Computational Linguistics and Chinese Language Processing
Vol. 17, No. 2, June 2012, pp. 37-54

Frequency, Collocation, and Statistical Modeling of Lexical Items: A Case Study of Temporal Expressions in Two Conversational Corpora[1]

Sheng-Fu Wang*, Jing-Chen Yang*, Yu-Yun Chang*,

Yu-Wen Liu+, and Shu-Kai Hsieh*

Abstract

This study examines how different dimensions of corpus frequency data may affect the outcome of statistical modeling of lexical items. Our analysis mainly focuses on a recently constructed elderly speaker corpus that is used to reveal patterns of aging people's language use. A conversational corpus contributed by speakers in their 20s serves as complementary material. The target words examined are temporal expressions, which might reveal how the speech produced by the elderly is organized. We conduct divisive hierarchical clustering analyses based on two different dimensions of corporal data, namely raw frequency distribution and collocation-based vectors. When different dimensions of data were used as the input, results showed that the target terms were clustered in different ways. Analyses based on frequency distributions and collocational patterns are distinct from each other. Specifically, statistically-based collocational analysis generally produces more distinct clustering results that differentiate temporal terms more delicately than do the ones based on raw frequency.

[1] Acknowledgement: Thanks Wang Chun-Chieh, Liu Chun-Jui, Anna Lofstrand, and Hsu Chan-Chia for their involvement in the construction of the elderly speakers' corpus and the early development of this paper.

* Graduate Institute of Linguistics, National Taiwan University, 3F, Le-Xue Building, No. 1, Sec. 4, Roosevelt Rd., Taipei Taiwan, 106
E-mail: {sftwang0416; flower75828; june06029}@gmail.com; shukaihsieh@ntu.edu.tw

+ Department of English, National Taiwan Normal University, No. 162, He-ping East Road, Section 1, Taipei, Taiwan, 106
E-mail: Yw_L7@hotmail.com

Keywords: Clustering, Collocation, Corpus Linguistics, Temporal Expression, Gerontology

1. Introduction

The study of gerontology is gaining wide attention, as the aging population is becoming a major issue in society. Research has noted that aging causes not only physiological changes for elderly people, but also affects their language production (Burk & Shafto, 2004), cognitive load (Wilson, 2008), context processing speed (Rush, Barch, & Braver 2006), language performance patterns compared to younger individuals (Veliz, Riffo, & Arancibia, 2010), *etc*. To study gerontology from a linguistic point of view, Green (1993) proposed that the phenomenon of gerontology could be studied through discourse analysis. Therefore, we use conversations the elderly and from younger speakers as our speech corpora and take these corpora as an input to exemplify the procedures and usage of lexical modeling.

As for the examination of temporal terms, we postulated that people's social roles, especially the elderly's, might be embedded in conversations where speakers share personal experiences or judgment of the past (Kuo, 2008) and the present. Thus, we presume that temporal expressions might serve as the anchoring points in a conversation-based corpus and might reveal a certain aspect of the speech behavior patterns of the elderly.

Statistical modeling can serve to describe a given set of data, be it diachronic subsets, register, or lexical units. Statistical models often take the so-called "bottom-up" approach that suits most corpus linguists' empirical state of mind. Moreover, nice and neat visualization is often a feat in such modeling techniques, to an extent that some of the models are called "graph models" (Widdows & Dorow, 2002). When the proper behavior of lexical units and the structure of the lexicon are applied, statistical modeling may help us develop NLP-oriented lexicographic modules in the form of dictionaries, thesauri, and ontologies (Mitrofanova, Mukhin, Panicheva, & Savitsky, 2007).

A glimpse at relevant studies would reveal that the most prominent kind of data input is related to the distributional patterns of the lexical items in corpora. The distributional data could be in the form of word frequencies and variability of frequencies (Gries & Hilpert, to appear) or the distribution of n-grams as a whole (Gries, Newman, & Shaoul, 2011). The distributional pattern or dependency with syntactic patterns is also a prominent source of data input (Cimiano, Hotho, & Staab, 2004a, 2004b; Lin, 1998; Pereira & Tishby 1992). Target lexical items' dependency and co-occurrence with particular word types also has been taken as the basis of lexical modeling in some studies (Redington, Chater, & Finch, 1993). Moreover, statistically-based collocational patterns are used for modeling similarities among lexical units of interest (Chen, 2009; Gries & Stefanowitsch, to appear).

The above-mentioned different methods, or rather, different data inputs, are considered to fall somewhere between raw distributional data and relational data, or between lexical items and syntactic patterns. In our study, we aim to compare the two endpoints of this methodological continuum, namely the "frequency distributional data" input and "collocation data," in order to see how these different types of inputs may result in different data in lexical modeling. The former is the analysis based on the raw distributional data of lexical items, and the latter is based on more fine-grained examination of the relationship between lexical items, such as a particular item's collocational pattern with other lexical units in a corpus. By comparing the two kinds of analyses, we hope to reveal the extent to which these two methods demonstrate different patterns, thereby making contributions to the evaluation and selection of research methods in modeling lexical items.

This paper is organized as follows: Section 2 introduces the corpora used in this study, including data collection, guidelines for transcription, and annotation standards. Section 3 reports basic corporal information and preliminary analysis of six selected temporal expressions from the corpus. Section 4 demonstrates the methods and results of statistical modeling of temporal expressions, as well as a meta-analysis on different models. Section 5 and Section 6 summarize our findings and suggestions for further research respectively.

