B-2

A Reflection on the Clustering in Corpus Linguistics

Akitaka Yamada, Georgetown University

ay314@georgetown.edu

Agglomerative clustering is one of the promising unsupervised methods that give us a good insight on how the language is used. As is often the case with any quantitative methods in statistics, however, this technique needs a good adjustment when it is applied to a particular research field, based on the nature of the data collected in that specific field. Especially, the following two problems are discussed in this presentation --- issues not well-discussed in introductory textbooks of corpus linguistics (Baayen 2008; Johnson 2008; Gries 2013). First, this study shows how we compare different distance measures and draw a robust conclusion. Second, the study also discusses the information lost in the clustering algorithm and shows some auxiliary ways to supplement clustering analysis. These claims are exemplified in two case-studies from different research traditions in linguistics in order to show how unsupervised studies guide and affect the research in theoretical linguistics.

1. Introduction

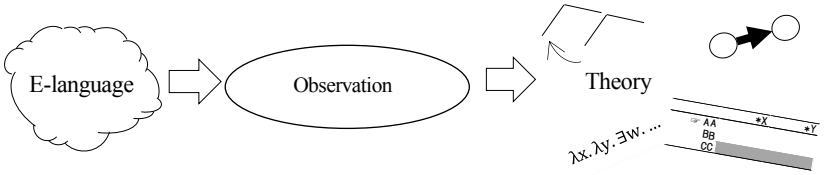
1.1 Linguistic inquiry

Explicitly or implicitly, any study of theoretical linguistics consists in the following two stages.

(1) Two steps in linguistic inquiry

a. To carry out observations

b. To build a theory



1.2 Need for exploratory quantitative linguistics

Sometimes, our intuition does not always give us an accurate observation due to VARIATION in the way the observation is made; *i.e.*, interpersonal variation (dialectal variation, … *etc.*) and intrapersonal variation (style, age, … *etc.*). The data, thus, needs to be quantified in an appropriate way. Corpus linguistics, by looking at many instances, tries to identify the general trend hidden in the data. The information extracted is usually summarized in the form of histogram (Figure 1).

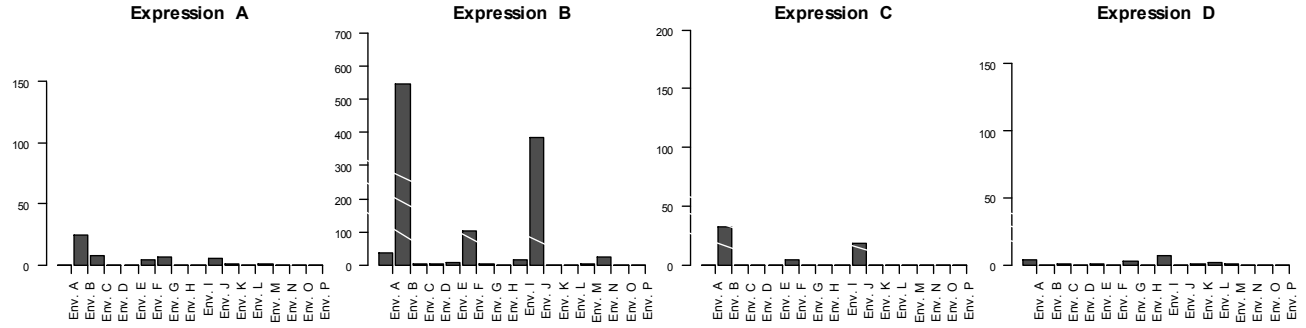


Figure 1. Histograms.

2. Hierarchical agglomerative clustering analysis

2.1 Examples: research questions

(2) **Example 1**: Tense and aspect *semantics/pragmatics*
What is the difference among the present, the present perfect, and the simple past in English?

(3) **Example 2**: Classification of complement-taking verbs *syntax (c-selection)*
What kind of verbs can take *-ka-to verb* construction in Japanese?

