# Frequency, Collocation, and Statistical Modeling of Lexical Items: A Case Study of Temporal Expressions in an Elderly Speaker Corpus[1]

王聖富　Sheng-Fu Wang
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
sftwang0416@gmail.com

楊靜琛　Jing-Chen Yang
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
flower75828@gmail.com

張瑜芸　Yu-Yun Chang
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
june06029@gmail.com

劉郁文　Yu-Wen Liu
國立臺灣師範大學英語學系
Department of English[2]
National Taiwan Normal University
Yw_L7@hotmail.com

謝舒凱　Shu-Kai Hsieh
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

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

## Abstract

This study examines how different dimensions of corpus frequency data may affect the outcome of statistical modeling of lexical items. The corpus used in our analysis is an elderly speaker corpus in its early development, and the target words are temporal expressions, which might reveal how the speech produced by the elderly is organized. We conduct divisive hierarchical clustering based on two different dimensions of corpus data, namely raw frequency distribution and collocation-based vectors. Results show when different dimensions of data were used as the input, the target terms were indeed clustered in different ways. Analyses based on frequency distributions and collocational patterns are distinct from each other. Specifically, statistically-based collocational analysis produces more distinct clustering results that differentiate temporal terms more delicately than do the ones based on raw frequency.

Keywords: clustering, collocation, corpus linguistics, temporal expression, gerontology

## 1. Introduction

The study of gerontology has gained globe wide attention as the aging population becomes a grave issue in our society nowadays. Much research has noted that aging caused not only physiological changes for elderly people, but also effects on their language production [1], cognitive load [2], context processing speed [3], language performance patterns compared to younger individuals [4], etc. To research gerontology from a linguistic viewpoint, Green [5] proposed that the phenomenon of gerontology could be studied through discourse analysis. Therefore, we collect conversations from the elderly as our speech corpus, and take the corpus as input to exemplify the procedures and usage of lexical modeling.

The social roles of elderly people may be embedded in the conversation when they share personal experience or judgment of the past [6] and the present. Thus, we presume that some temporal expressions might pervade as the anchoring points in the conversation-based aging corpus and might help us reveal a certain aspect of the speech behavior pattern the elderly have.

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 which 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” [7]. 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 forms of dictionaries, thesauruses, and ontologies [8].

A glimpse on relevant studies would reveal that the most prominent kind of data input is related to the distributional patterns of the lexical items in corpora, no matter whether the

lexical items themselves are the target of the modeling or not. The distributional data could be in the form of words' frequencies and variability of frequencies [9] or the distribution of n-grams as a whole [10]. Distributional pattern or dependency with syntactic patterns is also a prominent source of data input [11]-[14]. Target lexical items' dependency and co-occurrence with particular word types may also be taken as the basis of lexical modeling in some studies [15]. Moreover, statistically-based collocational patterns are used for modeling similarities among lexical units of interest [16], [17].

The abovementioned different methods, or rather, different data inputs, are considered falling 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 input may result in different data in lexical modeling. With preliminary research like ours, we hope to make contributions to the path to full understandings of universal linguistic and cognitive patterns in the elderly's speech act.

This paper is organized as follows: Section 2 introduces the construction of the elderly speaker corpus, including data collection, guidelines for transcription, and annotation standards. Section 3 reports basic corpus 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 is the summary, including some implications of our findings.

## 2. Corpus construction

### 2.1 Data collection

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 with each other 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 digit recorder in the format of WAV. The total length of the speech samples is 145 minutes.

### 2.2 Transcription

Speech samples collected from the elderly's conversations were then transcribed into Chinese characters, following Du Bois' transcription standards for discourse analysis [18]. Because prosodic features and vocal qualities of the intonation units (IUs) were not the main interest in this study, the aforementioned information was excluded from the transcription. 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

Sometimes speech overlap happened during the conversation. These speech overlaps were indicated by square brackets, as shown in example (2). In order to indicate on the transcription when and where utterances overlap, the left brackets of the overlapping speakers' speech are aligned vertically. Double square brackets were used for more overlaps occurring in a rapid succession 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

As bilinguals, the subjects might shift from Mandarin, which dominated the conversation, to another language. Such utterance of code-switching was enclosed in square brackets and labeled with *L2* as well as the code for the non-Mandarin language. Example (3) demonstrates the transcription for code-switching, where the language code TSM represents Taiwanese Southern Min.

