

Low-density Language Strategies for Persian and Armenian

Karine Megerdooian

The MITRE Corporation, McLean, Virginia, USA

Abstract. This paper presents research on the feasibility and development of methods for the rapid creation of stopgap language technology resources for low-density languages. The focus is on two broad strategies: (i) *related language bootstrapping* can be used to port existing technology from a resource-rich language to its associated lower-density variant; and (ii) clever use of linguistic knowledge can be employed to *scale down* the need for large amount of training or development data. Based on Persian and Armenian languages, the paper illustrates several methods that can be implemented in each instance in the goal of reducing human effort and avoiding the scarce data issue faced by statistical systems.

Keywords. low-resource languages, machine translation, linguistic development, Persian, Armenian

Introduction

Low-density languages, for which few online or computational resources exist¹, raise difficulties for standard natural language processing approaches that depend on machine learning techniques. These systems require large corpora, typically aligned parallel text or annotated documents, in order to train the statistical algorithms. As most of the languages in the world are considered to be low-density [1], there is an urgent need to develop strategies for rapidly creating new resources and retargeting existing technologies to these languages.

Recent methodologies have been developed for using web data to automatically create language corpora, mine linguistic data, or build lexicons and ontologies, while other approaches have focused on creating more efficient and robust techniques for identifying and locating existing web-based data for low-density languages [2]. Researchers have also exploited the application of available resources for developing systems or tools for low-resource languages by eliciting a corpus or language patterns [3,4], by bootstrapping resources for other languages [5,6], or by developing methods that require a smaller set of annotated data ([7,8], among others). This paper argues that different low-density languages require distinct strategies in order to rapidly build computational resources. By studying three specific cases – Tajiki Persian, conversational Iranian Persian found in weblogs and forums, and Eastern Armenian – we illustrate methodologies for cleverly reusing existing resources for these new

¹ The terms *low-density*, *lesser used*, *lesser studied*, and *minority languages* are often used interchangeably in the literature. These terminologies are not necessarily equivalent as certain majority languages commonly used in a society may still lack online resources and technologies (cf. Section 3). The terms *sparse-data*, *resource-poor* or *low-resource languages* are better suited to describe the languages discussed in this paper.

languages. The focus of this paper is on non-probabilistic methods for system development; however, the main argument that a more intimate knowledge of the context and characteristics of each language should be taken into account prior to development is also relevant for statistical approaches.

1. Strategies for low-density languages

As Maxwell and Hughes [1] point out, the obvious solution for dealing with the data acquisition bottleneck for low-density languages is to concentrate on the creation of more annotated resources. This is, however, an extremely time-consuming and labor-intensive task. A complementary approach, therefore, is for the research community to improve the way the information in smaller resources is used. To accomplish this goal, Maxwell and Hughes suggest two possible strategies:

- (i) Scaling down: Develop algorithms or methods that would require less data; and
- (ii) Bootstrapping: Transfer relevant linguistic information from existing tools and resources for resource-rich languages to a lower-density language.

In the case of statistical systems, scaling down could consist of downscaling state of the art algorithms by reducing the training data required for various tasks such as POS tagging, named entity recognition, and parsing. One such approach is active learning, where the annotation is performed and enhanced on samples that will best improve the learning algorithm thus requiring less annotation effort [9]. In addition, bootstrapping approaches have been implemented in cross-language knowledge induction, sometimes using comparable rather than parallel data (see [10] and references therein). In this paper, we introduce novel methods using non-probabilistic techniques and addressing both of these strategies. Bootstrapping is explored for related language pairs, where the existing resources and systems developed for a higher-density language can be used with little effort to build resources for the low-density variant. This approach is combined with the development of linguistic knowledge components that do not require large corpora and are thus especially suitable for low-resource languages. However, the paper advocates proper analysis of the linguistic context prior to actual development and illustrates methods for minimizing the human effort involved by focusing on linguistic properties that will provide the most gain for the new language system.

The paper targets three scenarios: Section 2 focuses on Tajiki Persian, which is a lower density variant of standard Iranian Persian. These languages have developed independently due to historical and political reasons and use distinct writing systems, yet the literary written forms of the two related languages remain almost identical. In Section 3, we look at the effect of *diglossia* in Iran where two distinct and significantly different variants of the language coexist. The “literary” form of Persian has traditionally been used in almost all forms of writing, while the “conversational” variant typically used in oral communication, is nowadays appearing more and more frequently in weblogs and forums. Existing computational systems for Persian have been developed for the literary language and face challenges in processing the conversational variant. Finally, Eastern Armenian is considered in Section 4. The

computational resources for this language are extremely scarce and it is unrelated to other resource-rich languages.

The paper argues that in each instance, a different strategy should be implemented to obtain the most beneficial results. This requires some preliminary analysis of context, language relatedness, and availability of existing resources for the related languages. In the first two instances consisting of Tajiki Persian and conversational Iranian Persian, a form of *related language bootstrapping* can be employed, with an eye on the existing gaps and specific characteristics of the low-density language. In the case where no related language resources can be located as in the case of Eastern Armenian, there is a need to build a system based on linguistic knowledge. In this instance, however, the portability and modularity of the language processing system is crucial as we are now able to reuse components and tools to create and extend existing resources.

2. Tajiki Persian

There exist three distinct main varieties of Persian spoken in Iran (sometimes referred to as *Farsi*), Afghanistan (also known as *Dari*), and *Tajik* spoken in Tajikistan as well as by the substantial Tajik minority within Afghanistan. There is currently a rich set of computational resources for Iranian Persian such as online corpora, parallel text, online lexicons, spellcheckers, morphological analyzers, machine translation engines, speech processing systems, and entity extraction tools. The online resources for Tajiki Persian, however, are extremely scarce and computational systems have not been developed for this lower-density variety of Persian. Iranian Persian and Tajiki Persian have developed independently, resulting in linguistic differences especially in the domains of pronunciation and lexical inventory. In addition, Iranian Persian is written in an extended version of the Arabic script, referred to as the Perso-Arabic writing system, whereas Tajiki Persian uses an extended version of the Cyrillic script. The literary written forms of these two languages, however, are almost identical. It is therefore possible to take advantage of the relatedness of these languages in order to create certain resources and build stopgap systems for Tajiki Persian with very little effort.

This section presents recent work that attempts to build a preliminary Tajik-to-English machine translation system by building a mapping transducer from Tajik in Cyrillic script to its Perso-Arabic equivalent, which is then fed through an existing Iranian Persian MT engine [11]. The mapping correspondences between these two writing systems, however, are nontrivial and the distinct patterns of language contact and development in Tajiki Persian and Iranian Persian give rise to a number of ambiguities that need to be resolved.

2.1. The Writing Systems of Persian

Iranian Persian (henceforth IP) uses an extended version of the Arabic script; it includes, in addition, the letters for پ /p/, گ /g/, ژ /zh/ and چ /ch/. Although Persian has maintained the original orthography of Arabic borrowings, the pronunciation of these words have been adapted to Persian which lacks certain phonemes such as interdental and emphatic alveolars. Hence, the three distinct letters س, ص, and ث are all pronounced /s/. One of the main characteristics of the script is the absence of

capitalization and diacritics (including certain vowels) in most written text, adding to the ambiguity for computational analysis. Further ambiguities arise due to the fact that in online text, certain morphemes can appear either attached to the stem form or separated from it by an intervening space or control character.