2.2 Example: results

Let us begin with the first example in (2) using the information extracted from COCA corpus. In order to classify verbs with respect to three tense/aspect forms (the present tense, the past tense and the present perfect), frequencies of these forms are counted for each verb, which are then transformed into the relative frequency After choosing an appropriate metric and method, hierarchical agglomerative clustering can be implemented. The following diagram is the result of the classification based on the Euclidean distance and the Ward's method.

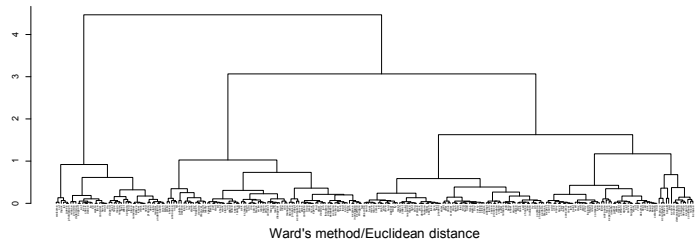


Figure 2. Dendrogram.

Most data points are merged to form a cluster within a short distance, leading to a high agglomerative coefficient, *i.e.*, 0.995, as if to say that this is an excellent result for a clustering. Despite this apparent success in agglomerative coefficient, however, the silhouette coefficient of this classification is not as high, *i.e.*, 0.39 (with $k = 3$).

2.3 Research question

As is often discussed, there is no ultimate tree in clustering analysis and researchers are required to be aware of the sensitivity of the results to prior assumptions that they have adopted. A natural question that arises in practice is (i) how we should cope with the subjectivity in metric selection, keeping in mind the nature of our data in corpus linguistics (Research Question A) and (ii) how we recover the information lost during clustering algorism (Research Question B).

3. Information geometry

In order to answer these questions, we need to formulate the general situation in mathematical terms. Consider a random variable *X* which takes *1, 2, …, n* and has a probabilistic distribution *P* (*i.e.*, $P(X = i) = P_i$). Suppose that we collect all possible probabilistic distributions and call it Space ***S***. Within this space, we can think of a set of probabilistic distributions we can get by changing the value of the parameter $\boldsymbol{\xi}$ and let this subset be $S = \{p_\xi\}$.

(4) $S = \{p_\xi | \boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_{n-1}) \in \Xi\}; \Xi \subseteq \mathbb{R}^{n-1}$

When we have observed a data, *i.e.*, a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, where x_i represents the relative frequency of the category i (*s.t.*, the two conditions in (5) are met), this empirical probabilistic distribution has a unique corresponding point on *n-1* dimensional space of S.

(5) For a vector $\boldsymbol{x}$, the *i-th* member x_i:

a. $0 \leq x_i$

b. $\sum_i x_i = 1$ (where $i \in \{1, \ldots, n\}$)

For example, when *n* is set to 3, due to the constraints given in (5), data points cannot distribute in a random manner as shown in (6)a. A set of possible *n*-1 dimensional multinomial probabilistic distributions have a corresponding region on the 2-dimensional plane, as seen in (6)b and (7)c (= S). Our goal is to scrutinize metrics that are suitable for this manifold.

(6) Example

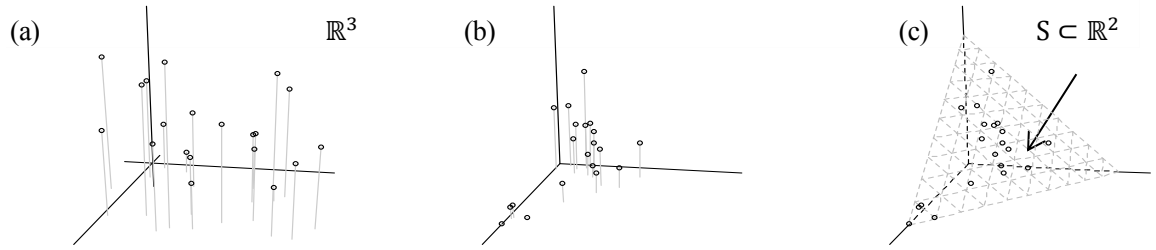


Figure 3. A set of probability distributions considered as a manifold with *n-1* dimension.

