How Should a Large Corpus Be Built?–A Comparative Study of Closure in Annotated Newspaper Corpora from Two Chinese Sources, Towards Building A Larger Representative Corpus Merged from Representative Sublanguage Collections

John Kovarik (kovariks@worldnet.att.net)
U.S. Department of Defense

Abstract

This study measures comparative lexical and syntactic closure rates in annotated Chinese newspaper corpora from the Academica Sinica Balanced Corpus and the University of Pennsylvania's Chinese Treebank. It then draws inferences as to how large such corpora need be to be representative models of subject-matter-constrained language domains within the same genre. Future large corpora should be built incrementally only by combining smaller representative sublanguage collections.

1 Prior Work

Practically speaking, earlier attempts at building corpora, such as the IBM/Lancaster approach, have taken an all-inclusive perspective toward text selection proposing that (Garside and McEnery, 1993) raw texts for parsed corpora should come from a variety of sources. The IBM/Lancaster group used the Canadian Hansards collection of parallel parsed English and French sentences as a base of English parsed sentences and then focused on the Computer Manuals domain, in which they attempted to randomly select texts with some additional non-Computer Manual material selected as "a measure of 'light relief' " supposedly for the benefit of the annotators. A broad approach was also used in the Hong Kong element of the International Corpus of English (ICE) project (Greenbaum, 1992) which sought to assemble a range of both spoken and written texts along with a range of both formal and informal situations to provide a reasonably large, well-documented and detailed snapshot of the use of educated English. (Bolt, 1994) Both the IBM/Lancaster approach and the ICE project build millions of tokens worth of corpora.

But from a more principled perspective, Douglas Biber, in speaking on representativeness in corpus design, pointed out that the linguistic characterization of a corpus should include both its central tendency and its range of variation. (Biber, 1993) Similarly Geoffrey Leech has stated that a corpus, in order to be representative, must somehow capture the magnitude of languages not only in their lexis but also in their syntax. (Leech, 1991) *This suggests we should build corpora focusing on how well they can approach lexical and syntactic closure, rather than by merely fixating on ever larger amounts of text. To build representative corpora why not first select representative texts constrained by genre of writing?*

In general one possible technique for methodical corpus selection would be to build large corpora out of representative sub-collections constrained by genre and subject matter. Zelig Harris said, "Certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it." (Harris, 1968) Calling this an inductive definition of sublanguage Satoshi Sekine has embarked on studies involving new trends in the analysis of sublanguages (Sekine, 1994). Both Harris and Sekine recognized that sublanguages are an efficient way to observe and measure the properties of natural language in smaller, representative blocks.

Getting down to specifics, McEnery and Wilson (McEnery and Wilson, 1996) have hypothesized that genres of writing, such as the style used in newspapers and similar printed publications to report news stories, represent a constrained subset of a natural language. Thus newspaper texts constitute a sublanguage – a version of a natural language which does not display all of the creativity of that natural lan-

guage. The newspaper sublanguage can be further constrained by subject matter to divide it into smaller, more manageable subsets.

A key mathematical feature of a sublanguage is that it will show a high degree of closure at various levels of description, setting it apart from unconstrained natural language. This closure property of a sublanguage is analogous to the mathematical property of transitive closure. McEnery and Wilson used the closure property to measure and compare rates of lexical and syntactic closure in three corpora: the IBM computer manual corpus, the Canadian Hansards, and the American Printing House for the Blind corpus. To date, however, there has been little work in a similar vein in other languages.

2 Overview

This work applies the methodology of McEnery and Wilson to examine closure rates in a comparative study of all available tagged Chinese newspaper corpora. First I define lexical and syntactic closure for this study in section 3. Then, section 4 begins this study with an examination of all the newspaper texts of the Academica Sinica Balanced Corpus (ASBC). Section 5 extends this study to an examination of the newspaper texts of the UPenn Chinese Treebank (CTB). Section 6 presents my findings and section 7 discusses some implications for future corpus building.

