# Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation


**Kemal Oflazer**[†,‡]
[†]Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
oflazer@sabanciuniv.edu

**İlknur Durgar El-Kahlout**[‡]
[‡] Faculty of Engineering and Natural Sciences
Sabancı University
Istanbul, Tuzla, 34956, Turkey
ilknurdurgar@su.sabanciuniv.edu



## Abstract

We investigate different representational granularities for sub-lexical representation in statistical machine translation work from English to Turkish. We find that (i) representing both Turkish and English at the morpheme-level but with some selective morpheme-grouping on the Turkish side of the training data, (ii) augmenting the training data with "sentences" comprising only the content words of the original training data to bias root word alignment, (iii) reranking the n-best morpheme-sequence outputs of the decoder with a word-based language model, and (iv) using model iteration all provide a non-trivial improvement over a fully word-based baseline. Despite our very limited training data, we improve from 20.22 BLEU points for our simplest model to 25.08 BLEU points for an improvement of 4.86 points or 24% relative.


## 1 Introduction

Statistical machine translation (SMT) from English-to-Turkish poses a number of difficulties. Typologically English and Turkish are rather distant languages: while English has very limited morphology and rather fixed SVO constituent order, Turkish is an agglutinative language with a very rich and productive derivational and inflectional morphology, and a very flexible (but SOV dominant) constituent order. Another issue of practical significance is the lack of large scale parallel text resources, with no substantial improvement expected in the near future.

In this paper, we investigate different representational granularities for sub-lexical representation of parallel data for English-to-Turkish phrase-based SMT and compare them with a word-based baseline. We also employ two-levels of language models: the decoder uses a morpheme based LM while it is generating an n-best list. The n-best lists are then rescored using a word-based LM.

The paper is structured as follows: We first briefly discuss issues in SMT and Turkish, and review related work. We then outline how we exploit morphology, and present results from our baseline and morphologically segmented models, followed by some sample outputs. We then describe discuss model iteration. Finally, we present a comprehensive discussion of our approach and results, and briefly discuss word-repair – fixing morphologicaly malformed words – and offer a few ideas about the adaptation of BLEU to morphologically complex languages like Turkish.

## 2 Turkish and SMT

Our previous experience with SMT into Turkish (Durgar El-Kahlout and Oflazer, 2006) hinted that exploiting sub-lexical structure would be a fruitful avenue to pursue. This was based on the observation that a Turkish word would have to align with a complete phrase on the English side, and that sometimes these phrases on the English side could be discontinuous. Figure 1 shows a pair of English and Turkish sentences that are aligned at the word (top) and morpheme (bottom) levels. At the morpheme level, we have split the Turkish words into their lexical morphemes while English words with overt morphemes have been stemmed, and such morphemes have been marked with a tag.

The productive morphology of Turkish implies potentially a very large vocabulary size. Thus, sparseness which is more acute when very modest



*Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32,
Prague, June 2007. 

