# Multilingual Noise-Robust Supervised Morphological Analysis using the WordFrame Model

**Richard Wicentowski**
Swarthmore College
Swarthmore, Pennsylvania, USA 19081
richardw@cs.swarthmore.edu

## Abstract

This paper presents the WordFrame model, a noise-robust supervised algorithm capable of inducing morphological analyses for languages which exhibit prefixation, suffixation, and internal vowel shifts. In combination with a näive approach to suffix-based morphology, this algorithm is shown to be remarkably effective across a broad range of languages, including those exhibiting infixation and partial reduplication. Results are presented for over 30 languages with a median accuracy of 97.5% on test sets including both regular and irregular verbal inflections. Because the proposed method trains extremely well under conditions of high noise, it is an ideal candidate for use in co-training with unsupervised algorithms.

## 1 Introduction

This paper presents the WordFrame model, a novel algorithm capable of inducing morphological analyses for a large number of the world's languages. The WordFrame model learns a set of string transductions from inflection-root pairs and uses these to transform unseen inflections into their corresponding root forms. These string transductions directly model prefixation, suffixation, associated point-of-affixation changes and stem-internal vowel shifts. Though not explicitly modeled, patterns extracted from large amounts of noisy training data can be highly effective at aligning inflections with roots in languages which exhibit vowel harmony, agglutination, and partial word reduplication.

The WordFrame model contains no language-specific parameters. While we make no claims that the model works equally well for all languages, its ability to analyze inflections in 32 diverse languages with a median accuracy of 97.5% attests to its flexibility in learning a wide range of morphological phenomena.

The effectiveness of the model when trained from noisy data makes it well-suited for co-training with low-accuracy unsupervised algorithms.

## 2 Previous Work

The development of the WordFrame model was motivated by work originally presented in Yarowsky and Wicentowski (2000). In that work, a suite of unsupervised learning algorithms and a supervised morphological learner are co-trained to achieve high accuracies for English and Spanish verb inflections. The supervised learner employed a naïve approach to morphology, only capable of learning word-final stem changes between inflections and roots. This "end-of-string model" of morphology was used again in Yarowsky et al. (2001) where it was applied to English, French and Czech. (More complete details of the end-of-string model are presented in Section 3.3.1.)

Though simplistic, this end-of-string model is robust to noise, especially important in co-training with low-accuracy unsupervised learners. However, the end-of-string model relied heavily upon externally provided, noise-free lists of affixes in order to correctly align inflections to roots. The WordFrame model allows, but does not require, such affix lists, thereby eliminating direct human supervision.

Much previous work has been done in automatically acquiring such affix lists, most recently the generative models built by Snover and Brent (2001) which are able to identify suffixes in English and Polish. Schone and Jurafsky (2001) use latent semantic analysis to find prefixes, suffixes and circumfixes in German, Dutch and English. Baroni (2003) treats morphology as a data compression problem to find English prefixes.

Goldsmith (2001) uses minimum description length to successfully find paradigmatic classes of suffixes in a number of European languages, including Dutch and Russian, though the approach has been less successful in handling prefixation.

The Boas project (Oflazer et al., 2001), (Hakkani-Tür et al., 2000), and (Oflazer and Nirenburg, 1999) has produced excellent results bootstrapping a morphological analyzer, but rely on direct human supervision to produce two-level rules (Koskenniemi,

1983) which are then compiled into a finite state machine.

## 3 The WordFrame Algorithm

### 3.1 Motivation

The supervised morphological learner presented in Yarowsky and Wicentowski (2000) modeled lemmatization as a word-final stem change plus a suffix taken from a (possibly empty) list of potential suffixes. Though effective for suffixation, this end-of-string (EOS) based model can not model other morphological phenomena, such as prefixation.

By including a pre-specified list of prefixes, we can extend the EOS model to handle simple prefixation: For each inflection, an analysis is performed on the original string, plus on each substring resulting from removing exactly one matching prefix taken from the list of prefixes. While effective for some simple prefixal morphologies, this extension cannot model word-initial stem changes at the point of prefixation. In contrast, the WordFrame (WF) algorithm can isolate a potential prefix and model any potential point-of-prefixation stem changes directly, *without* pre-specified lists of prefixes.

