A connective differentiation of textual production in interaction networks

Renato Fabbri\textsuperscript{1,a) (Dated: 24 December 2014)}

Instituto de Física de São Carlos, Universidade de São Paulo (IFSC/USP)

This paper explores textual production in interaction networks, with special emphasis on its relation to topological measures. Four email lists were selected, in which measures were taken from the texts participants wrote. Peripheral, intermediary and hub sectors of these networks were observed to have discrepant linguistic elaborations. For completeness of exposition, correlation of textual and topological measures were observed for the entire network and for each connective sector. The formation of principal components is used for further insights of how measures are related.

PACS numbers: 89.75.Fb,05.65.+b,89.65.-s

Keywords: complex networks, natural language processing, social network analysis, statistical physics

I. INTRODUCTION

Textual production has received considerable attention from the social network analysis community. Sentiment analysis and vocabularies related to different parties are among a number of examples\textsuperscript{10}. The relation of topological and textual measures is the subject of this article, for the following reasons:

- This relation has been set aside in literature, with scattered and vague suggestions of mutual implications of the text produced and topological characteristics of the agents in the network\textsuperscript{10}.
- This results eases understanding of human interaction, which is useful for both psychological and anthropological typologies (personality and cultural “types”)\textsuperscript{7}.
- There are interesting hypothesis about verbal differentiation of network sections and groups, derived from a previous article by the same author\textsuperscript{11}, some of which were herein confirmed.

Next section exposes materials used for this research, its textual and network facets. Section II A explains the analysis roadmap, with the measures chosen and methods for understanding data. Section II A is dedicated to detailing results and discussion. Section II A has concluding remarks and further works envisioned. Appendices hold information about mailing lists, tables, results still to interpret and directions on data and scripts.

II. MATERIALS

Eighty thousand messages were analysed, twenty thousand from each email list (see Appendix B). This data was accessed online through the GMANE database\textsuperscript{14}. Each message has an ID, the ID of the message it is a response to (if any), an author, a “date and time” field registering the moment the message was sent, and the textual content. Other fields are also available, but plays no direct role in the work here presented. Basic information of messages, threads and authors are summarized in Table II.

A. Network formation

FIG. 1. Formation of interaction network. Edges are directed as information flows, from an original message’s author to the observed responder. Further information is given in Section II A.

Message-response pairs yield interaction networks, such as shown in Figure I. Each participant is represented by a vertex, and each response is considered evidence that information emitted by the first author was received by the responder (that had to read, process its contents and render a relevant textual response). Therefore, an edge from first author to the second author (responder) is considered. This is the “information network” of the system. Edges can be considered in the reverse order, from the responder to the original sender, representing status attribution, as the responder considered what the sender said worthy of responding and is directing his attention to him. This is the “status network.”
As these networks are virtually equivalent, one considers but one of them, usually the information network.

B. Verbal observations

Each message has a textual content. Analysis this content can consider author, network section or community, or messages independently. As these are informal communities, there are typos, leet and invented words. This diversity and informality poses some challenges, by which the methodology was shaped.

III. METHODOLOGY

An article was written for reporting stability in such networks from the topological viewpoint\(^1\). This article is dedicated to reporting differentiation in the textual production of the network as connectivity changes. Here, the observance of primary textual statistics is needed, and both overall incidences, and correlation to topological aspects, were tackled.

It is coherent to have participants as vertexes and as references for the messages sent, for the text produced and for activity (related to time and date). Thus, to observe the text produced in a certain section, one might gather all text produced by all participants on that section. To observe correlation of textual and topological characteristics, one can take measures on each vertex.

A. Network measurements and partitioning

Basic network measures of connectivity, in the same networks, were observed in a previous article\(^1\). The present article uses the same topological measures to observe correlations, PCA formation and network sectioning in peripheral, intermediary and hub sectors, through strength measure. As described in that article, the “exclusivist criteria” for such partitioning is found to be the closest to literature predictions (5% of hubs, 15% of intermediary and 80% of peripheral vertex). Even so, strength-based criteria is simpler and yields reasonable results (5-10%, 5-25%, 65-90%). Beyond that, changing the sectioning to a degree or a compound criteria did not significantly change the presented results.

Consequently, herein is considered a strength partitioning each sector (periphery, intermediary, hubs) is regarded as a primitive sector of the primitive partitioning.

B. Textual measures

An infinitude of textual measures can be drawn from texts. This work focuses on the simplest of them, as they proved sufficient for current step. These measures include frequency of individual letters and punctuations (Tables [I]), of words and tokens (Table [III]), sizes of tokens, sentences and messages (Table [IV] and [VI] and POS (Part-Of-Speech) tags (Table [VII]). Other measures envisioned are in subsection V A.

This choice is based on: 1) the lack of such information in literature, as far as authors know; 2) potential relations of these incidences with topological aspects, such as connectivity; 3) the interdependence of textual artifacts suggests that simple measures should reflect complex behaviors subtle aspects. A preliminary study, with all the work from Machado de Assis\(^1\) made clear that these measures vary with respect to style. Considered measures are:

- Frequency of characters: letters, vowels, punctuations and uppercase. Table [I] is dedicated to such measures.
- Number of tokens, frequency of punctuations, of known words, of words that has wordnet synsets, of tokens that are stopwords, of words that return synsets and are stop words, etc. Table [III] is dedicated to measures of this kind.
- Mean and standard deviation for word and token sizes. Table [IV] is dedicated to these measures.
- Mean and standard deviation of sentence sizes. Table [V] is dedicated to this sort of measures.
- Mean and standard deviation of message sizes. Table [VI] presents some of these measures.
- Fraction of morphosyntactic classes, such as adverbs, adjectives and nouns, represented by POS (Part-Of-Speech) tags. Table [VII] displays such measures.

For sections (hubs, intermediary and peripheral), all messages written by authors in each section were considered together. For the histograms, independent messages were considered from each sector.

C. Topological measures

Degree (in, out and total), strength (in, out and total), betweenness centrality and clustering coefficient were measured for each vertex in the interaction network. This served two purposes:

- Obtaining a sound partitioning of the network in peripheral, intermediary and hub sectors. This was developed in a previous article by the same author\(^1\).
- Observance of correlation with textual measures and principal components formation.

These measures are not developed here extensively as they are very consolidated, simple, and was the core of a previous article this subject by the same author\(^1\).
D. Relating text and topology

Key observations for a deeper insight about network structure depend on theoretical background and intentions. For this article, we considered:

1. Incidences of linguistic traces in hub, intermediary and peripheral network sectors.

2. Correlation of measures of each vertex, easing pattern detection involving topology of interaction and language used.

3. PCA to gain further insights about how measures combine in principal components formation.

Criteria for this choice include integration with previous topological results, lack of concise results in literature (as far as author knows) that could substantiate correlations of topological and textual traces, and common sense as a long-time member of these networks.

First task, of textual production in hubs, intermediary and peripheral sectors, is observed by Tables II-XV. An adaptation of the Kolmogorov-Smirnov test was used to observe differences in textual productions, as exposed in Appendix A.

Second task is addressed by the correlation matrix with both textual and topological measurements of each participant, in Tables XVI-XVIII. Third, principal components composition are in Tables XIX-XXIII.

IV. RESULTS AND DISCUSSION

Although the results drawn from experiments and statistics were diverse, some fundamental insights can be obtained by going through tables and figures in the Appendix B. Most importantly: connectivity has strong influence in textual production of participants in the network. For example: hubs use more contractions, more adjectives, more common words, and less punctuation if compared to the rest of the network, specially the peripheral sector. Fewer threads are created in proportion to total messages sent by the hubs, while threads created by peripheral are twice as frequent as general peripheral messages. This suggests a symbiosis of peripheral diversity and hub activity.

Also, comparing lists with a fixed number of messages, the number of threads created seem to increase as the number of participants decrease. These information is condensed in Table II with further details.

