Computational Linguistics and Chinese Language Processing
vol. 2, no. 2, August 1997, pp. 47-62.

Segmentation Standard for Chinese Natural Language Processing

Chu-Ren Huang*, Keh-jiann Chen+, Feng-yi Chen+, and Li-Li Chang+

Abstract

This paper proposes a segmentation standard for Chinese natural language processing. The standard is proposed to achieve linguistic felicity, computational feasibility, and data uniformity. Linguistic felicity is maintained by a definition of segmentation unit that is equivalent to the theoretical definition of word, as well as a set of segmentation principles that are equivalent to a functional definition of a word. Computational feasibility is ensured by the fact that the above functional definitions are procedural in nature and can be converted to segmentation algorithms as well as by the implementable heuristic guidelines which deal with specific linguistic categories. Data uniformity is achieved by stratification of the standard itself and by defining a standard lexicon as part of the standard.

1. Introduction

One important feature of Chinese texts is that they are character-based, not word-based. Each Chinese character stands for one phonological syllable and in most cases represents a morpheme. The fact that Chinese writing does not mark word boundaries poses the unique question of word segmentation in Chinese computational linguistics (e.g. Sproat and Shih 1990, and Chen and Liu 1992).[1] Since words are the linguistically significant basic elements that are entered in the lexicon and manipulated by grammar rules, no language processing can be done unless words are identified. This applies to psychological studies as well (Zhou et al. 1992, and Bates et al. 1993). In theoretical terms, the successful establishment of a segmentation standard means that word boundaries are psychologically real in Chinese and hence verifies the status of a

* Institute of Linguistics, Academia Sinica, Nankang, Taipei, Taiwan, ROC. E-mail: hschuren@ccvax.sinica.edu.tw

+ Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan, ROC.

1. As pointed out by a reviewer of CLCLP, languages such as Japanese and Thai have segmentation problem, too. However, the Chinese language has a homogeneous writing system composed of Chinese characters (i.e. Hanji), thus rich writing system of Japanese, including hanji, hiragana, and katagana, encodes partial word segmentation information already. In other words, Chinese poses the unique problem of segmentation without any explicitly encoded boundary information.

word as a primary linguistic construct. The primacy of the concept of word can be more firmly established if its existence can be empirically supported in a language that does not mark it conventionally in texts. In computational terms, no serious Chinese language processing can be done without segmentation. No efficient sharing of electronic resources or computational tools is possible unless segmentation can be standardized. Evaluation, comparisons and improvements, are also impossible in Chinese computational linguistics without standardized segmentation.

Since the proposed segmentation standard is intended for Chinese natural language processing, it is very important that it reflects linguistic reality as well as computational applicability. In other words, there are two possible pitfalls that we must avoid. The first is when the standard is a set of ad hoc rules that allow clean and straightforward computational solution but do not consistently define units of linguistic information. The second is when the standard is a set of abstract linguistic concepts that do not lend themselves to any consistent prediction of segmentation units when applied to natural language processing. Hence we stipulate that the proposed standard must be linguistically felicitous, computationally feasible, and must ensure data uniformity.

1.1 Components of the Segmentation Standard

Our proposed segmentation standard consists of two major components to meet the goals discussed above. The modularization of the components will also facilitate revisions and maintenance in the future.

The two major components of the segmentation standards are the segmentation criteria and the (standard) reference lexicon. The tripartite segmentation criteria consist of a definition of segmentation unit, two segmentation principles, and a set of heuristic guidelines. The segmentation lexicon contains a list of Mandarin Chinese words and other linguistic units that the heuristic guidelines must refer to, hence the name reference.

In what follows, we will introduce the definition, the principles, the guidelines, and the lexicon in different sections. We will also define the levels of application of segmentation. A set of linguistically interesting data will be studied to illustrate the standard, and comparisons between our proposal and the national standard of the People's Republic of China will be discussed before the final concluding section.

2. Segmentation Standard Part I: Segmentation Criteria

2.1 A Definition of the Segmentation Unit

Given Bloomfield's (1933) definition of words as `the smallest units of speech that

can meaningfully stand by their own,' they should be the most natural units for segmentation in language processing as well. However, as Chao (1968) observes, sociological words and linguistic words very often do not match up. Chao's bifurcation can be further elaborated according to more recent developments of linguistic theories. In Sproat et al.'s (1996) succinct discussion, the term orthographic words is roughly equivalent to Chao's sociological words. In addition, it is observed that the notion of linguistic words can be further refined to include at least the following five notions: phonological word, morphological word, syntactic word, semantic word, and lexical word. Each different notion of word conceptually differentiates a set of significant linguistic units. Since adopting different notions of words will lead to different segmentation results, we need to examine the entailed segmentation results to decide on which notion of words is the most appropriate.