The EOS model also fails to capture word-internal vowel changes found in many languages. The WF model directly models stem-internal vowel changes in order to to learn higher-quality, less sparse, transformation rules.

| training pair | EOS analysis | WF analysis |
|---|---|---|
| acuerto→acortar | uerto→ortar | ue→o |
| apruebo→aprobar | uebo→obar | ue→o |
| muestro→mostrar | uestro→ostrar | ue→o |

Table 1: The above Spanish examples are misanalyzed by the EOS algorithm, which results in learning rules with low productivity. The WF algorithm is able to identify the productive *ue→o* stem-internal vowel change.

### 3.2 Required and Optional Resources

a. Training data of the form <*inflection,root*> is required for the WordFrame algorithm. Ideally, this data should be high-quality and noise-free, but the algorithm is robust to noise, which allows one to use lower-quality pairs extracted from unsupervised techniques.

b. Pre-specified lists of prefixes and suffixes can be incorporated, but are not required.

c. Precision can be improved (at the expense of coverage) by providing a list of potential roots extracted from a dictionary or large corpus.

d. In order to allow for word-internal vowel changes, the WordFrame model requires a list of the vowels of the language.

### 3.3 Formal Presentation

The WordFrame model is constructed explicitly as an extension to the end-of-string model proposed by Yarowsky and Wicentowski (2000); as such, we first give a brief presentation of the model, then introduce the WordFrame model.

In the discussion below, if affix lists are not explicitly provided, they are assumed to contain the single element $\epsilon$ (the empty string).

#### 3.3.1 The end-of-string model

The end-of-string model makes use of two optional externally provided sets: a set of acceptable suffixes, $\Psi'_s$, and a set of "canonical root endings", $\Psi_s$. The inclusion of a list of canonical root endings is motivated by languages where verb roots can end in only a limited number of ways (e.g. *-er*, *-ir* and *-re* in French).

From inflection-root training pairs, a deterministic analysis is made by removing the longest matching suffix ($\psi'_s \in \Psi'_s$) from the inflection, removing the longest matching canonical ending ($\psi_s \in \Psi_s$) from the root, and removing the longest common initial substring ($\gamma$) from both words. The remaining strings represent the word-final stem change ($\delta'_s \rightarrow \delta_s$) necessary to transform the inflection ($\gamma\delta'_s\psi'_s$) into the root ($\gamma\delta_s\psi_s$). The word-final stem changes are stored in a hierarchically-smoothed suffix trie representing $P(\delta'_s \rightarrow \delta_s|\gamma\delta'_s)$.

A simple extension allows the EOS model to handle purely concatenative prefixation: the analysis begins by removing the longest matching prefix taken from a given set of prefixes ($\psi'_p \in \Psi'_p$), then continuing as above. This changes the inflection to $\psi'_p\gamma\delta'_s\psi'_s$, and leaves the root as $\gamma\delta_s\psi_s$. (See Table 2 for an overview of this notation.)

Given a previously unseen inflection, one finds the root that maximizes $P(\gamma\delta_s\psi_s|\psi'_p\gamma\delta'_s\psi'_s)$. By making strong independence assumptions and some approximations, and assuming that all prefixes and suffixes are equally likely, this is equivalent to:[1]

$$P(\gamma\delta_s\psi_s|\psi'_p\gamma\delta'_s\psi'_s) = \max_{\psi'_p,\gamma\delta'_s,\psi'_s} P(\delta'_s \rightarrow \delta_s|\gamma\delta'_s)$$

Note we are using a slightly different, but equivalent, notation to that used in Yarowsky and Wicentowski (2000). Simply, we use $\psi'_s$ rather than $\sigma$, and we use $\delta'_s \rightarrow \delta_s$ rather than $\alpha \rightarrow \beta$. This change was made in order to make the formalization of the WF model more clear.

