Unit 4

Designing a Software for Improving Writing Literacy

DESIGNING A SOFTWARE FOR PERSIAN ORTHOGRAPHICAL FEATURES

Pre-reading Activities

In this unit, you will

- improve your understanding of the target technical words.
- learn about various supporting topic sentences in writing.
- learn how to preview a reading comprehension passage through pre-reading questions to improve comprehension.
- be familiar with designing Persian orthographical features.

I. Target Academic Vocabulary

Check out the meanings and functions of the target academic words in a monolingual and bilingual dictionary.

Orthographical (adj)

Cognate (adj)

Transform (v)

Contrary (adj)

Classify (v)

Detect (v)

Pop up (v)

Substitute (v)

II. Writing development

Enumerators

So far, we have learned how to limit a topic sentence and support it with examples, details, facts and statistics. Now, we should logically and cohesively structure our supporting sentences. The most common

method for developing a paragraph is 'enumeration', that is, an author starts a paragraph with a general idea and then breaks it down by listing some, most or all parts. For examples,

- Computer (General idea)
- Hardware & Software (parts)
- University (General idea)
- Humanity and social science college, Law and political science college, Engineering college, Art and Education college (parts)

Note: Some authors prefer to use a variety of enumerators beside *kinds, types* or *parts*. The most common ones are: classes, elements/ factors, characteristics, aspects, divisions, subdivision, and categories.

III. Pre-reading questions:

Read and respond to the questions below, and then discuss them in pair/group.

1.Do you have any ideas how to improve writing literacy digitally?

2.Did you ever read about digitally improving the Persian orthography improvement?

---

3. What could be new quick software to indicate Persian orthographical errors?

IV. Reading comprehension passage

This passage discusses designing software to improve Persian orthographical features in a computer.

DESIGNING SOFTWARE FOR PERSIAN ORTHOGRAPHICAL FEATURES

With the advance of natural language processing techniques and the expanding use of computers, researchers have examined many languages to find out orthographical and structural errors in a text. Examining orthographical correctness and morphological consistency of words involve the application of natural language processing techniques. Little research has so far been undertaken in computationally analyzing Persian language.

Computational lexicon is among the most important resources needed to design a system that checks the orthography and morphology of words. In a language with a rich morphology, such as Persian and Arabic, the lexicon is expected to provide enough information to enable the system to process intricate inflections correctly.

The software system detects context-independent misspellings and checks

the morphological consistency of words in Persian language context, and provides isolated-word error correction. The system assists a user by offering a set of candidate corrections that are close to the incorrect word. For example, when the system receives the word "پصر", it not only recognizes the misspelling, but also suggests "پسر" as a replacement. In addition, if a word like "کتابان" appears instead of "کتابها", the system detects a morphological error, and gives an appropriate message to correct it. Figure 1 shows a block diagram of the system.

