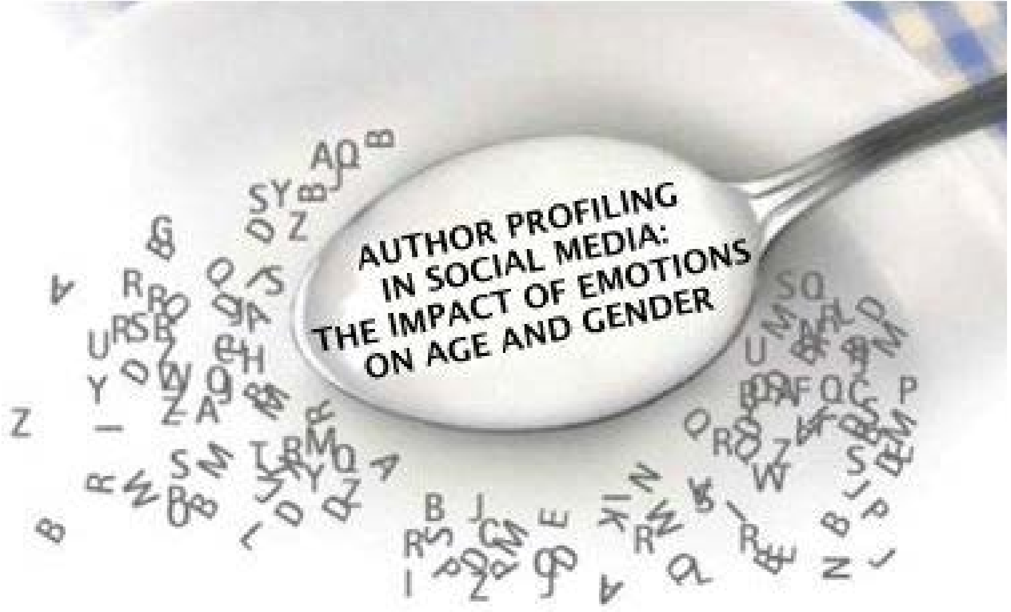


Paolo Rosso

Pattern Recognition and Human Language Technology

Research Center https://www.prhlt.upv.es/

Universitat Politècnica de València

Swiss Text 2016, Zurich 8th June

Outline

- Author profiling: gender and age
- Author profiling in social media: shared tasks @ PAN
- EmoGraph: The impact of emotions

Author Profiling

- Distinguishing between classes of authors, rather than individual authors
- **Marketing**, **Forensic Linguistics, Security**
  - Gender
  - Age
  - Personality profile: Big five personality traits
  - Native language
  - Language variety
  - Ideological/organizational affiliation
  - Etc.

Author Profiling: Gender and Age

Moshe Koppel, Bar-Illan University

J. W. Pennebaker, The University of Texas Austin

…

Which is Male/Female?

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]

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

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

Which is Male/Female?

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.

Which is Male/Female?

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.

Which is Male/Female?

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.

Which is Male/Female?

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.

Which is Male/Female?

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.

Teen **Twenties** **Thirties**

Male **Female**

Example 1

Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotton, and I wanted to cry, but...it's ok.

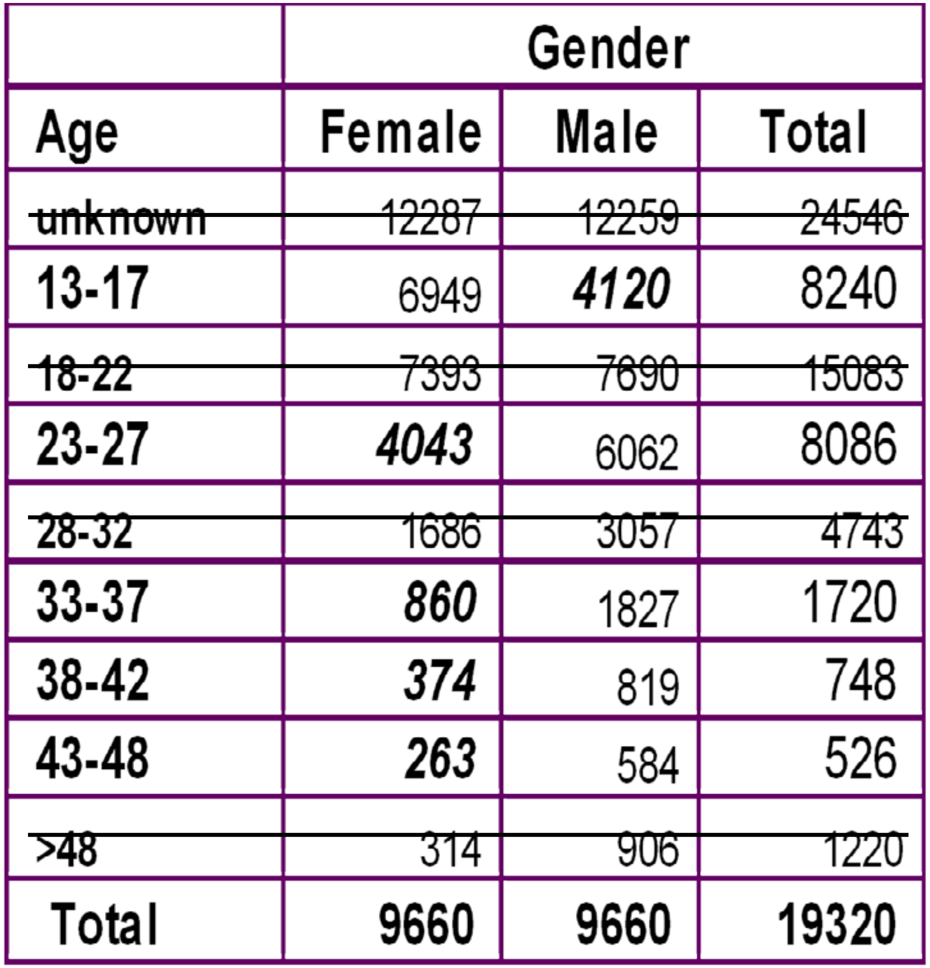
Teen Twenties Thirties

Male **Female**

Example 1

Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotton, and I wanted to cry, but...it's ok.

