[From: *International Journal of Translation*, vol.13, no.1-2, Jan-Dec 2001]

Dictionary Organization in Linguistic Automaton for Oriental Languages

RAJMUND PIOTROWSKI, YURI ROMANOV, YURI TOVMACH,
NATALIA ZAITSEVA, MICHAEL BLEKHMAN

The central problem for natural language processing (NLP) systems dealing with non-Indo-European ("Oriental") languages is how to develop automatic dictionaries (AD) and dictionary entry (DE) schemes. The point is that the need of Oriental language industrial NLP has been felt for some time. It has acquired additional urgency with the rapid growth of business contacts between Russia and the nations of the Middle East and the Pacific Rim. The very notions of such language items as root, stem, word form (w/f) and text word (t/w), which are so essential in designing an AD, are quite distinct in each of the Oriental languages and fundamentally different from what we are used to treat as a root, a t/w etc. in the Indo-European languages. If an Oriental language AD is to be integrated into a multimodular linguistic automaton and the system has to retain its basic structure, this project requires development of various forms of sub-lexicon databases. The structure of Arabic and Hebrew t/w requires elaboration of four versions of DE while the differentiation of full and structural words in Chinese provides two versions. An agglutinative word structure model, such as Turkic and Finno-Ugric, requires a tree-structured database and special procedures of access.

INTRODUCTION

There is an increasing theoretical and practical interest in the belief, perception and expectations which scholars, specialists in informatics and businessmen bring to the NLP of non Indo-European ("Oriental") languages.

The beginning of the new millenium is marked by intensification and expansion of the types of information flows that are processed by linguistic automata (LA) - cf. Piotrowski 1999, p. 140 f. The list of languages used for engineering, business, and political information is expanding. Aside from Japanese, which has established itself in the information industry since the late 1970s, national languages of the new industrial "tigers" - China, Indonesia, Malaysia, Singapore, South Korea, and Taiwan are becoming involved in NLP. At the same time, Arabic, Finnish, Hebrew, Hungarian, and the Turkic languages have also entered the arena of language engineering and machine translation (MT). Therefore in addition to Germanic, Romance and Slavic languages, the international Speech Statistics group (SpStG) has been handling text processing of a number of "exotic" languages (Andrezen et al. 1992; cf. McCarty and Prince 1990).

Years of experience with NLP development suggest that the core module of an LA should be an automatic dictionary combined with elementary morphologic analysis and synthesis. That is why the SpStG begins the development of NLP systems, including MT systems for Oriental languages, with compilation of a larger AD, which stores encyclopedic, lexical and morphological information.

Our purpose is to characterize and compare lexical-morphologic models of AD entries for some non Indo-European languages in accordance with their typology. In contrast to inflective-analytic European patterns, these languages are characterized by the following word structures:

- the root pattern, which involves the use of syntactic words and lexical stems with no morphology,
- the interflectional pattern, where the grammatical meaning is expressed by an internal inflection, i.e., the change of sounds/letters of the stem,
- agglutinative patterns, where word-forming and word-changing affixes are attached to a lexical stem on the right according to a sequence prescribed by a strict morphotactics (Table 1).

English	Root pattern (Chinese)	Interflectional pattern (Arabic)	Aggultinative pattern (Turkish)
a/the sultan	蘇丹 sūdān	سلطان sultān	sultan
(the) sultans	蘇丹們 sūdānmen	سلاطين salātīn	sultanlar
to (the) sultans	給蘇丹們 gei sūdānmen	لسلاطين lisalātīn	sultanlara

Table 1. Forms of the noun sultan in Chinese, Arabic and Turkish/ Azerbaijani

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 autosemantic hieroglyphs and their combinations, the so-called full words (F/W), are entered into the lexicon as unanalyzable lexical items, whereas multiple grammar formants are treated as free structural words (S/W). High degree of lexical ambiguity making disambiguation a must, and the fact that “word” boundaries are not explicitly marked in the text are well-known problems with the Chinese text analysis (Sproat et al. 1996, pp. 378 f.). Here the grammatical and logico-semantic relations in the text are expressed by S/W, word order, and semantic valences. In addition to their role of the labels for syntactic units (predicate, direct and indirect objects, etc.), the S/W function as delimitators singling out "word" and phrases.

