Running head: ANN MULTILINGUAL TEXT PATTERN RECOGNITION

Unsupervised artificial neural networks for pattern recognition in multilingual text

Carolyn Evans (Corresponding Author)
Department of Communication
University at Buffalo, The State University of New York
359 Baldy Hall, Buffalo, NY 14260
Fax: 716-645-2086, Phone: 716-645-2141, c5a@buffalo.edu

Hao Chen
Department of Communication
University at Buffalo, The State University of New York
359 Baldy Hall, Buffalo, NY 14260
Fax: 716-645-2086, Phone: 716-645-2141, hchen4@buffalo.edu

Brenda Battleson
Department of Library and Information Studies
University at Buffalo, The State University of New York
522 Baldy Hall, Buffalo, NY 14260
Fax: 645-3775, Phone: 716-645-2412, x1210, blb@buffalo.edu

Joseph K. Wölfel
Talkhouse, LLC,
18 5th Ave.
Watertown, Massachusetts, 02472
Phone 617-393-0170, joe@talkhouse.com

Joseph Woelfel
Department of Communication, University at Buffalo
The State University of New York
339 Baldy Hall, Buffalo, NY 14260
Fax: 716-645-2086, Phone: 716-645-2141 x1188, listiac@aol.com

We are grateful for the assistance of H. Chun, P. Dastidar, J. Golzy, C. Chung, K. Kwon, D. Lim, S. Moon, and V. Tickoo

Rah Press

written in 2008

Abstract

Neural networks are able to discover patterns in data; it need not be textual data. When an artificial neural network is used to analyze textual data, however, it does not need grammar, heuristics, parsing or other linguistic artifacts to recognize, store, and retrieve clusters. This is because what the system searches for is not words per se, but rather, recurrent patterns in the bitstream that represents the words. Unlike software that examines ASCII coded documents, the present study utilizes software that processes Unicode documents. Examining Unicode documents extends the possibility of computer-aided text analysis to languages such as Chinese, Farsi, Hindi, and Korean that are pictorial rather than alphabetic.

Artificial Neural Networks for Pattern Recognition in Multilingual Text

An Artificial Neural Network (ANN) is able to discover patterns in data; it need not be textual data. When an ANN is used to analyze textual data, however, it does not need grammar, heuristics, parsing or other linguistic artifacts to recognize, store, and retrieve clusters; the ANN merely searches for recurrent patterns in the bitstream. This allows the system to consider not words per se, but rather, patterns in the stream of bits that represents the words (Harvey, 2005). This is consistent with Whorf's idea that the patterns of sentence structure are more important than individual words (1956). Further, since linguistic rules, grammatical rules, and heuristics are particular to the specific language for which they have been written, not considering these factors allows useful analysis of multiple languages.

The present study utilizes software able to process Unicode files. This is important as American Standard Code for Information Interchange (ASCII) is "an encoding technique which assigns a number to each of the 128 characters used most frequently in American English" (Indiana University, 2008) whereas Unicode assigns a unique number for each character in every alphabet, as well as accommodating non-alphabetic characters (Indiana University, 2008; The Unicode Consortium, 2009). Examining Unicode files thus extends the possibility of text analysis to languages such as Chinese, Farsi, Hindi, and Korean that are pictorial rather than alphabetic.

This ability to analyze diverse languages is important because if language reflects culture, or culture is shaped by language, then different cultures may use language differently. This may be suggested by discrepancies in the language patterns discovered when analyzing different languages. It was hoped that by using an ANN able to examine multilingual texts, different

languages could be compared with each other to explore this idea of discrepancy between languages and thus, potentially, cultures.

Originally designed to mimic the basic functions of organic neurological processes and simulate human pattern recognition, storage, and retrieval, Wölfpak was not primarily developed to conduct text analysis. It was developed, as was its predecessor Catpac that analyzes ASCII coded text, as a pure research project into the way the brain functions. Their usefulness for text analysis was evident only after the fact and does not support conventional text analysis theory and method. Indeed professional text analysts may note that ANNs often don't easily support many of the procedures they normally utilize (for example precoding techniques such as stemming).

For researchers not committed to conventional text analysis theory and method, however, or researchers with a large amount of text, an ANN such as Wölfpak may be the best choice.[1] In addition to allowing inspection of a greater amount of text than generally possible when utilizing the expertise of human coders, utilizing an ANN is also felt to be useful to guard against bias. A researcher may potentially be biased and not realize it; the choice of coding scheme may reflect a researcher's background and obscure unexpected results, thus potentially dismissing patterns an ANN can detect. At such times a neural network program may thus discover patterns not otherwise considered. An ANN may also be used to confirm a coding scheme when a researcher must work alone.

Artificial Neural Networks

Both supervised and unsupervised ANNs have been used to consider a wide variety of topics and textual inputs such as online chat (Kucukyilmaz, Cambazoglu, Aykanat, & Can, 2008;

[1] For a more complete discussion of text analysis in general, especially as related to computer aided text analysis, see http://www.textanalysis.info/

Rosen, Woelfel, Krikorian, & Barnett, 2003), organizational structure (Doerfel & Barnett, 1999), authorship (Bosch & Jason, 1998; Tweedie, Singh, & Holmes, 1996), online mass media (Tian & Stewart, 2005; Wölfel, Hsieh, et al., 2005), open-ended survey questions (Zywica & Danowski, 2008), and intercultural communication (Arasaratnam & Doerfel, 2005; Wölfel, Chen, et al., 2005). The primary advantages of single pass unsupervised ANNs[2] such as Catpac and Wölfpak are speed, independence from analyst bias, no pre-coding requirement, and a variety of display options for results (e.g., dendograms, perceptual maps, network diagrams, lists) (Wölfel, 1998).

Many programs and algorithms for computerized text analysis exist; it is therefore convenient to divide them into two broad types: rule-based analysis and propinquity analysis. These methods differ among themselves widely but share the notion that the correct interpretation of language depends on some rules or schemes, either learned or genetic (Brill & Mooney, 1997; Chang, Dell, & Bock, 2006; Majewski & Zurada, 2008). Rule based text analysis consists of those types of analysis based on linguistic, syntactic, grammatical, or other schemes of analysis.

Researchers favoring propinquity based analysis, on the other hand, consider grammar, syntax, rules and the like to be devices invented by analysts rather than the basis by which individuals interpret language. These researchers assume instead that words tend to become associated in meaning simply because they frequently occur "close" to each other in discourse (Danowski, 2007). This is consistent with a Hebbian learning model and was demonstrated biologically by Doty (as recounted by Kandel) in an experiment that showed connection "simply requires the pairing of two stimuli" (Kandel, 2006, p. 161). The idea that when President Nixon

[2] http://www.ncgia.ucsb.edu/giscc/units/u188/u188.html is a good source for details on the difference between supervised and unsupervised networks if this terminology is unfamiliar; "single pass" means that the ANN reads through the document once.

said "(T)he American People deserve to know whether their president is a crook. Well, I am not a crook…" the "not" was meaningless and the concepts of "Nixon" and "crook" were associated thereafter illustrates this viewpoint (Woelfel, 2008).

The oldest form of propinquity analysis is co-occurrence analysis and the most elementary form of that type of analysis is what Danowski (2007) calls the "bag of words" approach[3]. In this form of analysis, the number of times words co-occur in the same "bag" (e.g., document, page, episode, utterance, etc.) is counted, and a matrix of frequencies of co-occurrence is computed. This co-occurrence matrix is the basis of all further analysis.

Examples of such programs are: WORDij (Danowski, 1982), which was used to analyze co-occurrences of words in a Computer Bulletin Board; Newton (Newton, Buck, & Woelfel, 1986), which constructed co-occurrence matrices based on the co-occurrence of behaviors in 15 second intervals of prime time TV shows in five countries; and Catpac (Woelfel, 1993b; Wölfel, 1998), which has been used extensively in diverse subject areas worldwide (Allen, 2005; Awe & Crawford, 2003; Rezaei-Moghaddam, Karami, & Woelfel, 2006; Rosen, et al., 2003; Ryan, 1998; Tian & Stewart, 2005). Initially Catpac computed co-occurrences in text using a "bag of numbers" approach, with the beginning and end of each "bag" determined by numeric codes embedded in the text. Later, Danowski's sliding window method was implemented and then, in the late 1980's, an interactive activation and competition (IAC) ANN was added.

