One Tokenization per Source

Jin GUO
Kent Ridge Digital Labs
21 Heng Mui Keng Terrace, Singapore 119613

Abstract

We report in this paper the observation of *one tokenization per source*. That is, the same critical fragment in different sentences from the same source almost always realize one and the same of its many possible tokenizations. This observation is demonstrated very helpful in sentence tokenization practice, and is argued to be with far-reaching implications in natural language processing.

1 Introduction

This paper sets to establish the hypothesis of *one tokenization per source*. That is, if an ambiguous fragment appears two or more times in different sentences from the same source, it is extremely likely that they will all share the same tokenization.

Sentence tokenization is the task of mapping sentences from character strings into streams of tokens. This is a long-standing problem in Chinese Language Processing, since, in Chinese, there is an apparent lack of such explicit word delimiters as white-spaces in English. And researchers have gradually been turning to model the task as a general lexicalization or bracketing problem in Computational Linguistics, with the hope that the research might also benefit the study of similar problems in multiple languages. For instance, in Machine Translation, it is widely agreed that many multiple-word expressions, such as idioms, compounds and some collocations, while not explicitly delimited in sentences, are ideally to be treated as single lexicalized units.

The primary obstacle in sentence tokenization is in the existence of uncertainties both in the notion of words/tokens and in the recognition of words/tokens in context. The same fragment in different contexts would have to be tokenized differently. For instance, the character string *todayissunday* would normally be tokenized as "*today is sunday*" but can also reasonably be "*today is sun day*".

In terms of *possibility*, it has been argued that no lexically possible tokenization can not be grammatically and meaningfully realized in at least some special contexts, as every token can be assigned to bear any meaning without any orthographic means. Consequently, the mainstream research in the literature has been focused on the modeling and utilization of local and sentential contexts, either linguistically in a rule-based framework or statistically in a searching and optimization set-up (Gan, Palmer and Lua 1996; Sproat, Shih, Gale and Chang 1996; Wu 1997; Guo 1997).

Hence, it was really a surprise when we first observed the regularity of *one tokenization per source*. Nevertheless, the regularity turns out to be very helpful in sentence tokenization practice, and to be with far-reaching implications in natural language processing. Retrospectively, we now understand that it is by no means an isolated special phenomenon but another display of the postulated general law of *one realization per expression*.

In the rest of the paper, we will first present a concrete corpus verification (Section 2), clarify its meaning and scope (Section 3), display its striking utility value in tokenization (Section 4), and then disclose its implication for the notion of words/tokens (Section 5), and associate the hypothesis with the general law of one realization per expression through examination of related works in the literature (Section 6).

2 Corpus Investigation

This section reports a concrete corpus investigation aimed at validating the hypothesis.

2.1 Data

The two resources used in this study are the Chinese *PH* corpus (Guo 1993) and the *Beihang* dictionary (Liu and Liang 1989). The Chinese *PH* corpus is a collection of about *4* million morphemes of news articles from the single source of China's *Xinhua* News Agency in 1990 and 1991. The *Beihang* dictionary is a collection of about *50,000* word-like tokens, each of which occurs at least *5* times in a balanced collection of more than 20 million Chinese characters.

What is unique in the *PH* corpus is that all and only *unambiguous* token boundaries with respect to the *Beihang* dictionary have been marked. For instance, if the English character string *fundsandmoney* were in the *PH* corpus, it would be in the form of *fundsand/money*, since the position in between character *d* and *m* is an unambiguous token boundary with respect to normal English dictionary, but *fundsand* could be either *funds/and* or *fund/sand*.

There are two types of fragments in between adjacent unambiguous token boundaries: those which are dictionary entries on the whole, and those which are not.

2.2 Dictionary-Entry Fragments

We manually tokenized in context each of the dictionary-entry fragments in the first 6,000 lines of the *PH* corpus. There are 6,700 different fragments which cumulatively occur 46,635 times. Among them, 14 fragments (Table 1, Column 1) realize different tokenizations in their 87 occurrences. 16 tokenization errors would be introduced if taking majority tokenizations only (Table 2).

Also listed in Table 1 are the numbers of fragments tokenized as single tokens (Column 2) or as a stream of multiple tokens (Column 3). For instance, the first fragment must be tokenized as a single token for 17 times but only for once as a token-pair.

