Phonological networks have been evidenced to share similar structural network properties at the macro level (Arbesman, Strogatz & Vitevitch, 2010). Example of macro level properties include a small-world structure, power-law degree distribution, as well as similarities in connectivity and clustering. However, the shared similarities between different language have only been observed at the macro level, i.e., they only deal with the global characteristics and the overall structure and connectivity of the given network. To better establish the notion of similarity between languages and their network structure, there is a need to dive deeper and investigate network structure at the meso and micro levels. At the micro level, the analysis is interested in the individual nodes and the relationships between them (i.e., edges). At the meso level, the aim is to examine the phenomenon known as clustering and bunching of nodes based on certain shared similarities, and how these groups or clusters interact with and relate to one another.

## Community Structure

At the meso level, Siew (2013) found that a robust community structure does exist among words in the English language. Community structure refers to the presence of distinct groups or clusters of nodes within a network. These clusters typically share certain phonetic similarities. These includes more than just rhyming words as the phonological similarities could exist in any part of the word. The behaviour and properties of phonological segments at the meso level have been observed to be similar to that of the macro level, such as a power law distribution and small-world structure.

## Aim and Hypotheses

However, these findings related to community structure has not yet been established in other languages. Thus, we were motivated to determine whether community structure exists in other languages as well. If community structure is evidenced, and the underlying patterns and properties exhibited are similar to what has been found in English, we can then conclude that even at the meso level, languages do share a similar network structure. The languages chosen to be investigated were Dutch, French, German, and Spanish. These languages were chosen because they come from the same language family and thus share numerous similarities with English. For example, these languages are all romanised and use the Latin alphabet, which means that they would have similar vowel and consonant sounds, compared to other languages such as Mandarin Chinese. Another reason would be that the lexical properties of words in these languages are more readily available and interpretable than others. The data for these languages were all sourced from the CLEARPOND database (Marian et al., 2012), to maintain consistency and improve comparability between the data sets.

## Method

### Phonological Networks

A total of four separate phonological networks in four different languages (Dutch, French, German, Spanish) were constructed in this study. Each individual phonological networks was constructed in a similar manner to what was set out by Vitevitch (2008). Approximately 20,000 words were obtained from the CLEARPOND database (Marian, Bartolotti, Chabal & Shook, 2012) for each of the four languages. CLEARPOND is an open-source platform with freely accessible databases containing information on phonological and orthographical neighbourhood densities of words in various languages.

In each of the phonological network, each node corresponded to a word's phonological transcription in International Phonetic Alphabet (IPA). Then, if any two word were considered to be phonologically similar, the two nodes (representing the two words) would be linked together via an undirected (or unweighted) edge.

In psycholinguistics, words are traditionally considered phonologically similar if their phonological transcriptions have an edit distance of one, i.e., if adding, subtracting, or replacing a phoneme of a valid given word results in another valid word, the two words will be considered phonologically similar (Greenberg and Jenkins, 1964; Landauer and Streeter, 1973; Luce and Pisoni, 1998). The validity of such a method in measuring phonological similarity in psychology has also been justified, where it has been evidenced that when participants were tasked to produce similar sounding words to a given word, they would produce words that differed by just one phoneme from the given word (Luce & Large, 2001). For example, the phonological neighbours of the word “dog” /dɒɡ/ would include words such as “log” /lɒɡ/, “hog” /hɒɡ/, or “dig” /dɪɡ/. The nodes of these phonological neighbours would be linked together via undirected edges to the node representing the word /dɒɡ/.

Each network consisted of a giant component containing thousands of interconnected words, numerous lexical islands (smaller components containing interconnected words that are not linked to the giant component), and lexical hermits (words that are not connected to any other word, otherwise known as “isolates” in the network science literature). In each of the present analyses, the community structure of the giant component in each language was extracted and used. Islands and hermits were excluded from the analyses as each island and hermit theoretically represents its own “community”, and thus any community detection analyses performed would not be likely to yield any statistically meaningful results.

### Modularity and Community Detection

Modularity, *Q*, measures the density of links inside communities as compared to links between communities ([Newman, 2006](https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00553/full#B51); [Fortunato, 2010](https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00553/full#B25)), and is mathematically defined as

where Aij represents the adjacency matrix of the weights of the edge between nodes *i* and *j*, ki is equal to the sum of the weights of the edges attached to node *i*, ci is the community to which node *i* is assigned, cj is the community to which node *j* is assigned, the δ function δ(ci, cj) is 1 if ci = cj and 0 otherwise, and . Since the present network has unweighted edges, Aij is simply reduced to a matrix with constants, and ki and kj is equal to the number of edges attached to node *i* and node *j* respectively (Siew, 2013).

Modularity is also often optimised and used to determine the quality of communities that have been partitioned in a network (Fortunato, 2010). The closer the *Q* values are to 1, the higher the quality of communities that have been partitioned. is where the density of links within communities is high relative to the density of links between communities. *Q* values typically range from 0.42 to 0.78, regardless of network size (Blondel et al., 2008). Although there is variance among the different types of complex networks, community structure is relatively robust within these networks and the partitions defining its communities are very distinct, as evidenced from the prevalent positive *Q* values (Siew, 2013).

