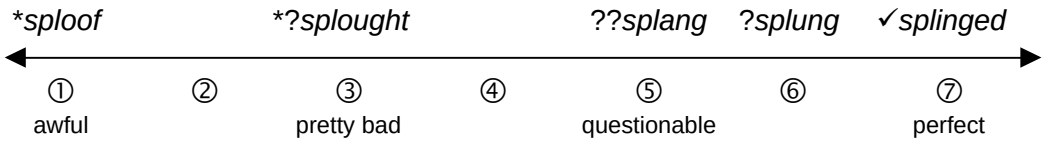
The lexical bases of morphological well-formedness

Adam Albright
University of California, Los Angeles

1. Introduction

Not all words, real or novel, are created equal — some sound better than others, either for phonological or for morphological reasons. That is, well-formedness is a gradient notion. One simple way to measure gradient well-formedness is through acceptability ratings. For example, native English speakers generally agree that there are several conceivable past tenses for the made-up verb *spling*, but not all of these competing possibilities are equally plausible or well-formed:

(1) "How good is _____ as the past tense of *spling*?"



Gradient well-formedness has been documented in a number of different domains, and for a number of different languages. Within morphology, several studies have shown that novel English irregular past tenses are more acceptable when they resemble existing irregulars (Bybee & Moder 1983, Bybee & Slobin 1982, Prasada & Pinker 1993). Ullman (1999) found further that the acceptability of *existing* irregular English past tense forms depends on the behavior of similar verbs in the lexicon. Albright (1999) showed that the acceptability of both regular and irregular conjugation classes in Italian depends on similar existing verbs. These are just a few results from a large and growing body of evidence suggesting that gradient well-formedness is a product of statistical patterns within the lexicon.

A more controversial issue is *how* exactly gradient well-formedness is derived from the lexicon. Bybee (1995) argues that the strength of a morphological pattern is related to its type frequency — i.e., the number of words which take the pattern. In the case of the hypothetical verb *spling*, we would look to the English lexicon and find that there are ten other *ing* ~ *ung* verbs, making the pattern a relatively robust one. Connectionist models, on the other hand, are influenced by both type and token frequency of similar words. For a connectionist network, then, *splung* is a plausible past tense of *spling* not only because there are ten other *ing* ~ *ung* verbs, but because in addition, some of them are quite common.

Do type and token frequency really both play a role in shaping morphological well-formedness intuitions? In spite of the intensity of the debate between proponents of connectionist vs. symbolic models, few studies have actually taken on this question directly. Bybee (1995) reviews some arguments that type frequency is the most important consideration. Additional support for this view comes from the fact that individual high-frequency words do not seem to improve the productivity of isolated irregular patterns — for example, English has a high frequency verb *say* ~ *said*, but the novel verb *shay* could not have a past tense **shed*.

The goal of the current study is to provide a more rigorous comparison of type and token frequency, through computational modeling of experimentally obtained

morphological well-formedness ratings. It is organized as follows: first, I will present an assortment of lexical statistics which could plausibly form the basis of morphological well-formedness ratings. I will then describe an automated procedure for collecting these statistics from the lexicon, and producing predicted acceptability ratings. Finally, I will compare the relative effectiveness of these different statistics in modeling experimentally obtained acceptability ratings for two different morphological processes: past tense formation in English, and verbal conjugation class assignment in Italian. For both of these languages, I will show that a model based on type frequency provides the closest match to human intuitions, and employing token frequency does not improve the performance of the model.

2. Predicting well-formedness from lexical statistics

2.1 An assortment of lexical statistics

What kinds of statistics can be computed from lexicon? Consider, for example, the morphological change from [ɪ] → [æ] in the structural phonological environment / X [–syll,+cont] __ŋ# (where X is an unrestricted variable, standing for any amount of phonological material). There are a variety of statistics which we can collect for each such morphological change in a given phonological environment:

(2) Possible lexical statistics

a. *Scope(types)*: the number of words that meet the structural description of the rule
- 10 English verbs contain the environment $X\begin{bmatrix}\text{-syll}\\ \text{+cont}\end{bmatrix}$__ŋ# (*bring*, *cling*, *fling*, *ring*, *sling*, *spring*, *string*, *swing*, *wing*, and *wring*)

b. *Scope(tokens)*: the combined token frequency of all words that meet the structural description
- The 10 verbs listed in (a) have a combined lemma count of 561 in Francis & Kučera (1982)

c. *Hits (types)*: the number of words that meet the structural description of the rule and also participate in the same morphological change
- 6 verbs containing the environment $X\begin{bmatrix}\text{-syll}\\ \text{+cont}\end{bmatrix}$__ŋ# actually form their past tense by the change [ɪ] → [æ] (*swing*, *fling*, *wring*, *cling*, *sling*, and *string*)

d. *Hits (tokens)*: the combined token frequency of the "hits" for the environment
- the 6 verbs in (c) have a lemma count of 43

e. *Raw reliability (types)*: the ratio of the *Hits(types)* to the *Scope(types)* — i.e., the reliability of the change in this particular environment
- 6 out of 10 verbs containing environment $X\begin{bmatrix}\text{-syll}\\ \text{+cont}\end{bmatrix}$__ŋ# form past tenses by the change [ɪ] → [æ], so the reliability of [ɪ] → [æ] in this environment is 0.6

f. *Raw reliability (tokens)*: the ratio of *Hits(tokens)* to *Scope(tokens)*
- for [ɪ] → [æ] / $X\begin{bmatrix}\text{-syll}\\ \text{+cont}\end{bmatrix}$__ŋ#, 43/561 = 0.0766

g. *Adjusted reliability*: the statistical lower confidence limit of the reliability ratios (type and token), as suggested by Mikheev (1997).

(Rationale: we are more confidence about generalizations when there is more data — i.e., when an environment contains more words)

- the 75% confidence adjustment of 0.6 = 0.4825

h. *Type × token*: a measure taking both type and token frequency into account by multiplying them (*Adjusted type reliability × Adjusted token reliability*)

- for [ɪ] → [æ] / X$\left[\begin{smallmatrix}\text{-syll}\\ \text{+cont}\end{smallmatrix}\right]$__ŋ#, 0.6 × 0.766 = 0.04596

i. *Reward for length*: reward generalizations with more segments fully specified, by multiplying *Adjusted type reliability* × 1.2^n, for *n* shared segments. (Rationale: shared segments make the similarity between words more salient, and could help to increase the productivity of a pattern by inducing analogy to existing forms.)

- X$\left[\begin{smallmatrix}\text{-syll}\\ \text{+cont}\end{smallmatrix}\right]$__ŋ# has 1 full segment specified; 0.6 x 1.2 = .72

The measures in (2) are clearly only a small fraction of the possible ways to compute lexical statistics about the morphological behavior of words within a phonological environment; however this list provides a reasonable starting point.

But what environments should we collect lexical statistics for? In order to answer this, it is useful to consider what types of phonological environments can condition morphological alternations. The most restricted morphological process is suppletion, in which the environment for the change is limited to just one word. Morphological alternations may also be conditioned by more general phonological environments. For example, in English, null-marking for past tenses occurs only in [t]-final roots (*cut ~ cut*, *quit ~ quit*, etc.); in Toba Batak, *-um-* is prefixed only before vowels, labial consonants, and nasals (Crowhurst 1998). Morphological alternations can also be conditioned by rather general phonological environments, such as the alternation of the Korean nominative marker: *–ka* after vowel-final stems and *–i* after consonant-final stems. Finally, morphological processes can be completely insensitive to the phonological environment, as is the case for the invariant Hungarian accusative marker *–t*.

Thus it seems that if we want to find the correct generalization about the phonological environment where a morphological process occurs, we must consider environments at *all* levels of generality, from word-specific to context-free. There are several ways we could do this. We could, for instance, start by listing all logically possible structural descriptions for the language, and then see which ones are actually instantiated by members of each inflectional class. A more efficient way, however, is to use an automated discovery procedure to construct the list of relevant environments directly from the lexicon; I turn next to the description of one such procedure.

2.2 An algorithm for exploring environments.

One algorithm for exploring the phonological environments surrounding a morphological alternation is the 'minimal generalization' algorithm of Albright & Hayes (1998). The algorithm takes as its input pairs of morphologically related words. It starts by considering each word as a very specific environment for a morphological rule. For example, given the English (present, past) pair (*sip*, *sipped*), it posits a morphological rule which suffixes [t] in the phonological environment /sɪp__#:

(3) ∅ → t / sɪp__#

It then generalizes by seeking pairs of words that involve the same morphological change (in this case ∅ → [t]). When it finds such a pair, it compares the phonological environments to discover what material they share, and what material is unique to just one of the forms. It then posits a new rule with the shared material as its environment, converting the residue to a variable. For example, comparing (*sip*, *sipped*) and (*grip*, *gripped*), it would hypothesize that the morphological change ∅ → [t] can occur not just after *sip* and *grip*, but after any word ending in a coronal continuant + [ɪp]:

(4)

	change	residue	shared features	shared segments	change location
Comparing:	∅ → [t] /		s	ɪp	____
with:	∅ → [t] /	g	r	ɪp	____
yields:	∅ → [t] /	X	[+consonantal +coronal ...]	ɪp	____

The example in (4) shows that when we consider pairs of similar words, the resulting generalizations will be quite specific. When we consider pairs of dissimilar words, however, the shared material is minimal, and we can arrive at quite general — even context-free — generalizations. When the process of pairwise comparison is iterated across the entire lexicon, the result is a comprehensive list of thousands of phonological environments where each morphological process may occur in existing words.