Table 1: Dictionary-entry fragments realizing different tokenizations in the PH corpus.

(1)	(2)	(3)	(1)	(2)	(3)
不断	17	1	所在	3	1
城市	13	1	冬天	2	1
党中央	7	1	先生	2	1
个人	7	1	好的	1	3
才能	6	1	大气	1	1
日前	5	1	高等	1	1
年底	3	3	上层	1	1

Table 2: Statistics for dictionary-entry fragments.

(0) *Fragment*	*(1)* *All*	*(2)* *Multiple*	*(3)=(2)/(1)* *Percentage*
Occurrences	46635	87	0.19
Forms	6700	14	0.21
Errors	46635	16	0.03

In short, 0.19% of all the different dictionary-entry fragments, taking 0.21% of all the occurrences, have realized different tokenizations, and 0.03% tokenization errors would be introduced if forced to take one tokenization per fragment.

2.3 Non-Dictionary-Entry Fragments

Similarly, we identified in the *PH* corpus all fragments that are not entries in the *Beihang* dictionary, and manually tokenized each of them in context. There are *14,984* different fragments which cumulatively occur *49,308* times. Among them, only 35 fragments (Table 3) realize different tokenizations in their 137 occurrences. 39 tokenization errors would be introduced if taking majority tokenizations only (Table 4).

Table 3: Non-dictionary-entry fragments realizing different tokenizations in the PH corpus.

白银矿	北山区	从属于	从小学
大厂矿	大都会	大路上	大学校
多少年	防治水	个人大	个人中
工作为	化工作	加强调	列车队
美国会	牧草地	南通过	其中东
人大会	日前作	日夜里	上工人
十一日	十一日至	团拜会	新安放
要好的	一时间	游客人	着重要
政协会	中共同	有计划生育	

Table 4: Statistics for non-dictionary entry fragments.

(0) *Fragment*	*(1)* *All*	*(2)* *Multiple*	*(3)=(2)/(1)* *Percentage*
Forms	14984	35	0.23
Occurrences	49308	137	0.28
Errors	49308	39	0.08

In short, 0.23% of all the non-dictionary-entry fragments, taking 0.28% of all occurrences, have realized different tokenizations, and 0.08% tokenization errors would be introduced if forced to take one tokenization per fragment.

2.4 Tokenization Criteria

Some readers might question the reliability of the preceding results, because it is well-known in the literature that both the inter- and intra-judge tokenization consistencies can hardly be better than 95% but easily go worse than 70%, if the tokenization is guided solely by the *intuition* of human judges.

To ensure consistency, the manual tokenization reported in this paper has been independently done twice under the following three criteria, applied in that order:

(1) *Dictionary Existence*: The tokenization contains no non-dictionary-entry character fragment.

(2) *Structural Consistency*: The tokenization has no *crossing-brackets* (Black, Garside and Leech 1993) with at least one correct and complete structural analysis of its underlying sentence.

(3) *Maximum Tokenization*: The tokenization is a *critical tokenization* (Guo 1997).

The basic idea behind is to regard sentence tokenization as a (shallow) type of (phrase-structure-like) morpho-syntactic parsing which is to assign a tree-like structure to a sentence. The tokenization of a sentence is taken to be the single-layer bracketing corresponding to the highest-possible cross-section of the sentence tree, with each bracket a token in dictionary.

Among the three criteria, both the criterion of dictionary existence and that of maximum tokenization are well-defined without any uncertainty, as long as the tokenization dictionary is specified.

However, the criterion of structural consistency is somewhat under-specified since the same linguistic expression may have different sentence structural analyses under different grammatical theories and/or formalisms, and it may be read differently by different people.

Fortunately, our tokenization practice has shown that this is not a problem when all the controversial fragments are carefully identified and their tokenizations from different grammar schools are purposely categorized. Note, the emphasis here is not on producing a unique "correct" tokenization but on managing and minimizing tokenization inconsistency[1].

3 One Tokenization per Source

Noticing that all the fragments studied in the preceding section are *critical fragments* (Guo 1997) from the same *source*, it becomes reasonable to accept the following hypothesis.