Tajiki Persian is based on the Cyrillic alphabet. It also includes several additional characters that represent Persian sounds not existent in Russian. These are $\text{х} = /h/$, $\text{ҷ} = /j/$, $\text{қ} = /q/$, $\text{ғ} = /gh/$, $\text{ӯ} = /ö/$, $\text{ӣ} = /i/$. Tajiki text is much less ambiguous than its corresponding IP script as all the vowels are generally represented in this writing system and capitalization is used for proper names and at the beginning of sentences. The orthography corresponds more directly to the Persian language pronunciation. For instance, the sounds $/s/$ and $/t/$ are represented with the Cyrillic character ‘с’ and ‘т’ respectively, regardless of the original spelling. The divergent pronunciation of the two language variants is also represented in the writing. Hence, the two distinct pronunciations of *shir* ‘milk’ and *sheyr* ‘lion’ in Tajiki Persian are also depicted in the orthography as *шир* and *шеп*, respectively, preserving a distinction previously held in Classical Persian, while in Modern Iranian Persian they are both written and pronounced identically as شیر (*shir*). On the other hand, IP makes a distinction between *pul* ‘money’ and *pol* ‘bridge’, whereas Tajiki Persian pronounces both as пул (*pol*) [12].



Figure 1. Sample Tajiki and Iranian Persian writing systems (source: BBC Persian)

2.2. Issues in Mapping

The correspondence between Tajiki and Iranian Persian scripts is not always trivial. In certain instances, a basic letter correspondence can help achieve a correct map from Tajik into Iranian Persian as shown in Table 1. Consonants typically display a one-to-one correspondence in the two scripts. In addition, the most frequent representation of the $/a/$ sound is the letter ‘o’ in Tajik and the *alef* character ‘ا’ in IP as shown.

Table 1. Direct mapping of Tajiki to Farsi script

китобҳо	کتابها	<i>ketabha</i>	‘books’
коршиносони	کارشناسان	<i>karshenasane</i>	‘experts of’
мардум	مردم	<i>mardom</i>	‘people’
вокунише	واکنشی	<i>vakoneshi</i>	‘a reaction’
корманди давлати	کارمند دولتی	<i>karmande dowlati</i>	‘government worker’

However, ambiguities arise at several levels. For instance, the Iranian Persian writing system includes three distinct letters representing the /s/ sound, four characters corresponding to /z/, two letters for /t/, and two different letters pronounced as /h/, due to the original orthography of the borrowed Arabic words. Hence, a basic mapping to the most common character results in divergences from standard orthography. For instance, the Tajik word *Фурсаат* ‘opportunity’ may be mapped into the Perso-Arabic script as *فرست* (with a *sin* character), as *فرنت* (with a *se*) or as *فرصت* (with a *sat*), but only the latter is the correct Iranian Persian spelling. This word is actually more ambiguous than shown since the /t/ sound, the last character, is itself ambiguous between *te* ‘ت’ or *ta* ‘ط’; thus this Tajiki word has six possible mappings, of which only one is correct.

Another major divergence comes from the distinct representations of the diacritic vowels – /æ/, /e/ and /o/ – in everyday writing. These vowels can be written in many ways in Perso-Arabic script. The Tajiki letter ‘и’, for instance, generally maps to the /e/ diacritic in Persian (also known as *zīr*) which is often not represented in the written form, hence in the word *китоби* only the four letters ‘к’, ‘т’, ‘о’ and ‘б’ will be mapped. However, ‘и’ can also be mapped to ‘ی’ (*ye*) in the IP script as in *фаронсавиҳо* ‘the French’ which is written as *فرانسویها* (*faeransæviha*) in Perso-Arabic.

Certain positional cues, however, can help disambiguate the character in Perso-Arabic script. For instance, the /æ/ sound is typically represented as ‘a’ in Tajik but is not written in Iranian Persian as can be seen in the transliteration of the Perso-Arabic orthography of the first example in Table 2. Yet, it can also appear as an *alef* in Perso-Arabic script if it appears in the beginning of the word as in the second example shown, or as a ‘h’ if it is at the end of the word as illustrated in the third example in the table.

Table 2. Contextual cues in mapping

Tajik	Perso-Arabic	Transliteration	English
пайомадҳои	پیامدهای	<i>pyamdhay</i>	‘consequences of’
анҷуман	انجمن	<i>anjmn</i>	‘organization’
қоғида	قاعده	<i>qaedh</i>	‘regulation’

There are also factors beyond the level of the word. In written IP, if a suffix follows a word ending in the sound /e/ (which is written with the letter *he* ‘ه’), it can never be attached to the preceding word. The suffixes in Tajiki Persian, however, appear attached to the end of the word. Examples are the plural morpheme /ha/ written attached in Tajik (*қоғидаҳо*) and detached in Iranian Persian (*قاعدهها*), or the auxiliary verb /æst/ ‘is’ again represented attached to the verb in Tajik (*шудааст*) and written independently in IP (*شده است*). Even more problematic is the fact that a number of compound nouns are written as a single unit in Tajiki Persian while their subparts remain detached in the Perso-Arabic script in IP. For instance, the compound noun *riyasæt-jomhuri* ‘the presidency’ (literally “the directorship of the republic”) is

represented in Tajik as *раёсатҷумҳурии*, whereas it consists of two independent words in IP: *ریاست جمهوری*. Furthermore, Iranian and Tajiki Persian have differing patterns of contact, which in turn leads to different patterns of borrowed words. The choice of orthography makes a difference, as well: whereas Western terms borrowed into Iranian Persian must be reformulated in Perso-Arabic, the use of Cyrillic in Tajik allows for Russian terms (as well as other languages in contact from former Soviet republics, such as Uzbek) to be preserved in the original orthography instead of adapting to the Tajiki pattern. For instance, the month October in Iranian Persian is a borrowing from French and is represented as *اکتبر* /oktoʔbr/ while it is written as *октябрь* in Tajiki Persian.

Further ambiguities arise if the input source does not take advantage of the extended Tajiki script. For instance, BBC Persian documents written in Tajiki Persian use the same character ‘r’ to represent both /g/ and /gh/ (the latter is written as *ғ* in the extended Tajiki script). The unavailability of the full extended script inevitably gives rise to further ambiguities in mapping.

The issues discussed in this section suggest the need for an intelligent mapping algorithm and strategies for disambiguating the obtained results. In addition, a morphological analyzer component is needed to handle the segmentation issues presented.

2.3. System description

Based on the abovementioned descriptive study of the correspondences between the Tajiki Persian and Iranian Persian writing systems, a proof-of-concept Tajik system can be developed based on existing IP tools in order to serve as a stopgap measure until language-specific resources can be built. To begin with, an extensive finite-state transducer (FST) is written that converts Tajik text to Perso-Arabic script. The point of such an FST is to overgenerate, since as described above, many segments may represent several potential spellings in the target script. The potential combinatorial explosion is controlled using contextual rules that help to disambiguate the output, exemplified in Figure 2, as well as available Iranian Persian resources (lexicon, morphological analyzer).

```
# Add alef under diacritic at the beginning of the word
define initialA [(Aa) <- a || .#._];

# Represent the /a/ sound at the end of the word (marked
# by WD tag) as ‘he’
define silentH [h <- a %^WD];
```

Figure 2. Contextual rules for mapping Tajiki ‘a’

Where there is still ambiguity between forms even after a lookup, a variety of disambiguation strategies such as statistical language modeling using Iranian Persian corpora could also be employed. Lexical divergences such as borrowings from Russian will need to be handled in a pre-processing step of looking up Cyrillic terms in a special lexicon with their corresponding Persian terms. As these terms are merged in with the FST/lookup table output, the results of the transformation improve. Finally, the output is run through a commercially available Persian to English machine translation engine. The final end-to-end system thus results in a rapidly developed Tajik to English

MT system without the benefit of Tajik-language resources. The rest of this section provides further details on the various components in the system.

