# Probabilistic ParaMor

Christian Monson, Kristy Hollingshead, and Brian Roark
Center for Spoken Language Understanding
Oregon Health & Science University
monsonc@ohsu.edu

## Abstract

The ParaMor algorithm for unsupervised morphology induction, which competed in the 2007 and 2008 Morpho Challenge competitions, does not assign a numeric score to its segmentation decisions. Scoring each character boundary in each word with the likelihood that it falls at a true morpheme boundary would allow ParaMor to adjust the confidence level at which the algorithm proposes segmentations. A sliding threshold on segmentation confidence would, in turn, permit a trade off between precision and recall that could optimize $F_1$ or other metrics of interest. Our submission to Morpho Challenge 2009 enriches ParaMor with segmentation confidences by training an off-the-shelf statistical natural language tagger to mimic ParaMor's morphological segmentations. For a given word, the tagger's probabilistic confidence that ParaMor would propose the character, c, as the first character of a new morpheme serves as the numeric score of the candidate morpheme boundary that immediately precedes c. We have trained a ParaMor tagger mimic over a development data set of 500,000 unique Hungarian word types. By adjusting the threshold above which the ParaMor mimic proposes morpheme boundaries, we improve ParaMor's $F_1$ score for Hungarian by 5.9% absolute, from 41.4% to 47.3%. Moreover, by training a probabilistic tagger to emulate the segmentations of a second unsupervised morphology induction system, Morfessor, we are able to combine ParaMor's segmentation decisions with Morfessor's to form a single joint segmentation of each word. Our joint ParaMor-Morfessor tagger mimic enhances $F_1$ performance on our Hungarian development set by a further 3.4% absolute, ultimately achieving an $F_1$ score of 50.7%.

## Categories and Subject Descriptors

I.2 [Artificial Intelligence]: I.2.7 Natural Language Processing

## General Terms

Experimentation

## Keywords

Natural language morphology, Unsupervised learning, Morphology induction, Statistical Mimic

## 1. Introduction

Unsupervised morphology induction is the task of learning the morphological analyses of the words of an unknown natural language from nothing more than a raw corpus of unannotated text. Analyzing words down to the morpheme level has helped a range of natural language processing tasks including machine translation (Oflazer and El-Kahlout, 2007), information retrieval (Kurimo and Turunen, 2008), and speech recognition (Creutz, 2006). But building a morphological analysis system by hand can take person-months of time—hence the need for automatic methods for morphology induction.

A wide variety of approaches to unsupervised morphology induction have been proposed in recent years. Techniques inspired by Zellig Harris' early work (Harris, 1955), measure the probabilities of word-internal character transitions to identify likely morpheme boundaries (Bernhard, 2008). Other systems rely on the minimum description length principle to pick out a set of highly descriptive morphemes (Goldsmith, 2001; Creutz, 2006). Recent work on unsupervised morphology induction for Semitic languages has focused on estimating robust statistical models of morphology (Snyder and Barzilay, 2008; Poon et al., 2009). And this paper extends a morphological induction system called ParaMor that leverages morphological paradigms as the inherent structure of natural language morphology (Monson, 2009).

## 1.1. ParaMor

The ParaMor algorithm is a linguistically motivated unsupervised morphology induction system. By counting the frequency of word-final strings on shared word-initial strings in a list of unannotated words, ParaMor automatically builds sets of suffixes that model the paradigm structure found in inflectional morphology. For example, from one corpus of Spanish newswire, ParaMor discovers that each member of a set of 41 word-final strings that includes *a, aba, aban, acion, aciones, ación, ada, adas, ado, ador, ...* attaches to a set of candidate stems that covers the forms *aboy, celebr, desarroll,* and *genera.* Although ignorant of syntactic and lexical features, ParaMor has discovered the set of verbal suffixes that attach to Spanish *ar* verbs.

The ParaMor algorithm competed in both the 2007 and 2008 Morpho Challenge Competitions, both solo and in a joint submission with a second unsupervised morphology induction system Morfessor. Setting aside the joint ParaMor-Morfessor system, the solo ParaMor system placed first in the Turkish Linguistic competition of Morpho Challenge 2008, at 46.5% $F_1$, and second in English, with an $F_1$ score of 52.5%. Meanwhile the joint ParaMor-Morfessor system placed first overall in the 2008 Linguistic competitions for German, Finnish, Turkish, and Arabic.