Blog Corpus



	Gender		
Age	**Female**	**Male**	**Total**
~~unknown~~	~~12287~~	~~12259~~	~~24546~~
13-17	6949	***4120***	8240
~~18-22~~	~~7393~~	~~7690~~	~~15083~~
23-27	***4043***	6062	8086
~~28-32~~	~~1686~~	~~3057~~	~~4743~~
33-37	***860***	1827	1720
38-42	***374***	819	748
43-48	***263***	584	526
~~>48~~	~~314~~	~~906~~	~~1220~~
Total	**9660**	**9660**	**19320**

Final balanced corpus:

- 19,320 total blogs
  - – 8240 in “10s”
  - – 8086 in “20s”
  - – 2994 in “30s”

- 681,288 total posts
- 141,106,859 total words

J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pages 199–205. AAAI, 2006.

The lifecycle of the common blogger...

Word	10s	20s	30s
maths	105	3	2
homework	137	18	15
bored	384	111	47
sis	74	26	10
boring	369	102	63
awesome	292	128	57
mum	125	41	23
crappy	46	28	11
mad	216	80	53
dumb	89	45	22

The lifecycle of the common blogger...

Word	10s	20s	30s
maths	105	3	2
homework	137	18	15
bored	384	111	47
sis	74	26	10
boring	369	102	63
awesome	292	128	57
mum	125	41	23
crappy	46	28	11
mad	216	80	53
dumb	89	45	22

Word	10s	20s	30s
semester	22	44	18
apartment	18	123	55
drunk	77	88	41
beer	32	115	70
student	65	98	61
album	64	84	56
college	151	192	131
someday	35	40	28
dating	31	52	37
bar	45	153	111

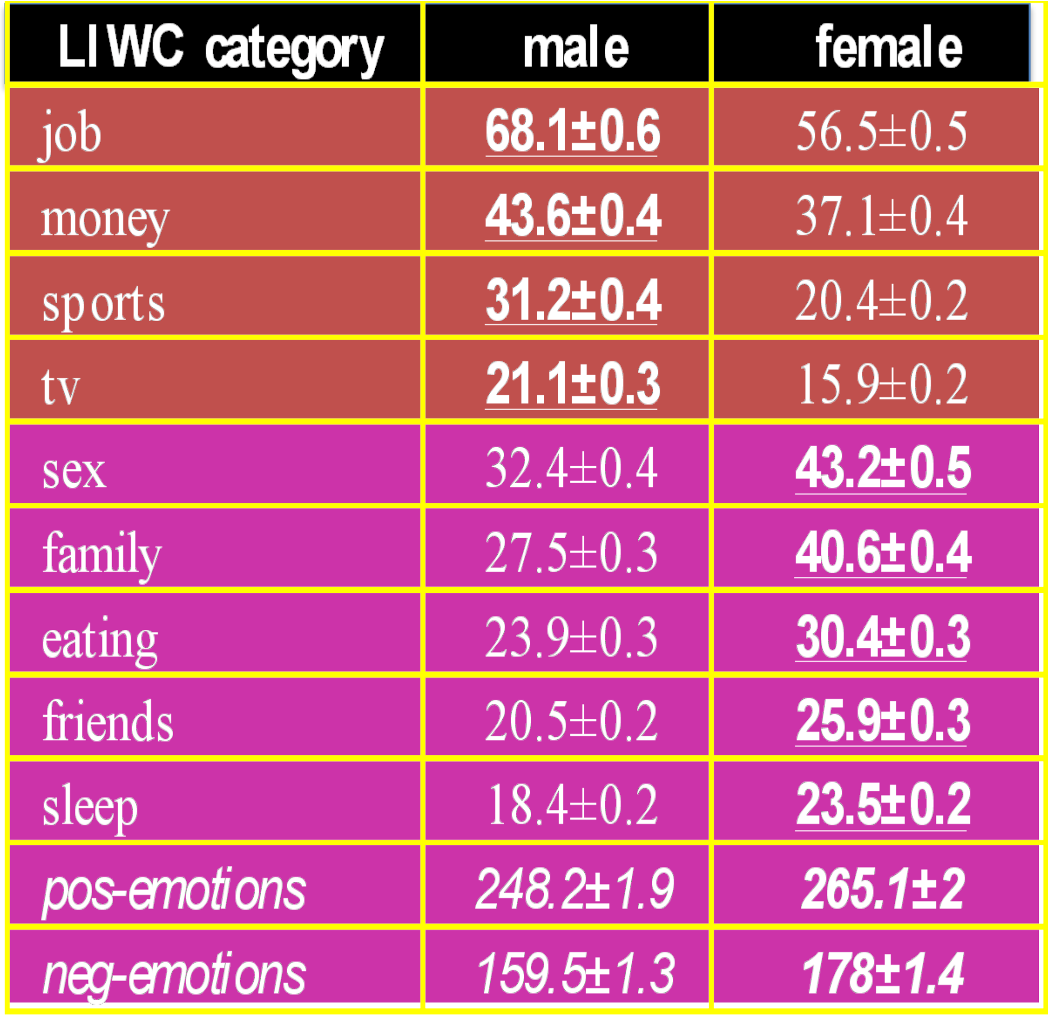
The lifecycle of the common blogger...

Word	10s	20s	30s
maths	105	3	2
homework	137	18	15
bored	384	111	47
sis	74	26	10
boring	369	102	63
awesome	292	128	57
mum	125	41	23
crappy	46	28	11
mad	216	80	53
dumb	89	45	22

Word	10s	20s	30s
semester	22	44	18
apartment	18	123	55
drunk	77	88	41
beer	32	115	70
student	65	98	61
album	64	84	56
college	151	192	131
someday	35	40	28
dating	31	52	37
bar	45	153	111

Word	10s	20s	30s
marriage	27	83	141
development	16	50	82
campaign	14	38	70
tax	14	38	72
local	38	118	185
democratic	13	29	59
son	51	92	237
systems	12	36	55
provide	15	54	69
workers	10	35	46

Men are from Mars...
Women are from Venus...



