Projecting Farsi POS Data To Tag Pashto

Mohammad Khan, Eric Baucom, Anthony Meyer, Lwin Moe
Indiana University
{khanms,eabaucom,antmeyer,lwinmoe}@indiana.edu

Abstract

We present our findings on projecting part of speech (POS) information from a well resourced language, Farsi, to help tag a lower resourced language, Pashto, following Feldman and Hana (2010). We make a series of modifications to both tag transition and lexical emission parameter files generated from a hidden Markov model tagger, TnT, trained on the source language (Farsi). Changes to the emission parameters are immediately effective, whereas changes made to the transition information are most effective when we introduce a custom tagset. We reach our best results of 70.84% when we employ all emission and transition modifications to the Farsi corpus with the custom tagset.

1 Introduction

State-of-the-art work in computational linguistics typically requires heavy investment in language-specific resources. Large-scale resources, in the form of corpora with part of speech (POS), syntactic, or semantic annotation schemes, are used in nearly all statistically driven natural language processing applications. For global languages like English, these resources are already present in at least some form, but, for less commonly taught languages like Pashto, they are not.

Work has already been done exploring how to rapidly develop resources for less commonly taught languages. Feldman and Hana (2010) present a method utilizing Hidden Markov Model (HMM) POS tagging information from a well resourced language (Czech) to help tag a lower resourced language (Russian). They perform various modifications on the two kinds of parameter files generated by the HMM, lexical and transition, in order to make a closer fit between the source and target languages. Following the same basic approach, we perform various syntactic transformations, or "Pashtifications", on the training input in order to improve the tag transition information from the source language, and also develop a tagset that is suitable to both source and target. To improve lexical emission information from the source language, we use "cognate" analysis (employing minimum edit distance), rudimentary morphological analysis (based on suffixes and focusing on verbs), along with enrichment of the source lexicon (by adding closed class words of the target language).

We perform a series of experiments involving different combinations of these strategies and evaluate on a small hand-tagged test set. The aim of this project is to rapidly develop a resource for a lower-resourced language using as little language-specific information as possible. In theory, this could be performed without any in-depth knowledge of the target language, though in our case we did have a native Pashto speaker to assist in tagging our gold standard for evaluation.

The rest of the paper is structured as follows: we begin by discussing related work in section 2; then in section 3 we give some background on Pashto morphosyntax; next we discuss the corpora and tagsets used in our experiments (section 4), followed by the experiments and results themselves (section 6); finally, we offer conclusions and discuss future work in section 7.

2 Related Work

2.1 POS Tagging for Low-resourced Languages

Feldman and Hana (2010) provide the basic approach that we followed in our method. They use annotated corpora from comparatively well resourced languages to provide information about morphological/POS tagging in related, under-resourced languages. In their HMM model for POS tagging, they use transitional probabilities from the source language (with or without modifications) along with lexical emission probabilities for the target language derived through various means.

For their transition probabilities that they de-

Proceedings of the Student Research Workshop associated with RANLP 2011, pages 25–32,
Hissar, Bulgaria, 13 September 2011.

rived from Czech, Feldman and Hana (2010) introduced some "Russifications" to the Czech training data to make it more similar to Russian syntax (target language). Most of the changes from the source involved changing a particle to an affix, or vice versa.

For lexical emission probabilities, Feldman and Hana (2010) combine different methods to obtain the best results. They obtained "cognates" from source languages by looking at Levenshtein distance and used the gold POS tags to count word and tag maximum likelihood estimate (MLE) frequencies. They also used a morphological analyzer, developed by hand with the help of language experts, to inform the lexical probabilities.

2.2 Pashto POS Tagging

Rabbi et al. (2009) present a rule-based POS tagger for Pashto. Their method is to manually tag a lexicon and then use that lexicon and Pashto specific rules to tag unknown tokens. They reach an accuracy of 88% with 100 000 tagged words in the lexicon and 120 Pashto specific rules. This approach achieves good results but requires a very large manually tagged lexicon and a large manually created set of language specific rules in order to do so. Operating within a resource-light paradigm, our aim is to reach comparable results using less time and effort.

3 Pashto Morphology and Syntax

Pashto has a rich morphology. It uses three forms of affixation: prefixes, infixes, and suffixes. The morphology represents gender, number, case, tense, and aspect. There is also ambiguity among the morphemes. For example, the suffix *-wo* is used as an oblique plural marker for nouns and adjectives, and as a past tense maker in one class of verbs.

