Stylometry in plagiarism detection and author profiling

Paolo Rosso

PRHLT Research Center

Universitat Politècnica de València

http://www.dsic.upv.es/~prosso/

Tehran, 25/01/2017

Outline

- Plagiarism
- Intrinsic plagiarism detection
- Author profiling

Plagiarism

- Verbatim
- Paraphrasing
- Ideas
- Cross-language

- Source code

Plagiarism detection

- External : external evidence
- Intrinsic: intrinsic evidence (style analysis)

- Cross-language: translated plagiarism

Intrinsic plagiarism detection

- Insertion of text from a different author into a document causes style and complexity irregularities

Stylometry: Intrinsic plagiarism detection

- The study of linguistic style applied to written language
- Quantifying writing style irregularities:

Text readability: Gunning fog, Flesch–Kincaid, …

Vocabulary richness: types/tokens ratio

Basic statistics: avg. sentence length, avg. word length, word avg. word classes

n-grams profiles statistics: character level statistics

Gunning fog index

IG = 0.4 (|words|/|sentences|+
100*(|complex_words|/|words|))

Complext words: words with three or more syllables

IG(comics) = 6

IG(Newsweek) = 10

An example

In this work, we have carried out some research on the influence that mineral salts on the mood of people. For this research I have worked with 5 people who have taken water with different amount of mineral salts. Our theory is that the more minerals are in the water, the more moody people are. […]

Mineral salts are inorganic molecules of easy ionization in presence of water in living beings they appear by precipitation as well as dissolved mineral salts. […] Dissolved mineral salts are always ionized. These salts have structural function and pH regulating functions, of the osmotic pressure and of biochemical reactions, in which specific ions are involved.

It seems to me that the results are good. […]

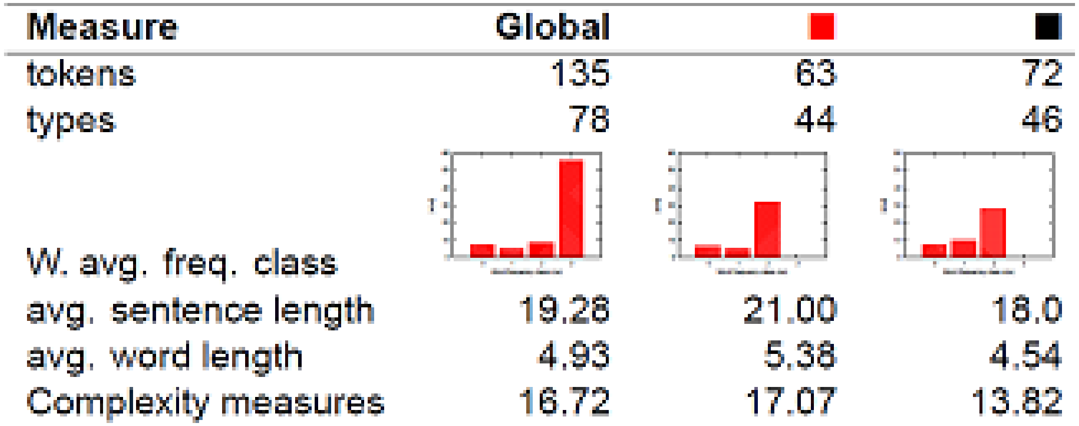
An example

In this work, we have carried out some research on the influence that mineral salts on the mood of people. For this research I have worked with 5 people who have taken water with different amount of mineral salts. Our theory is that the more minerals are in the water, the more moody people are. […]

Mineral salts are inorganic molecules of easy ionization in presence of water in living beings they appear by precipitation as well as dissolved mineral salts. […] Dissolved mineral salts are always ionized. These salts have structural function and pH regulating functions, of the osmotic pressure and of biochemical reactions, in which specific ions are involved.

It seems to me that the results are good. […]

An example



Measure	Global	🟥	⬛
tokens	135	63	72
types	78	44	46
W. avg. freq. class			
avg. sentence length	19.28	21.00	18.0
avg. word length	4.93	5.38	4.54
Complexity measures	16.72	17.07	13.82

Intrinsic plagiarism detection @ PAN

- char n-grams (Stamatatos)
- word freq. class + text frequencies (Zechner et al.)
  (Mahgoub et al. @ AraPlagDet)
- Kolmogorov complexity measure (Seaward & Matwin)

  …

Intrinsic plagiarism detection @ PAN

- char n-grams (Stamatatos)
- word freq. class + text frequencies (Zechner et al.) (Mahgoub et al. @ AraPlagDet)
- Kolmogorov complexity measure (Seaward & Matwin)

  …

char n-gram classes based on frequency of n-grams (Bensaleme et al., EMNLP 2015)

Outline

- Plagiarism
- Intrinsic plagiarism detection
- Author profiling

Gender: which is female/male?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

[examples: Moshe Koppel]

British National Corpus

- 920 documents labelled for
  - author gender
  - document genre

- Used 566 controlled for genre

	Male	Fem
Fiction (prose)	132	132
Non-fiction	151	151
Arts (general)	8	8
Arts (acad.)	12	12
Belief/Thought	12	12
Biography	27	27
Commerce	5	5
Leisure	8	8
Science (gen.)	13	13
Soc. Sci. (gen.)	26	26
Soc. Sci. (acad.)	19	19
World Affairs	21	21

M. Koppel, S. Argamon, and A. R. Shimoni. Automatically categorizing written texts by author gender. Literary and linguistic computing 17(4), 2002.

Distinguishing features: male vs. female style

Males use more

- Determiners
- Adjectives
- *of* modifiers (e.g. *pot of gold*)

Informational features

Females use more

- Pronouns *
- *for* and *with*
- Negation
- Present tense

Involvedness features

J. W. Pennebaker. The Secret Life of Pronouns: What Our Words Say about Us. Bloomsbury USA, 2013.

Gender: which is female/male?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Gender: which is female/male?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Gender: which is female/male?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Gender: which is female/male?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Gender: which is Female/Male?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

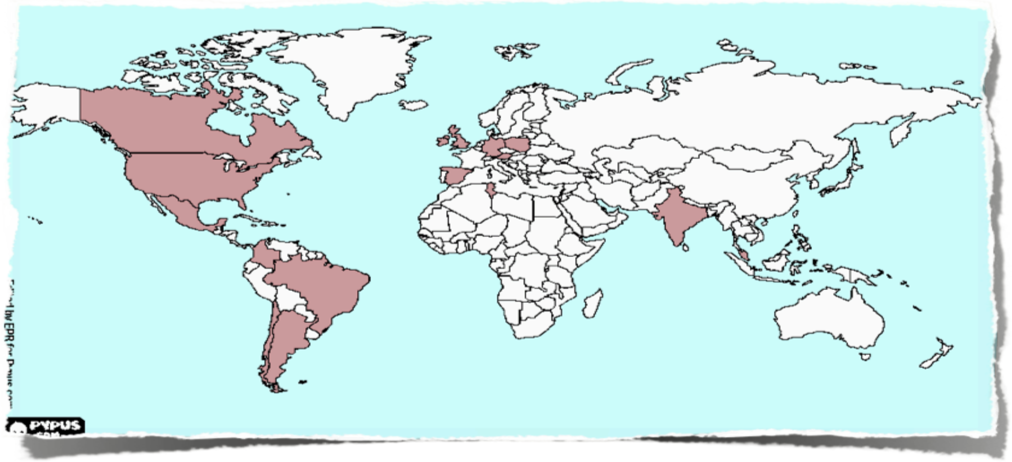
The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Gender & age identification

AUTHOR	COLLECTION	FEATURES	RESULTS	OTHER CHARACTERISTICS
Argamon et al., 2002	British National Corpus	Part-of-speech	Gender: 80% accuracy	
Koppel et al., 2003	Blogs	Lexical and syntactic features	Gender: 80% accuracy	Self-labeling
Schler et al., 2006	Blogs	Stylistic features + content words with the highest information gain	Gender: 80% accuracy Age: 75% accuracy	
Goswami et al., 2009	Blogs	Slang + sentence length	Gender: 89.18 accuracy Age: 80.32 accuracy	
Zhang & Zhang, 2010	Segments of blog	Words, punctuation, average words/sentence length, POS, word factor analysis	Gender: 72.10 accuracy	
Nguyen et al., 2011 y 2013	Blogs & Twitter	Unigrams, POS, LIWC	Correlation: 0.74 Mean absolute error: 4.1 - 6.8 years	Manual labeling Age as continuous variable
Peersman et al., 2011	Netlog	Unigrams, bigrams, trigrams and tetagrams	Gender+Age: 88.8 accuracy	Self-labeling, min 16 plus 16,18,25

Author profiling: PAN @CLEF 2013

- Teams submitting results: 21 (Registered teams: 64)



- (Towards) **big data**: 400,000 social media texts

including **chat lines of potential pedophiles** (task in 2012)

- **Age classes**: 10s (13-17), 20s (23-27), 30s (33-48)
- **Languages**: English and Spanish

http://pan.webis.de/

Approaches: Features

- Stylistic features: frequency of punctuation marks, capital letters,…
- Part of Speech
- Readability measures
- Dictionary-based words, topic-based words
- Collocations
- Character or word n-grams
- Slang words, character flooding
- Emoticons
- Emotion words

F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. Overview of the Author Profiling Task at PAN 2013 - Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings Vol. 1179. 2013.

Author Profiling @ PAN-14 : Features

- Similar features of AP@PAN-13:

  content (bag of words, word n-grams) and stylistic features

- frequency of words related to different psycholinguistic concepts, extracted from: LIWC and MRC psycholinguistic database

F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkman, B. Stein, B. Verhoeven, and W. Daelemans. Overview of the 2nd Author Profiling Task at PAN 2014—Notebook for PAN at CLEF 2014. CEUR Workshop Proceedings Vol. 1180, pp. 898-927, 2014.

Stylometry: Author profiling

- **Term frequency (F)**: terms with character flooding; terms starting with capital letter; terms in capital letters…
- **Punctuation marks (P)**: frequency of use of dots, commas, colon, semicolon, exclamations and question marks
- **Part-Of-Speech**: frequency of use of each grammatical category
- **Emoticons (E)**: number of different types of emoticons representing emotions
- **Spanish Emotion Lexicon (SEL)**: terms co-occurring with the six basic Ekman's emotions: happiness, anger, fear, sadness, disgust, surprise

EmoGraph

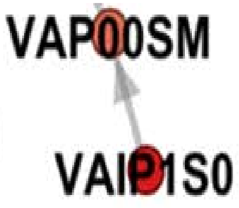
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.

EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

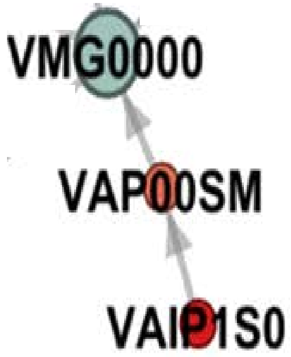
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

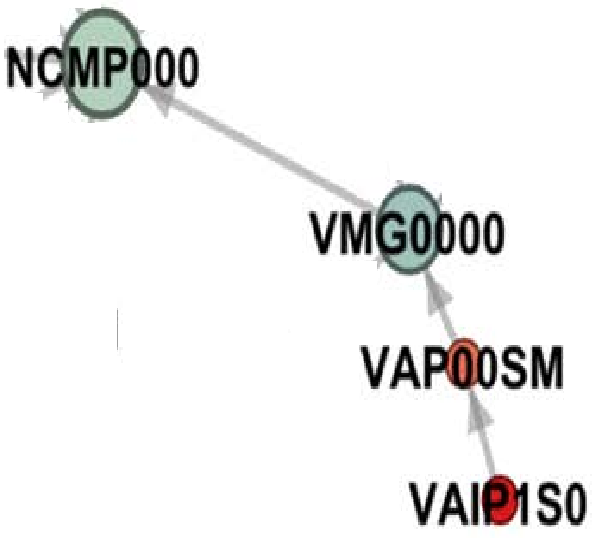
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

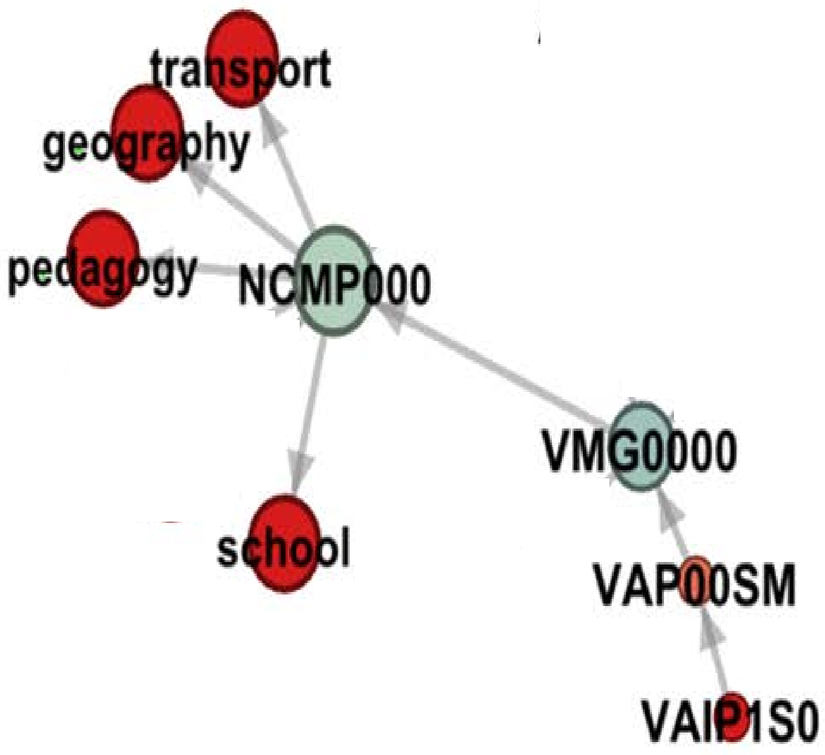
(I) have been taking online **courses** about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