The Louvain method is a community detection method which adopts modularity (“greedy”) optimisation approach that comprises two iteratively repeating phases. In the first phase, each node is split into its own community, such that the number of communities equal the number of nodes. Modularity is then optimised by determining which of its neighbours, *j*, would produce the highest increase in modularity, *Q*, when a node, *i*, is linked to each of its neighbours. Thereafter, the node will be classified into the same community as said neighbour, and the process repeats for all nodes within the network. In the second phase, a new network is formed where the nodes represent each community found in the first phase. This two-phase process is repeated until the maximum possible value of *Q* has been achieved. While the order that each node is being processed in this method will affect the final output, the differences in permutations of communities that will possibly be attained do not have any significant impact on the optimisation process and quality of communities partitioned (Blondel et al., 2008).

While other methods of community methods (e.g., centrality, betweenness) and dynamic methods (e.g., clique perlocation; Derényi et al., 2005) exist, they often involve more complex and convoluted algorithms. Whereas, the Louvain method involves a significantly quicker, simpler, and much more intuitive algorithm, just as how complex networks are believed to operate naturistically (Blondel et al., 2008). Furthermore, it even integrates the notion of hierarchy, as communities are being built one node at a time (Blondel et al., 2008). Thus, although community structure in networks can be determined by numerous other techniques [see Porter et al. (2009) or Fortunato (2010) for a review of these algorithms], the Louvain method was selected for its high quality in detecting communities within relatively short computational times (Blondel et al., 2008). Researchers have shown that the majority of community detection techniques produce findings that are comparable with each other despite varying in even the very smallest features, e.g., whether or not communities can overlap (Porter et al., 2009), which makes it rather unlikely that the communities partitioned by the Louvain method to merely be a by-product of the use of a specific community detection technique.

### Random Communities

Random communities were generated to provide a baseline scenario for comparing between the lexical properties of each community partitioned from the giant component of each of the four original networks. The random communities were constructed by randomly reallocating words from the giant component of each network to the *same number* of communities with the *same sizes* as what had been obtained via the Louvain method. This is a popular randomisation techniques amongst research studies that have examined the community structure in various complex networks to create communities for comparison (e.g., Traud et al., 2008). By doing this, it is possible to compare the real communities that were created using the community detection method against the random communities that were created by randomly grouping words into communities of the same sizes in a fruitful and impartial manner.

## Results

### Communities in the Networks

Communities were detected, partitioned and constructed via the “*igraph*” package on R. The Louvain method was applied onto the giant component of the phonological networks of each of the four languages, and it found a total of 44 communities in Dutch, 90 communities in French, 54 communities in German, and 81 communities in Spanish (refer to Table 1). For easier reference, the communities have been reordered in all tables to be in ascending order, and the community numbers have been relabelled accordingly. The communities of the words from the original giant component of each language’s phonological network will be referred to as “real communities” from here on. Modularity, *Q*, was also measured to be relatively large values that fall neatly within the range of modularity values as asserted by Blondel et al. (2008) (as in Table 2).

In order to assess if the communities detected were genuine features of the phonological network, 10 Erdõs-Renyi (ER) graphs were generated per language to compare against the real networks. The average modularity scores of randomly generated ER networks were much smaller than that of the real phonological networks across all four languages (as in Table 2). This suggests that the communities extracted from these random ER graphs are of lower quality than the communities extracted from the real phonological networks. A larger modularity score (as *Q* approaches 1) implies that the community structure exists within the network in a highly robust manner (Blondel et al., 2008). The high *Q* of the real phonological networks when compared to the random ER networks posits that phonological networks are present in tightly knitted communities (Siew, 2013).

### Lexical Characteristics in the Communities

A series of random communities were then created by randomising the community membership of words in the real communities of the giant component in each language’s phonological network. These random communities would be termed as such and serve as baseline communities to compare against.

As each community would have certain traits that set them apart from other communities, a multiple linear regression was performed to investigate what certain lexical characteristics could tell us about each community. The raw beta coefficients can be found in Table 3. This was done by comparing the standardised beta scores for the multiple linear regression of the real communities for each of the four languages. The same analyses were run for the random communities of each language, but this time, we expected there to be no significant results as the effects of the lexical characteristics should be totally random. This would allow us to conclude that our findings for the real communities did not occur by chance. The standardised beta scores can be found in Table 4. As a reminder, the community numbers of the real communities have been reordered and relabelled to be in ascending order.

#### Phoneme length

Phoneme length was defined by the number of phonemes present in each word. The multiple linear regression analysis revealed that phoneme length was statistically significant for all real communities but not significant for all random communities across all languages, with the exception of random French community (p = 0.044). Here, we can conclude that some communities that contain more words with higher average number of phonemes while some other communities contain more words with lower average number of phonemes, and it did not occur by chance.