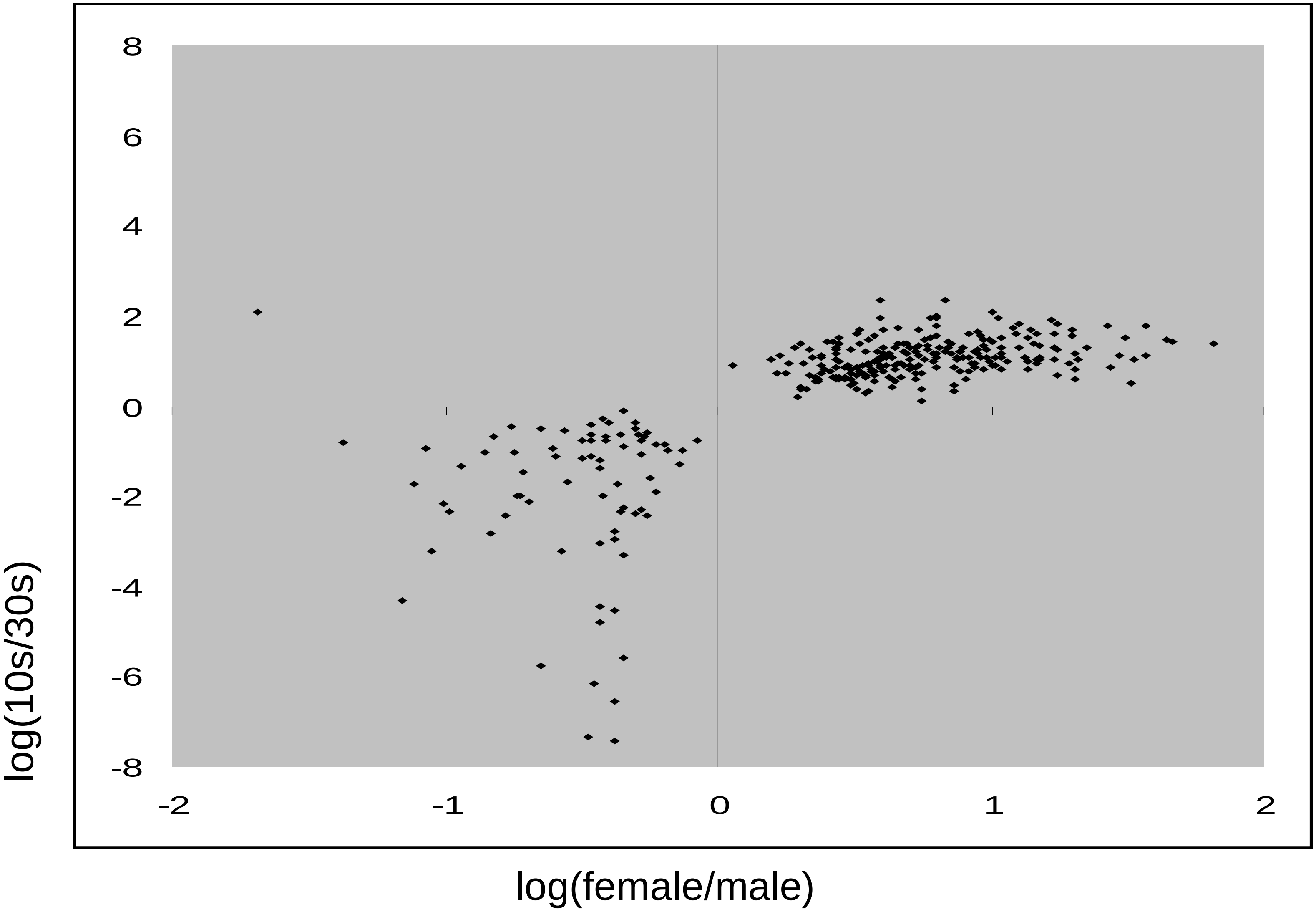
LIWC category	male	female
job	**68.1±0.6**	56.5±0.5
money	**43.6±0.4**	37.1±0.4
sports	**31.2±0.4**	20.4±0.2
tv	**21.1±0.3**	15.9±0.2
sex	32.4±0.4	**43.2±0.5**
family	27.5±0.3	**40.6±0.4**
eating	23.9±0.3	**30.4±0.3**
friends	20.5±0.2	**25.9±0.3**
sleep	18.4±0.2	**23.5±0.2**
pos-emotions	*248.2±1.9*	***265.1±2***
neg-emotions	*159.5±1.3*	***178±1.4***