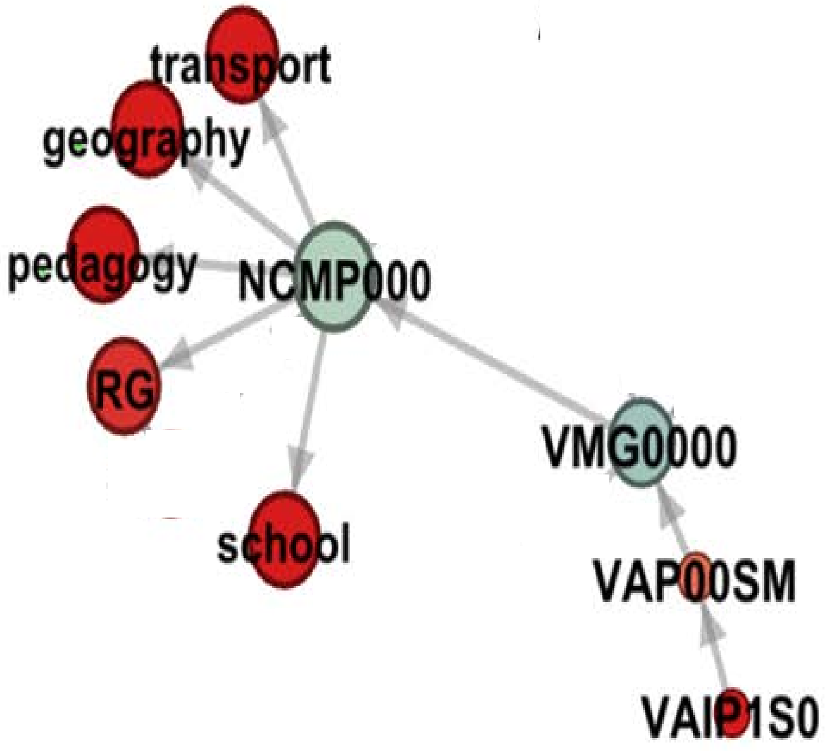
(I) have been taking online **courses** about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

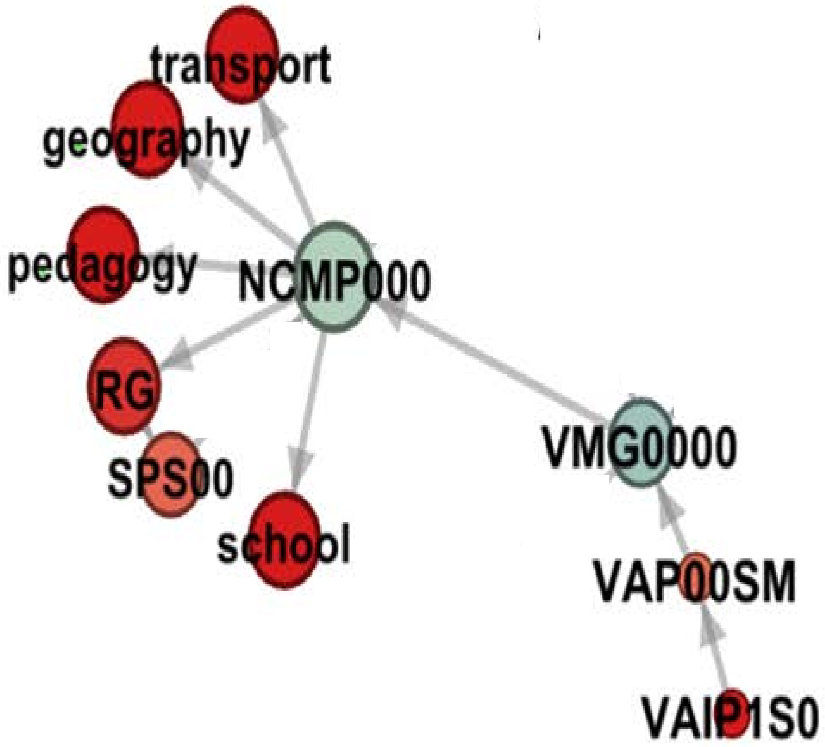
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

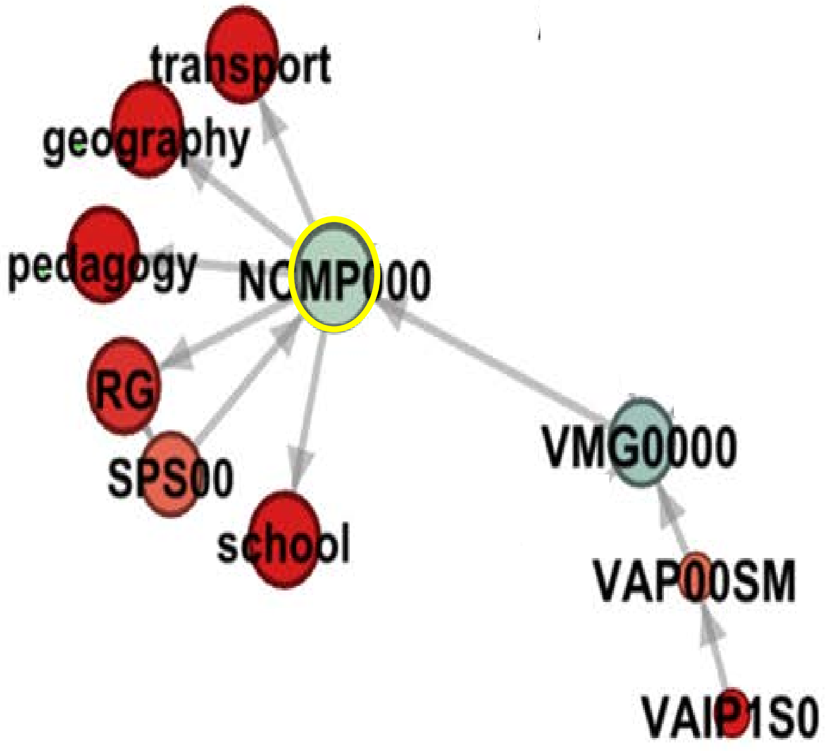
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

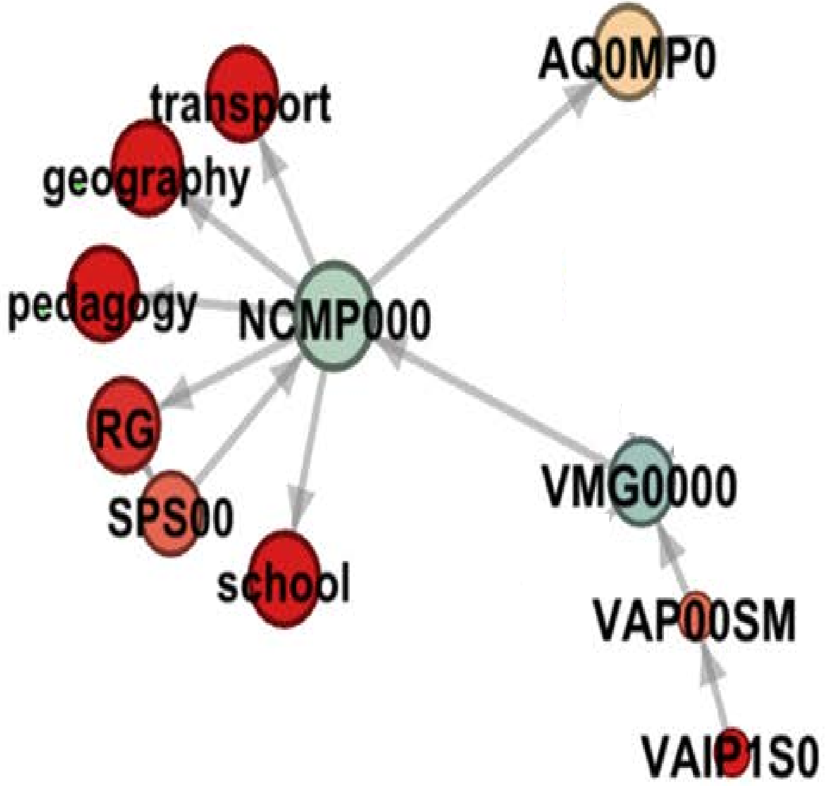
(I) have been taking online courses about valuable **subjects** that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

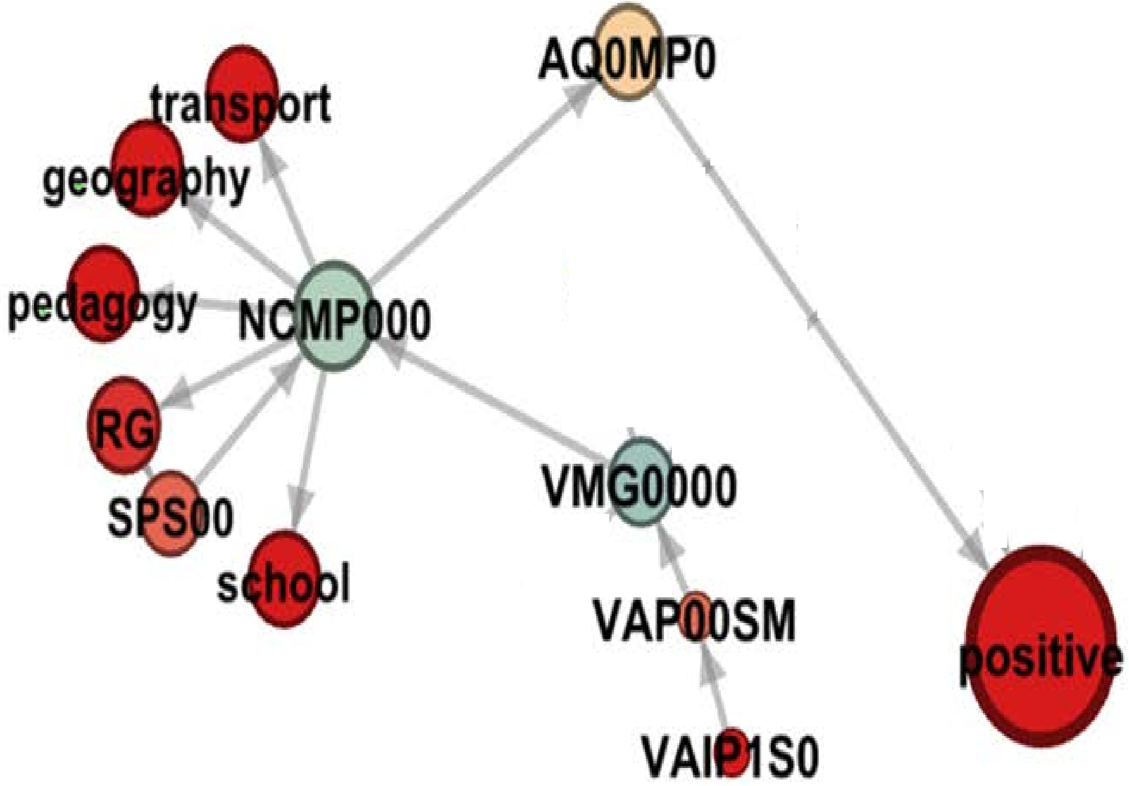
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

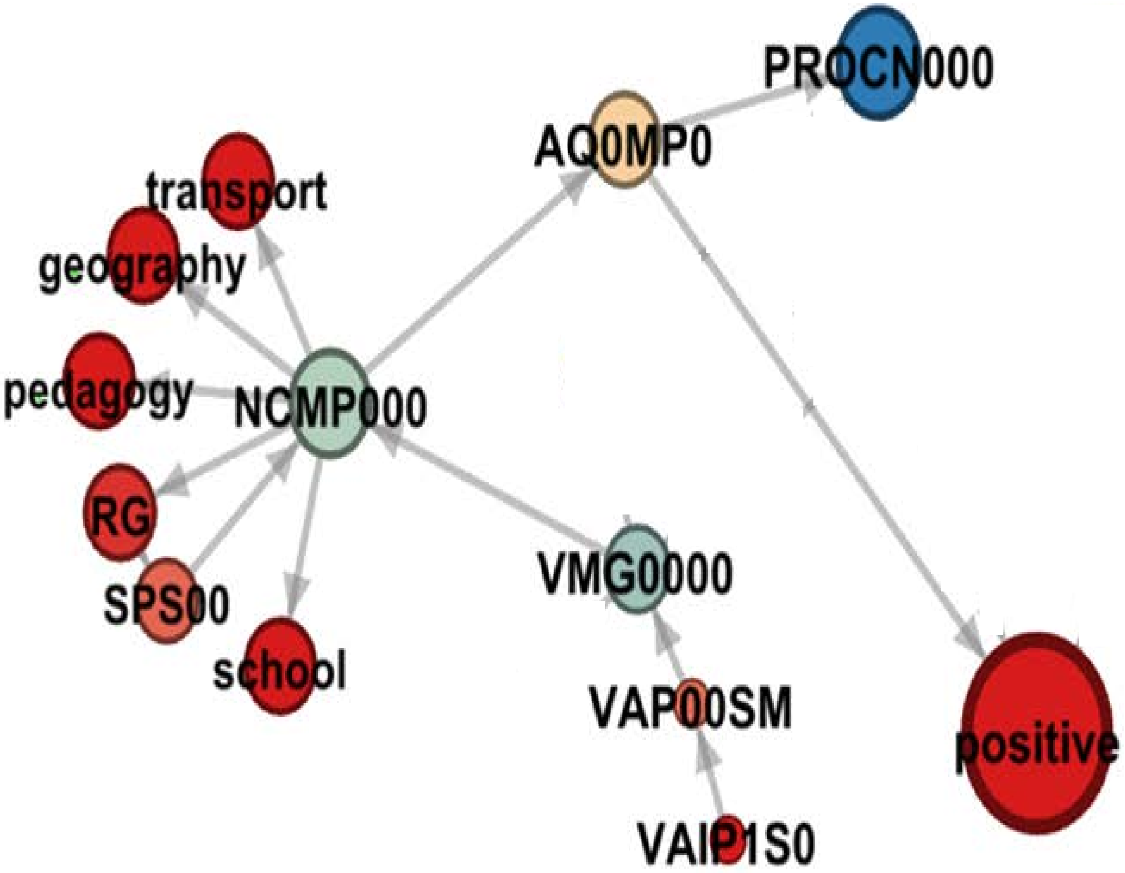
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