The transliteration is developed using Xerox Finite-State Technology [13]. A basic grammar is written allowing any combinations of Tajiki characters to form a word. The grammar is compiled into a finite-state transducer (FST) where the lower side consists of the input string and the upper side provides the transliterated form of the word. A number of contextual rules are composed on the FST, as exemplified in Figure 2, thus performing the required orthographic and phonological alternations on the word forms based on the position of the character within the word. If contextual cues are unable to produce a single mapped output, the transducer creates all possible results for each input token, which is then disambiguated at the next stage in the process. For each input token, the resulting transliterated Iranian Persian words undergo morphological analysis and lexicon look-up to determine possible lexical items [14]. If an analysis is found, then the form is used. If there is no analysis, the word is matched against an unstemmed wordlist culled from various Persian corpora. If still no match is located, a number of “rules of thumb” are employed to select a likely alternative based on letter frequencies. Figure 3 shows the results of disambiguation when the morphological analyzer/lexicon combination works successfully.

1 alternatives (12 originally)	سخنگوی	sxngv+Noun+sg+ez	[speaker;spokesman;]
1 alternatives (1 originally)	بانک	bank+Noun+sg	[bank;]
1 alternatives (32 originally)	تاجیکستان	taJykstan+PropN	[Tajikistan;]
1 alternatives (12 originally)	پرداخته است	prdaxtn+Verb+ind+perf.past+3sg	[pay;attend;]

Figure 3. Disambiguated analyses

2.4. Evaluation

The results of the preliminary evaluation show that the current system is able to achieve 89.8% accuracy for an input corpus that uses the extended version of the Tajiki script. The transliterated document can then be used with the Language Weaver Persian-to-English MT system to create translations of the original Tajiki text.

Our current test corpus consists of approximately 500,000 words from articles taken from Radio Ozodi (the Tajik broadcast of Radio Free Europe). As a beginning testbed, this seemed ideal, since the domain largely matches the training corpora of commercial Persian MT systems such as Language Weaver Persian, and unlike several other sources, the full range of Tajik diacritics are used; later refinements will have to take into account defective orthography used by many electronic sources. At this early stage, a small test set of 6,156 tokens was run through the morphological analyzer and lexical lookup and evaluated against a golden truth corpus. The results show that the current system is able to achieve 89.8% accuracy in transliterating a document in Tajiki

script to its Iranian Persian equivalent. In other words, in the case of 89.8% of input tokens, there was at least one correct transliterated form which was used as input to the MT component. The average token returned with 6.27 alternative spellings. Further analysis on the larger corpus is needed to determine the accurate level of precision and recall for various input documents.²

2.5. Summary

This section presented a methodology for the rapid creation of language technology resources for Tajiki Persian by taking advantage of existing resources and systems developed for the higher-density variety of Iranian Persian. It is expected that in the long term, stopgap systems like the one proposed here will be replaced with fully-developed MT based on the cultivating of resources, parallel corpora, rule development, and so forth. In the meantime, a comprehensive finite-state transducer is developed based on a preliminary study of the similarities and differences of correspondences in the two writing systems which, combined with simple scripts to integrate the results with existing Iranian Persian resources, provides a first draft Tajik to English machine translation system. The results documented are still at the preliminary stages; nevertheless, the approach has been proven effective for rapidly building translation capabilities for a language with scarce resources, in case a related higher-density language with a distinct writing system is available. The transliteration transducer requires very little human effort and a very small corpus is needed for testing purposes. In addition, it is hypothesized that this methodology can be used across a variety of unevenly dense languages with distinct scripts, such as Hindustani (Hindi, Urdu), the Turkic languages (Turkish, Azeri, Uzbek, Uighur), and Kurdish (Kurmanji and Sorani).

3. Persian Weblogs

Since its beginnings in 2001, the Persian blogosphere has undergone a dramatic growth making Persian one of the top ten languages of the global blog community in 2007 [15]. The Persian blogosphere³ has opened the door to journalists, intellectuals, and University students who use blogs to evade government censorship or social and political restrictions, as well as conservative individuals who discuss various religious or political topics online [16]. This new medium has also provided a forum for bloggers to express their opinions and thoughts in their everyday speech rather than the traditional literary language. This creates a new challenge for the analysis of Persian language websites as current grammars and academic textbooks of Persian focus mainly on the literary dialect and existing text-based computational systems often fail to analyze or process conversational Persian.

3.1. Language of blogs

The *diglossic* situation of Persian, whereby two distinct varieties of the language coexist in the society, is also reflected in the language found in the Iranian blog

² For a more detailed discussion of results, see [11].

³ The focus of this paper is on weblogs in Iran and among the Iranian expatriate community and the proliferation of Persian language blogs in Afghanistan is not studied.

community. Traditionally, Persian literature and news media have been written in the literary dialect which holds a higher prestige over the conversational form of the language. Although the latter has been used in some works of modern literature, its usage is generally limited to the informal, conversational domains and is rarely seen in written form. With the advent of blogs, the restrictions against the use of the conversational dialect in writing have been challenged and, despite strong criticisms from intellectuals and professional journalists, bloggers often use the conversational Persian variant in their posts.

Preliminary exploration of the language of Persian blogs shows parallels with English Blogspeak. As noted by Crystal [17] for English, the content of a site (e.g., information, political opinion, education, personal diary) strongly influences the general character of the language being used leading to linguistic variation on the Internet. This observation holds for the Persian language websites as well. Hence in both English and Persian, the language of blogs that address personal thoughts, opinions, and issues has been characterized as a conversational style in writing. Non-standard spelling that reflects the colloquial pronunciation of words is often used. Blog entries are usually written in short sentences and include a large number of hyperlinks. Deviant spelling is common and standard orthography is often ignored, opting instead for a more intimate style. Emotions are expressed with emoticons, ellipsis, repetition of letters and punctuation marks, and emphasis is shown with capitals and special symbols. Jargons and neologisms abound in Blogspeak, especially based on technical or computer-related terms.

Persian Blogspeak differs from that of English, however, due to its strong diglossic situation. While syntactic or grammatical variation is less frequent in English, the distinction between the literary and conversational language is especially poignant in Persian, affecting morphology and syntax as well. Persian Blogspeak often includes properties corresponding to the conversational language such as shortened verbal stems, frequent use of attached pronoun forms, and affixes that are not part of the standard formal grammar. There are more instances of free word order, idiomatic expressions, loan words, and an inordinate amount of orthographic variance partly due to the flexibility and ambiguity of the Perso-Arabic script. This section presents an overview of some of these characteristics and the ambiguities and challenges they raise for computational processing.