B. Characters

Peripheral vertex use more punctuation characters, digits and uppercase letters. Hubs use more letters and vowels among letters. The use of space does not seem to have any relation to connectivity, with the exception that the intermediary presented a slightly lower incidence of spaces than both peripheral and hub sectors.

Total number of characters in ELE list, in the 20 thousand messages, is more than three times what other lists exhibited. This suggests peculiarities related to communication conventions and style (see Appendix B1).

Further information is given in Table II.

C. Tokens and words

Largest average size of tokens is with the most wordy list (ELE). This implies that is has more characters, tokens, and characters per token in comparison to the other lists. Longer words used by hubs might be related to the use of a specialized vocabulary. Although the token diversity \( \frac{|\text{tokens}|}{|\text{tokens}|} \) found in peripheral sector is far greater, this result has the masking artifact that the peripheral sector corpus is smaller, yielding a larger token diversity. This can be noticed by the token diversity of the whole network, which is lower than in the sections. This same discussion applies to the lexical diversity \( \frac{|\text{kw}|}{|\text{kw}|} \).

Punctuations among tokens are less abundant in hubs, and discrepancies here are larger that with characters comparisons (subsection V B). Known words are used more frequently by hubs.

ELE and CPP both exhibit intermediaries with the more frequent production of punctuation, less frequent production of known words, the highest incidence of words with wordnet synsets among known words. This suggests some peculiarity in network structure, such as the intermediary be strong authorities in such networks, using smaller sentences and a larger jargon.

Words with synsets, among known English words, are less frequent in hubs further evidencing the jargon hubs develop.

Further information is given in Table III.

D. Sizes of tokens and words

Sizes of known words are smaller for hubs, which suggests its use of more common words, although some of the
previous results suggests that hubs have a very differentiated and specialized vocabulary. Larger words seems to
be related to intermediary sector, which might be related
to cultured vocabulary.

Further information is given in Table IV.

E. Sizes of sentences

Hubs present the lowest average sentence size, both in characters and in tokens. Also, the incidence of usual
known words seems to decay with connectivity, as does the number of known words with synsets.

Further information is given in Table V.

F. Messages

Regarding characters and tokens, connectivity was related to smaller messages. ELE list displayed an inverse
situation: the more connected the sector, the longer the
messages are. This was considered a peculiarity of the
culture bonded with the political subject of ELE list, to
be further verified. Regarding sentences, the size of mes-
sages seem to hold steady until hubs are reached.

Further information is given in Table VI.

G. POS tags

Lower connectivity delivers more nouns and less adjectives, adverbs and verbs. This suggests that the networks
collect issues important to the world by the peripheral
sector. These issues are qualified, elaborated about, by
the more connected participants.

Further information is given in Table VII.

H. Differentiation of measures

The Kolmogorov-Smirnov test was adapted for our
need to compare measures. Results suggests that the
texts produced by each sector is very distinct. Counterin-
tuitively, intermediary sector sometimes yields a greater
difference from periphery and hubs than these extreme
sectors themselves (Tables VIII and XII).

At the core of the results presented on this article, are
two strong and immediate interpretations that follows
Tables VIII-XV:

- Differences of textual production of the primitive
  sectors are extreme, as can be noticed from the
  values on these tables, beyond reference values
  used for considering the null hypothesis (see Ap-
  pendix A).

- Differences between sectors on the same network
  (Tables VIII, X, XII and XIV) are bigger than
differences between same sector from distinct lists
  (Tables IX, XI, XIII and XV).

I. Correlation of measures

Correlation of degree (how many participants the par-
ticipant related to) and strength (how many interactions)
measures is substantially smaller for intermediary sector.
This raises interesting inquiries, to which the reader is
invited, along with further analysis of Tables XVI, XVII
and XVIII as detailing their interpretation goes beyond
the scope of this article. Noteworthy is the negative cor-
relation of degree and message size (number of characters,
tokens or sentences) that intermediaries presented.

J. Formation of principal components

Principal components formation seem to be the less
stable of all features considered. First component, with
≈ 25% of dispersion, relies heavily on POS tags, and
slightly on sizes of tokens, sentences and messages. Sec-
ond component, with almost 12% of dispersion, blends
topology, POS tags and size measures. Third compo-
nent, with about 8.5% is based on nouns frequency and
size measures. Fourth and fifth components present less
than 5% of total dispersion, but are included for com-
pleteness of exposition.

Tables XIX, XXIII exhibit these and further insights.

K. Results still to be interpreted

These networks yield diverse characteristics, some of
which were not of core importance for this step of the
research. Even so, at least one of these characteristics
was found interesting enough to be considered a result
and an example of interesting artifacts found.

Histogram differences of incident and existent word
sizes were found constant. That is, in each list, when a
histogram of word sizes were made with all words written,
and another histogram made with sizes of all different
words, the cumulative positive difference of the two his-
tograms were found constant for all lists analysed. When
all known English words were considered, the difference
was always ≈ 1.0. When stopwords were discarded, the
difference found was different, but still constant, slightly
above 0.5. When only stopwords were considered, the dif-
ference was ≈ 0.6. When only known English words that
does not have wordnet synsets are used, this difference is
≈ 1.2.

These results currently lacks substantial interpreta-
tion, which is provocative and should lead at least to
a research note. Appendix C and Figures 2-6 are dedi-
cated to this histogram differences.
V. FINAL REMARKS

Human interaction networks yield diverse linguistic peculiarities reported by its members. This is a first systematic exploration of such peculiarities with primitive connective sectors (hubs, intermediary and periphery) in mind, as far the author knows. Results were regarded as stronger than envisioned from start, which poses diverse and intriguing questions. This results, confluent with recent research and development, some by the current author, are of core importance for social technologies and transformations, such as collection and diffusion and information, resource recommendation in linked data contexts, and open processes of legal documents refinement.

All the data used is public, all scripts used are online (see Appendix D).

A. Further work

Results suggest that less connected participants bring external content and concepts, while hubs qualify the content. This hold mainly as periphery uses more nouns while hubs present more adjectives. This should be further verified, maybe with a dedicated article.

Similarity measures of texts in message-response threads have been thought about by the author, and some results are being organized. These are two hypothesis obtained from recent experiments:

- Existence of information “ducts”, observable through similarity measures. These might coincide with asymmetries of edges between vertexes pairs, with homophily or with message-response threads, to point just a few possibilities.

- Valuable insights can be driven from self-similarity of messages by same authors, of messages sent at the same period of the day, etc. This includes incidences of word sizes, incidences of tags and morphosintactic classes, incidences of particular wordnet synset characteristics and wordnet word distances.

Given current results, diversity and self-similarity should vary with respect to connectivity. Literature usually assumes that periphery holds greater diversity, which should be further verified.

Other directions for next steps are:

- Word sets are very useful to derive and confirm hypothesis. As an example, one can observe most incident or most basic words and word types in the English language, curses or words related to food.

- Interpretation of various unveiled results, such as the one exposed in Appendix C and Figures 2, 3, 4, 5 and 6.

- Extend word class observations to include plurals, gender, common prefixes and suffixes, etc.

- Date and time should also be addressed in textual production of interaction networks, as potentially linked to participation habits and purposes (e.g. low dispersion of sent time). This was tackled by the author for the topological characterization of interaction networks, but left aside in this article.

- Balance token diversity with corpus size, as pointed in section IV C.

- The textual features distributions are likely to be have more than one peak or other non-trivial characteristic. Therefore, further analysis should be made for comparing measures of interest.

- Extend analysis to the windowed approach used in the article where hub, peripheral and intermediary sectors where topologically characterized.

- For ELE list, the more connected the sector, the longer the messages are. This is the inverse of what was found in the other lists, and was considered a peculiarity of the culture bonded with the political subject of ELE list, to be further verified.

- Tackle Portuguese analysis of interaction networks, as this research have ongoing implications in Brazil.

- Analyse other lists.

- Analyse interaction networks from other platforms such as Twitter, Facebook, LinkedIn, Diaspora, etc.