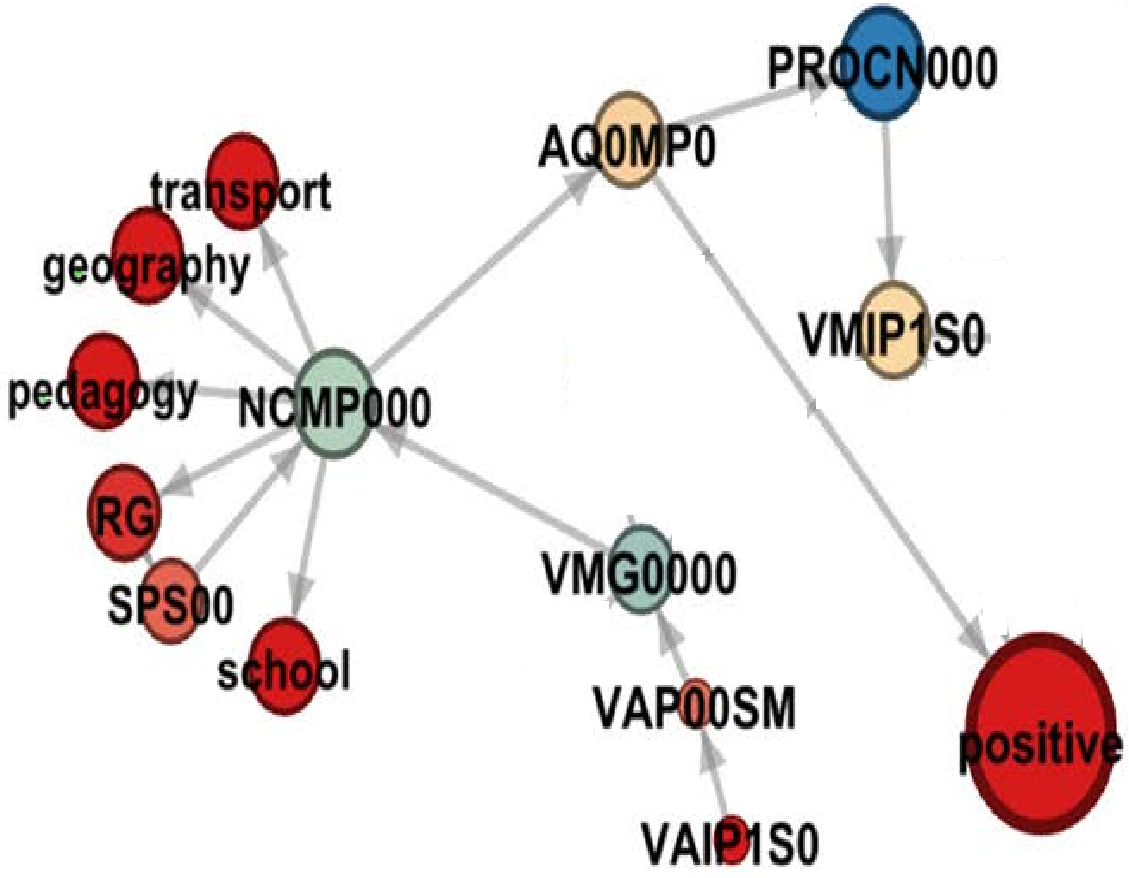
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

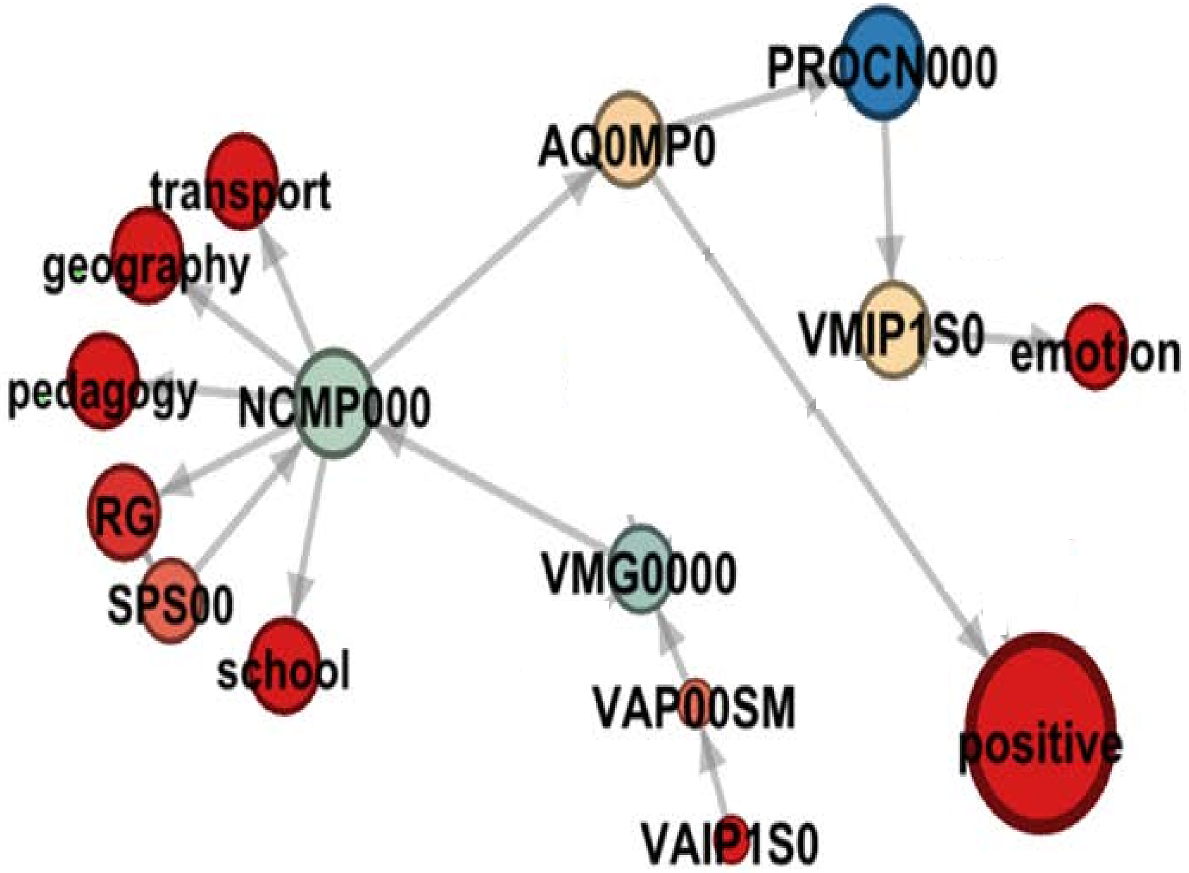
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

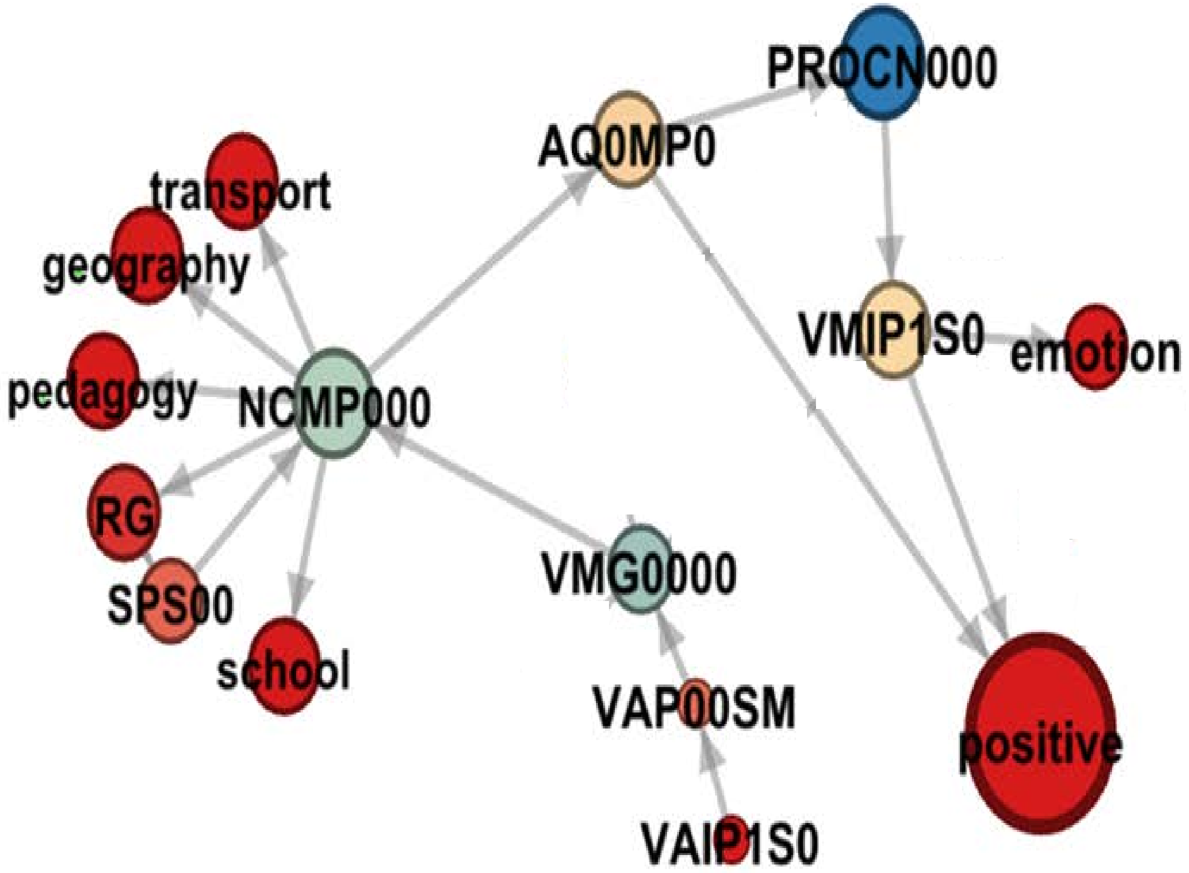
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

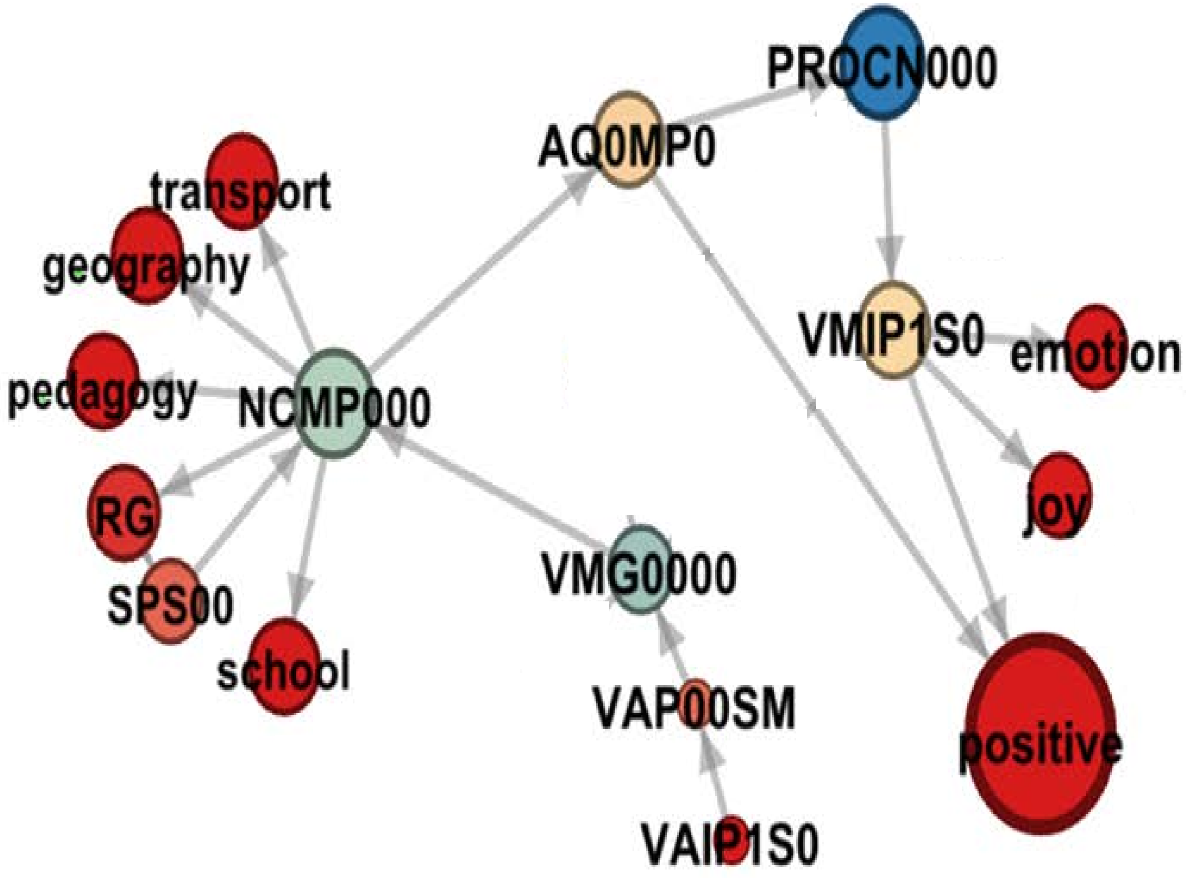
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