Persian morphology is affixal, consisting mainly of suffixes and prefixes, which generally follow a regular morphotactic order. Ambiguities arise in a computational analysis due to the use of the Arabic script since certain vowels are not marked in written text and spacing between words and morphemes is sometimes inconsistent. Furthermore, some affixes can represent different morphemes. For instance, the suffix *-i* as in مردی (*mærdi*) can be an indefinite article (a man), a relativizing particle (the man (that)), the second person singular form of the copula verb 'to be' (you are a man), or a derivational form creating adjectives out of nouns (manhood, manliness)⁴. In addition, the lack of capitalization and short vowels can add to the ambiguity since the word can also be analyzed as the Nigerian province 'Maradi' or the verbal form *mordi* 'you died'.

Conversational forms of morphemes give rise to further ambiguity. For instance, *zanha* 'women' would be pronounced as *zana* in the conversational variant. Since the /æ/ vowel is not usually written in Persian script, it would have the form زنا [zna] in a

⁴ There is less ambiguity in speech since the stress pattern can distinguish some of the constructions.

text. Without the overt vowel in the first syllable, however, the word is now ambiguous between *zāna* ‘women’ and *zēna* ‘adultery’. Another instance of ambiguity arising from the conversational orthographic form can be found with words that end in the sound /e/ which is written as a ‘h’. In conversational speech, the word-final /e/ assimilates (or merges) with the following /æ/ sound of some affixes. For example, the word *khune* ‘house’ when used with the pronominal suffix *-æm* becomes *khunæm* in the spoken language, meaning ‘my house’. It is often represented as in خونم [xvnm] in the written form on blogs, which becomes ambiguous with the word *khunæm* meaning ‘my blood’ written exactly the same way. There is no ambiguity in the spoken language since the stress pattern of the words distinguish the two constructions: *khun.Em* ‘my house’ vs. *khUnæm* ‘my blood’; yet the orthographic forms remain ambiguous.

The use of conversational language in writing introduces a number of affixes that can never be found in traditional literary text, such as the definite suffix ‘e’ as in *ketabe* ‘the book’ or *forushændehe* ‘the salesperson’. The area that has undergone the most change is probably the verbal domain, where not only the inflectional endings are modified but many of the verbal stems are also shortened. Hence, the literary form *miguyænd* ‘they say’ has become *migæn*, or *miændazæd* ‘he/she throws’ is pronounced *mindaze*, making it impossible for a system trained on literary text to analyze the conversational forms. In addition, conversational language makes more use of affixes where the literary language would use separate tokens. This is illustrated in the examples in Table 3 contrasting literary and conversational forms for the same sentences, with contrasting elements shown in bold.

Table 3. Conversational and literary corresponding forms in Persian

Literary form	Conversational form
استرس امتحان مرا گرفته است (<i>estrese emtehan mæra gerefte æst</i>) stress-of exam me caught is ‘The stress of the exam has got me’	استرس امتحان گرفته تم (<i>estrese emtehan gereftætæm</i>) stress-of exam caught-3sg-me ‘The stress of the exam has got me’
گرگ او را می خورد (<i>gorg u ra mikhoraed</i>) wolf him Obj is-eating ‘The wolf is eating him’	گرگه میخورتش (<i>gorge mikhoraetæsh</i>) wolf-def is-eating-3sg-him ‘The wolf is eating him’

Since the writing in blogs more directly reflects the way people speak, changes in the pronunciation of Persian are represented in blog text as well. Examples include:

- the alternation of /an/ to /un/ in words like *nan* ‘bread’ → *nun*, *zendani* ‘prisoner’ → *zenduni*, or *tehran* ‘Tehran’ → *te:run*⁵
- the assimilation of /n/ to /m/ before /b/ as in *shænbe* ‘Saturday’ → *shæmbe* or *tænbael* ‘lazy’ → *tæmbael*
- the elimination of /t/ in /st/ clusters in words like *bæstæni* ‘ice cream’ → *bæssæni*, or *kojast* ‘where is it?’ → *kojas*

Blogs contain a large number of loanwords, especially from English which currently exerts a big influence on Persian because of computers and technology. Scientific and technological terms are widely used on blogs. This is in particular

⁵ /:/ represents a lengthened vowel.

intensified when the Iranian government tries to crack down on the blogs, and bloggers begin posting ways to break the filtering technologies and provide technical support to each other online. These words of course follow the morphological rules of Persian and can take affixes as in *filteringeshun* (فیلترینگشون) = *filtering* ‘filtering’ + *eshun* ‘their’ (‘their [attempts at/act of] filtering’). Examples of technical terms borrowed from English are: *anlayn* (آن لاین), *pabliš* (پابلیش), *chatrum* (چتروم), *imeyl* (ایمیل), *monitowr* (مونیتور), *es-em-es* (اس ام اس), *di-vi-di* (دی وی دی), *vindoż* (ویندوز), *afis* (آفیس), *fotoshap* (فتوشاپ), *kibord* (کیبورد). Other English and older French loans are also quite common: *pazel* (پازل) ‘puzzle’, *partner* (پارتنر) ‘partner’, *holokast* (هولوکاست) ‘holocaust’, *nostalzhi* (نوستالژی) ‘nostalgia’, *seksualite* (سکسوالیته) ‘sexuality’.

One of the more striking aspects of Blogspeak, however, is the amount of neologisms or new words created by bloggers, typically using a loan word as the base combined with Persian language word-formation patterns. They include frequent words such as *linkduni* (لینکدونی) which literally means a storage place for links and corresponds to the English term ‘blogroll’, *tabusazi* (تابوسازی) meaning the act of making something taboo, or *filtershekæn* (فیلترشکن) literally meaning filter-breaker and referring to anti-filter software. Similarly, a number of new verbs have been formed by combining a loanword with a light verb such as *kærdæn* ‘to do’ or *zædæn* ‘to hit’⁶. Examples of these new constructions are *chat kærdæn* (چت کردن) ‘to chat’, *hæk kærdæn* (هک کردن) ‘to hack’, and *imeyl zædæn* (ایمیل زدن) ‘to email’. More recently, verbs are formed on the simple verb construction pattern instead of the compound forms above by adding the *-idæn* infinitival morpheme, forming the verbs *klikidæn* (کلیکیدن) ‘to click’, *danlodidæn* (دانلودیدن) ‘to download’, and *lagidæn* (لاگیدن) ‘to blog’.

Misspellings are very common in weblogs, but sometimes they are done on purpose as for the many forms of spelling the word *seks* (سکس) ‘sex’ in order to be able to discuss a very taboo subject in Iranian society among the youth without being subject to filtering by the government. Bloggers write this word with the various characters representing the /s/ sound in Persian: *سکص*, *سکث*, *سکس*. In addition, the spelling of some words of Arabic origin are being modified by bloggers to represent their current pronunciation in Persian. Examples are the words that end in a “tanvin” as in *lutfæn* [lTfn] ‘please’ which is now sometimes spelled with a /n/ as in *lutfæn* [lTfn] to reflect the actual pronunciation *lotfen*. Similarly, *mosy* [mvsy] ‘Moses’ and *hty* [Hty] ‘even’ are now often spelled as *mosa* [mvsa] and *hta* [Hta] to represent their pronunciations as *musa* and *haeta*, respectively.

As these examples show, the variant forms introduced by writing the conversational form of Persian add to the complexity and ambiguity of computational processing. As discussed earlier, there are also syntactic distinctions between the literary and conversational variants of Persian as conversational text contains more instances of scrambling (permutations of word order), topicalization, idiomatic expressions, and cultural inferences. This diversity is accentuated by variants of orthographic forms found online as each social group has defined its own standards or writing approaches: (i) the traditional orthography taught at school and recommended by the Persian Language Academy has strict rules of spelling and spacing; (ii) journalist and intellectual bloggers have recently proposed their own guidelines for

⁶ Light verb constructions are very pervasive in Persian and consist of a preverbal element (noun, adjective or preposition) followed by a verb that is somewhat bleached in meaning called a “light verb”.