The IAC network in Catpac creates a set of artificial neurons, one for each word in the text. Then it "activates" those neurons whose associated words are in the sliding window at each iteration. The "connections", that is, the degree of closeness or propinquity, of those neurons

[3] Woelfel refers to the bags as "cases" or "episodes", and the Catpac software and manual refer to the "bag of words" method as "case delimited mode."

which are co-present in the window are then incremented. As the network grows and connections are established among the neurons, activation of neurons in the window can result in the activation of other neurons, not in the window, that are positively connected to those in the window. Connections among these neurons are also incremented[4].

A matrix of weights of the strength of these connections is then generated. In this manner, the propinquity of nodes is not established simply on the basis of pairwise co-occurrences, but on the basis of both direct pairwise relations and complete n-way indirect relationships among all the nodes. These deeper, more finely detailed relationships among the nodes are shown by more detailed dendograms and perceptual maps than those generated by co-occurrence (Chen, Evans, Battleson, Zubrow, & Woelfel, 2008).

As mentioned previously, grammar as such is irrelevant although individual languages may display regular "grammatical" patterns. The set of activation values of the nodes of a network at any given time may be considered a "pattern." If a given pattern of activations is frequently presented to the network as input, the nodes that make up that pattern will become fairly tightly connected. One result of these interconnections is that activating a large enough subset of nodes in a pattern will usually cause the remaining nodes to become active as well. This means that the total pattern may be stored in the network and retrieved later by activating only parts of the total pattern, rather than the entire pattern (Woelfel, 1993a). Indeed it is largely this trait and the ability to learn that distinguishes ANNs from co-occurrence models.

Overall Research Task

[4] For more details on the complete operation of IAC networks (including forgetting, normalization, and other issues not discussed here) see Woelfel 1993a, 1993b, and http://www.itee.uq.edu.au/~cogs2010/cmc/chapters/IAC/

Since ANNs can recognize underlying patterns in whatever is being represented and display those patterns in graphic ways to assist human interpretation, this study utilized an unsupervised IAC neural network for analysis of text. Different language versions of the United Nations' Universal Declaration of Human Rights were compared with each other in an attempt to discover whether language difference affected the translations. Observations were presented graphically. It was hoped that one could determine whether a single document translated into multiple languages said the same thing in the same way in each language; also, whether the particular computer program utilized (Wölfpak) could detect such similarities and differences if they existed.

Method

Unicode text files available at http://unicode.org/udhr/ were downloaded and initially evaluated by Wölfpak without using an exclude file (an exclude file is a list of words the researcher does not wish the neural network to process during analysis). Bengali and Tamil translations were evaluated but due to technical problems experienced by the consultants in India that portion of the project was discontinued. Greek and Russian translations were also initially evaluated but not explored further at this time.

Chinese, English, Farsi, Hindi, and Korean translations were originally chosen for primary analysis in the present study. Native language speakers were consulted to assist with the creation of appropriate exclude files, then the text was again evaluated with Wölfpak utilizing those exclude files. This process was done repeatedly in many of the languages. Since the Farsi native speaker became unavailable after the initial exclude file was developed, evaluation for that language ended at that time.

Results

Radial (see Figures 1-4), tree map (see Figures 5-8), and dendogram (see Figures 9-12) output was then generated for the remaining four primary languages. The radial output labels in each language were slightly modified to allow better viewing of all clusters and labels. Final output was saved as images from the files created by Wölfpak (file extension .wpak) and resized in Adobe Photoshop[5].

The .wpak files created by Wölfpak are formatted as XML files; accordingly, they can be used with any software capable of working with XML files. It should be noted that both the radial and tree map output are interactive when viewed within the Wölfpak program. The radial view will rearrange labels in relation to a chosen label; the tree map view will display the word associated with any particular box the mouse is hovering over.

Discussion

Initial evaluation of the output does not suggest any clear finding. The number of total clusters observed in the radial output is similar for all five languages, as are the translations of some of the clusters in the dendogram output (see Table 1), yet the languages do seem to cluster somewhat differently. Whether this is due to language, however, remains unclear at this time. It is now felt that the United Nations' Universal Declaration of Human Rights was not ideal for this project since it is short and written formally, rather than in narrative format. As noted by Woelfel (1993a), "…every text may not contain sufficient information to produce a useful cluster analysis or perceptual map, regardless of the sophistication of the analytic tools" (p. 78).

Preliminary analysis was subsequently conducted using newspaper articles and short stories in Chinese and English, a selection from the Bhagavad-Gita in Hindi and English, and a website

[5] GIMP, an open source program, or PAINT may be substituted for Photoshop.

available in both Korean and English. These documents appeared to form more intelligible clusters than those generated in response to the United Nations' Universal Declaration of Human Rights.

In addition to considering the same documents translated into different languages, preliminary analysis was also conducted on short narratives containing multiple languages within the same document using a single exclude file containing terms from both languages (specifically Korean & English and Hindi & English). Intelligible clusters were successfully generated for both languages under such circumstances.

In the future it is recommended that long narratives, or groups of narratives, be investigated as such documents appear to display more useful clustering patterns than those generated by the United Nations' Universal Declaration of Human Rights. The assistance of native speakers during the analysis phase is also suggested.

References

Allen, S. (2005). Using perceptual maps to communicate concepts of Sustainable Forest Management: Collaborative research with the Office of the Wet'suwet'en Nation in British Columbia. *The Forestry Chronicle, 81*(1), 381-386.

Arasaratnam, L. A., & Doerfel, M. L. (2005). Intercultural communication competence: Identifying key components from multicultural perspectives. *International Journal of Intercultural Relations, 29*(2), 137-163.

Awe, C., & Crawford, S. Y. (2003). Perceptions of campus experiences by African-American pharmacy students based on institutional type. *American Journal of Pharmaceutical Education, 67*(1), 80-90.

Bosch, R. A., & Jason, A. S. (1998). Separating hyperplanes and the authorship of the disputed Federalist Papers. *The American Mathematical Monthly, 105*(7), 601-608.

Brill, E., & Mooney, R. (1997). An overview of empirical natural language processing. *AI Magazine, 18*(4), 13-24.

Chang, F., Dell, G., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234-272.

Chen, H., Evans, C., Battleson, B., Zubrow, E., & Woelfel, J. (2008, January). *Procedures for the precise analysis of very large textual datasets.* Paper presented at the International Network for Social Network Analysis (INSNA) conference, St. Pete Beach, FL.

Danowski, J. (1982). A network-based content analysis methodology for computer-mediated communication: An illustration with a computer bulletin board. *Communication Yearbook, 6*, 904-925.

Danowski, J. (2007, May). *Comparisons of word-network software*. Paper presented at the International Network for Social Network Analysis (INSNA) conference, Corfu, Greece.

Doerfel, M., & Barnett, G. (1999). A semantic network analysis of the International Communication Association. *Human Communication Research, 25*(4), 589-603.

Harvey, C. (2005, May). Unicode Issues. Retrieved April 3, 2009 from http://languagegeek.org/issues/unicode.html

Indiana University (2008, April). What are the differences between ASCII, ISO 8859, and Unicode? Retrieved March 30, 2009, from http://kb.iu.edu/data/ahfr.html

Kandel, E. (2006). *In search of memory: The emergence of a new science of mind*. New York: W. W. Norton & Company.

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2008). Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management, 44*(4), 1448-1466.

Majewski, M., & Zurada, J. (2008). Sentence recognition using artificial neural networks. *Knowledge-Based Systems, 21*(7), 629-635.

Newton, B. J., Buck, E. B., & Woelfel, J. (1986). Metric multidimensional scaling of viewers' perceptions of TV in five countries. *Human Organization, 45*(2), 162.

Rezaei-Moghaddam, K., Karami, E., & Woelfel, J. (2006). The agricultural specialists' attitudes toward alternative sustainable agricultural paradigms: A Galileo method analysis. *Journal of Food, Agriculture & Environment, 4*(2), 310-319.

Rosen, D., Woelfel, J., Krikorian, D., & Barnett, G. (2003). Procedures for analyses of online communities. *Journal of Computer-Mediated Communication, 8*(4).