2. Corpora

2.1 The Elderly Speaker Corpus

Speech data were collected from four pairs of elderly people. Each pair consisted of one male and one female speaker. All subjects are native speakers of Mandarin and Taiwanese Southern Min. One pair is from Changhua while the others are from Taipei. The mean age of the subjects is 65.75 years old (SD = 6.16). Each pair of speakers was asked to do a face-to-face conversation in Mandarin for 30 to 40 minutes. The designated conversational topic was the speakers' life experience in the past and the present. During the recording, other participants, such as the subject's relatives or the observer, might also be involved in the talk. All files were recorded by a digital recorder in the WAV format. The total length of the speech samples is 145 minutes.

Speech samples collected from the elderly people's conversations were then transcribed into Chinese characters, following Du Bois *et al.*'s transcription standards for discourse analysis (Du Bois, Schuetze-Coburn, Cumming, & Paolino, 1993). Prosodic features and vocal qualities of the intonation units (IUs) were excluded from the transcription since they are not the main interest in this study. A short guideline of transcription standards is provided below.

Conversation samples were manually processed into several IUs. Each IU was labeled with a number on the left, as shown in Example (1).

(1)

34 SM: a 你 看 這 個 做工 的
P. you see this CL. do.work DE
35 ...(1.3) 那 個 有--
that CL. have
36 有夠 重
have.enough heavy

Speech overlap occurring during the conversation was indicated by square brackets, as shown in Example (2). To indicate overlap, brackets were vertically aligned where the overlaps began. Double square brackets were used for numerous overlaps occurring within a short stretch of speech, with their left brackets displaying temporal alignment.

(2)

70 SF: ...都 [送 人家]
all give others
71 SM: [送 人家] [[撫養 la]]
give others to raise P.
72 SF: [[撫養]]
to raise

Sometimes, the subjects switched between Mandarin and other languages when speaking. Such cases of code-switching were enclosed in square brackets and labeled with *L2* as well as the code for that non-Mandarin language. Example (3) demonstrates a case where a speaker switched between Mandarin and Taiwanese Southern Min (TSM). It should be noted that, since this study targeted elderly people's Mandarin speech performance, code-switching to languages other than Mandarin was excluded from our analysis.

(3)

268 SF: [L2 TSM 單輪車 TSM L2]
single wheeler

Laughter was also identified in the transcription. Each syllable of laughter was labeled with one token of the symbol @ (see Example 4a). Longer laughter was indicated by a single symbol @ with the duration in the parentheses (see Example 4b). Two @ symbols were placed at each end of an IU to show that the subject spoke while laughing (see Example 4c).

(4)

a. 163 F1: @@@@@
b. 200 SM: @(3.3)
c. 828 O: @沒 那麼 嚴重 la@
not that serious P.

The occurrence and duration of a pause in discourse were transcribed. Pauses were represented by dots: two dots for short pauses of less than 0.3 seconds, three dots for medium pauses between 0.3 and 0.6 seconds, and three dots with the duration specified in parentheses for pauses longer than 0.7 seconds. Example (5) below is the instance for pauses.

(5)

40 SF: ..以前 o..是--
before P. is

41 SF: ...eh ..都 是..父母...(0.9)做 X
P. all is parents do X

Particles were transcribed in phonetic transcription to avoid disagreement on the employment of homophonic Mandarin characters, as Example (6) shows. Phonetic transcriptions for the particles included *la, hoNh, a, o, le, haNh, hioh,* and *ma.*

(6)

26 SM: hoNh.. a 我們 二十 幾 歲 結婚
P. P. we twenty more age get.married

The recorded utterances were not always audible or clear enough for the transcribers to identify what was being said. Each syllable of uncertain hearing was labeled with a capital X, as shown in Example (5) above. Last but not least, truncated words or IUs were represented by double hyphens --, as shown in Examples (1) and (5).

The transcription was automatically segmented and POS (part of speech) tagged through the CKIP Chinese Word Segmentation System provided by the Chinese Knowledge Information Processing (CKIP) group at Academia Sinica (2004). The segmentation and POS standards were based on the Sinica Corpus guidelines (1998). The annotated language samples then were checked manually. The procedure is described below.

First, every segmentation result derived from CKIP was examined and corrected if wrong, as in the following examples. Example (7a) is the original IU before segmentation and tagging. Through CKIP, we attain the result in Example (7b), which is falsely processed. Example (7c) shows the proper segmentation after manual correction.

(7)

a. 我爸爸是他媽媽的哥哥
"My father is his mother's brother."

b. *我 爸爸 是 他媽 媽的 哥哥
I father is he.mom mom.DE brother

c. 我 爸爸 是 他 媽媽 的 哥哥
I father is he mom DE brother

Second, POS tags were viewed as correct if the main word classes were correct, while the details of their sub-classes were not of primary concern. For instance, in Example (8), the main word class of each POS tag (in this case, *N*, *DE*, *V*, or *D*) was examined, but not the sub-class tagging, as we give less consideration for the subcategories they belong to.

(8)

他(**N**h)	的(**DE**)	腦筋(**N**a)	動(**V**AC)	得(**DE**)	比較(**D**fa)	快(**V**H)
he	DE	brains	act	DE	more	fast

"He gets new ideas faster."

Third, particles' tags were manually corrected to *I* for IU-initial particles[2], and *T* for IU-final particles. If an IU contained nothing but particles, then the particles were tagged as *I*.

Finally, POS tags were removed for truncations (*e.g.* 這--), uncertain hearing (*i.e.* X), and code-switching. Given that truncations are not generally viewed as lexical items, they were not suitable for analysis at the lexical level.

2.2 The NTU Conversational Corpus

This corpus contains speech data collected by Master's students of the Graduate Institute of Linguistics at the National Taiwan University. The transcription follows Du Bois *et al.*'s standards for discourse analysis (1993). The data was collected by graduate students in their early 20s, and most of the recruited speakers were similar in age to the data collectors. In other words, this corpus contains mostly "youth speech," which is suitable as a complementary corpus to the elderly speakers' corpus for our analysis.