Verbs are ergative in Pashto, i.e. they agree with the subjects in present imperfective cases, while in the past tense, the verb agrees with the object regardless of the aspect. Pashto has "subject object verb" word order, but that order is relatively flexible if compared to English. There are two types of verbs in Pashto: compound verbs and non-compound verbs. Compound verbs are derived by adding a light verb (similar to "be," "do," etc.) to a noun or adjective. Non-compound verbs, which are less common, are not derived. For example, the word for "sharpening" is *terə kawəl* ("sharp"+"do"), while the word for "to go" *tləl* is not derived from a noun or adjective. Compound verbs are written as one or two words depending on the phonotactic properties of the compounding elements.

As compared to Farsi, noun-noun compounding is relatively less common in Pashto. In spoken Farsi, such nouns end with an audible vowel affix known as the *harf-e-izafat*, but this suffix is not actually written in Farsi text. Such compounding is formed by using prepositions in front of the first noun in Pashto. For example "computer table" is formed from *də kamputʂər mez* meaning "of computer table" in Pashto. Pashto uses postpositions as well. For example, the phrase for "in the stream" is formed as *pə wyalə ke* ("the stream in").

Adjectives that modify a noun precede the modified nouns, and the intensifiers precede the adjectives, as in English.

4 Corpora and Tagsets

4.1 Farsi

Farsi is a sister language of Pashto spoken in the same geographical area. It is also the official language of Iran. Farsi has a large lexical similarity with Pashto. It shares a large number of cognates and borrowed terms (from Arabic). The syntaxes of the two languages do differ to some extent. However, Farsi is the only language we found that is close enough to Pashto and has enough resources to be useful in our task.[1] We therefore used Farsi as the source language in our experiments.

The Bijankhan Corpus[2] (Oroumchian et al., 2006) is a freely available Farsi corpus. This corpus was manually tagged for POS at the University of Tehran in Iran. The corpus is a collection of 4 300 articles from the daily news and other common texts. It has 2.6 million tokens. The tagset used to tag the corpus consists of 550 different POS tags and is described further in section 4.3.

4.2 Pashto

Test corpora We use two different corpora for testing and development. The first corpus is a hand-tagged corpus of spoken Pashto, which consists of dialogues between an English speaker and a Pashto speaker mediated by an interpreter. The

[1] Urdu is another close language but, compared to Farsi, is not as resource rich.

[2] `http://ece.ut.ac.ir/dbrg/bijankhan/`

spoken corpus consists of 708 tokens and is based on news data. We hand-annotated 375 tokens of news articles.

Web corpus In order to improve our lexical emission probabilities in the tagger, we needed to conduct both morphological and cognate analysis. A large amount of raw text in our target language, Pashto, was needed for the two processes. Since we could not find any such resource that was both readily available and in the appropriate domain, we decided to obtain our own corpus from the web.

We used `BootCaT` (Baroni and Bernardini, 2004) with appropriate seeds (such as words containing one or more of eight Pashto-specific characters, unique closed class words, etc.) to find Pashto websites. We then used `wget` to obtain a web-corpus of 473 MBs (text only) in size. We then extracted a Pashto lexicon of more than a million words. This lexicon was used in the morphological analysis and cognate detection.

4.3 Tagsets

4.3.1 The BijanKhan Corpus Tagset

The Bijankhan tagset, containing 550 tags, has a hierarchical structure, with most full tags comprising three or more tiers. The first tier specifies the coarse, primary word class; the second tier specifies either a word subclass or a piece of morphological information; and the third tier often expresses information of a semantic nature. For example, the tag `N_SING_LOC` means that the word in question is 1. a noun, 2. singular, and 3. a location of some kind (e.g. Bloomington). Note that delimiter between the tiers is the underscore symbol ("_"). In other tags, the third and fourth tiers express some grammatical or morphological, rather than semantic, nuance, as in `N_SING_CN_GEN` where the fourth-tier subtag `GEN` indicates *harf-e-izafat*, which is also used as a genitive marker in Farsi.

Because the original, or "extended" BijanKhan tagset of 550 tags can become impractical for NLP purposes and lead to data sparsity issues, Amiri et al. (2007) devise a systematic, if somewhat simplistic, method for dramatically reducing the tagset size. Their method essentially consists of the following steps: 1. for any tag with of three or more tiers, they eliminate any subtag past the second tier; 2. for two-tier tags, they remove the second tier if it is used rarely; and 3. they discard any whole tag that occurs rarely. They use this method to derive a tagset of just 40 tags.