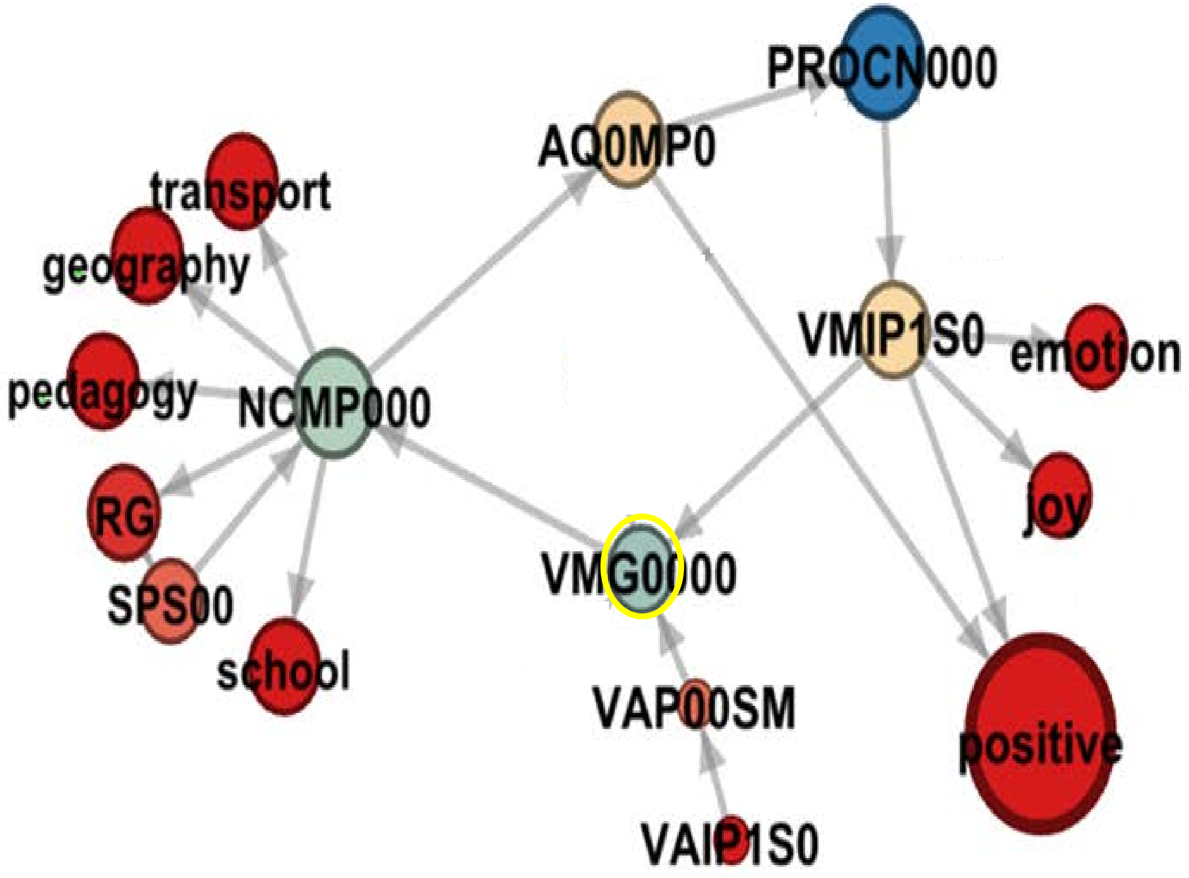
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

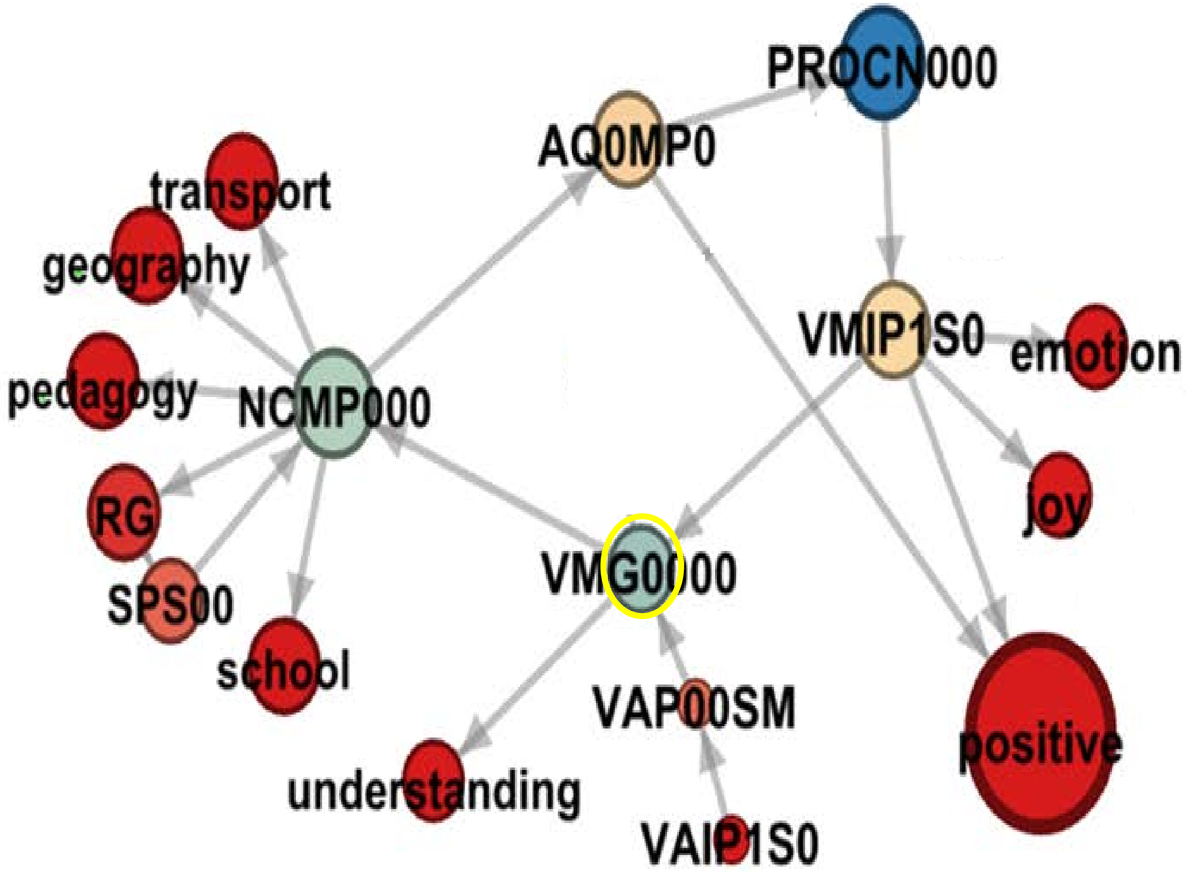
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

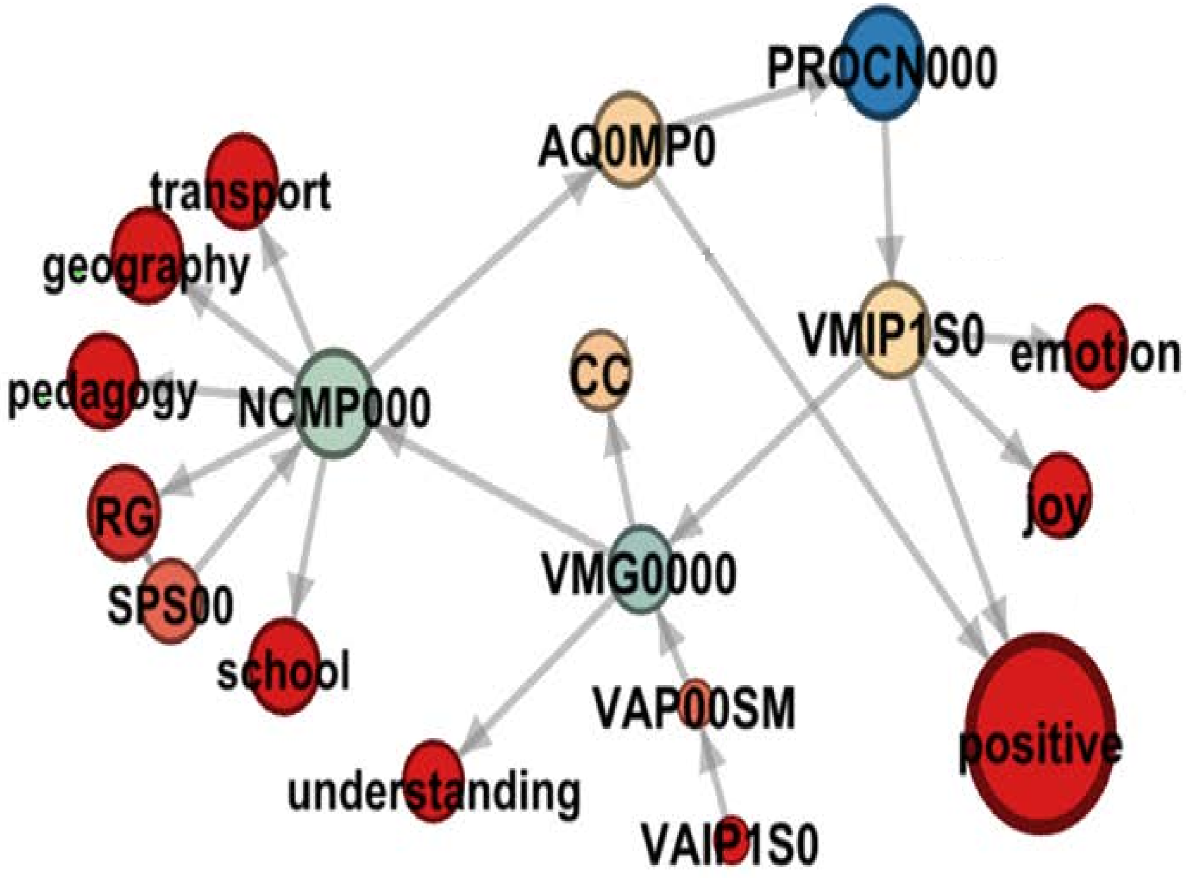
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

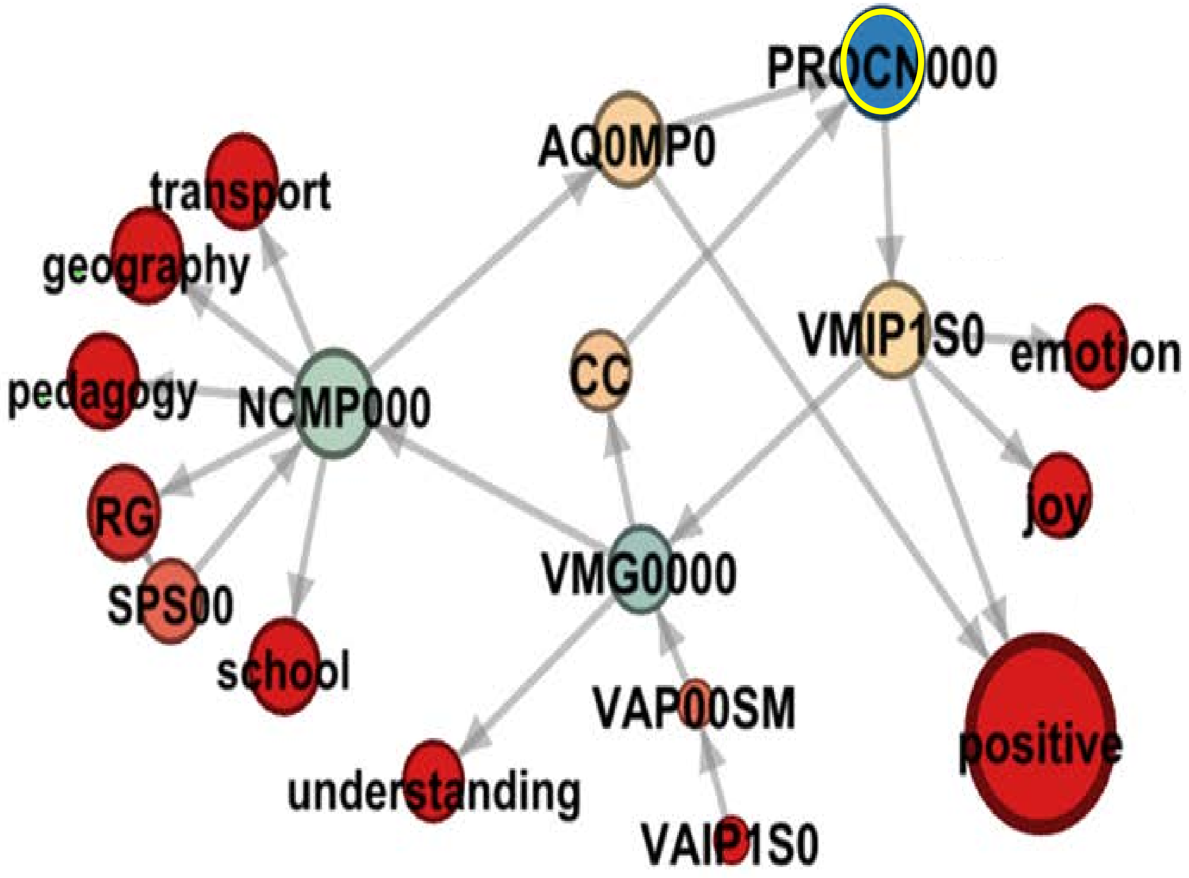
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