[1] Full details available in (Wicentowski, 2002).

| | | prefix | point-of-prefixation change | secondary common substring | vowel change | primary common substring | point-of-suffixation change | suffix/ ending |
|---|---|---|---|---|---|---|---|---|
| Extended EOS | inflection | $\psi'_p$ | | | | $\gamma_s$ | $\delta'_s$ | $\psi'_s$ |
| | root | | | | | | $\delta_s$ | $\psi_s$ |
| WordFrame | inflection | $\psi'_p$ | $\delta'_p$ | $\gamma_p$ | $\delta'_v$ | $\gamma_s$ | $\delta'_s$ | $\psi'_s$ |
| | root | | $\delta_p$ | | $\delta_v$ | | $\delta_s$ | $\psi_s$ |

Table 2: Overview of the analyzed components of the inflection and root using the end-of-string (EOS) model extended to allow for simple prefixation, and the WordFrame model. If lists of prefixes, suffixes and endings are not specified, the prefix, suffix and ending are set to $\epsilon$.

### 3.3.2 The WordFrame model

The WordFrame model fills two major gaps in the EOS model: the inability to model prefixation without a list of provided prefixes, and the inability to model stem-internal vowel shifts.

While not required, the WordFrame model does allow for the inclusion of lists of prefixes, and when provided, can automatically discover the point-of-prefixation stem change, $\delta'_p \rightarrow \delta_p$. When a list of prefixes is not provided, the word-initial stem change will model both the prefix and stem change.

Formally, this requires the inclusion of the point-of-prefixation stem change into the notation used in the EOS model. When presented with an inflection-root pair, the longest common substring in the inflection and root, $\gamma$, is assumed to be the stem. The string preceding the stem is the prefix and point-of-prefixation stem change, $\psi'_p\delta'_p$; the string following the stem is the suffix and point-of-suffixation stem change, $\psi'_s\delta'_s$. Combining these parts, the inflection can be represented as $\psi'_p\delta'_p\gamma\delta'_s\psi'_s$, and the root as $\delta_p\gamma\delta_s\psi_s$.

In addition, the WordFrame model allows for a single word-internal vowel change within the stem. To accommodate this, the longest common substring of the inflection and root, $\gamma$, is allowed to be split in a single location to allow the vowel change $\delta'_v \rightarrow \delta_v$ where $\delta'_v$ and $\delta_v$ are taken from a predetermined list of vowels for the language.[2] The portions of the stem located before and after the vowel change are now $\gamma_p$ and $\gamma_s$, respectively.

Both $\delta'_v$ and $\delta_v$ may contain more than vowel, thereby allowing vowel changes such as *ee*→*e*. However, as presented here, the WF model does not allow for the insertion of vowels into the stem where there were no vowels previously; more formally, both $\delta'_v$ and $\delta_v$ must contain at least one vowel, or they both must be $\epsilon$. Though this restriction can be removed, initial results (not presented here) indicated a significant drop in accuracy when entire vowels clusters could be removed or inserted. In addition, the vowel change must be *internal* to the stem, and cannot be located at the boundary of the stem; formally, unless both $\delta'_v$ and $\delta_v$ are $\epsilon$, both portions of the split stem ($\gamma_p$ and $\gamma_s$) must contain at least one letter. This prevents confusion between "stem-internal" vowel changes and stem-changes at the point of affixation.

As with the EOS model, a deterministic analysis is made from inflection-root training pairs. If provided, the longest matching prefix and suffix are removed from the inflection, and the longest matching canonical ending is removed from the root.[3] The remaining string must then be analyzed to find the longest common substring with at most one vowel change, which we call the WordFrame.

The WordFrame ($\gamma_p\delta'_v\gamma_s$, $\gamma_p\delta_v\gamma_s$) is defined to be the longest common substring with at most one internal vowel cluster ($V^* \rightarrow V^*$) transformation. Should there be multiple "longest" substrings, the substring closest to the start of the inflection is chosen.[4] In practice, there is rarely more than one such "longest" substring.

The remaining strings at the start and end of the common substring form the point-of-prefixation and point-of-suffixation stem changes.

The final representation of the inflection-root pair in the WF model is shown in Table 2.

Given an unseen inflection, one finds the root that maximizes $P(\delta_p\gamma_p\delta_v\gamma_s\delta_s\psi_s|\psi'_p\gamma_s\delta'_s\psi'_s)$. If we make the simplifying assumption that all prefixes, suffixes and endings are equally likely and remove

[2] If one wishes to model arbitrary internal changes, this "vowel" list could be made to include every letter in the alphabet; results are not presented for this configuration.