One tokenization per source: *For any critical fragment from a given source, if one of its tokenization is correct in one occurrence, the same tokenization is also correct in all its other occurrences.*

The linguistic object here is a critical fragment, i.e., the one in between two adjacent critical points or unambiguous token boundaries (Guo 1997), but not an arbitrary sentence segment. The hypothesis says nothing about the tokenization of a non-critical fragment. Moreover, the hypothesis does not apply even if a fragment is critical in some other sentences from the same source, but not critical in the sentence in question.

The hypothesis does *not* imply context independence in tokenization. While the correct tokenization correlates decisively with its source, it does not indicate that the correct tokenization has no association with its local sentential context. Rather, the tokenization of any fragment has to be realized in local and sentential context.

It might be arguable that the *PH* corpus of 4 million morphemes is not big enough to enable many of the critical fragments to realize their different readings in diverse sentential contexts. To answer the question, 10 colleagues were asked to tokenize, *without* seeing the context, the most frequent 123 non-dictionary-entry critical fragments extracted from the *PH* corpus. Several of these fragments[2] have thus been marked "context dependent", since they have "obvious" different readings in different contexts. Shown in Figure 1 are three examples.

219 [c< 先进 水 >< 先 进水 >]
180 [c< 主要 是 >< 主 要是 >]
106 [< 人参 加 >c< 人 参加 >]

Figure 1: Critical fragments with "obvious" multiple readings. Preceding numbers are their occurrence counts in the PH corpus.

[1] For instance, the Chinese fragment 中小学 (secondary primary school) is taken as "[secondary (and) primary] school" by one school of thought, but "[secondary (school)] (and) [primary school]" by another. But both will never agree that the fragment must be analyzed differently in different context.

[2] While all fragments are lexically ambiguous in tokenization, many of them have received consistent unique tokenizations, as these fragments are, to the human judges, *self-sufficient* for comfortable ambiguity resolution.

We looked all these questionable fragments up in a larger corpus of about 60 million morphemes of news articles collected from the same source as that of the *PH* corpus in a longer time span from 1989 to 1993. It turns out that all the fragments each always takes one and the same tokenization with no exception.

While we have not been able to specify the notion of *source* used in the hypothesis to the same clarity as that of critical fragment and critical tokenization in (Guo 1997), the above empirical test has made us feel comfortable to believe that the scope of the source can be sufficiently large to cover any single domain of practical interest.

4 Application in Tokenization

The hypothesis of one tokenization per source can be applied in many ways in sentence tokenization. For tokenization ambiguity resolution, let us examine the following strategy:

Tokenization by memorization: *If the correct tokenization of a critical fragment is known in one context, remember the tokenization. If the same critical fragment is seen again, retrieve its stored tokenization. Otherwise, if a critical fragment encountered has no stored tokenization, randomly select one of its critical tokenizations.*

This is a pure and straightforward implementation of the hypothesis of one tokenization per source, as it does not explore any constraints other than the tokenization dictionary.

While sounds trivial, this strategy performs surprisingly well. While the strategy is universally applicable to any tokenization ambiguity resolution, here we will only examine its performance in the resolution of critical ambiguities (Guo 1997), for ease of direct comparison with works in the literature.

As above, we have manually tokenized[3] all non-dictionary-entry critical fragments in the *PH* corpus; i.e., we have known the correct tokenizations for all of these fragments. Therefore, if any of these fragments presents somewhere else, its tokenization can be readily retrieved from what we have manually done. If the hypothesis holds perfect, we could not make any error.

The only weakness of this strategy is its apparent inadequacy in dealing with the *sparse data problem*. That is, for unseen critical fragments, only the simplest tokenization by random selection is taken. Fortunately, we have seen on the *PH* corpus that, on average, each non-dictionary-entry critical fragment has just two (100,398 over 49,308 or 2.04 to be exact) critical tokenizations to be chosen from. Hence, a tokenization accuracy of about 50% can be expected for unknown non-dictionary-entry critical fragments.

The question then becomes that: what is the chance of encountering a non-dictionary-entry critical fragment that has not been seen before in the *PH* corpus and thus has no known correct tokenization? A satisfactory answer to this question can be readily derived from the Good-Turing Theorem[4] (Good 1953; Church and Gale with Kruskal 1991, page 49).