Ryan, C. (1998). Saltwater crocodiles as tourist attractions. *Journal of Sustainable Tourism, 6*(4), 314-327.

The Unicode Consortium (2009, January). What is Unicode? Retrieved March 30, 2009, from http://unicode.org/standard/WhatIsUnicode.html

Tian, Y., & Stewart, C. M. (2005). Framing the SARS crisis: A computer-assisted text analysis of CNN and BBC online news reports of SARS. *Asian Journal of Communication, 15*(3), 289-301.

Tweedie, F. J., Singh, S., & Holmes, D. I. (1996). Neural network applications in stylometry: The *Federalist Papers*. *Computers and the Humanities, 30*, 1-10.

Whorf, B. (1956). *Language, Thought, and Reality* (5th ed.). Cambridge: The MIT Press.

Woelfel, J. (1993a). Artificial neural networks in policy research: A current assessment. *Journal of Communication, 43*(1), 63-80.

Woelfel, J. (1993b). *Galileo*CATPAC: User manual and tutorial*. Amherst, NY: Terra Research.

Woelfel, J. (2008). *Procedures for precise comparisons of text: Orthogonal procrustean rotation of Galileo spaces generated by a neural network*. Unpublished manuscript, University at Buffalo, The State University of New York, Amherst, New York.

Wölfel, J. K. (1998). *User's guide Catpac II: version 2.0*. Amherst, NY: RAH Press.

Wölfel, J. K., Chen, H., Kim, J., Murero, M., Sharma, B., Woelfel, J., et al. (2005, October). *Artificial neural networks for the analysis of text across cultures and languages*. Paper presented at the International Communication Association (ICA) conference, New York, NY.

Wölfel, J. K., Hsieh, R., Chen, H., Hwang, J., Cheong, P., Rosen, D., et al. (2005, February). *Wölfpak: A neural network for multilingual text analysis*. Paper presented at the International Society for Network Analysis (INSNA) conference, Los Angeles, CA.

Zywica, J., & Danowski, J. (2008). The faces of facebookers: Investigating Social Enhancement and Social Compensation hypotheses; predicting Facebook™ and offline popularity from sociability and self-esteem, and mapping the meanings of popularity with semantic networks. *Journal of Computer-Mediated Communication, 14*(1), 1-34.

Figure 1. Radial view output for Chinese.

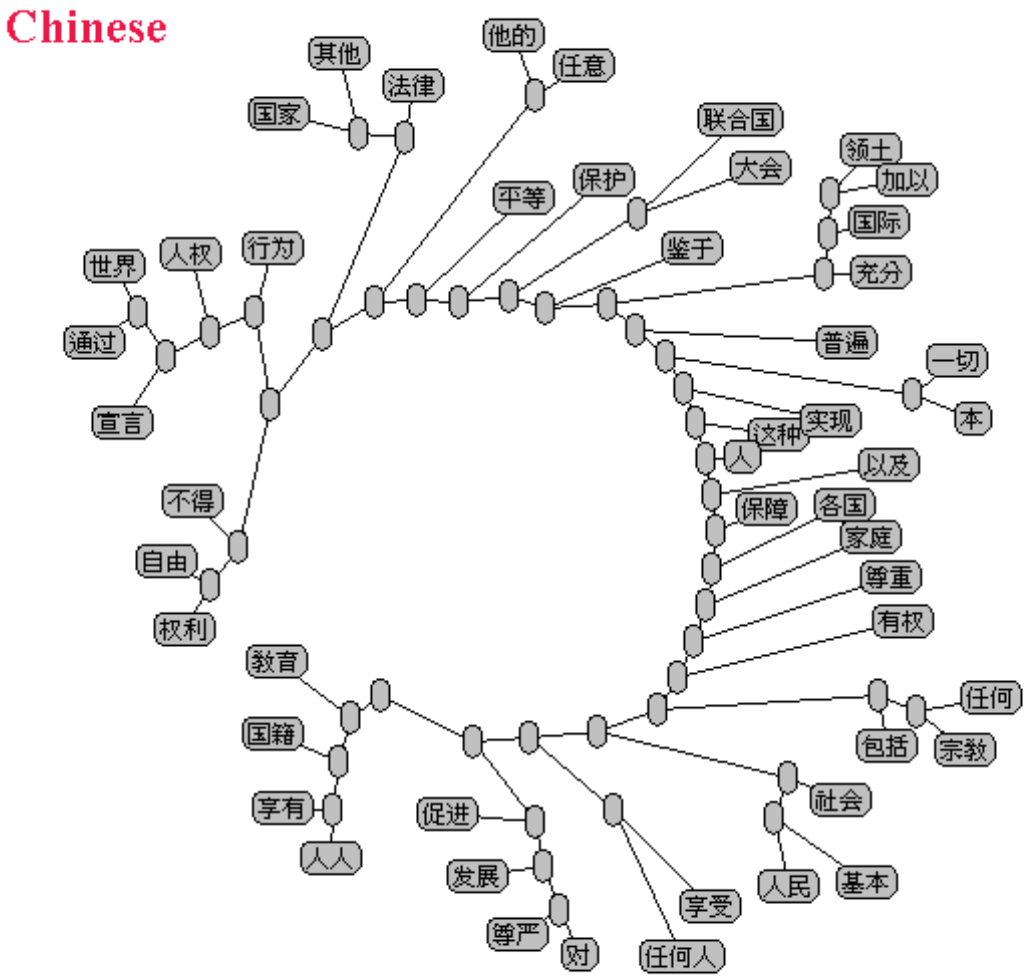

Figure 2. Radial view output for English.