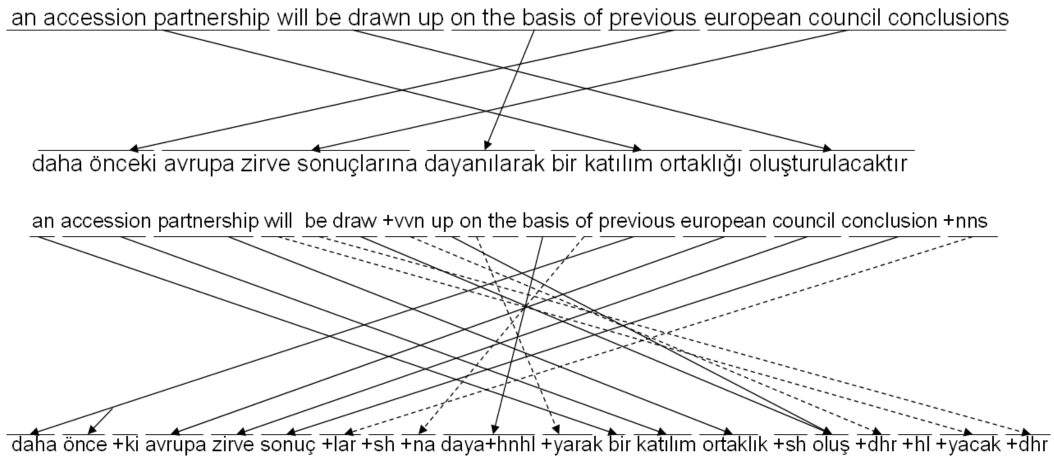


Figure 1: Word and morpheme alignments for a pair of English-Turkish sentences

parallel resources are available becomes an important issue. However, Turkish employs about 30,000 root words and about 150 distinct suffixes, so when morphemes are used as the units in the parallel texts, the sparseness problem can be alleviated to some extent.

Our approach in this paper is to represent Turkish words with their morphological segmentation. We use lexical morphemes instead of surface morphemes, as most surface distinctions are manifestations of word-internal phenomena such as vowel harmony, and morphotactics. With lexical morpheme representation, we can abstract away such word-internal details *and* conflate statistics for seemingly different suffixes, as at this level of representation words that look very different on the surface, look very similar.[1] For instance, although the words *evinde* 'in his house' and *masasında* 'on his table' look quite different, the lexical morphemes except for the root are the same: `ev+sH+ndA` vs. `masa+sH+ndA`.

We should however note that although employing a morpheme based representations dramatically reduces the vocabulary size on the Turkish side, it also runs the risk of overloading distortion mechanisms to account for *both* word-internal morpheme sequencing and sentence level word ordering.

The segmentation of a word in general is not unique. We first generate a representation that contains both the lexical segments and the morphological features encoded for all possible segmentations and interpretations of the word. For the word *emeli* for instance, our morphological analyzer generates the following with lexical morphemes bracketed with `(..)`:

`(em)em+Verb+Pos(+yAlH)^DB+Adverb+Since`
*since (someone) sucked (something)*
`(emel)emel+Noun+A3sg(+sH)+P3sg+Nom`
*his/her ambition*
`(emel)emel+Noun+A3sg+Pnon(+yH)+Acc`
*ambition (as object of a transitive verb)*

These analyses are then disambiguated with a statistical disambiguator (Yüret and Türe, 2006) which operates on the morphological features.[2] Finally, the morphological features are removed from each parse leaving the lexical morphemes.

Using morphology in SMT has been recently addressed by researchers translation from or into morphologically rich(er) languages. Niessen and Ney (2004) have used morphological decomposition to improve alignment quality. Yang and Kirchhoff (2006) use phrase-based backoff models to translate words that are unknown to the decoder, by morphologically decomposing the unknown source word. They particularly apply their method to translating *from* Finnish – another language with very similar structural characteristics to Turkish. Corston-Oliver and Gamon (2004) normalize inflectional morphology by stemming the word for German-English word alignment. Lee (2004) uses a morphologically analyzed and tagged parallel corpus for Arabic-English SMT. Zolmann et al. (2006) also exploit morphology in Arabic-English SMT. Popovic and Ney (2004) investigate improving translation qual-

[1] This is in a sense very similar to the more general problem of lexical redundancy addressed by Talbot and Osborne (2006) but our approach does not require the more sophisticated solution there.

[2] This disambiguator has about 94% accuracy.

ity from inflected languages by using stems, suffixes and part-of-speech tags. Goldwater and McClosky (2005) use morphological analysis on Czech text to get improvements in Czech to English SMT. Recently, Minkov et al. (2007) have used morphological postprocessing *on the output side* using structural information and information from the source side, to improve SMT quality.