Table 5: Occurrence distribution of non-dictionary-entry critical fragments in the PH *corpus.*

r	1	2	3	4	5
Nr	9587	2181	939	523	339
r	6	7	8	9	≥9
Nr	230	188	128	94	775

Table 4 and Table 5 show that, among the 14,984 different non-dictionary-entry critical fragments and their 49,308 occurrences in the *PH* corpus, 9,587 different fragments each occurs exactly once. By the Good-Turing Theorem, the chance of encountering an arbitrary non-dictionary-entry critical fragment that is not in the *PH* corpus is about 9,587 over 49,308 or slightly less than 20%.

In summary, if applied to non-dictionary-entry critical fragment tokenization, the simple strategy of tokenization by memorization delivers virtually 100% tokenization accuracy for slightly over 80% of the fragments, and about 50% accuracy for the rest 20% fragments, and hence has an overall tokenization accuracy of better than 90% (= 80% x 100% + 20% x 50%).

[3] This is not a prohibitive job but can be done well within one man-month, if the hypothesis is adopted.

[4] The theorem states that, when two independent marginally binomial samples B_1 and B_2 are drawn, the expected frequency r^* in the sample B_2 of types occurring r times in B_1 is $r^*=(r+1)E(N_{r+1})/E(N_r)$, where $E(N_r)$ is the expectation of the number of types whose frequency in a sample is r.

What we are looking for here is the quantity of $r^*E(N_r)$ for $r=0$, or $E(N_1)$, which can be closely approximated by the number of non-dictionary-entry fragments that occurred exactly once in the *PH* corpus.

This strategy rivals all proposals with directly comparable performance reports in the literature, including[5] the representative one by Sun and T'sou (1995), which has the tokenization accuracy of 85.9%. Notice that what Sun and T'sou proposed is not a trivial solution. They developed an advanced four-step decision procedure that combines both *mutual information* and *t-score* indicators in a sophisticated way for sensible decision making.

Since the memorization strategy complements with most other existing tokenization strategies, certain types of hybrid solutions are viable. For instance, if the strategy of tokenization by memorization is applied to known critical fragments and the Sun and T'sou algorithm is applied to unknown critical fragments, the overall accuracy of critical ambiguity resolution can be better than 97% (= 80% + 20% x 85.9%).

The above analyses, together with some other more or less comparable results in the literature, are summarized in Table 6 below. It is interesting to note that, the best accuracy registered in China's national 863-Project evaluation in 1995 was only 78%. In conclusion, the hypothesis of one tokenization per source is unquestionably helpful in sentence tokenization.

Table 6:Tokenization performance comparisons.

Approach	Accuracy (%)
Memorization	90
Sun et al. (1996)	*85.9*
Wong et al. (1994)	71.2
Zheng and Liu (1997)	81
863-Project 1995 Evaluation (Zheng and Liu, 1997)	78
Memorization + Sun et al.	**97**

5 The Notion of Tokens

Upon accepting the validness of the hypothesis of *one tokenization per source*, and after experiencing its striking utility value in sentence tokenization, now it becomes compelling for a new paradigm. Parallel to what Dalton did for separating physical mixtures from chemical compounds (Kuhn 1970, page 130-135), we are now suggesting to regard the hypothesis as a *law-of-language* and to take it as the proposition of what a word/token must be.

The Notion of Tokens: *A stretch of characters is a legitimate token to be put in tokenization dictionary if and only if it does not introduce any violation to the law of one tokenization per source.*

Opponents should reject this notion instantly as it obviously makes the law of one tokenization per source a *tautology*, which was once one of our own objections. We recommend these readers to reexamine some of Kuhn's (1970) arguments.

Apparently, the issue at hand is not merely over a matter of definition of words/tokens. The merit of the notion, we believe, lies in its far-reaching implications in natural language processing in general and in sentence tokenization in particular.

For instance, it makes the separation between words and non-words operational in Chinese, yet maintains the cohesiveness of words/tokens as a relatively independent layer of linguistic entities for rigorous scrutiny. In contrast, while the paradigm of "mutual affinity" represented by measurements such as mutual information and t-score has repetitively exhibited inappropriateness in the very large number of intermediate cases, the paradigm of "linguistic words" represented by terms like syntactic-words, phonological-words and semantic-words is in essence rejecting the notion of Chinese words/tokens at all, as compounding, phrase-forming and even sentence formation in Chinese are governed by more or less the same set of regularities, and as the whole is always larger than the simple sum of its parts. We shall leave further discussions to another place.