J. W. Pennebaker - LIWC: Linguistic Inquiry and Word Count

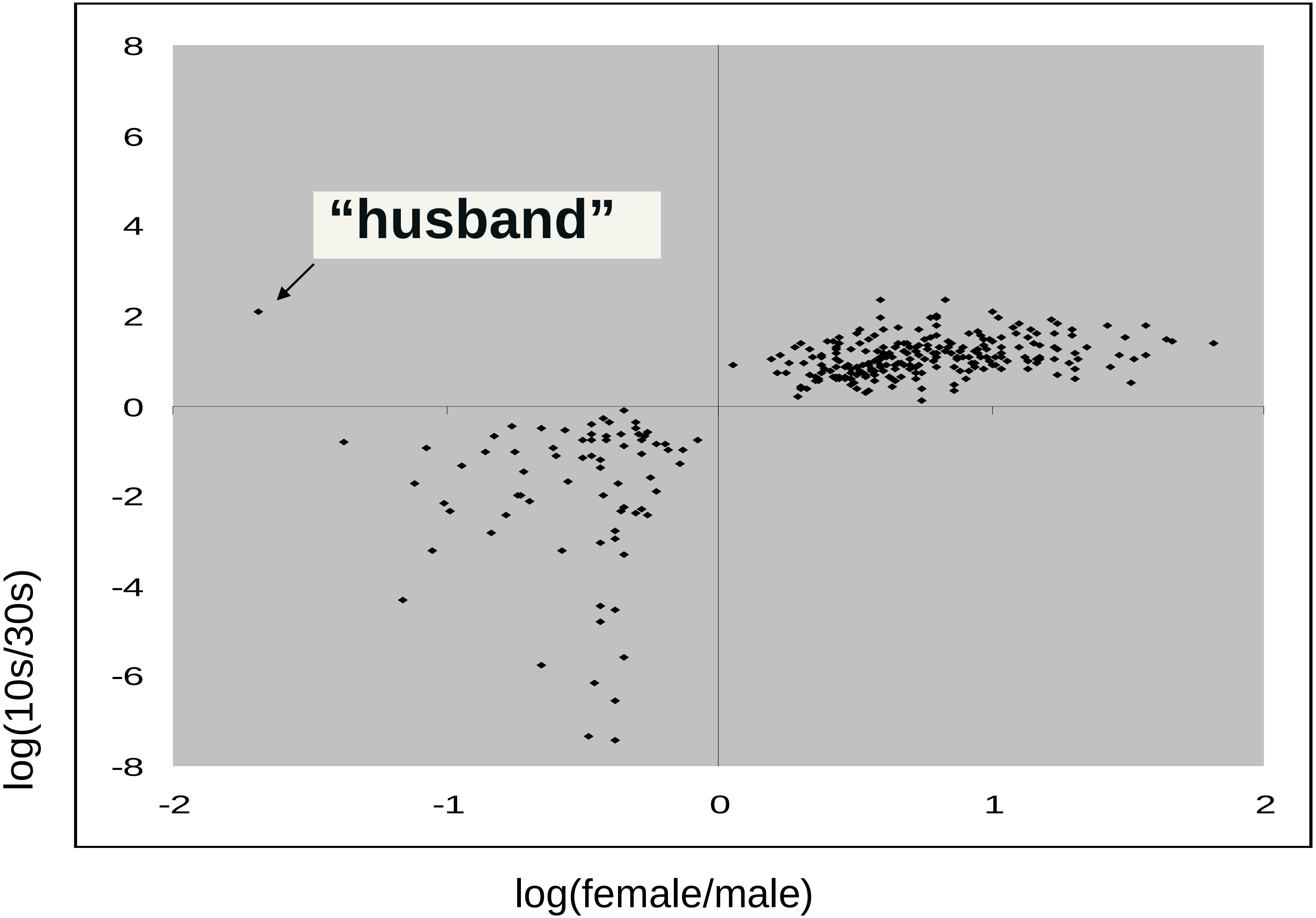
Relating Age & Gender

- Now…is there a linguistic connection between age and gender?
- Consider the most distinctive words for both Age and Gender:
  - Intersect the 1000 words with **highest Age information gain** and the 1000 words with **highest Gender information gain**
  - Total of 316 words
  - Plot log(30s/10s) vs. log(male/female)

Relating Age & Gender



Relating Age & Gender



“husband”

Author Profiling in Social Media: Shared Tasks @ PAN

Francisco Rangel, Autoritas Consulting

Paolo Rosso, Universitat Politècnica de València

…

Uncovering Plagiarism, Authorship, and Social Software Misuse

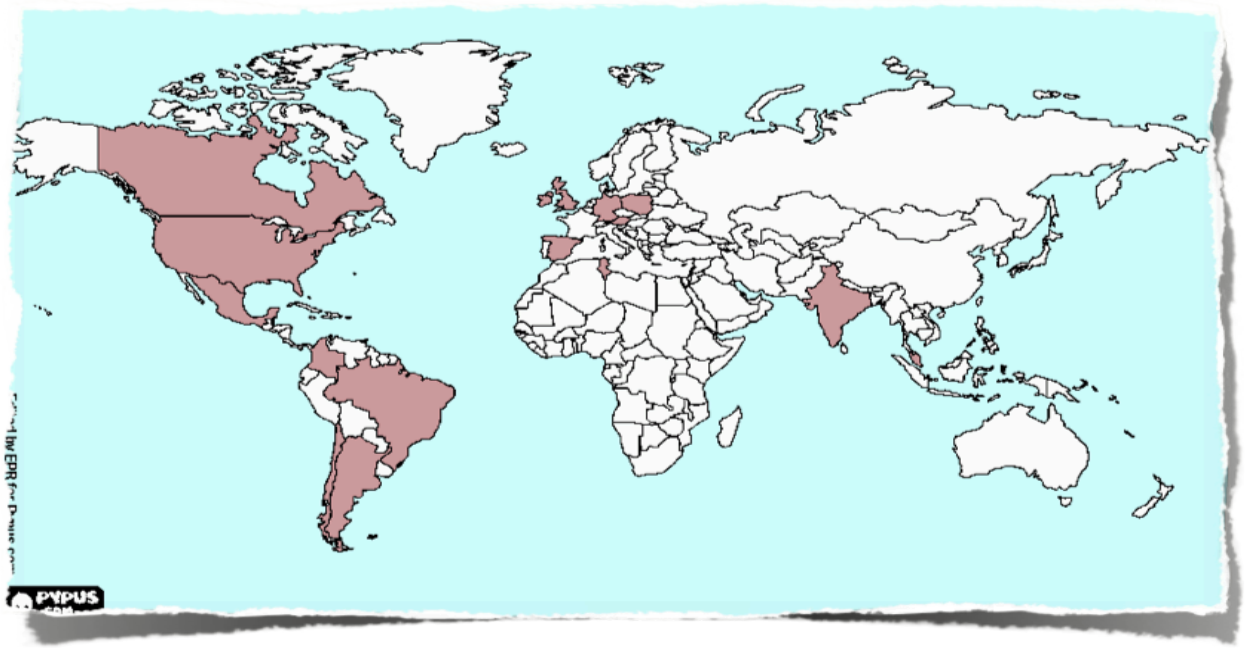
Since 2007 as workshop (SIGIR, ECAI); since 2009 organising benchmark activities: since 2010 as **PAN** Lab @ Conference and Labs of the Evaluation Forum: (**CLEF**) - http://pan.webis.de/

- Plagiarism detection (since 2009)
- Author identification (since 2011)
- **Author profiling** (since 2013)
- Online sexual predator (in 2012)
- Author obfuscation (since 2016)

Author Profiling @ PAN-13

- Teams submitting results: 21 (Registered teams: 64)
- (Towards) **big data**: 400,000 social media texts