#### Word length

Word length was defined by the number of letters present in each word. The multiple linear regression analysis again revealed that word length was statistically significant for all real communities but not significant for all random communities across all languages. Thus, we can conclude that some communities more words that have a higher average number of letters per word while some other communities contains more words that have a lower average number of letters per word, and it did not occur by chance.

#### Word frequency

Word frequency was defined by how often each word occurs in a given language, and log-base 10 of the raw frequency values were used in the present analyses (Kučera & Francis, 1967). The multiple linear regression analysis revealed that phoneme length was statistically significant for all real communities but not significant for all random communities across all languages. Here, we can conclude that some communities that contain more words that occur more often while some other communities contain more words that occur less often, and that this did not occur by chance.

#### Neighbourhood density

Neighbourhood density was defined by the number of words that are phonologically similar to each word (Luce and Pisoni, 1998). Words are considered to be phonologically similar when they differ from each other by an edit distance of 1, i.e., the substitution, addition, or deletion of one phoneme from any given word would result in its phonological neighbour being produced. This method of connecting words with an edit distance of 1 is identical to how the metric *degree* is derived in network analyses by connecting two nodes via an edge. The multiple linear regression analysis revealed that phoneme length was statistically significant for all real communities but not significant for all random communities across all languages. Here, we can conclude that some communities that contain more words from high density neighbourhoods while some other communities contain more words from low density neighbourhoods, and this did not occur by chance.

#### Neighbourhood frequency

Neighbourhood frequency was defined as the word frequency of the phonological neighbours of a given word. Similarly, log-base 10 values of word frequency were derived, based on Kučera and Francis (1967). The multiple linear regression analysis revealed that neighbourhood frequency was statistically significant for all real communities but not significant for all random communities across all languages. Here, we can conclude that some communities contained more words that have phonological neighbours with high frequency while some other communities contained more words that have phonological neighbours with low frequency, and the results did not occur by chance.

#### Relative Importance of Lexical Characteristics

To uncover whether these lexical characteristics are truly important predictors in accounting for the variance between communities, a relative importance analysis was performed for each language separately. This was done by comparing between the standardised beta coefficients of the lexical characteristics between real and random communities for each language. The analyses revealed that the standardised beta coefficients for lexical characteristics of real communities were much larger than that of random communities. We wanted to further our understanding of the characteristics of each real community, so we sought out to investigate whether communities contained words which share identical phonological segments, i.e., biphones, as posited by Siew (2013).

#### Raw Biphone Counts

We wanted to further our understanding of the characteristics of each real community, so we sought out to investigate whether communities contained words which share identical phonological segments, i.e., biphones. Raw biphone counts, defined as the total number of times the biphone is found in each community, were measured for each of the real and random communities across all languages. Noe that raw biphone counts are not position-specific and they do not represent the overall frequencies in a given language.

Two-way Kolmogorov-Smirnov (K-S) tests were conducted to compare between the raw biphone counts found in the real and random communities of the same size in each of the languages (as in Table 5). The results from K-S tests were mixed across the four languages. Notably, K-S tests were more likely to be significant as community size increased beyond a certain size. This is evidenced in figures 13-16, where we see the same pattern in the visual plots of K-S test significance against community size across all languages. This is further supported by visual plots of K-S test significance against K-S D-statistic, where there is the K-S D-statisctic is found to be significant only when it passes a certain threshold that lies between 0.05 and 0.10 (Figures 9-12). This suggests that when communities are too small, there may not be enough data to tell us anything meaningful about those communities. If we only consider communities of this given size, the results from K-S tests would then now be mostly significant. These results indicate that raw biphone counts of communities obtained using community detection methods differed significantly from the raw biphone counts of randomly generated communities, when communities are sufficiently large. Thus subsequent analyses were done with this in mind.

Figures 1-8 show the distribution of raw biphone counts from an arbitrarily selected real and random community for each of the four languages. From these figures, the distribution of biphones in the random communities differ largely from the distribution of biphones in the real communities, which suggests that communities in a given phonological network share certain specific phonological segments.

] It is interesting to note that there are relatively few biphones that occur frequently within a community, and a large number of biphones that occur rarely. This was observed generally across all real communities in all four languages, but not in randomly generated communities of any of the four languages, which is similar to what has been established in the literature regarding word frequency (Zipf, 1935). Even though the most frequent biphones in each real community are different from one another, the general trends in biphone distributions at the community level are in line with the overall distribution and trend in general.

Interestingly, the most frequently occurring biphones in each real community are very closely related to one another and can be easily combined to form longer phonological semgents. For example, in the Dutch real community 43, the five most common phonemes are “Nk”, “aN”, “Ik”, “lI”, and “IN”. Here, you can observe that the phonemes present in each biphone are very similar, and they can easily be combined together by connecting two biphones via a shared phoneme, e.g., “aN” + “Nk” = “aNk”. Across the four languages, the most frequently occurring biphones in real communities share such particular phonological features, such that words in a community may mostly be phonological variants of one another to some extent.