Recall that in computational linguistic terms, the primary goal of segmentation is to identify units to access lexical information (i.e. dictionary lookup). This is parallel to the psycholinguistic assumption of words as units of lexical access and acquisition. Also recall that in a modular representation of grammatical information, the lexicon is the only location where knowledge of different modules exist simultaneously given that the essence of modular representaton requires that grammatical information not be accessible from other modules. The above assumptions require that segmentation units be useful in accessing all linguistic information: phonological, morphological, syntactic, and semantic. This will be the premise of our evaluation of the different notions of linguistic words. First, phonological words are defined as the domain of application of phonological rules. Hence they are natural units in applications such as text-to-speech. However, a phonological word often involves more than one syntactic or semantic units, thus parsing and interpretation will be difficult if segmentation reflects phonological words only. Second, even though syntactic words as smallest unit in syntax seems to be a good candidate for segmentation, the necessity for lemmatization in many languages attests to the fact that some units that cannot occur independently in syntax may have independent grammatical function and meaning and need to be treated as basic units in language processing (e.g. Sproat 1992). Last, similarly, morphological and semantic words also focus on only one aspect of linguistic behavior and cannot be the optimal unit for lexical access. In sum, we found that the notion of words must integrate the modular knowledge of phonology, morphology, syntax, and semantics. The lexicon as the knowledge-base of all linguistic knowledge is exactly the locus where such an integrated notion of words exist. Hence we propose that lexical words are the optimal notion for defining segmentation units. Lexical words are defined as entries in the lexicon of each language. They will not always coincide with the notions of phonological, morphological,

syntactic, semantic words etc. However, a lexical word will contain enough information such that boundaries of all other linguistic words, e.g. phonological, morphological etc., can be surmised. Segmentation units is thus defined as the optimal unit of linguistic information.

Since linguistic modules may interface differently in grammars of different languages, the above position entails that compositions of lexical words may vary from language to language. In other words, lexicon and thus segmentation units may require language-dependent rules to identify. In English, a sociological/ orthographic word can be defined by the delimitation of blanks in writing. It is nevertheless not uncommon for a lexical word such as a compound to be composed of more than one sociological words, such as `the White House.' Since these cases represent only a relatively small portion of English texts, it has been uncontroversial to take the orthographic marking as the default while identify the idiosyncratic words with additional morpho-lexical processes in computational linguistics. In other words, sociological words are taken as the default standard for segmentation units as well as a reasonable approximation to lexical words in English natural language processing.

Chinese, on the other hand, takes characters as its sociological/orthographic words. It is worth noting that Chinese words may be made up of one or more characters. In terms of types of lexical entries, one-character words represent only slightly less than 10% of all entries (in comparison, two-character words take up more than 65% of lexical entries). In terms of tokens, one-character words are estimated to represent roughly 50% of all words in Chinese (Chen et al., 1993). Since the notion of sociological word (i.e. one-word-per-character) is not a good working hypothesis for lexical words, and since there is no fixed length for words, a crucial step is to take the definition of lexical words directly as the standard for segmentation unit.

We follow the above findings and define the standard segmentation unit as a close approximation of lexical words with emphasis on functional rather than phonological or morphological independence.

(1) **Segmentation Unit**$_{\text{def}}$ is the smallest string of character(s) that has both an independent meaning and an identifiable and constant grammatical function.

There are three points worth remarking involving the above definition. First, no technical linguistic terms are used. Even though we risk being imprecise, the choice of non-technical terms is deliberate such that even developers in information industries with little or no linguistic background could follow this standard. Second, it follows from this definition that most of the so-called particles will be treated as segmentation units. They

include *le* 了 'perfective marker', and *de* 的 'relative clause marker' etc. These particles show various levels of linguistic dependencies but represent invariant grammatical functions. Lastly, homomorphic words that are either syntactically or semantically ambiguous (i.e. has more than one syntactic categories or meanings) will be segmentation units. In other words, each unique orm/meaning/syntactic-function pairing will be a segmentation unit, even though segmentation result can only show form differences and not meanging/function variations.