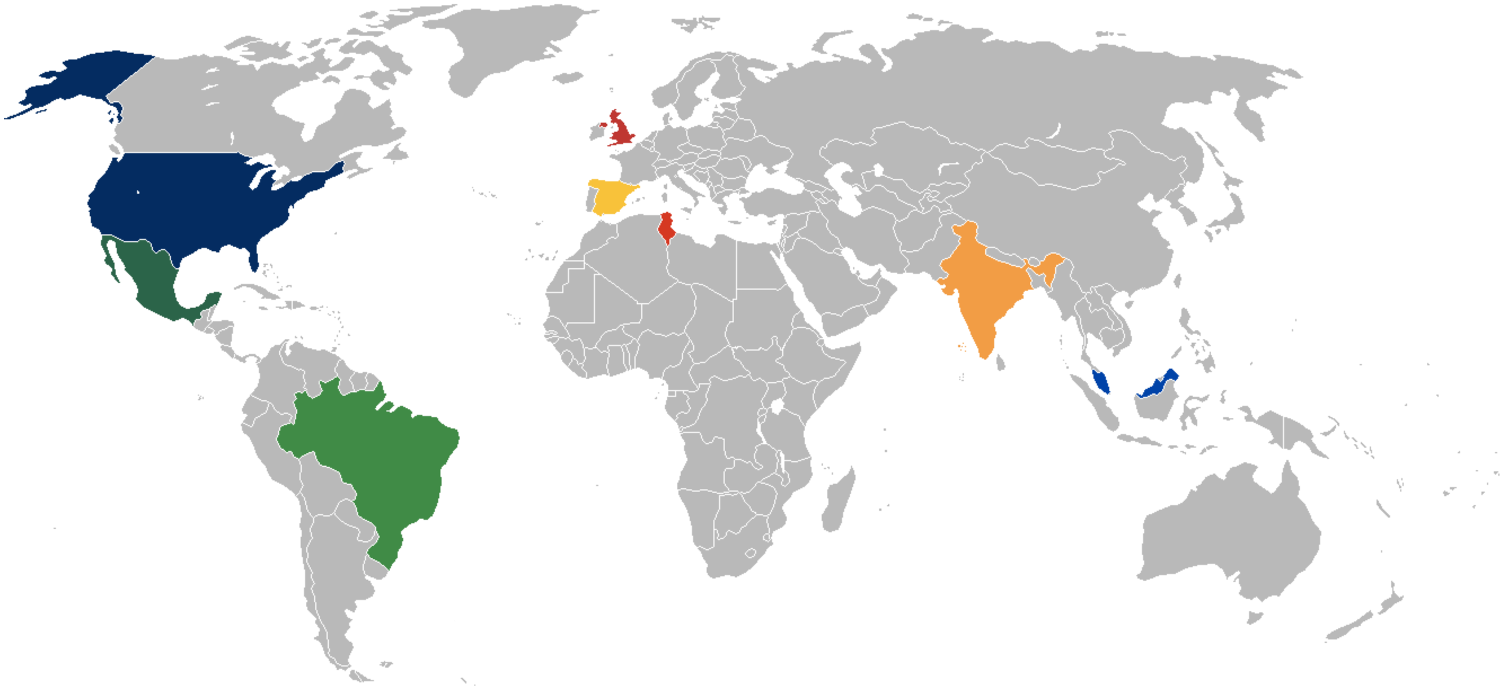
including **chat lines of potential pedophiles** (task @ PAN-12)



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

Author Profiling @ PAN-14

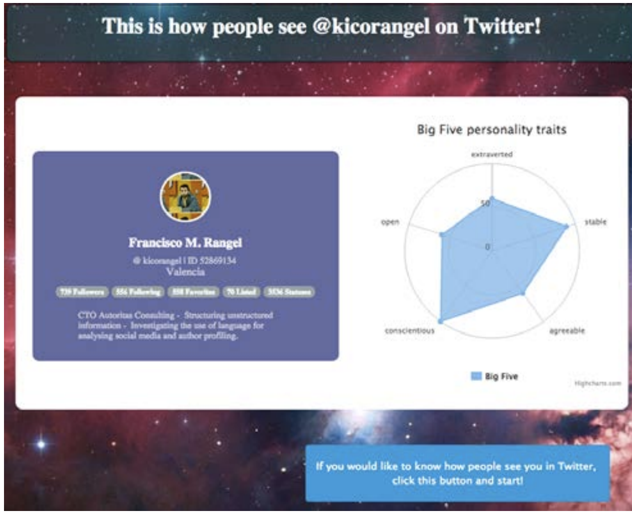
- Teams submitting results: 10
- **Social media + blogs + Twitter + TripAdvisor**



- **Age classes**: 18-24, 25-34, 35-49, 50-64, 65+
- Languages: English and Spanish

Author Profiling @ PAN-15

- Teams submitting results: 22
- Age classes: 18-24, 25-34, 35-49, 50+
- Gender, age and **personality** in **Twitter**
- http://your-personality-test.com/
- Languages: English, Spanish, Italian, Dutch



This is how people see @kicorangel on Twitter!

Francisco M. Rangel

@ kicorangel | ID 52869134

Valencia

739 Followers · 556 Following · 858 Favorites · 70 Listed · 3536 Statuses

CTO Autoritas Consulting - Structuring unstructured information - Investigating the use of language for analysing social media and author profiling.

If you would like to know how people see you in Twitter, click this button and start!

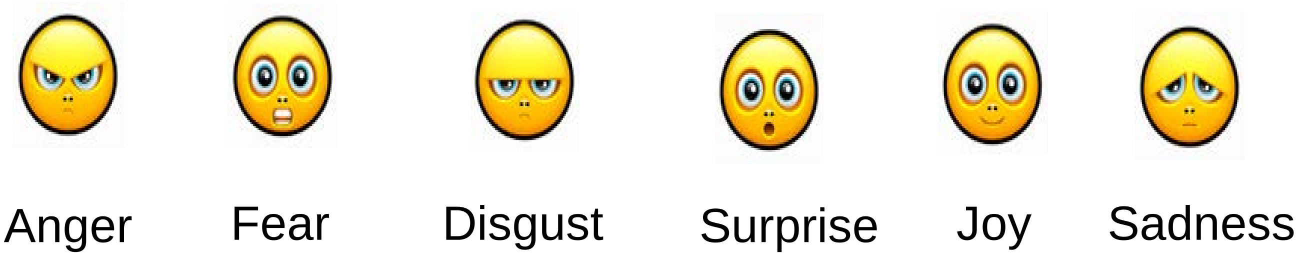
Author Profiling @ PAN-16

- Teams submitting results: 22
- **Cross-genre** gender and age (train: **Twitter**; e.g. test: **blogs**)
- Age classes: 18-24, 25-34, 35-49, 35-49, 50-64, 65+
- Languages: English, Spanish, Dutch