4.3.2 Reducing the Extended Tagset

In our experiments, we used two tagsets. First, we used the reduced tagset of Amiri et al. (2007), rather than trying to work with full 550-tag tagset. Preliminary experiments then showed that this tagset did not provide enough morphosyntactic information and resulted in low accuracies. The problem in our case lay in our cross-lingual application of Amiri et al.'s (2007) tagset. The morphological and syntactic differences between Farsi and Pashto are such that much information pertinent to Pashto is destroyed by Amiri et al.'s (2007) simplistic tagset-reduction technique. The more nuanced information found in the extended tags' third and fourth tiers is often necessary for relating Farsi morphosyntactic categories and POS-tag sequences to those of Pashto. For instance, the tag `N_SING_LOC_GEN` (followed by another noun tag) is indicative of the Farsi noun-noun compound construction, i.e. N_1N_2. The equivalent Pashto expression requires the explicit use of the preposition *də* "of" (a stand-alone word), in addition to the reversal of the ordering of the two nouns, so that N_1N_2 (Farsi) $\rightarrow$ *də* N_2N_1 (Pashto). Several of our Pashtification rules involve noun phrases of this type, but their application is nearly impossible without access to the original extended tags.

We therefore decided to build our own tagset. We mapped the original extended tags to a reduced tagset of our own design, dubbing it the Pashto Extended Reduced Tagset (PERT). By starting with the extended tags from the Farsi corpus, we can provide the Pashtification rules with the fine-grained information they require. We also ensure that the final set of reduced tags is equally applicable to both Farsi and Pashto. Our goal was to remain as close to the original Farsi tagset nomenclature and design as possible. Our reduced tagset consists of the 39 tags presented in table 1. In the design of this tagset we could not include all the necessary categories for Pashto. For example, such important categories as gender, case, and aspect are missing, which could be a third tier of information added to an existing category. We could not add these because Farsi does not possess such grammatical categories. Similarly, we did not want to have categories such as `NP` for Pashto because its nonstandard orthography makes this category especially problematic, but needed to have

ADJ	Adjective
ADJ_INO	Participle
ADJ_ORD	Cardinal numbers
ADJ_SUP	Superlative adjective
ADV	Adverb
ADV_EXM	Adverb of examples
ADV_I	Interrogative adverb
ADV_LOC	Adverb of location
ADV_NEGG	Adverb of negation
ADV_NI	Negative interrogative adverb
ADV_TIME	Temporal adverb
AR	Arabic (foreign language)
CON	Conjunction
DET	Determiner
IF	Conditional if
MORP_SING	Singular morpheme
MORP_PL	Plural morpheme
QUA	Quantifier
MS	Mathematics symbol
N_PL	Plural noun
N_SING	Singular noun
NN	Numeric date
NP	Noun phrase
OH	Addressee
OHH	Addresser
P	Preposition
PP	Det. + Preposition
PRO	Pronoun
PS	Whole phrase
V_PRS	Present tense verb
MOD	Modal
V_PA	Past tense verb
V_IMP	Imperative verb
CL	Clitic
P_POS	Postposition
V_SUB	Subjunctive verb
INF	Infinitive
NEGG	Negation particle
DELM	Delimiter (e.g. commas, period)

Table 1: Pashto Extended Reduced Tagset (PERT)

them to stay consistent with Farsi tagset, a language with a more standardized orthographic convention. The syntactic distribution of the subcategories of adverbs vary from one subcategory to another in both languages. We therefore chose six subcategories of adverbs whose inclusion results in better transition probabilities.

5 Cross-language Projection

We now discuss the modifications made to the parameter files generated by TnT (Brants, 2000), our HMM POS tagger, after being trained on Farsi. The tag transition parameter file was modified via "Pashtification" in order to more closely model the POS tag sequences and morphosyntactic structure of Pashto. We discuss these modifications in section 5.1. The lexical emission parameter file was modified directly by adding closed class words and their POS tags and by adding other Pashto words with hypothesized POS tags based on analyses of our development corpora. We discuss these modifications in section 5.2.

5.1 Pashtification

To improve the transition probabilities obtained from the source language, we performed various syntactic modifications, or "Pashtifications", on the Farsi corpus. The changes were based on systematic syntactic differences between Pashto and Farsi, but did not require extensive Pashto knowledge.