[3] A canonical *prefix* is not included in the model because we knew of no language in which this occurred; introducing it to the model would be straight-forward.

[4] This places a bias in favor of end-of-string changes and is motivated by the number of languages which are suffixal and the relative few that are not; this could be adjusted for prefixal languages.

END-OF-STRING

| | | $\psi'_p$ | $\delta'_p \rightarrow \delta_p$ | $\gamma_p$ | $\delta'_v \rightarrow \delta_v$ | $\gamma_s$ | $\delta'_s \rightarrow \delta_s$ | $\psi'_s \rightarrow \psi_s$ |
|---|---|---|---|---|---|---|---|---|
| English | kept→keep<br>sang→sing | | | | | ke<br>s | p→ep<br>ang→ing | t→ $\epsilon$ |
| Spanish | acuerto→acortar<br>muestro→mostrar | | | | | ac<br>m | uert→ort<br>uestr→ostr | o→ar<br>o→ar |
| German | gestunken→stinken<br>gefielt→gefallen | | | | | gef | gestunk→stink<br>iel→all | en→en<br>t→en |

WORDFRAME

| | | $\psi'_p$ | $\delta'_p \rightarrow \delta_p$ | $\gamma_p$ | $\delta'_v \rightarrow \delta_v$ | $\gamma_s$ | $\delta'_s \rightarrow \delta_s$ | $\psi'_s \rightarrow \psi_s$ |
|---|---|---|---|---|---|---|---|---|
| English | kept→keep<br>sang→sing | | | k<br>s | e→ee<br>a→i | p<br>ng | | t→ $\epsilon$ |
| Spanish | acuerto→acortar<br>muestro→mostrar | | | ac<br>m | ue→o<br>ue→o | rt<br>str | | o→ar<br>o→ar |
| German | gestunken→stinken<br>gefielt→gefallen | | ge→ $\epsilon$ | st<br>gef | u→i<br>ie→a | nk<br>l | $\epsilon$ →l | en→en<br>t→en |

Table 3: End-of-string and WordFrame analysis of training data assuming no provided lists of prefixes. The EOS analysis yields non-productive rules such as *gestunk→stink*. The WF analysis captures the productive Spanish vowel change *ue* → *o*, the German prefix *ge*, and English vowel changes *e→ee* and *a→i*.

the longest possible affixes deterministically, this is equivalent to:

$$P(\delta_p\gamma_p\delta_v\gamma_s\delta_s|\delta'_p\gamma_p\delta'_v\gamma_s\delta'_s)$$
$$= P(\delta'_v \rightarrow \delta_v, \delta'_p \rightarrow \delta_p, \delta'_s \rightarrow \delta_s|\delta'_p\gamma_p\delta'_v\gamma_s\delta'_s)$$

This can be expanded using the chain rule. As before, the point-of-suffixation probabilities are implicitly conditioned on the applicability of the change to $\delta'_p\gamma_p\delta'_v\gamma_s\delta'_s$, and are taken from a suffix trie created during training. The point-of-prefixation probabilities are implicitly conditioned on the applicability of the change to $\delta'_p\gamma_p\delta'_v\gamma_s$, i.e. once $\delta'_s$ has been removed, and are taken from an analogous prefix trie. The vowel change probability is conditioned on the applicability of the change to $\gamma_p\delta'_v\gamma_s$. In the current implementation, this is approximated using the conditional probability of the vowel change $P(\delta_v|\delta'_v)$ without regard to the local context. This is a major weakness in the current system and one that will be addressed in future work.

The WordFrame model's ability to capture stem-internal vowel changes allows for proper analysis of the Spanish examples from Table 1, and also allows for the analysis of prefixes without the use of a pre-specified list of prefixes, as shown in Table 3.

## 4 Experimental Evaluation

All of the experimental results presented here were done using 10-fold cross-validation on the training data. The majority of the training data used here was obtained from web sources, although some has been hand-entered or scanned from printed materials then hand-corrected. All of the data used were inflected verbs; there was no derivational morphology in this evaluation.[5] Unless otherwise specified, all results are system accuracies at 100% coverage – Section 5.3 addresses precision at lower coverages.

| | |
|---|---|
| point-of-prefixation change | $ge \rightarrow \epsilon$ |
| point-of-suffixation change | $\epsilon \rightarrow l$ |
| vowel changes | $u \rightarrow i$<br>$ie \rightarrow a$ |

Table 4: String transductions derived from the German examples listed in Table 3.