EmoGraph:
The Impact of Emotions

Francisco Rangel, Autoritas Consulting

Paolo Rosso, Universitat Politècnica de València



Style + Six Basic Emotions

- **Word frequency**: words with character flooding; words starting with capital letter; words in capital letters…
- **Punctuation marks**: frequency of use of dots, commas, colon, semicolon, exclamations and question marks
- **Part-Of-Speech**: frequency of use of each grammatical category
- **Emoticons**: number of different types of emoticons representing emotions
- **Spanish Emotion Lexicon**: words co-occurring with each emotion: **happiness, anger, fear, sadness, disgust, surprise**

Morpho-syntactic Analysis with Freeling

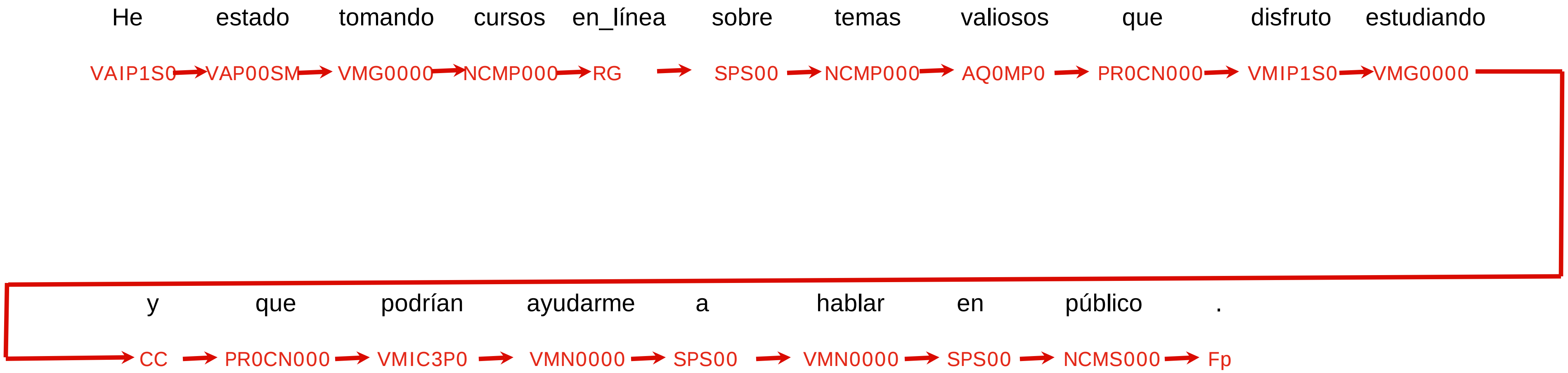
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.
I have been taking online courses about valuable subjects that I enjoy studying and might help me to speak in public.

He	estado	tomando	cursos	en_línea	sobre	temas	valiosos	que	disfruto	estudiando
VAIP1S0	VAP00SM	VMG0000	NCMP000	RG	SPS00	NCMP000	AQ0MP0	PR0CN000	VMIP1S0	VMG0000

y	que	podrían	ayudarme	a	hablar	en	público	.
CC	PR0CN000	VMIC3P0	VMN0000	SPS00	VMN0000	SPS00	NCMS000	Fp

POS sequence - Nodes - Edges creation

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 might help me to speak in public.



He estado tomando cursos en_línea sobre temas valiosos que disfruto estudiando

VAIP1S0 → VAP00SM → VMG0000 → NCMP000 → RG → SPS00 → NCMP000 → AQ0MP0 → PR0CN000 → VMIP1S0 → VMG0000

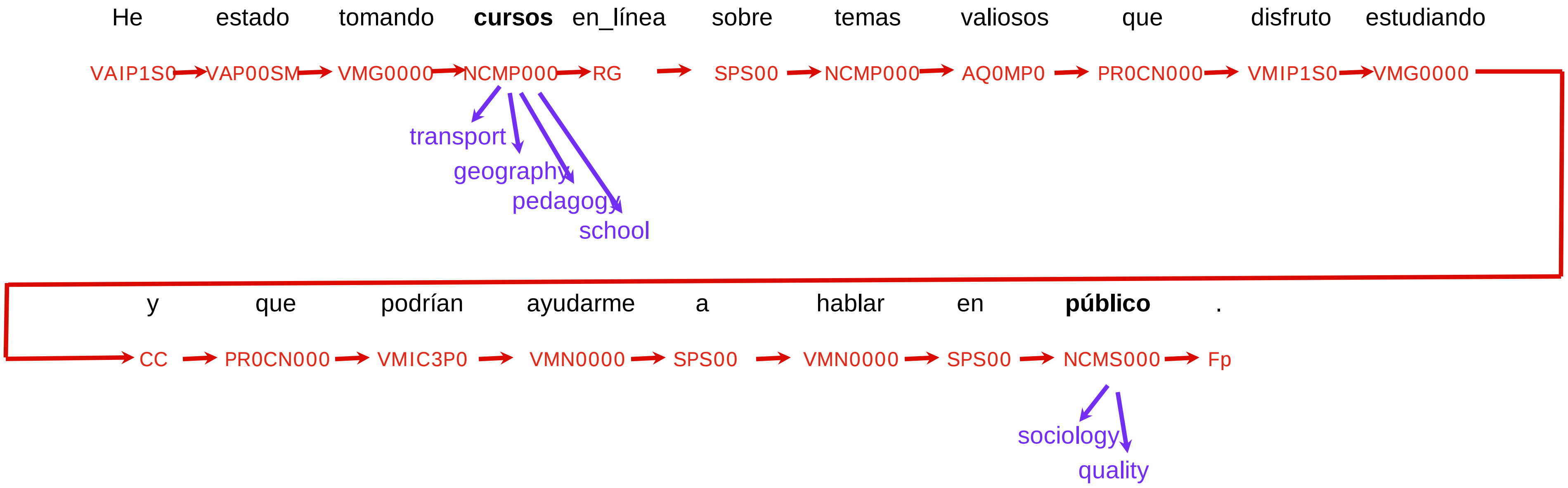
y que podrían ayudarme a hablar en público .

→ CC → PR0CN000 → VMIC3P0 → VMN0000 → SPS00 → VMN0000 → SPS00 → NCMS000 → Fp

Take into account that this sequence, when converted to graph, there are repeated nodes such as NCMP000 that create loops

Topics with Wordnet Domains

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 <u>courses</u> about valuable subjects that I enjoy studying and might help me to speak in <u>public</u>.