4. Distance between the Euclidean distance and our intuition

4.1 Euclidean distance

(7) Euclidian distance: $\mathrm{D_E}(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_j |x_j - y_j|^2\right)^{1/2}$

a. spherical b. symmetric

c. treating all dimensions equally d. sensitive to outliers

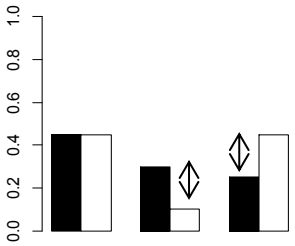


Figure 4. The Euclidean distance.

4.2 Example

For each pair of any data points (*i.e.*, in our case *verbs*), we need to define a particular discrepancy measure. For example, we need to decide whether the distance between SMILE and ANNOUNCE is greater or smaller than the distance between DECLINE and PUBLISH (*e.g.*, see Figure 5). The Euclidean distance claims that the degrees of discrepancy between the verbs in (a), (b) and (c) are almost the same, whereas the distance between INCREASE and EVOLVE is around three times the length of these three pairs. Suppose, however, that we did not observe the

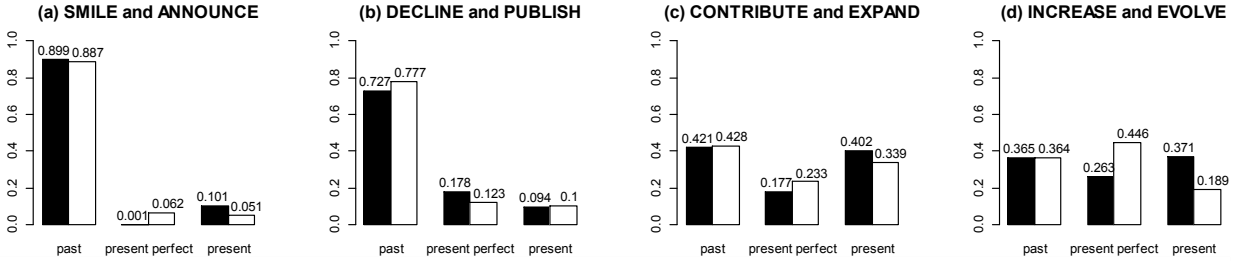


Figure 5. Comparison of relative frequency (I). *Three different environments are taken into account; the past tense, the present perfect and the present tense.*

frequency in the past tense (Figure 7). Under the Euclidean perspective, the distance in (d) is 0.53 times the length of the one in (a). Because the Euclidean distance respects all the dimensions equally and having a new dimension changes the values of the other dimensions, an addition of one dimension to the existing table would change the results to a great extent. For example, the preponderance of the past tense of SMILE and ANNOUNCE concealed their opposite tendency in present and present perfect uses. Though, in some cases, the nature of this metric is exactly what we want, in other cases (or, in most cases), we want to appreciate such a difference in (a) much more than we do for (b) and (c).

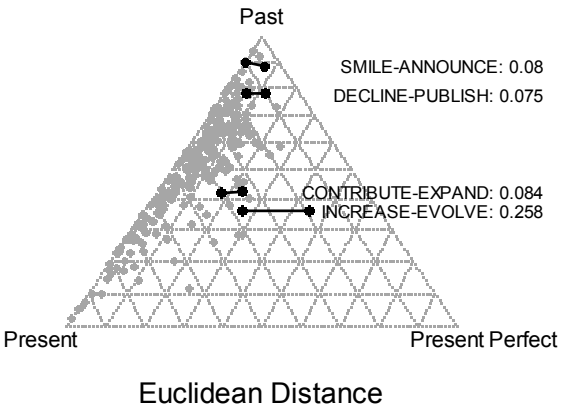


Figure 6. Discrepancy with respect to the Euclidean metric. *Notice that the first three pairs look almost the same, while the last pair shows the largest discrepancy.*