Persian orthography that differ from the traditional rules; and (iii) bloggers using conversational language, mostly the youth, write the words as they are pronounced in spoken Persian and do not have any set standards.

3.2. Computational Processing of Blogspeak

As most existing computational systems of Persian have been developed for formal writing usually found in news reports, we expect that they would fail to successfully process the elements found in conversational blog text. This section describes the results of an evaluation performed on an existing morphological analyzer to confirm that the presence of conversational forms strongly hinders computational processing of Persian text. It is reasonable, however, to take advantage of these existing resources for building tools to process Persian Blogspeak which includes both literary and conversational variants. Although syntactic structures are also affected, the fundamental distinction between modern conversational Persian and literary Persian lies at the word level, ranging from choice of lexical items and inflection of word forms to orthographic variance. This suggests that it may be advantageous to extend an existing morphological analyzer for Persian to cover the conversational forms of the language.

Table 4. Ground truth files: Percent of conversation forms in each post

Topic	Total Entries	Conversational Entries in Text						% of Conv. in text
		Verb	Non-Verbal	Closed Class	Loans	Foreign Words	Interj	
News	271	0	0	0	0	0	0	0.00%
News	141	0	0	0	0	0	0	0.00%
Politics	219	8	0	3	1	0	0	5.5%
Politics/Book	467	45	13	20	3	0	2	17.8%
Politics	340	31	19	20	3	0	2	22.1%
Journal	391	33	16	34	0	0	5	22.5%
Journal	725	55	51	49	13	1	4	23.9%
Tech/Book	326	25	14	15	20	5	2	24.9%
Journal	16	0	0	4	0	0	0	25.0%
Tech/ Blogs	147	20	5	18	8	1	0	35.4%
Tech/ Blogs	150	14	8	20	11	3	0	37.3%
TOTAL NUMBERS:	3193	231	126	183	59	10	15	19.5%

A morphological analysis tool that can process conversational text found in blogs and offer the literary Persian equivalent will be able to associate blog vocabulary with dictionary forms which could ultimately benefit applications from Part-of-Speech (POS) tagging to machine translation and entity extraction. Such a task, however, may require a lot of human effort given the large amount of variance in conversational text. It is therefore beneficial to first perform a study to determine the most efficient way of extending the morphological system, allowing us to cleverly bootstrap existing resources. This section describes the results of such a study.

For the purpose of the evaluation, we selected a unification-based Persian morphological analyzer that was developed for processing literary Persian text [14]. The study was run on a small sample of blog posts collected from popular Persian blog

sites⁷ containing various topics ranging from personal journals to discussions of societal issues, to technical and computer-related subjects. Each post was manually annotated for Part of Speech information in order to develop a ground truth totaling 3,193 entries (including compounds). The number of entries in each document displaying conversational text characteristics was computed and can be seen in Table 4. As can be seen from the table, the two news-related posts did not contain any conversational forms at all. Some conversational morphology was found in the files related to politics, but the posts categorized as journals have a higher rate of conversational morphology and lexical items. Discussions of technical issues (often related to filtering and ways to avoid them) contain a large number of loans and foreign words. The results of the evaluation were judged based on accuracy and ambiguity level and are illustrated in Figure 4.

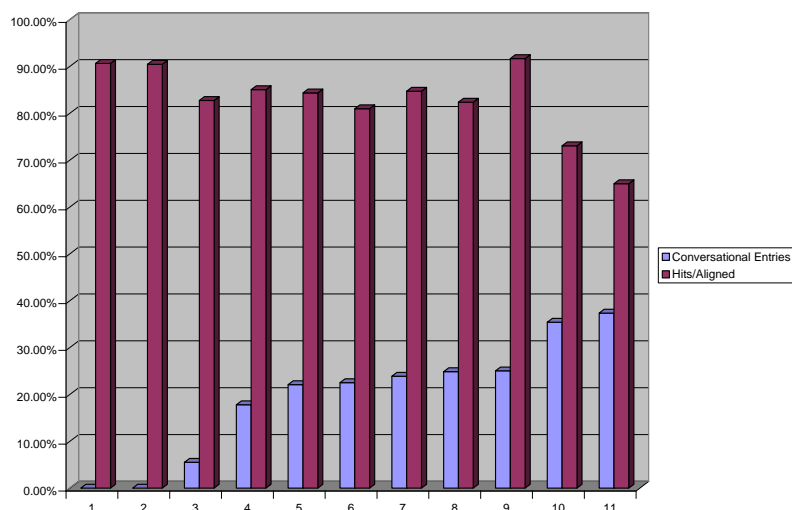


Figure 4. Correlation of conversational entry and accuracy of POS tag per correctly aligned entry

For each post, the system results were compared to the ground truth files by taking into account the number of correct alignments of the word entries, the number of correct POS tags, and the ambiguity the system generated in each instance – where ambiguity is defined as the number of POS tags in the system output, divided by the total number of entries in the output. The results shown in Figure 4 indicate a clear correlation between the percent of conversational entries in a blog post vs. the hits per correctly aligned entries in the document. The figure illustrates that the morphological analyzer tends to get better scores for the news items that have very low or null conversational entries. The lowest accuracy scores were obtained for the last two files that have the highest number of conversational forms. The only obvious exception to this pattern is file #9, which had a relatively high number of conversational forms but showed high accuracy scores; this file was a very short blog post consisting of only two sentences (16 entries). Figure 5 also shows that, as the level of conversational forms in the text increases, the ambiguity also seems to increase. This indicates that the morphological system encounters more unknowns in the text and thus generates more

⁷ The sites were www.khorshidkhanoom.com, www.z8un.blogfa.com, and www.4shanbe.blogfa.com.

guesses as to their POS tag. The correlations noted in the results thus suggest a direct correspondence between the number of conversational forms in a text and the difficulty of analysis for the system.

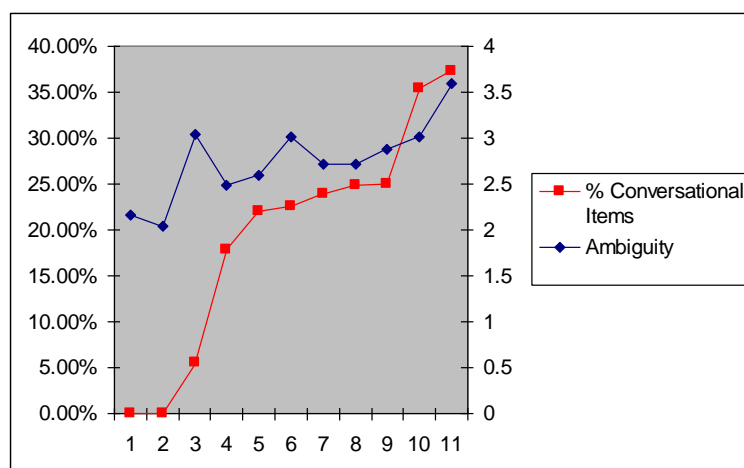


Figure 5. Correlation of conversational entries and ambiguity per system analysis

The total accuracy score obtained for all blog posts was 84%. A close examination of the results showed that a large majority of the correctly tagged entries (97%) were of literary form, while only 3% were conversational entries which were guessed correctly by the system. From the mistagged entries, on the other hand, a majority of 78% were conversational entries, as illustrated in Table 5.