## 3 Exploiting Morphology

Our parallel data consists mainly of documents in international relations and legal documents from sources such as the Turkish Ministry of Foreign Affairs, EU, etc. We process these as follows: (i) We segment the words in our Turkish corpus into lexical morphemes whereby differences in the surface representations of morphemes due to word-internal phenomena are abstracted out to improve statistics during alignment.[3] (ii) We tag the English side using TreeTagger (Schmid, 1994), which provides a *lemma* and a *part-of-speech* for each word. We then remove any tags which do not imply an explicit morpheme or an exceptional form. So for instance, if the word *book* gets tagged as *+NN*, we keep *book* in the text, but remove *+NN*. For *books* tagged as *+NNS* or *booking* tagged as *+VVG*, we keep *book* and *+NNS*, and *book* and *+VVG*. A word like *went* is replaced by *go +VVD*.[4] (iii) From these morphologically segmented corpora, we also extract for each sentence, the sequence of roots for open class content words (nouns, adjectives, adverbs, and verbs). For Turkish, this corresponds to removing *all* morphemes and any roots for closed classes. For English, this corresponds to removing all words tagged as closed class words along with the tags such as *+VVG* above that signal a morpheme on an open class content word. We use this to augment the training corpus and bias content word alignments, with the hope that such roots may get a chance to align without any additional "noise" from morphemes and other function words.

From such processed data, we compile the data sets whose statistics are listed in Table 1. One can note that Turkish has many more distinct word forms (about twice as many as English), but has much less number of distinct content words than English.[5] For language models in decoding and n-best list rescoring, we use, in addition to the training data, a monolingual Turkish text of about 100,000 sentences (in a segmented and disambiguated form).

| **Turkish** | **Sent.** | **Words (UNK)** | **Uniq. Words** |
|---|---|---|---|
| **Train** | 45,709 | 557,530 | 52,897 |
| **Train-Content** | 56,609 | 436,762 | 13,767 |
| **Tune** | 200 | 3,258 | 1,442 |
| **Test** | 649 | 10,334 (545) | 4,355 |
| **English** | | | |
| **Train** | 45,709 | 723,399 | 26,747 |
| **Train-Content** | 56,609 | 403,162 | 19,791 |
| **Test** | 649 | 13,484 (231) | 3,220 |

| **Turkish** | **Morphemes** | **Uniq. Morp.** | **Morp./ Word** | **Uniq. Roots** | **Uniq. Suff.** |
|---|---|---|---|---|---|
| **Train** | 1,005,045 | 15,081 | 1.80 | 14,976 | 105 |
| **Tune** | 6,240 | 859 | 1.92 | 810 | 49 |
| **Test** | 18,713 | 2,297 | 1.81 | 2,220 | 77 |

Table 1: Statistics on Turkish and English training and test data, and Turkish morphological structure

A typical sentence pair in our data looks like the following, where we have highlighted the content root words with bold font, coindexed them to show their alignments and bracketed the "words" that evaluation on test would consider.

- **T:** [**kat**$_1$ +hl +ma] [**ortaklık**$_2$ +sh +nhn] [**uygula**$_3$ +hn +ma +sh] [,] [**ortaklık**$_4$] [**anlaşma**$_5$ +sh] [**çerçeve**$_6$ +sh +nda] [**izle**$_7$ +hn +yacak +dhr] [.]
- **E:** the **implementation**$_3$ of the **accession**$_1$ **partnership**$_2$ will be **monitor**$_7$ +vvn in the **framework**$_6$ of the **association**$_4$ **agreement**$_5$ .

Note that when the morphemes/tags (starting with a +) are concatenated, we get the "word-based" version of the corpus, since surface words are directly recoverable from the concatenated representation. We use this word-based representation also for word-based language models used for rescoring.

We employ the phrase-based SMT framework (Koehn et al., 2003), and use the Moses toolkit (Koehn et al., 2007), and the SRILM language modelling toolkit (Stolcke, 2002), and evaluate our decoded translations using the BLEU measure (Papineni et al., 2002), using a *single* reference translation.

[3]So for example, the surface plural morphemes *+ler* and *+lar* get conflated to *+lAr* and their statistics are hence combined.

[4]Ideally, it would have been very desirable to actually do derivational morphological analysis on the English side, so that one could for example analyze *accession* into *access* plus a marker indicating nominalization.