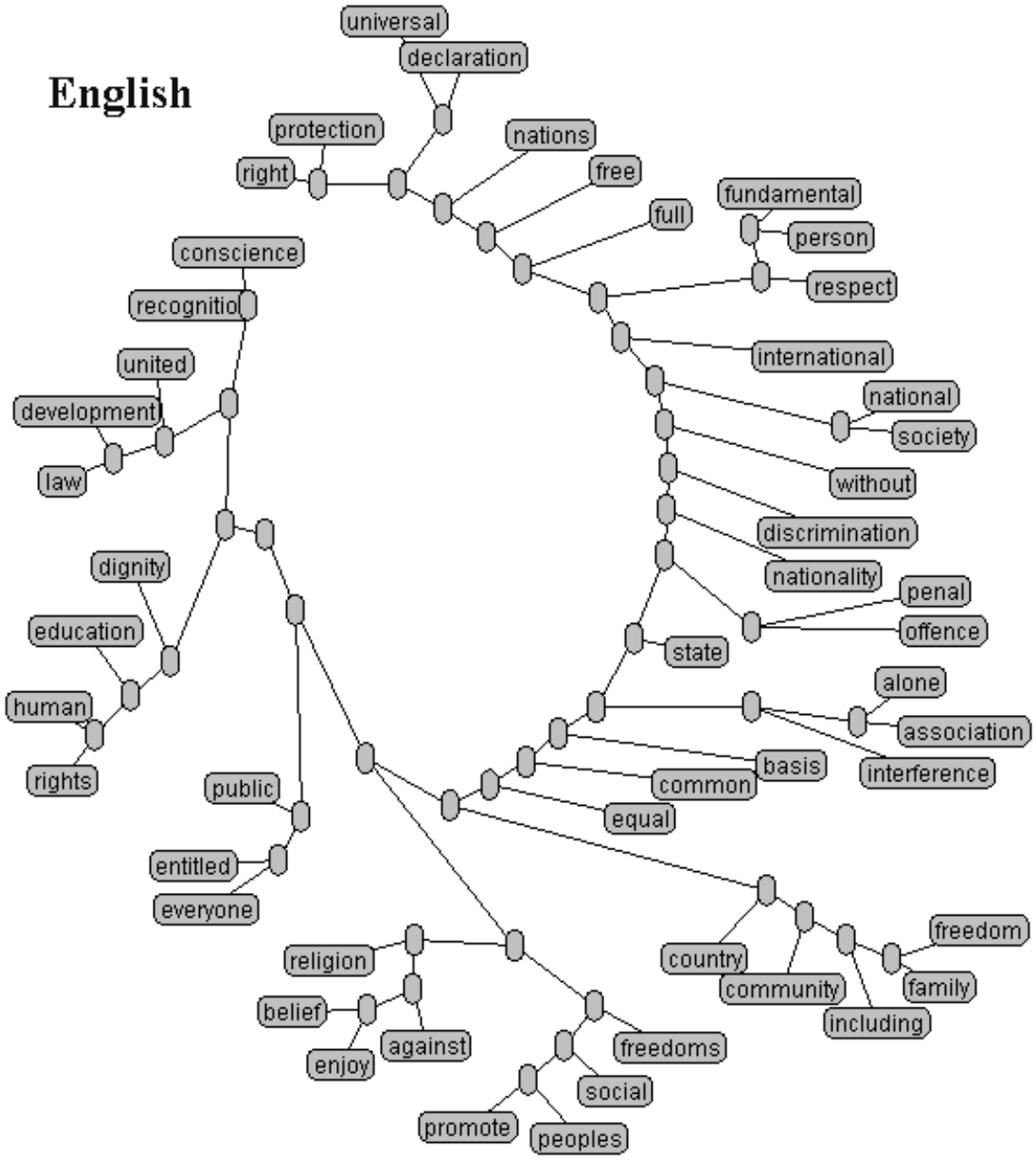

Figure 3. Radial view output for Hindi.

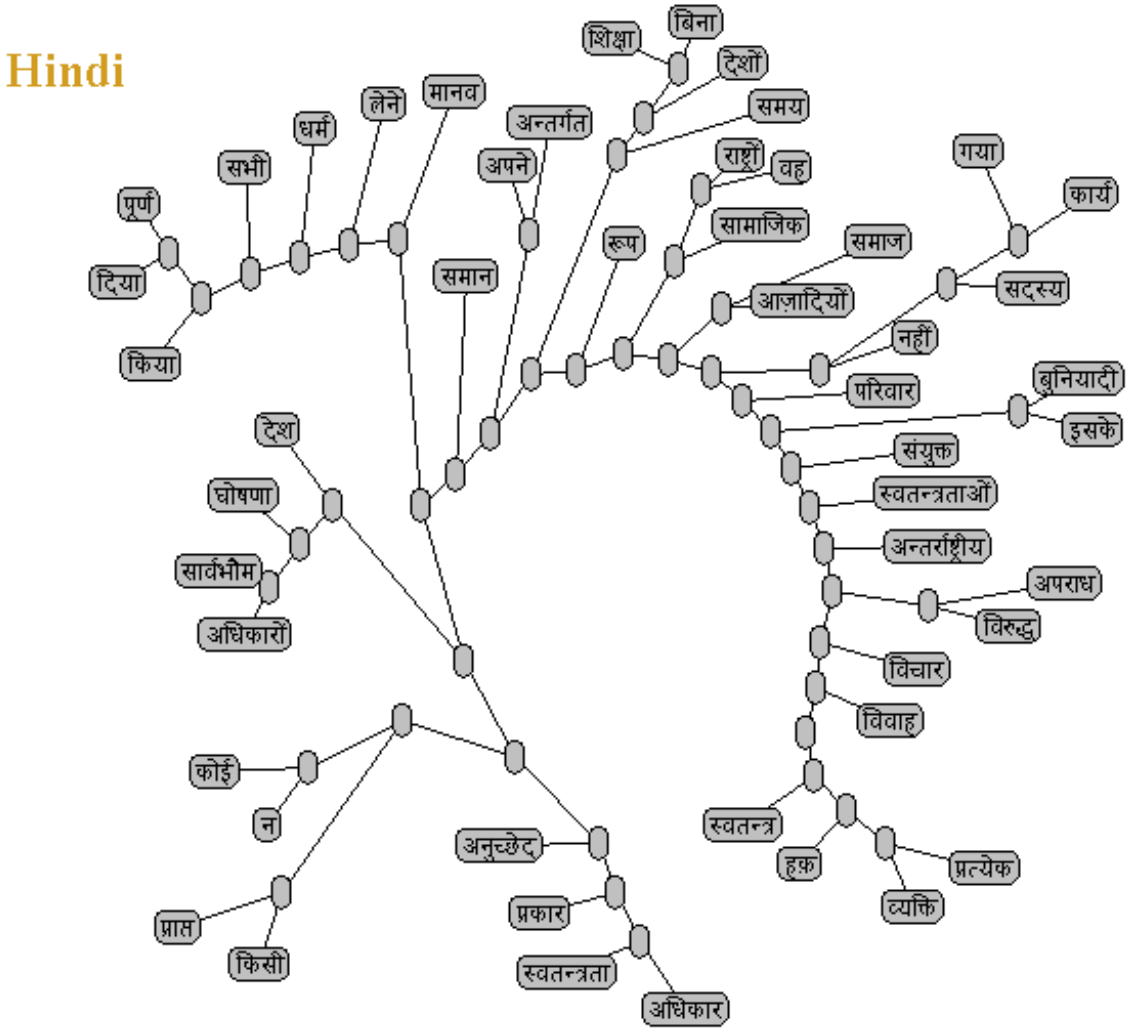

Figure 4. Radial view output for Korean.