Once the relevant phonological environments have been collected, we can calculate for each one the statistics described in section 2.1. The last remaining step, then, is to use these statistics to predict the well-formedness of novel words, in order to model human intuitions.

2.3 Predicting well-formedness from lexical statistics

The statistics described above in (2) pertain to phonological environments, not to particular words. When we gather well-formedness intuitions from people, however, we typically ask them about entire words. Unfortunately, each word contains many environments simultaneously — for instance, *glip* is [glɪp]-final, [lɪp]-final, [ɪp]-final, [p]-final, bilabial-final, stop-final, etc. Which environment do we look at for lexical statistics to predict the acceptability of *glipped*? What I will assume here is that we should try all applicable environments, and let the one with the highest score determine the predicted well-formedness. Example (5) shows four phonological environments contained within the novel verb *glip*, along with hypothetical reliability values. In this case, we would use the second environment (5b), to predict an acceptability value of 95% for the outcome *glipped*.

(5)
a. ∅ → [t] / X lɪp__# .89
b. ∅ → [t] / X ɪp__# .95 ← use this one
c. ∅ → [t] / X p__# .67
d. ∅ → [t] / X [+LAB] __# .65

This "best foot forward" convention provides a mechanism for predicting the well-formedness of novel words the score of the form is the score of the best rule that can

derive it. Note that any of the lexical statistics described above in (2) could be used in this way as the basis for predicted well-formedness ratings. In the remainder of this paper, I will compare the fit between predictions based on different lexical statistics and experimentally obtained well-formedness ratings from human speakers.

3. English past tenses

Prasada & Pinker (1993) presented 60 novel verbs to English speakers in a present context ("*John likes to* cleef"). Participants then rated the acceptability of potential past forms, on a scale of one to seven:

(6) Yesterday, John *cleefed*	1	2	3	4	5	6	7
Yesterday, John *cleft*	1	2	3	4	5	6	7

Prasada & Pinker claim that ratings of novel irregular past tense forms like *cleft* were influenced by the phonological form of the word, while ratings of regular past tense forms like *cleefed* showed no such effect. If this is true, then we should be able to predict the ratings of novel irregulars using some version of the lexical statistics described above, while the ratings of novel regulars should not be predictable based on lexical statistics.

In order to test this hypothesis, I applied the automatic environment-exploring algorithm to a database of 2,181 (present, past) pairs of English verbs, in phonetic transcription. This database was based on a file taken from Brian MacWhinney's web site[1], augmented slightly to include all of the irregular verbs of English. The database also included the token frequency for each verb, as listed in Francis and Kučera (1982). The result was a comprehensive list of all of the phonological environments surrounding each change used to express the present/past distinction in English (including both the regular and irregular patterns). Statistics were then calculated for each environment, using both type and token frequency. Finally, each of the novel verbs from Prasada & Pinker's study was submitted to the system, in order to determine their predicted ratings under each of the different bases for well-formedness proposed in (2).

The first result was that there were highly significant correlations between the actual well-formedness ratings and *all* versions of the predicted well-formedness ratings which are based on ratios ($p < 10^{-11}$ in all cases, much more significant for some measures).[2]

(7) Correlation of predicted to actual acceptability ratings

Basis of predictions	Correlation (Pearson's *r*) (*d.f.* = 58)
raw reliability (type)	0.6109
raw reliability (token)	0.5950
adjusted reliability	0.7319
type × token	0.7230
reward for length	0.7321

As can be seen, predictions based on type and token frequency both do well in modeling the actual acceptability ratings, with type frequency slightly ahead of token frequency; adjustments based on confidence limits, or the specificity of the phonological environment, also help the model considerably.

Prasada & Pinker mention a possible confound in their data, however. They point out that when subjects are asked to rate morphological well-formedness, they may have difficulty factoring out the independent effect of phonological well-formedness. Therefore, they also collected ratings of phonological well-formedness, which can be

used to correct for this confounding influence. In particular, we can perform partial correlations, factoring out the phonological well-formedness ratings to leave what should be a purely morphological effect.