(3)

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

Laughter was also marked 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 was transcribed. Pauses are represented by dots: two dots for short pauses that are less than 0.3 seconds, three dots for medium pauses between 0.3 and 0.6 seconds, and three dots for pauses longer than 0.7 seconds with its duration specified in parentheses. 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 what 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 previous example (1) and (5).

**2.3 Annotation**

After all recorded samples were transcribed, the transcription would be automatically segmented and tagged with POS (part of speech) through the CKIP Chinese Word

Segmentation System provided by the Chinese Knowledge Information Processing (CKIP) group at the Academia Sinica [19]. The segmentation and POS standards were based on the Sinica Corpus guidelines [20]. The annotated language samples were then manually checked. The procedure is described below.

Firstly, every segmentation result derived from CKIP was examined, and corrected if wrong, as in the following examples. Example 7*a* is the original IU before segementation and tagging. Through CKIP, we get the result in example 7b, which is falsely processed. Example 7c shows the right 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

Secondly, POS tags were viewed as correct only 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*) is examined, but not the sub-class tagging, as we give less consideration for whether the POS tags should be N*a* or N*h*.

(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."

Thirdly, particles were identified as FW (for foreign word) in the CKIP system. These tags were manually corrected to *I* for IU-initial particles[3], and *T* for IU-final particles. If an IU contained nothing but particles, then the particles were tagged as *I*.

Lastly, POS tags were removed for truncations (e.g. 這--), uncertain hearing (i.e. X) and code-switching. Given that truncations were not generally viewed as lexical items, they were not suitable to be analyzed at lexical level. Considering this study targeted the elderly's Mandarin speech performance, code-switching phenomena were of less value for our analysis. Therefore, those tags were removed in these cases.

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

## 3. Corpus information & Preliminary analysis

This 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 (male: 2,267 IUs; female: 1,472 IUs), and there are 18,076 word tokens in total (male: 11,383 word tokens; female: 6,693 word tokens).

The corpus processing tool used here is R [21], 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 hold some vital clues to the elderly's speech pattern, so the following analyses will focus on the subjects' use of temporal expressions. By looking at word frequency, we first find that except for function words and pronouns, temporal expressions such as 現在 (now) and 以前 (before) are of high frequencies. This result is possibly influenced by the theme of the conversation assigned to the subjects. The term 現在 (now) expresses the speakers' concept of "the present," while 以前 (before) reveals their idea of "the past." We are interested in how elderly people use these two terms and other temporal expressions (tagged as Nd) to frame the present- and the past-related concept.

Six temporal expressions are selected for the analysis. Terms for the present-related concept are 現在 (now) and 最近 (recently); those for the past-related concept are 以前 (before), 小時候 (in one's childhood), 民國 (R.O.C. year), and 當初 (back then). Examining their frequency, we see that 現在 (now) and 以前 (before) appear most frequently, whereas other terms are seldom used by elderly speakers in this corpus. Table 1 lists the frequency of the six target temporal expressions.

Table 1. The frequency of six temporal expressions from elderly speakers in the corpus.

| Term | Frequency | Ranking |
|---|---|---|
| 現在(now) | 169 | 1 |
| 以前(before) | 169 | 2 |
| 小時候(in one's childhood) | 12 | 3 |
| 民國(R.O.C. year) | 11 | 4 |
| 當初(back then) | 9 | 5 |
| 最近(recently) | 6 | 6 |

## 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. It differs from

agglomerative hierarchical clustering in that a group of entities is first divided into large groups and then smaller groups are classified. Such a method is useful for finding a few clusters large in size [22]. We would like to find out whether the terms for "the present" and "the past" can really be grouped into clusters different in temporality. Thus, divisive hierarchical clustering serves our need.

We execute a series of hierarchical clustering with different data input. The first analysis is run with the frequencies of the temporal terms across different files/texts in our corpus. Such an input is 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 conversation topic. 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 any 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 corpus.