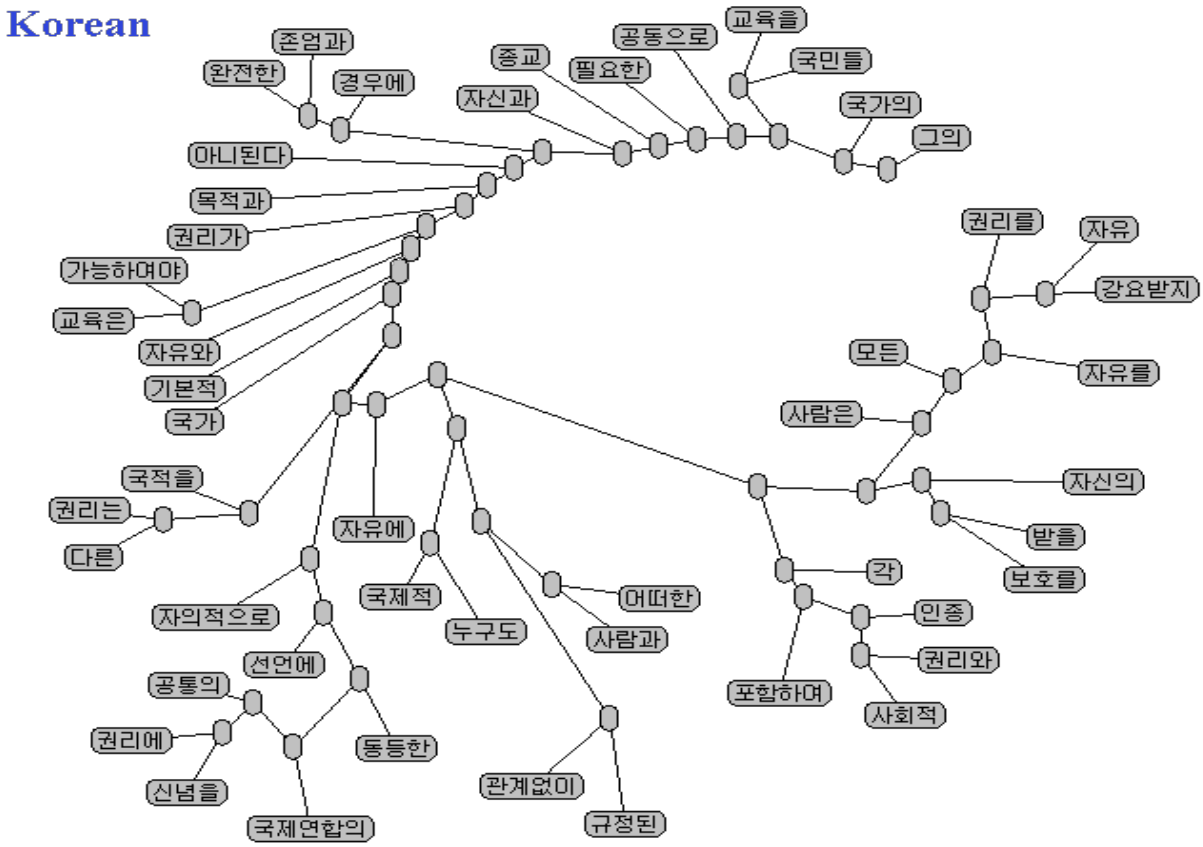

Figure 5. Tree map output for Chinese

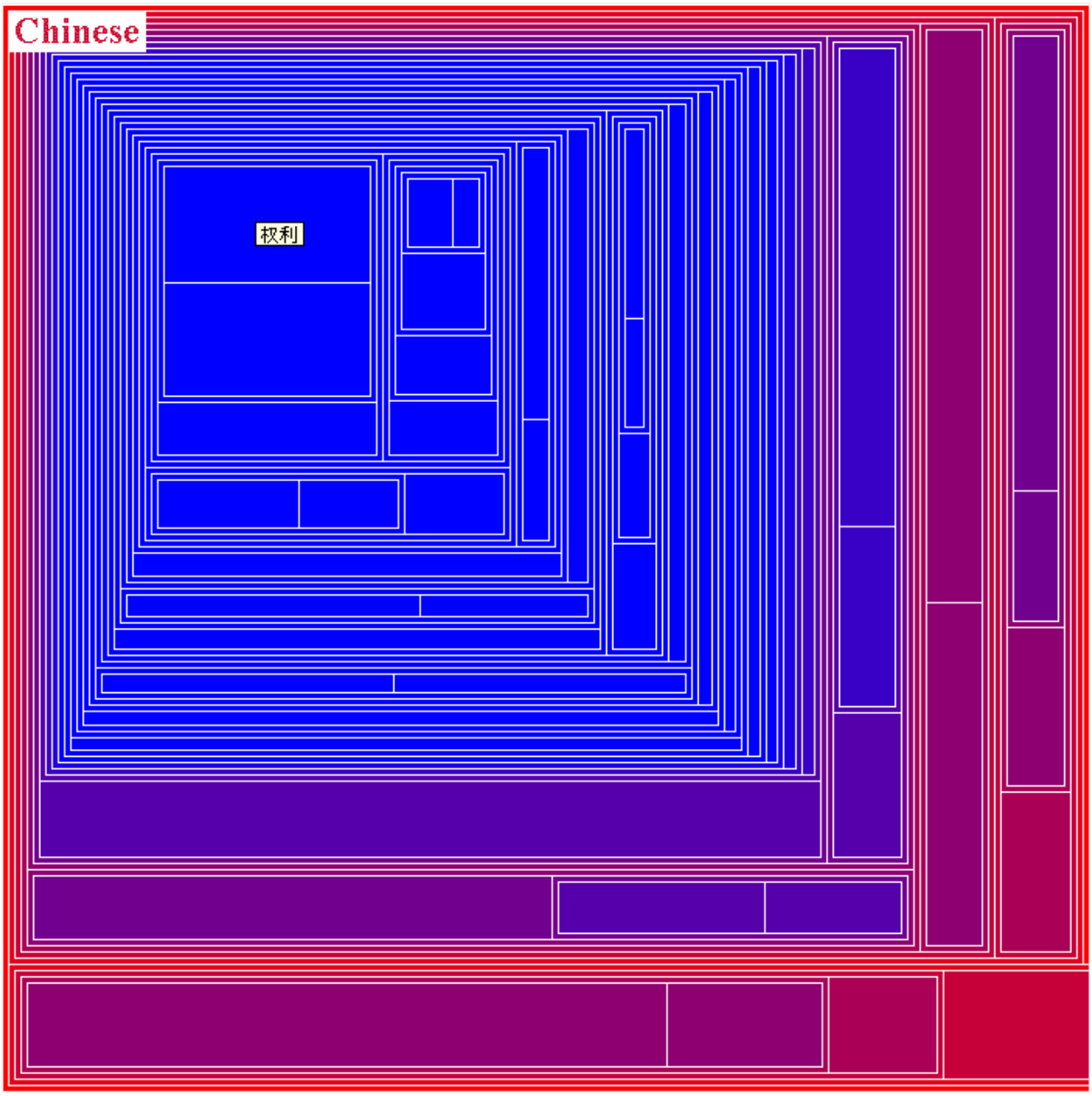

Figure 6. Tree map output for English.

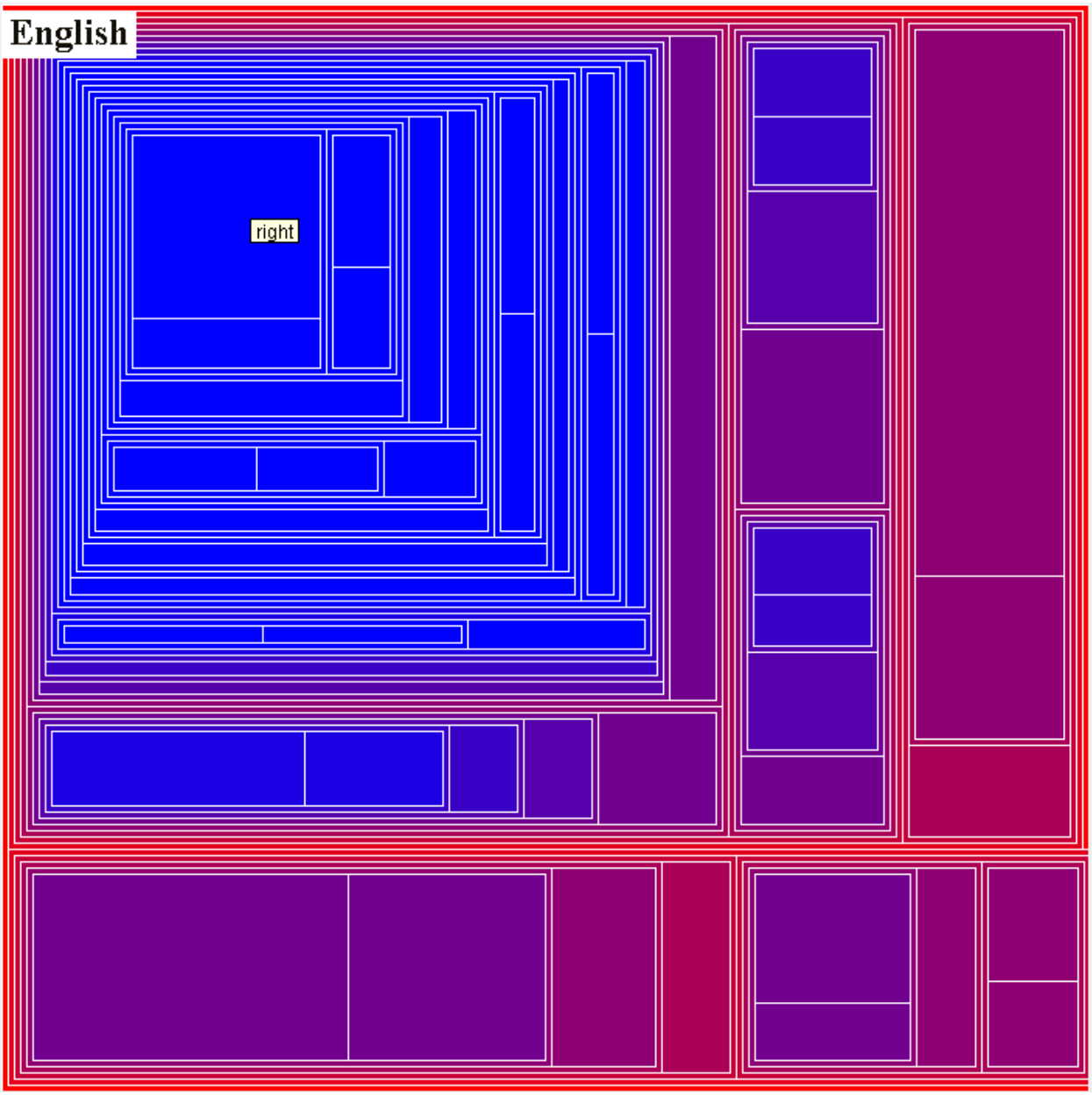

Figure 7. Tree map output for Hindi.