2.2 Segmentation Principles

Based on the definition of segmentation units, we propose two segmentation principles to elaborate on how the two crucial elements, i.e. independent meaning and constant grammatical function, can be determined. The principles also provide a functional/procedural algorithm for identifying segmentation units.

(2) **Segmentation Principles**
(a) A string whose meaning cannot be derived by the sum of its components should be treated as a segmentation unit.
(b) A string whose structural composition is not determined by the grammatical requirements of its components, or a string which has a grammatical category other than the one predicted by its structural composition should be treated as a segmentation unit.

Notice again that non-technical terms are chosen whenever possible so that the standard can be followed by people of different backgrounds. This definition has been examined and accepted by a work-task committee, more than half of whose members come from non-linguistic background. Whether it will actually be effective among non-technical users remains to be tested in large-scale implementation. Also take note that characters are the basic processing units we start with when segmentation is involved. Thus the two principles address the question of which strings of characters can be further combined to form a segmentation unit. Principles (2a) and (b) elaborate on the semantic (independentmeaning) and syntactic (constant function) components of the definition of segmentation unit. Because of their procedural nature, they also provide the basis for segmentation algorithm. The conversion to actual segmentation process can be illustrated with the two conditions in (2b). Since a character could be a lexical or sub-lexical element, the basic decision in segmentation is whether the relation between two characters are morpholexical or syntactic. With a VN sequence such as *lai-dian* 來電 come-electricity 'to strike a chord with, to mutually attract', the first part of principle (2b) applies to predict that it is a segmentation unit since *lai* 來 is an intransitive verb and cannot take an object. With VV sequences such as [*churu*]n 出入 exit-enter 'discrepancy' and [*kushi*]vt 哭濕 cry-wet 'to cause to become wet [by shedding tears on]',

the second part of principle 2b) predicts that they are segmentation units since their respective categories, noun and transitive verb, cannot be inherited from the conjunctive compound structure.[2]

2.3 Segmentation Guidelines

Even though the above principled ways of defining segmentation units provide a broad direction for standardized segmentation, they lack the nuance for guiding actual segmentation. The definition of segmentation units and the segmentation principles are essentially language independent formalization of information units (i.e. words). Thus they will not vary with linguistic change, and need not be revised for specific applications. However, this universal nature also prevents them from referring to specific details. This is most obvious when the actual data does not allow a clearcut theoretical classification. Hence we propose that a set of Segmentation Guidelines be included in our segmentation standard to reflect heuristic knowledge that is dependent on actual linguistic data. In other words, these guidelines can be added, deleted, or altered as necessitated by the kind of linguistic data we are dealing with. Since all essential linguistic knowledge is encoded in the lexicon, it follows that the guidelines will have to refer to a Mandarin lexicon. In contrast, the broad linguistic concepts in the definition and principles do not refer to specific lexical information. Last, we also envision that the guidelines are heuristic and quantifiable. They are heuristic because segmentation decisions depend on consulting the lexical information listed in the reference lexicon, and because fulfilling the conditions of one guideline alone does not necessarily qualifies a string as a segmentation unit. It is quantifiable since a string is more likely to be a segmentation unit when it satisfies the requirements of more guidelines.

(3) **Segmentation Guidelines**

(a) Bound morphemes should be attached to neighboring words to form a segmentation unit when possible.

(b) A string of characters that has a high frequency in the language or high co-occurrence frequency among the components should be treated as a segmentation unit when possible.

(c) String separated by overt segmentation markers should be segmented.

(d) Strings with complex internal structures should be segmented when possible.

3. Segmentation Standard Part II: The Reference lexicon

2. As observed by a CLCLP reviewer, guideline 2a) also applies to *lai-dian*, since its meaning is not compositional. It is also worthwhile to note that *kushi* illustrates why Li's (1990) a priori assumption that VR compounds are headed by V fails. That *kushi* is transitive cannot be predicted from the property of the intransitive verb *ku*.