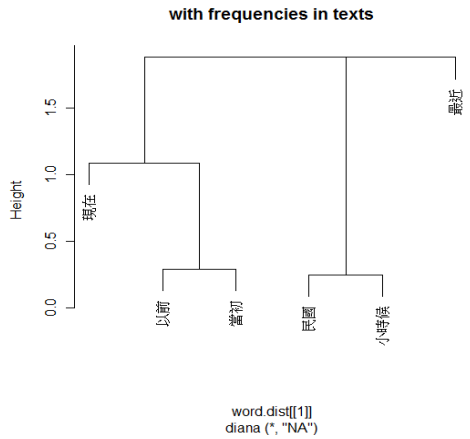


Figure 1. Clustering based on frequencies in texts

Next, four clustering analyses are made based on the frequency data across subsets of different sizes. The sizes chosen for producing subsets are 10, 50, 200, and 500 words respectively. 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. The results are shown in Figure 2. As we can see in the four graphs below, 現在 (now) and 以前 (before) are classified in the same small cluster. It is worth noting that 最近 (recently) is clustered independently with a subset size up to 200, which shows only when the subset is big enough can we see it grouped with terms related to the past.

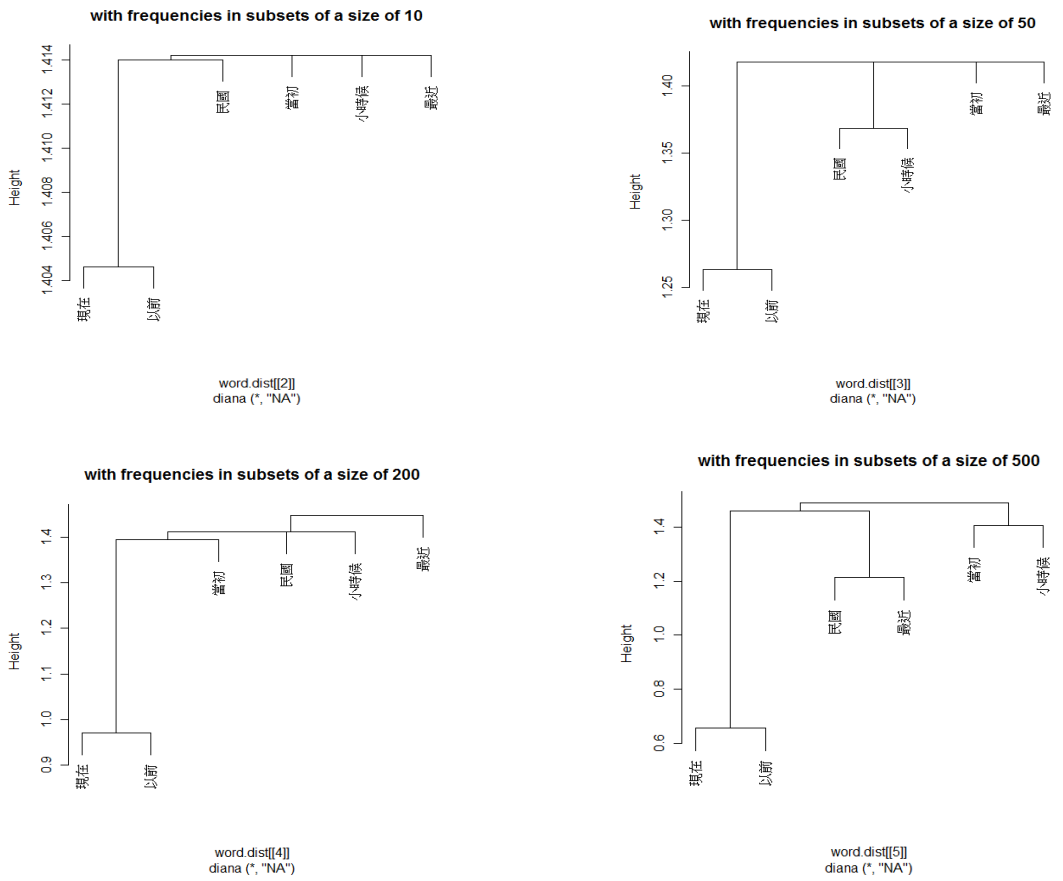


Figure 2. Clustering based on frequencies across subsets. Upper left, with subsets of a size of 10 words. Upper right, of 50 words. Lower left, of 200 words. Lower right, of 500 words.