Space limits the number of results that can be presented here since most of the evaluations have been carried out in each of the 32 languages. Therefore, in comparing the models, results will only be shown for only a representative subset of the languages. When appropriate, a median or average for all languages will also be given. Table 10 presents the final results for all languages.

[5] Examples of derivational morphology, as well as nominal and adjectival inflectional morphology, are excluded from this presentation due to the lack of available training data for more than a small number of well-studied languages.

### 4.1 End-of-string vs. WordFrame

The most striking difference in performance between the EOS model and WordFrame model comes from the evaluation of languages with prefixal morphologies. The EOS model cannot handle prefixation without pre-specified lists of prefixes, so when these are omitted, the WF model drastically outperforms the EOS model (Table 5).

| | w/o Affixes | | w/ Affixes | |
|---|---|---|---|---|
| Language | EOS | WF | EOS | WF |
| Tagalog | 1.6% | **89.9%** | 92.0% | **96.0%** |
| Swahili | 2.9% | **96.8%** | 93.8% | **96.9%** |
| Irish | 45.7% | **89.5%** | - | - |
| Spanish | **94.7%** | 90.2% | **96.5%** | 95.2% |
| Portuguese | 97.4% | **97.9%** | 97.3% | **97.5%** |

Table 5: Accuracy of the EOS model vs the WF model without and with pre-specified lists of affixes (if available for that language).

Table 5 also shows that the simple EOS model can sometimes significantly outperform the WF model (e.g. in Spanish). Making things more difficult, predicting which model will be more successful for a particular language and set of training data may not be possible, as illustrated by the fact that EOS model performed better for Spanish, but the closely-related Portuguese was better handled by the WF model. Additionally, as illustrated by the Portuguese example, it is not always beneficial to include lists of affixes, making selection of the model problematic.

Lists of prefixes and suffixes were not available for all languages.[6] However, for the 25 languages where such lists were available, the WordFrame model performed equally or better on only 17 (68%). Evidence suggests that this occurs when the affix lists have missing prefixes or suffixes. Since these lists were extracted from printed grammars, such gaps were unavoidable.

Regardless of whether or not affix lists were included, the WordFrame model only outperformed the EOS model for just over half the languages. An examination of the output of the WF model suggests that the relative parity in performance of the two models is due to the poor estimation of the vowel change probability which is approximated without regard to the contextual clues.

[6]The affix lists used in this evaluation were hand-entered from grammar references and were only available for 25 of the 32 languages evaluated here; therefore, the results presented in this section omit these seven languages: Norwegian, Hindi, Sanskrit, Tamil, Russian, Irish, and Welsh.

## 5 WordFrame + EOS

One of our goals in designing the WordFrame model was to reduce or eliminate the dependence on externally supplied affix lists. However, the results presented in Section 4.1 indicate that the WF model outperforms the EOS model for just over half ($17/32$) of the evaluated languages, even when affix lists are included.

Predicting which model worked better for a particular language proved difficult, so we created a new analyzer by combining our WordFrame model with the end-of-string model. For each inflection, the root which received the highest probability using an equally-weighted linear combination was selected as the final analysis.

This new combination analyzer outperformed both stand-alone models for 21 of the 25 languages with significant overall accuracy improvements as shown in Table 6(a).

(a)

| w/o Affixes | EOS | WF | Combined |
|---|---|---|---|
| Average | 79.2% | 91.0% | **93.0%** |
| Median | 93.6% | 95.9% | **97.4%** |

(b)

| w/ Affixes | EOS | WF | Combined |
|---|---|---|---|
| Average | 95.1% | 95.0% | **96.8%** |
| Median | 96.7% | 96.7% | **97.6%** |

Table 6: Average and median accuracy of the individual models vs. the combined model (a) with and (b) without affix lists.

When affix lists *are* available, combining the WordFrame model and the end-of-string model yielded very similar results: the combined model outperformed either model on its own for 23 of the 25 languages. Of the two remaining languages, the stand-alone WF model outperformed the combined model by just one example out of 5197 in Danish, and just 4 examples out of 9497 in Tagalog. As before, the combined model showed significant accuracy increases over either stand-alone model, as shown in Table 6(b).