As mentioned above, the reference lexicon is so-called because both segmentationprinciples and guidelines must refere to it. Entries in this lexicon, i.e. lexical words or lexemes, should include non-derivational words as well as productive morpho-lexical affixes. It will also contain the list of mandatory segmentation markers, such as the end of sentence marker (.), (o) etc. It is obvious that bound morphemes (including derivational and inflectional affixes) and segmentation markers can only be standardized when they are exhaustively listed in a lexicon. With appropriate morpho-lexical information attached, these entries will also cover all derivational processes. Non-derivational words, on the other hand, are trickier. Since neither their forms nor their meanings can be predicted with generative rules, the only way to verify that they are segmentation units is to consult a lexical list. However, neologism constantly add new forms and meanings to words in a language and old forms and meanings do become obsolete. In other words, the lexicon of a language is always in a flux and a reference lexicon that faithly reflects the current states of the language is extremely difficult if not impossible to maintain. We will deal with the diffulty of updating the reference lexicon later in this paper. We will first postulate that a reference lexicon be the basic knowledge-base of the segmentation standard, where all algorithmatic rules must refer to.

The definition of the reference lexicon, i.e. the theoretical models determining how the entries are selected, calls for a separate paper to explicate. It suffices to unerline here that selection of lexical entries must meet both the necessary conditions of the segmentation standard and the sufficient conditions defined on real language use, i.e. an entry is included only when it qualifies as a segmentation unit. The segmentation definition and principles are the same definition and principles that entries in the reference lexicon must conform to. And guidelines (3a)-(3c) are also useful heuristic guidelines for selecting lexical entries. The crucial issue here is then what prevents the proposed standard from being vacuously circular, since the basic reference knowledge base for the standard is also governed by the standard.

The answer lies in that the reference lexicon must be compiled empirically based on data of actual language use. In other words, with the selection of each form-meaning pair as an entry in the lexicon, we are solving the empirical question of whether a certain form or meaning exits in a language. In order for this solution, as well as the whole lexicon, to be scientifically sound, it is crucial that the decision be verifiable empirically. Since the actual use of any language cannot be enumerated within finite time, the empirical verification must be done based on a reliable sampling of the language, i.e. a reference corpus. Actually, that the same abstract principles and guidelines apply is expected since we approached the segmentation problem by identifying the definition of segmentation

units with that of lexical words.

A reference corpus is a corpus that represent the core uses of current language uses. In other words, generalizations extracted from the reference corpus should be applicable generally to the language. As mentioned above, the emiprical question of whether a lexical form exists in a language cannot be reliably anwsered without reliable corpus data. Thus, our reference corpus will be balanced (Chen et al. 1996) to represent different genres, styles, topics etc. Entries of the reference lexicon must be extracted from the reference corpus by a set of heuristic principles, not by the arbitrary decision of any human. The reference corpus must be periodically updated and renewed to reflect constant language changes and lexical shifts, such that new words and new usages can be empirically determined. Note that a corpus is critical to the segmentation standard since information such as frequency or collocational frequency must be obtained from a corpus. Changes in such distributional attributes of the language can also be easily traced by monitoring different versions of the reference corpora. After being exhaustively segmented according to the segmentation standard, the reference corpus will also serve as the testing and/or training data for segmentation algorithms developed according to the segmentation standard.

4. Three Levels of Segmentation Standard

A central concern in proposing any standard is whether this standard can be successfully and consistently followed. To put it more bluntly, a standard, regardless of its theoretical value, is meaningless unless it can be consistently followed. We took into consideration of the state of art of automatic segmentation in Chinese NLP as well as the technology level of information industries dealing with Chinese natural languages and proposed the following stratification of three levels of instantiations for the Segmentation Standard. It is hoped that this stratification will ensure successful standardization as well as lead to eventual identification of segmentation units with linguistics words.

(4) **Levels of Segmentation Standard**

(a) **Faithful[*xin4*]** 信 : All segmentation units listed in the reference lexicon should be successfully segmented. This will be the default segmentation level for the exchange of electronic texts.

(b) **Truthful[*da2*]** 達 : All segmentation units identified at the Faithful level as well as all segmentation units derivable by morphological rules should be successfully segmented. This will be the level for most natural language processing applications.

(c) **Graceful[*ya3*]** 雅 : All linguistic words are successfully identified as segmentation units. This is the ideal goal of segmentation and will be the segmentation level for fully automated language understanding.