When phonological well-formedness is factored out in this way, the correlations between the predicted ratings and the observed ratings actually go up slightly. As before, type frequency is a slightly (but non-significantly) better predictor than token frequency. Furthermore, a significant correlation is observed even when regular and irregular forms are considered separately:

(8) Partial correlations, factoring out phonological well-formedness

Basis of predictions	Correlation (all forms)	Correlation (regs only)	Correlation (irregs only)
raw reliability (type)	0.6310	0.4798	0.3526
raw reliability (token)	0.6141	0.4170	0.3971
adjusted reliability	0.7443	0.3904	0.5729
type × token	0.7455	0.4033	0.5516
reward for length	0.7401	0.2583	0.5741

Thus, *contra* Prasada & Pinker, even novel regulars show a significant effect of lexical statistics. Note that this could *not* emerge if ratings for novel regulars were uniform for all words, as is predicted by the dual mechanism hypothesis, since significant correlations can only be achieved when there is adequate variance in the data. In fact, it seems that when phonological well-formedness is factored out in a partial correlation, there is *more* variance between the items, and this variance is captured well by the predictions of lexical statistics.

The evidence from English can thus be summarized as follows: all bases for lexical statistics perform relatively well, with type frequency slightly better than token frequency. Furthermore, lexical statistics can account for differences not only between novel irregulars, but also between novel regulars.

4. Italian conjugation classes

Italian has four conjugation classes, differing in theme vowel and stress in the infinitive:

(9)

Vowel	Stress	Infinitive suffix	Sample root	Sample infinitive	Gloss
[a]	*suffix*	-are	sed-	seˈdare	‘sedate’
[e]	*root*	-ere	led-	ˈledere	‘harm’
[e]	*suffix*	-ere	sed-	seˈdere	‘sit’
[i]	*suffix*	-ire	sped-	speˈdire	‘send’

Although the four classes are distinct in the infinitive, the distinction is neutralized in various ways in other inflections; the 1sg form, in particular, is ambiguous between all four conjugation classes.

Albright (1999) presented novel verbs to consultants in the 1sg form, and asked them to rate the acceptability of potential infinitives:

(10) a. *Oggi* rabado *con mio fratello.*
today *rabad*.1sg with my brother

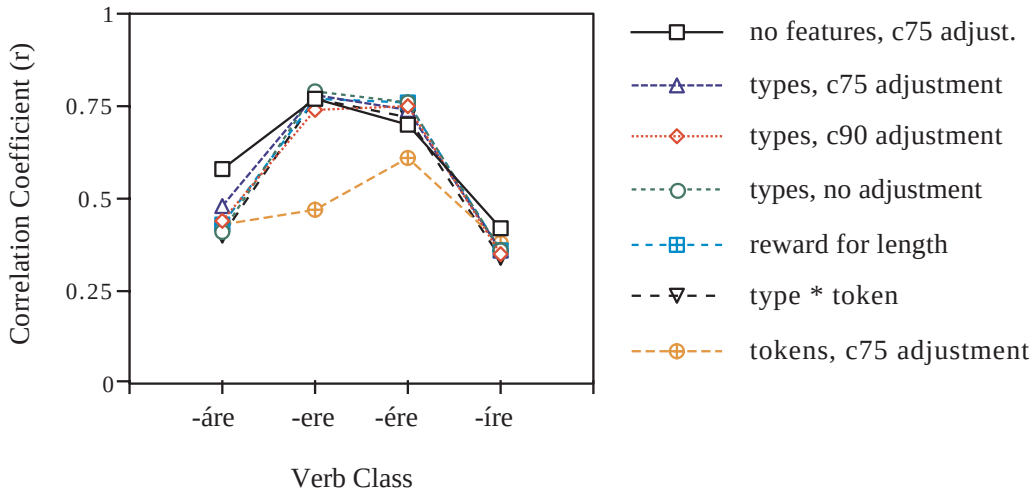
b.	*Mi piace* rabadare	1	2	3	4	5	6	7
	Mi piace rabadere	1	2	3	4	5	6	7
	Mi piace rabadère	1	2	3	4	5	6	7
	Mi piace rabadire	1	2	3	4	5	6	7
	"I like to *rabad*"							

As in Prasada & Pinker's study, participants also rated the phonological well-formedness of the verb, in this case in its 1sg form.