- Emotion classification has not been done and was considered out of the scope for this stage of development, but should be addressed in a near future.

Wordnet synsets incidences was studied as well, as a potentially useful way to characterize networks and sectors. Core aspects understood as useful for this research include:

- Incidence of hypernyms, hyponyms, holonyms and meronyms.

- Use and development of similarity measures of words, phrases and messages, by use of semantic criteria (Wordnet) and bag of words.

VI. ACKNOWLEDGMENTS

Renato Fabbri is grateful to CNPq (processo 140860/2013-4, project 870336/1997-5); Postgraduate Committee of the IFSC-USP; Prof. Dr. Leonardo Paulo Maia for the insights which lead to the adaptation of the Kolmogorov-Smirnov test presented in Appendix A.
GMANE developers and maintainers; participants of the mailing lists analysed.

1. Canonical Kolmogorov-Smirnov test

Let $F_{1,n}$ and $F_{2,n'}$ two empirical distribution functions, where $n$ and $n'$ are the number of observations on each sample. The two-sample Kolmogorov-Smirnov test rejects the null hypothesis if:

$$D_{n,n'} > c(\alpha) \sqrt{\frac{n + n'}{nn'}}$$ (A1)

where $D_{n,n'} = \sup_x |F_{1,n} - F_{2,n'}|$ and $c(\alpha)$ is given for each level of $\alpha$:

| $\alpha$  | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 |
|----------|-----|------|-------|------|-------|-------|
| $c(\alpha)$ | 1.22 | 1.36 | 1.48 | 1.63 | 1.73 | 1.96 |

2. Adaptation

We need to compare empirical distribution functions, so $D_{n,n'}$ is given, as are $n$ and $n'$. Therefore, as all terms in equation [A1] are positive and $c(\alpha)$ can be isolated:

$$c(\alpha) < \frac{D_{n,n'}}{\sqrt{\frac{n + n'}{nn'}}} = c'(\alpha)$$ (A2)

Tables VIII-XV are populated with values for $c'(\alpha)$. When $c'(\alpha)$ is high, low values of $\alpha$ are possible for the test to reject the null hypothesis. Therefore, when $c'(\alpha)$ is greater than $\approx 1.7$, it is reasonable to assume that $F_{1,n}$ and $F_{2,n'}$ differ.

Appendix B: Support information

1. Brief description of the email lists chosen

GMANE is a public email list database with some tenths of thousand of list.[14] Four email lists were selected, in a similar fashion developed in[11], but with MET substituted by ELE list so that all lists are in English. The lists are:

- CPP, the development list of the standard C++ library.[14] Dominated by specialized computer programmers.
- LAD: Linux Audio Developers list.[2]
- LAU: Linux Audio Users list.[3]
- ELE: list for discussion of the election reform.[3]

Table 1 has an overview of these lists, in terms of participants, threads and messages in each of the primitive connective sectors.
2. **Meaning of achronims and abbreviations in the following tables**

| symbol | meaning |
|--------|---------|
| \(|x|\) | the number of times \(x\) was found |
| \(kw\) | known word |
| \(|x \neq|\) | number of different \(x\) found |
| \(kwss\) | known word with (wordnet) synset |
| \(kwsw\) | known word that is a stopword |
| \(ukwsw\) | unknown word that is a stopword |
| \(nsssw\) | word without (wordnet) synset that is a stopword |

Other symbols are explained on the tables itself. Some concepts, such as *contractions*, *token* and *char* are standard in natural language processing, and the reader is invited to visit [8].

3. **Tables**
TABLE I. Columns date1 and dateM have dates (month/day/year) of first and last messages from the 20,000 messages considered. N is the number of participants (number of different email addresses). M is number of messages. Γ is the number of threads (count of messages without antecedent). −M is messages missing in the 20,000 collection, 100 \frac{\text{M}}{\text{M}_{27000}} = 0.27/100 in the worst case. ELE notably has the fewest participants and the largest number of threads. This relation holds for pairs of lists considered: as the number of participants increase, the number of threads decrease. A similar role is observed in MET list described in \cite{12}, suggesting that 1) Non-technical topics gathers fewer participants and yields shorter threads; 2) MET technopolitical characteristic is confirmed by having intermediary relation, between ELE (politics) and LAD (highly technical - GNU/Linux and music). These results should be further investigated in future research (see section IV and subsection IV A). The number of threads started by hubs is significantly lower than activity for all list, this suggests creative exploitation is done by hubs, i.e. hubs acquire/absorb creativity. ∆Y is number of years involved in the first 20,000 messages of each list. Dates of first and last message is: Mar/13/2002 and Aug/25/2009 for CPP; Jun/30/2003 and Oct/07/2009 for LAD; Jun/29/2003 and Jul/23/2005 for LAU; finally, Abr/18/2002 and Aug/31/2011 for ELE. See section IV and subsection IV A for further directions.

|   | CPP | LAD | LAU | ELE |
|---|-----|-----|-----|-----|
|   | g. | p. | i. | h. | g. | p. | i. | h. | g. | p. | i. | h. | g. | p. | i. | h. |
| date1 | 4/13/2 | - | - | - | 6/30/3 | - | - | - | 08/29/3 | - | - | - | 10/07/9 | - | - | - |
| dateM | 8/25/9 | - | - | - | 10/07/9 | - | - | - | 07/23/5 | - | - | - | 07/23/5 | - | - | - |
| N | 1052 | 84 | 163 | 55 | 1268 | 936 | 210 | 122 | 1183 | 904 | 155 | 124 | 302 | 225 | 36 | 41 |
| N% | - | 79.28% | 15.4% | 5.23% | - | 73.82% | 16.56% | 9.62% | - | 76.42% | 13.16% | 10.48% | - | 74.50% | 11.92% | 13.56% |
| M | 19993 | 1654 | 2673 | 1566 | 19996 | 2331 | 3542 | 14123 | 19995 | 3018 | 2882 | 14995 | 19996 | 1821 | 2413 | 15712 |
| M% | - | 8.27% | 13.37% | 78.33% | - | 11.65% | 17.71% | 70.61% | - | 15.09% | 14.41% | 70.47% | - | 9.11% | 12.06% | 78.56% |
| Γ | 4506 | 924 | 702 | 2880 | 3113 | 812 | 670 | 1631 | 3373 | 1121 | 675 | 1577 | 6070 | 782 | 1072 | 4216 |
| Γ% | - | 20.51% | 15.58% | 63.91% | - | 26.08% | 21.52% | 52.39% | - | 33.23% | 20.01% | 46.75% | - | 12.88% | 17.66% | 69.46% |
| −M | 7 | - | - | - | 4 | - | - | - | 5 | - | - | - | 54 | - | - | - |
| ΔY | 7.44 | - | - | - | 6.25 | - | - | - | 2.08 | - | - | - | 9.37 | - | - | - |

TABLE II. Measures based on characters of the text produced by network participants, fairly stable. Suggested relations are: 1) punctuations of CPP, that can be expected by its programming language focus and dots and semicolon abundance in such parlance; 2) greater number of letters on ELE is expected by its political disposition; 3) not statistically clear, but higher percentage of vowels might be a sign of erudition. Most of all, number of characters incident in ELE 20,000 messages are more than the other three lists added. MET has an intermediary value of 13,137,042 characters considered: as the number of participants increase, the number of threads decrease. A similar role is observed in MET list described in \cite{12}. This builds up to a dichotomic typology of networks: technical (more participants, fewer and longer threads, e.g. CPP) – political (less participants, more and shorter threads, e.g. ELE). Higher incidence of digits and lower incidence of letters seem to be associated to technical subjects. See subsection IV B for further discussion and context.
TABLE III. Basic measures on tokens, known English words, stopwords, words with and without synset. Lexical diversity is higher in LAU and LAD, probably linked to these lists hybrid technical interests (music and GNU/Linux). Larger known words and tokens are incident in ELE and LAD. ELE also exhibits larger incidence of stopwords without synsets (prolixity?). Stronger use of words with synsets that are not stopwords is held by CPP. Stopwords that have synset account for ≈ 25% of all known words, which might be an indicative of language complexity (not same as good writing though). See subsection ?? for further discussion and context.
TABLE IV. Sizes of tokens and words. Practically all sizes are greater for ELE. See subsection [IVD] for discussion and context.