ParaMor's successes are particularly remarkable given that ParaMor is a rule-based system incapable of measuring the confidence of the morphological segmentations it proposes. Without a confidence measure on individual segmentation decisions, it is impossible to increase or decrease the number of segmentation points that ParaMor proposes so as to optimize ParaMor's precision-recall performance for a given task. A tradeoff between precision and recall is inherent in any classification task, including morphological segmentation. If system A proposes segmenting a corpus at a superset of the character boundaries at which system B proposes segmentations, some of system A's additional proposed boundaries will match true morpheme boundaries but others will not—increasing recall but decreasing precision.

The ability to trade off precision against recall is clearly relevant for a morphological analysis system. A morphological analysis system embedded in a patent search application, for example, should likely favor stem recall over precision: seeking to return as many documents that potentially match a query as possible. On the other hand, a morphology system that enhances the language model of a text-input system for cell-phones should likely focus on precise and accurate input—lowering the number of incorrect word proposals that the user must correct by hand. And a system designed to perform well in the linguistic competition of Morpho Challenge, should balance precision and recall of morphemes. Morpho Challenge is evaluated by $F_1$, the harmonic mean of precision and recall, and the harmonic mean imposes a strict penalty when the gap between the precision and recall scores in large. Consequently, $F_1$ is nearly always maximum when precision and recall are balanced. Our submission to Morpho Challenge 2009 imbues ParaMor's segmentation decisions with probabilistic confidence scores by enlisting the help of a natural language tagger.

## 2. Probabilistic ParaMor

At each character, *c*, in each word, a morphology segmentation algorithm, such as ParaMor, makes a binary decision to either place or to not place a morpheme boundary before *c*. We view the morphology segmentation task as a labeling problem akin to part-of-speech tagging. In part-of-speech tagging each sequential word in a sequence must be labeled with its part of speech, V, N, Adj, etc. In morphological segmentation, each sequential character in a word must be labeled as beginning a new morpheme or as continuing the current one.

There are two advantages to reformulating segmentation as a tagging problem. First, taggers are a proven and well-understood natural language processing technique that have been adapted to a variety of problems beyond part-of-speech labeling. Taggers have been used for named entity recognition (Tjong Kim Sang, 2002) and NP-chunking (Tjong Kim Sang and Buchholz, 2000); to flag words on the periphery of a parse constituent (Roark and Hollingshead, 2009); as well as to segment written Chinese into words (Xue, 2003)—a task closely related to morphology segmentation. Second, standard natural language taggers are statistical models that can output a probability distribution over all possible labels for each tagged item. Thus, a statistical tagger trained to label morpheme boundaries outputs a probabilistic confidence score that any particular character begins a morpheme.

Just one problem remains: Statistical taggers are *supervised* induction methods while the Morpho Challenge competitions explicitly forbid supervised induction. Where unsupervised induction methods learn from unlabeled examples, i.e. unadorned words for morphology induction, supervised methods require labeled training data. In the case of morphological segmentation, labeled training data would consist of a set of words with each character labeled as the start of or as the continuation of a morpheme.

While data labeled with the truth is forbidden in Morpho Challenge, we can construct *artificial* training data from the segmented output of an unsupervised morphology induction algorithm such as ParaMor; and then use the artificially labeled data to train a statistical tagger to *mimic* the segmentations that the unsupervised method produces. The probabilistic segmentation scores that the tagger mimic assigns to each character can then serve as a numeric confidence of original unsupervised method.

## 2.1. Training a Tagger Mimic

Using the unsupervised morphology induction algorithm ParaMor as a source of labeled data, we trained a finite-stage tagger (Hollingshead et al., 2005) to identify, for each character, *c*, in a given word, whether or not ParaMor would place a morpheme boundary immediately before c. Additionally we had our tagger learn whether each proposed boundary began a stem or an affix. Thus we trained a statistical model to "tag" each character as beginning a new stem morpheme, beginning a new suffix morpheme, or as not occurring at the left edge of a morpheme. The feature set used in the tagger consisted of the surrounding sequences of characters and morpheme- tags. The character sequences are represented by character n-grams up to three characters on either side of the current character. Thus in a word like "quickly", the character-features for tagging the letter 'c' would be: 'quic', 'uic', 'ic', 'c', 'ck', 'ckl', and 'ckly'. The morpheme-tag features are represented as unigram, bigram, and trigram morpheme-tags (i.e., tags from the current and two previous characters).