## Discussion

## References

Auer, E. T., and Luce, P. A. (2005). “Probabilistic phonotactics in spoken word recognition,” in *The Handbook of Speech Perception*, eds D. B. Pisoni and R. E. Remez (Oxford: Blackwell), 610–630. doi: 10.1002/9780470757024.ch25

Cortese, M. J., and Khanna, M. M. (2008). Age of acquisition ratings for 3, 000 monosyllabic words. *Behav. Res. Methods* 40, 791–794. doi: 10.3758/BRM.40.3.791

Derényi, I., Palla, G., and Vicsek, T. (2005). Clique percolation in random networks. *Phys. Rev. Lett*. 94:160202. doi: 10.1103/PhysRevLett.94.160202

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep*. 486, 75–174. doi: 10.1016/j.physrep.2009.11.002

Frauenfelder, U. H., Baayen, R. H., and Hellwig, F. M. (1993). Neighborhood density and frequency across languages and modalities. *J. Mem. Lang*. 32, 781–804. doi: 10.1006/jmla.1993.1039

Ghyselinck, M., Lewis, M. B., and Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: a review of the literature and a new multi-task investigation. *Acta Psychol*. 115, 43–67. doi: 10.1016/j.actpsy.2003.11.002

Greenberg, J. H., and Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American english. *Word* 20, 157–177.

Kučera, H., and Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. (2012). Age-of-acquisition ratings for 30, 000 english words. *Behav. Res. Methods* 44, 978–990. doi: 10.3758/s13428-012-0210-4

Landauer, T. K., and Streeter, L. A. (1973). Structural differences between common and rare words: failure of equivalence assumptions for theories of word recognition. *J. Verbal Learn. Verbal Behav*. 12, 119–131. doi: 10.1016/S0022-5371(73)80001-5

Luce, P. A., and Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Lang. Cogn. Process*. 16, 565–581. doi: 10.1080/01690960143000137

Luce, P. A., and Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear Hear*. 19, 1–36. doi: 10.1097/00003446-199802000-00001

Marian, V., Bartolotti, J., Chabal, S., Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE 7(8): e43230.*doi:10.1371/journal.pone.0043230

Maxwell, S. E., and Delaney, H. D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. New York, NY: Psychology Press Taylor and Francis Group.

Nusbaum, H. C., Pisoni, D. B., and Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: measuring the familiarity of 20, 000 words. *Res. Speech Percept. Prog. Rep*. 10, 357–376.

Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Not. AMS* 56, 1082–1097.

Ravasz, E., and Barabási, A. L. (2003). Hierarchical organization in complex networks. *Phys. Rev. E* 67:026112. doi: 10.1103/PhysRevE.67.026112

Siew, C. S. (2013). Community structure in the phonological network. *Frontiers in psychology*, *4*, 553.

Steyvers, M., and Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cogn. Sci*. 29, 41–78. doi: 10.1207/s15516709cog2901\_3

Vitevitch, M. S., and Luce, P. A. (1998). When words compete: levels of processing in perception of spoken words. *Psychol. Sci*. 9, 325–329. doi: 10.1111/1467-9280.00064

Vitevitch, M. S., and Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behav. Res. Methods* 36, 481–487. doi: 10.3758/BF03195594

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval. *J. Speech Lang. Hear. Res*. 51, 408. doi: 10.1044/1092-4388(2008/030)

Zipf, G. K. (1935). *The Psycho-Biology of Language*. Boston, MA: Houghton-Mifflin.