| - | CPP | LAD | LAU | ELE |
|---|-----|-----|-----|-----|
| $\mu$ (size of known word = $\sigma$) | 4.51 | 4.44 | 4.35 | 4.46 |
| $\sigma$ (known word) | 2.39 | 2.35 | 2.25 | 2.54 |
| $\mu$ (number of known words) | 7.52 | 7.54 | 7.43 | 7.92 |
| $\sigma$ (number of known words) | 2.57 | 2.53 | 2.51 | 2.62 |
| $\mu$ (size of known word) | 4.92 | 4.82 | 4.70 | 5.11 |
| $\sigma$ (size of known word) | 2.54 | 2.50 | 2.40 | 2.69 |
| $\mu$ (number of tokens) | 7.56 | 7.57 | 7.47 | 7.94 |
| $\sigma$ (number of tokens) | 2.54 | 2.49 | 2.48 | 2.58 |
| $\mu$ (size of tokens) | 2.89 | 2.85 | 2.86 | 2.88 |
| $\sigma$ (size of tokens) | 1.06 | 1.05 | 1.04 | 1.09 |
| $\mu$ (number of words) | 3.92 | 3.97 | 3.97 | 3.97 |
| $\sigma$ (number of words) | 1.60 | 1.60 | 1.60 | 1.68 |
| $\mu$ (size of words) | 3.01 | 2.97 | 2.99 | 2.99 |
| $\sigma$ (size of words) | 1.25 | 1.25 | 1.25 | 1.25 |
| $\mu$ (number of sentences) | 6.32 | 6.65 | 6.48 | 7.37 |
| $\sigma$ (number of sentences) | 3.07 | 3.07 | 3.07 | 3.37 |

TABLE V. Sizes of sentences in characters and in tokens. Hubs produce the smallest sentences and, at the same time, present the lowest incidence of known words and of known words with synsets. See subsection [IVE] for discussion and context.

| - | CPP | LAD | LAU | ELE |
|---|-----|-----|-----|-----|
| $\mu$ (size of sentences in characters) | 106086 | 113033 | 111703 | 325399 |
| $\sigma$ (size of sentences in characters) | 110.15 | 116.16 | 105.15 | 117.67 |
| $\mu$ (size of sentences in tokens) | 118.31 | 110.52 | 105.15 | 117.67 |
| $\sigma$ (size of sentences in tokens) | 125.02 | 116.16 | 107.55 | 126.06 |
| $\mu$ (number of tokens per sentence) | 250.34 | 148.98 | 208.63 | 127.57 |
| $\sigma$ (number of tokens per sentence) | 312.02 | 243.78 | 386.51 | 120.50 |
| $\mu$ (size of tokens per sentence) | 64.74 | 33.44 | 38.11 | 34.48 |
| $\sigma$ (size of tokens per sentence) | 81.47 | 51.90 | 51.39 | 27.38 |
| $\mu$ (number of words per sentence) | 13.88 | 15.15 | 14.11 | 17.03 |
| $\sigma$ (number of words per sentence) | 16.09 | 15.76 | 14.39 | 17.76 |
| $\mu$ (size of words per sentence) | 6.90 | 7.26 | 6.67 | 8.19 |
| $\sigma$ (size of words per sentence) | 8.13 | 7.79 | 7.06 | 8.60 |

Textual differentiation in interaction networks
|        | CPP     | LAD     | LAU     | ELE     |
|--------|---------|---------|---------|---------|
|        | g.      | p.      | i.      | h.      | g.      | p.      | i.      | h.      | g.      | p.      | i.      | h.      | g.      | p.      | i.      | h.      |
| µ (|chars| msg) | 632.81  | 883.15  | 841.05  | 570.09  | 628.49  | 763.32  | 655.59  | 599.39  | 591.12  | 697.59  | 623.79  | 561.61  | 1934.43 | 1638.41 | 1996.38 | 1993.42 |
| σ (|chars| msg) | 1761.57 | 1247.79 | 3896.49 | 1101.55 | 836.23  | 1136.90 | 826.08  | 770.30  | 831.47  | 1194.85 | 982.59  | 686.75  | 2642.25 | 1737.49 | 1992.88 | 2819.96 |
| µ (|tokens| msg) | 143.35  | 202.36  | 194.09  | 128.28  | 135.99  | 164.49  | 141.88  | 129.81  | 131.37  | 153.18  | 139.27  | 125.01  | 406.39  | 347.64  | 383.28  | 417.36  |
| σ (|tokens| msg) | 444.20  | 287.17  | 940.83  | 304.37  | 178.11  | 237.80  | 172.03  | 165.98  | 173.89  | 213.52  | 212.91  | 152.35  | 557.29  | 365.05  | 435.87  | 593.08  |
| µ (|sents| msg) | 5.71    | 6.39    | 7.09    | 5.40    | 6.12    | 6.55    | 6.11    | 6.04    | 6.08    | 6.23    | 6.23    | 6.01    | 17.22   | 13.74   | 14.79   | 18.05   |
| σ (|sents| msg) | 16.36   | 6.29    | 41.76   | 6.55    | 6.75    | 7.51    | 6.67    | 6.61    | 6.58    | 8.03    | 8.87    | 6.18    | 23.97   | 14.06   | 17.01   | 25.80   |

TABLE VI. Mean and standard deviation of message sizes. Greater size of messages from ELE list reflects domain of interest, as does its hubsi sector, which produces the largest texts. See subsection IV F for discussion and context.
### TABLE VII. Incidence of Brown Tags. Used Brill tagger with \( \approx 85\% \) of correctly identified tags on the Brown Corpus. Most explicit is the peripheral incidence of nouns and hubs incidence of adjectives, adverbs and verbs. See subsection IV C for discussion and context.

|       | CPP | LAD | LAU | ELE |
|-------|-----|-----|-----|-----|
|       | g.  | p.  | i.  | h.  |
| NN    | 28.17 | 30.38 | 31.13 | 27.19 |
| NNS   | 2.51  | 2.32 | 2.56 | 2.53 |
| NNP   | 0.72  | 0.75 | 1.03 | 0.65 |
| NNPS  | 0.01  | 0.01 | 0.00 | 0.01 |
|       | 31.41 | 33.46 | 34.73 | 30.38 |
| JJ    | 4.83  | 4.60 | 4.72 | 4.89 |
| JJR   | 0.45  | 0.37 | 0.38 | 0.48 |
| JJS   | 0.17  | 0.15 | 0.14 | 0.17 |
| RB    | 6.43  | 5.29 | 5.73 | 6.76 |
| RBR   | 0.11  | 0.08 | 0.09 | 0.12 |
| RBS   | 0.02  | 0.01 | 0.01 | 0.02 |
| RP    | 0.35  | 0.30 | 0.27 | 0.37 |
|       | 12.36 | 10.79 | 11.34 | 12.82 |
| VB    | 6.25  | 6.24 | 6.31 | 6.25 |
| VBZ   | 3.94  | 3.89 | 3.80 | 3.97 |
| VBP   | 3.17  | 3.07 | 3.17 | 3.18 |
| VBN   | 2.00  | 2.14 | 2.06 | 1.97 |
| VBD   | 1.52  | 1.64 | 1.49 | 1.50 |
| VBG   | 1.50  | 1.66 | 1.41 | 1.50 |
| MD    | 2.20  | 1.78 | 2.09 | 2.28 |
|       | 20.58 | 20.42 | 20.32 | 20.66 |
| IN    | 12.60 | 12.49 | 12.08 | 12.73 |
| DT    | 10.76 | 10.96 | 10.33 | 10.82 |
| PRP   | 3.62  | 2.83 | 3.02 | 3.87 |
| PRPS  | 0.73  | 0.85 | 0.56 | 0.75 |
| PDT   | 0.08  | 0.08 | 0.07 | 0.08 |
| TO    | 2.93  | 2.94 | 2.87 | 2.94 |
| CC    | 2.77  | 2.97 | 2.54 | 2.79 |
| WRB   | 0.58  | 0.68 | 0.56 | 0.56 |
| WDT   | 0.54  | 0.53 | 0.55 | 0.54 |
| WP    | 0.32  | 0.28 | 0.29 | 0.33 |
| WP$   | 0.00  | 0.00 | 0.00 | 0.00 |
|       | 34.93 | 34.62 | 32.87 | 35.42 |
| CD    | 0.38  | 0.37 | 0.36 | 0.38 |
| EX    | 0.27  | 0.28 | 0.29 | 0.27 |
| UH    | 0.07  | 0.04 | 0.08 | 0.07 |
| FW    | 0.01  | 0.03 | 0.00 | 0.00 |
|       | 0.72  | 0.71 | 0.74 | 0.72 |
| list | measure | H-P | H-I | I-P |
|------|---------|-----|-----|-----|
| CPP  | 5.58    | 2.54| 7.82|     |
| LAD  | 7.67    | 2.07| 8.35|     |
| LAU  | 6.23    | 1.63| 5.98|     |
| ELE  | 3.42    | 0.77| 2.81|     |