3 Lexical and Syntactic Closure

3.1 Tokenization in Chinese

It should be pointed out that Chinese is an agglutinative, not an inflected language. Moreover, while Chinese tokens can concatenate, Chinese has no extensive morphology like many Indo-European languages. Chinese, of course, has no white space separating lexemes, as a result, all Chinese text must first be segmented into word lengths. However, once a text has been segmented, no stemming is needed so each segmented Chinese word can be counted as it occurs without the need of finding its lemma.

3.2 Lexical Closure

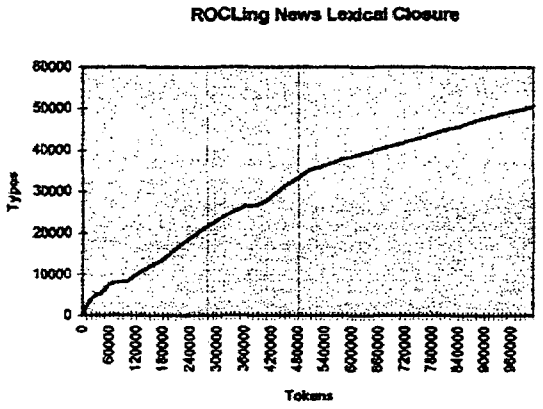
Lexical closure is that property of a collection of text whereby given a representative sample, the number of new lexical forms seen among every additional 1,000 new tokens begins to level off at a rate below 10 per cent.

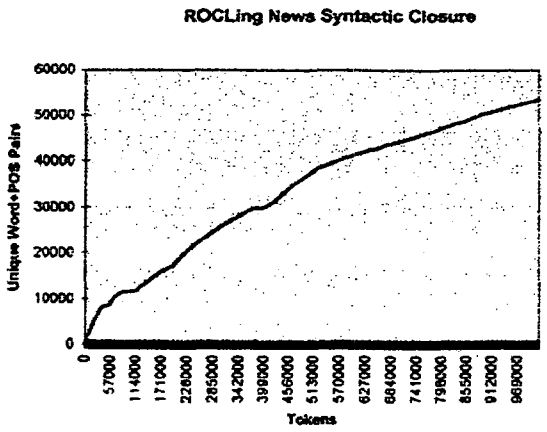
3.3 Syntactic Closure

Syntactic closure is that property of a collection of text whereby given a representative sample of a type of text, then the number of new syntactic forms seen among every additional 1,000 new tokens begins to level off. A syntactic form is the combination of token plus type. Thus syntactic closure approaches as the number of new grammatical uses for a previously observed token plus the number of new tokens, regardless of syntactic use, level off to a growth rate below 10 per cent.

4 Academica Sinica Balanced Corpus (ASBC)

While it is common practice to attempt to build huge annotated corpora, it is of course very tedious, very expensive, and especially challenging for annotators to maintain consistency over such a huge task. Consequently one must hope that once an annotated corpus of newspaper texts is created, it can be statistically measured and confirmed to be a representative sample. I first measured lexical and syntactic closure rates in all ASBC newspaper texts but found that when viewed as a whole this newspaper sub-collection of the ASBC does not approach closure (see graphs below).





This raises the question *how can we hope for NLP applications to learn on large corpora if they themselves never approach statistical closure, never approach being statistically confirmed as a representative model of the language?*

I then focused downward on subsections of the newspaper corpus–grouping them by similar filename. I searched the ASBC corpus looking for files of annotated newspaper text and found a total of 57 files (18.7 Mb); my findings are summarized in the following table.