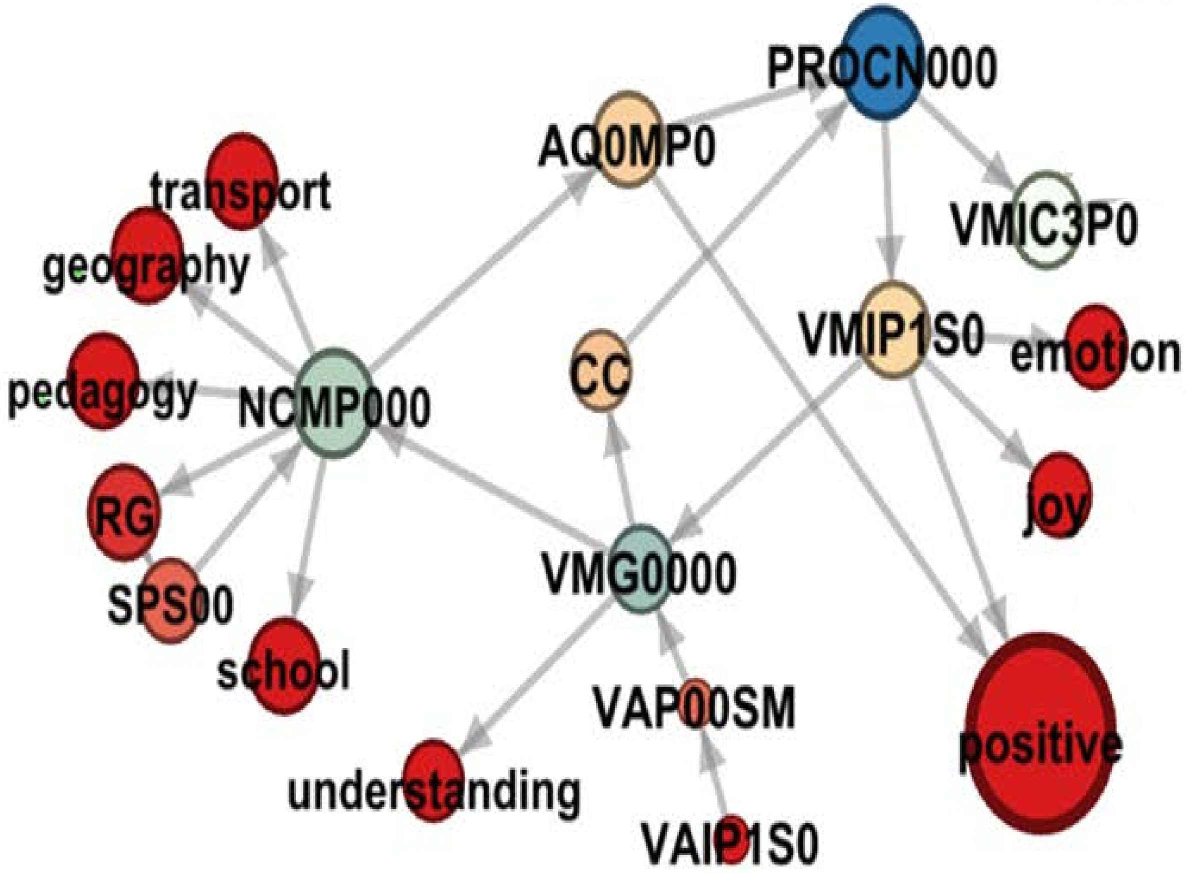
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

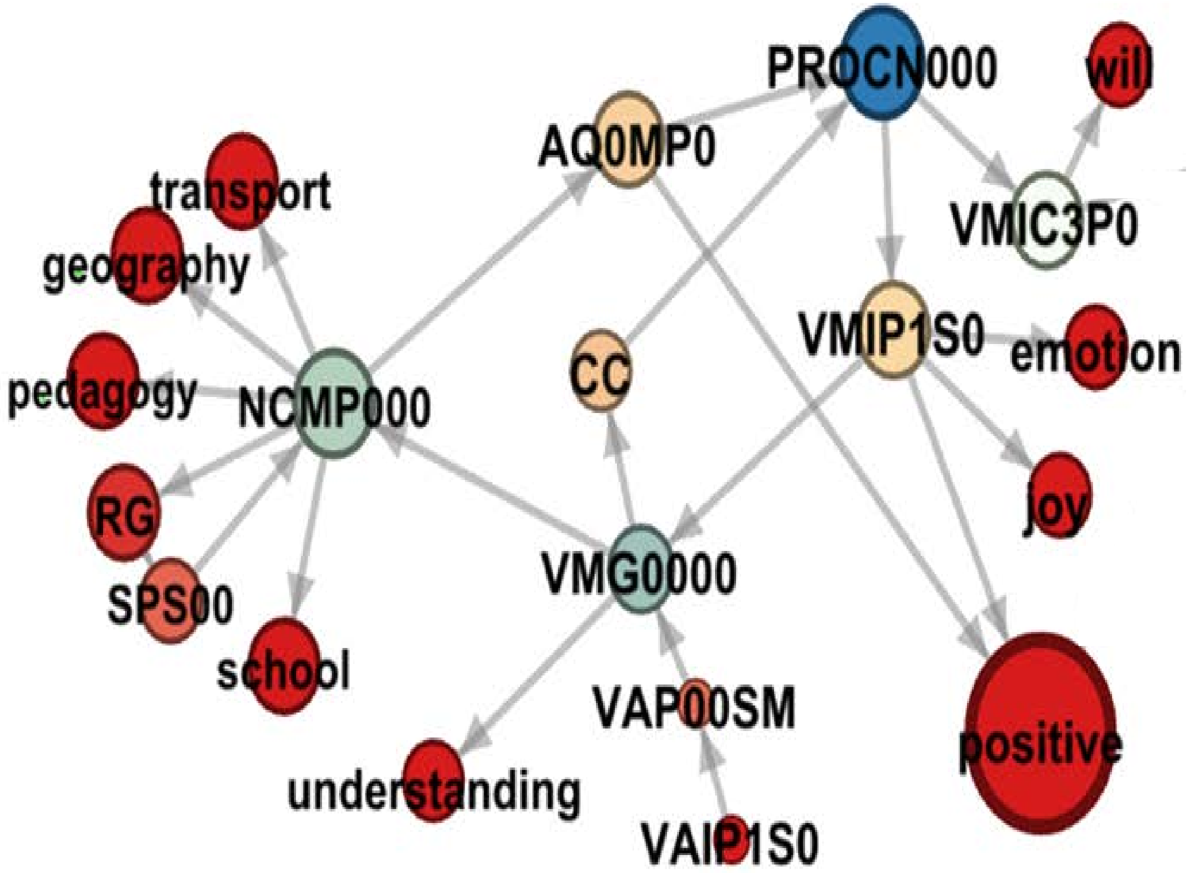
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

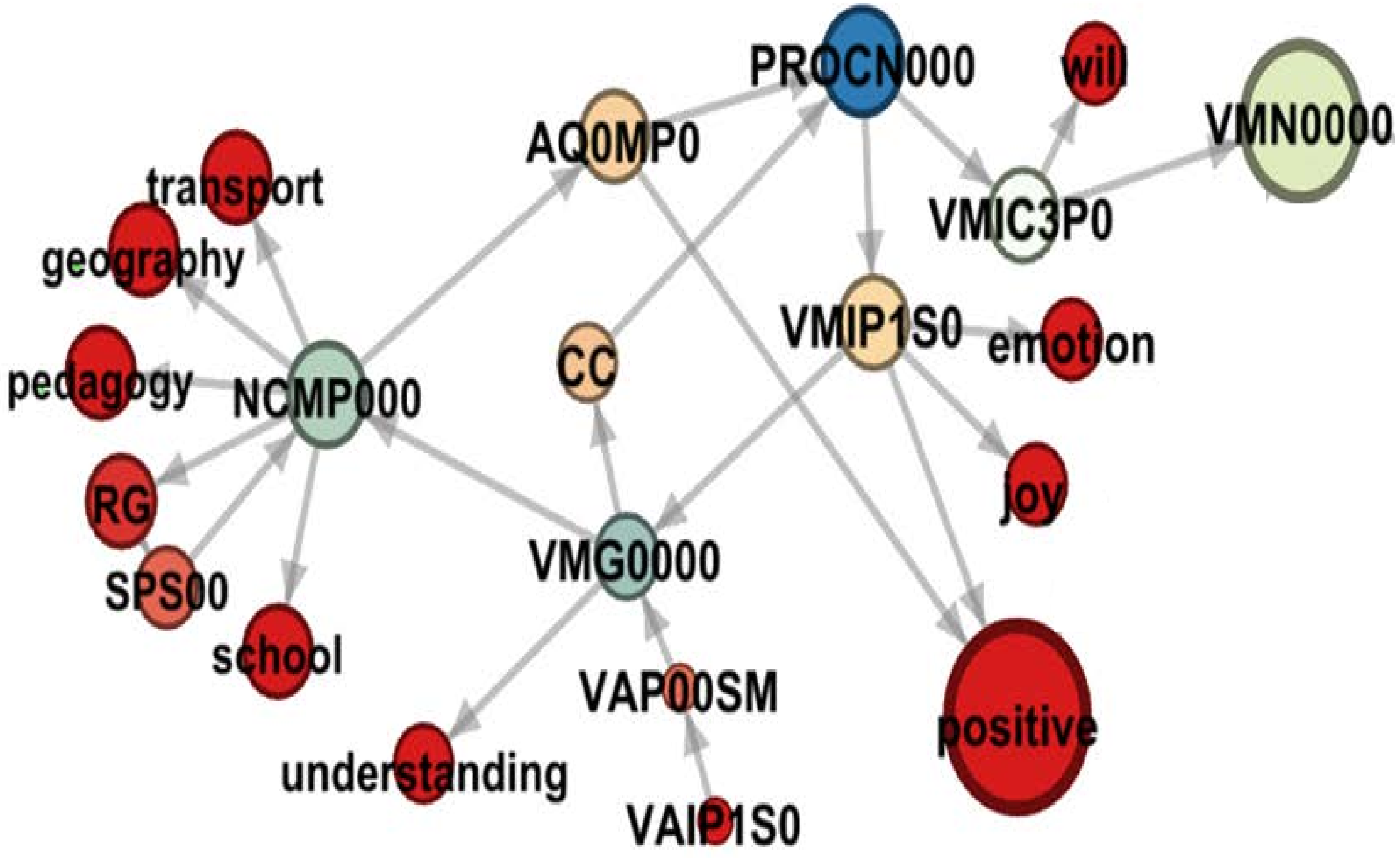
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

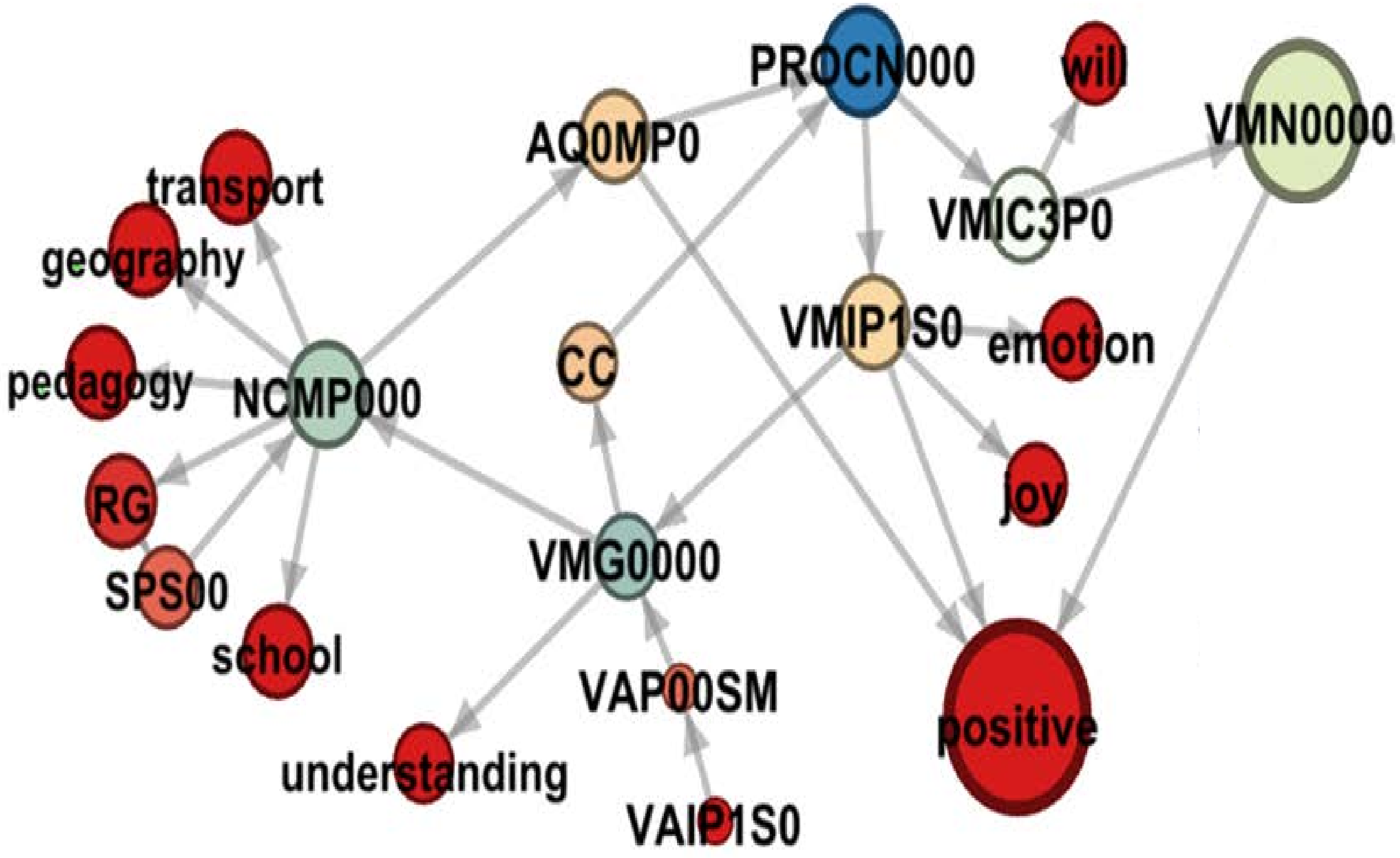
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