**TABLE VIII.** Kolmogorov $c(\alpha)$ values for substantives. See subsection [IV.H] for discussion and directions.

| list | measure | H-P | H-I | I-P |
|------|---------|-----|-----|-----|
| CPP  | 1.53    | 0.89| 1.45|     |
| LAD  | 2.32    | 0.97| 2.09|     |
| LAU  | 2.10    | 0.78| 1.68|     |
| ELE  | 1.51    | 1.32| 1.15|     |

**TABLE XIV.** Kolmogorov $c(\alpha)$ values for punctuations/char. See subsection [IV.H] for discussion and directions.

| list | measure | H-P | H-I | I-P |
|------|---------|-----|-----|-----|
| CPP  | 1.53    | 0.89| 1.45|     |
| LAD  | 2.32    | 0.97| 2.09|     |
| LAU  | 2.10    | 0.78| 1.68|     |
| ELE  | 1.51    | 1.32| 1.15|     |

**TABLE XII.** Kolmogorov $c(\alpha)$ values for stopwords. See subsection [IV.H] for discussion and directions.

| list | measure | H-P | H-I | I-P |
|------|---------|-----|-----|-----|
| CPP  | 5.74    | 4.88| 8.28|     |
| LAD  | 3.23    | 2.49| 4.16|     |
| LAU  | 2.49    | 1.87| 4.02|     |
| ELE  | 2.49    | 1.87| 4.02|     |

**TABLE XV.** Kolmogorov $c(\alpha)$ values for punctuations/char. See subsection [IV.H] for discussion and directions.

### Subsection [IV.H] for discussion and directions.
|                        | CPP | LAD | LAU | ELE |
|------------------------|-----|-----|-----|-----|
| \(d - d_o\)           | 0.9972 | 0.8188 | 0.9477 | 1.0162 |
| \(d - d_s\)           | 0.9932 | 0.8517 | 0.9381 | 1.0126 |
| \(d - s\)             | 0.9572 | 0.9167 | 0.8598 | 0.9835 |
| \(d - s_t\)           | 0.9539 | 0.7715 | 0.8329 | 0.9893 |
| \(d - s_o\)           | 0.9547 | 0.7662 | 0.7574 | 0.9692 |
| \(d - bc\)            | 0.9698 | 0.5773 | 0.7471 | 0.9921 |
| \(d - \text{triangles}\) | 0.9716 | 0.7773 | 0.9342 | 0.9742 |
| \(d_t - d_o\)         | 0.9787 | 0.3936 | 0.7615 | 1.0031 |
| \(d_t - s\)           | 0.9595 | 0.7348 | 0.8066 | 0.9757 |
| \(d_t - s_t\)         | 0.9601 | 0.9315 | 0.8802 | 0.9838 |
| \(d_t - s_o\)         | 0.9523 | 0.3664 | 0.6185 | 0.9587 |
| \(d_t - bc\)          | 0.9780 | 0.4765 | 0.7036 | 0.9970 |
| \(d_t - \text{triangles}\) | 0.9599 | 0.5910 | 0.8621 | 0.9634 |
| \(d_s - s\)           | 0.9413 | 0.7934 | 0.8052 | 0.9866 |
| \(d_s - s_o\)         | 0.9457 | 0.8894 | 0.8084 | 0.9770 |
| \(d_s - bc\)          | 0.9452 | 0.4872 | 0.6967 | 0.9749 |
| \(d_o - \text{triangles}\) | 0.9756 | 0.7021 | 0.8903 | 0.9825 |
| \(s - s_t\)           | 0.9985 | 0.7926 | 0.9230 | 1.0162 |
| \(s - s_o\)           | 0.9971 | 0.8764 | 0.9345 | 1.0146 |
| \(s - \text{triangles}\) | 0.9298 | 0.6961 | 0.8118 | 0.9518 |
| \(s_t - s_o\)         | 0.9886 | 0.3980 | 0.7088 | 1.0062 |
| \(s_t - \text{triangles}\) | 0.9227 | 0.5365 | 0.7822 | 0.9552 |
| \(s_o - \text{triangles}\) | 0.9321 | 0.6209 | 0.7191 | 0.9410 |
| \(be - \text{triangles}\) | 0.9055 | 0.4769 | 0.6933 | 0.9031 |
| \(IC - IP\)           | -1.0010 | -1.0012 | 0.0000 | 0.0000 |

