AUTOMATIC DICTIONARY ORGANIZATION IN NLP SYSTEMS FOR ORIENTAL LANGUAGES

V.Andrezen, L.Kogan, W.Kwitakowski, R.Minvaleev, R.Piotrowski, V.Shumovsky, E.Tioun, Yu.Tovmach

Dept. of Applied Linguistcs
Hertzen Pedagogical University
48, Moika, 191186
St.-Petersburg, USSR

Abstract

This paper presents a description of automatic dictionaries (ADs) and dictionary entry (DE) schemes for NLP systems dealing with Oriental languages. The uniformity of the AD organization and of the DE pattern does not prevent the system from taking into account the structural differences of isolating (analytical), agglutinating and internal-flection languages.

The "Speech Statistics" (SpSt) project team has been designing a linguistic automaton aimed at NL processing in a variety of forms.

In addition to Germanic and Romance languages the system under development is to handle text processing of a number of Oriental languages. The strategy adopted by the SpSt group is characterized by a lexicalized approach: the NLP algorithms for any language are entirely AD dependent, i.e., a large lexicon database has been provided, its entries being loaded with information including not only lexical, but also morphological, syntactic and semantic data. This information concentrated in dictionary entries (DEs) is essential for both source text analysis and target (Russian) text generation.

The DE structure is largely determined by the typological features of the source language. The SpSt group has hitherto had to deal with European languages and it was for these languages (inflective and inflective - analytical) that the prototype entry schemes were elaborated and adopted. No doubt, the typological characteristics of Oriental languages required certain modifications to be made to tne basic scheme. Hence in the present paper each of the language types is given consideration. Agglutinating languages proved to be the most suitable to process according to the SpSt strategy. But an isolating language will be the first to be proposed for discussion.

1. The AD organization for an isolating language: Chinese

For the purposes of NLP it is plausible to assume written Chinese as exclusively isolating language where affixation is virtually non-existent. The few inflective word-forms are entered into the lexicon as unanalizable lexical items, whereas multiple grammar formants are treated as free structural elements. High degree of lexical ambiguity making syntactic disambiguation a must, and the fact that word boundaries are not explicitly marked in the text are well-known problems with Chinese text analysis. (Actually, in the MULTIS project elaborated by the SpSt group Chinese characters are transformed into 4-digit strings in conformity with Chinese Standard Telegraph Code).

Thus grammatical and logico - semantic relations in the text are expressed by word order, structural words and semantic valencies. In addition to their role of the labels for syntactic units (predicate, direct and indirect objects, etc.) the structural words function as delimitators singling out word-forms and phrases. A separate sub-lexicon for structural words is accordingly provided within the whole lexicon database of Chinese as source language. The file of notional words comprises lexical items of various lengths ranging from one-character items to eight-character ones, no differentiation being made among one-stem words, composite words and phrases. A distinct version of the DE scheme is assigned to each of the two classes of lexical items: notional words (N/W) and structural words (S/W).

The DE scheme for N/W includes, along with syntactic and semantic, the following data: 1) Part - of - speech assignment; 2) Information on the lexical ambiguity. Thus, by way of example, the one-stem word sudan 'sultan' and composite beida 'Beijing University' are coded N00, where N denotes noun, while qianding 'to sign a treaty' is coded 0S0 where S denotes verb/noun lexical ambiguity (to be eventually disambiguated by syntactic means).

As to the DE schemes for S/Ws, each of these should include positional characteristics of the lexical item and provide information on the way the given particle affects formation of the Russian equivalent. E.g., in the grammatical coding of the verbal aspect S/W le and nominal S/Ws de and ba the following points are marked: 1) part - of - speech dependence; 2) position (pre- or post-position with respect to the N/W); 3) Russian matching; 4) syntactic function.

The information placed in a DE may be used in translating sentences as illustrated below:

苏丹把和约签订了。
Sudan ba heyue qianding le.
The sultan the peace treaty signed

In carrying out the lexico - syntactical analysis of this sentence two word groups are delimitated : nominal group ba heyue and verbal group quanding le. In the ba-DE there are data to define ba as a S/W in preposition to a direct object which is equivalent to a Russian noun in the Accusative Case. In the le-DE there are data to define le as a verbal index in a post-position to a verbal predicate and indicating the completion of an action, equivalent to a Russian verb in the Past Tense, Perfective. (For the sake of simplicity the polyvalent and polysemantic nature of these particles is ignored in this example).