The analyses above are obtained provided with the temporal terms' frequencies of occurrence in different parts of the corpus. In addition to this method, we can also do 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 are given based on this assumption. The first analysis is done by using each word type's collocational pattern (span = 3) with the six temporal terms as input. The second analysis is achieved through the dependency patterns of sentential particles (i.e. lah, hoNh, ah, oh, le, haNh, hioh, mah, as described by [23]), taking the temporal terms as its input. There are two reasons for the inclusion of particle collocation. Firstly, in regard to methodology, running more than one collocational test allows one to see whether collocational analyses with different approaches generate similar results. Secondly, 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 may be similar, regardless of the actual word types of their collocates. Noteworthily, 民國 (R.O.C. year) and 最近 (recently) are clustered together from other terms.

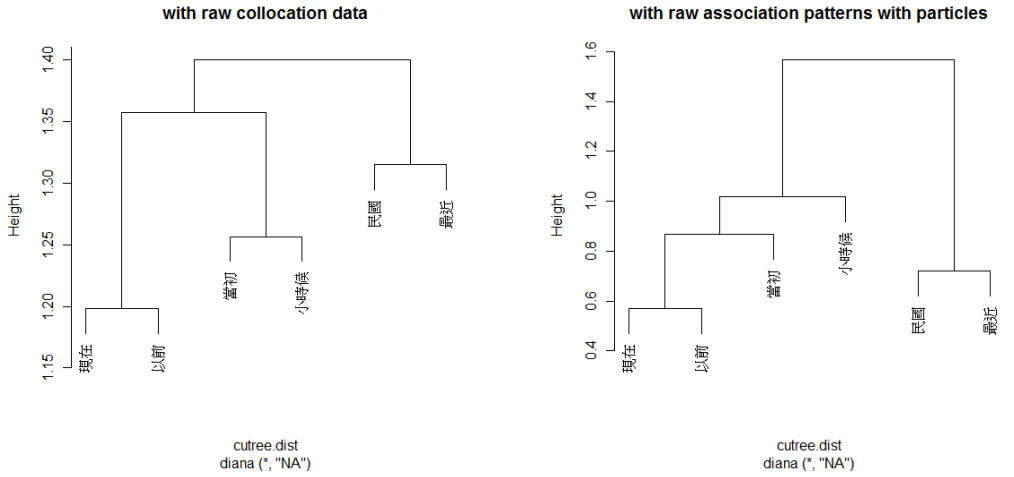


Figure 3. Clustering based on association/collocation frequencies. Left, with all word types in the corpus. Right, with particles.

Potentially, there is a problem 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 [24], [25], 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 is calculated and used as input for clustering analysis. The output is shown in Figure 4.

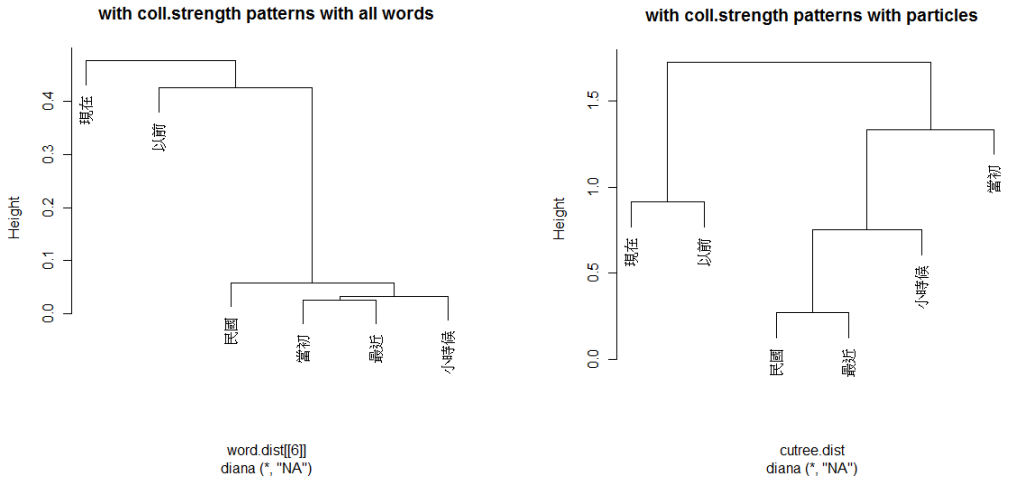


Figure 4. Clustering based on coll.strength patterns. Left, with all word types in the corpus. Right, with particles.