TABLE XVI. Correlation of topological measures. See subsection IV I for discussion and directions.
|                | CPP   |       | LAD   |       | LAU   |       | ELE   |       |
|----------------|-------|-------|-------|-------|-------|-------|-------|-------|
|                | g.    | p.    | i.    | h.    | g.    | p.    | i.    | h.    |
| nc-nt          | 1.000 | 0.978 | 0.992 | 1.018 | 1.000 | 0.994 | 1.001 | 1.008 |
| np/(nc-ne)-nt  | 0.934 | 0.936 | 0.930 | 0.983 | 0.893 | 0.890 | 0.956 | 0.932 |
| nt-ntd         | 0.927 | 0.870 | 0.837 | 0.988 | 0.943 | 0.918 | 0.954 | 0.967 |
| Nwss/Nkw-Nwss/Nkw_ | 0.805 | 0.862 | -0.107 | -0.401 | 0.877 | 0.922 | 0.882 | -0.040 |
| Nwsw/Nkw-Nwssw/Nwss | 0.890 | 0.882 | 0.960 | 0.995 | 0.903 | 0.899 | 0.935 | 0.941 |
| mtkw-mtkwnsw   | 0.855 | 0.868 | 0.386 | 0.388 | 0.941 | 0.943 | 0.943 | 0.769 |
| mtkw-mtkwnsw_  | 0.849 | 0.878 | 0.447 | 0.125 | 0.915 | 0.939 | 0.929 | 0.426 |
| mtkw-mtams     | 0.855 | 0.867 | 0.434 | 0.450 | 0.942 | 0.944 | 0.946 | 0.786 |
| mtkw-dtkw      | 0.962 | 0.969 | 0.739 | 0.612 | 0.979 | 0.984 | 0.942 | 0.660 |
| mtkw-dtkwnsw   | 0.851 | 0.854 | 0.788 | 0.814 | 0.927 | 0.926 | 0.942 | 0.920 |
| mtkw-mtams     | 0.848 | 0.850 | 0.825 | 0.815 | 0.929 | 0.929 | 0.945 | 0.921 |
| mtkw-mtams_    | 0.887 | 0.887 | 0.882 | 0.805 | 0.929 | 0.928 | 0.928 | 0.962 |
| mtkw-mtams_    | 0.826 | 0.838 | 0.784 | 0.555 | 0.910 | 0.925 | 0.921 | 0.457 |
| mtkw-dtkw      | 0.867 | 0.875 | 0.610 | 0.506 | 0.911 | 0.916 | 0.914 | 0.607 |
| mtkw-mtkwnsw_  | 0.871 | 0.907 | 0.912 | 1.007 | 0.913 | 0.941 | 0.964 | 0.993 |
| mtkw-mtams_    | 0.863 | 0.899 | 0.901 | 1.008 | 0.912 | 0.941 | 0.964 | 0.993 |
| mtkw-mtams_    | 0.823 | 0.773 | 0.753 | 0.743 | 0.889 | 0.861 | 0.876 | 0.790 |
| mtkw-mtams_    | 0.838 | 0.768 | 0.774 | 0.897 | 0.901 | 0.867 | 0.871 | 0.856 |
| mtkw-dtkw      | 0.821 | 0.829 | 0.598 | 0.598 | 0.915 | 0.917 | 0.908 | 0.632 |
| mtkw-dtkwnsw   | 0.896 | 0.901 | 0.687 | 0.518 | 0.940 | 0.941 | 0.942 | 0.625 |
| mtkw-mtkwnsw_  | 0.851 | 0.860 | 0.765 | 0.752 | 0.920 | 0.929 | 0.935 | 0.696 |
| mtkw-dtkwnsw_  | 0.929 | 0.930 | 0.935 | 0.992 | 0.951 | 0.951 | 0.993 | 0.989 |
| mtkw-dtkwnsw_  | 0.822 | 0.829 | 0.641 | 0.623 | 0.917 | 0.919 | 0.909 | 0.625 |
| mtkw-dtkwnsw_  | 0.877 | 0.882 | 0.679 | 0.531 | 0.933 | 0.934 | 0.939 | 0.628 |
| mtkw-dtkwnsw_  | 0.845 | 0.853 | 0.770 | 0.752 | 0.922 | 0.931 | 0.936 | 0.693 |
| mtkw-dtkwnsw_  | 0.914 | 0.914 | 0.929 | 0.994 | 0.945 | 0.944 | 0.991 | 0.989 |
| mtkw-mtkwnsw_  | 0.940 | 0.968 | 0.754 | 0.633 | 0.972 | 0.990 | 0.959 | 0.567 |

Textual differentiation in interaction networks
|        | CPP                      | LAD                      | LAU                      | ELE                      |
|--------|--------------------------|--------------------------|--------------------------|--------------------------|
|        | g. | p. | i. | h. | g. | p. | i. | h. | g. | p. | i. | h. | g. | p. | i. | h. |
| mtkwnsw- | 0.985 | 0.985 | 0.986 | 0.994 | 0.998 | 0.998 | 1.001 | 1.004 | 0.997 | 0.997 | 1.005 | 1.003 | 0.999 | 1.000 | 1.025 | 1.014 |
| mtkwnsw- | 0.930 | 0.957 | 0.734 | 0.620 | 0.970 | 0.988 | 0.958 | 0.565 | 0.965 | 0.982 | 0.980 | 0.603 | 0.911 | 0.971 | 0.761 | 0.539 |
| mtkwmsw- | 0.959 | 0.968 | 0.661 | 0.459 | 0.979 | 0.985 | 0.939 | 0.579 | 0.973 | 0.981 | 0.941 | 0.520 | 0.943 | 0.967 | 0.678 | 0.497 |
| dtkwsw-  | 0.988 | 0.988 | 0.993 | 1.007 | 0.994 | 0.994 | 1.001 | 1.005 | 0.992 | 0.992 | 1.003 | 1.002 | 0.997 | 0.999 | 1.019 | 1.012 |
| dtkwsw-  | 0.951 | 0.960 | 0.645 | 0.441 | 0.973 | 0.978 | 0.937 | 0.573 | 0.966 | 0.973 | 0.942 | 0.533 | 0.943 | 0.963 | 0.695 | 0.527 |
| mtkwsw-  | 0.938 | 0.965 | 0.772 | 0.630 | 0.969 | 0.987 | 0.955 | 0.562 | 0.966 | 0.983 | 0.978 | 0.606 | 0.891 | 0.959 | 0.768 | 0.533 |
| mtkwsw-  | 0.993 | 0.992 | 1.000 | 1.018 | 0.998 | 0.999 | 1.004 | 1.007 | 0.998 | 0.998 | 1.006 | 1.007 | 1.001 | 1.001 | 1.027 | 1.024 |
| mtkwmsw- | 0.946 | 0.954 | 0.648 | 0.481 | 0.973 | 0.978 | 0.935 | 0.579 | 0.965 | 0.973 | 0.938 | 0.514 | 0.940 | 0.963 | 0.666 | 0.490 |
| mtkwmsw- | 0.990 | 0.990 | 0.991 | 1.010 | 0.995 | 0.995 | 1.002 | 1.000 | 0.994 | 0.993 | 1.005 | 1.004 | 0.999 | 1.000 | 1.018 | 1.018 |
| mtmmsw-  | 0.945 | 0.973 | 0.764 | 0.620 | 0.971 | 0.990 | 0.956 | 0.562 | 0.968 | 0.986 | 0.979 | 0.602 | 0.903 | 0.969 | 0.763 | 0.542 |
| dtmmsw-  | 0.958 | 0.967 | 0.650 | 0.466 | 0.979 | 0.984 | 0.936 | 0.579 | 0.972 | 0.980 | 0.941 | 0.530 | 0.949 | 0.968 | 0.690 | 0.516 |
| mtsT-mtsT | 0.885 | 0.885 | 0.840 | 0.494 | 0.957 | 0.957 | 0.980 | 0.894 | 0.967 | 0.965 | 0.997 | 0.826 | 0.989 | 0.990 | 0.904 | 0.920 |
| mtsT-mtsT | 0.901 | 0.885 | 0.904 | 0.808 | 0.952 | 0.952 | 0.942 | 0.783 | 0.961 | 0.959 | 0.980 | 0.825 | 0.906 | 0.967 | 0.796 | 0.740 |
| mtsT-mtsT | 0.820 | 0.746 | 0.871 | 0.970 | 0.848 | 0.784 | 0.836 | 0.932 | 0.841 | 0.779 | 0.914 | 0.938 | 0.930 | 0.855 | 0.943 | 0.968 |
| mtsT-mtsT | 0.977 | 0.977 | 0.979 | 1.009 | 0.981 | 0.982 | 0.989 | 0.990 | 0.871 | 0.873 | 0.987 | 1.002 | 0.970 | 0.972 | 0.985 | 0.988 |
| mtsT-mtsT | 0.979 | 0.980 | 0.976 | 1.010 | 0.956 | 0.957 | 0.956 | 0.990 | 0.889 | 0.905 | 0.926 | 0.992 | 0.962 | 0.949 | 1.010 | 1.019 |
| mtsT-mtsT | 0.962 | 0.962 | 0.967 | 0.953 | 0.968 | 0.969 | 0.980 | 0.966 | 0.961 | 0.961 | 0.965 | 0.998 | 0.974 | 0.976 | 0.984 | 0.981 |
| mtsT-mtsT | 0.969 | 0.967 | 0.981 | 1.003 | 0.973 | 0.975 | 0.959 | 0.925 | 0.948 | 0.945 | 0.976 | 0.955 | 0.956 | 0.966 | 0.952 | 0.973 |
| mtsT-mtsT | 0.962 | 0.957 | 0.996 | 1.001 | 0.991 | 0.991 | 0.997 | 1.002 | 0.877 | 0.872 | 0.995 | 1.000 | 0.995 | 0.996 | 1.016 | 1.023 |
| mtsT-mtsT | 0.989 | 0.976 | 0.997 | 1.015 | 0.982 | 0.980 | 0.984 | 0.994 | 0.874 | 0.863 | 0.960 | 0.996 | 0.992 | 0.997 | 1.007 | 0.994 |
| mtsT-mtsT | 0.804 | 0.809 | 0.792 | 0.648 | 0.852 | 0.852 | 0.889 | 0.823 | 0.841 | 0.837 | 0.925 | 0.854 | 0.904 | 0.909 | 0.802 | 0.906 |
| mtsT-mtsT | 0.994 | 0.995 | 1.004 | 1.016 | 0.997 | 0.997 | 1.001 | 1.006 | 0.997 | 0.997 | 1.004 | 1.005 | 1.000 | 1.001 | 1.026 | 1.023 |
| mtsT-mtsT | 0.996 | 0.997 | 1.001 | 1.016 | 0.999 | 0.999 | 1.003 | 1.000 | 0.999 | 0.999 | 1.006 | 1.003 | 1.002 | 1.003 | 1.021 | 1.022 |