We used the averaged perceptron algorithm, as presented in Collins (2002), to train the tagger. During training, the decoding process is performed using a Viterbi search with a second-order Markov assumption. At test-time, we use the forward-backward algorithm, again with a second-order Markov assumption, to output the perceptron-score of each morphological tag for each character in the word. The main benefit of decoding in this manner is that, by normalizing the scores at each character (using softmax due to the log linear modeling), we can extract the probability of each tag at each character rather than just the single perceptron-preferred solution for the entire word.

## 2.2. Efficacy of the ParaMor Tagger Mimic

Using our finite state tagger, each character, *c*, in each word is scored with the likelihood that ParaMor would treat *c* as the first character in a new morpheme. Consider the segmentation that results from placing morpheme boundaries before each character that is tagged as the start of a new morpheme, stem or affix, with a probability greater than 0.5. This baseline mimic segmentation, although trained to emulate ParaMor's segmentations, will not be fully identical with ParaMor's original segmentation of a set of words. Figure 1 summarizes our tagging accuracy at emulating segmentations for the five languages and six data sets of the Linguistic completion of Morpho Challenge 2009. Tagging accuracy is calculated over all tagged characters, averaging over the held-out test-folds during 10-fold cross-validation. For all the test languages and scenarios, our tagger successfully emulates ParaMor at an accuracy above 93%, with particularly strong accuracy for German, 96.6%, and English 97.6%.

| | English | German | Finnish | Turkish | Arabic -V | Arabic +V |
|---|---|---|---|---|---|---|
| **Linguistic** | 97.6% | 96.6% | 93.5% | 93.6% | 93.3% | 93.7% |

**Figure 1**: The tagging accuracy of our finite-state tagger at mimicking ParaMor's morphological segmentations over the data from the Linguistic competition of Morpho Challenge 2009.

The mimic tagger's departures from the original ParaMor segmentation may either hurt or improve the segmentation quality. On the one hand, when the mimic tagger deviates from the ParaMor segmentation, the mimic may be capturing some real generalization of morphological structure that is hidden in the statistical distribution of ParaMor's original segmentation. On the other hand, a disagreement between the original and the mimic ParaMor segmentations may simply be a failure of the tagger to model the irregularities inherent in natural language morphology.

To evaluate the effectiveness of using a tagger to mimic ParaMor's segmentations, we performed a development evaluation over a Hungarian dataset. We used Hunmorph (Trón et al., 2005), a hand-built Hungarian morphological analyzer, to produce an morphological answer key containing 500,000 unique Hungarian word types from the Hunglish corpus (Varga et al., 2009). Our Hungarian ParaMor tagger mimic actually outperforms ParaMor's original segmentations at $F_1$: Where the original ParaMor attained an $F_1$ of 41.4%, the ParaMor tagger mimic improved $F_1$ to 42.7% with the help of a slightly higher recall.

## 2.3. Optimizing $F_1$

Having retained ParaMor's underlying performance quality by training a natural language tagger to mimic ParaMor's segmentations, we next seek to increase the tagger-mimic's $F_1$ further by leveraging the probabilistic scores that the tagger mimic assigns to each segmentation decision. As discussed in section 1.1, numeric scores for each segmentation decision are the key to trading off precision for recall, reducing the gap between precision and recall, and raising $F_1$.

To optimize $F_1$, we proceed as follows:

1. For each character, *c*, that does not begin a word, record the tagger mimic's probability that *c* begins a morpheme.
2. Sort this list of probabilities smallest to largest and count the number of probability scores that are larger than 0.5, assigning *k* to be this count.
3. For a given positive factor, $\alpha$, consult the list of sorted probabilities to identify the probability score, *S*, above which $\alpha k$ of the probabilities lie.
4. Segment at all characters which receive a probabilistic segmentation score above *S*.

In prose, *k* is the number of word-internal morpheme boundaries that the default ParaMor mimic proposes. To trade off recall against precision adjust the number of morpheme boundaries that the ParaMor mimic proposes. And to increase or decrease the number of morpheme boundaries that the ParaMor mimic proposes by a factor $\alpha$, we move the probability threshold from 0.5 to that value which will permit □□ segmentations.