| Table 1: *Community Sizes in Different Languages* | | | | |
| --- | --- | --- | --- | --- |
| **Community** | *Dutch N* | *French N* | *German N* | *Spanish N* |
| **1** | 4 | 3 | 4 | 6 |
| **2** | 5 | 4 | 4 | 7 |
| **3** | 5 | 6 | 4 | 7 |
| **4** | 5 | 6 | 4 | 7 |
| **5** | 5 | 6 | 4 | 7 |
| **6** | 6 | 6 | 5 | 8 |
| **7** | 7 | 6 | 6 | 8 |
| **8** | 8 | 7 | 7 | 8 |
| **9** | 11 | 7 | 7 | 8 |
| **10** | 13 | 9 | 8 | 9 |
| **11** | 14 | 10 | 8 | 10 |
| **12** | 14 | 10 | 8 | 11 |
| **13** | 14 | 11 | 9 | 11 |
| **14** | 16 | 12 | 9 | 11 |
| **15** | 22 | 12 | 10 | 13 |
| **16** | 23 | 13 | 12 | 14 |
| **17** | 26 | 13 | 13 | 15 |
| **18** | 26 | 13 | 15 | 17 |
| **19** | 29 | 13 | 17 | 17 |
| **20** | 31 | 14 | 17 | 19 |
| **21** | 31 | 15 | 18 | 22 |
| **22** | 34 | 15 | 21 | 23 |
| **23** | 41 | 16 | 27 | 24 |
| **24** | 45 | 16 | 54 | 26 |
| **25** | 72 | 16 | 55 | 27 |
| **26** | 74 | 16 | 56 | 27 |
| **27** | 85 | 17 | 72 | 28 |
| **28** | 133 | 17 | 74 | 30 |
| **29** | 219 | 17 | 78 | 30 |
| **30** | 226 | 17 | 87 | 37 |
| **31** | 245 | 17 | 94 | 38 |
| **32** | 260 | 18 | 126 | 38 |
| **33** | 319 | 18 | 133 | 38 |
| **34** | 367 | 19 | 139 | 40 |
| **35** | 372 | 21 | 142 | 42 |
| **36** | 449 | 22 | 186 | 43 |
| **37** | 463 | 22 | 211 | 43 |
| **38** | 487 | 23 | 214 | 47 |
| **39** | 488 | 24 | 227 | 49 |
| **40** | 535 | 26 | 246 | 51 |
| **41** | 579 | 26 | 325 | 59 |
| **42** | 739 | 27 | 424 | 60 |
| **43** | 748 | 29 | 454 | 60 |
| **44** | 1232 | 30 | 460 | 64 |
| **45** |  | 30 | 474 | 65 |
| **46** |  | 32 | 481 | 66 |
| **47** |  | 32 | 502 | 77 |
| **48** |  | 33 | 503 | 81 |
| **49** |  | 33 | 511 | 93 |
| **50** |  | 36 | 548 | 97 |
| **51** |  | 46 | 556 | 103 |
| **52** |  | 47 | 640 | 104 |
| **53** |  | 48 | 813 | 105 |
| **54** |  | 49 | 864 | 106 |
| **55** |  | 50 |  | 106 |
| **56** |  | 50 |  | 107 |
| **57** |  | 57 |  | 110 |
| **58** |  | 58 |  | 112 |
| **59** |  | 58 |  | 113 |
| **60** |  | 62 |  | 113 |
| **61** |  | 63 |  | 136 |
| **62** |  | 69 |  | 146 |
| **63** |  | 71 |  | 146 |
| **64** |  | 81 |  | 165 |
| **65** |  | 84 |  | 172 |
| **66** |  | 84 |  | 183 |
| **67** |  | 91 |  | 188 |
| **68** |  | 98 |  | 215 |
| **69** |  | 117 |  | 217 |
| **70** |  | 180 |  | 220 |
| **71** |  | 188 |  | 285 |
| **72** |  | 199 |  | 299 |
| **73** |  | 216 |  | 300 |
| **74** |  | 284 |  | 310 |
| **75** |  | 300 |  | 342 |
| **76** |  | 344 |  | 342 |
| **77** |  | 355 |  | 418 |
| **78** |  | 454 |  | 462 |
| **79** |  | 517 |  | 508 |
| **80** |  | 579 |  | 603 |
| **81** |  | 672 |  | 622 |
| **82** |  | 717 |  |  |
| **83** |  | 741 |  |  |
| **84** |  | 761 |  |  |
| **85** |  | 810 |  |  |
| **86** |  | 1004 |  |  |
| **87** |  | 1031 |  |  |
| **88** |  | 1216 |  |  |
| **89** |  | 1256 |  |  |
| **90** |  | 1727 |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 2: *Modularity Scores* | | | | | | |
|  |  | % of Total | Real Network | | Random (ER) Network | |
| **Language** | Size | Words | Q | Degree | Q | Degree |
| **Dutch** | 8527 | 0.30727 | 0.74336 | 7.23232 | 0.33196 | 7.23232 |
| **French** | 15695 | 0.56557 | 0.71524 | 17.3515 | 0.19803 | 17.3515 |
| **German** | 9986 | 0.35984 | 0.73350 | 7.27899 | 0.33001 | 7.27899 |
| **Spanish** | 8990 | 0.32395 | 0.81339 | 5.41151 | 0.41412 | 5.41151 |

Table 3: Raw Beta Coefficients

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Dutch | D-Rdm | French | F-Rdm | German | G-Rdm | Spanish | S-Rdm |
| Phone | -0.339^ | 0.075 | -3.756\* | 0.240# | -0.647\* | -0.036 | -6.955\* | 0.148 |
| Letters | -0.692\* | -0.127 | -0.127# | -0.048 | -0.342\* | -0.011 | 3.018\* | -0.298 |
| Freq | -0.056 | 0.074 | -0.286^ | 0.020 | -0.340^ | 0.112 | -1.509\* | 0.211 |
| Ph\_D | 0.097\* | 0.014 | 0.052\* | 0.009 | 0.221\* | -0.022 | 0.534\* | 0.001 |
| Ph\_F | -0.001\* | >-0.001 | -0.005\* | <0.001 | >-0.001 | <0.001 | -0.002\* | >-0.001 |

Significance levels: \* - (p < .001) , ^ - (p < .01) , # - (p < .05) .