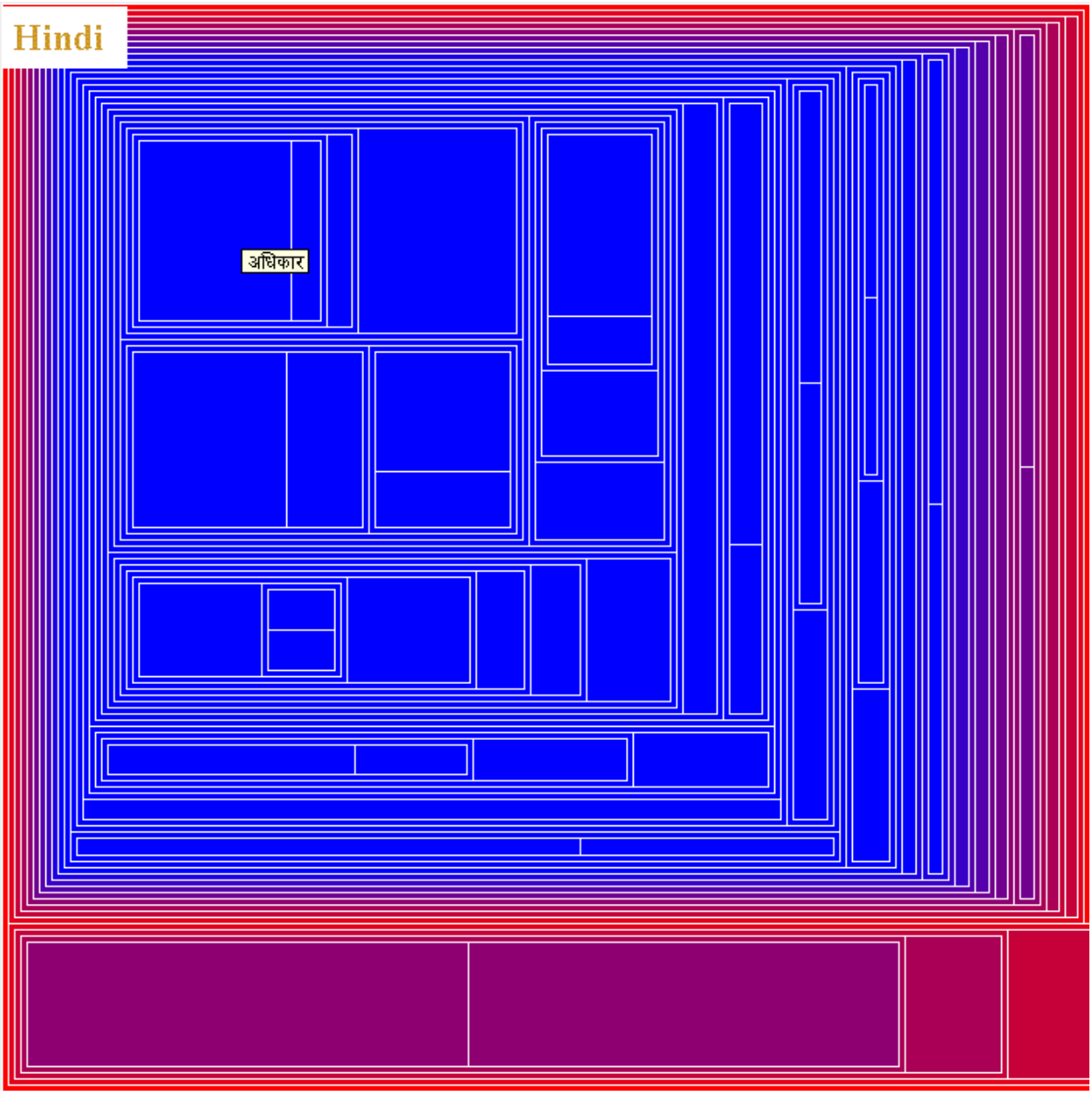

Figure 8 Tree map output for Korean.