Table 5. Breakdown of accuracy results

	Literary	Conversational
Correctly Tagged	97%	3%
Incorrectly Tagged	22%	78%

The results suggest that the presence of conversational forms in text does play a significant role in morphological analysis for tools developed primarily based on literary Persian. The statistical results suggest a direct correlation between the number of conversational forms and reduced performance. In addition, a closer examination shows that the majority of mistagged elements are of conversational form. A system that provides guesses based on word forms can provide better results although the ambiguity is increased as more (unknown) conversational forms are encountered in text. Based on these results, one can conclude that the presence of conversational forms negatively impacts the output of the morphological analyzer. The study also suggests that starting from an existing morphological analyzer for Persian, an analyzer can be developed to recognize the various conversational forms encountered in blog text. A closer examination, however, shows that in order for the extended system to be efficient, the additions can be performed in stages beginning with modifications that bear the most result. Table 6 provides a breakdown of the literary and conversational system tags for a small representative sample based on Part-of-Speech category. Note that these represent the number of entries tagged with a particular POS in the ground truth; hence we have counted each entry whether it was previously encountered or not

(e.g., the postposition mistagged 7 times was the same one, namely the object marker *ro* in its conversational form).

Table 6. System tags analyzed per POS category

POS	Literary Matched	Literary Mismatched	Conversational Guessed	Conversational Mismatched
Noun	100	6	5	1
Verb	35	0	1	26
Adjective	22	1	0	3
Adverb	27	1	1	1
Proper Name	9	5	0	0
Pronoun	9	0	0	3
Preposition	42	0	1	3
Postposition	0	0	0	7
Conjunction	43	1	0	0
Determiner	11	0	0	1
Numeral	7	0	0	1
Quantifier	5	0	0	1
Relativizer	13	0	0	0
Question Word	4	0	0	0
Interjection	0	0	0	2
Loan Word (Noun)	0	0	3	0

Table 6 shows that the POS category missed most often is the verb in conversational form, which indicates that adding verbal conjugation rules for conversational language would provide the most value. On the other hand, nouns are typically guessed correctly by the system based on their inflectional morphology. The table also shows that the addition of frequent conversational items, such as the postposition *ro* and certain closed class items, could significantly improve the analysis results for blog text containing conversational language. Such a study helps us develop a more efficient strategy for extending the existing system as it is clear that we can improve the system stage by stage as some changes provide more value than others.

3.3. Summary

This section described a second instance of related languages, in this case the literary and conversational forms of the same language, and discussed a strategy for bootstrapping existing resources to improve processing of the new low-density variety. It was argued that a preliminary analysis of the distinctions and similarities between the language variants and a test study to identify a hierarchy of challenges raised by the low-density variant can be extremely beneficial in developing a bootstrapping strategy.

A descriptive study of conversational Persian found in weblogs identified a number of important differences at the word level in particular – such as the creation of new words, an extended set of borrowings especially in the technical domain, and orthographic representations that more directly reflect the pronunciation of words – which affect both lexical elements and morphological affixes. Thus, a focus on the morphological component would arguably provide the best results in capturing the new

conversational forms encountered in text. As no previous research had been carried out to show whether conversational variants do in fact hinder morphological analysis, an evaluation was performed on a small corpus of blog text. A closer study using an existing system, however, helped identify specific gaps showing that the addition of some knowledge (e.g., verbal paradigm, frequent items, certain lexical categories) provides more value than others. This allows the extension of existing rules or lexicons to be performed in stages to provide rapid improvements with least human effort.

With the advent of social media such as blogs, forums and chatrooms, more and more people tend to use their spoken language in writing, which may show enormous differences with the traditional written form. We expect that the strategy proposed for Persian blog text can also be applied successfully to a large number of languages that display strong forms of diglossia as in the case of Arabic languages or the many languages in India.

4. Eastern Armenian

The third case scenario is Eastern Armenian, an Indo-European language with a severe lack of existing computational systems or tools. In addition, Eastern Armenian is written in the Armenian alphabet and therefore does not share the writing system of existing higher density languages. There are no computational grammars developed for this language and traditional grammars, as is often the case, are very prescriptive and incomplete. Several computational resources (e.g., a morphological analyzer and a corpus) are currently being developed for Western Armenian, a related language albeit with significant linguistic differences. It may therefore be possible to take advantage of existing Western Armenian tools to bootstrap these computational components for Eastern Armenian in the near future. There exists no syntactic component, as far as we are aware, for either variant of Armenian. Hence, in this paper we treat Eastern Armenian (henceforth EA) as an instance of a low-density language that is not in any way related to a resource-rich language with respect to parsing components. The main reusable elements in this case become the existing tools such as the segmenter, morphological analysis system, lexicon look-up tool, parser system, etc. This shows the absolute importance of modular and language-independent tools for development. In addition, the section presents a methodology for linguistic development for partial parsing.

4.1. Linguistic development

A common sentiment in natural language processing is that the development of knowledge-base systems is labor-intensive and time-consuming: “Statistical NLP models have a distinct advantage over rule based approaches to [rapidly retarget existing technologies to new languages], as they require far less manual labor” [19]. With very few exceptions, these claims are not substantiated by empirical evidence (cf. [20] but also [21,22]).

It is indeed true that, given pre-annotated or pre-aligned corpora, probabilistic technologies can be developed much more rapidly than a knowledge-base system. However, the creation of corpora that can be used to train statistical systems are extremely labor-intensive as illustrated in the following paragraph describing the development of a Cebuano corpus [23, p. 82]:

The production of “parallel text” tells that story well. The University of Maryland produced nearly a million words of verse-aligned parallel text the first day, by the simple expedient of obtaining a Cebuano bible and aligning the verse numbers with those in an English bible that was already at hand. The University of Southern California Information Sciences Institute (USC-ISI) hired native speakers of Cebuano to produce translations, producing several thousand words within a week. But it was not until the team at Carnegie Mellon University found the newsletter of the Philippine Communist Party on the Web in both Cebuano and English that a large amount of truly representative example translations became available.

In particular in the case of low-density languages, for which very few resources exist, the development of parallel text and annotated corpora from scratch is rather a daunting task. On the other hand, we argue in this section that for resource-poor languages, development of knowledge-base components by a trained linguist, such as shallow parsers, with emphasis on providing generalizations for the most important structures in the language is highly valuable and can result in a first draft system with little effort. This section offers a blueprint for development of a partial parser without the need for a large set of hand-constructed rules.

Partial parsing refers to techniques used for recovering syntactic fragments instead of providing all of the information contained in a traditional syntactic analysis. As described in Abney [24], “The idea is to factor the parse into those pieces of structure that can be reliably recovered with a small amount of syntactic information, as opposed to those pieces of structure that require much larger quantities of information, such as lexical association information.” A partial parser typically recognizes the key elements of a clause such as clausal boundary markers and simplex (i.e., non-recursive) clauses such as noun phrases (NP), preposition or postposition phrases (PP), and possibly simple verb phrases. Although partial parsers may also detect subjects and predicates, deeper analysis such as attachment resolution is not included at this stage of processing.