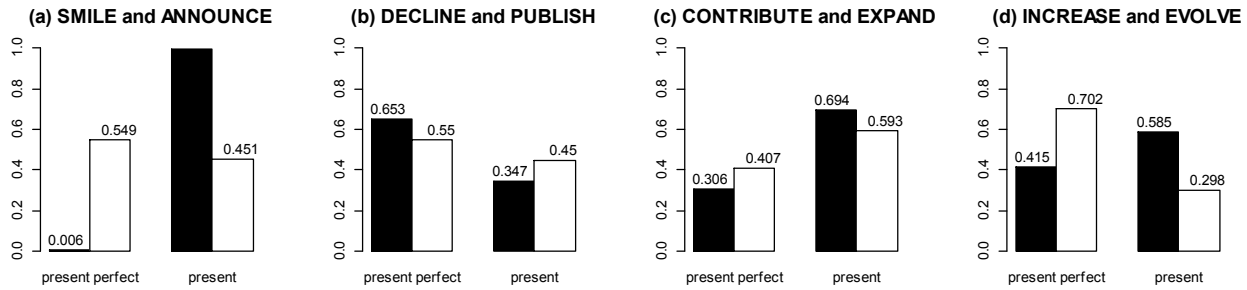


Figure 7. Comparison of relative frequency (II). *This is the results we would get if the past was not observed. Before looking at the use in the past tense, we conclude, based on the Euclidean distance, that SMILE and ANNOUNCE have the largest distance, which is completely denied by the interpretation we have in Figure 5.*

4.3 Hellinger distance

The Hellinger distance --- an alternative metric not extensively discussed in corpus linguistics --- is an instance of the α-divergence (with α set to 0) and, therefore, an instance of the *f*-divergence (Amari and Nagaoka 2000).

(8) Hellinger distance: $\mathrm{D_H}(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_j \left|\sqrt{x_j} - \sqrt{y_j}\right|^2\right)^{1/2}$

Intuitively speaking, this metric calculates the distance in the same way as the Euclidean distance does. However, it does this only after every single point on S is mapped onto a member in another set S' by a function *f* which gives the square root in every dimension (Figure 8).

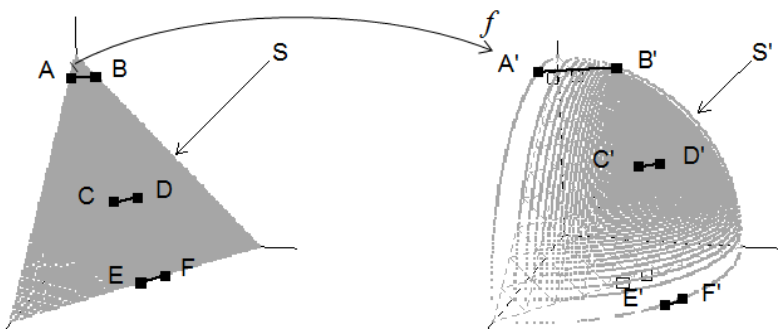


Figure 8. From the Euclidean distance to the Hellinger Distance. *For every point in S, there exists only one corresponding point in S' and vice versa. Nevertheless, the distance relation between any two data points in S is affected by this transformation. For instance, the distance between A' (= f(A)) and B' (=f(B)) is much larger compared to the distance between C' and D' and the distance between E' and F'.*

As we discussed above, even though SMILE and ANNOUNCE look very similar in one dimention, *i.e.*, the past tense, they show the opposite behavior with respect to the other two tense/aspect. This intuition is captured by the Hellinger distance. On the surface of S', the data points distribute as displayed in Figure 9. Unlike the Euclidean distance (= Figure 6), the Hellinger distance appreciates the contrast between SMILE and ANNOUNCE in the present and the present perfect and endows a larger distance to this pair than to the pair of DECLINE-PUBLISH and the pair of CONTRIBUTE-EXPAND. In fact, the distance is slightly larger than the distance between INCREASE and EVOLVE.