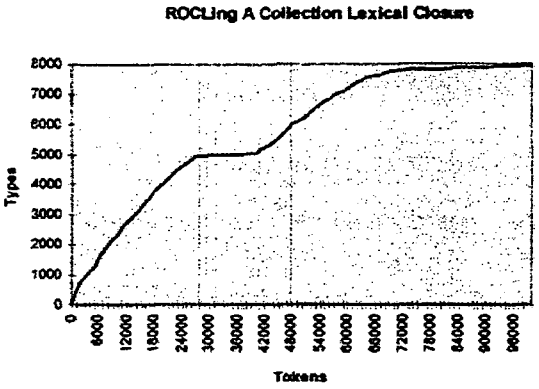
Academica Sinica Balanced Corpus News

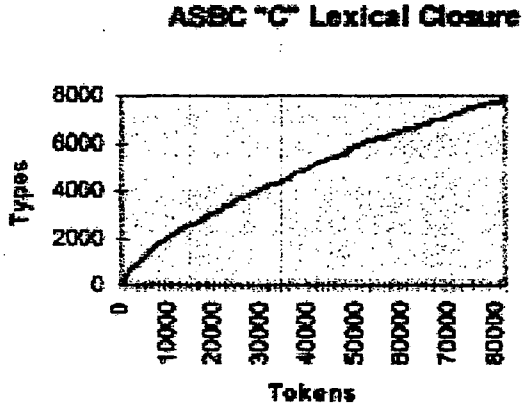
Files	Size	Filenames	Subject Matter
05	01.9Mb	A....	'94 Academics
16	01.5Mb	C....	'93 Various News
01	00.2Mb	SSSLA	'91 Politics etc.
36	15.1Mb	T....	'91-'95 Sports etc.

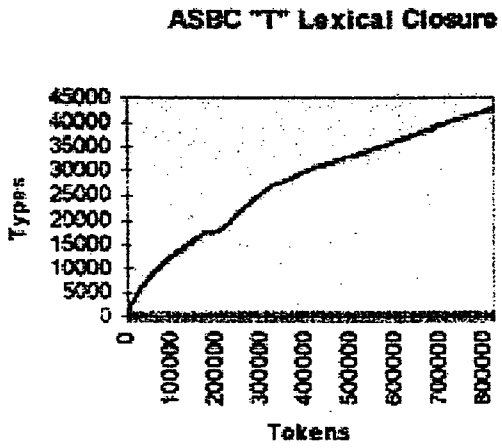
The large single file, named "SSLA", dealt with a wide assortment of subject matter and thus was significantly different from the other 3 newspaper collections. Not only was its individual file size rather large; it was not even close to the size and homogeneity of the other three newspaper multi-file collections. I rejected it from further study.

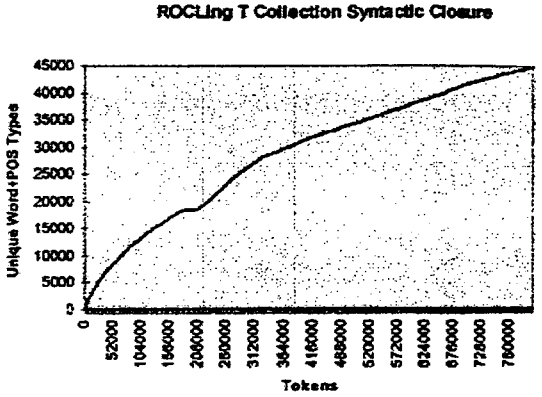
The other sub-collections were more similar. Topically speaking, the ASBC "A" newspaper collection was focused primarily on news (77 per cent) while at the same time focusing narrowly on academic events in 1994. The ASBC "C" newspaper collection was less narrowly focused on news (73.5 per cent) but expanded its focus to other than academia while limiting itself to events of 1993. The ASBC "T" newspaper collection, however, spanned the period 1991 through 1995 and dealt with many different subjects, the most frequent of which were sports, news, and domestic politics, but even each of these most frequent subjects only represented 9 per cent of the whole.

Let us consider the three ASBC newspaper sub-collections ("A", "C", and "T" filenames) to be potentially representative sublanguages. If we can observe relatively high degrees of closure at various levels of description, we can propose that such sub-collections are representative sublanguages within the newspaper genre of Chinese natural language. Conversely, those which do not have a high degrees of closure are definitely not sublanguage corpora and not of further interest for this study. The following graphs depict the observed lexical and syntactic closure rates of the three ASBC newspaper sub-collections under study.