The names of the three levels of standard are adopted from the three levels of translation described by Yan Fu, the first major Chinese translator of Western texts. In the original usage, *xin*4 means that all the elements of the original text are faithfully represented, *da*2 means that the meaning of the original text is truthfully transferred, and *ya*3 means all the literary nuances, including metaphors, stylistic variations, etc., are gracefully preserved. We follow the spirit of this division and give it new interpretation in terms of segmentation for NLP. The goal of the Faithful level is to define a segmentation standard such that uniformity of electronic texts can be achieved even when they are prepared with the lowest possible computational sophistication. In other words, the standard must be as easy to follow as the convention of inserting blanks at wordbreaks in English text processing. Thus we stipulate that the Faithful standard requires only that all entries in the reference lexicon be properly segmented. Thus, unless an entry is listed in the lexicon, a string will simply be segmented by individual characters. Notice that this is NOT a trivial level since possible ambiguous segments take up as high as 25% of Chinese texts (Chen and Liu 1992). For instance, the string *ba3 shou3* 把手 has an entry meaning 'handle,' but it could also be segmented as two units 'prep.+hand' depending on its context. We believe that reasonably high consistency of ambiguity resolution can be achieved since unknown words, i.e. words not listed in the lexicon, are not involved. Various automatic segmentation programs have reported over 96% precision rate when unknown words are not taken into account (Chen and Liu 1992, Chiang et al.1992).

The goal of the Truthful level is to define a segmentation standard for most computational linguistic applications. The coverage of the Faithful level is too low for most NLP applications. For instance, unknown words can be left unidentified for data exchange but not for machine translation. Unknown words can be classified into three types of words that cannot be listed in the lexicon (Wang et al. 1995). The first type are the words that are generated by morphological rules. They are productive and cannot be exhaustively listed in the lexicon. The second type are the derived words whose derivation is either context-dependent or do not seem to fall into the more familiar types of morphological rules. A good example is the *suoxie* 縮寫 abbreviation where a character from each compound or phrase component is selected to form a new word (Huang et al. 1993), such as deriving *hua2hang2* 華航 from *zhong1hua2 hang2kong1* 中華航空 'China Airlines.' The third type are the unknown words which are not derived by any rules. Proper names in Chinese are a good example of this type since any characters in the language can be used in a given name (Chen et al. 1994, and Sun et al. 1994). We feel that only the first type of unknown words can be comfortably dealt with by current Chinese NLP technology; while more in-depth linguistic research need to be

carried out on the last two types of unknown words to identify generalizations for automatic language processing. Thus, at the Truthful level of segmentation, we stipulate that all lexical entries as well as all morphologically derivable unknown words should be properly segmented. The applicable morphological rules will be exhaustively listed in the reference lexicon under the affixes involved (following the theoretical architect of LFG and HPSG). This level will offer a wide enough coverage for most NLP applications and yet a reasonably high consistency can be achieved with current automatic segmentation technology. Since a finite state machine simulating the morphological rules on top of a finite lexicon listing can easily generate all the segmentation units, the only technical challenge would be to resolve ambiguities among the above units.

Lastly, the Graceful level of segmentation standard will have to deal with the two remaining types of unknown words, i.e. the suoxie type and the type which are not derivable from morphological rules. Current researchers are already tackling some of the problems involved in these two types of unknown words. It may not be too long before the research matures and reasonable consistency can be achieved at this level of standard for fully automated language understanding.

5. Illustration

In this section, we will discuss two difficult cases for segmentation and show how our Segmentation Standard offers straightforward solutions.

5.1 Telescopic Compounds

We refer to the first set of data as telescopic compounds. They are conjunctive compounds with internal ellipsis. What makes them even harder than other compounds to segment is that the elliptical parts are simply the elements that two conjuncts share regardless of their morpho-syntactic status. In (6), we show that the 'folded' (i.e. shared) part of the compound could be the ending 6(a), the beginning 6(b,c) or both the ending and the beginning (6d).

(6) Telescopic Compounds
(a. 父母親 fu-mu-qin fa-mo-ther 'father and mother, parents'
(b. 青少年 qing-shao-nian green-little-age 'youths [qingnian] and teenagers [shaonian]'
(c. 青少女 qing-shao-nu green-little-woman 'young women [*qingnu] and teen-age girls [shaonu]'
(d. 中山南北路 Zhongshan-nan-bei-lu Zhongshan-south-north-road 'South Zhongshan road [Zhongshannanlu] and North Zhongshan road

[Zhongshanbeilu]'

The definition of segmentation unit and segmentation principles do not offer clearcut result for the telescopic compounds. Even though they seem to be semantically and syntactically compositional, their composition is atypical since some of the constituents are missing. Thus we have to rely on the applicable heuristic guidelines (4a) and (4b). From (4a), we find that, if segmented, these compounds will leave dangling bound forms, such as *qin*, 青 *qing* 親 , etc. From (4b), we find that these compounds occur frequently and the MI values between the components are higher than 2 (Sproat and Shih 1990). Thus, the guidelines indicate that these compounds are segmentation units. Whether they will be segmented at the Faithful or higher levels depend on if a specific compound is frequent enough to be listed in the lexicon.