Figure 2 plots the precision, recall, and $F_1$ of the ParaMor tagger mimic as the number of word-internal morpheme boundaries varies between one half and four times the baseline *k* number of word-internal boundaries. As Figure 2 shows, adjusting $\alpha$ allows for a smooth tradeoff between precision and recall. $F_1$ reaches its maximum value of 47.5% at $\alpha = 4/3$. As expected 4/3 is near the location where recall overtakes precision. The improvement in $F_1$ for the ParaMor tagger mimic of 4.8% is statistically significant at a 95% confidence value if we assume that $F_1$ is normally distributed.

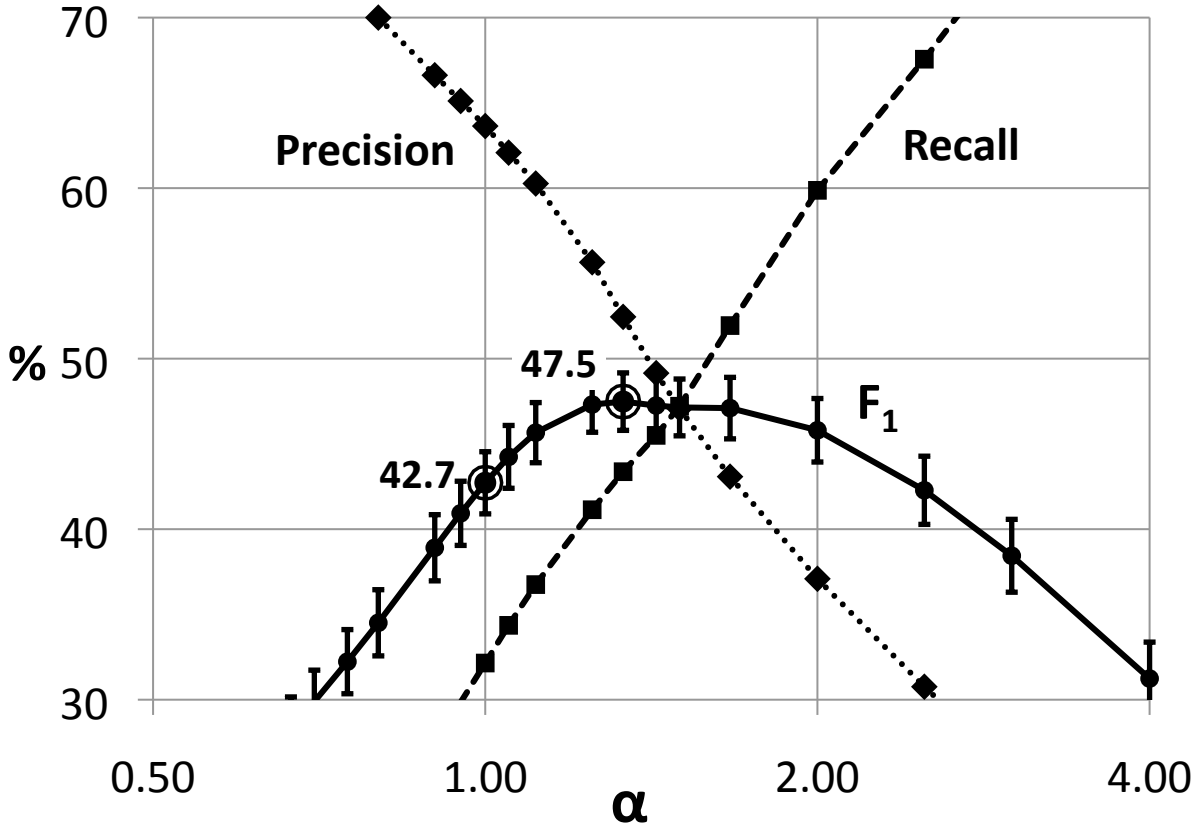


**Figure 2:** Precision, Recall, and $F_1$ of the ParaMor tagger mimic as $\alpha$ moves between 0.5 and 4.0. When $\alpha$ is 1, $F_1$ lies at 42.7%. But by increasing the number of morpheme boundaries that the tagger mimic proposes by a third, to $\alpha = 4/3$, the gap between precision and recall decreases and $F_1$ rises by 4.8% absolute to reach 47.5%. The error bars are 95% confidence intervals on each $F_1$ value, assuming the $F_1$ measurements are normally distributed.