**TABLE XVII:** Correlation of textual measures. See subsection IV I for discussion and directions.
| -                           | CPP   | LAD   | LAU   | ELE   |
|-----------------------------|-------|-------|-------|-------|
| nc-d<sub>i</sub>           | 0.929 | 0.377 | 0.434 | 0.942 |
| nc-s                        | 0.951 | 0.441 | 0.359 | 0.962 |
| nc-s<sub>i</sub>            | 0.946 | 0.258 | 0.207 | 0.961 |
| nc-s<sub>b</sub>            | 0.951 | 0.458 | 0.448 | 0.957 |
| nc-tri                      | 0.935 | 0.312 | 0.276 | 0.941 |
| nt-d                        | 0.926 | 0.348 | 0.244 | 0.925 |
| nt-d<sub>i</sub>            | 0.919 | 0.205 | 0.144 | 0.912 |
| nt-d<sub>b</sub>            | 0.926 | 0.369 | 0.320 | 0.938 |
| nt-s                        | 0.946 | 0.424 | 0.335 | 0.956 |
| nt-s<sub>i</sub>            | 0.941 | 0.240 | 0.195 | 0.956 |
| nt-s<sub>b</sub>            | 0.945 | 0.447 | 0.415 | 0.950 |
| nt-bc                       | 0.865 | 0.247 | 0.085 | 0.845 |
| nt-tri                      | 0.933 | 0.295 | 0.268 | 0.938 |
| ntd-d                       | 0.905 | 0.430 | 0.402 | 0.903 |
| ntd-d<sub>i</sub>           | 0.882 | 0.267 | 0.292 | 0.892 |
| ntd-d<sub>b</sub>           | 0.925 | 0.443 | 0.468 | 0.912 |
| ntd-s<sub>i</sub>           | 0.851 | 0.527 | 0.537 | 0.919 |
| ntd-s<sub>b</sub>           | 0.833 | 0.322 | 0.372 | 0.914 |
| ntd-s<sub>b</sub>           | 0.867 | 0.536 | 0.612 | 0.919 |
| ntd-bc                      | 0.811 | 0.243 | 0.195 | 0.819 |
| ntd-tri<sub>i</sub>         | 0.923 | 0.363 | 0.427 | 0.930 |
| ntd-tri<sub>b</sub>         | 0.523 | 0.036 | 0.019 | 0.451 |
| ntd-sector                  | 0.686 | 0.000 | 0.000 | 0.000 |
| ntd/nt-sector               | -0.547 | 0.000 | 0.000 | 0.000 |
| mntsw2-sector               | 0.555 | 0.000 | 0.000 | 0.000 |

**TABLE XVIII:** Correlation of textual and topological measures. See subsection [IV](#) for discussion and directions.
### TABLE XIX. Composition of first component (threshold: $|val| > 0.05$). See subsection IV.J for discussion and directions.

|    | CPP | LAD | LAU | ELE |
|----|-----|-----|-----|-----|
| **$\lambda$** | 17.71 | 18.46 | 19.44 | 30.20 |
| mtkwsw_ | 0.09 | -1.13 | -2.89 | 0.94 |
| mts_ | -0.34 | 1.56 | 0.86 | -0.09 |
| mtsT | 0.35 | -1.20 | -2.17 | 3.16 |
| dtswT | -0.45 | 1.20 | -1.49 | 0.63 |
| mtsTskw | -0.11 | -2.19 | 0.59 | -2.85 |
| dtsT | 0.77 | 8.15 | -3.14 | 0.39 |
| dttT | -0.43 | 2.51 | 2.88 | 1.45 |
| ntsSwT | -0.45 | -2.04 | 1.25 | -1.22 |
| dtswT | -3.39 | 1.44 | -1.29 | 0.49 |
| NN | -2.70 | 0.60 | -4.54 | -0.45 |
| JFR | 0.54 | 3.13 | -0.51 | -0.51 |
| JJS | -8.43 | 0.16 | -3.34 | 0.59 |
| RB | 4.07 | 1.34 | 2.53 | 0.59 |
| RBS | 0.49 | -1.54 | -0.87 | 0.66 |
| VBD | 1.48 | 0.09 | -0.80 | -1.11 |
| VBG | 0.19 | -0.68 | 1.10 | -1.29 |
| IN | -1.74 | 0.32 | -0.56 | -1.24 |
| PRPS | 6.51 | 0.22 | 0.29 | 0.41 |
| PDT | 5.12 | 0.11 | 0.24 | 0.41 |

### TABLE XX. Composition of second component (threshold: $|val| > 0.05$). See subsection IV.J for discussion and directions.

|    | CPP | LAD | LAU | ELE |
|----|-----|-----|-----|-----|
| **$\lambda$** | 11.48 | 10.25 | 12.23 | 15.72 |
| nc | -5.21 | -3.60 | 3.06 | -4.12 |
| Nwssw/Nws | 0.43 | 0.34 | 0.35 | -0.66 |
| dtswT | -5.57 | 0.29 | 1.79 | -1.41 |
| JFR | 0.22 | -4.19 | 0.73 | 1.98 |
| RB | -0.68 | -2.74 | 0.90 | 0.22 |
| IN | 0.46 | 0.98 | -0.35 | -0.54 |
| WP$|$ | -0.15 | 1.52 | 2.10 | 0.00 |
| CD | -5.85 | 0.15 | 1.02 | 0.00 |
| mtaH | 0.51 | -0.93 | -3.67 | 0.00 |
| dtamH | 0.10 | -0.35 | -1.14 | 0.21 |
| mprof | -3.63 | 2.72 | -1.61 | 0.21 |
| dprof | -0.73 | 1.16 | -2.33 | 0.24 |
| $d_o$ | -0.02 | -0.01 | -4.02 | 1.06 |
| $s_o$ | -0.71 | 2.39 | -1.52 | 0.25 |
| bc | 0.70 | 1.55 | 0.11 | -0.18 |
| tri | -0.07 | 8.58 | 0.00 | -0.45 |
| in cent | 15.09 | -0.00 | 0.00 | 0.53 |
### TABLE XXI. Composition of third component (threshold: |val| > 0.05). See subsection IVJ for discussion and directions.

|       | CPP | LAD | LAU | ELE |
|-------|-----|-----|-----|-----|
|       | g.  | p.  | i.  | h.  |
| λ     | 3.81| 4.00| 6.75| 5.52|
| ncont | 2.30| 2.60| -2.75| -2.62|
| dtsw  | 0.58| -0.12| -0.84| -1.52|
| WRB   | -1.39| 1.62| -5.39| 0.00|
| WP    | -2.35| 1.63| -1.44| 0.00|
| WP$   | -1.35| 2.81| -4.42| 0.00|
| EX    | -7.04| 8.54| -0.88| 0.00|
| UW    | 0.99| -1.17| -7.02| 0.00|
| FW    | -1.35| 5.80| 1.68| 0.00|
| mlwss | 0.53| 1.41| -0.79| 0.00|
| dlwss | 2.83| 0.44| -1.10| 0.00|
| dprof | -1.24| 1.04| -0.39| -0.81|
| d     | 0.43| -0.91| -1.59| -0.81|
| s     | 0.07| -0.50| -0.46| 0.80|
| si    | -0.34| 0.18| -0.54| 1.27|
| bc    | -0.17| 0.06| 0.03| -0.71|
| tri   | 1.09| 0.11| 0.00| -0.31|
| cv    | -0.24| 0.24| 0.00| -0.25|