Table 4: Standardised Beta Coefficients

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Dutch | D-Rdm | French | F-Rdm | German | G-Rdm | Spanish | S-Rdm |
| Phone | -0.071 | 0.015 | -0.347 | 0.019 | -0.109 | -0.005 | -0.616 | 0.011 |
| Letters | -0.158 | -0.027 | -0.016 | -0.005 | -0.069 | -0.002 | 0.264 | -0.022 |
| Freq | -0.007 | 0.008 | -0.015 | 0.001 | -0.031 | 0.009 | -0.067 | 0.008 |
| Ph\_D | 0.111 | 0.015 | 0.086 | 0.013 | 0.233 | -0.020 | 0.158 | <0.001 |
| Ph\_F | -0.087 | -0.004 | -0.066 | 0.003 | -0.018 | 0.006 | -0.061 | -0.012 |

| Table 5: *Summary of Kolmogorov-Smirnov Tests* | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Dutch* | | | | *French* | | | | *German* | | | | *Spanish* | | |
| ***Community*** | D | | p | | D | | p | | D | | p | | D | p | |
| ***1*** | 0.01 | | 1.00 | | 0.01 | | 1.00 | | 0.00 | | 1.00 | | 0.03 | 0.99 | |
| ***2*** | 0.01 | | 1.00 | | 0.01 | | 1.00 | | 0.01 | | 1.00 | | 0.05 | 0.68 | |
| ***3*** | 0.02 | | 1.00 | | 0.01 | | 1.00 | | 0.01 | | 1.00 | | 0.02 | 1.00 | |
| ***4*** | 0.01 | | 1.00 | | 0.01 | | 1.00 | | 0.01 | | 1.00 | | 0.03 | 0.97 | |
| ***5*** | 0.02 | | 1.00 | | 0.01 | | 1.00 | | 0.01 | | 1.00 | | 0.04 | 0.92 | |
| ***6*** | 0.02 | | 1.00 | | 0.01 | | 1.00 | | 0.01 | | 1.00 | | 0.04 | 0.74 | |
| ***7*** | 0.02 | | 0.99 | | 0.01 | | 1.00 | | 0.01 | | 1.00 | | 0.03 | 0.95 | |
| ***8*** | 0.02 | | 1.00 | | 0.01 | | 1.00 | | 0.02 | | 1.00 | | 0.03 | 0.99 | |
| ***9*** | 0.03 | | 0.94 | | 0.04 | | 0.66 | | 0.02 | | 1.00 | | 0.04 | 0.79 | |
| ***10*** | 0.04 | | 0.66 | | 0.01 | | 1.00 | | 0.02 | | 1.00 | | 0.06 | 0.42 | |
| ***11*** | 0.04 | | 0.70 | | 0.02 | | 1.00 | | 0.02 | | 0.94 | | 0.04 | 0.88 | |
| ***12*** | 0.03 | | 0.86 | | 0.01 | | 1.00 | | 0.02 | | 0.98 | | 0.04 | 0.74 | |
| ***13*** | 0.03 | | 0.96 | | 0.02 | | 1.00 | | 0.02 | | 0.98 | | 0.06 | 0.37 | |
| ***14*** | 0.04 | | 0.49 | | 0.03 | | 0.94 | | 0.02 | | 0.98 | | 0.05 | 0.52 | |
| ***15*** | 0.06 | | 0.17 | | 0.01 | | 1.00 | | 0.02 | | 1.00 | | 0.06 | 0.29 | |
| ***16*** | 0.06 | | 0.08 | | 0.02 | | 1.00 | | 0.04 | | 0.60 | | 0.07 | 0.17 | |
| ***17*** | 0.06 | | 0.14 | | 0.01 | | 1.00 | | 0.02 | | 0.94 | | 0.07 | 0.20 | |
| ***18*** | 0.07 | | 0.04 | | 0.01 | | 1.00 | | 0.03 | | 0.86 | | 0.06 | 0.33 | |
| ***19*** | 0.07 | | 0.05 | | 0.03 | | 0.99 | | 0.04 | | 0.45 | | 0.08 | 0.11 | |
| ***20*** | 0.07 | | 0.03 | | 0.02 | | 1.00 | | 0.03 | | 0.79 | | 0.08 | 0.13 | |
| ***21*** | 0.07 | | 0.04 | | 0.03 | | 0.96 | | 0.05 | | 0.22 | | 0.09 | 0.07 | |
| ***22*** | 0.08 | | 0.02 | | 0.02 | | 0.99 | | 0.04 | | 0.60 | | 0.10 | 0.02 | |
| ***23*** | 0.09 | | 0.01 | | 0.03 | | 0.98 | | 0.04 | | 0.38 | | 0.07 | 0.26 | |
| ***24*** | 0.08 | | 0.02 | | 0.04 | | 0.71 | | 0.10 | | 0.00 | | 0.12 | 0.00 | |
| ***25*** | 0.14 | | 0.00 | | 0.02 | | 0.99 | | 0.10 | | 0.00 | | 0.12 | 0.00 | |
| ***26*** | 0.12 | | 0.00 | | 0.04 | | 0.75 | | 0.11 | | 0.00 | | 0.12 | 0.00 | |
| ***27*** | 0.13 | | 0.00 | | 0.03 | | 0.84 | | 0.13 | | 0.00 | | 0.09 | 0.03 | |
| ***28*** | 0.22 | | 0.00 | | 0.03 | | 0.84 | | 0.13 | | 0.00 | | 0.12 | 0.00 | |
| ***29*** | 0.22 | | 0.00 | | 0.02 | | 1.00 | | 0.12 | | 0.00 | | 0.13 | 0.00 | |
| ***30*** | 0.21 | | 0.00 | | 0.03 | | 0.98 | | 0.13 | | 0.00 | | 0.12 | 0.00 | |
| ***31*** | 0.19 | | 0.00 | | 0.01 | | 1.00 | | 0.14 | | 0.00 | | 0.16 | 0.00 | |
| ***32*** | 0.18 | | 0.00 | | 0.04 | | 0.75 | | 0.15 | | 0.00 | | 0.15 | 0.00 | |
| ***33*** | 0.19 | | 0.00 | | 0.03 | | 0.91 | | 0.16 | | 0.00 | | 0.13 | 0.00 | |
| ***34*** | 0.24 | | 0.00 | | 0.03 | | 0.84 | | 0.17 | | 0.00 | | 0.14 | 0.00 | |
| ***35*** | 0.25 | | 0.00 | | 0.05 | | 0.39 | | 0.17 | | 0.00 | | 0.14 | 0.00 | |
| ***36*** | 0.24 | | 0.00 | | 0.04 | | 0.57 | | 0.24 | | 0.00 | | 0.11 | 0.01 | |
