.g., *fell-FALL*)
  - **English strong verbs**: Allen & Badecker 2002, Crepaldi et al. 2010, Kielar et al. 2008, Kielar & Joanisse 2009, Marslen-Wilson et al. 1997, Pastizzo & Feldman 2002, Stockall & Marantz 2006
  - **German *-n* past participles**: Smolka et al. 2007
  - **French irregular verbs**: Meunier & Marslen-Wilson 2004
- The effect is not driven by semantic and/or orthographic similarity
  - **Masked priming**: semantic and orthographic similarity inhibit recognition (Baayen et al. 2007, Drews & Zwitserlood 1995, Grainger et al. 1991, Henderson et al. 1984)

Productivity and the rank hypothesis

Irregularly inflected words also have an advantage over frequency-matched regularly inflected words:

- **German past participles**: adults (Fleischhauer & Clahsen 2012) and children (Clahsen et al. 2004) produce irregular *-n* (*rufen-gerufen* 'shout-shouted') faster than regular *-t* (*fragen-gefragt* 'ask-asked')
- **German noun plurals**: irregular *-er* (*Ei-Eier* 'eyes') recognized faster than regular *-s* (*Auto-Autos* 'car-cars') (Penke & Krause 2002)
- **Idioms behave similarly**: *kick the bucket* processed faster than frequency-matched semantically-compositional expressions like *lift the bucket* (Fraser 1974, Swinney & Cutler 1979)

Proposed models of irregular processing

Irregularity represented as list of patterns (e.g., *SPE*, Yang 2005)

☐ Pattern 1

☐ Pattern 2

…

☐ Default

Predictions:

- Same-pattern irregulars acquired together (Yang, 2002)
- Irregulars faster than regulars of similar frequency
- Rank effect over patterns

Grouping irregulars

- Irregulars that follow same pattern grouped into *classes*
  - e.g., {bought, brought, caught, sought, taught, thought}
- *Class frequency* is sum of word frequencies of all class members
- *Class rank* is the rank of class frequencies

Results

- Class rank predicts recognition latencies of irregulars better (*AIC* = 3372) than word frequency (*AIC* = 5404)
- Class rank contributes significant information beyond that of word frequency ($p = .00165$)
- Conclusion: rank hypothesis and organization of irregulars into classes provide better model than frequency alone

IV. Conclusions

Extensions

- Different ways of organizing irregulars into classes
- Other languages with freely-available Lexicon Projects:
  - Producing reliable theory-neutral morphological analysis is the challenge
- Evaluating confounded predictors such as:
  - *family size* (Moscoso del Prado Martín et al. 2005)
  - *inflectional entropy* (Baayen 2005)

Conclusions

- Evidence for obligatory decomposition of regulars and irregulars in all frequency ranges
- Frequency effects for both best explained by a single access mechanism as per Rank Hypothesis
- Big (and good) data and simple models help align the atoms of lexical memory with those of morphological theory

Thanks!