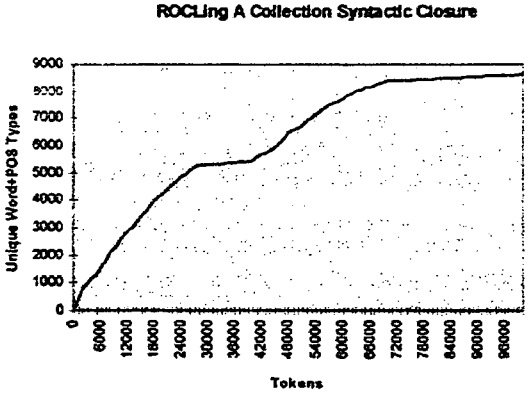


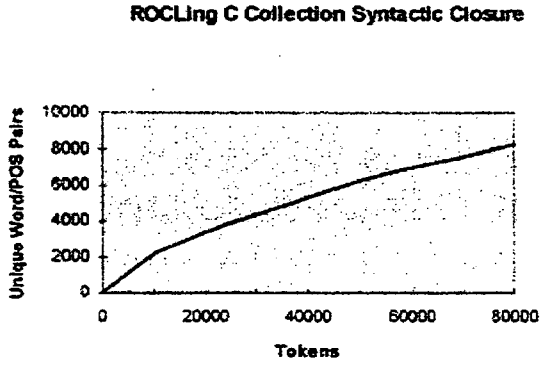
It appears that the ASBC "A" newspaper sub-collections does approach lexical closure; while the "C" and "T" newspaper sub-collections definitely do not.

It appears that the ASBC "A" newspaper sub-collection also approaches syntactic closure; while the "C" and "T" newspaper sub-collections do not.

5 UPenn Chinese Treebank

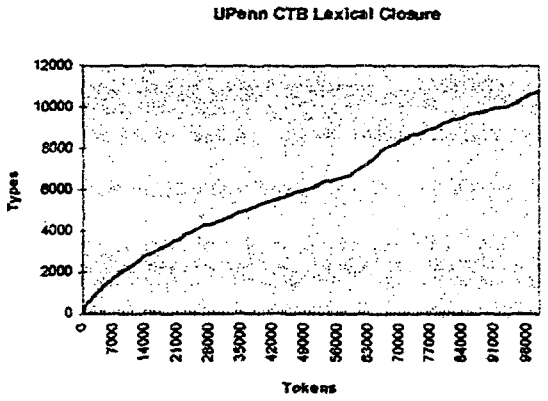
I next applied the same measures on the UPenn Chinese Treebank corpus. I wanted to compare the rate at which the UPenn collection approaches lexical and syntactic closure with that of the ASBC "A" and "T" sub-collections. The 329 Xinhua newswire documents in the UPenn Chinese Treebank annotated corpus came from two sub-collections and total 3,289 sentences averaging 27 words or 47 characters per sentence excluding newspaper headlines which are characteristically highly abbreviated clauses. The Perl script written to do this analysis is freely available at *http://home.att.net/~kovariks/closure.htm.*

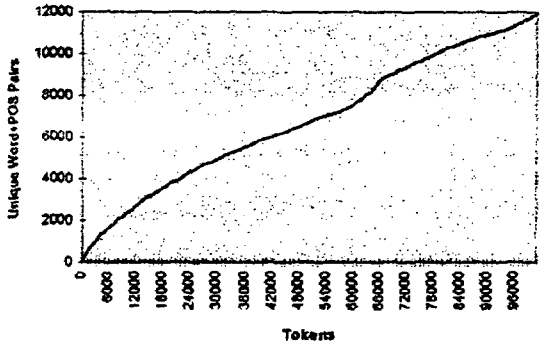




UPenn Chinese Treebank Newspaper Texts

Files	Size	Sub-Collection	Subject
121	.322Mb	One	'94 Economics
040	.114Mb	Two	'96 Economics
076	.406Mb	Two	'97 Economics
054	.181Mb	Two	'98 Economics
038	.060Mb	One	General