[5]The training set in the first row of 1 was limited to sentences on the Turkish side which had at most 90 tokens (roots and bound morphemes) in total in order to comply with requirements of the GIZA++ alignment tool. However when only the content words are included, we have more sentences to include since much less number of sentences violate the length restriction when morphemes/function word are removed.

| Moses Dec. Parms. | BLEU | BLEU-c |
|---|---|---|
| Default | 16.29 | 16.13 |
| dl = -1, -weight-d = 0.1 | 20.16 | 19.77 |

Table 2: BLEU results for baseline experiments.
BLEU is for the model trained on the training set
BLEU-C is for the model trained on training set augmented with the content words.

### 3.1 The Baseline System

As a baseline system, we trained a model using default Moses parameters (e.g., maximum phrase length = 7), using the word-based training corpus. The English test set was decoded with both default decoder parameters and with the distortion limit (*-dl* in Moses) set to *unlimited* (-1 in Moses) and distortion weight (*-weight-d* in Moses) set to a very low value of 0.1 to allow for long distance distortions.[6] We also augmented the training set with the content word data and trained a second baseline model. Minimum error rate training with the tune set did not provide any tangible improvements.[7] Table 2 shows the BLEU results for baseline performance. It can be seen that adding the content word training data actually hampers the baseline performance.

### 3.2 Fully Morphologically Segmented Model

We now trained a model using the fully morphologically segmented training corpus *with* and *without content word parallel corpus augmentation*. For decoding, we used a 5-gram *morpheme-based* language model with the hope of capturing *local morphotactic ordering* constraints, and perhaps some sentence level ordering of words.[8] We then decoded and obtained 1000-best lists. The 1000-best sentences were then converted to "words" (by concatenating the morphemes) and then rescored with a 4-gram word-based language model with the hope of enforcing more distant *word sequencing* constraints. For this, we followed the following procedure: We tried various linear combinations of the word-based language model and the translation model scores on the *tune* corpus, and used the combination that performed best to evaluate the *test* corpus. We also experimented with both the default decoding parameters, and the modified parameters used in the baseline model decoding above.

The results in Table 3 indicate that the default decoding parameters used by the Moses decoder provide a very dismal results – much below the baseline scores. We can speculate that as the constituent orders of Turkish and English are very different, (root) words may have to be scrambled to rather long distances *along with* the translations of functions words and tags on the English side, to morphemes on the Turkish side. Thus limiting maximum distortion and penalizing distortions with the default higher weight, result in these low BLEU results. Allowing the decoder to consider longer range distortions and penalizing such distortions much less with the modified decoding parameters, seem to make an enormous difference in this case, providing close to almost 7 BLEU points improvement.[9]

We can also see that, contrary to the case with the baseline word-based experiments, using the additional content word corpus for training actually provides a tangible improvement (about 6.2% relative (w/o rescoring)), most likely due to slightly better alignments when content words are used.[10] Rescoring the 1000-best sentence output with a 4-gram word-based language model provides an additional 0.79 BLEU points (about 4% relative) – from 20.22 to 21.01 – for the model with the basic training set, and an additional 0.71 BLEU points (about 3% relative) – from 21.47 to 22.18– for the model with the augmented training set. The cumulative improvement is 1.96 BLEU points or about 9.4% relative.

### 3.3 Selectively Segmented Model

A systematic analysis of the alignment files produced by GIZA++ for a small subset of the training sentences showed that certain morphemes on the

[6]We arrived at this combination by experimenting with the decoder to avoid the almost monotonic translation we were getting with the default parameters.

[7]We ran MERT on the baseline model and the morphologically segmented models forcing *-weight-d* to range a very small around 0.1, but letting the other parameters range in their suggested ranges. Even though the procedure came back claiming that it achieved a better BLEU score on the tune set, running the new model on the test set did not show any improvement at all. This may have been due to the fact that the initial choice of *-weight-d* along with *-dl* set to 1 provides such a drastic improvement that perturbations in the other parameters do not have much impact.