Finally, we build the WordFrame+EOS classifier, by combining all four individual classifiers (EOS with and without affix lists, and WF with and without affix lists) using a simple equally-weighted linear combination. This is motivated from our initial observation that using affix lists does not always improve overall accuracy. Cumulative results are shown below in Table 7, and results for each individual language is shown in Table 10.

| WF + EOS | w/o Affixes | w/ Affixes | Combined |
|---|---|---|---|
| Average | 93.0% | 96.8% | **97.2%** |
| Median | 97.4% | 97.6% | **97.9%** |

Table 7: Accuracy of the combined models, plus a combination of the combined models in the 25 languages for which affix lists were available.

### 5.1 Robustness to Noise

The WordFrame model was designed as an alternative to the end-of-string model. In Yarowsky and Wicentowski (2000), the end-of-string model is trained from inflection-root pairs acquired through unsupervised methods. None of those previously presented unsupervised models yielded high accuracies on their own, so it was important that the end-of-string model was robust enough to learn string transduction rules even in the presence of large amounts of noise.

In order for the WF+EOS model to be an adequate replacement for the end-of-string model, it must also be robust to noise. To test this, we first ran the WF+EOS model as before on all of the data using 10-fold cross-validation. Then, we introduced noise by randomly assigning a certain percentage of the inflections to the roots of other inflections. For example, the correct pair *menaced-menace* became the incorrect pair *menaced-move*. The results of introducing this noise are presented in Table 9 and Figure 1.

| Noise | 0% | 10% | 25% | 50% | 75% |
|---|---|---|---|---|---|
| English | 99.1% | 98.6% | 98.6% | 98.4% | 97.6% |
| French | 99.6% | 99.5% | 99.5% | 99.3% | 98.9% |
| Estonian | 96.8% | 94.7% | 94.3% | 92.0% | 87.0% |
| Turkish | 99.5% | 98.5% | 98.2% | 97.1% | 91.4% |

Table 9: The combined WordFrame and EOS model maintains high accuracy in the presence noise. Above, up to 75% of the inflections in the training data have been assigned incorrect roots.

As one might expect, the effect of introducing noise is particularly pronounced for highly inflected languages such as Estonian, as well as with the vowel-harmony morphology found in Turkish[7]. However, languages with minimal inflection (English) or a fairly regular inflection space (French) show much less pronounced drops in accuracy as noise increases.

[7] All of the data is inflectional verb morphology, making the Turkish task substantially easier than most other attempts at modeling Turkish morphology.

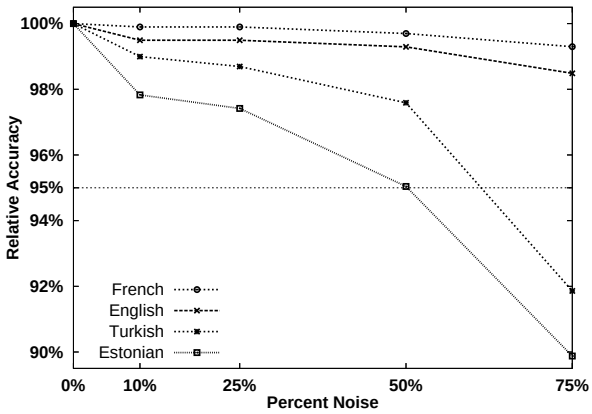


Figure 1: The WF+EOS algorithm's robustness to noise yields only a 5% reduction in performance even when 50% of the training samples are replaced with noise.

It is important to point out that the incorrect pairs were not added *in addition* to the correct pairs; rather, they replaced the correct pairs. For example, the Estonian training data was comprised of 5932 inflection-root pairs. When testing at 50% noise, there were only 2966 correct training pairs, and 2966 incorrect pairs. This means that real size of the training data was also reduced, further lowering accuracy, and making the model's effective robustness to noise more impressive.