Although constant effort has been made in data collection of the NTU conversational corpus for more than ten years, little effort has been devoted to data organization and preprocessing. For this study, we selected a subset of face-to-face conversations between speakers (mostly students) less than 30 years old. The topics of these conversations were mostly everyday life experiences. The chosen subset, containing around 66,000 words, was tokenized and annotated the same way as the elderly speakers' corpus for further analysis.

3. Corpus Information & Preliminary Analysis

The elderly speaker corpus contains 4,982 IUs of Mandarin utterances and 22,090 word tokens produced by all speakers. Elderly people's production in Mandarin contains 3,739 IUs, and there are 18,076 word tokens in total. The subset of the NTU conversational corpus used in this study contains 15,863 IUs and 65,742 word tokens.

---

[2] According to the standards provided by Sinica Corpus, *I* represents "interjections" which usually occur in the IU-initial position.

The corpus processing tool used here is R (R. D. C. Team, 2010), which allows us to perform tasks, including preprocessing, word frequency, KWIC (KeyWord In Context) extraction, and statistical modeling.

We assume that time-related words may provide some vital clues to the elderly's and the youth's speech patterns, so the following analyses will focus on the subjects' use of temporal expressions. We are interested in how elderly people use 現在 (now) and 以前 (before), as well as other temporal expressions (tagged as Nd), to frame the present- and the past-related concepts, and how their use differs from the younger generations' employment of temporal expressions. Thus, the six most frequent temporal expressions from each corpus were selected for the analysis. Table 1 lists the frequency of the six target temporal expressions. As shown in the rankings, "now" is the most frequent temporal expression in both corpora.

Table 1. The frequency of the most frequent temporal expressions in the two corpora.

Elderly Speaker Corpus			NTU (Youth) Corpus		
Rank.	**Term**	**Freq.**	**Rank.**	**Term**	**Freq.**
1	現在(now)	169	1	現在(now)	137
2	以前(before)	169	2	後來(later)	76
3	小時候(in one's childhood)	12	3	今天(today)	52
4	民國(R.O.C. year)	11	4	以前(before)	52
5	當初(back then)	9	5	昨天(yesterday)	34
6	最近(recently)	6	6	今年(this year)	31

4. Statistical Modeling of Temporal Expressions

In this section, we will present quantitative analyses with the help of hierarchical clustering, a data-driven approach, to see how the temporal terms of interest are grouped together with the frequency data extracted from our corpus.

The clustering method employed here is divisive hierarchical clustering. This differs from agglomerative hierarchical clustering in that a group of entities is first divided into large groups before smaller groups are classified. Such a method is useful for finding a few clusters that are large in size (Rush, Barch, & Braver, 2006). We would like to find out whether the terms for "the present" and "the past" can really be grouped into clusters that are different in temporality. Thus, divisive hierarchical clustering serves our need. In our current study, the clustering was conducted with the help of the *dist()* function in R (2010), which takes a table of correlations between the vectors containing different temporal expressions' frequency data or collocational data in the corpus and transforms it into a "table of distances" based on the square distance between these vectors.

4.1 Modeling Results from the Elderly Speaker Corpus

We executed a series of hierarchical clustering with different data input. The first analysis was run with the frequencies of the temporal terms across different files/texts in our corpus. Such an input was expected to capture the co-occurrence pattern of these temporal terms affected by individual speaker's style or idiolect, as well as by differences in the topic of conversation. The output is presented in Figure 1, where 現在 (now) is separate from 以前 (before) under a major cluster on the left. Also, 最近 (recently) stands independently from the other expressions, suggesting that temporal terms within a particular time domain are more likely to occur in the same text, which is really a conversational event in our corpora.

Next, four clustering analyses were made based on the frequency data across subsets of different sizes. The sizes chosen for producing subsets were 10, 50, 200, and 500 words. Smaller subsets may reflect linguistic patterns in a few clauses, and larger subsets may reflect patterns in a larger unit, such as major or minor topics in the flow of conversation. As we can see in Figure 2, 現在 (now) and 以前 (before) are classified in the same small cluster.

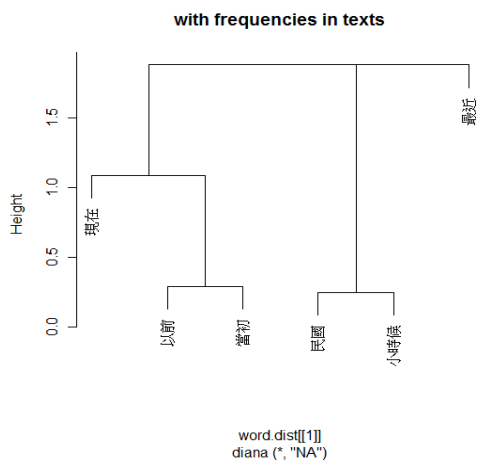

Figure 1. Clustering based on frequencies in texts in the elderly speaker corpus. "Height" in the y-axis represents the furthest distance between the entities under a particular node in terms of the distance, based on the correlation of data vectors of these expressions. Thus, it is sensitive to how far apart the entities in question are.

