

CHANGE OF BASIS METHODS

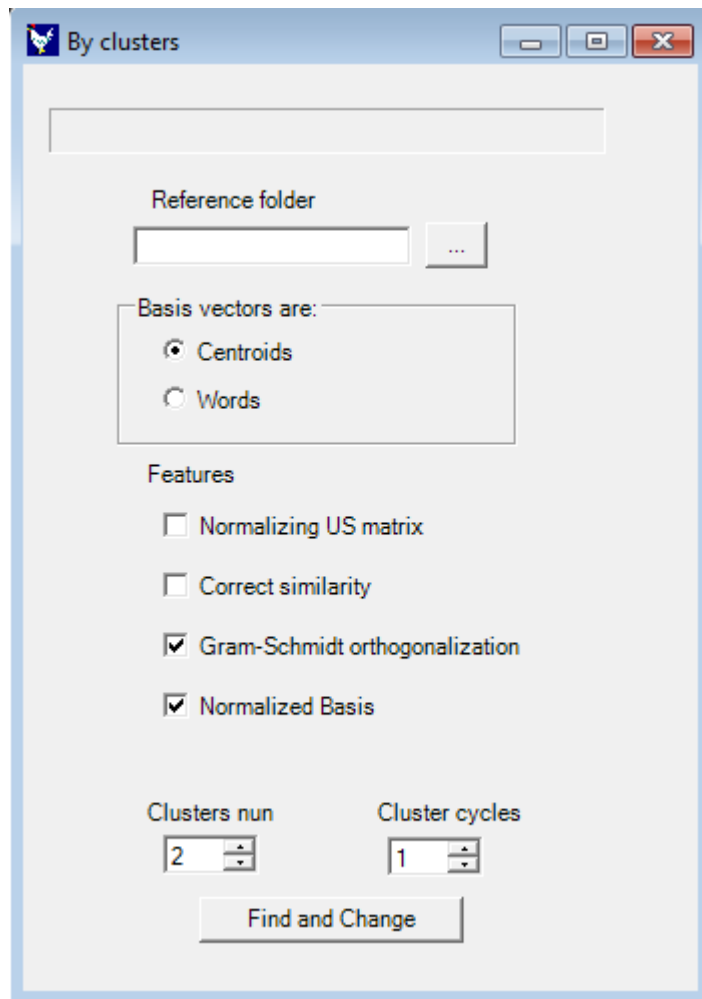
The *change of basis* procedure included in **Gallito 2.0** makes it possible to change the coordinate basis from which to represent the latent semantic space. As is well known, the dimensions of the latent semantic space lack an interpretation: they are pure abstractions, derived from the mathematical application *Singular Value Decomposition*. These arbitrary dimensions which arise by default account for most of the (weighted) variance in the occurrence matrix, but have no interpretation. **Gallito 2.0** provides users with a very interesting feature: changing the perspective of the latent semantic space and convert the abstract dimensions into real words.

There are two procedures to carry out a change of basis in Gallito: *by clusters* and *by predefined words*.

(1) By clusters

To effect a change of basis by clusters, go the following menu: Spaces → Change of basis → by clusters.

The idea underlying this procedure is the selection of a certain number of clusters with which to represent the latent semantic space.



The user must first choose the directory where the new matrices and the new files will be generated. To do so, the user must click on the button in **Reference folder**. Then he/she must choose the folder where said files will be generated. This will be the working folder.

Basic vectors are:

Basis vectors are Centroids: when selecting this option, **Gallito 2.0** runs a k -means cluster analysis on the terms in the semantic space with the selected number of clusters (see **Clusters num** later on). If k clusters are chosen, the procedure will group the terms in the semantic space into k groups.

The new vectors that will constitute the new basis are precisely the terms that represent those clusters or the vectors for those very same clusters (depending on the parameters set).

For example, if the first of the k clusters is constituted by the terms “*politics*”, “*president*”, “*government*”, and “*party*”, the vector that will serve to represent the new semantic space is generated by calculating the average for the vectors of those four terms.

With this option, the new dimensions with which the semantic space is represented can be based on semantically grouped words (not on a specific word). However, the cluster is assigned the label of the most representative term of the cluster, and this will be the term appearing in all the generated files (the term with the highest correlation to the average vector which does not belong to another cluster).

The number of clusters (k) to be used for the basis change must be specified in **Cluster num**.

This number can range between 2 and the number of dimensions in the term matrix. When a number of clusters is lower than the number of dimensions in the term matrix, the rest of the vectors of the basis will be filled in with vectors from the canonical basis.

Cluster cycles specifies the number of iterations performed by the k -means procedure to assign words to clusters. When the number is 1 the first assignation will be the one given. It should be taken into account that increasing the number of cycles increases the calculation time.

Basis vectors are Words: if this option is chosen instead of **Basis vectors are Clusters**, the procedure will work in the same way, but instead of creating the basis from average vectors (centroids) the words representing the clusters will be used directly.