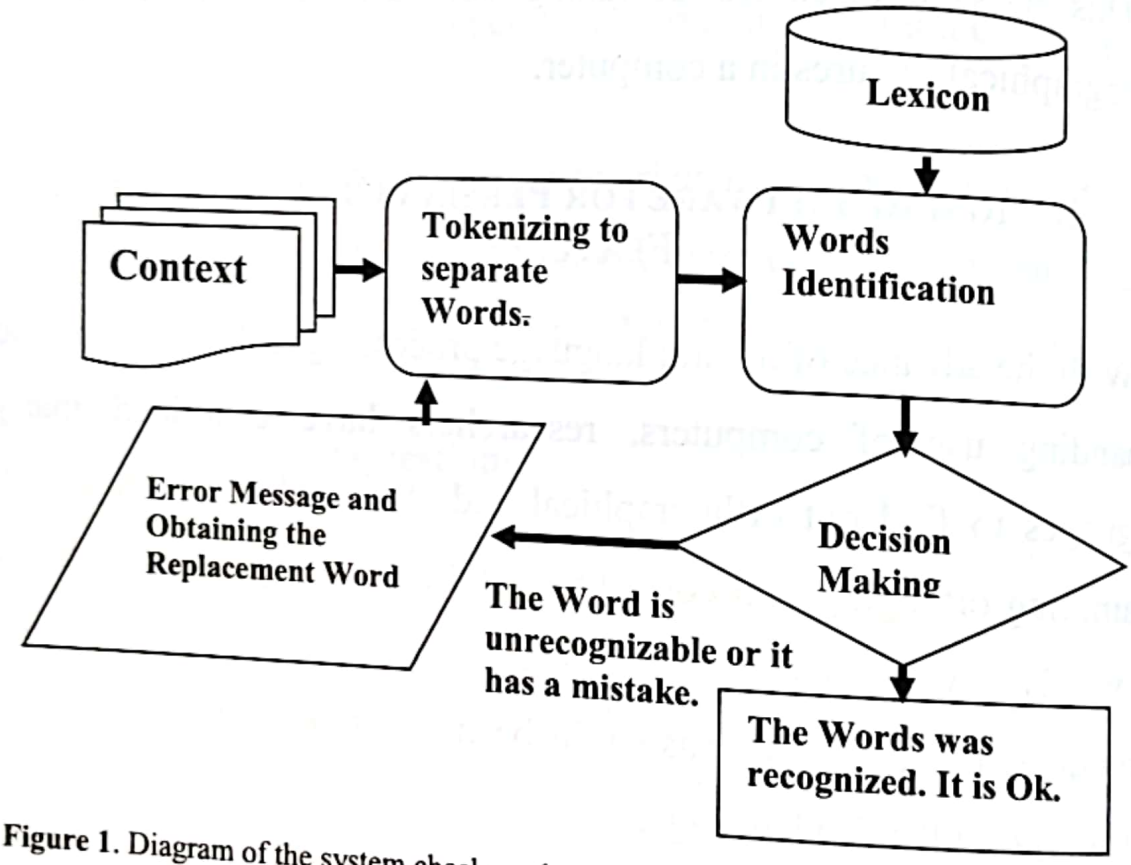

Figure 1. Diagram of the system checks orthographical consistency of words in the context.

1. Parsing and identifying the words

The presented system isolates words in the text using the blank space between two consecutive words. Then, it evaluates the orthographical and morphological correctness of the words by means of the lexicon. If the system can find the exact word in the lexicon, it confirms the

orthography and morphology of the word. Hence, the more the words in the lexicon lead to more accuracy for the performance of the system.

Consequently, all derivatives of a word in the lexicon are needed. This point would cause the size/volume of the lexicon to be dramatically large. Therefore, a method is required for optimizing the size of the lexicon and hence improving the system performance is presented.

2. Implementing the lexicon

To reduce the size of a lexicon, the stem within the lexicon replaces the whole set of words, which can be extracted from the same stem. In order to obtain all of the derivative words from a stem existing in the lexicon, the morphological information for each of the stem words should be there. Hence, a code is inserted in front of each word containing information regarding its grammatical characteristics.

For each word an eight-bit code is sufficient to store all of its morphological information. Designing such a code system is a subtle task that is explained below.

Words in Persian can be classified into seven morphological groups: noun, verb, preposition, adjective, adverb, pronoun, interjection. This classification has been used because different grammatical groups have their own rules to produce cognate words. For instance, verbs and interjections cannot be in plural form, but common nouns may appear in plural form. In addition, common nouns like "کتاب" can be pluralized into "کتابها"; "دوست" into "دوستان" and"دوستها"; and "امتحان" into "امتحانها" and "امتحانات". This kind of information must be provided by the

characteristic code of words in the lexicon.

The characteristic code contains two parts. The first part indicates the group (type) of the word, and the second part indicates permissible operations that can be implemented on the word (see Figure 2). The characteristic codes have a fixed length, equal to eight bits, but the length of their two constituent parts varies depending on the word group. As the number of grammatical rules applicable to different word groups may not be the same, the second part of the variable code does not need a fixed length. For example, interjections have the same grammatical features in Persian language, but nouns and/or verbs have different grammatical features. Thus, for recognizing interjections, only one code is enough, that is, the length of the operation code for interjections is zero.

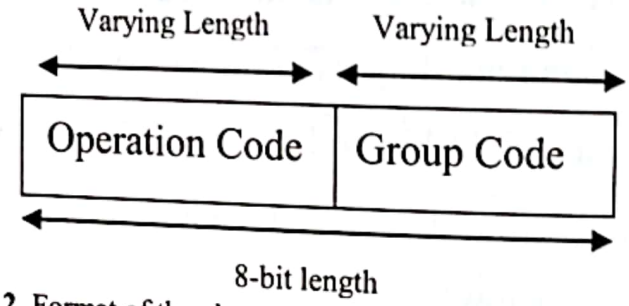

Figure 2. Format of the characteristic code for words in the lexicon.