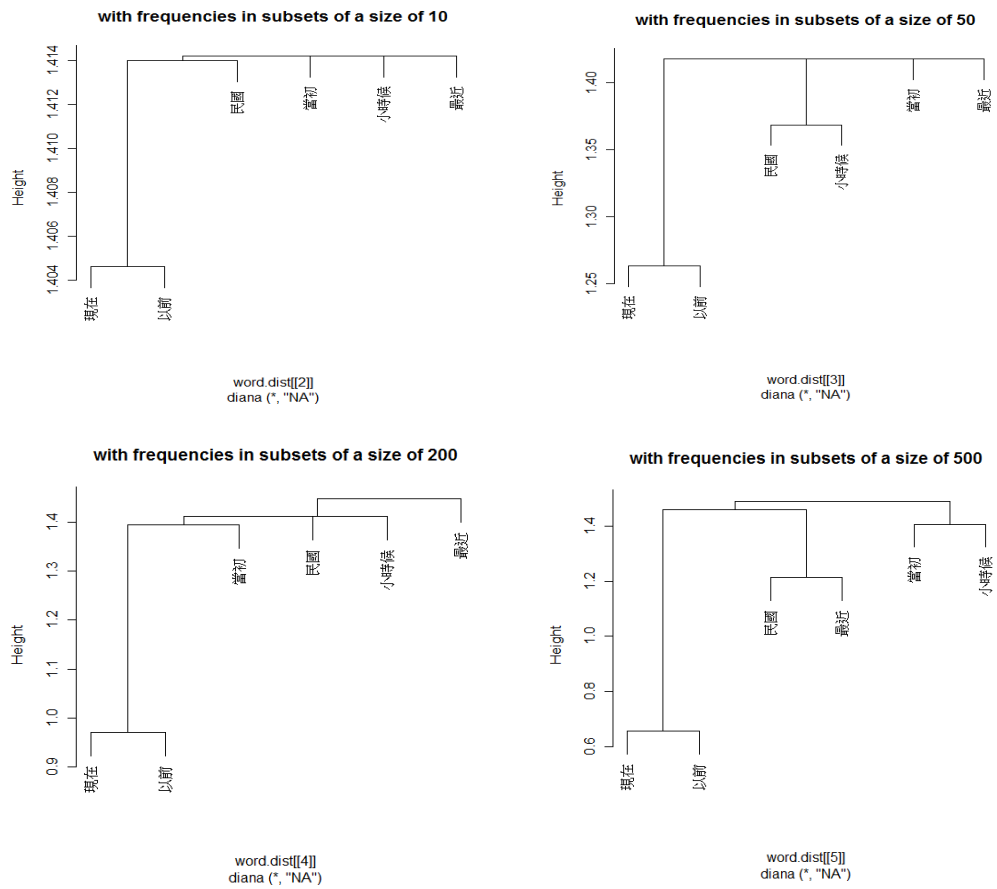

Figure 2. Clustering based on frequencies across subsets in the elderly speaker corpus. Upper left, with subsets of a size of 10 words. Upper right, of 50 words. Lower left, of 200 words. Lower right, of 500 words.

We can also conduct clustering analysis according to how these terms collocate with other words in the corpus, on the premise that collocational patterns should reveal some characteristics of lexical items. Thus, two more analyses based on this assumption were given. The first analysis was done using each word type's collocational pattern (span = 3) with the six temporal terms as input. The second analysis was achieved through the dependency patterns of sentential particles (*i.e.* lah, hoNh, ah, oh, le, haNh, hioh, mah, as described by Li, 1999), taking the temporal terms as input. There are two reasons for the inclusion of particle collocation. First, with regard to methodology, running more than one collocational test allows one to see whether collocational analyses with different approaches generate similar results. Second, sentential particles' dependency patterns might help us understand how the "referent" of each temporal expression is conceived and presented in discourse. The outcome is

illustrated in Figure 3. Again, 現在 (now) and 以前 (before) are clustered closely, showing that their collocational patterns might be similar.

There is a potential problem in using raw frequencies in studying collocates. Collocates with high frequencies might simply be high frequency words, rather than being "exclusively close" to the terms of interest. Thus, we bring forth collexeme analysis (Gries, Hampe, & Schönefeld, 2005; Gries, 2007), a statistical method developed for finding "true collocates," that is, collocates with strong collocational strength (coll.strength hereafter). The coll.strength of each word type and particle was calculated and used as input for clustering analysis. The output is shown in Figure 4.

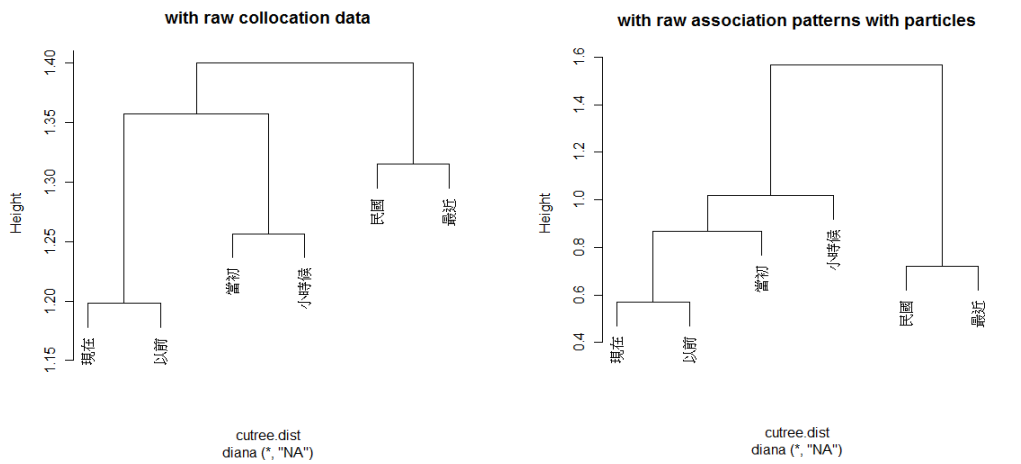

Figure 3. Clustering based on association/collocation frequencies. Left, with all word types in the elderly speaker corpus. Right, with particles.