## 2.4. ParaMor in Morpho Challenge 2009

Our ParaMor tagger mimic competed in all the language scenarios of Morpho Challenge 2009. For all languages of the Linguistic, Information Retrieval, and Machine Translation competitions of Morpho Challenge we set $\alpha$ at 4/3, the setting which produced the highest $F_1$ on our Hungarian development set. Figure 3 contains the 2009 linguistic competition's precision, recall, and $F_1$ scores for both the original ParaMor, which competed in Morpho Challenge 2008, and for the ParaMor tagger mimic on the non-Arabic languages. Most likely due to the small size of the Arabic data sets, all versions of ParaMor suffered from extraordinarily low recall in both the vowelized and unvoweled Arabic scenarios. But in the four non-Arabic languages of the Morpho Challenge 2009 Linguistic competition, the gap between precision and recall is smaller for the ParaMor Mimic than it is for the 2008 ParaMor system. And in all languages but English, the reduced precision-recall gap results in a higher $F_1$ score.

The increase in $F_1$ for German, Finnish, and Turkish is more modest than the Hungarian results had led us to hope—about one percentage point in each case. Two reasons likely limited the improvements in $F_1$. First, the performance rose by a smaller amount for the Challenge test languages than they did for our Hungarian devel-

| | English | | | German | | | Finnish | | | Turkish | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| Original | 63.3 | 52.0 | 57.1 | 57.0 | 42.1 | 48.4 | 50.0 | 37.6 | 42.9 | 57.4 | 45.8 | 50.9 |
| Mimic | 53.1 | 59.0 | 55.9 | 50.8 | 47.7 | 49.2 | 47.2 | 40.5 | 43.6 | 49.5 | 54.8 | 52.0 |

**Figure 3:** The precision, recall, and $F_1$ of the original ParaMor, which competed in Morpho Challenge 2008, and the ParaMor tagger mimic in the non-Arabic languages of the Linguistic competition of Morpho Challenge 2009. In all languages but English, the mimic system improves on the original ParaMor's $F_1$ score.

opment set because we were explicitly tuning our α parameter to Hungarian. Second, it may be atypical that the tagger mimic outperformed the baseline ParaMor system on the Hungarian data. Time and resource constraints forced us to train the tagger mimics over subsets of the full Morpho Challenge data, anecdotally lowering tagging mimic accuracy by about a percentage point. To ascertain the quality of the tagger mimics across the range of Morpho Challenge languages, we plan to ask the Morpho Challenge 2009 Committee to evaluate the baseline ParaMor tagger mimic systems with $\alpha$ set to 1.

## 3. Joining ParaMor with Morfessor

In addition to the ParaMor tagger mimic system, we submitted two systems to Morpho Challenge 2009 which join segmentations derived from ParaMor with segmentations obtained from the freely available unsupervised morphology induction system Morfessor (Creutz, 2006), see section 3.1. Our joint ParaMor-Morfessor systems differ substantially from the ParaMor-Morfessor systems that the lead author submitted in the 2007 and 2008 Challenges. In particular, both joint systems submitted to the 2009 Challenge combine the ParaMor and the Morfessor segmentations of each word into a *single* analysis of that word.

### 3.1. Morfessor

In brief, the unsupervised morphology induction system Morfessor Categories-MAP searches for a segmentation of a corpus that maximizes the corpus probability score according to a specific generative probability model. The Morfessor system then further refines the morphological segmentations it proposes by restricting morpheme sequences with a Hidden Markov Model (HMM) which permits only (prefix* stem suffix*)+ sequences.

The Morfessor system serves as a strong baseline system in Morpho Challenge. In the 2008 Challenge, Morfessor placed first in the Arabic Linguistic competition at 34.0% $F_1$, and second in Turkish at 38.5%. Both Morfessor's underlying recursive search strategy and its HMM structure are designed to handle agglutinative morphology where a single word consists of several morphemes in sequence. In general, Morfessor attains a higher precision than recall. In the 2008 Morpho Challenge, Morfessor's lowest precision score in the linguistic competition was 67.2%, for German, while the highest recall score it achieved was 36.8%, also for German. Although, Morfessor's precision scores for all the other languages of the 2008 linguistic competition lie above 70%, Morfessor's more balanced precision and recall scores for German lead to Morfessor's highest $F_1$ score for any language.