### 5.2 Regular vs. Irregular Inflections

For 13 of the languages evaluated, the inflections were classified as either *regular*, *irregular*, or *semi-regular*. As an example, the English pair *jumped-jump* was classified as regular, the pair *hopped-hop* was semi-regular (because of the doubling of the final-*p*), and the pair *threw-throw* was labeled irregular.[8]

Table 8 shows the accuracy of the WF+EOS model in each of the three categories, as well as for all data in total.[9] As expected, the WF+EOS model performs very well on regular inflections and reasonably well on the semi-regular inflections for most languages.

The performance on the irregular verbs, though clearly not as good as on the regular or semi-regular verbs, was surprisingly good, most notably in French, and to a lesser extent, Spanish and Ital-

[8] These classifications were assigned by the provider of our training pairs, not by us.

[9] The small discrepancy between the data in Table 8 and Table 10 is due to the fact that some of the inflection-root pairs were not labeled. The "All" column of Table 8 reflects only labeled inflections.

| | All | | Regular | | Semi | | Irregular | |
|---|---|---|---|---|---|---|---|---|
| Language | accuracy | types | accuracy | types | accuracy | types | accuracy | types |
| Spanish | 97.28% | 58589 | 97.60% | 52709 | 95.38% | 1665 | 93.40% | 3861 |
| Catalan | 90.65% | 4066 | 96.31% | 2898 | 84.35% | 230 | 74.73% | 938 |
| Occitan | 93.39% | 7583 | 98.46% | 6096 | 97.55% | 654 | 52.58% | 795 |
| French | 99.58% | 63644 | 99.79% | 57255 | 99.95% | 2221 | 97.00% | 3866 |
| Italian | 98.43% | 62920 | 98.75% | 54643 | 99.58% | 3335 | 93.64% | 4496 |
| Romanian | 97.84% | 24000 | 98.95% | 21237 | 94.78% | 920 | 85.36% | 1660 |
| English | 98.95% | 3703 | 99.45% | 3073 | 99.50% | 597 | 40.62% | 32 |
| Danish | 97.87% | 4185 | 98.59% | 3760 | 95.00% | 220 | 87.80% | 205 |
| Norwegian | 95.85% | 1954 | 97.57% | 1731 | 90.62% | 96 | 76.38% | 127 |
| Icelandic | 92.58% | 3692 | 97.78% | 2884 | 97.79% | 226 | 64.78% | 582 |
| Hindi | 84.77% | 256 | 98.58% | 212 | 33.33% | 9 | 14.29% | 35 |
| Turkish | 99.46% | 29131 | 99.95% | 26134 | 95.66% | 2811 | 88.71% | 186 |
| Welsh | 88.55% | 45812 | 89.27% | 44060 | 86.69% | 1180 | 32.84% | 536 |

Table 8: Accuracy of WF+EOS on different types of inflections

ian. This is due in large part because our test set included many irregular verbs which shared the same irregularity. For example, in French, the inflection-root pair *prit-prendre* is irregular; however, the pairs *apprit-apprendre* and *comprit-comprendre* both follow the same irregular rule. The inclusion of just one of these three pairs in the training data will allow the WF+EOS model to correctly find the root form of the other two. Our French test set included many examples of this, including roots that ended *-tenir*, *-venir*, *-mettre*, and *-duire*.

For most languages however, the performance on the irregular set was not that good. We propose no new solutions to handling irregular verb forms, but suggest using non-string-based techniques, such as those presented in (Yarowsky and Wicentowski, 2000), (Baroni et al., 2002) and (Wicentowski, 2002).

### 5.3 Accuracy, Precision and Coverage

All of the previous results assumed that each inflection *must* be aligned to exactly one root, though one can improve precision by relaxing this constraint. The WF+EOS model transforms an inflection into a new string which we can compare against a dictionary, wordlist, or large corpus. In determining the final inflection-root alignment, we can down-weight, or even throw away, all proposed roots which are are not found in such a wordlist. While this will adversely affect coverage, precision may be more important in early iterations of co-training. Given a sufficiently large wordlist, such a weighting scheme *cannot* discard correct analyses. In addition, a large majority of the incorrectly analyzed inflections are proposed roots which are not actually words. By excluding all proposed roots which were not found in a broad coverage wordlist (available for 19 languages), median coverage fell to 97.4%, but median precision increased from 97.5% to 99.1%.