2. The AD organization for an agglutinating language: Turkish

The agglutinative word-formation technique is characterized by ordered addition of affixes to the stem to produce formant strings of various lengths. An outstanding feature of agglutinating languages is that these word-forms are not reproduced ready-made in speech but are constructed by the speaker actually 'ad hoc' according to definite rules. Each of the limited set of affixes imparts 'a semantic quant' or represents a grammatical category. E.g., see the following patterns where the stem 'sultan' and some of its derivatives are presented:

sultan	' sultan '
sultanlar	' sultans '
sultanlariniz	' our sultans '
sultanlarinizdan	' from our sultans '

Word-formation in Turkish is carried out in accordance with either of the two prototypes: nominal or verbal. Of one nominal stem it is theoretically possible to derive an infinite number of word-forms (actually,though, only some 200 as registered in the corpora). As to the verbal paradigms, of each stem it is potentially possible to form more than 11 thousand word-forms.

Clearly, Turkish lexicon database would include, besides stems, sub-lexicons of postpositive affixes. Along with the stems of which both nouns and verbs may be derived there are those assigned to only one definite part of speech class (e.g. gel 'come'), also unproductive lexemes such as ve 'and' zaten 'generally'.

Each DE contains coded information indicating: 1) the lexeme's part of speech class and type of lexical ambiguity (e.g., for the stem insan 'man' it is noun/adjective ambiguity, that is, NA); 2) the lexeme's semantic class (e.g., for the stem insan 'man' there is an indication that it belongs to Subject (S) semantic class, and, consequently, may function as the subject of a sentence; 3) Russian equivalent (the address of the "machine" stem with necessary lexico-grammatical information).

Turkish affixes are structured so as to form four connected graphs constructed to the rules of the grammar of orders. Graph 1 presents simple noun morphology; Graph 2, finite verb form morphology; Graph 3, non-finite verb forms; Graph 4, nominal predicate;. The word stem is assigned to the graph root while affixes (their allomorph variants) are assigned to graph nodes, each node corresponding to definite grammar rank.

Recognition and lexico-morphological analysis of the Turkish text word is accomplished as follows: 1. Stem recognition and affixes delimitation by means of the AD search. If this results in the recognition of the input text word, the task is fulfilled, and the target word equivalent is passed to the output unit (e.g., the text word 'Ankara'). 2. If no recognition is acknowledged, the system goes on with lexico-morphological analysis.It is performed by consecutive superposition of affixes on the end segments of the string, the affixes being fed by access to an appropriate graph. The operation is accomplished by mask matching method proceeding from right to left, from junior order to senior order affixes. All possible affixes having been identified, the initial part of the text word that remains is treated as the hypothetic stem and is eventually searched in the AD.The search may result in different situations.3. If the hypothetic stem is identified as one of AD stems and its part of speech assignment coincides with that of affixes, then the task is considered to be fulfilled. E.g., in analysing the text word tutanaklarinin the noun stem tutanak 'protocol' is revealed:it is ajoined by nominal affixes larinin.

The target equivalent with its grammatical characteristics is passed to the syntactic module.

4. In case of failure (that is, when the stem is not found) the string is recovered in its original form (identified affixes are 'glued' back), and the

analysis restarts with access to Graph 2 on the assumption that the input text word is a finite verb form, etc. This sequential access to graphs does not take place at random but has been programmed according to the frequency data received by a preliminary quantitative analysis of some text corpora.

3. The AD organization for internal-flection languages: Arabic and Hebrew.

The word morphology of Arabic and Hebrew is not only characterized by internal flection but also by a rather wide use of agglutinative formants and external flection. Taking into account these features of the Semitic word-form structure three different approaches to AD design seem plausible.