Partial parsing is especially useful for languages that display free word order, i.e., the clausal constituents do not always appear in a strict relative order in the sentence. In addition, the results of partial parsing can be used for bootstrapping – extracting information from a partially parsed corpus for use by more sophisticated parsers – or they could be used in applications such as entity extraction. This technique is therefore sometimes used as a preprocessing step. In what follows, we will focus on some important aspects of partial parsing development that are common in many low-density languages in order to build a first phase syntactic parser.

4.1.1. Grammar Books

When faced with a new language, clearly the first place to begin is with reference and grammar books. Unfortunately, in most instances, traditional grammars fail to provide all the relevant information needed for computational purposes. These grammar books oftentimes tend to define prescriptive rules (sometimes based on an older, literary version of the language) rather than providing descriptions of the modern language. The paradigms introduced are often incomplete for system development (e.g., not all lexical elements following a certain paradigm would be listed or the grammar book may not list certain irregular forms or exceptions). In addition, many grammar books fall short of making the correct generalizations and in most cases, grammarians do not explore distinctions between the spoken and written forms of the language under study.

More importantly, grammar books that have not been developed for text analysis fail to discuss orthographic issues and variances which are crucial for computational analysis. It is therefore important to locate speakers of the language and a set of relevant corpora combined with keyword searches online to complement grammatical descriptions in reference books.

4.1.2. Phrasal boundaries

A first step in partial parsing is to reliably detect phrasal boundaries in text. One general method is to use function words to delimit clauses. In addition, a number of lexical items or morphological elements tend to appear at the beginning or end boundaries of certain phrases. For instance, pronouns in Persian within a noun phrase always occupy the last position at the NP boundary. In addition, certain affixes or function words are used to link the phrasal constituents in languages. These elements can easily be identified following a study of the basic NP and PP structures in the low-density language. This section presents a contrastive examination of the linking element in Tajiki Persian, Iranian Persian, and Eastern Armenian.

One of the important distinctions between the Tajiki and Iranian Persian writing systems involves the recognition of phrasal boundaries. Boundary recognition is a significant problem in Iranian Persian which uses the Perso-Arabic script, as there is no capitalization and the main morpheme linking the elements of a noun phrase is pronounced as /e/ which, being a diacritic, is typically not represented in orthography. As expected, this gives rise to very ambiguous results in applications such as MT and entity recognition which involve some level of phrasal parsing. In Tajiki Persian, however, the linking morpheme is represented in text, clearly indicating phrasal boundaries in a sentence. This distinction is illustrated in Ex. (1).

- (1) نشست سران کشورهای ساحلی خزر شروع شد
 ‘The session of the heads of the coastal countries of the Caspian (Sea) began’

The nominal elements in the sentence are linked to each other with the so-called “ezafe” morpheme, which is pronounced as /e/ after consonants and /ye/ after vowels as shown in the transcribed version in Ex. 18.

- (2) *neshæst-e særan-e keshværho-ye saheli-e xæzær*
 session-**ez** heads-**ez** countries-**ez** coastal-**ez** Caspian

When a word in the noun phrase does not carry this affix, it marks the phrasal boundary. Hence, in this example, *xæzær* ‘Caspian’ is the end of the NP as shown in the parsed version in (3). However, the /e/ morpheme is typically not written in text, resulting in parsing ambiguity as any of the nouns may present a potential NP boundary for the system.

- (3) [نشست سران کشورهای ساحلی خزر] [شروع شد]
 [_{NP} session-**ez** heads-**ez** countries-**ez** coastal-**ez** Caspian][_{VP} beginning became]

Tajiki Persian orthography, on the other hand, explicitly writes the “ezafe” morpheme (ن). Ex. (4) illustrated this for the same sentence, clearly demarcating the phrasal boundary.

- (4) Нишастии сарони кишварҳои соҳили Хазар шуруъ шуд
session-ez heads-ez countries-ez coastal-ez Caspian beginning become

Hence, Tajiki Persian documents provide information on capitalization and boundary recognition which is not available to systems dealing with Iranian Persian text. In the case of Eastern Armenian, a language with very rich inflectional morphology, the linking elements within the NP constituents are also clearly demarcated by use of the genitive case morpheme as shown in Ex. (5).

- (5) Եվրոպական երկրների հանդիպումը տեղի ունեցավ
 European-gen countries-gen meeting-nom place had
 ‘The meeting of the European countries took place.’

4.1.3. NP and PP structure

In order to determine the basic NP structure for partial parsing purposes, we would need to establish where the various constituents of the noun phrase appear relative to each other. Noun phrases, even in languages displaying relatively free word order, are quite rigid in structure. For instance, NPs in Persian (both the Tajiki and Iranian variants) can be described as in the schema shown in Figure 6 where each element except for the head noun is optional. This schema can represent the sentence ‘*in do ta ketab-e kheyli kohne-ye to*’ [lit. this two (unit) book-of very old-of you] which can be translated into English as “these two very old books of yours”.

Determiner	Number (Classifier) OrdinalNum SuperlativeAdj Quantifier	Noun	[(Adverb) Adjectives]	Possessor
------------	---	------	------------------------	-----------

Figure 6. NP structure for Persian

The relative ordering of the elements within a noun phrase can be determined by any trained linguist by identifying the elements that appear in complementary distribution. Note that these are basic structures that cover most syntactic constructions encountered in a corpus and do not necessarily cover more complex realizations of the noun phrase which can be added at a later stage. The schema in Figure 7 shows the basic NP structure in Eastern Armenian. This schema can be used to parse the sentence in Ex. (6).

Determiner Possessor	Number (Classifier) OrdinalNum Quantifier	[(Adverb) Adjectives]	Noun
-------------------------	---	------------------------	------

Figure 7. NP structure for Eastern Armenian

- (6) մեր երկու լավագույն աշակերտները
 our two excellent students-nom
 ‘our two excellent students’

Languages that have prepositions or postpositions generally tend to form PPs by combining the preposition or postposition with a noun phrase. This is indeed the case

for Persian where preposition phrases are simply formed as P + NP structures. Eastern Armenian, on the other hand, has a mixed system consisting of prepositions, postpositions and case markers. These are exemplified in Table 7.

Table 7. Eastern Armenian preposition/postposition phrases and oblique cases

Preposition	դեպի համալսարան	toward university	'towards the University'
Postposition	սեղանի վրայ	table-gen on	'on the table'
Locative case	համալսարանում	university-loc	'in the University'
Ablative case	համալսարանից	university-abl	'from the University'
Instrumental case	դանակով	knife-inst	'with the knife'

4.1.4. Word order

In addition to the elements and structures described, preliminary word order analysis is required to be able to develop simple rules (e.g., regular expression rules) for recognizing simple sentences containing a subject, an object, and a verbal element. For instance, both Persian and Armenian are verb-final languages and follow the word order 'subject-object-verb' for simple sentences, although Eastern Armenian writing style typically follows the 'subject-verb-object' order. However, it is easier to identify the arguments and their functions within an Armenian text as the latter generally uses distinct affixation (e.g., case marking) to distinguish subjects, objects, and oblique nouns. Additional items, such as basic adverbials and indirect objects may also be included for the purposes of partial parsing.

4.2. Language Technology Characteristics

In order for language technology to be valuable for low-density languages, it needs to be designed following the basic principles of reusability and portability. Agirre et al. [18] state that "if we want [Human Language Technology] to be of help for more than 6000 languages in the world, and not a new source of discrimination between them, the portability of HLT software is a crucial feature." For this purpose, natural language processing tools should be modular and language-independent allowing them to be combined for building a text analysis system for a new language.