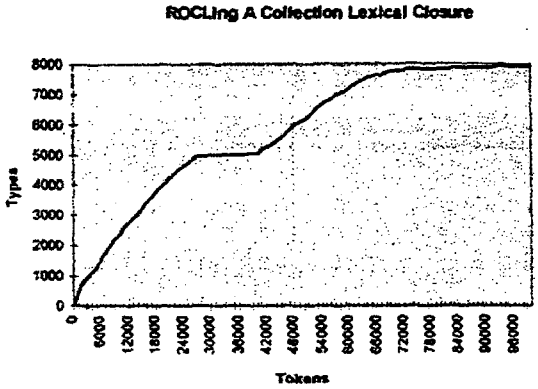


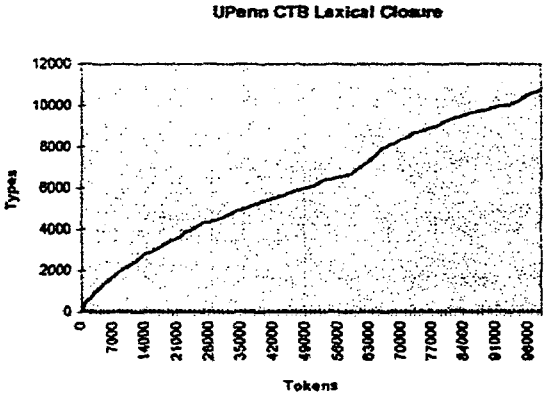
6 Findings

The UPenn data initially approaches lexical and syntactic closure at a rate which can be favorably compared with the ASBC "A" Chinese newspaper sub-corpus. By the time 59,000 tokens of the UPenn corpus were tagged, only 56 new token+tag combinations were observed in the preceding 1,000 tokens and 27 of those were new proper nouns. In comparison, at this point the closure rate in the ASBC "A" corpus was not quite as good (see table below). But continuing to the 69,000 token mark, the ASBC "A" closure rate had overtaken that of the UPenn CTB.

Tokens	Corpus	New token+tags
59,000	UPenn CTB	56 (27 NR)
	ASBC A	77 (34 N...)
69,000	UPenn CTB	61 (36 NR)
	ASBC A	67 (49 N...)

Interestingly the graphs for both the ASBC "A" and the UPenn CTB data reveal two-humped curves. At approximately the 60,000 token mark on both UPenn CTB graphs the curve was starting to flatten only to suddenly shift into a sharper climb.





An investigation of the UPenn CTB data revealed that the vast majority of the documents at this point dealt with international aspects of the Chinese economy [1], whereas previously the vast majority of documents had focused on Chinese domestic economic growth [2]. And similarly on the ASBC "A" collection closure graphs around the 30,000 token mark, both

[1] Headlines of articles in UPenn CTB from 66,800-68,300 tokens– 66809: 英宣布参加入侵海地的多国部队 ; 66976: 西班牙外交大臣反对美国封锁古巴 ; 67114: 法批评美国对古巴的禁运政策 ; 67282: 德、俄将加强反核走私行动 ; 67439: 日本科学试验卫星入轨无望 ; 67920: 俄货运飞船与"和平"号轨道站 ; 68298: 法将不参与对海地的军事干预 .

[2] Topics of articles in UPenn CTB from 28,000-32,000 tokens– 28071: 中保财险公司 ; 28345: 中国税务部门规定: ; 28745: 广东科技产出指标 ; 29197:

curves start to flatten only to suddenly around the 40,000 token mark shift back into a climb until about 70,000 tokens. An investigation of the ASBC data also showed a subtle shift in subject matter, most of the documents from 29,000 to 39,000 tokens were short notes and simple bulletins on shifts in academic positions, whereas the later data had more long stories on subjects of greater complexity [3], whereas the later data had more long stories on subjects of greater complexity [4].