In the SILOD MT system, developed by the SpStG, a separate sub-lexicon for S/W is accordingly provided within the whole lexicon database of Chinese as a source language. The F/W file comprises lexical items of various lengths ranging from one-hieroglyph items

to eight-hieroglyph 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 F/W and S/W files. The DE scheme for F/W (Table 2) includes the syntactic and semantic data needed for the Chinese input text analysis.

Table 2. SILOD DE scheme of Chinese

F/W 蘇丹 sudan

	Lexical/grammatical class (part of speech) and possibility of gramma tical homography	Semantic and functional Features	Russian equivalent *султан* *
Codes	N Ø Ø	s c Ø a Ø p	S n 2 a Ø Ø
Positions in the DE	1 2 3	4 5 6 7 8 9	10 11 12 13 14 15

Thus, by way of example, the composite 北大 *beidà* ‘Beijing University’ is coded as N Ø Ø, where the N in the first position in the above scheme denotes a noun, while 簽訂 qianding ‘to sign (a treaty)’ or ‘signing’ is coded as Ø S Ø, where the S from the second scheme position indicates the verb/noun lexical ambiguity (to be eventually disambiguated by syntactic means).

As to the DE schemes for S/W, 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 of S/W

了 le and of the nominal S/W 的 de or 把 b a

the following points are marked:

1) part-of-speech dependence;
2) position (pre- or post-position with respect to the A/W);
3) Russian equivalents;
4) syntactic function.

Let us consider the analysis and translation procedure of the following Chinese text:

蘇丹把和約簽訂了
Sudan ba héyue qianding le
'(The) sultan the treaty signed'.

The first phase involves step-by-step search and identification of text hieroglyphs and their combination with their counterparts in the both vocabulary files. After that all information from DE is extracted and transferred into the text frame. As a result, one obtains a word-for-word or phrase-for-phrase translation.

In carrying out the lexical-syntactical analysis of this sentence, three word groups are delimitated:

two nominal groups	蘇丹	Sudan	(the) sultan'
and ,	把和約	ba héyue	'(the) peace treaty'
and the verbal phrase	簽訂了	Qiandìng le	'signed'.

In the 把*ba-DE* there are data to define it as a S/W in preposition to a direct object which is equivalent to a Russian noun in the Accusative Case.

In the 了 le-DE there are data to define it as a verbal index in a post-position to a verbal predicate and indicating the comipletion 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). As a result of the described procedure, our LA generates a correct Russian text with a changed word order: *султан подписал договор.*

2. THE AD ORGANIZATION FOR INTERNAL-FLEXION LANGUAGES: ARABIC AND HEBREW.

The Semitic morphology is characterized by not only the internal flection but also by a rather wide use of agglutinative formants and the external flection (Kataja, Koskenniemi 1988; Beesley 1988; Kiraz 2000). Taking into account these features, three different approaches to the Semitic AD and its DE seem plausible:

1. Representation of lexicon items by w/f listed in alphabetical order. In this case, the following Hebrew words would have three independent entries:

סלטן *SiLTWoN* **sultan**
סלטנים *SiLTWoNiM* **sultans**
סלטני *SuLTWoNei* **sultans**
(status constructus)

2. An alphabetical arrangement of machine stems, as it has been made for European languages. In this case the above Hebrew

סלטן - *SiLTWoN.*

3. Designing the source lexicon as a lexicon of roots; all above mentioned Hebrew w/f would then be represented by the root

סלטן - *SLTN*

w/f may be reduced to only one item

supplemented with lists of internal and external affixes.