1. Representation of lexicon items by word-forms listed in alphabetical order; in this case the following Hebrew words would have three independent DEs:

סלטן	SiLTWoN	sultan
סלטנים	SiLTWoNJiM	sultans
	SuLTWoNei	sultans
סלטני	(status constructus)	

2. An alphabetical arrangement of machine stems as has been made for European languages; in this case the above Hebrew wordforms may be reduced to only one item.

3. Designing the source lexicon as a lexicon of roots; all above-mentioned Hebrew word-forms would then be representated by the root ס.ל.ט SLTN supplemented with lists of internal and external affixes.

Since word-formation and word-building in Semitic languages are practically limitless the option of the first or second approaches would cause a dimension crisis with respect to the lexicon size: the AD would surpass the critical storage capacity while the dictionary search would be strongly impeded.

With root-based AD organization the root-originated word-form development process follows the order: "root-derivation - internal flection types - rules of combination with definite external affixes". Unfortunately, this kind of AD organization requires, for the purposes of the text-word lexico-grammatical analysis, a multiple access to the hard disk, and this would again cause a dimension crisis, now with respect to the system operating speed.

To relieve the dimension crisis a trade-off may be suggested: combined root-based and alphabetic approach to construction, operation and maintenance of the AD. With this approach five lists (sub-lexicons) of linguistic units are distinguished.

1. List of roots actually in use (some 500 for Hebrew, 200 for Arabic).
2. List of internal flections (some 600 for Hebrew, 900 for Arabic).
3. Alphabetic list of words with regular word-formation (nouns, adjectives, also basic forms of verbs);
4. List of words of Semitic origin with irrigular word-formation and borrowed words (i.e., those where triliteral scheme is not observed). E.g.,

יום 'day', ימים days (this is one of the few Hebrew nouns where the internal flection is changed with word-formation).

טלגרף 'to telegraph' (a borrowed word, not subject to conjugation rules).

לקח 'to take' root לקח, two root letters are omitted Arabic: 'father'

... أبي - Genitive; أبا Accusative

Nominative... أبو (an anomaly in declension).

لَيْسَ 'not to be', a verb having only Perfect forms.

5. List of external affixes (prefixes, suffixes, circumfixes). These are compiled having in view their ability to form combinations. Lists 1 and 2 being of a limited length are included into the RAM: this allows for the possibility to analyze the text word without accessing the hard disk. The rest of the lists are entered into the disk database. Accesing to these lists is to take place after the primary root - affix identification of the text word has been done.

Stems of other lists may be assigned various entries. Irregular word-forms are specified as paradigms where each word-form is supplied with the target language equivalent.

Recognition and lexico-morphological analysis of the Semitic text words goes on by the following steps:

1. The root is singled out and recognized according to List 1. The operation performed is in fact a combinatorial-probabilistic analysis of possible consonant combinations within the input text word. The operation is based on the actual consonants being used exclusively in roots (so-called root consonants) or in both roots and affixes (structural consonants).

2. Internal flection types (derivations) and their versions are identified with the models included in List 2.

3. The roots recognized are reduced to lexicon forms as in List 3: this allows one to get the target language equivalent of the item. The final synthesis of the target text word is performed on the basis of the information of the internal and external flections of the given source text word. The external flections are determined by the types and versions of the internal flection: singling out an internal flection automatically identifies the corresponding external one with one of the models in List 5.

If the system fails to recognize the given text word, which may be caused by the irregular word-formation, this word is translated with the aid of List 4. Besides, the lexico-morphological analysis certainly makes use of the dictionary of phrases though its structure is not considered in this paper.

Conclusion

As is evident, the very notion of the text word, which is so essential in designing automatic dictionaries, is quite distinct in each of the Oriental languages and fundamentally different from what we are used to treat as a text word in Indo-European languages, inflecting or inflecting-analytical. If an Oriental language AD is to be integrated into a multimodular NLP system (such as MULTIS elaborated by the SpSt group) and the system has to retain its basic structure, this project requires development of various forms of sub-lexicon databases. As we have seen the most complicated structure of Arabic and Hebrew text word required elaboration of four versions of DE while the differentiation of notional and structural words in Chinese required two versions. An agglutinative word structure model such as in Turkish, though the most suitable for the SpSt grammar, required a tree-structured database and special procedures of access.