Thus the ASBC "A" collection, more so than the UPenn CTB corpus, did eventually seem to approach closure. By the time we reach 80,000 tokens, the ASBC "A" collection only saw 14 new token+tag combinations in the last thousand tokens of new text(see table below). Six of those 14 new combinations were nouns, seven were verbs, and one was a preposition [5], having nearly reached lexical closure on newspaper articles regarding academics. In contrast the CTB collection instead logged 146 new token+tag combinations at the 80,000 token mark and its new vocabulary ranged widely across 12 different parts of speech [6]. While the majority of this new vocabulary were nouns, this continuing influx of new words was due primarily to the late inclusion of international news in the CTB's collection of newspaper articles regarding Chinese economics.

天津开发区近百家外资企业 ; 29444: 中国三家企业获 ; 29948: 江苏开放型经济 ; 31003: 陕西将向海内外 ; 31499: 台湾一经济学家指出 ; 31890: 陕西引进外资 ; 32246: 甘肃经济 .

[3]Topics of articles in ASBC "A" from 29,300-32,000 tokens- 29306: 29337: 29382: 讯息: 康乐会消息 ; 29537: 29581: 讯息: 小启 ; 29636: 环保: 读者来文 ; 30087: 讯息: 动态报导 ; 30153: 讯息: 「兰屿观点」在法国获佳作奖 ; 30342: 30402: 30449: 30478: 30539: 讯息: 人事动态 ; 30590: 30677: 30772: 30822: 30944: 讯息: 小启 ; 30976: 31093: 讯息: 学术研讨会 .

[4]Topics of articles in ASBC "A" from 56,000-57,000 tokens- 56183: 人事: 历史语言研究所徵组助理 ; 56232: 讯息: 康乐会活动 ; 56274: 讯息: 房屋出租 ; 56323: 讯息: 「台湾近代经济史」座谈会 ; 56475: 讯息: 「亲子互动与伤害」研讨会 ; 56635: 人物: 动态报导 ; 56798: 人事: 李院长获聘为高普考典试委员长 ; 56852: 讯息: 欧美所邀贾化乐博士来访 ; 56930: 讯息: 生医所协办「台湾问题」研讨会 ; 57019: 讯息: 「第七届傅斯年汉学讲座」.

[5]ASBC Subcollection A* at 80,000 tokens, 14 New token+types observed: Na 检, Na 民间团体, Na 天文学家, Na 语汇词, Na 正教授, Nc 五组, P 据, VA 过路, VC 奖, VC 签, VC 现存, VF 奉, VF 助, VJ 费; Total tags: 7, Total new items: 14.