[8]Given that on the average we have almost two bound morphemes per "word" (for inflecting word classes), a morpheme 5-gram would cover about 2 "words".

[9]The "morpheme" BLEU scores are much higher (34.43 on the test set) where we measure BLEU *using decoded morphemes* as tokens. This is just indicative and but correlates with word-level BLEU which we report in Table 3, and can be used to gauge relative improvements to the models.

[10]We also constructed phrase tables only from the actual training set (w/o the content word section) *after* the alignment phase. The resulting models fared slightly worse though we do not yet understand why.

| Moses Dec. Parms. | BLEU | BLEU-c |
|---|---|---|
| Default | 13.55 | NA |
| dl = -1, -weight-d = 0.1 | 20.22 | 21.47 |
| dl = -1, -weight-d = 0.1<br>**+ word-level LM rescoring** | 21.01 | 22.18 |

Table 3: BLEU results for experiments with fully morphologically segmented training set

Turkish side were almost consistently never aligned with anything on the English side: e.g., the compound noun marker morpheme in Turkish (+sh) does not have a corresponding unit on the English side since English noun-noun compounds do not carry any overt markers. Such markers were never aligned to anything or were aligned almost randomly to tokens on the English side. Since we perform derivational morphological analysis on the Turkish side but not on the English side, we noted that most verbal nominalizations on the English side were just aligned to the verb roots on the Turkish side and the additional markers on the Turkish side indicating the nominalization and agreement markers etc., were mostly unaligned.

For just these cases, we selectively attached such morphemes (and in the case of verbs, the intervening morphemes) to the root, but otherwise kept other morphemes, especially any case morphemes, still by themselves, as they almost often align with prepositions on the English side quite accurately.[11]

This time, we trained a model on just the content-word augmented training corpus, with the better performing parameters for the decoder and again did 1000-best rescoring.[12] The results for this experiment are shown in Table 4. The resulting BLEU represents 2.43 points (11% relative) improvement over the best fully segmented model (and 4.39 points 21.7% compared to the very initial morphologically segmented model). This is a very encouraging result that indicates we should perhaps consider a much more detailed analysis of morpheme alignments to uncover additional morphemes with similar status. Table 5 provides additional details on the BLEU scores for this model, for different ranges of (English source) sentence length.

[11]It should be noted that what to selectively attach to the root should be considered on a per-language basis; if Turkish were to be aligned with a language with similar morphological markers, this perhaps would not have been needed. Again one perhaps can use methods similar to those suggested by Talbot and Osborne (2006).

[12]Decoders for the fully-segmented model and selectively segmented model use different 5-gram language models, since the language model corpus should have the same selectively segmented units as those in the training set. However, the word-level language models used in rescoring are the same.

| Moses Dec. Parms. | BLEU-c |
|---|---|
| dl = -1, -weight-d = 0.1<br>**+ word-level LM rescoring**<br>**(Full Segmentation (from Table 3))** | 22.18 |
| dl = -1, -weight-d = 0.1 | 23.47 |
| dl = -1, -weight-d = 0.1<br>**+ word-level LM rescoring** | **24.61** |

Table 4: BLEU results for experiments with selectively segmented and content-word augmented training set

| Range | Sent. | BLEU-c |
|---|---|---|
| 1 - 10 | 172 | 44.36 |
| 1 - 15 | 276 | 34.63 |
| **5 - 15** | **217** | **33.00** |
| 1 - 20 | 369 | 28.84 |
| 1 - 30 | 517 | 27.88 |
| 1 - 40 | 589 | 24.90 |
| All | 649 | 24.61 |

Table 5: BLEU Scores for different ranges of (source) sentence length for the result in Table 4

## 4 Sample Rules and Translations

We have extracted some additional statistics from the translations produced from English test set. Of the 10,563 words in the decoded test set, a total of 957 words (9.0 %) were not seen in the training corpus. However, interestingly, of these 957 words, 432 (45%) were actually morphologically well-formed (some as complex as having 4-5 morphemes!) This indicates that the *phrase-based translation model is able to synthesize novel complex words.*[13] In fact, some phrase table entries seem to capture morphologically marked subcategorization patterns. An example is the phrase translation pair

`after examine +vvg` ⇒
`+acc incele+dhk +abl sonra`

which very much resembles a typical structural transfer rule one would find in a symbolic machine translation system

`PP(after examine +vvg` $NP_{eng}$`)` ⇒
`PP(`$NP_{turk}$`+acc incele+dhk +abl sonra)`

in that the accusative marker is tacked to the translation of the English NP.