## 6 Conclusions

We have presented the WordFrame model, a noise-robust supervised morphological analyzer which is highly successful across a broad range of languages. We have shown our model effective at learning morphologies which exhibit prefixation, suffixation, and stem-internal vowel changes. In addition, the WordFrame model was successful in handling the agglutination, infixation and partial reduplication found in languages such as Tagalog without explicitly modeling these phenomena. Most importantly, the WordFrame model is robust to large amounts of noise, making it an ideal candidate for use in co-training with lower-accuracy unsupervised algorithms.

## References


M. Baroni, J. Matiasek, and T. Harald. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the Workshop on Morphological and Phonological Learning*, pages 48–57.

M. Baroni. 2003. Distribution-driven morpheme discovery: A computational/experimental study. *Yearbook of Morphology*, pages 213–248.

J. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

| Language | Accuracy | Wordlist Entries | Training Data Roots | Infls |
|---|---|---|---|---|
| Spanish | 97.3% | 32895 | 1190 | 57224 |
| Portuguese | 97.9% | 30145 | 584 | 22135 |
| Catalan | 90.7% | - | 103 | 4058 |
| Occitan | 93.4% | - | 180 | 7559 |
| French | 99.6% | 27548 | 1829 | 63559 |
| Italian | 98.5% | 27221 | 1582 | 62658 |
| Romanian | 97.9% | 25228 | 1070 | 24877 |
| Latin | 91.4% | - | 279 | 26818 |
| English | 99.1% | 264075 | 1218 | 4915 |
| Danish | 97.9% | 51351 | 1062 | 5197 |
| Norwegian | 95.9% | - | 547 | 2489 |
| Swedish | 98.5% | 46009 | 4035 | 13871 |
| Icelandic | 92.6% | - | 314 | 3987 |
| Hindi | 84.8% | - | 15 | 255 |
| Sanskrit | 89.5% | - | 867 | 1968 |
| Tamil | 91.0% | - | 24 | 602 |
| Estonian | 96.9% | 344 | 147 | 5932 |
| Finnish | 97.5% | - | 1434 | 79734 |
| Turkish | 99.5% | 25497 | 87 | 29130 |
| Uzbek | 99.5% | - | 434 | 27296 |
| Basque | 96.1% | 33020 | 1185 | 5842 |
| Czech | 98.7% | 29066 | 5715 | 23786 |
| Polish | 97.6% | 42005 | 601 | 23725 |
| Russian | 90.8% | 42740 | 191 | 3068 |
| Greek | 100% | 35245 | 9 | 201 |
| German | 98.0% | 45779 | 1213 | 14120 |
| Dutch | 98.4% | 41962 | 1016 | 5768 |
| Irish | 95.5% | - | 54 | 1376 |
| Welsh | 88.6% | - | 1053 | 44295 |
| Tagalog | 97.5% | - | 212 | 9479 |
| Swahili | 97.0% | 24985 | 818 | 27773 |
| Klingon | 100% | 2114 | 699 | 5135 |

Table 10: For each language, the accuracy of the WordFrame model combined with the end-of-string model, the number of wordlist entries available (Section 5.3), and the total training size used for cross-validation.


D. Hakkani-Tür, K. Oflazer, and G. Tür. 2000. Statistical morphological disambiguation for agglutinative languages. In *18th International Conference on Computational Linguistics*.

K. Koskenniemi. 1983. *Two-level morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, Department of Linguistics, University of Helsinki, Finland.

K. Oflazer and S. Nirenburg. 1999. Practical bootstrapping of morphological analyzers. In *Conference on Natural Language Learning*.

K. Oflazer, S. Nirenberg, and M. McShane. 2001. Bootstrapping morphological analyzers by combining human elicitation and maching learning. *Computational Linguistics*, 27(1):59–84.

P. Schone and D. Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *Proceedings of the North American Chapter of the Association of Computational Linguistics*.

M. Snover and M. R. Brent. 2001. A bayesian model for morpheme and paradigm identification. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, volume 39, pages 482–490.

R. Wicentowski. 2002. *Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework*. Ph.D. thesis, The Johns Hopkins University.

D. Yarowsky and R. Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, pages 207–216.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the Human Language Technology Conference*, pages 161–168.