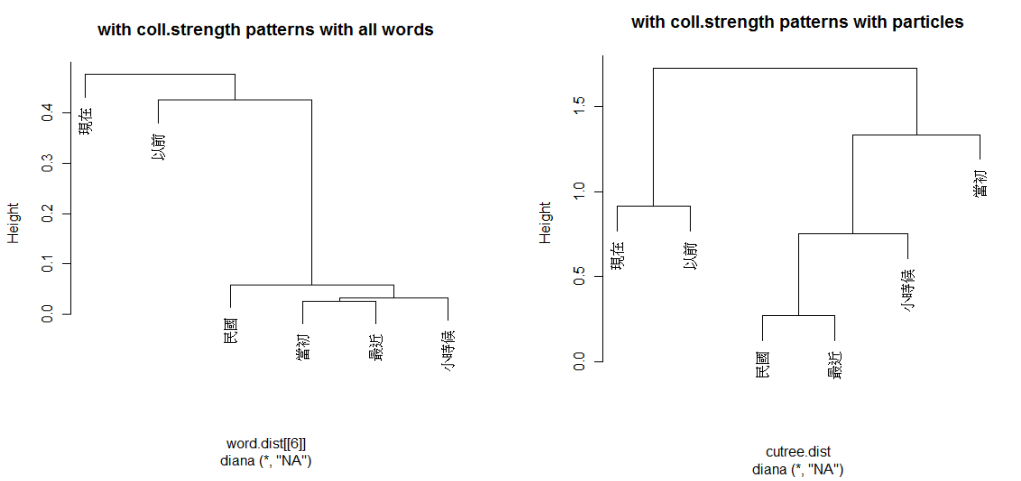

Figure 4. Clustering based on coll.strength patterns. Left, with all word types in the elderly speaker corpus. Right, with particles.

To sum up, 現在 (now) and 以前 (before) seem to intertwine, concerning their occurrences in different subsets of the corpus. This may suggest that, when elderly speakers talk about the past, the present follows as a contrast in time regarding the same subject matter, and *vice versa*. Only the by-text analysis shows a difference between the terms for the present and those for the past, suggesting that there are some conversations featuring more present than the past, and *vice versa*. Collocational strength analysis is another approach revealing a difference between 現在 (now) and 以前 (before), showing that, although the two terms usually are used closely, they still attract different words with different strengths. It should be noted that association patterns with particles do not distinguish between the present and the past. A possible explanation for this is that such a difference in pragmatic and discourse meaning is too fine-grained to be shown with information based on quantitative data. In other words, it shows that a quantitative method with corpus data has its limitation, especially when the annotation only functions at the basic POS level. Such findings of the temporal terms may in turn suggest that modeling lexical items is not a simple matter of looking for any types of analyzable data input. In addition to surface frequencies, taking collocational patterns into account, especially those based on statistical analyses, seems to be a requirement to capture the nuances among lexical items.

4.2 Modeling Results from the NTU (Youth) Conversational Corpus

We performed similar analyses for the selected portion of the NTU conversational corpus, with a few modifications. First, the analysis based on the frequencies of temporal expressions in corpus subsets was only conducted with subset sizes of 10, 50, and 200 words, since the individual texts (conversations) in the NTU corpus are mostly not as long as those in our elderly speakers' corpus. For certain conversations, a 500-word window can cover almost all of the words in the text, making the analysis too similar to the by-text analysis we conducted. Second, the analyses of temporal expressions' raw collocation and collocational strength with particles were not conducted for the NTU corpus because the manual annotation on collocation between particles and temporal terms has not yet been completed.

Figure 5 shows the results of clustering analysis based on raw frequencies of the temporal expressions in the NTU conversational corpus. Temporal scope seems to be the factor determining the clustering patterns. Expressions that have a less definite time frame, such as 後來 (later) and 以前 (before), are grouped together, whereas expressions denoting a specific temporal scope, such as 今天 (today) and 昨天 (yesterday), are grouped under the same node.

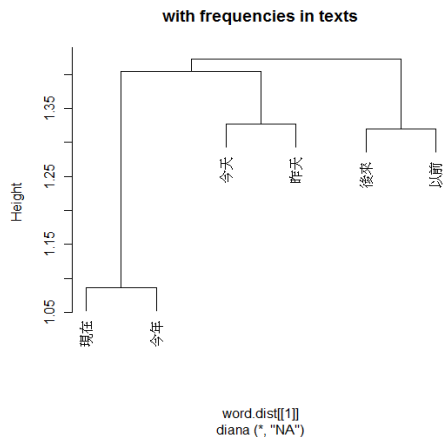

Figure 5. Clustering based on frequencies in texts in the NTU conversational corpus.