A case in point is that verbs can be transitive or intransitive and some verbs can be transformed from intransitive to transitive ones on the basis of their rules. In addition, there are different ways to transform present root to an infinitive form.

Different rules must be used to make infinitive words like "دویدن", "رفتن", "شنیدن", "آزمودن" and "آموختن" from their present root verbs "دو", "رو", "شنو", "آزما" and "آموز" respectively. These examples and the examples

provided earlier indicate that verbs and nouns in lexicon need more than one code.

Table 1 illustrates codes, which are used for seven groups of words in this system. The codes are in binary, and "X" indicates that the related bit can be "1" or "0" in which the former shows a particular grammatical feature for the word. For example, a code like "00100000" is used for the proper adverbs such as "never" and "sure"; and the code "01100000" is used for the common adverbs such as "year" and "time". Since these common adverbs may appear plural, such as "years" and "times", contrary to proper adverbs, this distinction has been made. It should be noted that there are different types of adverbs in Persian grammar. However, they are morphologically classified into two groups: proper adverbs and common adverbs. Hence, the two-bit pattern is enough for

TABLE 1. The Binary Characteristic Codes for Different Types of Words. Xs represent the operation code.

Type of Word	Characteristic Code
Verb	XXXXXXX1
Noun	XXXXXX 1 0
Adjective	XXXXX 1 0 0
Pronoun	XXXX 1 0 00
Preposition	XXX 1 0 000
Adverb	XX 1 0 0000
Interjection	0 1 0 00000

this coding. Therefore, if the system received a word in a sentence that appears plural and the word was introduced as a proper adverb, the word contained a structural error.

3. Morphological orthographical analysis and orthographical errors

When the system finds a word in the lexicon, regardless of its meaning in the context, it considers the word as correct in orthographical and morphological respects. Otherwise, the system considers the word as an extended word (its stem exists in the lexicon); hence, it tries to detect the stem. In detecting the stem, the system may need to pass through a few steps. In each step, the possible added prefixes and/or suffixes are removed from the front and/or back of the word respectively, until the word can be found in the lexicon. Then, the characteristic code of the detected word is used to judge whether the morphology of the original word is acceptable. In other words, if the characteristic code does not let the stem have the specified prefix or suffix, the system pops up the message, "The word has a morphological error". For instance, after the system receives a word "کتابان", it initially searches for the word in the lexicon, and if it doesn't find it, it then recognizes "ان" at the end of the word as a sign of a plural form in some nouns in Persian language. Finally, it searches for its stem "کتاب" in the lexicon. According to Persian grammar, the sign of plural is "ها" for this word. Hence, the message system first pops up, "This word is inaccurate in the plural form", and then suggests the "ها" as a replacement.

Finding the stem of extended words involves a number of steps.

Each step relates to one grammatical group where the system tries to find incorporated morphemes. For example, in the step that relates to the noun group, the system tries to find morphemes such as [ها,ات,ان, or ندـ,يدـ,يمـ,دـ,يـ,وـ,م] or a compound form of them on the end of the word. Then it deletes those morphemes and searches for the rest of the word in the lexicon. The system will go to the next step, if the word does not have any morphemes related to the current group, or by deletion of the morphemes, the system cannot find the word in the lexicon.

Finally, if the system cannot determine the word in any of the above steps, the word will be considered as orthographical errors. Since the orthographical errors are presented in the homophone letters, they can be classified into [ع,ا,آ], [غ,ق], [ط,ت], [ث,ص,س], [ظ,ض,ذ,ز], and [ه,ح]. If any of these letters are present in a word, they will be transformed to other homophone letters from the same group and following that the word is searched in the lexicon.

4. Implementing the system

The lexicon file used in this system contains more than 12000 word stems. A logical record is constructed for each word in the lexicon. Since the length of different words may not be the same, records with varying length are considered in the database. To speed up the searching, a three-level index has been used for a lexicon. In this system, a user can retrieve or add a word in the lexicon.

To evaluate the performance of the system, texts with orthographical and/or morphological errors have been tested on the system. Except for cases that stem of words didn't exist in the lexicon, the system in all

cases could successfully recognize any morphological or orthographical errors.

It should be noted that implementing this system requires a word with an orthographical error having one letter replaced with one of its homophone's letters. For instance, the word "دست" may appear as "دسط" or "دصت"; but not "دشط", or "دصط". This system can process and detects an orthographical error in a word only if one letter of the word has been substituted by one of its homophone letters.