On the other hand, these compounds sometimes occur with segmentation markers between characters, such as *qing,shaonu* 青、少女 . In this case, guideline (4d) applies at the two lower levels of standard and the compounds will be segmented at the marker. This ensures computational feasibility and allows the solution of the difficult question of incorporating segmentation markers as part of a word to be postponed for later work.

5.2 Strings Containing Foreign Words

Strings containing foreign words and/or other non-Chinese character symbols are common in electronic textual data nowadays. These may or may not be words. Even if the string in question is a word, it is often not listed in the monolingual Chinese dictionary that a segmentation standard refers to. Listing all foreign words in a standard lexicon is of course impractical. There is a very practical solution provided by our segmentation standard though.

(3) **Segmentation Guidelines**
(c) String separated by overt segmentation markers should be segmented.

(3c) stipulates that overt segmentation markers should be followed. We consider code-switching (i.e. switching from one language to another) as clear and overt segmentation marker. Thus, all foreign words, as well as mathematic or scientific symbols, will be segmented from the neighboring Chinese words. Once these foreign word strings are segmented, special lexicons could be referred to for lookup. These words include an English lexicon, or a lexicon of computer science terms. Similarly, mathematic or scientific equations, as well as arabic numerals, will be segmented and dealt with in a different module. Last, but not the least, there are growing uses of code-mixing even at

the morphological level. For instance, the following sequence is used as a unit in Taiwan Mandarin: *k-shu* K 書 'to hit the book'. Our claim is that this item has already been lexicalized and has to be listed in the lexicon. Thus it should be identified as a word and not governed by (3c).[3]

6. Comparison to the PRC National Standard for Segmentation for Chinese Information Processing

The Segmentation Standard proposed here originated from the Segmentation Standard adopted by the Computational Linguistic Society of R.O.C. for the NLP research community in Taiwan in 1989. The current standard integrates the experience this research group gained since then by manually tagging a 5 million word corpus and compiling a 80 thousand entry lexicon. It also incorporates discussions with three working groups composed of linguists, computer scientists, and information industrialists respectively. This proposal is being submitted as a draft for national standard to the government of R.O.C.

During the same time period, scholars in mainland China started their discussion of a segmentation standard in 1987 (Liang 1989). A draft of the standard was publicized in 1990 (Liang 1991). A national standard, i.e. GB13715, was announced and implemented in 1993 (Liu et al. 1994).

Given the geo-political differences, it may be impossible to unify the two proposed standards in the near future. Even if a unified standard is reached, it would still be necessary to maintain separate lexicons to reflect the widening differences between words used on both sides of the Taiwan Strait. However, from a purely academic point of view, the two sets of proposals do represent very different design philosophies. A comparative study could shed light on future development of standards for information processing.

It is interesting to note that both standards clearly specify that they are designed for natural language processing but stipulate their relation with the linguistic notion of word differently. The PRC standard underlines that a segmentation unit is different from a (linguistic) word and says nothing more about it. Our current standard takes a version of the definition of word as the definition of a segmentation unit. Our principles and

3. For instance, the newest edition of Xiandai Hanyu Cidian includeed 39 entries that start with a Western alphabet, though not in the main body of the dictionary. Recognition of the fact that mix-coding is allowed at the lexical level poses a dilemma for Chinese lexicography. That is, the language-specific and more informative layout of lexicon based on Chinese characters connot accomodate these entries.

guidelines are motivated by this definition, even though the three-level implementation of the standard allows deviation from the theoretical notion for most current practical applications. We think our approach is better equipped to deal with possible conflictsamong rules, to accommodate novel data, and to adapt to future technological and theoretical advancements.

First, our approach has a unifying definition which can resolve possible conflicts in lower-level heuristic rules (i.e. the segmentation guidelines). On the other hand, all the rules in the PRC standard are same-level application rules, thus it would be difficult to resolve possible conflicts in a non-ad-hoc way that would affect rule interaction.