The next question is: How do we evaluate all 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: Each ‘tree’ is quantified in terms of which cluster an item is clustered to. We collect the data for all the trees shown above and execute clustering as meta-analysis. The outcome is shown in Figure 5.

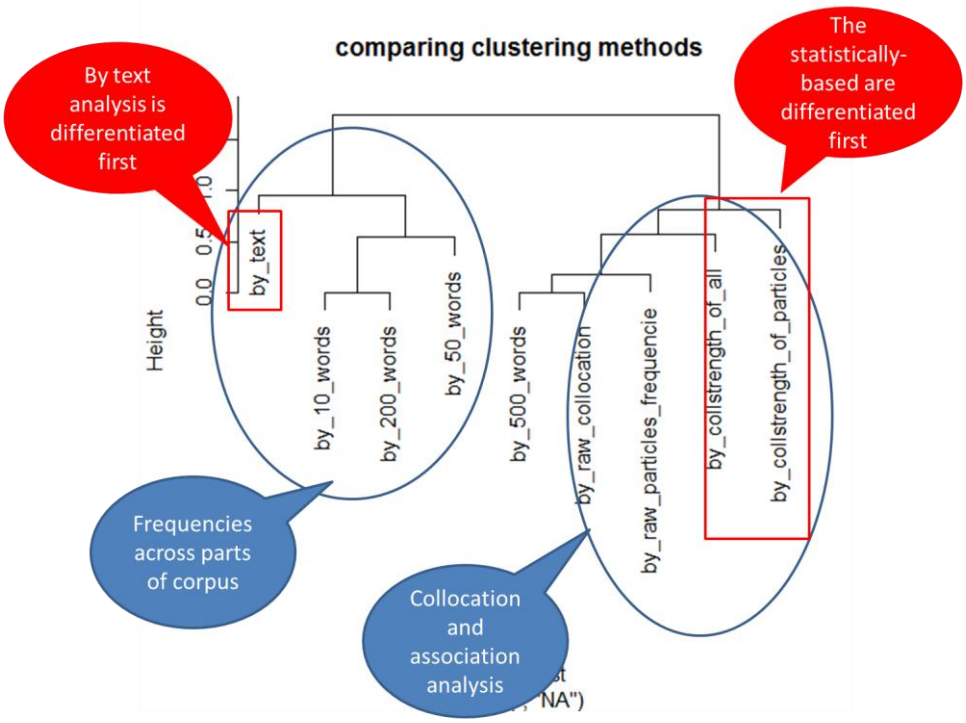


Figure 5. Clustering of various results with different types of input data

An interesting pattern shows up. There are two major clusters. The left one is based on frequency patterns of temporal terms, and the right one basically contains analyses regarding 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 being singled out. This conforms to our impression that temporal terms are clustered differently, with 現在 (now) and 最近 (recently) placed relatively away from other past-related expressions. Moreover, in the right major cluster, the analyses with coll.strength are the first ones being differentiated from the others. Again, it 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 5 is that, first, different types of data input certainly influence the outcome of clustering analysis, and second, the results of quantitative analysis can also be evaluated through quantitative analysis, just as how we use hierarchical clustering to analyze and evaluate results of hierarchical clustering.

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 that for the past, suggesting that some elderly might tend to converse about the present more than the past, or vice versa. Collocational strength analysis is another approach revealing a difference between 現在 (now) and 以前 (before), showing that although usually used closely, the two terms still attract different words with different strengths. It should be noted that association patterns with particles invested in the qualitative analysis 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 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 finding 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 nuance among lexical items.

## 5. Conclusion

Statistical modeling based on different types of data input does display 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. collocation 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 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, hierarchical modeling proposed here may help avoiding utilizing methods that are in fact very similar.

According to our analysis on temporal terms, the findings suggest that 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 in [24]. These findings can be seen as a pilot result on the linguistic pattern of aging people.

## 6. Future work

Our prime purpose of this study is 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. 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. At the present time, the elderly speaker corpus does not yet reach a big scale, and its expansion is desirable as the outcome of our statistical modeling could be altered if the corpus size increases, which might give us other insight into our study. Furthermore, we can work on the comparisons of younger speakers’ speech and that of the elderly’s, and combine what we find with theories and research in other fields of study, such as sociology and cognitive science, hoping to discover the relationship between language and aging in Taiwanese society.