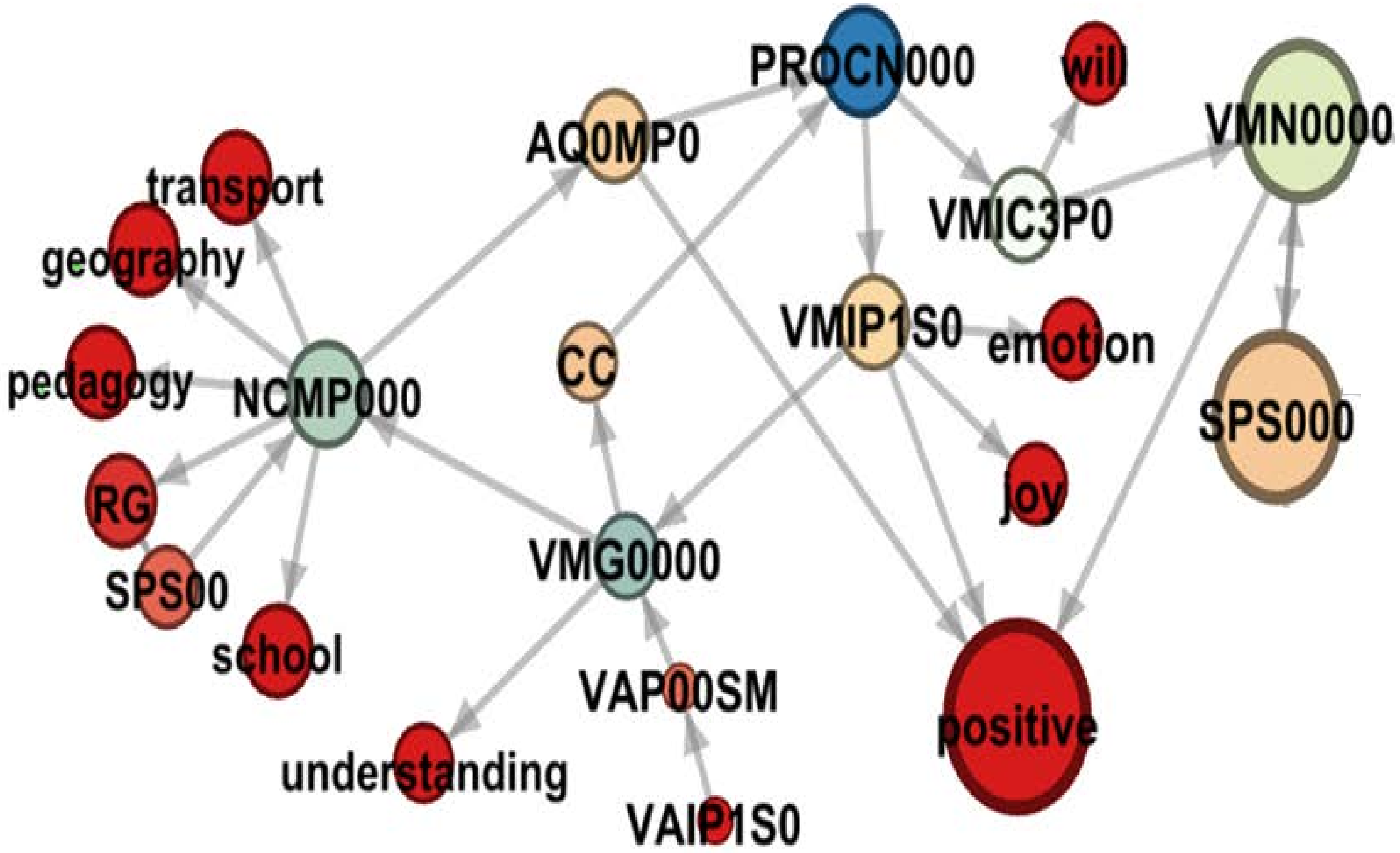
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

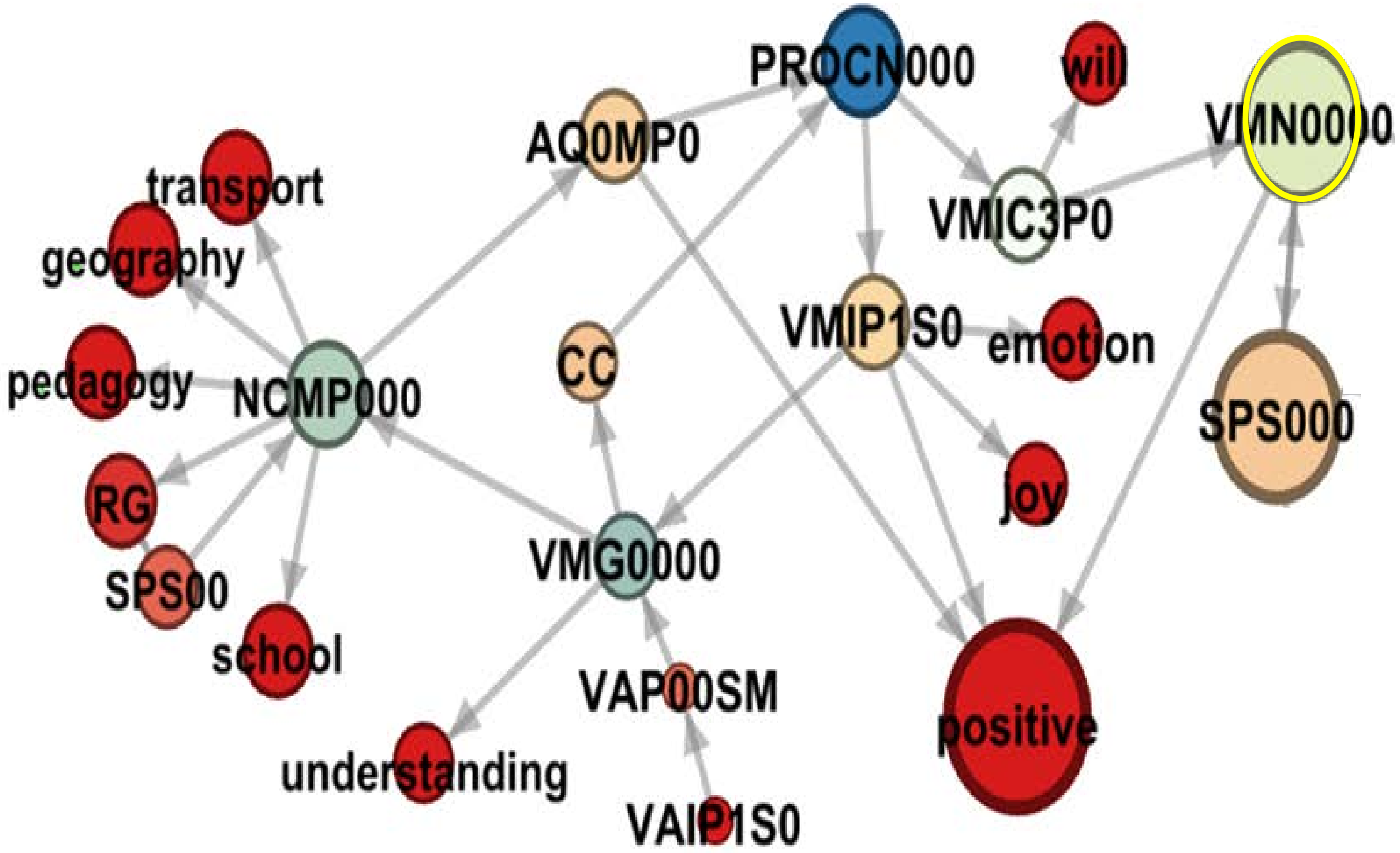
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

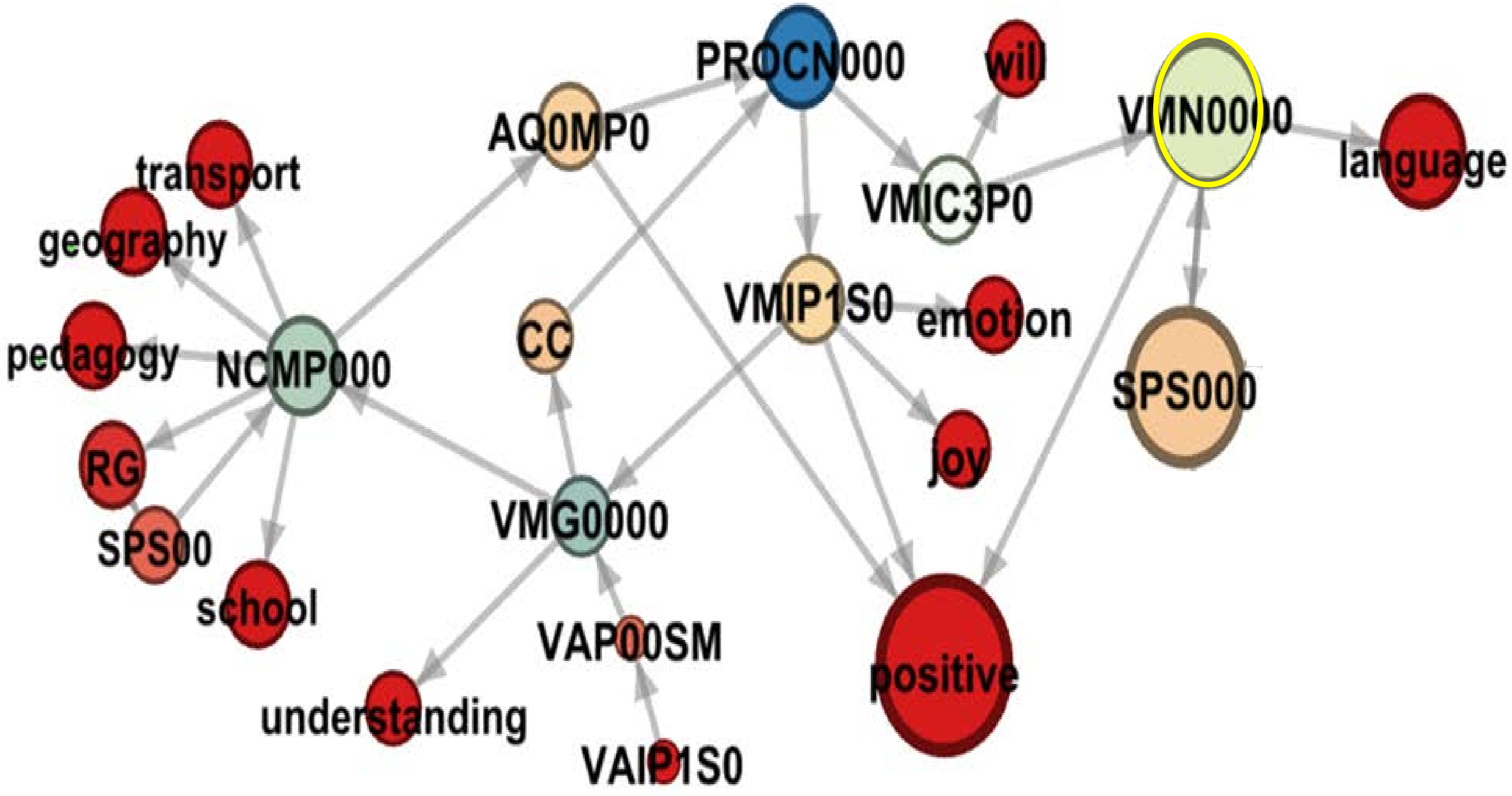
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

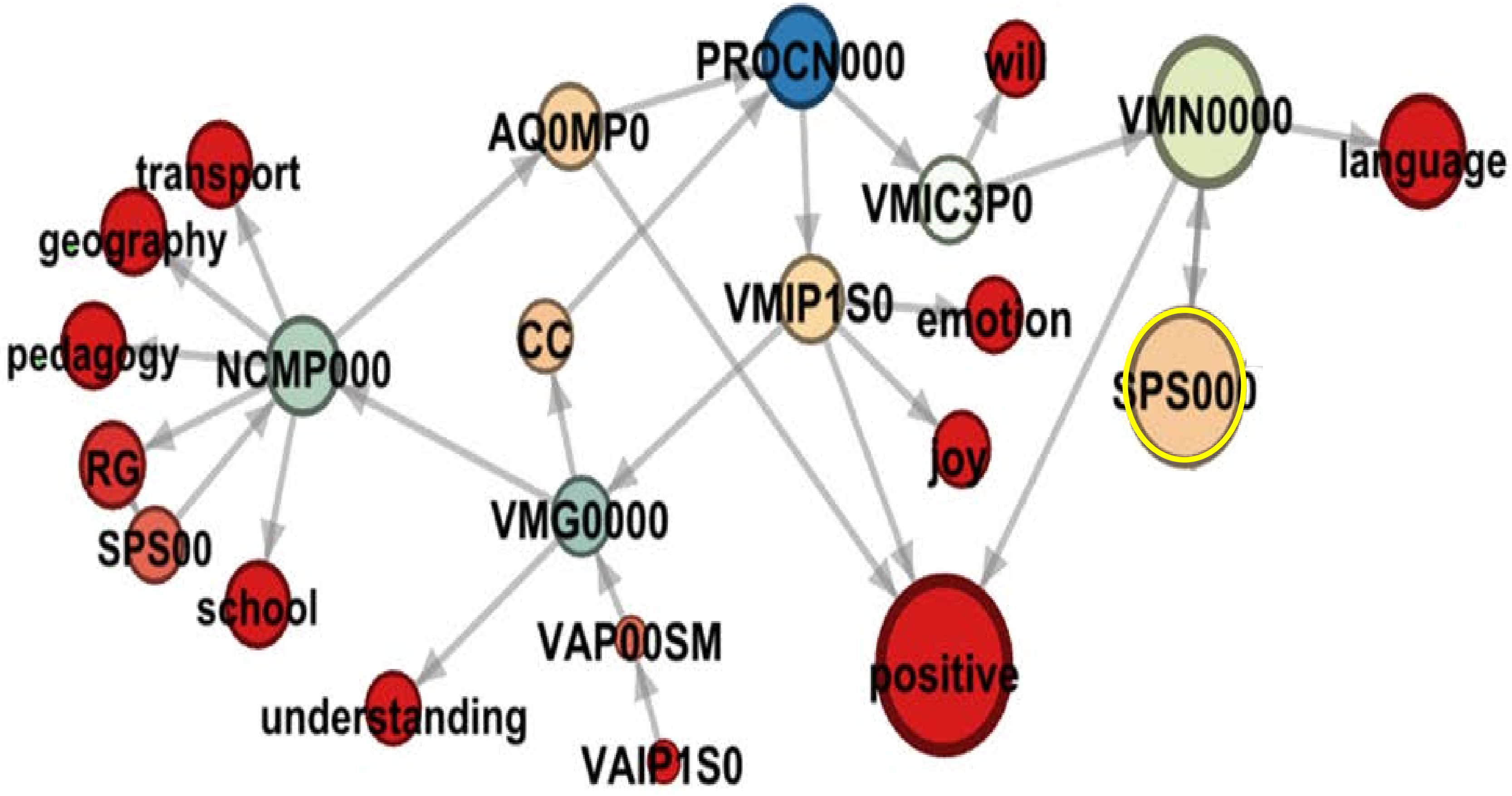
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