Second, our approach can easily account for new data. The PRC standard would call for additional local rules in order to account for facts not previously specified in the standard, such as the telescopic compounds discussed above. In our proposal, no addition of rules will be necessary. The high-level definition and principles should cover all segmentation facts conceptually, while the low-level guidelines, especially the use of frequency, should apply to all segmentation data.

Third, our three-level implementation allows us to easily change with the future development of computational technologies or linguistic theories. We have set an ideal level of segmentation standard where segmentation units can be unified with linguistics words. By adding to the Truthful level any previously unsolved linguistic facts whenever the technology is mature enough, we will be able to keep improving our segmentation standard with the development of Chinese computational linguistics. In the mean time, the Faithful value will ensure that a basic level of electronic data exchange is always consistently maintained. The PRC standard did its best to stipulate the current states, but will have problem being exhaustive or always up-to-date.

Last but not the least, continuous maintenance and updating of the reference lexicon is crucial to the reusability of the segmentation standard. This is a crucial prerequisite for the implementation of our segmentation standard as well as a lesson learned by the less then successful implementation of the PRC standard. The research group of Liang et al. has disbanded after the successful application of the PRC national standard GB/T 13715-92. However, since the reference lexicon is the crucial basis for any segmentation algorithm where lexical changes are registered and accounted for, it needs to be maintained and updated continuously. Even though other research groups have proposed principled methods to update the original small lexicon (e.g. Sun and Zhang 1997, Lin and Miao 1997), the discontinuity has made it quite difficult for wider and practical adoption of the standard. Thus we emphasize that a segmentation standard must also include a standard reference lexicon shared by the NLP community as well as a mech-

anism for periodical and continuous updates.

7. Concluding Remarks

In this paper, we propose a Segmentation Standard for Chinese language processing. We propose that the standard should be composed of two distinct parts: (a) the language and lexicon-independent definition and principles, and (b) the lexicon-dependent guidelines. The definition and principles offer the conceptual basis of segmentation and will be the unifying idea to resolve possible local heuristic conflicts. The lexicon-dependent guidelines as well as the data-dependent lexicon allows the standard to be easily adaptable to linguistic as well as sub-language changes.

Bibliography

Bates, E., S. Chen, P. Li, M. Opie, O. Tzeng. "Where is the Boundary between Compounds and Phrases in Chinese?" *Brain and Language*, 45 (1993):94-107.

Bloomfield, L. *Language*. New York: Holt, Rinehart, and Winston, 1993.

Chang, J.-S., S.-D. Chen, S.-J. Ker, Y. Chen, and J. S. Liu. "A Multiple-Corpus Approach to Recognition of Proper Names in Chinese Texts." *Computer Processing of Chinese and Oriental Languages*. 8.1 (1994): 75-85.

Chao, Y. R. *A Grammar of Spoken Chinese*. Berkeley:University of California Press, 1968.

Chen, C.-Y., S.-F. Tseng, C.-R. Huang and K.-j. Chen. "Some Distributional Properties of Mandarin Chinese -- A Study Based on the Academia Sinica Corpus." *Proceedings of the First Pacific Asia Conference on Formal and Computational Linguistics*. Taipei, 1993, pp. 81-95.

Chen, H.-H., and C.-C. Li. "Recognition of Text-based Organization Names in Chinese." [In Chinese.] *Communications of COLIPS*. 4.2 (1994):131-142.

Chen, K.-J. and S.-H. Liu. "Word Identification for Mandarin Chinese Sentences." *COLING-92*, Nantes, France, 1992, pp. 101-105.

____, C.-R. Huang, L.-P. Chang, and H.-L. Hsu. "SINICA CORPUS: Design Methodology for Balanced Corpora." In B.-S. Park and J.-B. Kim Eds. *Language, Information, and Computation*. *Selected Papers from the 11th PACLIC*. Seoul: Kynung Hee University, 1996, pp. 167-176.

Chiang, T.-H., J.-S. Chang, M.-Y. Lin, and K.Y. Su. "Statistical Models for Word Segmentation and Unknown Word Resolution." *Proceedings of ROCLING V*, 1992, pp.121-146.

Chinese Knowledge Information Processing Group. *ShouWen JieZi - A Study of Chinese Word*

Boundaries and Segmentation Standard for Information Processing [In Chinese]. CKIP Technical Report 96-01. Taipei: Academia Sinica, 1996.