### 3.2. Two Methods for System Combination

#### 3.2.1. Union

The first of our two joint submissions to Morpho Challenge 2009 fuses a single morphological segmentation from the disparate segmentations proposed by the ParaMor and Morfessor systems by segmenting each word at *every* location that either ParaMor or Morfessor suggests. Hence, this submission is the union of all segmentation points that are proposed by ParaMor and Morfessor. As an example union segmentation take the English word *polymers'* from the Linguistic competition of this year's Challenge. ParaMor segments *polymers'* as *polym +er +s'*, Morfessor as *polymer +s +'*, while the union analysis is *polym +er +s +'*.

#### 3.2.2. A Joint ParaMor-Morfessor Mimic

The second of our joint ParaMor-Morfessor submissions builds on the idea of tagger mimics that was proposed in section 2. While Morfessor has itself a statistical model that internally scores individual morphological segmentations with probabilities, the final segmentations that Morfessor proposes are not by default annotated with confidences. Hence, we followed the procedure outlined in section 2 to train a natural language tagger to mimic Morfessor's morphological analyses. It is encouraging that our technique for inducing a probabilistic model through a mimic tagger immediately extends from a non-statistical system like ParaMor to the black-box scenario for Morfessor.

With mimic taggers for both ParaMor and Morfessor in hand we then joined, for each character, $c$, in each word, the tag probabilities from the ParaMor mimic with the corresponding probabilities from the Morfessor mimic. We weighted the probability scores from the ParaMor mimic and the Morfessor mimic equally. To obtain the final morphological segmentation of each word, our joint ParaMor-Morfessor mimic followed the methodology described in section 2.3 of optimizing $F_1$ against our Hungarian development set, with one caveat. Because we weighted the probabilities of ParaMor and Morfessor equally, any segmentation point that is strongly suggested by only one of the two systems receives an adjusted probability score just less than 0.5. Hence, we moved the baseline probability threshold from 0.5 to 0.49. With this single adjustment, the $\alpha$ factor that maximized Hungarian $F_1$ was 10/9, an 11% increase in the number of proposed morpheme boundaries.

## 4. The Performance of our Joint ParaMor-Morfessor Systems

Figure 4 summarizes the precision, recall, and $F_1$ performance of three joint ParaMor-Morfessor systems over the datasets for the non-Arabic languages from the Linguistic competition of Morpho Challenge 2009. The first two rows of Figure 4 give performance numbers for the Union and Tagger Mimic systems which we submitted

| | English | | | German | | | Finnish | | | Turkish | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| Union | 55.7 | 62.3 | 58.8 | 52.3 | 60.3 | 56.1 | 47.9 | 51.0 | 49.4 | 47.3 | 60.0 | 52.9 |
| Mimic | 54.8 | 60.2 | 57.4 | 51.1 | 57.8 | 54.2 | 51.8 | 45.4 | 48.4 | 48.1 | 60.4 | 53.5 |
| 2008 | 70.1* | 67.4* | 68.7* | 64.1 | 61.5 | 62.8 | 65.2 | 50.4 | 56.9 | 66.8 | 58.0 | 62.1 |

**Figure 4:** The precision, recall, and $F_1$ of three joint ParaMor-Morfessor morphological segmentation systems over the non-Arabic languages of the Linguistic competition of Morpho Challenge 2009.

*The best result, reported here, was from Morpho Challenge 2007

this year. The third row lists the performance numbers from the joint ParaMor-Morfessor system that was submitted by the lead author to Morpho Challenge 2008.

Although the union and tagger mimic joint systems do outperform at $F_1$ the solo ParaMor mimic (Compare Figure 3), it was disappointing that the simple union system outscored the ParaMor-Morfessor tagger mimic in three of the four relevant language scenarios. Particularly surprising is that the recall of the joint tagger mimic falls below the recall of the union system in every language but Turkish. With an $\alpha$ factor above 1, the joint tagger mimic is proposing all the segmentation points that either the ParaMor mimic or the Morfessor mimic hypothesize—effectively the union of the mimic systems. And yet recall is below the raw union. We tentatively conclude that the cumulative failure of the ParaMor mimic to emulate the original ParaMor segmentations on the one hand, and the Morfessor mimic to emulate Morfessor on the other, drags down the recall (and precision) of the joint mimic.