6 Discussion

Like most discoveries in the literature, when we first captured the regularity several years ago, we simply could not believe it. Then, after careful experimental validation on large representative corpora, we accepted it but still could not imagine

[5] The task there is the resolution of overlapping ambiguities, which, while not exactly the same, is comparable with the resolution of critical ambiguities. The tokenization dictionary they used has about 50,000 entries, comparable to the *Beihang* dictionary we used in this study. The corpus they used has about 20 million words, larger than the *PH* corpus. More importantly, in terms of content, it is believed that both the dictionary and corpus are comparable to what we used in this study. Therefore, the two should more or less be comparable.

any of its utility value. Finally, after working out ways that unquestionably demonstrated its usefulness, we realized that, in the literature, so many supportive evidences have already been presented. Further, while never consciously in an explicit form, the hypothesis has actually already been widely employed.

For example, Zheng and Liu (1997) recently studied a newswire corpus of about 1.8 million Chinese characters and reported that, among all the 4,646 different *chain-length-1 two-character-overlapping-type*[6] ambiguous fragments which cumulatively occur 14,581 times in the corpus, only 8 fragments each has different tokenizations in different context, and there is no such fragment in all the 3,409 different *chain-length-2 two-character-overlapping-type*[7] ambiguous fragments.

Unfortunately, due to the lack of a proper representation framework comparable to the critical tokenization theory employed here, their observation is neither complete nor explanatory. It is not complete, since the two ambiguous types apparently do not cover all possible ambiguities. It is not explanatory, since both types of ambiguous fragments are not guaranteed to be critical fragments, and thus may involve other types of ambiguities.

Consequently, Zheng and Liu (1997) themselves merely took the apparent regularity as a special case, and focused on the development of local-context-oriented disambiguation rules. Moreover, while they constructed for tokenization disambiguation an annotated "phrase base" of all ambiguous fragments in the large corpus, they still concluded that good results can not come solely from corpus but have to rely on the utilization of syntactic, semantic, pragmatic and other information.

The actual implementation of the weighted finite-state transducer by Sproat et al. (1996) can be taken as an evidence that the hypothesis of one tokenization per source has already in practical use. While the primary strength of such a transducer is its effectiveness in representing and utilizing local and sentential constraints, what Sproat et al. (1996) implemented was simply a token unigram scoring function. Under this setting, no critical fragment can realize different tokenizations in different local sentential context, since no local constraints other than the identity of a token together with its associated token score can be utilized. That is, the requirement of one tokenization per source has actually been implicitly obeyed.

We admit here that, while we have been aware of the fact for long time, only after the dissemination of the closely related hypotheses of *one sense per discourse* (Gale, Church and Yarowsky 1992) and *one sense per collocation* (Yarowsky 1993), we are able to articulate the hypothesis of *one tokenization per source.*

The point here is that, *one tokenization per source* is unlikely an isolated phenomenon. Rather, there must exist a general *law* that covers all the related linguistic phenomena. Let us speculate that, for a *proper linguistic expression* in a *proper scope,* there always exists the regularity of *one realization per expression.* That is, only one of the multiple values on one aspect of a linguistic expression can be realized in the specified scope. In this way, *one tokenization per source* becomes a particular articulation of *one realization per expression.*

The two essential terms here are the *proper linguistic expression* and the *proper scope* of the claim. A quick example is helpful here: part-of-speech tagging for the English sentence "*Can you can the can?*" If the linguistic expressions are taken as ordinary English words, they are nevertheless highly ambiguous, e.g., the English word *can* realizes three different part-of-speeches in the sentence. However, if "*the can*", "*can the*" and the like are taken as the underling linguistic expressions, they are apparently unambiguous: "the can/NN", "can/VB the" and the rest "can/MD". This fact can largely be predicted by the hypothesis of *one sense per collocation*, and can partially explain the great success of Brill's transformation-based part-of-speech tagging (Brill 1993).