Portability requires a more modular and flexible architecture rather than a hard-wired ordering of algorithms or linguistic knowledge coded directly into the software. One way of accomplishing this is to develop knowledge such as language rules within an independent component using a metalanguage, such as the Xerox finite state tools for morphological analysis, unification-based grammars for parsing, or regular expression rules. In addition, a number of tools can be built language-independently so that they can be applied to new languages, such as interface tools for lexicon development or scripts for corpus analysis. A modular or componentalized approach to language technology will therefore enable us to arrange and develop the natural language processing system in the best possible way for any given application or language.

In the case of low-resource languages, modularity, language-independence, and portability are crucial features in a toolkit allowing it to rapidly be applied to the

development of computational systems for these languages, thus minimizing both human effort and cost.

4.3. Summary

This section discussed the main linguistic elements to be considered for the development of a partial parser with a small regular-expression grammar to recover syntactic fragments quite reliably. It was argued that, in the absence of related higher-density languages, one can develop linguistic knowledge fairly quickly for basic components. In addition, the existence of portable, modular, and reusable tools is crucial for application to new low-resource languages. Eventually, patterns for detecting multiword expressions, the distinct word orders found in active and passive sentences, as well as verbal subcategorization information can be added to enhance the basic rules if needed. Alternatively, partially parsed text can be used to train statistical systems.

5. Conclusion

By building upon the characteristics of three low-density languages –Tajiki Persian, conversational variant of Iranian Persian, and Eastern Armenian – this paper argues that there does not exist one single method for rapidly developing computational systems for low-density languages. Instead, it was proposed that focused studies can be performed to detect relevant language-specific characteristics, which can then be used for applying linguistic knowledge, taking advantage of portable and modular components in order to create stopgap language technology resources. These goals can be achieved by emphasizing two broad strategies: (i) *related language bootstrapping* can be used to port existing technology from a resource-rich language to its associated lower-density variant; and (ii) clever use of linguistic knowledge can be employed to *scale down* the need for large amount of training or development data. The methods discussed can be implemented for low-density languages in the goal of reducing human effort and avoiding the scarce data issue faced by statistical systems.

Acknowledgements

I would like to thank the participants of the 2007 NATO Advanced Study Institute on Low-Density Languages for valuable comments and discussion. The work on Tajiki Persian described in this paper was done in collaboration with Dan Parvaz and was supported by an innovation grant by the MITRE Corporation. The study on Persian weblogs was supported by a sponsored project at MITRE.

References

- [1] M. Maxwell and B. Hughes, 2006. Frontiers in Linguistic Annotation for Lower-density Languages. In *Proceedings of COLING/ACL2006 Workshop on Frontiers in Linguistically Annotated Corpora*, 29-37. Association for Computational Linguistics.
- [2] B. Hughes, 2005. Towards Effective and Robust Strategies for Finding Web Resources for Lesser Used Languages. In *Proceedings of Lesser Used Languages and Computational Linguistics*. EURAC, Bolzano,.
- [3] K. Oflazer, S. Nirenburg, and M. McShane, 2001. Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. *Computational Linguistics*, 27(1).
- [4] Ch. Monson, A.F. Llitjós, R. Aranovich, E. Peterson, J. Carbonell and A. Lavie, 2006. Building NLP Systems for Two Resource-Scarce Indigenous Languages: Mapudungun and Quechua. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation*.
- [5] H. Somers, 2005. Faking it: Synthetic Text-to-Speech Synthesis for Under-Resourced Languages – Experimental Design. In *Proceedings of the Australasian Language Technology Workshop 2005*. Sydney, Australia, pp. 71--77.
- [6] Ch. Xi, and R. Hwa, 2005. A Backoff Model for Bootstrapping Resources for Non-English Languages. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*. Vancouver, Canada, pp. 851--858.
- [7] Ph. Resnik, 2004. Exploiting Hidden Meanings: Using Bilingual Text for Monolingual Annotation. In *Lecture Notes in Computer Science 2945: Computational Linguistics and Intelligent Text Processing*, Springer-Verlag, pp. 283-299.
- [8] D. Yarowsky, G. Ngai, and R. Wicentowski, 2001. Inducing Multilingual Text Analysis via Robust Projection across Aligned Corpora. In *Proceedings of the First International Conference on Human Language Technology Research (HLT)*, pp. 161–168.
- [9] S. Tong, 2001. *Active Learning: Theory and Applications*. PhD Dissertation, Stanford University.
- [10] A. Feldman, 2006. *Portable Language Technology: A Resource-Light Approach to Morpho-Syntactic Tagging*. PhD Dissertation, Ohio State University.
- [11] K. Megerdooomian and D. Parvaz, 2008. Low-density Language Bootstrapping: The Case of Tajiki Persian. In *Proceedings of LREC 2008*, Marrakech, Morocco.
- [12] J.R. Perry, 2005. *A Tajik Persian Reference Grammar*. Boston: Brill.
- [13] K.R. Beesley and L. Karttunen, 2003. *Finite-State Morphology: Xerox Tools and Techniques*. Palo Alto: CSLI Publications.
- [14] J.W. Amtrup, 2003. Morphology in Machine Translation Systems: Efficient Integration of Finite State Transducers and Feature Structure Descriptions. *Machine Translation*, 18(3), pp. 217--238.
- [15] D. Sifry, 2007. The Technorati State of the Live Web: April 2007. Available at <http://technorati.com/weblog/2007/04/328.html>
- [16] J. Kelly and B. Etling, 2008. Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere, Research Publication No. 2008-01. The Berkman Center for Internet and Society.
- [17] D. Crystal, 2001. *Language and the Internet*. Cambridge: Cambridge University Press.
- [18] E. Agirre, I. Aldezabal, I. Alegria, X. Arregi, J. Arriola, X. Artola, A. Díaz de Ilaraza, N. Ezeiza, K. Gojenola, K. Sarasola, and A. Soroa, 2002. Towards the Definition of a Basic Toolkit for HLT. In *LREC 2002: Workshop on Portability Issues in HLT*. Las Palmas, Canary Islands, Spain.
- [19] O. Kolak & Ph. Resnik, 2005. OCR Post-processing for Low Density Languages. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Pages: 867 - 874
- [20] G. Ngai and D. Yarowsky, 2000. Rule Writing or Annotation: Cost- efficient Resource Usage for Base Noun Phrase Chunking. In *Proceedings of ACL-2000*, Hong Kong, pp. 117-125.
- [21] J.P. Chanod and P. Tapanainen, 1995. Tagging French – Comparing a Statistical and Constraint-based Method. In *EACL-95*.
- [22] G. Labaka, N. Stroppa, A. Way and K. Sarasola, 2007. Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation. In *Proceedings of MT-Summit XI*, Copenhagen.
- [23] D.W. Oard, 2003. The Surprise Language Exercises. In *ACM Transactions on Asian Language Information Processing (TALIP)*, Volume 2 , Issue 2 Pages: 79 - 84
- [24] S. Abney, 1997. Part-of-Speech Tagging and Partial Parsing. In *Corpus-based Methods in Natural Language Processing*, Edited by S. Young and G. Bloothoof. Kluwer Academic Publishers, Dordrecht.