## 7. References


[1] D. M. Bruke and M. A. Shafto, "Aging and language production," *Current Directions in Psychological Science,* vol. 13, pp. 21-24, 2004.

[2] K. R. Wilson, "The effects of cognitive load on gait in older adults," Ph. D., Department of Communication Disorders, Florida State University, 2008.

[3] B. Rush, *et al.*, "Accounting for cognitive aging: context processing, inhibition or processing speed?," *Aging, Neuropsychology and Cognition,* vol. 13, pp. 588-610, 2006.

[4] M. Veliz, *et al.*, "Cognitive Aging and Language Processing: Relevant Issues," in *Revista de Lingüística Teórica y Aplicada*, 2010, pp. 75-103.

[5] B. S. Green, *Gerontolgy And The Social Construction of Old Age*. New York: Aldine De Gruyter, 1993.

[6] S.-H. Kuo, "Discourse and Aging: A Sociolinguistic Analysis of Elderly Speech in Taiwan," National Tsing Hua University, 2008.

[7] D. Widdows and B. Dorow, "A Graph Model for Unsupervised Lexical Acquisition," in *Proceedings of the 19th International Conference on Computational Linguistics*, 2002, pp. 1093-1099.

[8] O. Mitrofanova, *et al.*, "Automatic word clustering in Russian texts," in *Proceedings of the 10th international conference on Text, speech and dialogue*, 2007, pp. 85-91.

[9] S. T. Gries and M. Hilpert, "Variability-based Neighbor Clustering: A bottom-up approach to periodization in historical linguistics," To appear.

[10] S. T. Gries, *et al.*, "N-grams and the clustering of registers," *Empirical Language Research Journal*, vol. 5, 2011.

[11] P. Cimiano, *et al.*, "Comparing Conceptual, Divisive and Agglomerative Clustering for Learning Taxonomies from Text," in *Proceedings of the European Conference of Artificial Intelligence*, 2004, pp. 435-439.

[12] P. Cimiano, *et al.*, "Clustering concept hierarchies from text," in *Proceedings of LREC 2004*, 2004, pp. 1-4.

[13] D. Lin, "Automatic retrieval and clustering of similar words," in *Proceedings of the 17th international conference on Computational linguistics*, 1998, pp. 768-774.

[14] F. Pereira and N. Tishby, "Distributional Similarity, Phase Transitions and Hierarchical Clustering," Association for the Advancement of Artificial Intelligence1992.

[15] M. Redington, *et al.*, "Distributional Information and the Acquisition of Linguistic Categories: A Statistical Approach," in *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, 1993, pp. 848-853.

[16] C.-H. Chen, "Corpus, Lexicon, and Construction: A Quantitative Corpus Approach to Mandarin Possessive Construction," *International Journal of Computational Linguigstics and Chinese Language Processing*, vol. 14, pp. 305-340, 2009.

[17] S. T. Gries and A. Stefanowitsch, "Cluster analysis and the identification of collexeme classes," in *Empirical and Experimental Methods in Cognitive/Functional Research*, S. Rice and J. Newman, Eds., ed Stanford, CA: CSLI, To appear.

[18] J. Du Bois, *et al.*, *Outline of discourse transcription*. Hillsdale, NJ: Lawrence Erlbau, 1993.

[19] CKIP. (2004). *CKIP Chinese Word Segmentation System*. Available: Retrieved June 2, 2011, from http://ckipsvr.iis.sinica.edu.tw/

[20] CKIP, "Introduction to Sinica Corpus: A tagged balance corpus for Mandarin Chinese," Academia Sinica, Taipei, 1998.

[21] R. D. C. Team. (2010). *R: A language and environment for statistical computing*. Available: http://www.R-project.org

[22] R. H. Baayen, *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*: Cambridge University Press, 2008.

[23] C. I. Li, *Utterance-Final Particles in Taiwanese: A Discourse-Pragmatic Analysis*. Taipei: The Crane Publishing Company, 1999.

[24] S. T. Gries, *et al.*, "Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions," *Cognitive Linguistics*, vol. 16, pp. 635-676, 2005.

[25] S. T. Gries. (2007). *Collostructional analysis: Computing the degree of association between words and words/constructions*. Available: http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/coll.analysis.r