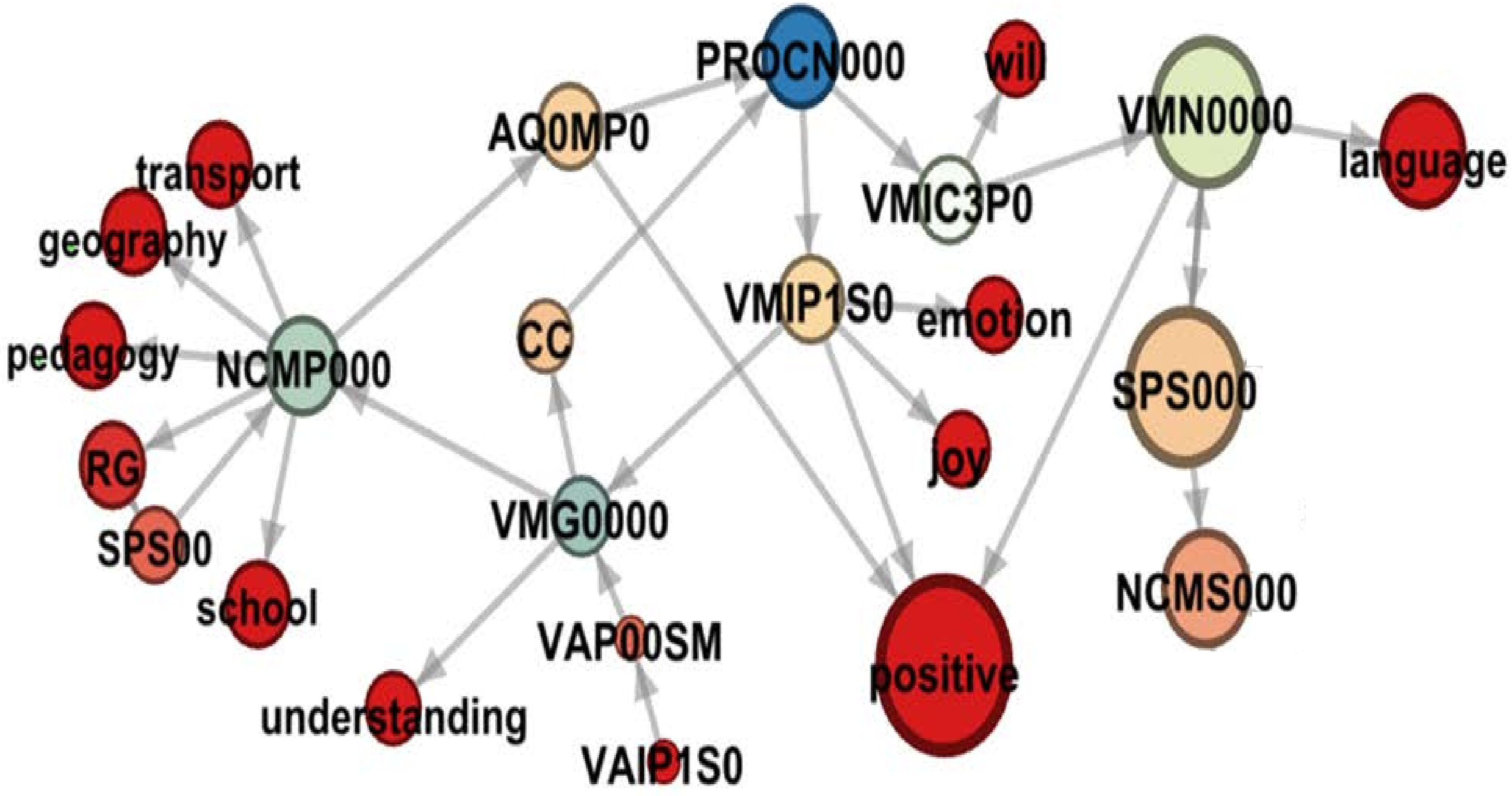
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

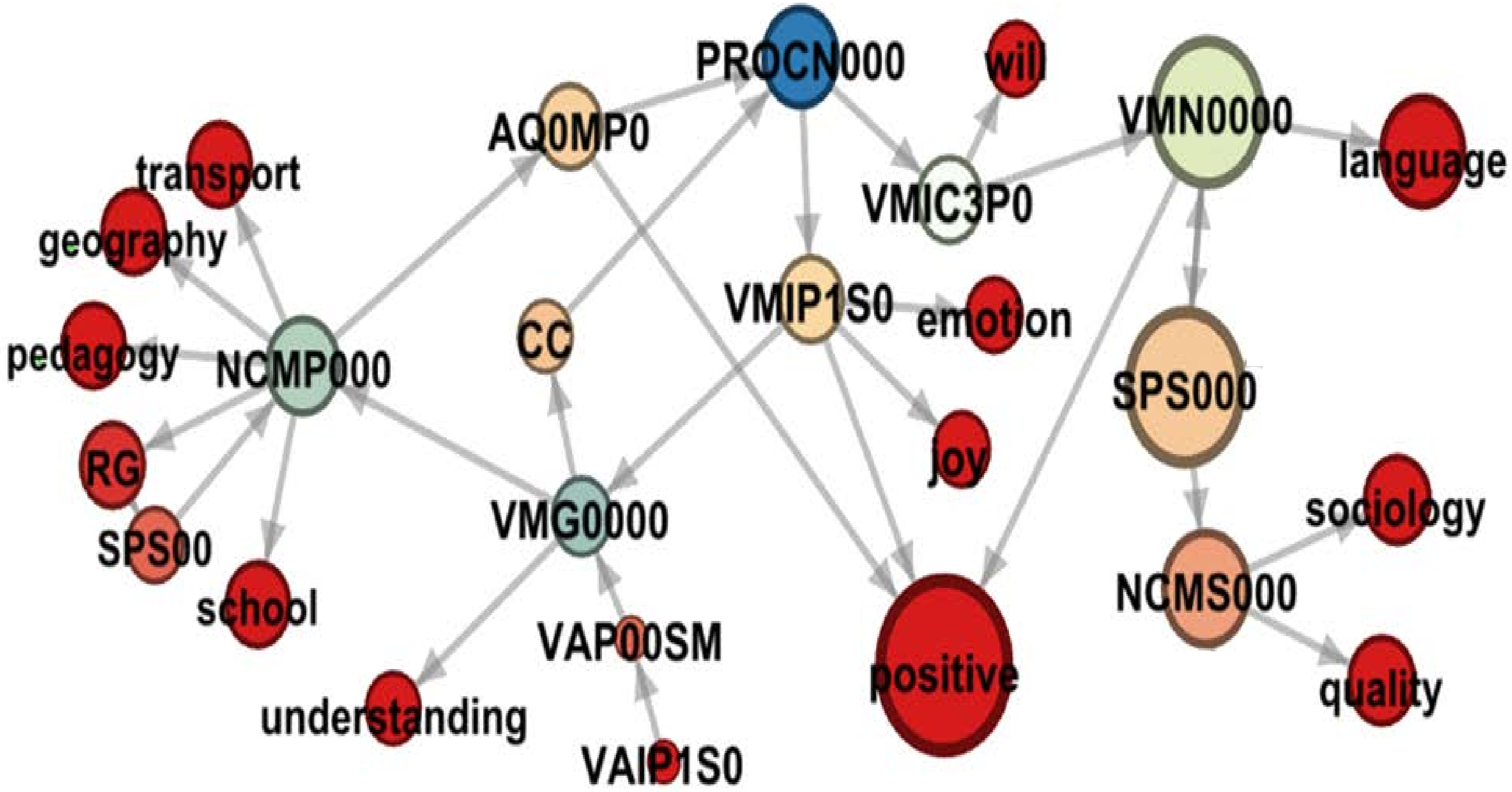
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

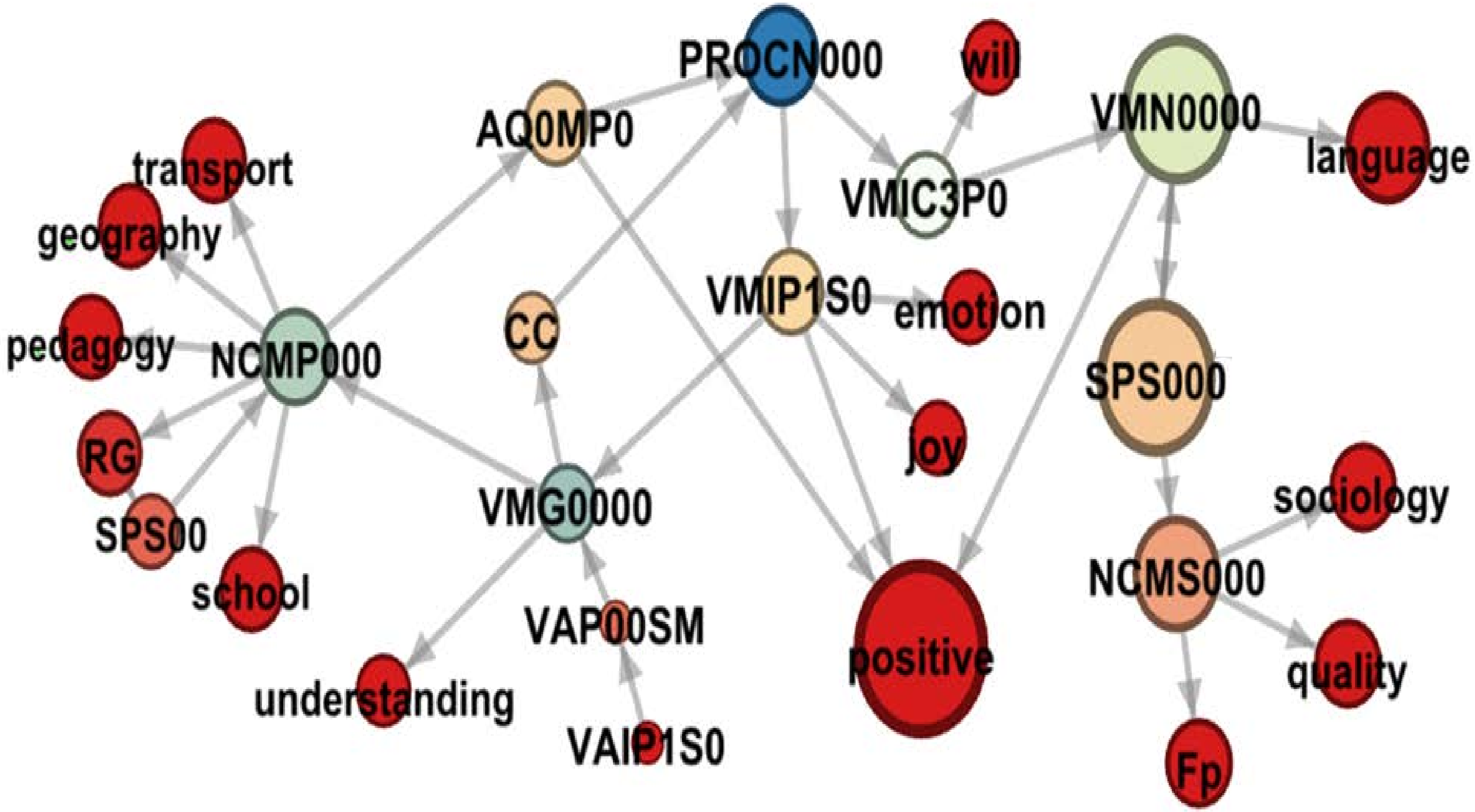
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



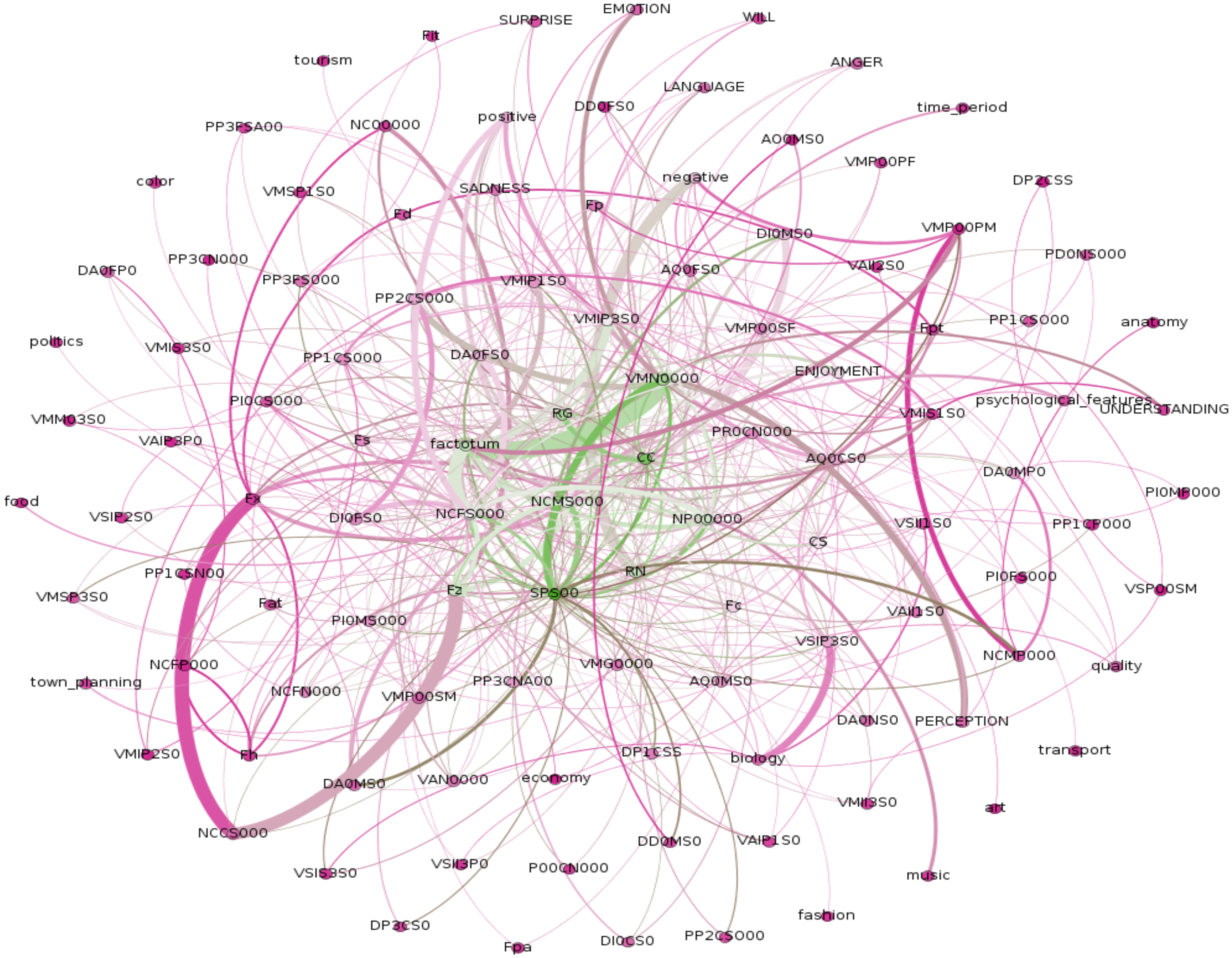
EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

