

Productivity and Reuse in Language: A Bayesian Framework

Timothy J. O’Donnell (timo@wjh.harvard.edu)

Department of Psychology, Harvard University

Introduction

Perhaps the most celebrated aspect of human language is its creativity. Language allows us to produce, comprehend—and perhaps even think—an unbounded number of thoughts. Creativity in language is made possible by computation. Novel expressions can be produced and understood because the linguistic system provides *productive* processes for generating linguistic structures. These productive processes operate via the combination of large numbers of stored, reusable units. A fundamental problem for linguistic and psycholinguistic theory, as well as for the language learner, is understanding which patterns in linguistic data should give rise to productive generalizations and which should give rise to stored, reusable structures.

We present a model of productivity and reuse which treats the problem as an inference in a Bayesian framework. The model is an example of the *rational analysis* approach in the sense of Anderson (1990). That is, the model can be seen as asking the question: what would an agent who was making optimal use of information conclude about productivity and reuse from the patterns in the input data? The specific model applied here, known as *Fragment Grammars*, is a generalization of a *Adaptor Grammars* (Johnson et al., 2007a). In other work we show how Fragment Grammars can be seen as a special case of an even more general framework which allows inferences about productivity and reuse to be integrated with *any* probabilistic generative model (O’Donnell, 2011).

A productive computation is one which can give rise to novel forms. In a Bayesian setting, if a system hypothesizes that some (sub)computation is productive, it must reserve probability mass for hitherto unseen structures. On the other hand, if a Bayesian system hypothesizes that some sequence of computations will be reused together, it must reserve probability mass to that particular sequence as a whole. Since there is only a finite budget of probability, this necessarily leads to a tradeoff: a probabilistic system hypothesizes reusability at the cost of generalization and productivity at the cost of reusability. The Fragment Grammar model can be seen as optimizing this tradeoff for a given dataset. Importantly, the currency which is optimized is *not* the cost of computation or memory, as is commonly discussed (e.g. Fraunfelder & Schreuder (1992)). Rather the system optimizes its ability to correctly *predict* which computations will generate novel structures and which sequences of computations will be reused as units.

The Models

In addition to the Fragment Grammar model, we systematically consider three other approaches. These have been chosen to represent the range of theoretical proposals that have been made in the literature: *full-parsing*, *full-listing*, *exemplar-based storage*. We have formulated and implemented all three of these proposals as probabilistic models which have otherwise been kept maximally similar to Fragment Grammars. Mathematical details can be found in O’Donnell (2011); O’Donnell et al. (2009).

Full-parsing In the first alternative model, *full-parsing*, all productive generalizations that are inherent in the starting state are maintained, and all structure is always computed (Figure 1). In such a setting each stored structure is highly reusable. However, any computation will involve choosing many small, abstract items and therefore a *particular* form may be highly improbable. Such an approach to reuse will be most effective when all of the *potential* productivity in the system is in fact *real* productivity. We formalize the full parsing model using *Multinomial-Dirichlet Probabilistic Context-Free Grammar* (MDPCFG) (Johnson et al., 2007b).

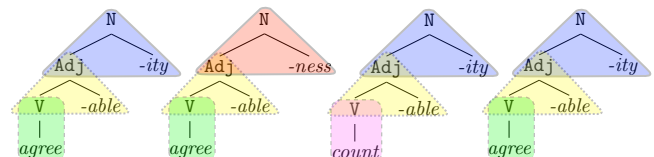


Figure 1: Full Parsing: This figure shows the consequences of the full-parsing theory. Subcomputations can be shared across many forms, as is shown by highlighting in the figure. However, each form requires many computational choices to generate.

Full-listing In the second alternative model, *full-listing*, once a structure is built, it is stored in its entirety. The underlying computational system can still account for productive generalization, but the system is very conservative, in that it assumes that anything built once is reusable as a whole (Figure 2). In such a setting, each stored structure is extremely specific, and therefore can only be reused in limited contexts. Such an approach to reuse will be most effective when the language consists of a small number of specific, but frequently used forms. To formalize the full-listing model we choose *Adaptor Grammar* (AG) (Johnson et al., 2007a).