One of our "Pashtifications" involved inserting Pashto prepositions into long noun-noun compounds in the Farsi corpus. Contrary to Pashto, Farsi allows intensive noun-noun compounding where the component nouns are joined by *harf-e-izafat* (spoken preposition). *Harf-e-izafat* is not written in Farsi and has multiple grammatical functions including genitive marking. But, as shown in section 3, the linking prepositions are made explicit in the Pashto orthography, so we inserted the Pashto linking prepositions into the original Farsi corpus.

We applied 47 "Pashtification" rules, most of which were related to verbs. Below is a synopsis.

Preposition insertion. Insert a preposition in the noun-noun chain whenever two or more nouns are tagged with a genitive subcategory.

Adjective-Noun inversion. As discussed earlier, adjectives precede nouns in Pashto, but follow their modified nouns in Farsi.

Indefinite article insertion. Indefinite articles are suffixed in Farsi adjectives or nouns. Pashto, on the other hand, uses a separate word before the noun or adjective. We applied a rule that makes this change by adding a determiner category before the noun or adjective.

Clitic insertion. Farsi uses personal affixes attached to nouns and adjectives to describe possession or belonging, such as "his book." Pashto, on the other hand, uses a clitic. We inserted Pashto clitics after the nouns tagged with the personal affixes.

Present tense verb rules. Verbs that are formed from an adjective or a noun root are marked as adjectivized or nominalized verbs in the Farsi corpus. Sometimes these verbs are written as two separate words in Farsi, yet they are tagged as one word. In Pashto, this type of construction almost always occurs as two separate words. For example, for an adjective followed by a present tense verb, we changed the one V_PRS_ADJ tag to two tags of ADJ and V_PRS. The same rule was applied to the verbs tagged as verb present, adverb, noun,

and pronoun. Also, in Farsi negation is inflected in verbs almost all the time while it is not the case with Pashto. Negative verbs were changed to negation plus verb in Pashto.

Past tense verb rules. We made similar changes to past tense verbs. Some participle plus verb constructions are treated simply as past tense in Farsi. These were tagged with a specific tag (V_PA_NAR_POS). We changed these to ADJ_INO (tag for participle) plus present as two different tags.

Auxiliary rules. The category AUX is extremely ambiguous in the BijanKhan corpus. It sometimes refers to a main verb in the matrix clause, or it refers to a modal such as *bayad* ("must"). We included several rules and new categories to exclude the need for an AUX category. Pashto does not use any auxiliary verbs other than for constructing participles. Participles are marked as ADJ_INO (accusative adjectives) in the BijanKhan corpus. Often the adjective and the verbal part are combined as one token despite the orthography showing two separate words. We used the subcategory portion of AUX tags to change the auxiliaries to present, past, or subjunctive categories if the auxiliary needed to be translated to a verb.

Imperative verb. Like other verbs, imperatives in Farsi are written with the negation inserted. In Pashto, the negation is not part of the verb. We applied a rule that separates the two.

***Ra* exclusion.** Farsi uses a direct object marker *ra* which we changed to P_POST (postposition). The Farsi accusative marker occurs in the same syntactic location as a P_POST in Pashto.

5.2 Lexical Modifications

5.2.1 Cognate Analysis

Farsi and Pashto share many "cognates", an umbrella term we are using to describe both true linguistic cognates (where the words share a common ancestor) and loan words (borrowed from the same language, in this case usually Arabic). We exploited this lexical similarity to improve our tagger by assuming that words we determined to be cognates would share similar tag distributions.

We used a normalized Levenshtein distance to detect cognates in Farsi and Pashto. Levenshtein distance is a measure of similarity between two strings (Levenshtein, 1966), so the intuition is that if words are spelled similarly, they will have similar meanings, or, crucially for our application, they will have the same POS tag distribution.

We first obtained a table listing all the edit distance scores of all possible word-word combinations between the Farsi corpus and our Pashto lexicon obtained from the web-based corpus. In order to avoid favoring shorter words, which have shorter edit distances by virtue of having fewer letters to permute, we normalized the score with the lengths of words in question. We chose the maximum length of the two words in consideration, and used that length to divide the Levenshtein distance to get the normalized score:

$$normalized\_score = \frac{Levenshtein\_distance}{maximum\_word\_length}$$

If the normalized score was below a certain threshold, then we added the Pashto version of the word to the lexicon, which was used by our tagger, along with the tag distribution from the Farsi word. If the Pashto word was similar to more than one Farsi words, we combined the tag distribution of all the Farsi words. If the word was already present in the lexicon, we simply used its tag distribution. For example, if Pashto word p is similar to Farsi word f, whose tag distribution is "ADJ 10, ADV 5", we added p with the tag distribution of f to our lexicon. If a Pashto word x already existed in our lexicon, the tag distribution of the Farsi cognate was the same, because the Pashto lexicon was originally generated from the Farsi corpus.