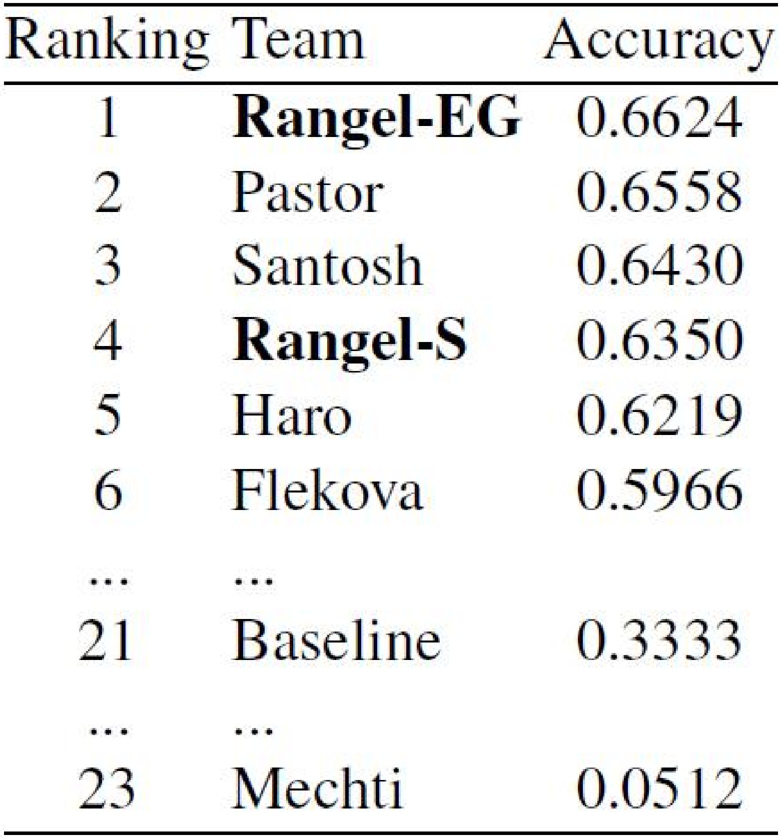
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.



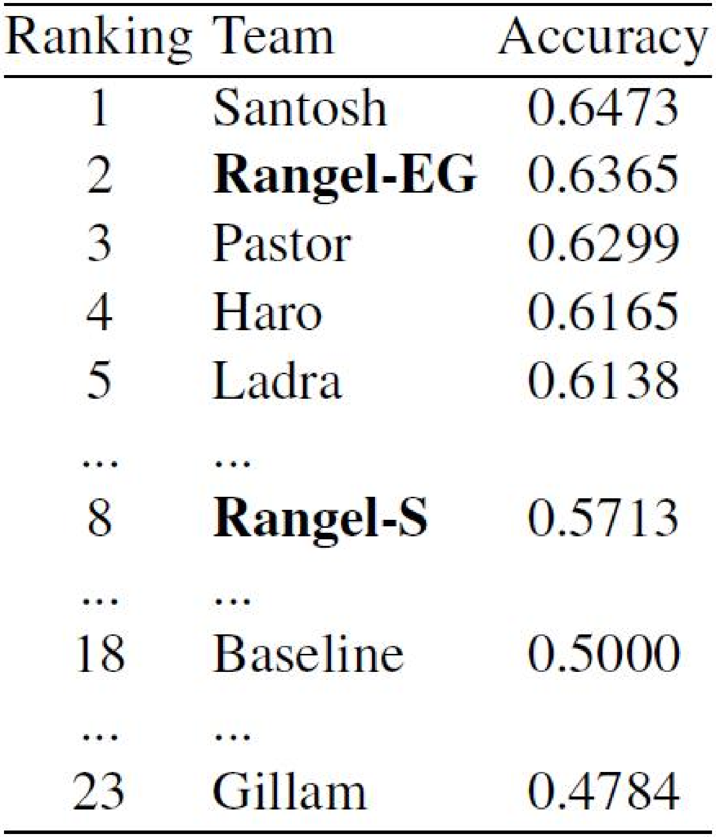
EmoGraph: author’s sentences



Style +SEL (S) vs EmoGraph (EG)



Ranking	Team	Accuracy
1	**Rangel-EG**	0.6624
2	Pastor	0.6558
3	Santosh	0.6430
4	**Rangel-S**	0.6350
5	Haro	0.6219
6	Flekova	0.5966
...	...	
21	Baseline	0.3333
...	...	
23	Mechti	0.0512



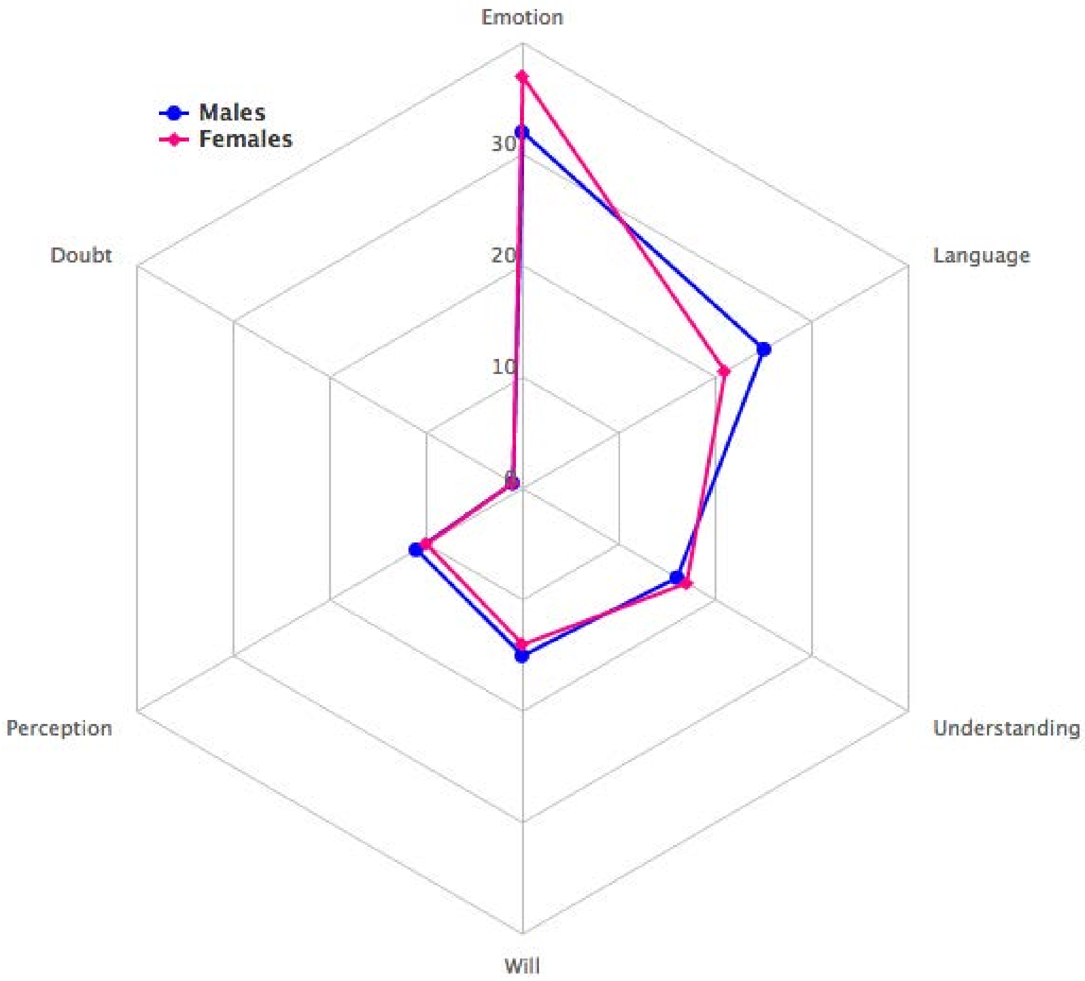
Ranking	Team	Accuracy
1	Santosh	0.6473
2	**Rangel-EG**	0.6365
3	Pastor	0.6299
4	Haro	0.6165
5	Ladra	0.6138
...	...	
8	**Rangel-S**	0.5713
...	...	
18	Baseline	0.5000
...	...	
23	Gillam	0.4784

Rangel F., Rosso P.. On the impact of emotions on author profiling. Information, Processing & Management, 52(1): 73-92, 2016

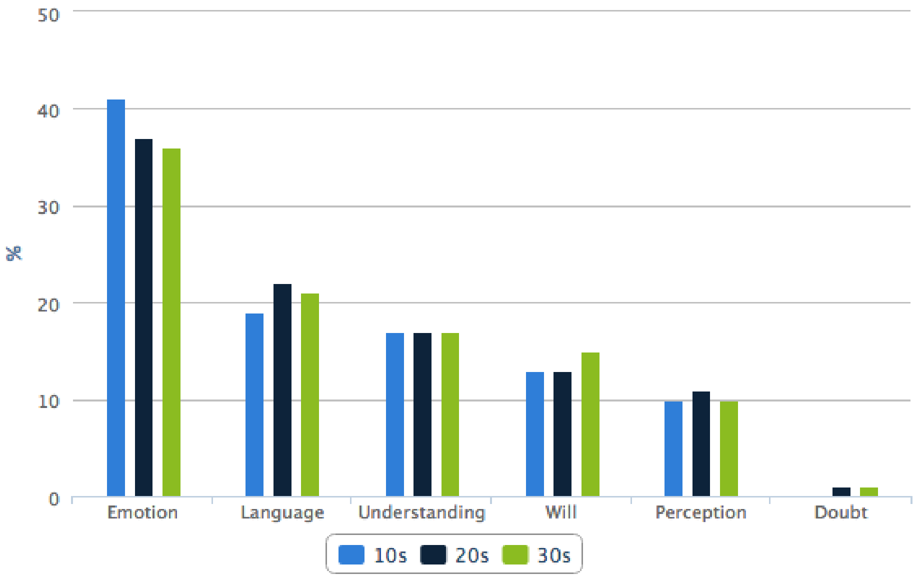
Levin's verb classes

Manual labelling of verbs in Spanish (158) and in English (172) by computational linguists (Autoritas)

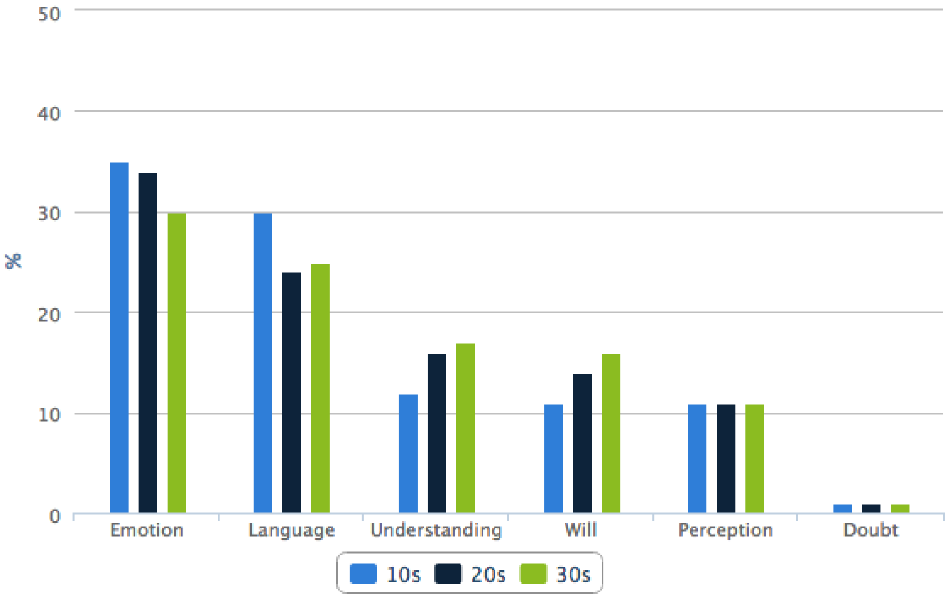
- **Emotion**: sentir, querer, amar…
- **Language**: decir, declarar, hablar…
- **Understanding**: entender, saber, conocer, pensar…
- **Perception**: oler, ver, escuchar…
- **Will**: deber, prohibir, permitir…
- **Doubt**: dudar, ignorar…



Levin’s verbs per gender & age



Females vs. Males



B. Levin. English Verb Classes and Alternations. University of Chicago Press, 1993.

To sum up on stylometry

- Plagiarism detection: when due to high paraphrasing it is difficult to provide an external evidence of plagiarism, then studying changes in writing style could be the only option

- Analysis of writing style could be useful also for tasks such as author profiling

Thanks

Paolo Rosso

prosso@dsic.upv.es