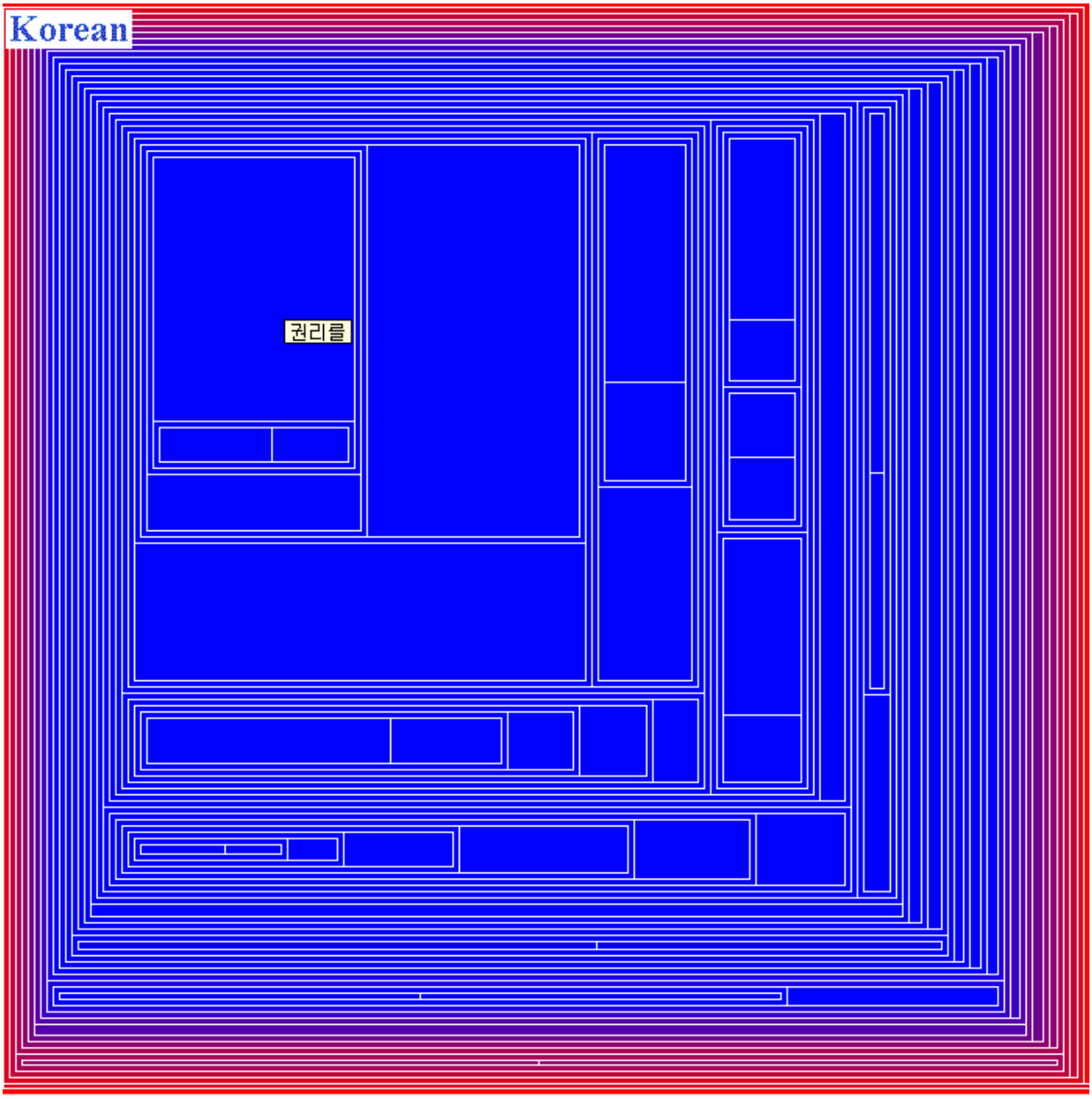

Figure 9. Dendogram output for Chinese

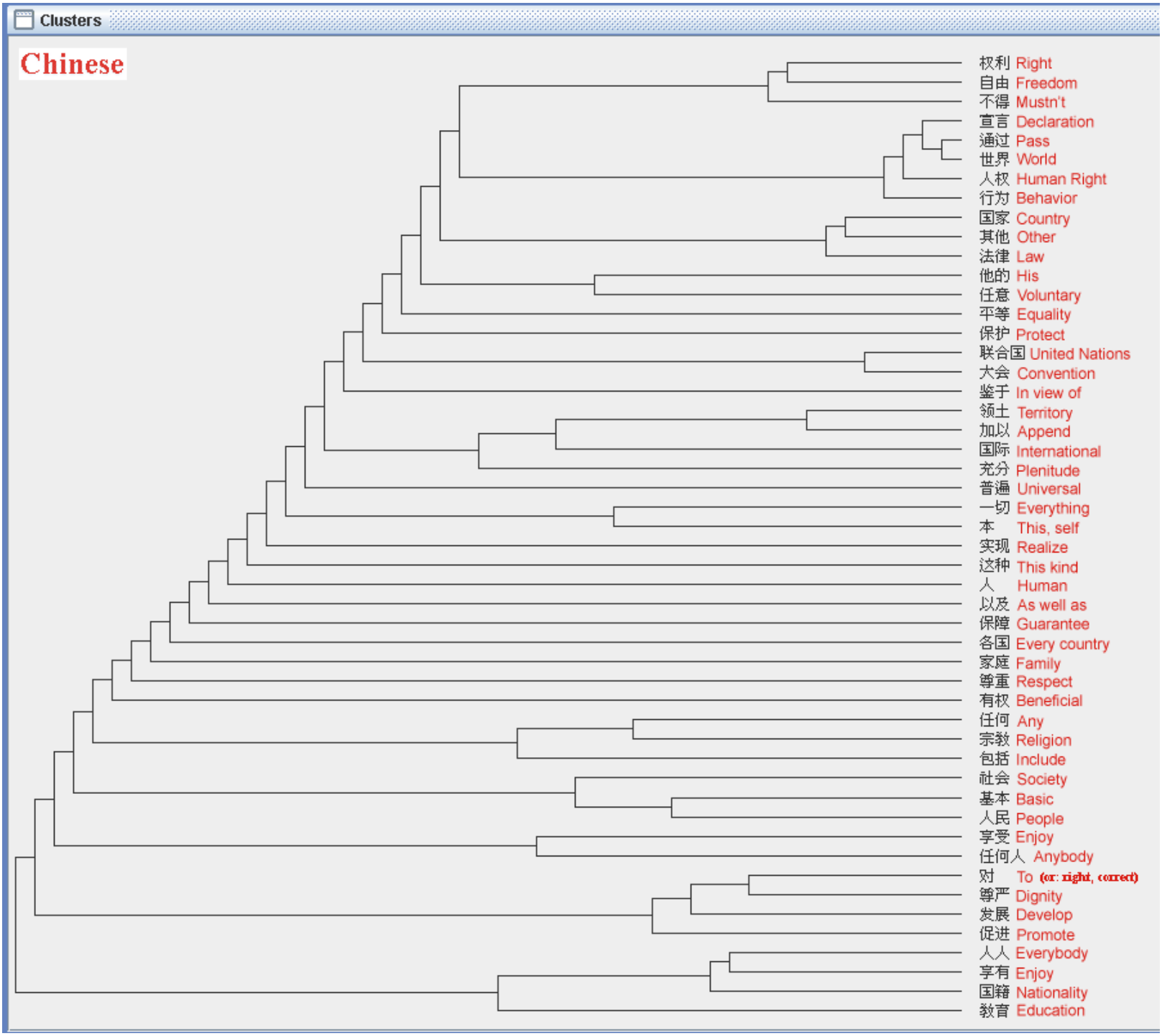

Figure 10. Dendogram output for English.

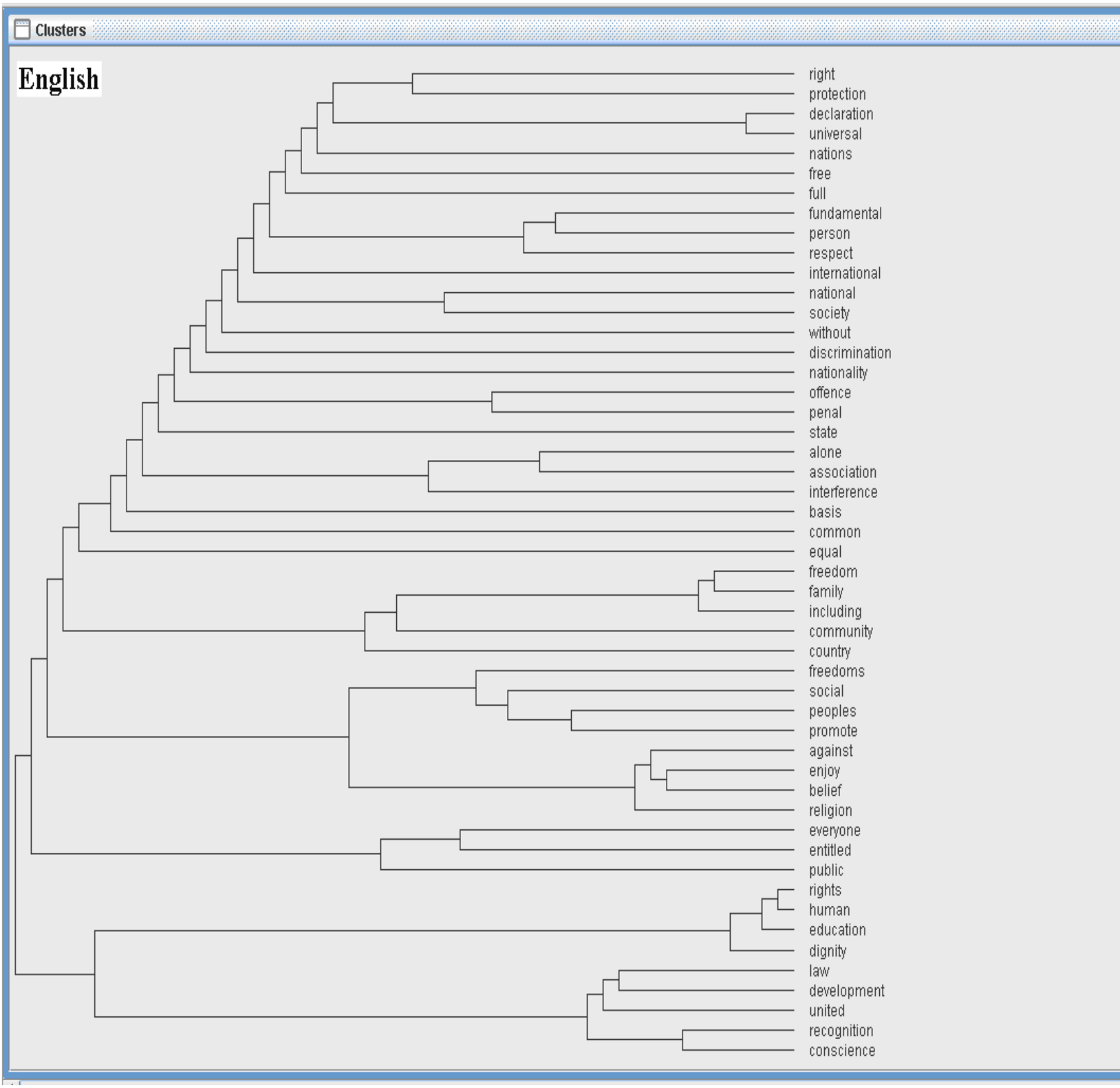

Figure 11. Dendogram output for Hindi.