### TABLE XXII. Composition of fourth component (threshold: |val| > 0.05). See subsection IVJ for discussion and directions.

|       | CPP | LAD | LAU | ELE |
|-------|-----|-----|-----|-----|
|       | g.  | p.  | i.  | h.  |
| λ     | 8.97| 7.53| 7.71| 7.19|
| Nkw/nt| 1.81| 1.26| -1.38| 0.76|
| mtsw2.| 2.46| 1.66| -1.23| 3.90|
| mtsTS | 0.42| -0.97| -1.76| 2.08|
| dtsTSkw| 1.66| 2.31| 1.25| 0.93|
| mtsTSpv| 0.71| 6.83| 1.68| -0.24|
| dtsTSpv| 5.50| -2.29| -2.81| 0.93|
| ntmT  | -2.90| 0.60| -0.74| 1.75|
| dtmT  | 1.64| -0.43| 0.52| 0.21|
| dttmT | 0.90| -0.20| -2.54| 2.69|
| ntsmT | -0.33| -5.56| 4.20| 1.39|
| dtsmT | -1.96| 0.77| 2.26| -0.26|
| NN    | 2.78| 0.13| 0.15| 1.10|

The table entries represent the composition of the third and fourth components, respectively, with values indicating the magnitude of each variable's contribution to the component. Variables are sorted by their contribution, with positive values suggesting a positive relationship and negative values a negative one. Threshold values are applied to filter contributions, ensuring only significant effects are highlighted.
|        | CPP     | LAD     | LAU     | ELE     |
|--------|---------|---------|---------|---------|
|        | g.      | p.      | i.      | h.      | g.      | p.      | i.      | h.      | g.      | p.      | i.      | h.      | g.      | p.      | i.      | h.      |
| \( \lambda \) | 3.48    | 3.38    | 4.54    | 4.16    | 3.42    | 3.74    | 4.60    | 4.41    | 2.97    | 3.14    | 3.83    | 5.33    | 3.32    | 3.58    | 5.95    | 7.29    |
| Nkw/nsssw/Nkw | -6.10   | -3.15   | 0.31    | 0.94    | -2.48   | -1.57   | -0.20   | 0.43    | 1.12    | 1.63    | 0.33    | 1.23    | -2.41   | -0.97   | -1.30   | 1.91    |
| mtsTS    | -4.04   | -1.67   | -1.79   | 1.54    | -1.90   | 0.84    | -5.02   | -0.15   | -1.53   | -1.19   | 0.50    | 0.67    | 1.56    | -0.20   | 0.89    | 0.24    |
| dtsTS    | -0.97   | 3.59    | -1.47   | 0.09    | 5.70    | 0.45    | 2.78    | 0.45    | -3.19   | 6.97    | 1.40    | 0.12    | -1.89   | -1.59   | -0.20   | 0.30    |
| dtsTSkw  | 0.03    | -1.95   | 0.87    | -1.14   | -5.87   | -2.23   | 0.06    | -2.74   | -6.87   | -4.49   | -0.07   | 1.36    | 2.53    | -0.85   | -0.44   | -0.20   |
| dtsTSpv  | 1.87    | -0.41   | 0.82    | 3.16    | 2.28    | -5.32   | -1.16   | 1.35    | 4.05    | -4.68   | 3.13    | -0.67   | 3.68    | 3.04    | -0.58   | -0.84   |
| sector   | 0.00    | 0.00    | 0.00    | 0.48    | 0.00    | 0.00    | 0.00    | 0.00    | 0.00    | 0.00    | 0.00    | 0.00    | -0.00   | -0.00   | 11.05   | 1.52    |

TABLE XXIII. Composition of fifth component (threshold: \(|val| > 0.05\)). See subsection IV J for discussion and directions.
Appendix C: Histograms of existent and incident words

See subsection IV. K and Figures 2-6 for discussion and directions.
FIG. 2. Size of words that are known in English. Crossing of incident and existential sizes is around 5 (Figure 3 shows a shift to length 6-7 when consider only non stopwords). Words with three letters have maximum incidence, while most words have 7 letters. See subsection IV K for discussion and directions.
FIG. 3. Size of words that are known in English and are not stopwords. Crossing of incident and existential sizes is around 6-7 (figure 2 shows a shift to length 5 when considered stopwords). In this case, words with 4 letters have maximum incidence, while most words still have 7 letters. Exception for ELE, which exhibits maximum incidence of words with 5 letters and most words having 8 letters, which might be associated with ELE network typology discussed in tables III and IVK. See subsection IVK for discussion and directions.
FIG. 4. Size of words that are known, are not stopwords and have synsets. Resembles figure 3. Stopword sizes histogram are in figure 5. Differences suggests $\approx 0.5$ might be constant. LAD and LAU exquisite vocabulary (GNU/Linux, programming, sound/signal processing, music) might be responsible for higher difference of distributions. See subsection IV K for discussion and directions.

FIG. 5. Size histogram of stopwords. Stopwords with two letters are the most frequent, while most of them have four letters. Differences in distribution seem stable around $\approx 0.6$. See subsection IV K for discussion and directions.
FIG. 6. Size histogram of known English words that are not stopwords and do not return synsets. Differences in distribution suggests less stable behavior, with high incidence of few words high number of existing words with many letters. Observe difference $\geq 1$, as observed only with all known words, but even higher. See subsection IV K for discussion and directions.
Appendix D: Online scripts, data and writing

All data can be accessed in the GMANE database, which consists of some tenths of thousands of mailing lists, with constant updates. All scripts used on this article are in http://sourceforge.net/p/labmacambira/fimDoMundo/ci/master/tree/python/toolkitGMANE/. The git repository of this article itself is https://github.com/ttm/artigoTextoNasRedes.git.

1 gmane.comp.gcc.libstdc++.-devel is list ID in GMANE archive.
2 gmane.linux.audio.devel is list ID in GMANE archive.
3 gmane.linux.audio.users is list ID in GMANE archive.
4 gmane.politics.election-methods is list ID in GMANE archive.
5 Ontologia de participação social. http://tinyurl.com/p2doueu
6 Produto 5 da consultoria PNUD/ONU de Renato Fabbri. https://github.com/ttm/pnud4/blob/master/latex/produto.pdf?raw=true
7 D. C. Antunes, A. A. Zui, et al. Do bullying ao preconceito: os desafios da barbarie à educação. Psicologia & Sociedade, 20(1):33–42, 2008.
8 S. Bird, E. Klein, and E. Loper. Natural language processing with Python. " O’Reilly Media, Inc.", 2009.
9 R. P. Cavalcante. Influência dos movimentos sociais na formação da agenda política: as jornadas de junho de 2013 no brasil. 2014.
10 D. Easley and J. Kleinberg. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press, 2010.
11 R. F. et al. Stability in human interaction networks: primitive typology of vertex, prominence of measures and activity statistics. May 2014. http://sourceforge.net/p/labmacambira/fimDoMundo/ci/master/tree/textos/evolutionSN/paper.pdf?format=raw
12 R. Fabbri. Ensaio sobre o auto-aproveitamento: um relato de investidas naturais na participação social. http://arxiv.org/abs/1412.6868
13 R. Fabbri. Incidência de letras na obra de machado de assis, 2012. http://sourceforge.net/p/labmacambira/rcpln/ci/master/tree/pln/trabLetras/resumoLetras.pdf?format=raw
14 K. Marek-Spartz, P. Chesley, and H. Sande. Construction of the gmane corpus for examining the diffusion of lexical innovations. 2012.