He estado tomando **cursos** en_línea sobre temas valiosos que disfruto estudiando

VAIP1S0 → VAP00SM → VMG0000 → NCMP000 → RG → SPS00 → NCMP000 → AQ0MP0 → PR0CN000 → VMIP1S0 → VMG0000

transport
geography
pedagogy
school

y que podrían ayudarme a hablar en **público** .

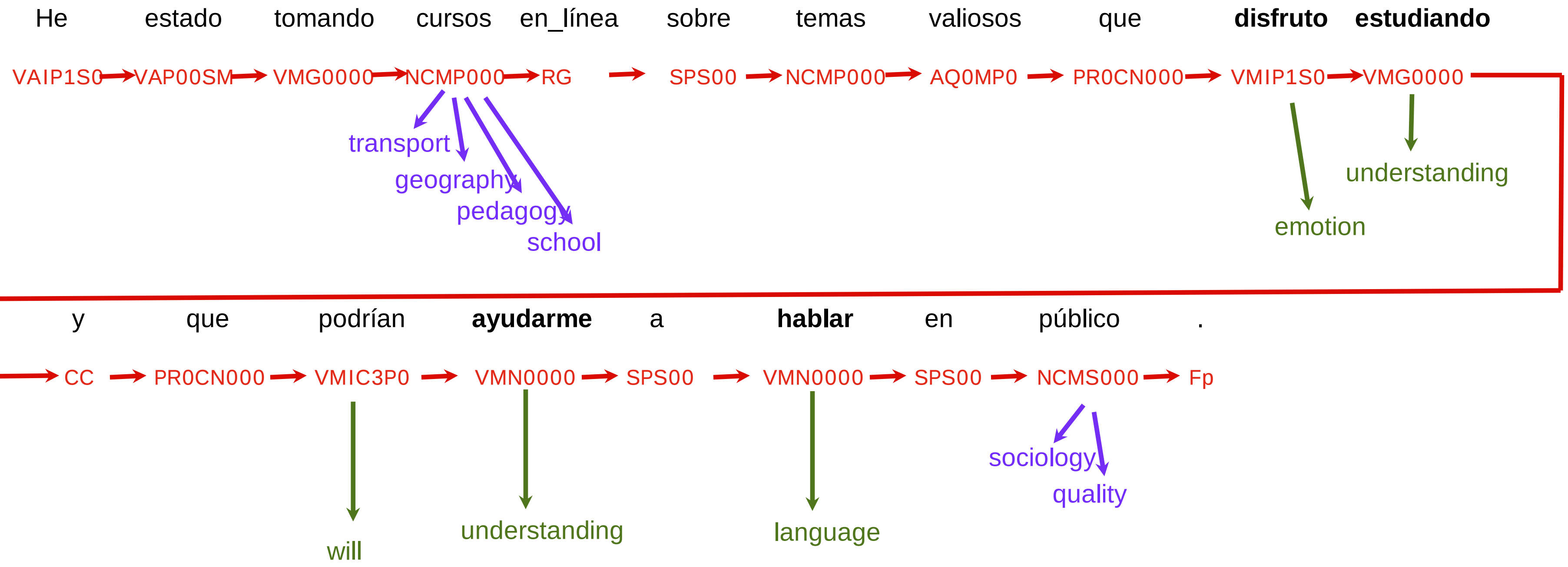
CC → PR0CN000 → VMIC3P0 → VMN0000 → SPS00 → VMN0000 → SPS00 → NCMS000 → Fp

sociology
quality

Semantic Classification of Verbs

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 might help me to speak in public.



He estado tomando cursos en_línea sobre temas valiosos que **disfruto** **estudiando**

VAIP1S0 → VAP00SM → VMG0000 → NCMP000 → RG → SPS00 → NCMP000 → AQ0MP0 → PR0CN000 → VMIP1S0 → VMG0000

transport

geography

pedagogy

school

understanding

emotion

y que podrían **ayudarme** a **hablar** en público .

→ CC → PR0CN000 → VMIC3P0 → VMN0000 → SPS00 → VMN0000 → SPS00 → NCMS000 → Fp

sociology

quality

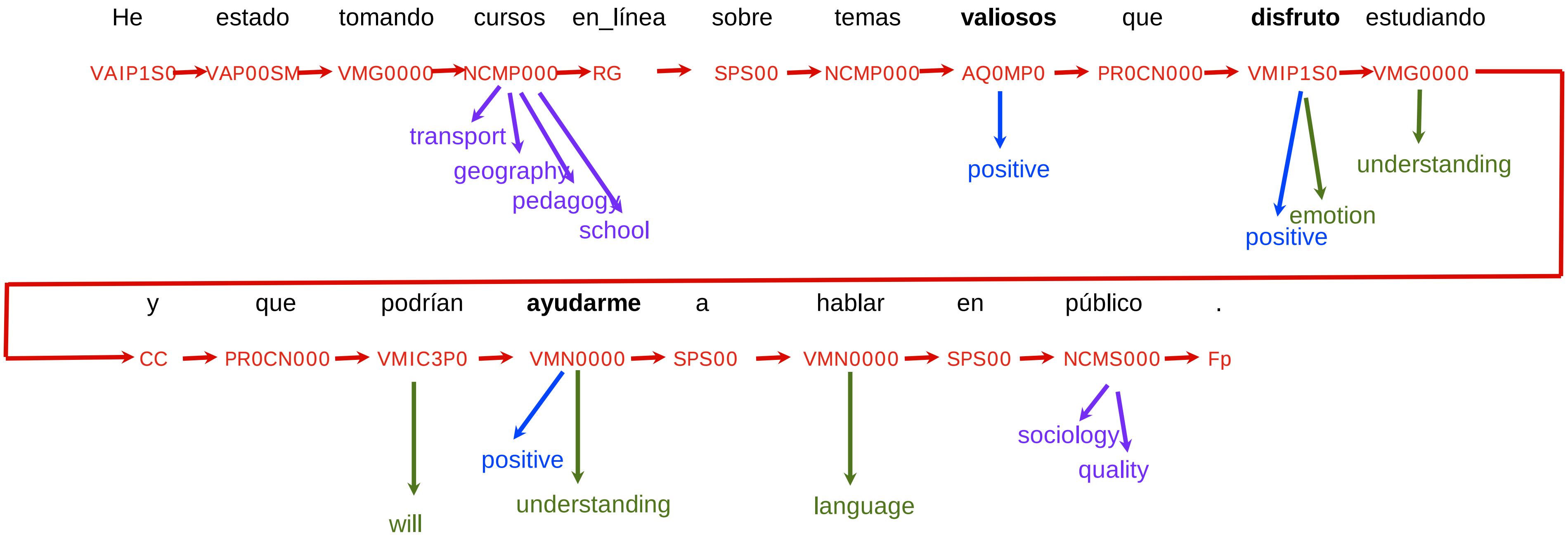
will

understanding

language

Polarity

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 <u>valuable</u> subjects that I <u>enjoy</u> studying and might <u>help</u> me to speak in public.