As to the hypothesis of *one tokenization per source*, it is now clear that, the theory of critical tokenization has provided the suitable means for capturing the proper linguistic expression.

[6] Roughly a three-character fragment *abc* where *a, b, c, ab,* and *bc* are all tokens in the tokenization dictionary.

[7] Roughly a four-character fragment *abcd*, where *a, b, c, d, ab, bc,* and *cd* are all tokens in the tokenization dictionary.

7 Conclusion

The hypothesis of *one tokenization per source* confirms surprisingly well (99.92% ~ 99.97%) with corpus evidences, and works extremely well (90% ~ 97%) in critical ambiguity resolution. It is formulated on the critical tokenization theory and inspired by the parallel hypotheses of *one sense per discourse* and *one sense per collocation*, as is postulated as a particular articulation of the general law of *one realization per expression*. We also argue for the further generalization of regarding it as a new *paradigm* for studying the twin-issue of token and tokenization.

Acknowledgements

Part of this paper, especially the Introduction and Discussion sections, was once presented at the November 1997 session of the monthly Symposium on Linguistics and Language Information Research organized by COLIPS (Chinese and Oriental Languages Information Processing Society) in Singapore. Fruitful discussions, especially with Xu Jie, Ji Donghong, Su Jian, Ni Yibin, and Lua Kim Teng, are gratefully acknowledged, as are the tokenization efforts by dozen of my colleagues and friends. However, the opinions expressed reflect solely those of the author.

References

Black, Ezra, Roger Garside, and Geoffery Leech (1993). Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach, Amsterdam: Rodopi Publishers.

Brill, Eric (1993). A Corpus-Based Approach to Language Learning, Ph.D Dissertation, Department of Computer and Information Science, University of Pennsylvania.

Church, Kenneth. W. and William A. Gale (1991). A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams, Computer Speech and Language, Vol. 5, No. 1, pages 19-54.

Gale, William A., Kenneth W. Church and David Yarowsky (1992b). One Sense Per Discourse, In: Proceedings of the 4th DARPA Workshop on Speech and Natural Language, pages 233-237.

Gan, Kok-Wee; Palmer, Martha; and Lua, Kim-Teng (1996). A Statistically Emergent Approach for Language Processing: Application to Modeling Context Effects in Ambiguous Chinese Word Boundary Perception. Computational Linguistics Vol. 22, No. 4, pages 531-553.

Good, I. J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika, Volume 40, pages 237-264.

Guo, Jin (1993). PH - A Free Chinese Corpus, Communications of COLIPS, Vol. 3, No. 1, pages 45-48.

Guo, Jin (1997). Critical Tokenization and its Properties, Computational Linguistics, Vol. 23, No. 4, pages 569-596.

Kuhn, Thomas (1970). The Structure of Scientific Revolutions. Second Edition, Enlarged. The University of Chicago Press. Chicago.

Liu, Yuan and Nanyuan Liang (1989). Contemporary Chinese Common Word Frequency Dictionary (Phonetically Ordered Version). Yuhang Press, Beijing.

Sproat, Richard, Chilin Shih, Villiam Gale, and Nancy Chang (1996). A Stochastic Finite-State Word-Segmentation Algorithm for Chinese, Computational Linguistics, Vol. 22, No. 3, pages 377-404.

Sun, Maosong and Benjemin T'sou (1995). Ambiguity Resolution in Chinese Word Segmentation, Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation (PACLIC-95), pages 121-126, Hong Kong.

Wong, K-F.; Pan, H-H.; Low, B-T.; Cheng, C-H.; Lum, V. and Lam, S-S. (1995). A Tool for Compute-Assisted Open Response Analysis, Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages, pages 191-198, Hawaii.

Wu, Dekai (1997). Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora, Computational Linguistics, Vol. 23, No. 3, pages 377-403.

Yarowsky, David (1993). One Sense Per Collocation, In: Proceedings of ARPA Human Language Technology Workshop, Princeton, pages 266-271.

Zheng, Jiaheng and Kaiying Liu (1997). The Research of Ambiguity Word-Segmentation Technique for the Chinese Text, In Chen, Liwai and Qi Yuan (editors). Language Engineering, Tsinghua University Press. Page 201-206.