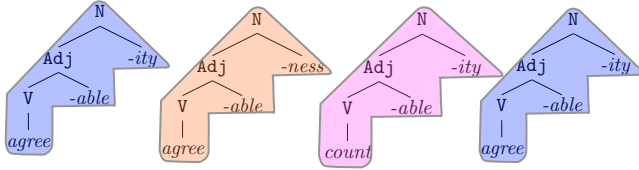


Figure 2: Full Listing: This figure shows the consequences of the full-listing model. Stored subcomputations are very specific, their substructures cannot be shared with other forms, however, they can be reused in their entirety with high probability.

Exemplar-based Inference The third alternative model, *exemplar-based inference*, stores *all* structures which are consistent with the data, both small and abstract, and large and specific (and all in between). This model differs from the earlier two in that it can represent both productive generalizations and specific structures. However, it differs from Fragment Grammars in that it spreads its beliefs across all possibilities, rather than committing to one analysis or another for each linguistic expression.

We formalize exemplar-based models using two different variants of the *Data-Oriented Parsing* (DOP) formalism for tree substitution grammar estimation (Bod et al., 2003). The first uses an estimator known as Data-Oriented Parsing 1 (DOP1) which assigns weights to fragments of structure which are proportional to their frequency in the input. The second is a variant of DOP which gives equal weight to each node in the input training corpus (Bod, 2003; Goodman, 2003). This latter estimator was proposed by Joshua Goodman and we will refer to it as the *Goodman Estimator* (GDMN).

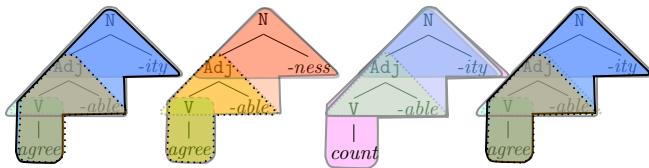


Figure 3: Exemplar-based Inference: This figure shows exemplar-based inference. Here every possible subtree consistent with the data is stored. Note that this leads to many overlapping analyses for each item.

Productivity and Reuse as an Inference Our model, Fragment Grammar, treats the problem of which subcomputations are productive and which should be stored for reuse as an inference. It is able to store abstract structures like the full-parsing (MDPCFG) approach, as well as large specific structures like the full-listing (AG) approach, and all intermediate structures as in the exemplar-based (DOP1/GDMN) approach. However, unlike the latter it must commit to a single analysis for each instance of a structure. The tradeoffs inherent in this arrangement are shown in Figure 4.

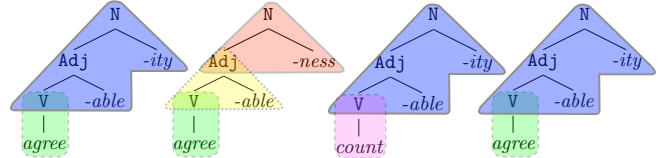


Figure 4: Productivity and Reuse as an Inference: This figure shows the consequences of inferring the set of subcomputations that best account for the data. In this example, more sharing is allowed on average than full-parsing with less computation on average than full-listing.

English Past Tense: Adult Data

In this section we present the results of applying the models to the well-known English Past tense dataset. We focus on one particular aspect of this dataset which is particularly well-suited to our model: the differential productivity of regular and irregular inflectional markers.

The English regular, *-ed*, past tense rule is rampantly productive. A large number of studies have shown generalization of the regular rule to novel stems in both production and rating tasks (e.g. (Albright & Hayes, 2003; Prasada & Pinker, 1993)). The regular rule can be applied to forms which are phonotactically odd for English and to stems derived from other morphological processes (Kim et al., 1991).

Irregular inflectional classes, on the other hand, are much less productive. While a number of studies have shown that irregular classes can be generalized, this generalization is much more limited than the case of the regular rule and is more sensitive to the phonological and semantic structure of the stem (Albright & Hayes, 2003; Bybee & Moder, 1983; Prasada & Pinker, 1993).