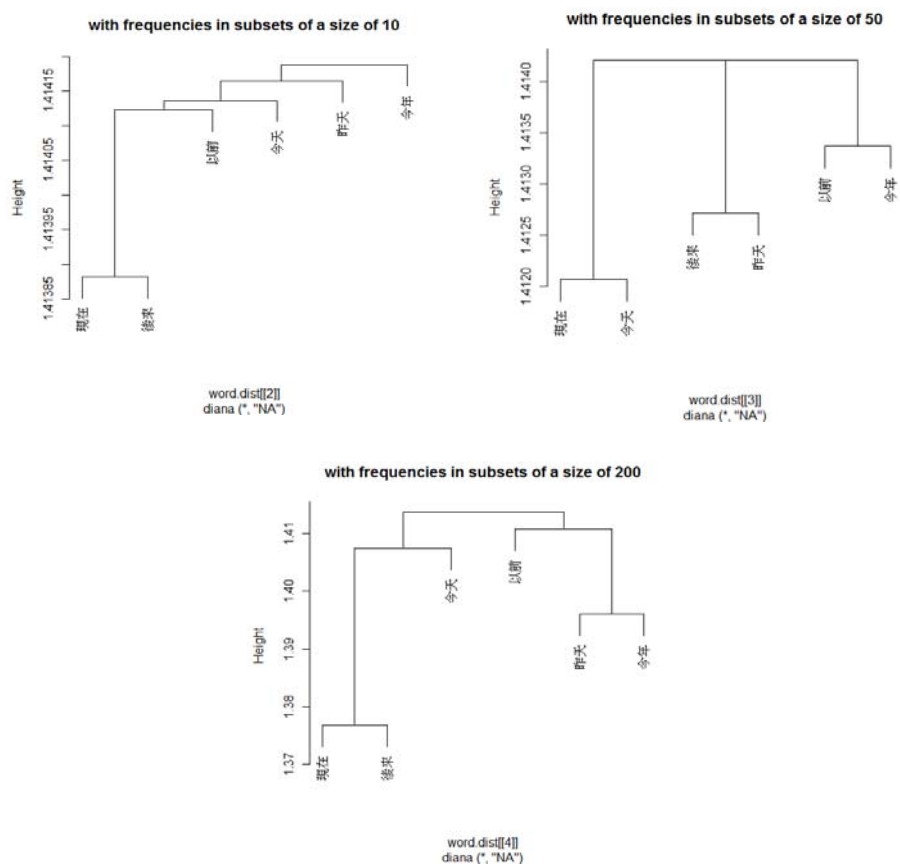

Figure 6. Clustering based on frequencies across subsets in the NTU conversational corpus. Left, with subsets of 10 words. Middle, of 50 words. Right, of 200 words.

Figure 6 shows clustering results based on the target expressions' frequencies in the subsets of the data. The results are not clear-cut, lacking a consistent pattern across the outputs from different inputs. One of the more consistent patterns may be that 現在 (now) and 後來 (later) are clustered together in two of the graphs above and are rather closely clustered in one of the two. This might imply that, within a context with a size smaller than the whole text, these two temporal expressions are commonly used together to form conversations.

As for the result of the target temporal expressions' collocates in raw frequencies, clustering patterns differentiate expressions for the past, the present, and those with different time scopes. In Figure 7, the expression 今年 (this year), which denotes the present within a bigger scope, is clustered away from other expressions. The expressions 現在 (now) and 今天 (today), both denoting the present within a smaller scope, are clustered together. Two expressions for the past, 以前 (before) and 昨天 (yesterday), are grouped under the same node. Finally, 後來 (later), an expression for denoting time sequences, is clustered alone. This indicates that different types of temporal expressions collocate with other words in the corpus in different manners.

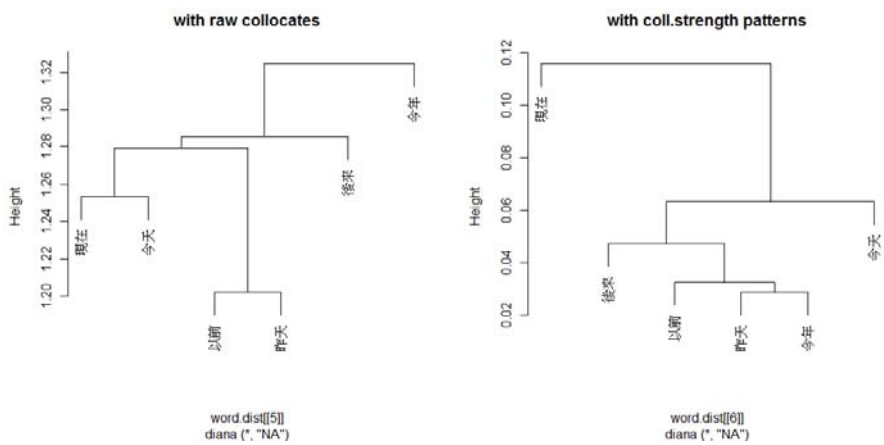

Figure 7. Left, clustering based on collocational frequencies with all the word types in the NTU conversational corpus. Right, clustering based on collocational strength patterns.

Yet, when the input data are collocational strength patterns, which should reveal these expressions' true collocates, the results of the clustering change. 現在 (now) and 今天 (today), two expressions about the present, stand out in two distinct clusters, while 後來 (later) also stands out to a certain extent. This reveals that expressions about the present are strongly and uniquely collocated with a certain group of words in the youth corpus, compared to other temporal terms. Such a result of modeling may imply that the youth use some particular patterns in structuring events or topics about the present.

4.3 Evaluation on Modeling Results

How do we evaluate all of these different results? The answer may not be surprising: We can do it with clustering analysis. The "clustering" package for R offers a function "cutree" for a simple quantification of different clustering, where each "tree" is quantified in terms of which cluster an item is clustered to. We collect the data for all of the trees shown above and execute clustering as meta-analysis. The outcome for the elderly speaker corpus is shown in Figure 8, and the one for the NTU conversational corpus is shown in Figure 9.