In order to model well-formedness of different conjugation class assignments, I created a database of 2,900 Italian verbs in the 1sg present and infinitive, with phonetic transcriptions and token frequencies (de Mauro et al. 1993). This database contained all of the verbs contained in a 500,000 word spoken corpus, plus all verbs in the Ispell electronic dictionary (Kuenning 1996). As before, predicted ratings were calculated for each novel item in each conjugation class, using the lexical statistics described above.

As with English, the results show similar performance from all metrics, with moderately strong correlations for all conjugation classes:

(11) Comparison of different metrics



Metrics based on token frequency perform a bit worse than those from type frequency, especially for the *–ere* and *–ère* classes. The explanation for this is probably that these classes contain relatively few verb types, but they have high token frequency. Therefore, it is in these classes that type and token frequency diverge most radically — and in this case, it is type frequency that seems to model human intuitions most closely.

5. Discussion

In both Italian and in English, type frequency and token frequency both do quite well at explaining human ratings. In fact, the similarity is not mysterious; predictions based on type frequency are themselves highly correlated with the predictions based on token frequency. The reason for this is that most words in a corpus have a token frequency of one. Therefore, predictions based on token frequency differ from type frequency only for those few neighborhoods that contain high-frequency verbs. Thus, we expect that using token frequency should only make a small difference — and to the extent that it makes a difference at all, it seems to make the predictions worse. It should also be noted that neither experiment actually included novel forms like *shay*, which would tease apart the predictions the most. In addition, the use of log frequencies would not help here,

because the lexical statistics employed here are all ratios, so it would be pointless to take the log of both the numerator and the denominator of the ratio. The fact that human intuitions are best modeled by type frequency may suggest that the statistics are calculated at the symbolic level — i.e., abstracted away from tokens. Furthermore, the fact that confidence statistics improved the accuracy of the predictions could possibly reflect reasoning behavior, and not simply the neuronal activation from similar words.

More generally, the method of computing lexical statistics described in section 2 was found to provide a good match to human ratings in both English and Italian. In both languages, this was true not only for irregular patterns, but also for the default/regular pattern. Taken together, these results support the view that morphological well-formedness intuitions for all patterns can be derived by a single, probabilistic model.

References

Albright, Adam. 1999 "Phonological subregularities in inflectional classes: Evidence from Italian". *UCLA Working Papers in Linguistics 1, Papers in Phonology 2*, ed. by Matthew Gordon, 1-47. http://www.humnet.ucla.edu/people/aalbrigh/papers.html

---------- & Bruce Hayes. 1998. "An automated learner for phonology and morphology". Ms., UCLA ms. Available at: http://www.humnet.ucla.edu/humnet/linguistics/people/hayes/learning/learning.htm

Bybee, Joan L. 1995. "Regular morphology and the lexicon". *Language and Cognitive Processes* 10:5.425-455.

---------- & Carol Lynn Moder. 1983. "Morphological classes as natural categories". *Language* 59:2.251-270.

Bybee, Joan L. and Dan I. Slobin. 1982. "Rules and schemas in the development and use of the English past tense". *Language* 58:2.265-289.

Crowhurst, Megan J. 1998. "*Um* infixation and prefixation in Toba Batak". *Language* 74:3.590-604.

de Mauro, Tullio and F. Mancini and M. Vedovelli and M. Voghera. 1993. *Lessico di frequenza dell'italiano parlato*. Milan: Etaslibri.

Francis W. Nelson and Henry Kučera. 1982. *Frequency Analysis of English Usage*. New York: Houghton Mifflin.

Kuenning, Geoff. 1996. "International Ispell". http://ficus-www.cs.ucla.edu/ficus-members/geoff/ispell.html

Mikheev, Andrei. 1997. "Automatic rule induction for unknown-word guessing". *Computational Linguistics* 23:3.405-423.

Prasada, Sandeep and Steven Pinker. 1993. "Generalization of regular and irregular morphological patterns". *Language and Cognitive Processes* 8.1-56.

Ullman, Michael T. 1999. "Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighborhood effects". *Language and Cognitive Processes* 14:1.47-67.

Adam Albright
Linguistics Dept, UCLA
Los Angeles, CA 90095-1543
Email: aalbrigh@ucla.edu Web: http://www.humnet.ucla.edu/people/aalbrigh/

[1]http://psyling.psy.cmu.edu/Brian/papers.html

[2] Measures such as *Hits(tokens)* that are not based on ratios predict the same well-formedness rating for all items — namely, the score of the largest, context-free generalization. Therefore, they can not be compared using correlations, because there is no variance in the values.