The simulations were performed on data extracted from the annotated version of the SwitchBoard corpus (Godfrey et al., 1992; M. P. Marcus et al., 1999). All verbs excluding forms of *be*, *have* and *do* were extracted from the corpus, lemmatized and paired with the appropriate inflection for the stem. The model was trained on the full English verbal paradigm (i.e. all tenses). Figure 5 shows the simple input representation used for all models. Note that this representation merely pairs stems with their correct inflectional information, it does not explicitly encode any phonological or semantic selectional restrictions. Any success that any of the models has learning contingencies between stems and inflections is the result of distributional information in the input.

Table 1 shows the probability that a past tense or past participle form sampled at random from the posterior generative model will be correctly inflected. These scores are broken down into regular and irregular forms that were in the training sample and a set of novel test cases.

Although the full-listing model AG is able to perform well on attested forms by memorizing the data, its gen-

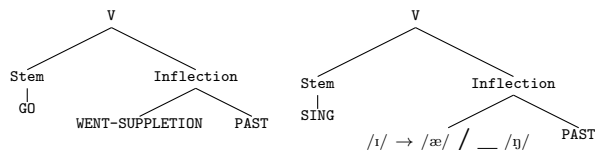


Figure 5: Example Trees for Past Tense: This figure shows examples of trees used as inputs to the past tense simulations.

eralization performance is limited. The DOP1 exemplar-based model by contrast performs well on the regular forms and generalization cases but fails to learn the irregular forms. Only Fragment Grammar is able to provide reasonable performance across all of the test sets.

Model	FG	MDPCFG	AG	DOP1	GDMN
Irr	0.999	0.042	0.995	0.057	0.175
Reg	0.999	0.488	0.995	0.998	0.662
Novel	0.894	0.446	0.790	0.920	0.480

Table 1: Performance of Models on Past Tense Dataset: This table shows the probability that a form randomly sampled from each trained model will be inflected correctly.

English Past Tense: Developmental Data

Perhaps the most well-known past tense developmental phenomenon is the occurrence of *overregularization* errors during development—leading to a U-shaped developmental trajectory. Under a U-shaped developmental trajectory overregularization is preceded by a period of correct performance on irregulars, and followed by recovery to correct usage on these forms. There has been much debate about the exact characteristics of U-shaped development for the English past tense. However, there is broad consensus on the following points (Hoeffner, 1996; G. F. Marcus et al., 1992). 1. There is a clear period of early correct performance before the first overregularization for some children and some verbs, although global early correct performance is more controversial. 2. Overall rates of overregularization are relatively low throughout development. 3. The onset of overregularization is not associated with any sudden discontinuities in the input data. 4. Overregularization occurs as performance on regular marking improves. 5. Recovery from overregularization is gradual.

The simulations were performed on data extracted from the CHILDES corpus (MacWhinney, 2000). All verbs which had available morphological information and child ages were extracted from adult (child-directed) speech and prepared in a similar manner to the SWITCHBOARD simulations. An empirical relative frequency distribution was constructed at each child age in months from 1;6 to 5;0. An age-specific training set of size 1000 was constructed from the relative frequency

distribution. The model was then trained on cumulative datasets for each age.

Figure 6 shows the developmental trajectory of the five models as well as the empirical trajectory over all data provided in the appendices of G. F. Marcus et al. (1992). The blue lines represent the performance on irregulars and the red lines represent performance of each model on generalization of the regular rule to a novel, *wug* stem (note that the equivalent empirical *wug* data is not available). What can be seen is that only the Fragment Grammar model provides a developmental trajectory which shows the correct overall pattern with low, but not non-existent, rates of overregularization. The only other model which performs reasonably well on both regulars and irregulars, AG, does so with essentially no overregularization errors.

English Derivational Morphology

In this section we turn to English derivational morphology which provides a richer system in which to study questions of productivity and reuse than the English past tense. We focus here on two aspects of the system. 1. *Productivity*: Derivational suffixes reside on a productivity cline from very productive (e.g. *-ness*) to very unproductive (e.g. *-th*). 2. *Ordering*: Only a small fraction of the suffix combinations which are possible in principle are attested in actual words. One theory which accounts for this fact is *complexity-based ordering* (CBO) (Hay, 2002; Plag & Baayen, 2009). CBO proposes that on average more separable suffixes should appear outside of less separable suffixes. A corollary is that more productive suffixes should be more likely to appear after less productive suffixes.

