Morphological Decomposition: Integrating Processing with Incremental Learning

Gaja Jarosz (jarosz@umass.edu), Department of Linguistics, University of Massachusetts

1. Background The relative frequency of morphological bases affects the productivity of phonological processes, including English aspiration, English stress shift, Turkish vowel harmony, and Japanese nasalization in derived forms [1-4]. Speakers rely on these relations when generalizing to novel words [2] and during online processing [4]. Token frequency can also affect the productivity of *morphological* processes [5-7]. Existing work incorporates some of these morphological effects into the phonological grammar using frequency-sensitive lexical constraints [7-9] and frequency-sensitive constraint evaluation [2].

2. Overview Hay's [5] race model of lexical access posits that base frequency modulates processing of derived forms: higher frequency favors decomposition and lower frequency favors holistic processing. Complementing previous modeling work, we explicitly formalize Hay's model within a Maximum Entropy Grammar framework, allowing learners to consider competing analyses for each derived form, one holistic and one decomposed. Although both analyses are compatible with the learning data, we show standard learning updates derive attested affix- and item-level effects of relative frequency: derived words with more frequent bases are more likely to be represented compositionally, and affixes occurring more often with decomposed derived forms are predicted to generalize more productively to novel items.

3. Methods Model predictions were evaluated on a toy language (Table 1) with the crucial item-level and affix-level token frequency manipulations [5-6]. The two affixes are matched on their type and token frequencies, but average base frequencies are higher for affix B than A. The model is tasked with learning all the base and derived forms, equipped with three families of constraints regulating stem, affix, and wholeform realization (Table 2). It considers compositional representations varying the stem and affix realizations as well as holistic representations for each derived form. Weights for all constraints are initialized to 0. The model ultimately learns to represent all X+A derived forms holistically and most X+B forms (S6-8) compositionally (Fig. 4.). At the item level, frequency of base forms modulates this process, however, with faster and stronger lexicalization for X+A forms with less frequent bases. Additionally, derived X+B forms with relatively low frequency bases can also be partially or wholly lexicalized, as illustrated by S5+B. At the affix level, the greater compositionality of X+B derived forms coincides with greater productivity of affix B.

4. Summary By directly incorporating Hay's competition model into learning, we find that incremental processing of learning tokens gives rise to representations that mirror a range of item-level and affix- level effects of usage frequency observed in work on lexical processing.

References

[1] Breiss, C., H. Katsuda & S. Kawahara. 2022. A Quantitative Study of Voiced Velar Nasalization in Japanese. In *PLC*.

[2] Breiss, Canaan. 2024. When bases compete: A voting model of lexical conservatism. *Language*. LSA 100(2).

[3] Grachek, Darby. 2025. Connection Between Lexical Processing and Phonological Regularity: A Comparison Between English and Turkish. *Proc. of CogSci* 47.

[4] Zuraw, Kie, Isabelle Lin, Meng Yang & Sharon Peperkamp. 2021. Competition between whole-word and decomposed representations of English prefixed words. *Morphology* 31(2).

[5] Hay, J. 2001. Lexical frequency in morphology: is everything relative? *Linguistics* 39(6).

[6] Hay, J., & Baayen, H. 2002. Parsing and productivity. *Yearbook of morphology*.

[7] Jarosz, Gaja, Cerys Hughes, Andrew Lamont, Brandon Prickett, Maggie Baird, Seoyoung Kim & Max Nelson. 2025. Type and token frequency jointly drive learning of morphology. *Journal of Memory and Language* 144.

[8] Smith, Brian W. & Claire Moore-Cantwell. 2017. Emergent idiosyncrasy in English comparatives. In *NELS 47*.

[9] Zuraw, Kie. 2000. *Patterned Exceptions in Phonology*. Ph.D. Dissertation, UCLA.

Figures and Tables

Table 1: **Toy Language**. 1-8 are stem morphemes. S1-S8 are their phonological exponents: the free-standing bases. A and B are the two affixes. All derived forms X+A and X+B are matched on token frequency, but the bases' frequencies vary within and across suffix classes.

	1	2	3	4	5	6	7	8
Base	S1	S2	S3	S4	S5	S6	S7	S8
Frequency	25	25	50	100	50	100	200	400
Base+Aff	S1+A	S2+A	S3+A	S4+A	S5+B	S6+B	S7+B	S8+B
Frequency	100	100	100	100	100	100	100	100

Figure 1: Results of Simulation

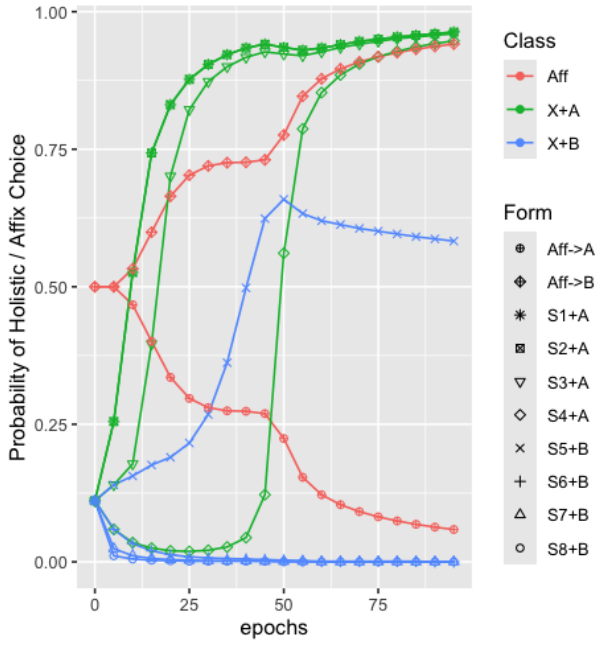

Table 2: Constraint Definitions & Examples

Family	*Definition & Examples*
Stem	Regulates realization of stem morphemes *e.g.* 1⇒S1, 1⇒S2, 1⇒S3,… 2⇒S1, 2⇒S2, 2⇒S3,…
Affix	Regulates realization of affix morphemes Aff⇒A, Aff⇒B
Holistic	Regulates realization of stored derived forms *e.g.* 1+Aff⇒S1+A, 2+Aff⇒ S2+A,… 5+Aff⇒S5+B,…