____ *A Frequency Dictionary of Written Chinese*. CKIP Technical Report no. 94-01. Taipei: Academia Sinica, 1994.

____ The CKIP Categorical Classification of Mandarin Chinese (In Chinese). CKIP Technical Report no. 93-05. Taipei: Academia Sinica, 1993.

Church, K., and P. Hanks. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics*. 16.1 (1990):22-29.

Huang, C.-R. "The Morpho-lexical Meaning of Mutual Information: A Corpus-based Approach Towards a Definition of Mandarin Words." Presented at the 1995 Linguistics Society of America Annual Meeting. January 5-8. New Orleans, 1995.

____, K. Ahrens, and K.-J. Chen. "A Data-driven Approach to Psychological Reality of the Mental Lexicon: Two Studies in Chinese Corpus Linguistics." *Proceedings of the International Conference on the Biological Basis of Language*. Chiayi: Center of Cognitive Science, National Chung Cheng University, 1993, pp. 53-68. Revised Version to Appear in *Bulletin of the Institute of History and Philology*. 1998.

____, K.-J Chen. F.-Y. Chen, W.-J. Wei, and L. Chang. "The Design Criteria and Content of the Segmentation Standard for Chinese Information Processing." [in Chinese]. *Yuyan Wenzi Ying-yong*. 1 (1997): 92-100.

Institute of Linguistics, Chinese Academy of Social Science, ed. *Xiandai Hanyu Cidian*. Revised Version. Beijing: Shangwu, 1996.

Li, Y. "On V-V Compounds in Chinese." *Natural Language and Linguistic Theory*. 8 (1990): 177-207.

Liang, N.-Y. "Research on Automatic Segmentation of Written Chinese and its Future Developments." [In Chinese.] Jisuanji Xinxibao**,** 1989.

Lin, X.G., and C.J. Miao. "Guifan+Cibiao yu Jinyen+Tongji." *Yuyan Wenzi Yingyong*. 1 (1997):87-91.

Liu, Y., Q. Tan, and X. Shen. *Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology*. Beijing: Qinghua University Press, 1994.

Sproat, R. *Morphology and Computation. Cambridge*: MIT Press, 1992.

____ and C. Shih. "A Statistical Method for Finding Word Boundaries in Chinese Text." *Computer Processing of Chinese and Oriental Languages*. 4.4 (1990):336-351.

____ C. Shih, W. Gale, and N. Chang. "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese." *Computational Linguistics*. 22.3 (1996):377-404**.**

Sun, M.S., C.N. Huang, H.Y. Gao, and J. Fang. "Automatic Recognition of Chinese Names." *Communications of COLIPS*. 4.2 (1994): 113-122.

____ and L. Zhang. "Renjibingcun, Zhiliangheyi -tantan zhiding xinxi chuliyong hanyu cibiao de celue." *Yuyan Wenzi Ying-yong*. 1 (1997):79-86.

Wang, M.-C., C.-R. Huang, and K.-J. Chen. "The Identification and classification of Unknown Words in Chinese: A N-gram-Based Approach." In A. Ishikawa and Y. Nitta Eds. *The Proceedings of the 1994 Kyoto Conference. A Festschrift for Professor Akira Ikeya.* Tokyo: The Logico-Linguistics Society of Japan, 1995, pp. 113-123.

Zhou, X., R. Ostrin and L. Tyler. "The Noun-Verb Problem and Chinese Aphasia: Comments on Bates et al. (1991)." *Brain and Languag*e, 45 (1993):86-93.

Acknowledgement

Research reported in this paper is partially supported by the Standardization Bureau of Taiwan, ROC. The authors are indebted to the following taskforce committee members for their invaluable contribution to the project: Claire H.H. Chang, One-Soon Her, Shuan-fan Huang, James H.Y. Tai, Charles T.C Tang, Jyun-shen Chang, Hsin-hsi Chen, Hsi-jiann Lee, Jhing-fa Wang, Chao-Huang Chang, Chiu-tang Chen, Una Y.L. Hsu, Jyn-jie Kuo, Hui-chun Ma, and Lin-Mei Wei. We would like to thank the three CLCLP reviewers for their constructive comments. We are also indebted to our colleagues at CKIP, Academia Sinica for their unfailing support as well as helpful suggestions. Any remaining errors are, of course, ours.