Input data for our derivational morphology simulations was derived from the CELEX database (Baayen et al., 1993). The morphological parses provided by CELEX were supplemented by applying a heuristic parsing algorithm to the remaining unparsed forms in the database. Approximately 10,000 of the resulting forms were then corrected by hand. We limit attention here to the set of suffixes studied in Plag & Baayen (2009). The input to the simulation consisted of trees like those shown in Figure 7.

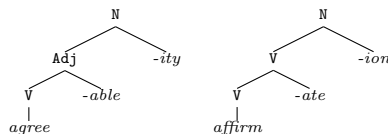


Figure 7: Example Trees for Derivational Morphology: This figure shows examples of the trees used as inputs to the derivational morphology simulations.

Productivity We first consider the productivity of suffixes as inferred by the various models. There is

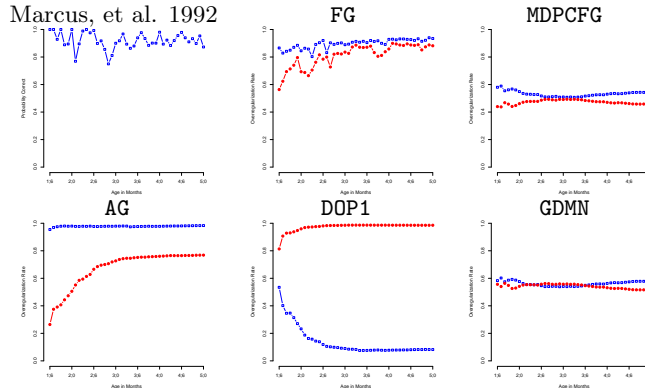


Figure 6: Developmental Trajectory: This figure shows the developmental trajectory of irregulars (blue) and *wug*-tests (red) for each model, as well as the empirical rates of correct irregular usage from G. F. Marcus et al. (1992). The x-axis is age in months, and the y-axis is probability correct.

no gold standard measure of productivity against which the models can be evaluated. However, two widely used empirical productivity statistics are Baayen’s \mathcal{P} and \mathcal{P}^* (e.g. Baayen (1992)). The former can be understood as an estimate of the probability that a particular suffix will be used to produce a novel form (i.e. $P(\text{NOVEL} \mid \text{SUFFIX})$). The latter can be understood as an estimate of the probability that a novel form will use a particular suffix (i.e. $P(\text{SUFFIX} \mid \text{NOVEL})$).

Model	FG	MDPCFG	AG	DOP1	ENDOP
\mathcal{P}	0.907	-0.0003	0.692	0.346	0.143
\mathcal{P}^*	0.543	0.500	0.485	0.478	0.495

Table 2: Correlation with Productivity Measures: The correlation between quantities computed from the trained models with empirical estimates of Baayen’s \mathcal{P} and \mathcal{P}^* given in Hay & Baayen (2002).

Table 2 shows the Pearson correlation between the quantities $P(\text{NOVEL} \mid \text{SUFFIX})$ and $P(\text{SUFFIX} \mid \text{NOVEL})$ computed from the various (posterior) models and empirical estimates of \mathcal{P} and \mathcal{P}^* given in Hay & Baayen (2002). Fragment Grammar provides the best fit to these quantities. Table 3 shows the five most productive suffixes learned by the Fragment Grammar model. On the other extreme the least productive suffixes include *-th:A>N*, *-dom:A>N*, and *-en:N>V*. These results match intuition.

Suffix	Category	Example
<i>-ly</i>	Adj>Adv	<i>quickly</i>
<i>-er</i>	V>N	<i>dancer</i>
<i>-ness</i>	Adj>N	<i>quickness</i>
<i>-er</i>	N>N	<i>jeweler</i>
<i>-ment</i>	V>N	<i>advancement</i>

Table 3: Most Productive Suffixes: This table shows the five most productive suffixes inferred by the Fragment Grammar model.

Ordering Plag & Baayen (2009) provide an empirical measure of ordering, based on graph theoretic tools, which gives an estimate of the *mean rank* of a suffix. The mean rank can be understood as a measure of how easily a particular suffix appears after other suffixes in complex words. To generate predictions from the model with regard to affix ordering, we computed the conditional probability that each suffix appears first or second in words with exactly two suffixes. Table 4 gives the Spearman rank correlation of the log odds that a suffix appears second (against first) with the mean rank statistics reported in Plag & Baayen (2009).