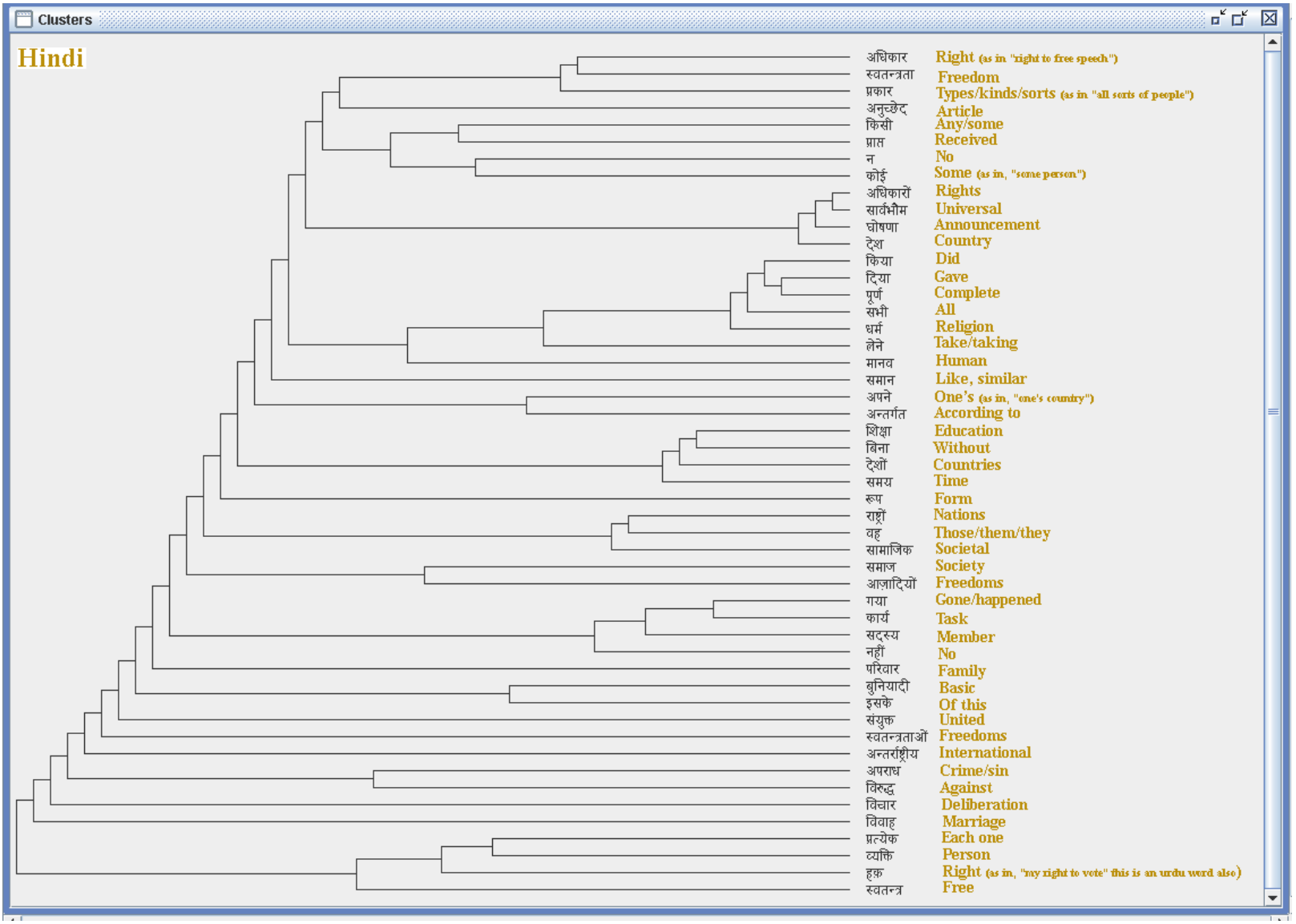

Figure 12. Dendogram output for Korean.

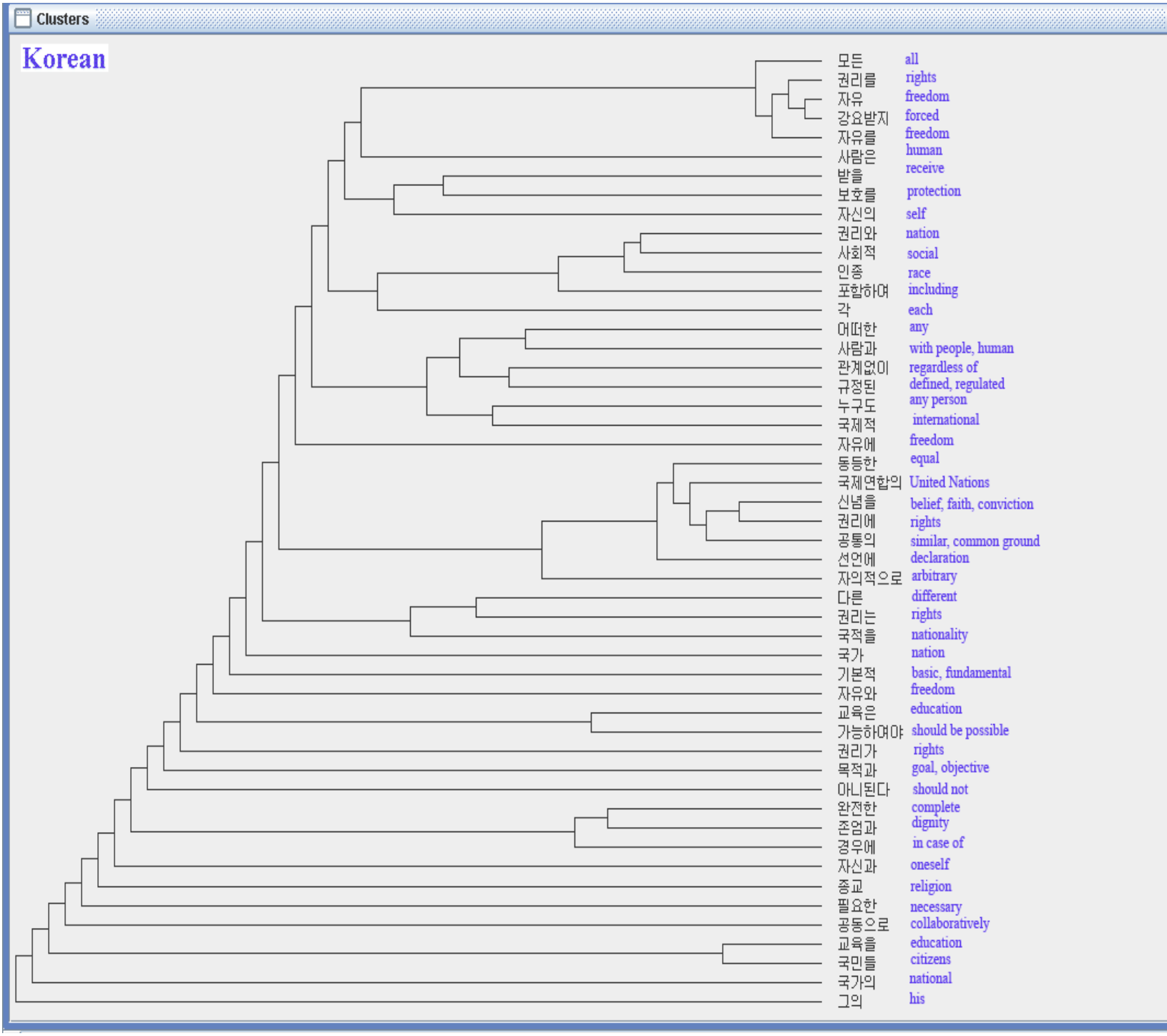

Table 1.

Clusters of 3 or more words for all five languages.

Chinese	English	Hindi	Korean
Right Freedom Mustn't	Freedoms Social Peoples promote	Right Freedom Types Article ("article" was excluded in other languages)	All Rights Freedom Forced freedom
Declaration Pass World Human right Behavior		Rights Universal Announcement Country	Equal United Nations Belief, Faith Rights Similar, common ground Declaration arbitrary
Country Other Law			
Territory Append International Plentitude			
Any religion include	Against Enjoy Belief religion	Did Gave Complete All Religion take human	
Society Basic People		Nations Those/them/they societal	Nation Social Race Including each
Right (correct) Dignity Develop Promote			Complete Dignity In case of

Everybody Enjoy Nationality Education	Rights Human Education dignity	Education Without Countries Time	Different Rights nationality
	Fundamental Person respect		Receive Protection self
	Alone Association Interference	gone/happened Task Member No	
	Freedom Family Including Community Country		
	Everyone Entitled Public	Each one Person Right Free	Any With people/human Regardless of Defined, regulated Any person international
	Law Development United Recognition conscience		
8 total clusters	8 total clusters	7 total clusters	7 total clusters

Note. All clusters were counted from radial output.