Figure 4 also highlights the relative success of the 2008 joint ParaMor-Morfessor system. In particular, the precision scores of the 2008 joint system are significantly above the precision scores of the joint systems we submitted to the 2009 Challenge. The 2008 (and 2007) joint system did *not* form a single unified segmentation for each word, but instead simply proposed the ParaMor analysis of each word alongside the Morfessor analysis—as if each word were ambiguous between a ParaMor and a Morfessor analysis. The evaluation procedure of Morpho Challenge performs a non-trivial averaging procedure over alternative segmentations of a word. We believe it is a shortcoming of the Morpho Challenge evaluation procedure that allows inflated precision scores when disparate systems' outputs are proposed as 'alternative' analyses.

## 5. The Next Steps

Having now built a sound methodology for assigning confidence scores to ParaMor's rule-based morphological segmentations, we will return our attention to improving ParaMor's modeling of morphological structure. Currently, the ParaMor algorithm cannot hypothesize that two distinct surface strings are simply allomorphs of the same underlying morpheme. Hence, ParaMor is unable, for example, to conflate the surface forms of Turkish and Finnish morphemes which differ only in their harmonized vowel. ParaMor is similarly limited in the types of morphological operations that ParaMor can analyze. Currently ParaMor only searches for suffixes. Broadening ParaMor to analyze prefixation, infixation, and perhaps the templatic morphology of Semitic languages will surely boost ParaMor's recall scores.

## Acknowledgements

This research was supported in part by NSF Grant #IIS-0811745 and DOD/NGIA grant #HM1582-08-1-0038. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF or DOD.

## References

Bernhard, Delphine. 2008. Simple Morpheme Labeling in Unsupervised Morpheme Analysis. *Lecture Notes in Computer Science: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Revised Selected Papers*, Budapest, Hungary, Springer, 5152/2008, 873-880.

Collins, Michael. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP).

Creutz, Mathias. 2006. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition*. Ph.D. Thesis, Computer and Information Science, Report D13, Helsinki, University of Technology, Espoo, Finland.

Goldsmith, John. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27.2, 153-198.

Harris, Zellig. 1955. From Phoneme to Morpheme. *Language*, 31.2, 190-222, Reprinted in Harris (1970).

Harris, Zellig. 1970. *Papers in Structural and Transformational Linguists*. Ed. D. Reidel, Dordrecht.

Hollingshead, Kristy, Seeger Fisher, and Brian Roark. 2005. Comparing and combining finite-state and context-free parsers. *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, BC, Canada.

Kurimo, Mikko and Ville Turunen. 2008. Unsupervised Morpheme Analysis Evaluation by IR Experiments – Morpho Challenge 2008. *Working Notes for the CLEF 2008 Workshop*.

Monson, Christian. 2009. *ParaMor: From Paradigm Structure to Natural Language Morphology Induction*. Ph.D. Thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

Oflazer, Kemal and İlknur Durgar El-Kahlout. 2007. Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation. *Statistical Machine Translation Workshop at ACL*.

Poon, Hoifung, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised Morphological Segmentation with Log-Linear Models. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*.

Roark, Brian and Kristy Hollingshead. 2009. Linear Complexity Context-Free Parsing Pipelines via Chart Constraints. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*.

Snyder, Benjamin and Regina Barzilay. 2008. Unsupervised Multilingual Learning for Morphological Segmentation. *Proceedings of ACL-08: HLT*.

Tjong Kim Sang, Eric F. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of CoNLL-2002*.

Tjong Kim Sang, Eric F. and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 Shared Task: Chunking. *Proceedings of the 4th Conference on Computational Natural Language Learning (CoNLL)*.

Trón, Viktor, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. 2005. Hunmorph: Open Source Word Analysis. *Proceedings of the ACL Workshop on Software*.

Varga, Dániel, Péter Halácsy, András Kornai, László Németh, Viktor Trón, Tamás Váradi, Bálint Sass, Gergő Bottyán, Enikő Héja, Ágnes Gyarmati, Ágnes Mészáros and Dávid Labundy. 2009. *Hunglish corpus*. <http://mokk.bme.hu/resources/hunglishcorpus>. Accessed on August 18, 2009.

Xue, Nianwen. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8.1, 29-47.