| ***37*** | 0.26 | | 0.00 | | 0.02 | | 0.99 | | 0.21 | | 0.00 | | 0.15 | 0.00 | |
| ***38*** | 0.23 | | 0.00 | | 0.06 | | 0.26 | | 0.24 | | 0.00 | | 0.19 | 0.00 | |
| ***39*** | 0.26 | | 0.00 | | 0.03 | | 0.84 | | 0.19 | | 0.00 | | 0.17 | 0.00 | |
| ***40*** | 0.24 | | 0.00 | | 0.04 | | 0.57 | | 0.21 | | 0.00 | | 0.15 | 0.00 | |
| ***41*** | 0.18 | | 0.00 | | 0.02 | | 0.99 | | 0.28 | | 0.00 | | 0.18 | 0.00 | |
| ***42*** | 0.19 | | 0.00 | | 0.06 | | 0.26 | | 0.24 | | 0.00 | | 0.16 | 0.00 | |
| ***43*** | 0.22 | | 0.00 | | 0.05 | | 0.36 | | 0.27 | | 0.00 | | 0.18 | 0.00 | |
| ***44*** | 0.16 | | 0.00 | | 0.06 | | 0.26 | | 0.24 | | 0.00 | | 0.16 | 0.00 | |
| ***45*** |  | |  | | 0.05 | | 0.39 | | 0.23 | | 0.00 | | 0.20 | 0.00 | |
| ***46*** |  | |  | | 0.06 | | 0.18 | | 0.22 | | 0.00 | | 0.22 | 0.00 | |
| ***47*** |  | |  | | 0.05 | | 0.43 | | 0.25 | | 0.00 | | 0.23 | 0.00 | |
| ***48*** |  | |  | | 0.04 | | 0.57 | | 0.30 | | 0.00 | | 0.23 | 0.00 | |
| ***49*** |  | |  | | 0.05 | | 0.36 | | 0.29 | | 0.00 | | 0.22 | 0.00 | |
| ***50*** |  | |  | | 0.05 | | 0.36 | | 0.24 | | 0.00 | | 0.22 | 0.00 | |
| ***51*** |  | |  | | 0.08 | | 0.05 | | 0.28 | | 0.00 | | 0.21 | 0.00 | |
| ***52*** |  | |  | | 0.07 | | 0.07 | | 0.24 | | 0.00 | | 0.28 | 0.00 | |
| ***53*** |  | |  | | 0.09 | | 0.01 | | 0.28 | | 0.00 | | 0.23 | 0.00 | |
| ***54*** |  | |  | | 0.10 | | 0.00 | | 0.29 | | 0.00 | | 0.24 | 0.00 | |
| ***55*** |  | |  | | 0.08 | | 0.04 | |  | |  | | 0.21 | 0.00 | |
| ***56*** |  | |  | | 0.09 | | 0.02 | |  | |  | | 0.21 | 0.00 | |
| ***57*** |  | |  | | 0.12 | | 0.00 | |  | |  | | 0.18 | 0.00 | |
| ***58*** |  | |  | | 0.12 | | 0.00 | |  | |  | | 0.20 | 0.00 | |
| ***59*** |  | |  | | 0.08 | | 0.05 | |  | |  | | 0.24 | 0.00 | |
| ***60*** |  | |  | | 0.11 | | 0.00 | |  | |  | | 0.24 | 0.00 | |
| ***61*** |  | |  | | 0.09 | | 0.01 | |  | |  | | 0.24 | 0.00 | |
| ***62*** |  | |  | | 0.08 | | 0.04 | |  | |  | | 0.24 | 0.00 | |
| ***63*** |  | |  | | 0.09 | | 0.01 | |  | |  | | 0.26 | 0.00 | |
| ***64*** |  | |  | | 0.12 | | 0.00 | |  | |  | | 0.24 | 0.00 | |
| ***65*** |  | |  | | 0.13 | | 0.00 | |  | |  | | 0.28 | 0.00 | |
| ***66*** |  | |  | | 0.10 | | 0.00 | |  | |  | | 0.24 | 0.00 | |
| ***67*** |  | |  | | 0.13 | | 0.00 | |  | |  | | 0.25 | 0.00 | |
| ***68*** |  | |  | | 0.15 | | 0.00 | |  | |  | | 0.28 | 0.00 | |
| ***69*** |  | |  | | 0.11 | | 0.00 | |  | |  | | 0.26 | 0.00 | |
| ***70*** |  | |  | | 0.12 | | 0.00 | |  | |  | | 0.25 | 0.00 | |
| ***71*** |  | |  | | 0.18 | | 0.00 | |  | |  | | 0.24 | 0.00 | |
| ***72*** |  | |  | | 0.19 | | 0.00 | |  | |  | | 0.18 | 0.00 | |
| ***73*** |  | |  | | 0.24 | | 0.00 | |  | |  | | 0.34 | 0.00 | |
| ***74*** |  | |  | | 0.22 | | 0.00 | |  | |  | | 0.22 | 0.00 | |
| ***75*** |  | |  | | 0.23 | | 0.00 | |  | |  | | 0.26 | 0.00 | |
| ***76*** |  | |  | | 0.23 | | 0.00 | |  | |  | | 0.24 | 0.00 | |
| ***77*** |  | |  | | 0.25 | | 0.00 | |  | |  | | 0.32 | 0.00 | |
| ***78*** |  | |  | | 0.23 | | 0.00 | |  | |  | | 0.25 | 0.00 | |
| ***79*** |  | |  | | 0.26 | | 0.00 | |  | |  | | 0.25 | 0.00 | |
| ***80*** |  | |  | | 0.24 | | 0.00 | |  | |  | | 0.24 | 0.00 | |
| ***81*** |  | |  | | 0.25 | | 0.00 | |  | |  | | 0.21 | 0.00 | |
| ***82*** |  | |  | | 0.24 | | 0.00 | |  | |  | |  |  | |
| ***83*** |  | |  | | 0.29 | | 0.00 | |  | |  | |  |  | |
| ***84*** |  | |  | | 0.18 | | 0.00 | |  | |  | |  |  | |
| ***85*** |  | |  | | 0.22 | | 0.00 | |  | |  | |  |  | |
| ***86*** |  | |  | | 0.25 | | 0.00 | |  | |  | |  |  | |
| ***87*** |  | |  | | 0.24 | | 0.00 | |  | |  | |  |  | |
| ***88*** |  | |  | | 0.20 | | 0.00 | |  | |  | |  |  | |
| ***89*** |  | |  | | 0.22 | | 0.00 | |  | |  | |  |  | |
| ***90*** |  | |  | | 0.22 | | 0.00 | |  | |  | |  |  | |
|  |  |  | |  | |  | |  | |  | |  | | |  |