Since word-formation and word-building in the Semitic languages are practically limitless, the option of the first or the second approach 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 the root-based AD organization, the root-originated w/f development process follows the order: "root-derivation - internal flexion types - rules of combination with definitive external affixes". Unfortunately, this kind of AD organization requires, for the purposes of the t/w 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 this crisis a trade-off may be suggested: a combined root-based and alphabetical approach to the 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 250 for Arabic, 300 for Hebrew),
2) List of internal flections (some 2500 for Arabic, 800 for Hebrew),
3) Alphabetic list of roots with regular word-formation (nouns, adjectives, also basic forms of verbs),
4) List of roots of the Semitic origin with an irregular 'word-formation.

E.g. arab. أب - آبَا *'AB - ABĀ'* 'a father, fathers' etc.;
hebr. יום - ימים *JOM - JaMIM* 'day - days' with the loss of *vav* in the internal flection,

5) List of external affixes (prefixes, suffixes, circumfixes), compiled with due account of combinations of these affixes with stems of various species.

Lists 1, 2 and 4 being of a limited length are included into the RAM: this allows for the possibility to analyze the t/w without accessing the hard disk. The rest of the lists are entered into the disk database. Accessing to these lists is to take place after the primary root - affix identification of the t/w has been done. Stems of other lists may be assigned to various entries. Irregular w/f are specified as paradigms where each w/f is supplied with the target language equivalent. The DE of each root, stem or affix is constructed in a way similar to that of the DE shown in Table 2.

Recognition and lexical-morphological analysis of the Semitic t/w goes on by the following procedure.

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 t/w.

The operation is based on the actual consonants being used exclusively in roots (the so-called root consonants) or in both roots and affixes (structural consonants).

2. Internal flexion 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 t/w is performed on the basis of the information of the internal and external flections of the given source t/w. The external flections are determined by the types and versions of the internal flexion: singling out an internal flexion automatically identifies the corresponding external one with one of the models in List 5.

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

3. The AD organization for an agglutinating language: Turkish & Azerbaijani

It has been known that the agglutinative word-formation technique is characterized by an ordered addition of affixes to the stem to produce formant strings of various lengths. Thus agglutinative w/fs are not reproduced ready-made in speech but are constructed by the speaker actually 'ad hoc' according to a definite morphotactic rules (Andrezen et al. 1992, p. 506; Oflazer 1994). Each of the limited sets of affixes imparts 'a semantic quant' or represents a grammatical category. As an example see the following patterns where the Turkish/Azerbaijanian stem sultan and some of its derivates are presented:

Sultan	'sultan'
Sultandan'	from (the) sultan'
Sultanlar	'sultans'
Sultanlarimiz	'our sultans'
Sultanlarimizdan	'from our sultans'

Word-formation in Turkish/Azerbaijani, along with all the Turkic languages, is carried out in accordance with either of the two paradigms: nominal or verbal. Of one nominal stem it is theoretically possible to derive an infinite number of w/fs (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 w/fs.

Clearly, the Turkic input AD is subdivided into two sub-lexicons. The first contains the stems of which both nouns and verbs may be derived, as well as those assigned to only one definite part-of-speech class (e.g., gel 'come'), and also unproductive lexemes, such as zaten 'generally'.

Each DE (cf. table 2) contains coded information indicating:

1) the lexeme's part of lexical ambiguity (e.g., for the stem insane 'man' it is noun/adjective ambiguity, that is NA);

2) the lexeme's semantic class (e.g., for the stem insane there is an indication that it belongs to the 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 lexical-grammatical information).

The second sub-lexicon includes word-changing and word-forming postpositive affixes (cf. S/W file for the Chinese syntactic words described in section 1). The Turkic affixes are structured so as to form four connected schemes constructed to the rules of the orders grammar. Scheme 1 presents simple noun morphology. Scheme 2 models finite verb form morphology. Scheme 3 represents non-finite verb forms. Scheme 4 depicts the nominal predicate structure.

Recognition and the lexical-morphological analysis of the Turkic t/w is accomplished as follows:

1. Stem recognition and affixes delimitation by means of the AD search. If this results in recognition of the input t/w, the task is fulfilled, and the target word equivalent is passed to the output unit (e.g., the t/w Ankara).