Model	FG	MDPCFG	AG	DOP1	ENDOP
Mean Rank	0.568	0.275	0.424	0.452	0.431

Table 4: Correlation of Ordering Probabilities and Ranks: The correlation between the probability that a particular suffix will appear first or second and the mean rank statistic for the suffix given in Plag & Baayen (2009).

These ordering results follow from the differential productivity of different suffixes. Productive affixes are represented by structures with a variable in their base position. All else being equal, this base position can be filled with a bare stem or a more complex structure which contains other suffixes. Less productive suffixes on the other hand tend to be stored together with their bases. They can easily be inserted into the variable position of a more productive suffix, but they follow other suffixes only with very low probability.

Acknowledgments

The Fragment Grammar model was developed in collaboration with Noah Goodman. We would like also to thank Marjorie Freedman, Jesse Snedeker, Manizheh Khan and Josh Hartshorne for detailed feedback on this work.

References

- Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2), 119–161.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology 1991* (pp. 109–149). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Baayen, R. H., Piepenbrock, R., & Rijn, H. van. (1993). *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania.
- Bod, R. (2003). An efficient implementation of a new DOP model. In *Proceedings of the 10th conference of the European chapter of the association for computational linguistics* (Vol. 1, pp. 19–26). Budapest, Hungary.
- Bod, R., Scha, R., & Sima'an, K. (Eds.). (2003). *Data-oriented parsing*. Stanford, CA: CSLI.
- Bybee, J. L., & Moder, C. L. (1983, June). Morphological classes as natural categories. *Language*, 59(2), 251–270.
- Frauenfelder, U. H., & Schreuder, R. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In *Yearbook of morphology 1991* (pp. 165–183). Dordrecht, The Netherlands: Springer.
- Godfrey, J., Holliman, E., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. *IEEE ICASSP*, 517–520.
- Goodman, J. (2003). Efficient parsing of DOP with PCFG-reductions. In *Data-oriented parsing*. Stanford, CA: CSLI Publications.
- Hay, J. (2002, September). From speech perception to morphology: Affix ordering revisited. *Language*, 78(3), 527–555.
- Hay, J., & Baayen, R. H. (2002). Parsing and productivity. In *Yearbook of morphology 2001* (Vol. 35, pp. 203–236). Dordrecht, The Netherlands: Springer.
- Hoeffner, J. (1996). *Are rules a thing of the past? A single mechanism account of English past tense acquisition and processing*. Unpublished doctoral dissertation, Carnegie Mellon University.
- Johnson, M., Griffiths, T. L., & Goldwater, S. (2007a). Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in neural information processing systems 19*. Cambridge, MA: MIT Press.
- Johnson, M., Griffiths, T. L., & Goldwater, S. (2007b). Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Proceedings of the North American conference on computational linguistics*. Rochester, New York.
- Kim, J. J., Pinker, S., Prince, A., & Prasada, S. (1991). Why no mere mortal has ever flown out to center field. *Cognitive Science*, 15, 173–218.
- MacWhinney, B. (2000). *The childes project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Marcus, G. F., Pinker, S., Ullman, M. T., Hollander, M., Rosen, T. J., & Xu, F. (1992). *Overregularization in language acquisition*. Chicago, IL: University of Chicago Press.
- Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., & Taylor, A. (1999). *Treebank-3* (Tech. Rep.). Philadelphia: Linguistic Data Consortium.
- O'Donnell, T. J. (2011). *Productivity and reuse in language*. Unpublished doctoral dissertation, Harvard University.
- O'Donnell, T. J., Goodman, N. D., & Tenenbaum, J. B. (2009). *Fragment grammars: Exploring computation and reuse in language* (Tech. Rep. No. MIT-CSAIL-TR-2009-013). Cambridge, Ma: MIT Computer Science and Artificial Intelligence Laboratory Technical Report Series.
- Plag, I., & Baayen, R. H. (2009, March). Suffix ordering and morphological processing. *Language*, 85(1), 109–152.
- Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8(1), 1–56.