A picture containing text

Description automatically generated

Figure 1. Raw Biphone Counts of Real Community 43 in Dutch

Chart

Description automatically generated with medium confidence

Figure 2. Raw Biphone Counts of Random Community 43 in Dutch

A picture containing histogram

Description automatically generated

Figure 3. Raw Biphone Counts of Real Community 88 in French

Timeline

Description automatically generated

Figure 4. Raw Biphone Counts of Random Community 88 in French

A picture containing text

Description automatically generated

Figure 5. Raw Biphone Counts of Real Community 48 in German

Chart, histogram

Description automatically generated

Figure 6. Raw Biphone Counts of Random Community 48 in German

Chart, histogram

Description automatically generated

Figure 7. Raw Biphone Counts of Real Community 72 in Spanish

Chart

Description automatically generated

Figure 8. Raw Biphone Counts of Random Community 72 in Spanish

Chart

Description automatically generated

Figure 9. Plot of K-S Test significane against D-statistic in DutchChart

Description automatically generated

Figure 10. Plot of K-S Test significane against D-statistic in French

Chart

Description automatically generated

Figure 11. Plot of K-S Test significane against D-statistic in German Chart

Description automatically generated

Figure 12. Plot of K-S Test significane against D-statistic in Spanish

Chart

Description automatically generated

Figure 13. Plot of K-S Test significane against Community Size in Dutch

Chart

Description automatically generated with medium confidence

Figure 14. Plot of K-S Test significane against Community Size in French

A picture containing text

Description automatically generated

Figure 15. Plot of K-S Test significane against Community Size in German

Text

Description automatically generated with medium confidence

Figure 16. Plot of K-S Test significane against D-Community Size in Spanish