Using the same example, if the words constituting the cluster are “*politics*”, “*president*”, “*government*”, and “*party*”, one of those words (the most representative one) will be used for the basis, not an average or centroid for those words. Representation of the dimension will no longer be a semantic grouping as is the case when choosing **Basis vectors are Clusters**, but rather a specific word. This makes it easier to interpret the dimension, but detracts slightly from cluster generality.

Other parameters:

Correct similarity: in order to assign the final label to the cluster, the cluster words can be previously weighted by their vector length (to favor labels that are more familiar in use).

Normalizing US Matrix: this option can be used to force the term matrix with which the cluster analysis is performed to be normalized. This means that all vectors have a norm of one, and words which appear more frequently or which have a longer vector are prevented from having a greater influence than the rest on the creation of the new basis.

Normalized Basis: this option (not to be confused with **Normalizing US Matrix**) normalizes the new final basis. Whereas the previous option is based on a normalized matrix to generate the new basis, this option forces the final basis to be normalized. Selecting it is practically mandatory.

Gram-Schmidt orthogonalization: by choosing Gram-Schmidt the user can force the final basis to be orthogonal (orthonormal if **Normalized Basis** has been previously selected). This generates an additional **GSreliability.txt** file in the previously specified

folder, the **Reference folder**. This file contains a column specifying which words in the new basis are reliable enough to serve as interpreters for all the terms in the space.

For example, when orthogonalization is forced by means of Gram-Schmidt, if the word “*tree*” is used to generate a dimension for the new semantic space, Gram-Schmidt will choose a term similar to “*tree*”, but not exactly *tree*. This is because the final basis is forced to be orthogonal. In order to find to what extent this term that is similar to *tree* is like the “old” *tree in the original space*, the GSreliability.txt file will specify the degree of correlation between the “old” tree and the substitute produced by Gram-Schmidt.

The advantage of Gram-Schmidt is that orthogonalization of the basis preserves 100% of the cosines in the former latent semantic space in the new latent semantic space. This is highly desirable if you want to calculate cosines, and not merely interpret the terms with respect to the new coordinates. When this procedure is used, the degree of de-virtualization of each substitute word should be taken into account for interpretation purposes. We recommend interpreting the dimension by means of the substitute word if it shares at least 50% of the variance with the real word (the equivalent of finding a minimum correlation of 0.70 in the GSreliability.txt file). See the following list as an example of the GSreliability.txt file for a basis created with 13 clusters on an 18-dimension semantic space. The items in this list will be the new dimensions for the new term matrix, and thus it will be possible to establish to what extent a given word is represented by the dimension “corruption”, “Spanish”, etc. The words which have been de-virtualized and thus do not serve to provide an interpretation are given in bold.

1 Corruption

0.953156796125899 spanish

0.966943010184085 water

0.923951103836911 ways

0.36566739446254 smoke

0.903705505768951 fuerteventura

0.776575604438863 economic

0.338408946494514 square

0.871168641428338 french

0.808075792310077 purchase

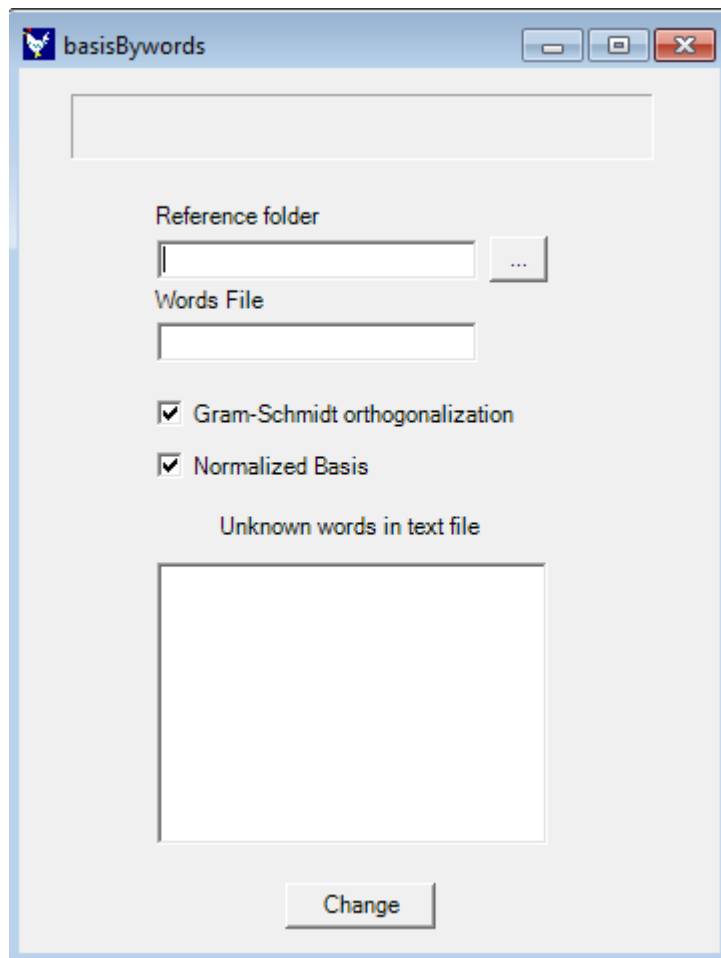
0.87014260844122 military


0.786023579624469 autonomy
0.77839435316233 suspension
0.344869326498221 ABSTRACT21
0.791818999023997 ABSTRACT22
0.947003051564806 ABSTRACT23
0.926487855328489 ABSTRACT24
0.874477881523237 ABSTRACT25