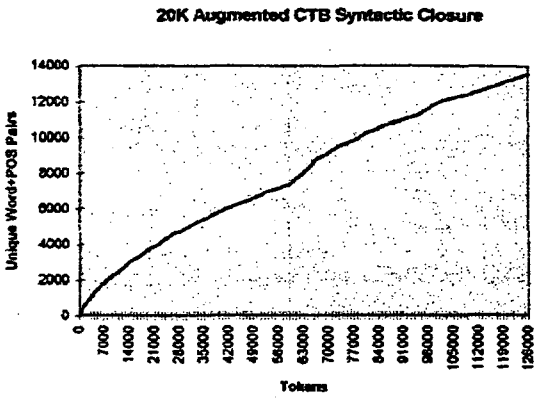
[6]UPenn CTB at 80,000 tokens, 146 New token+types observed: AD遂, AD为何, AD武装, AD行将, AD原则, AD正好, AD转瞬, CD１・２万, CD１７１６亿, CD２００, CD２００余, CD２１００多, CD２８００万, CD４００万, CD５５, JJ反法西斯, JJ军工, JJ老牌, JJ统一, JJ专属, LC东半部, LC东边, LC南端, LC西边, M等, M海里, NN阿方, NN阿军, NN捕鱼, NN参与者, NN得益者, NN电信, NN东西方, NN短评, NN对比, NN防务, NN风潮, NN副外长, NN附图, NN管辖权, NN管制, NN海域, NN合并, NN后人, NN后裔, NN环形, NN较量, NN尽头, NN巨人, NN句号, NN剧变, NN军政府, NN控制线, NN盟国, NN旗帜, NN前线, NN桥头堡, NN侵略者, NN热战, NN日, NN世人, NN势力, NN水域, NN牺牲, NN小岛, NN血泪, NN印方, NN英方, NN英国人, NN英军, NN渔牧业, NN渔业, NN渔业区, NN占领军, NN战, NN殖民地, NN中叶, NN终结, NN主岛, NN转折, NN组长, NR阿, NR柏林, NR大马尔维纳, NR大西洋, NR东北亚司, NR东德, NR东欧, NR东西欧, NR福克兰, NR华约, NR联秘希夫香卡尔・梅农, NR洛克希德, NR马丁, NR马丁・玛丽埃塔, NR斯坦利港, NR索莱达, NR唐家璇, NR王楠, NR西德, NR希特勒, NR夏治沔, NR新德里, NR雅尔塔, NR印, NT１１月, NT１８３３年, NT１８世纪, NT１９３９年, NT１９８２年, NT１９８６年, NT１９８９年, NT近来, PU布同¡, P据据, P凭借, VA安宁, VV波及, VV独立, VV对峙, VV遏制, VV分区, VV扶植, VV攻占, VV共睹, VV共管, VV合并, VV画上, VV划分, VV击败, VV解体, VV借重, VV进攻, VV剧变, VV看不到, VV冷战, VV沦为, VV难以为继, VV派, VV逝去, VV瓦解, VV危机重重, VV显而易见, VV一分为二, VV愈演愈烈, VV遭到; Total tags: 12, Total new items: 146.

Tokens	Corpus	New token+tags
80,000	UPenn CTB	146 (24 NR)
	ASBC A	14 (6 N...)

7 Corpus Building Implications

The fact that the UPenn Chinese Treebank data approaches lexical and syntactic closure at rates comparable to the ASBC "A" file newspaper collection suggests that if the UPenn data had been selected more narrowly, it might have reached closure for the economics domain in the newspaper genre even sooner. Some day corpus linguistics may only need much smaller collections of annotated corpora than is the practice today, relying on new directions in sublanguage research. In the case of the ASBC "A" collection, for example, a robust learning structure should be able to build a useable model on 100,000 words worth of data like this which exhibits strong tendencies to lexical and syntactic closure in the Chinese newspaper genre constrained to a given domain.

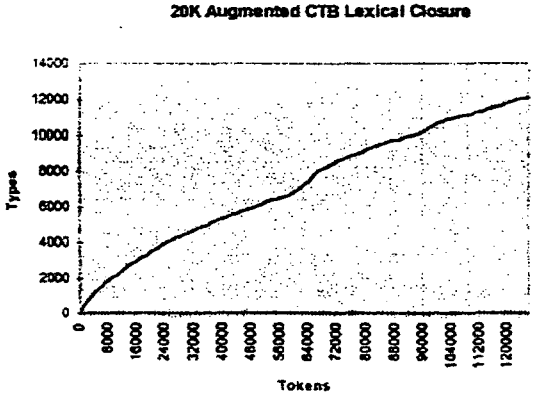
If the UPenn CTB were enlarged by the infusion of more news stories on international aspects of Chinese economic development, the CTB might better reach lexical and syntactic closure. The following graph shows that the blind addition of 20K additional Chinese economic news stories does not aid closure much. This additional data spanned many topics not seen in the original 100K collection. If it had been selected precisely to aid closure by measuring its potential contribution before extensive hand annotation, the result could have been better.