Post-reading Activities

I. Reading comprehension

Directions: Mark each statement as T (True), F (False), or NG (Not Given) to the information in the reading comprehension passage.

---------1.Enough research has been done in analyzing Persian language computationally.

---------2. The software system helps find context-dependent mis-spellings and consistency of words in Persian.

---------3. Code containing information on grammatical features is used for each word.

---------4. All transitive and intransitive verbs can be exchanged and used.

---------5. No difference exists between common and proper adverbs.

---------6. After finding the stem, the system needs to go through a few steps.

--------7. The first step focuses on grammatical group in cooperated with morphemes.

--------8. The number of word stems in this system is exactly 12000.

Questions ***9-15***: Choose the appropriate letter ***A-C.***

9. To find out orthographical and structural errors in a text, investigators……..

A. have done many experiments in the lab.

B. have done much research on various languages.

C. have invented many systems to track the errors.

10. What needs to be done before examining the orthographical and morphological correctness?

A. Finding an exact word in a lexicon.

B. Confirming orthography and morphology of a word.

C. Using a space between two words.

11. To reach the derivatives of a word stem, we should have access to……

A. phonological information.

B. morphological information.

C. the code information.

12. All the following options are TRUE but…….

A. The code feature has two parts.

B. The length of the two parts depends on the word group.

C. The code feature does not have a fixed length.

13. The difference between these two codes "00100000" and "01100000" 1is that .....

A. there is no important difference between them.

B. the former focuses on the proper adverbs and the latter focuses on the common adverbs.

C. there is a difference between them but it is not worth considering it.

14. According to the passage, the system pops up message when.....

A. a user wants to change a program.

B. a user makes an orthographical error.

C. a user considers an extended word.

15. This system provides an error message when .....

A. an extended word is replaced with a simple word.

B. a wrong letter in a word is detected.

C. a wrong word in a phrase is detected.

II. Vocabulary activities

Directions: Read each sentence on designing software for Persian orthographical features stated below. Circle the one word or phrase in parentheses ( ) that has the same meaning as the underlined word in the sentence. Compare your answers with a partner.

1. Examining orthographical correctness and morphological consistency (*ability/accuracy/proficiency*) of words involve the application of natural language processing.
2. Little research has so far been undertaken (*done/mentioned/enabled*) in computationally analyzing the Persian language.
3. For example, interjections (*excitements/interruptions/acceptances*) have the same grammatical feature in Persian language, but nouns and/or verbs have different grammatical features.
4. In each step, the possible added prefixes and/or suffixes are removed (*deleted/inserted/pasted*) from the beginning and/or end of the word respectively, until the word can be found in the lexicon.
5. The characteristic code of the detected (*rejected/accepted/targeted*) word is used to judge whether the morphology of the original word is acceptable.
6. The first step belongs to one grammatical group where the system tries to find incorporated (*extended/attached/informative*) morphemes.
7. The system assists (*improves/interests/helps*) the user by offering a set of candidate corrections that are close to the incorrect word.
8. In order to obtain all of the derivative words from a stem existing (*including/bringing/taking*) in the lexicon, the morphological information for each of the stem words should be there.
9. A case in point is that verbs can be transitive or intransitive and some verbs can be transformed (*taken/changed/systematized*) from intransitive to transitive ones on the basis of their rules.

10. As the number of grammatical rules applicable (*irrelevant/ logical/ appropriate*) to different word groups may not be the same, the second part of the variable code does not need a fixed length.

III. Writing development activities

Directions: Analyze the model paragraph below by filling in all the blank spaces provided.

Model paragraph 1

Words in Persian can be classified into seven morphological groups: noun, verb, preposition, adjective, adverb, pronoun, acoustics (sound). This classification has been used because different grammatical groups have their own rules to produce cognate words. For instance, verbs and sounds can't be in plural form, but common nouns may appear in plural form. In addition, common nouns like "کتاب" can be pluralised into "کتابها"; "دوست" into "دوستان" and"دوستها"; and "امتحان" into "امتحانها" and "امتحانات". This kind of information must be provided by the characteristic code of the word in the lexicon.

1. What is the general idea in the model paragraph above?

---

2. What are the enumerators used in the paragraph?

---

3. What types of supportive information do the authors use in the

paragraph (examples, facts, statistics or details)?