2. If no recognition is acknowledged, the system goes on with the lexical-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 a mask matching method proceeding from right to left, from the junior order to the senior order affixes. All possible affixes having been identified, the initial part of the text word that remains is treated as a 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 the AD stems and its part of speech assignment coincides with that of affixes, then the task is considered to be fulfilled. E.g., in analyzing the text word tutanaklarinin the noun stem tutanak 'protocol' is revealed: it is adjoined by the 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 Scheme 2 on the assumption that the input text word is a finite verb form, etc. This sequential access to above schemes 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. The algorithm for a morphological analysis of Turkic w/f has been detailed in the SpStG recent works, dedicated to Turkish-Russian (Mukhamedov, Piotrowski 1986, pp. 140 - 152) and Azerbaijani-Russian MT (Makhmudov 1982).

Statistical processing of larger corpora of Turkic texts done by the SpStG in the 1960s - 1980s (Bektaev 1978, pp. 26 f.; Mukhamedov, Piotrowski 1986, 84-136) revealed the fact that the transparent and consistently logical, with a rare exception, structure of the Turkic w/f makes it possible to build actually working generating algorithms of correct w/fs simultaneously for several Turkic languages. So, the SpStG has developed a program for an automatic synthesis of such w/fs. In it, the special analyzing/synthezing operators (ASO) take into account the presence or absence, in each particular language, of the palatal and labial vowel harmony, and also, of the consonant assimilation or dissimilation on the

morphological boundaries. When synthesizing w/fs, they took into account the ideas of the two-level morphology (Alam 1983; Koskenniemi 1986; Oflazer 1994; Seewald 1994, p. 8 f.), and they did it after the following pattern:

1) a Turkic stem (for nouns, it was the nominative case form) **was** introduced to the input of the LA together with the list of those grammatical meanings which must be present in the resulting synthesized w/f,

2) they indicated the language which the input stem belonged to and the corresponding w/f should be synthesized in,

3) they determined which synharmonic and assimilation/ dissimilation analyzers should be used with this particular Turkic language.

E.g., the LA receives the Turkic stem *с¥рг¥н* 'exile' and its task is to generate a w/f with the meaning of 'those who are in exile'. Using the palatal and labial vowel harmony ASO together with that of the consonant transformation on morphological boundaries, the LA generated the Kirghiz w/f *с¥рг¥ндθг¥лθр* which was grammatically correct and corresponding to the initial task. A similar task with the Kazakh stem *с¥ргін* (with the same meaning) the LA performed using the same program, but this time the labial vowel harmony ASO was switched off since there was no such a linguistic phenomenon in the Kazakh language. As a result, the LA generated the correct Kazakh w/f. *с¥ргіндегілер.* And at last, after the ASO had been switched off, the LA generated the correct Uzbek w/f surgundagilar. (As it is known, there is no vowel harmony and consonant assimilation/dissimilation on the morphological boundaries in Uzbek). This experiment was conducted on the basis of 300 Turkic stems and yielded about 90 per cent of correct results.

The algorithms of the type described above can be used only with such agglutinative languages in which vowel harmony and consonant assimilation/dissimilation on morphological boundaries work with a higher degree of regularity, as they do in the majority of the Turkic languages.

Unfortunately, in other agglutinative languages (for example in Estonian, Finnish, Hungarian, Japanese) there is a considerable amount of exceptions. Under certain circumstances, consonants in the stem or morphemes undergo irregular modifications and may sometimes be deleted. Similarly, a vowel in the root word, or in the affixed morphemes may also be lost. Thus the elaboration of the generating algorithms described above makes here no sense.