In addition, when using *Gram-Schmidt orthogonalization* a basisMatrixBeforeGsOrtog.txt file is obtained which contains the basis before application of Gram-Schmidt.

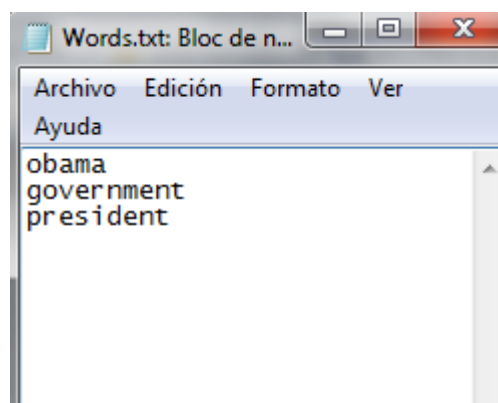
(2) By predefined words

The second method to change the basis is ***By predefined words***. This method is similar to the previous one insofar as the aim is to change the latent semantic space by means of a new basis. The difference lies in the fact that in this procedure the user chooses the words that will constitute the basis to calculate the new semantic space.




Reference folder: as before, the user must specify the folds where the matrices created by the procedure will be generated. To do this, click on the  button in Reference folder.

Words file: when using this option, the user must specify the name of the file containing the list of the words chosen by the user to constitute the new basis (this must be a plain text file, i.e., *.txt; *.dat). For example, if I specify a word list to constitute the new basis in a file with a .txt extension, this file must include the words, separated by a line break (see figure below).



In addition, it is important to accurately specify the file name and its extension in the *Words file* text box (the name must be without white spaces):



The image shows a screenshot of a software interface with two input fields. The first field is labeled 'Reference folder' and contains the text 'C:\Users\guille'. To the right of this field is a small square button with three dots '...', which is a standard file selection icon. The second field is labeled 'Words File' and contains the text 'words.txt'.

Finally, when words are included in the file, it is important for them to be written as they are in the latent semantic space. “*President*” is not the same thing as *president* (the program is case-sensitive).

Normalized Basis: this option normalizes the new final basis.

Gram-Schmidt orthogonalization: by choosing Gram-Schmidt the user can force the final basis to be orthogonal (orthonormal if **Normalized Basis** has been previously selected). Please see the description of this option in the previous section.

(3) Outputs of both procedures

- ***GSreliability.txt***

This file is only generated when the Gram-Schmidt option has been previously selected. In this file, the correlation between the vectors in the former basis and the vectors generated by the Gram-Schmidt procedure are stored. This file shows to what extent the Gram-Schmidt preserves the characteristics of the vector of the word chosen to create the new basis. The fact that the correlation between the former and the new vectors is not equal to one is due to the fact that Gram-Schmidt forces the vectors constituting the new basis to be orthogonal. As a recommendation, values lower than 0.70 should prevent the user from interpreting the new dimension by means of the chosen word, as this means that the vector generated by Gram-Schmidt shares less than 50% of the variance with the original word ($0.70^2 \approx 0.50$).

- ***basisMatrixBeforeGsOrtog.txt***

This file contains the vectors that constitute the new basis before the orthogonalization provided by Gram-Schmidt (it is a non-normalized basis). When selecting Gram-Schmidt this is not the basis used for generation of the new semantic space, but rather the basis that would have been used if the Gram-Schmidt option had not been selected.

- ***basisMatrix.txt***

This file contains the vectors that constitute the new basis, with which the latent semantic space has been generated. If Gram-Schmidt has been selected, this basis will be orthogonal. If *Normalized basis* has been selected, this basis will also be orthonormal (orthogonal vectors with modulo equal to one).

- ***newTermMatrix.txt***

This file represents the new latent semantic space. It contains all the processed terms in the corpus, with coordinates that correspond to the choice of the words specified (if the new basis has been established by means of a list of terms) or else to the set of clusters selected by means of the option of creating a basis from a cluster analysis.

- ***newTermMatrix.bnl***

This is the term file containing the new representation given by the basis, which can be loaded and used in **Gallito 2.0** after changing the basis.

- ***oldTermMatrix.txt***

This file contains the former latent semantic space whose coordinates are expressed in the dimensions provided by the *Singular Value Decomposition* procedure.

- *clusters.txt*

This file determines the assignation of each word to the clusters. It is only generated in the “**by clusters**” procedure.