Figure 2 shows how segments are translated to Turkish for a sample sentence. Figure 3 shows the translations of three sentences from the test data

[13]Though whether such words are actually correct in their context is not necessarily clear.

```
çocuk [[ child ]]
 hak+lar+sh +nhn [[ +nns +pos right ]]
 koru+hn+ma+sh [[ protection ]]
 +nhn [[ of ]]
 teşvik et+hl+ma+sh [[ promote ]]
 +loc [[ +nns in ]] ab [[ eu ]]
 ve ulus+lararasi standart +lar
    [[ and international standard +nns ]]
 +dat uygun [[ line with ]] +dhr .  [[ .]]
```

Figure 2: Phrasal translations selected for a sample sentence

**Inp.**: 1 . everyone's right to life shall be protected by law .
**Trans.**: 1 . herkesin yaşama hakkı kanunla korunur.
**Lit.**: everyone's living right is protected with law .
**Ref.**: 1 . herkesin yaşam hakkı yasanın koruması altındadır .
**Lit.**: everyone's life right is under the protection of the law.

**Inp.**: promote protection of children's rights in line with eu and international standards .
**Trans.**: çocuk haklarının korunmasının **ab ve uluslararası standartlar**a uygun şekilde geliştirilmesi.
**Lit.**: develop protection of children's rights in accordance with eu and international standards .
**Ref.**: **ab ve uluslararası standartlar** doğrultusunda çocuk haklarının korunmasının teşvik edilmesi.
**Lit.**: in line with eu and international standards promote/motivate protection of children's rights .

**Inp.**: as a key feature of such a strategy, an accession partnership will be drawn up on the basis of previous european council conclusions.
**Trans.**: bu stratejinin kilit unsuru bir katılım ortaklığı belgesi hazırlanacak kadarın temelinde , bir önceki avrupa konseyi sonuçlarıdır .
**Lit.**: as a key feature of this strategy, accession partnership document will be prepared ??? based are previous european council resolutions .
**Ref.**: bu stratejinin kilit unsuru olarak , daha önceki ab zirve sonuçlarına dayanılarak bir katılım ortaklığı oluşturulacaktır.
**Lit.**: as a key feature of this strategy an accession partnership based on earlier eu summit resolutions will be formed .

Figure 3: Some sample translations

along with the literal paraphrases of the translation and the reference versions. The first two are quite accurate and acceptable translations while the third clearly has missing and incorrect parts.

## 5 Model Iteration

We have also experimented with an iterative approach to use multiple models to see if further improvements are possible. This is akin to post-editing (though definitely not akin to the much more sophisticated approach in described in Simard et al. (2007)). We proceeded as follows: We used the selective segmentation based model above and decoded our English *training* data $\mathtt{E}_{Train}$ and English test data $\mathtt{E}_{Test}$ to obtain $\mathtt{T1}_{Train}$ and $\mathtt{T1}_{Test}$ respectively. We then trained the next model using $\mathtt{T1}_{Train}$ and $\mathtt{T}_{Train}$, to build a model that hopefully will improve upon the output of the previous model, $\mathtt{T1}_{Test}$, to bring it closer to $\mathtt{T}_{Test}$. This model when applied to $\mathtt{T1}_{Train}$ and $\mathtt{T1}_{Test}$ produce $\mathtt{T2}_{Train}$ and $\mathtt{T2}_{Test}$ respectively.