CONCLUSION

The need of industrial MT, automatic indexing and sense extracting of Oriental language texts has been felt for some time. It has acquired an additional urgency with the rapid growth of business contacts between Russia and the nations of the Middle East and the Pacific Rim. However, the notions of such language items as root, stem, text form and text word, which are so essential in designing automatic dictionaries, are quite distinct in each of the Oriental languages and fundamentally different from what we are used to treat as a word root, w/f, t/w etc. in the Indo-European languages. If an Oriental language AD is to be integrated into a multimodular linguistic automaton 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 an Arabic and a Hebrew text word provides elaboration of four versions of DE while the differentiation of full and structural words in Chinese requires two versions. An agglutinative word structure model, such as the Turkic one, provides a tree-structured database and special procedures of access. This model makes it possible to generate, using a computer, the correct agglutinate word forms for some Turkic languages.

Acknowledgements

We gratefully acknowledge the contribution to this work by Dr. Leonid Kogan, Rinat Minvaleev of St.Petersburg and Dr. Zhao Zhe of Changsha, as well as by other colleagues from the international Speech Statistics Group in Baky, Almaty, Shimkent and Tashkent.

REFERENCES

Alam, Y. S. A. 1983. two-level morphological analysis of Japanese. *Texas Linguistic Forum.* 22:229-252.

Andrezen, Vladimir/ Kogan, Leonid/ Kwiatkowski, Wladimir/ Minvaleev, Rinat/ Piotrowski, Rajmund/ Shumovsky, Vladislav/ Tioun, Elena/ Tovmach, Yurij. 1992. Automatic dictionary organization in NLP systems for Oriental languages. *Actes de Coling-92, Nantes, 23-28 aout 1992.* Nantes: Geta (IMAG) Association Champollion, 505-509.

Beesley, K. 1998. Arabic morphology using only finite-state operations Proceedings of the Workshop on Computational Approaches to Semitic Languages, Montreal, 50-57.

Bektaev, Khaldybai. 1978 *Statistic and informational typology of Turkic text* (in Russian). Alma-Ata: Nauka.

Blekhman, M., Bezhanova, O., Kursin, A., Rakova, A. 2000. Translation Software by Lingvistica '98 Inc.: The PARS Family of Machine Translation Systems. *International Journal of Translation.* 12:1-2.

Kataja, L. and Koskenniemi, Kimmo. 1988. Finite state description of Semitic morphology. In *COLING-88:* Papers Presented to the 12th International Conference on Computational Linguistics, Vol. 1:313-315.

Kiraz, George Anton. 2000. Multitiered Nonlinear Morphology Using Multiple Finite Automata: A Case Study on Syriac and Arabic. *Computational Linguistics.* 26.1: 77-105

Koskenniemi, Kimmo. 1986. Morfologisten kaksitasosääntöjen kääntäminen äärellisiksi automaateiksi. *STeP-86 Symposium Papers: Methodology* (Finnish Artificial Intelligence Symposium/Suomen Tekoälytutkimuksen Päivät). Otaniemi, Espoo, Finland. August 19-22, 1986. Otaniemi: Otapaino, 234

Makhmudov, Masud Akhmed ogly. 1982. *Elaborating a morphological analysis system for Turkic word form.* Illustrated by the Azerbaijanian text material (in Russian). Baky: Academy of Sciences of the Azerbaijani SSR

McCarthy, J. and Prince, A. 1990. Foot and word in prosodic morphology: The Arabic broken plural. *Natural Language and Linguistic Theory,* 8:209-283.

Mukhamedov, Sabit, Piotrowski, Rajmund. 1986. *Language engineering and systemic-statistical studies of Uzbek text* (in Russian). Tashkent: Fan.

Oflazer, Kemal. 1994. Two-level Description of Turkish Morphology. *Literary and Linguistic Computing.* 9.2:137-148.

Piotrowski, Rajmund. 1999. Linguistic automaton and its verbal-mental foundation (in Russian). Minsk: MGLU.

Seewald, Uta. 1994. Maschinelle morpho-semantische Analyse des Französischen 'Morze'. Eine Untersuchung am Beispiel des Wortschatzes der Datenverarbeitung. Tübingen: Max Niemeyer Verlag,.

Sproat, Richard Shih, Chilin, Gale, William, Chang, Nancy. A 1996. Stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics.* 22. 3:377-404.