We ran a series of experiments, evaluating on our hand-tagged test set, to determine the optimal threshold to use for deciding which Farsi and Pashto words were cognates. We first ran three experiments with 0.3, 0.5 and 0.8 as our threshold values. We found that 0.3 gave us the best results. We then ran another series of experiments with these values—0.24, 0.25, 0.26, 0.27, 0.28, 0.29, 0.30, 0.31, and 0.32. We finally chose the best value, 0.28, from the experiments.

5.2.2 Morphological Analysis

We also modified the original lexical emission file from Farsi by including information from a morphological analysis of our Pashto web corpus. The morphological analysis proceeded as follows: we developed a short list of affixes that typically occur

with various parts of speech categories in Pashto; we then looped through our Pashto web corpus and checked whether the current word appeared with any of the affixes anywhere else in the web corpus; if the word did occur with those affixes above a certain threshold, we could then assume with a measure of confidence that that word should be tagged as indicated by the suffix.

An example in English: imagine our suffixes are {`-ed, -ing, -s`} and our current word in the corpus is "work". We now check to see if "worked", etc., occur elsewhere in the corpus. If the threshold is met, we then alter the entry for the word plus suffix in the tagging lexicon, hopefully improving the lexical information for tagging.

5.2.3 Lexical Enrichment

To further improve the lexicon for the tagger, we added a set of closed class words. We chose the 200 most frequent words from the Bijankhan corpus for translation into Pashto. Our cognate detector was able to detect 91 of those. We therefore only had to translate 109 Farsi words into Pashto. We replaced these Farsi words with their Pashto equivalents. Not all the closed class elements were included in the two hundred frequent words—we therefore added 24 of the most common prepositions, postpositions, pronouns, and conjunctions to the training lexicon. We also included most (42) forms of light verbs. We believe that such language information does not constitute an intensive language resource. It can be obtained from any Pashto grammar resource in approximately 3-4 hours.

6 Experiments and Results

We performed two sets of experiments, in which we used two different tagsets (discussed in section 4.3.2). In the first set of experiments (section 6.1), we used the Amiri et al. (2007) tagset. In the second set of experiments (section 6.2), we used our custom tagset PERT.

6.1 Experiments with Amiri et al. (2007) Tagset

We ran a series of experiments combining different amounts and levels of information to see which provided the most help in tagging our Pashto test corpus using the Amiri et al. (2007) tagset. As a baseline, we determined that the most common tag in our test corpus was `N_SING`, the singular noun, and labeled every word with it. This naive approach was 16.62% accurate, meaning 16.62% of the words in the test corpus were singular nouns according to the gold standard.

Our biggest improvement over this baseline came from enriching the lexicon with closed class Pashto words (table 2, row 2). Other modifications like adding information from the morphological and cognate analyses did help, but not to the same degree.

The columns in table 2 correspond to different modifications made to the lexical information: "plain Farsi" is the lexicon obtained directly from Farsi, "+Cogs" includes information from the cognate analysis, "+MA" includes information from the morphological analysis, and "+Cogs MA" includes both types of information. The rows indicate whether the Farsi lexicon was enriched with Pashto vocabulary or not.

Across all trials, both cognate and morphological analysis information improved results, with the cognate information being more useful. The contribution of these lexical modifications has a greater effect in the experiments without the addition of Pashto closed class words where the Pashto-impoverished lexicons are introduced to at least some Pashto. The jump from adding basic closed class lexical information alone is substantial: from 16.91% to 62.65%, using an otherwise plain Farsi lexicon and plain Farsi transitions. The best results of 66.32% are achieved with a combination of closed class, cognate, and morphological analysis information. "Pashtifications" were not tested with this tagset due to poor preliminary performance.