| Step | BLEU |
|---|---|
| From Table 4 | 24.61 |
| Iter. 1 | 24.77 |
| Iter. 2 | **25.08** |

Table 6: BLEU results for two model iterations

We have not included the content word corpus in these experiments, as (i) our few very preliminary experiments indicated that using a morpheme-based models in subsequent iterations would perform worse than word-based models, and (ii) that for word-based models adding the content word training data was not helpful as our baseline experiments indicated. The models were tested by decoding the output of the previous model for original test data. For word-based decoding in the additional iterations we used a 3-gram word-based language model but reranked the 1000-best outputs using a 4-gram language model. Table 6 provides the BLEU results for these experiments corresponding to two additional model iterations.

The BLEU result for the second iteration, 25.08, represents a cumulative 4.86 points (24% relative) improvement over the initial fully morphologically segmented model using only the basic training set and no rescoring.

## 6 Discussion

Translation into Turkish seems to involve processes that are somewhat more complex than standard statistical translation models: sometimes words on the Turkish side are synthesized from the translations of two or more (SMT) phrases, and errors in any translated morpheme or its morphotactic position render the synthesized word incorrect, even though the rest of the word can be quite fine. If we just extract the root words (not just for content words but all words) in the decoded test set and the reference set, and compute *root word* BLEU, we obtain 30.62, [**64.6**/35.7/23.4/16.3]. The unigram precision score shows that we are getting almost 65% of the root words correct. However, the unigram precision score with full words is about 52% for our best model. Thus we are missing about 13% of the words *although we seem to be getting their roots* *correct*. With a tool that we have developed, *BLEU+* (Tantuğ et al., 2007), we have investigated such mismatches and have found that most of these are actually morphologically bogus, in that, although they have the root word right, the morphemes are either not the applicable ones or are in a morphotactically wrong position. These can easily be identified with the morphological generator that we have. In many cases, such morphologically bogus words are *one morpheme edit distance away* from the correct form in the reference file. Another avenue that *could* be pursued is the use of skip language models (supported by the SRILM toolkit) so that the content word order could directly be used by the decoder.[14]

At this point it is very hard to compare how our results fare in the grand scheme of things, since there is not much prior results for English to Turkish SMT. Koehn (2005) reports on translation from English to Finnish, another language that is morphologically as complex as Turkish, with the added complexity of compounding and stricter agreement between modifiers and head nouns. A standard phrase-based system trained with 941,890 pairs of sentences (about 20 times the data that we have!) gives a BLEU score of 13.00. However, in this study, nothing specific for Finnish was employed, and one can certainly employ techniques similar to presented here to improve upon this.

### 6.1 Word Repair

The fact that there are quite many erroneous words which are actually easy to fix suggests some ideas to improve unigram precision. One can utilize a morpheme level "spelling corrector" that operates on segmented representations, and corrects such forms to possible morphologically correct words in order to form a lattice which can again be rescored to select the contextually correct one.[15] With the *BLEU+* tool, we have done one experiment that shows that *if* we could recover all morphologically bogus words that are 1 and 2 morpheme edit distance from the correct form, the word BLEU score could rise to 29.86, [60.0/34.9/23.3/16.] and 30.48 [63.3/35.6/23.4/16.4] respectively. Obviously, these are upper-bound oracle scores, as subsequent candidate generation and lattice rescoring could make errors, but nevertheless they are very close to the root word BLEU scores above.

Another path to pursue in repairing words is to identify morphologically correct words which are either OOVs in the language model or for which the language model has low confidence. One can perhaps identify these using posterior probabilities (e.g., using techniques in Zens and Ney (2006)) and generate additional morphologically valid words that are "close" and construct a lattice that can be rescored.