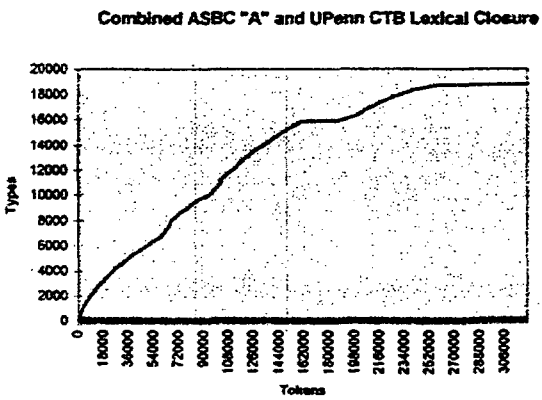


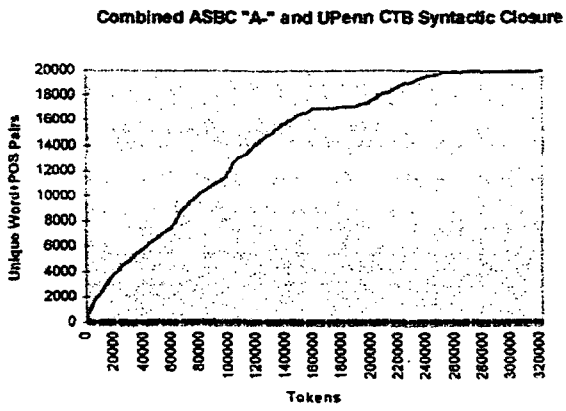
Nevertheless, this expanded CTB collection is sufficiently improved that the rate of closure toward of the expanded collection is better (see table below).

Tokens	New Tokens	New Token+Tags
121,000	55	64 (12 NR)
122,000	61	68 (17 NR)
123,000	61	76 (23 NR)
124,000	78	82 (18 NR)
125,000	48	63 (11 NR)
126,000	52	61 (16 NR)

Consequently, this expanded CTB collection is sufficiently improved that merging it with the ASBC "A" collection results in a far more measurably representative larger corpus as shown in the final two graphs below. The creation of such measurably representative large corpora out of such smaller, better focused sublanguage building blocks would be cheaper and faster without the resulting tools developed against such corpora suffering much degradation in speed or accuracy.







8 Discussion

Since only two Chinese tagged corpora are available at the present time, only about 200,000 words of Chinese corpora have been so studied. But to see more work in this vein, one need only consult McEnery and Wilson's study of three English corpora: the IBM computer manual corpus, the Canadian Hansards, and the American Printing House for the Blind corpus (McEnery and Wilson, 1996). Their study (exhaustively detailed in Chapter 6 of their book) spans more than 2.2 million words of tagged English text in three different domains. Consequently the total of 2.3 million words from McEnery and Wilson's results in tagged English texts when combined with these results in tagged Chinese newspaper texts should satisfy any who might argue that there is insufficient data upon which to draw some general conclusions.

This paper does not argue that the two closure measures used are the only measures possible. The argument here is simply that these two closure measures are used to spot when a sublanguage corpus approaches closure–that is, when the curve of new types and new combinations of type with token begins to flatten at a rate below ten percent. One can readily point out that no natural language corpus can ever guarantee closure. The best anyone can aspire to do today, given the current state of our art, is to only approach closure.

9 References

References

D. Biber. 1993. Representativeness in corpus design. volume 8(4), pages 243–257.

P. Bolt. 1994. The international corpus of english project–the hong kong experience. pages 15–24.

R. Garside and A. McEnery. 1993. Treebanking: the compilation of a corpus of skeleton parsed sentences. pages 17–35.

S. Greenbaum. 1992. A new corpus of english: Ice. pages 171–179.

Z. Harris. 1968. *Mathematical Structures of Language*. New York: John Wiley and Sons.

G. Leech. 1991. The state of the art in corpus linguistics. pages 8–29.

T. McEnery and A. Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

S. Sekine. 1994. A new direction for sublanguage nlp. In *Proceedings of the International Conference on New Methods in Language Processing, CCL, UMIST*, pages 123–129.