PERT baseline	25.89%
all Pashtifications	37.60%
translate frequent F words to P	51.77%
closed-class words added	61.85%
morphological analysis	68.66%
cognate analysis	**70.84**%

Table 3: Results with PERT

6.2 Experiments with PERT

In order to test our "Pashtifications", we ran experiments using PERT as well. The results are presented in table 3. Each row in the table represents one level of enhancement; with each level, more modifications are used to enhance the tagger's performance. Each level is also built upon the previous level, and thus includes the previous

Modifications	plain Farsi	+Cogs	+MA	+Cogs MA
plain Farsi	16.91%	22.50%	20.15%	24.56%
enriched	62.65%	65.88%	63.82%	**66.32**%

Table 2: Results using Amiri et al. (2007) tagset with different levels of modification to the lexicon.

level's performance boost.

Looking at table 3, we can see that merely changing the tagset to PERT for our baseline gets a performance boost of nearly 10 percentage points (cf. table 2). Seen in row 2, the application of the "Pashtification" rules results in a nearly 12 point accuracy boost to 37.60%. Rows 3-4 show the tagger's performance after successive levels of lexical enrichment. In row 3, the 109 most frequent words in the Farsi corpus have been translated into Pashto, leading to a 10 point increase in accuracy. In row 4, the lexicon is further enhanced through the direct addition of Pashto closed-class words and light verbs to the lexicon which results in a further boost to 61.85%. The addition of morphological analysis in row 5 brings the tagger's accuracy to nearly 68.66%. Finally, the addition of cognate detection takes the accuracy to our maximum accuracy of 70.84%. This accuracy is over 4 points higher than that achieved by the Amiri et al. (2007) tagset experiments (without Pashtifications).

The experiments with the two different tagsets shows that the level of detail captured by the choice in tagset can have a meaningful effect on the results, especially for any syntactic alterations. Indeed, implementing our Pashtification rules required a level of granularity that we were able to provide with PERT. Also, seen in the experiments with either tagset, the inclusion of the closed class elements (lexical enrichment) is key to achieving maximum results.

7 Conclusions and Future Work

We have presented a method of using the POS tagging information from Farsi, a relatively well-resourced language, to help automatically tag Pashto, a relatively lower-resourced language. We used a Hidden Markov Model trigram tagger, TnT, to generate the parameter files which we then modified through various means. Our modifications to the HMM parameter files proved very effective in boosting the tagger's performance on our hand-tagged Pashto test set, with lexical modifications (particularly closed class words) providing the largest boost, and transition modifications contributing substantially with a customized tagset. Ultimately, we improved a 16.62% baseline to 70.84%, which is a respectable number given Pashto's morphological complexity.

In the future, we plan to work on three points to improve this approach. First, we plan to build a better morphological analyzer (MA). Pashto is a morphologically rich language and a robust MA can help tag parts of speech more successfully. Second, we plan to increase the size of our test set to 3000 tokens. Lastly, we will use our automatically tagged test data as additional training and investigate the effect of iterative bootstrapping on the tagger's performance.We can use our 473 MB Pashto web corpus for this purpose.

Acknowledgments

We would thank to thank Sandra Kübler for her guidance and helpful comments, along with the three anonymous reviewers.

References

H. Amiri, H. Hojjat, and F. Oroumchian. 2007. Investigation on a feasible corpus for Persian POS tagging. In *Proceedings of the 12th international CSI computer conference, Iran*.

M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC*.

Thorsten Brants. 2000. TnT–a statistical part-of-speech tagger. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics and the 6th Conference on Applied Natural Language Processing (ANLP/NAACL)*, pages 224–231, Seattle, WA.

A. Feldman and J. Hana. 2010. A resource-light approach to morphosyntactic tagging. In C. Mair, C. F. Meyer, and N. Oostdijk, editors, *Language and Computers 70: Studies in Practical Linguistics*. Rodopi Press, Amsterdam-New York.

A. Hardie. 2003. Developing a model for automated part-of-speech tagging in Urdu. In *Proceedings of Corpus Linguistics*, pages 298–307.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707–710.

F. Oroumchian, S. Tasharofi, H. Amiri, H. Hojjat, and F. Raja. 2006. Creating a feasible corpus for Persian POS tagging. Technical report, University of Wollongong in Dubai. No. TR3/06.

I. Rabbi, A. M. Khan, and R. Ali. 2009. Rule-based part of speech tagging for Pashto language. In *Conference on Language and Technology, Lahore, Pakistan*.

Z. Xiao, A. M. McEnery, Paul Baker, and Andrew Hardie. 2004. Developing Asian language corpora: standards and practice. In *Proceedings of the 4th Workshop on Asian Language Resources, Sanya, China*, pages 1–8.