### 6.2 Some Thoughts on BLEU

BLEU is particularly harsh for Turkish and the morpheme based-approach, because of the all-or-none nature of token comparison, as discussed above. There are also cases where words with different morphemes have very close morphosemantics, convey the relevant meaning and are almost interchangeable:

- *gel+hyor* (geliyor - he is coming) vs. *gel+makta* (gelmekte - he is (in a state of) coming) are essentially the same. On a scale of 0 to 1, one could rate these at about 0.95 in similarity.
- *gel+yacak* (gelecek - he will come) vs. *gel+yacak+dhr* (gelecektir - he *will* come) in a sentence final position. Such pairs could be rated perhaps at 0.90 in similarity.
- *gel+dh* (geldi - he came (past tense)) vs. *gel+mhs* (gelmiş - he came (hearsay past tense)). These essentially mark past tense but differ in how the speaker relates to the event and could be rated at perhaps 0.70 similarity.

Note that using stems and their synonyms as used in METEOR (Banerjee and Lavie, 2005) could also be considered for word similarity.

Again using the *BLEU+* tool and *a slightly different formulation of token similarity* in BLEU computation, we find that using morphological similarity our best score above, 25.08 BLEU increases to 25.14 BLEU, while using only root word synonymy and very close hypernymy from Wordnet, gives us 25.45 BLEU. The combination of rules and Wordnet match gives 25.46 BLEU. Note that these increases are much less than what can (potentially) be gained from solving the word-repair problem above.

## 7 Conclusions

We have presented results from our investigation into using different granularity of sub-lexical representations for English to Turkish SMT. We have found that employing a language-pair specific representation somewhere in between using full word-forms and fully morphologically segmented representations and using content words as additional

[14]This was suggested by one of the reviewers.

[15]It would however perhaps be much better if the decoder could be augmented with a filter that could be invoked at much earlier stages of sentence generation to check if certain generated segments violate *hard-constraints* (such as morphotactic constraints) regardless of what the statistics say.

data provide a significant boost in BLEU scores, in addition to contributions of word-level rescoring of 1000-best outputs and model iteration, to give a BLEU score of 25.08 points with very modest parallel text resources. Detailed analysis of the errors point at a few directions such as word-repair, to improve word accuracy. This also suggests perhaps hooking into the decoder, a mechanism for imposing hard constraints (such as morphotactic constraints) during decoding to avoid generating morphologically bogus words. Another direction is to introduce exploitation of limited structures such as bracketed noun phrases before considering full-fledged syntactic structure.


## Acknowledgements

This work was supported by TÜBİTAK – The Turkish National Science and Technology Foundation under project grant 105E020. We thank the anonymous reviewer for some very useful comments and suggestions.


## References


Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June.

Simon Corston-Oliver and Michael Gamon. 2004. Normalizing German and English inflectional morphology to improve statistical word alignment. In *Proceedings of AMTA*, pages 48–57.

İlknur Durgar El-Kahlout and Kemal Oflazer. 2006. Initial explorations in English to Turkish statistical machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 7–14, New York City, June.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, British Columbia, Canada, October.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07) – Companion Volume*, June.

Philip Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, Phuket, Thailand.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004 - Companion Volume*, pages 57–60.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, June.

Sonja Niessen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntatic information. *Computational Linguistics*, 30(2):181–204.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, University of Pennsylvania.

Maja Popovic and Hermann Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1585–1588, May.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Proceedings of NAACL*, April.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the Intl. Conf. on Spoken Language Processing*.

David Talbot and Miles Osborne. 2006. Modelling lexical redundancy for machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 969–976, Sydney, Australia, July.

Cüneyd Tantuğ, Kemal Oflazer, and İlknur Durgar El-Kahlout. 2007. BLEU+: a tool for fine-grained BLEU computation. in preparation.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of EACL*, pages 41–48.

Deniz Yüret and Ferhan Türe. 2006. Learning morphological disambiguation rules for Turkish. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 328–334, New York City, USA, June.

Richard Zens and Hermann Ney. 2006. N-gram posterior probabilities for statistical machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 72–77, New York City, June. Association for Computational Linguistics.

Andreas Zollmann, Ashish Venugopal, and Stephan Vogel. 2006. Bridging the inflection morphology gap for Arabic statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 201–204, New York City, USA, June.