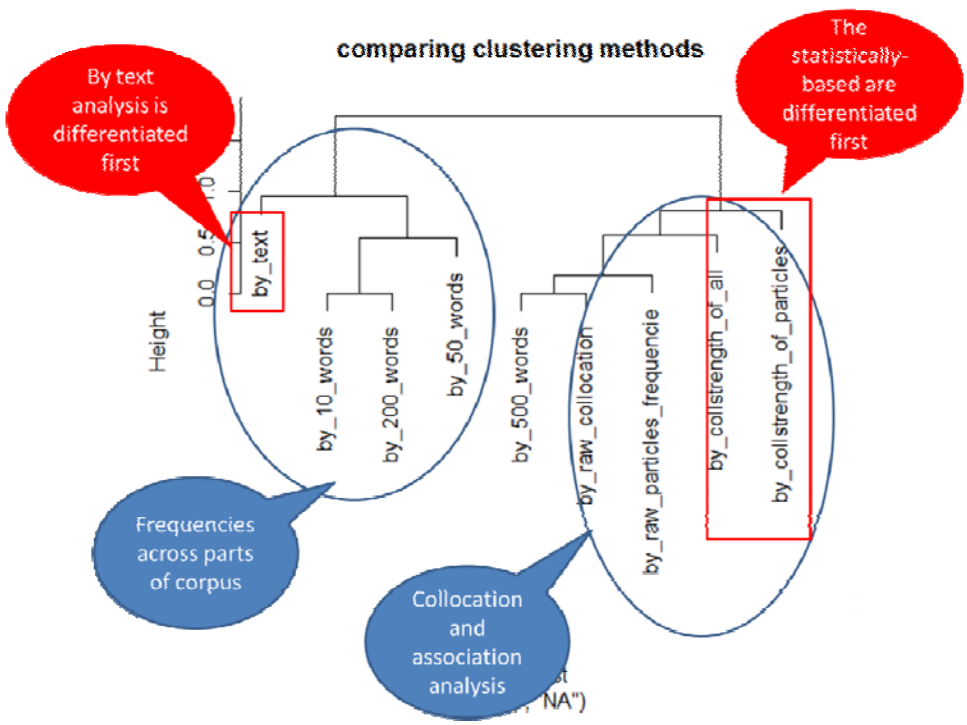

Figure 8. Clustering of various results from the elderly speaker corpus.

An interesting pattern shows up in Figure 8. There are two major clusters. The left one is based on frequency patterns of temporal terms, and the right one basically contains analyses of how these terms collocate or associate with other words or particles. Despite the curious occurrence of the "by-500-words" analysis in the right major cluster, the result of this meta-analysis seems to be able to characterize the major differences in terms of data input. More specifically, in the left major cluster, the "by-text" analysis is the first one singled out. This conforms to our impression that temporal terms are clustered differently, with 現在 (now) and 最近 (recently) placed relatively far away from other past-related expressions. Moreover, in the right major cluster, the analyses with coll.strength are the first ones differentiated from the others. Again, this reflects that statistically-based analyses produce

different patterns from the ones based on simple frequency values. What can be inferred from the patterns in Figure 8 is that different types of data input certainly influence the outcome of clustering analysis.

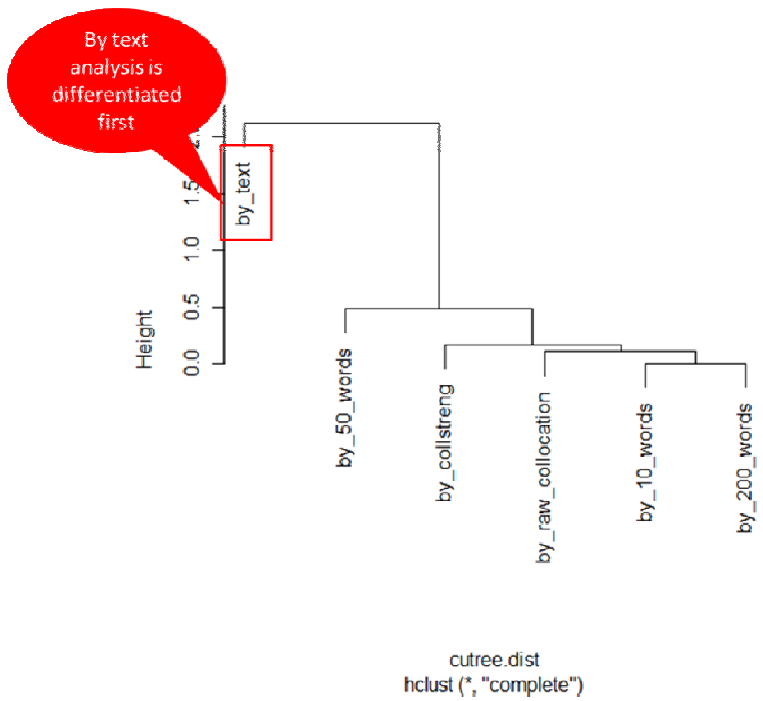

Figure 9. Clustering of various results from the NTU conversational corpus.

The evaluation of the clustering results of the NTU conversational corpus, shown in Figure 9, clearly is different from the one of the elderly speakers' corpus. The only similar pattern is that the by-text analysis still stands out as a particular kind of clustering. Other clustering analyses are lumped together with no clearly emerging pattern. There are several possible reasons for such a result. First, one may argue that the relationship between clustering methods and their results does not hold a consistent pattern. On the other hand, it could mean that there are some differences in the use of the temporal items between these two corpora, hence, between speakers different in age. At this stage, the two corpora are not equal in size, so any claims about how elderly speakers' linguistic patterns actually differ from those of younger speakers' would appear to be unsound. To sum up, the evaluation method we propose here is a technique that can possibly reveal differences among methods of modeling or even differences between various corpora, as shown in our preliminary research results. For further development, similar investigations on better-controlled and comparable corpora must be conducted.

5. Conclusion

Statistical modeling based on different types of data input displays different patterns, with modeling derived from frequencies and collocational patterns forming two major clusters (as revealed in the meta-analysis, which visualizes the difference between models based on quantitative data). In the "frequency" cluster, analysis based on distributional patterns is differentiated from the ones based on arbitrarily divided subsets. In the "collocation" cluster, statistically-oriented (*i.e.* collocational strength) analyses are distinguished from those based on surface collocational frequencies. For our present study, these findings are not overwhelmingly surprising because it is not hard to imagine the impact of the difference in texts and subsets on research, as well as the impact of surface frequencies and statistically-calculated relational patterns. Yet, when it comes to evaluating more types of modeling methods or inputs, meta-analysis of this kind provides a valuable means of choosing adequate methods. For instance, when researchers try to model different aspects of the lexical structure, the hierarchical modeling proposed here may help avoiding utilizing methods that are in fact very similar.