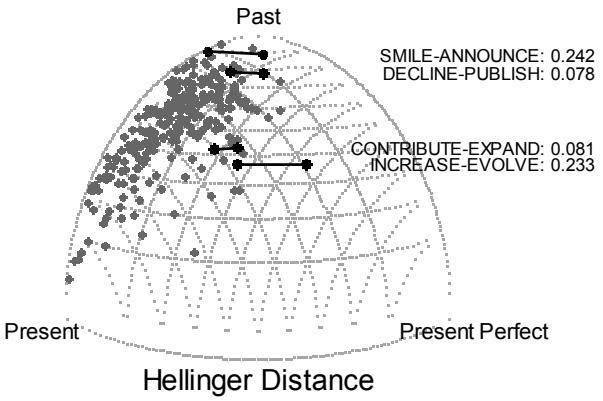


Figure 9. Discrepancy with respect to the Hellinger distance.

5. Example 1: Tense and aspect in English

The change in the metric results in a change in the dendrogram (Figure 10). The resulting change in partition is also shown in Figure 11. Clearly, under the Euclidean view, the spread along the PRESENT-PAST axis is the largest among the three arms of this triangle. The three major subclasses identified are (i) those used mostly in the present tense, (ii) those used mostly in the past tense and (iii) the rest. The transformation in the Hellinger distance, however, makes the difference in the present perfect use more salient, resulting in a different classification; (i) those with high relative frequency in the present tense, (ii) those with high relative frequency in the past and (iii) those whose use in the present perfect is relatively high among the verbs collected.

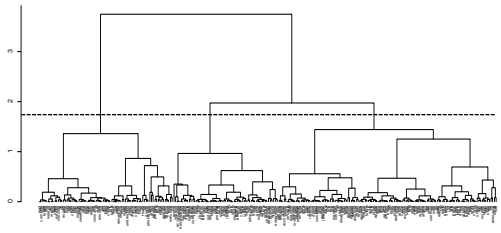


Figure 10. The Ward's method with the Hellinger distance.

The following lessons are worth our attention. First, the interpretation given above is drawn by looking at the scatterplot in Figure 11 and this interpretation is less easily drawn from the dendrogram. This is an answer to the second research question: *i.e.*, the scatterplots is an effective tool that recovers information lost in clustering analysis. Furthermore, the merest look at the scatterplot makes us realize some important extreme cases in distribution; *e.g.*, *evolve*, *guess* and *nod*. This realization is useful in building a theory in linguistics. For example, we can then ask why *evolve* is used so frequently in the present perfect whereas *nod* is reluctant to be used in the present perfect (Research Question B). Second, as we have seen, each metric is assigned a particular interpretation. The decision of one particular metric is the selection of a particular perspective from which we watch the E-language. Irrespective of the choice, there are some verbs which tend to be used in the past/present tense. This robust conclusion should be explained in whatever theory we are going to propose. In this way, comparison of different metrics tells us which verbs we should focus on in building a language theory (Research Question A).

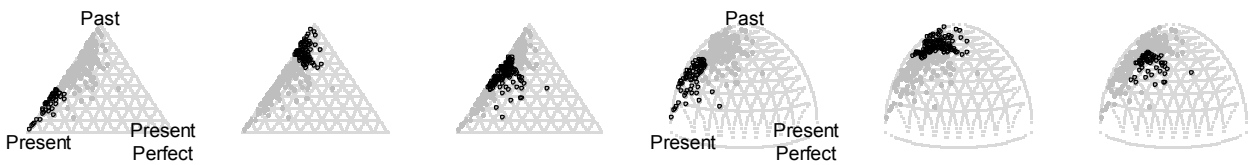


Figure 11. Change in partition. *The partition given by the Euclidean distance (left) and the partition by the Hellinger distance (right) based on the Ward's method.*

6. Example 2: *-ka-to* construction in Japanese

Unlike the example we have discussed so far, we, in practice, deal with data with a high dimension. For example, in order to answer the question in (3), 52 verbs are classified with respect to how often they are used in the following forms;*-ka+verb*, *-to+verb*, *-ka-to+verb* and *other postposition+verb* (the frequency information is extracted from BCCWJ). The presence of past tense, perfective, negation and politeness markers are also taken into account, leading to 112 features identified. In such a setting, one cannot visualize the data straightforwardly. A useful alternative approach is the MDS, a way of reconstructing a map that efficiently preserves the original distance but in a fewer dimension. In the scatterplot given below (ISO-map), the results of the six-way classification under each distance measure are visualized (with the Ward's used in the grouping process).