According to our analysis on temporal terms, the core expressions of the present and the past have very similar distributional patterns, showing that elderly speakers in the corpus tend to compare the present with the past in the same textual domains. The difference between these terms is disclosed only in models based on by-text frequency and statistical collocational analysis. The former shows that different speakers or conversation events may have their own preferred usage of temporal expressions. The latter indicates that these terms are still different in terms of their collocations, yet the difference can only be revealed through statistical tests on "true collocates" proposed by Gries (2007). These findings can be seen as a pilot result of the linguistic patterns of aging people and young people in comparison.

6. Future work

The primary purpose of this study was to attempt to highlight certain methodologies applicable to an elderly speaker corpus through several statistical approaches, rather than recklessly leaping to a conclusion that some universal elderly speech patterns are found in our corpus. The inclusion of the analysis on a younger speaker corpus helps us become more cautious with the claims we can make about our statistical approaches. Yet, to further explore the issue and confirm the validity of potential general linguistic patterns discovered in the current research, we must carefully conduct qualitative analyses of each temporal expression and interpret the results with the evidence from the quantitative methods we adopted previously. For example, the expansion of the size of the elderly speaker corpus does may alter the outcome of our statistical modeling.

Also, the inequality in terms of the size of our two corpora and the different ways of data collection make the direct comparison between the elderly speaker corpus and the NTU (youth) conversational corpus difficult. In the future, it might be advisable to collect speech materials from a small number of younger speakers by asking them to narrate personal experiences and stories just like what we asked the elderly speakers to do. By doing so, we can directly compare this small corpus contributed by younger speakers with our elderly corpus, to see whether we can prompt even similar high-frequency temporal expressions for comparing two corpora of two generations of speakers.

References

Baayen, R. H. (2008). Analyzing linguistic data. *A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Bruke, D. M. & Shafto, M. A. (2004). Aging and language production. *Current directions in psychological science,* 13, 21-24.

Chen, C.-H. (2009). Corpus, lexicon, and construction: A quantitative corpus approach to Mandarin possessive construction. *International Journal of Computational Linguigstics and Chinese Language Processing,* 14, 305-340.

Cimiano, P., Hotho, A., & Staab, S. (2004a). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. in *Proceedings of the European Conference of Artificial Intelligence*, 435-439.

Cimiano, P., Hotho, A., & Staab, S. (2004b). Clustering concept hierarchies from text. in *Proceedings of LREC 2004*, 1-4.

CKIP. (2004). CKIP Chinese word segmentation system. Retrieved June 2, 2011 from http://ckipsvr.iis.sinica.edu.tw/

CKIP. (1998). Introduction to Sinica Corpus: A tagged balance corpus for Mandarin Chinese. Taipei, Taiwan (R.O.C.): Academia Sinica.

Du Bois J. W., Schuetze-Coburn, S., Cumming, S. & Paolino, D. (1993). Outline of discourse transcription. In Edwards J. A. & Lampert M. D. (Eds.), *Talking data: Transcription and coding in discourse research.* Hillsdale: New Jersey, Lawrence Erlbau.

Green, B. S. (1993). *Gerontolgy and the social construction of old age.* New York: Aldine De Gruyter.

Gries, S. T. (2007). Collostructional analysis: Computing the degree of association between words and words/constructions. Retrived May 30, 2011 from: http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/coll.analysis.r

Gries, S. T., Hampe, B., & Schönefeld, D., (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. C*ognitive Linguistics*, 16, 635-676.

Gries, S. T. & Hilpert, M. (to appear). Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics. In Nevalainen, T. & Traugott, E. C.,

(Eds.), *Handbook on the history of English: Rethinking approaches to the history of English.* Oxford: Oxford University Press.

Gries, S. T., Newman, J. & Shaoul, C. (2011). N-grams and the clustering of registers. *Empirical language research*. Retrived May 28, 2011 from: http://ejournals.org.uk/ELR/article/2011/1

Gries, S. T. & Stefanowitsch, A. (to appear). Cluster analysis and the identification of collexeme classes. In S. Rice and J. Newman (Eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford, CA: CSLI.

Kuo, S.-H. (2008). Discourse and aging: A sociolinguistic analysis of elderly speech in Taiwan (NSC96-2411-H007-024). Taipei, Taiwan (R.O.C.): National Science Council.

Li, C. I. (1999). Utterance-final particles in Taiwanese: A discourse-pragmatic analysis. Taipei, Taiwan (R.O.C.): Crane Publishing Co.

Lin, D. (1998). Automatic retrieval and clustering of similar words. in *Proceedings of the 17th International Conference on Computational Linguistics*, 768-774.

Mitrofanova, O., Mukhin, A., Panicheva, P., & Savitsky, V. (2007). Automatic word clustering in Russian texts. *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, 85-91.

Pereira, F. & Tishby, N. (1992). Distributional similarity, phase transitions and hierarchical clustering. *Probabilistic approaches to natural language, papers from 1992 AAAI Fall Symposium*, 108-112.

R. D. C. Team. (2010). R: A language and environment for statistical computing. Retrieved June 1, 2010 from: http://www.R-project.org

Redington, M., Chater, N., & Finch, S. (1993). Distributional information and the acquisition of linguistic categories: A statistical approach. in *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, 848-853.

Rush, B. K., Barch, D. M., & Braver, T. S. (2006). Accounting for cognitive aging: Context processing, inhibition or processing speed? *Aging, Neuropsychology and Cognition*, 13, 588-610.

Veliz, M., Riffo, B., & Arancibia, B. (2010). Cognitive aging and language processing: Relevant issues. *Revista de Lingüística Teórica y Aplicada*, 75-103.

Widdows, D. & Dorow, B. (2002). A graph model for unsupervised lexical acquisition. in *Proceedings of the 19th International Conference on Computational Linguistics*, 1093-1099.

Wilson, K. R. (2008). *The effects of cognitive load on gait in older adults* (Doctoral dissertation). Florida State University, Tallahassee, FL.