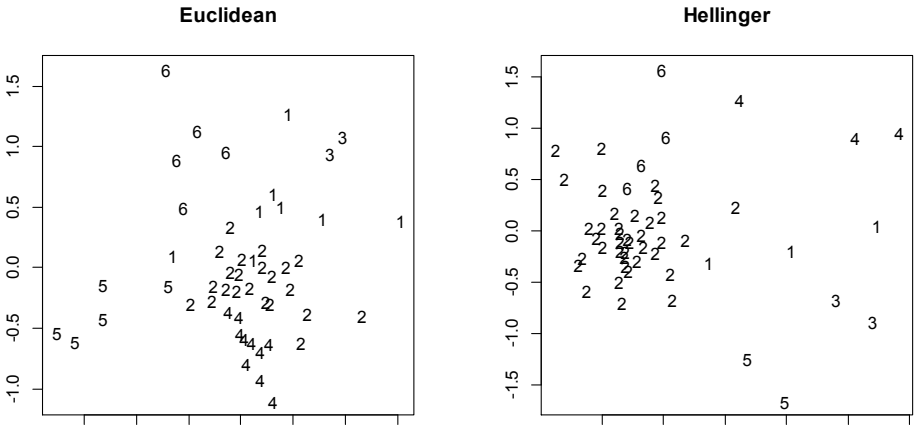


Figure 12. Multidimensional scaling. *Each plot shows six clusters identified based on the Euclidean and the Hellinger distance.*

The following table summarizes which verbs belong to which class under each decision. From this, we can understand the degree of robustness of clustering analysis.

	Euclidean	Hellinger		Euclidean	Hellinger		Euclidean	Hellinger		Euclidean	Hellinger
wakar-	1	1	*mir-*	2	2	*sinzur-*	2	2	*uttaer-*	4	2
sir-	1	1	*hanas-*	2	2	*omoikom-*	2	2	*iihar-*	4	2
sirer-	1	1	*kis-*	2	2	*kaisur-*	2	2	*tutomer-*	4	2
mier-	1	2	*katar-*	2	2	*minas-*	2	2	*iw-*	5	2
siraber-	1	3	*omoe-*	2	2	*mayow-*	3	4	*kanzir-*	5	2
tazuner-	1	3	*nober-*	2	2	*nayam-*	3	4	*ar-*	5	2
kakar-	1	5	*kak-*	2	2	*kiduk-*	4	2	*ie-*	5	2
simes-	1	5	*ronzur-*	2	2	*tuger-*	4	2	*sur- (archaic)*	5	2
kangaer-	2	2	*nar-*	2	2	*kokoromir-*	4	2	*komar-*	6	4
sur-	2	2	*kanzur-*	2	2	*tutaer-*	4	2	*omow-*	6	6
kimer-	2	2	*tok-*	2	2	*sator-*	4	2	*mitomer-*	6	6
kik-	2	2	*negaw-*	2	2	*tanom-*	4	2	*zonzur-*	6	6
osier-	2	2	*ossyar-*	2	2	*kotaer-*	4	2	*mous-*	6	6

Table 1. Difference in classification. *The Hellinger distance does not agree with the Euclidean distance in the shaded cells.*

Reference

Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry.* Oxford: Oxford University Press.

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

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction.* 2nd edition. Walter de Gruyter.

Johnson, K. (2008). *Quantitative methods in linguistics.* Blackwell Publishing.