He estado tomando cursos en_línea sobre temas **valiosos** que **disfruto** estudiando

VAIP1S0 → VAP00SM → VMG0000 → NCMP000 → RG → SPS00 → NCMP000 → AQ0MP0 → PR0CN000 → VMIP1S0 → VMG0000

transport
geography
pedagogy
school

positive

positive
emotion

understanding

y que podrían **ayudarme** a hablar en público .

→ CC → PR0CN000 → VMIC3P0 → VMN0000 → SPS00 → VMN0000 → SPS00 → NCMS000 → Fp

will

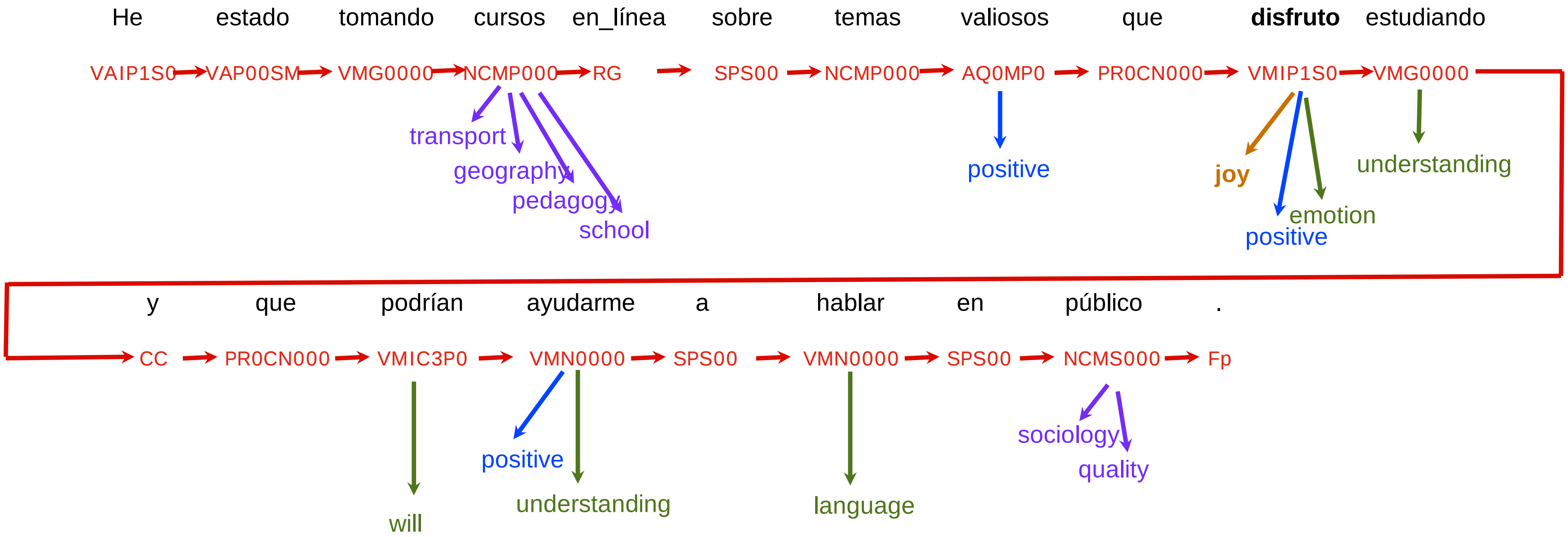
positive
understanding

language

sociology
quality

Emotions

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 might help me to speak in public.



He estado tomando cursos en_línea sobre temas valiosos que **disfruto** estudiando

VAIP1S0 → VAP00SM → VMG0000 → NCMP000 → RG → SPS00 → NCMP000 → AQ0MP0 → PR0CN000 → VMIP1S0 → VMG0000

transport
geography
pedagogy
school
positive
joy
emotion
positive
understanding

y que podrían ayudarme a hablar en público .

→ CC → PR0CN000 → VMIC3P0 → VMN0000 → SPS00 → VMN0000 → SPS00 → NCMS000 → Fp

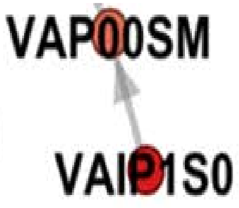
will
positive
understanding
language
sociology
quality

EmoGraph

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

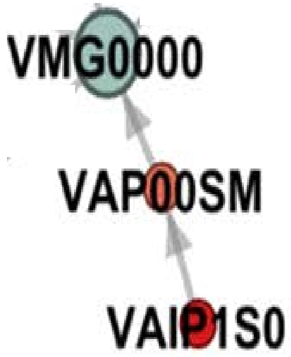
EmoGraph

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



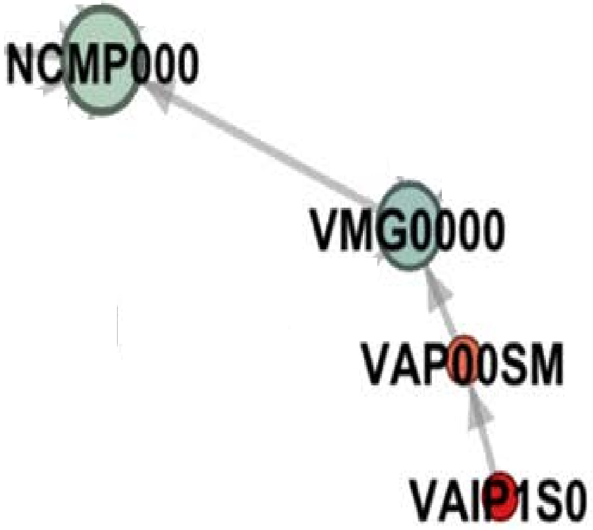
EmoGraph

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



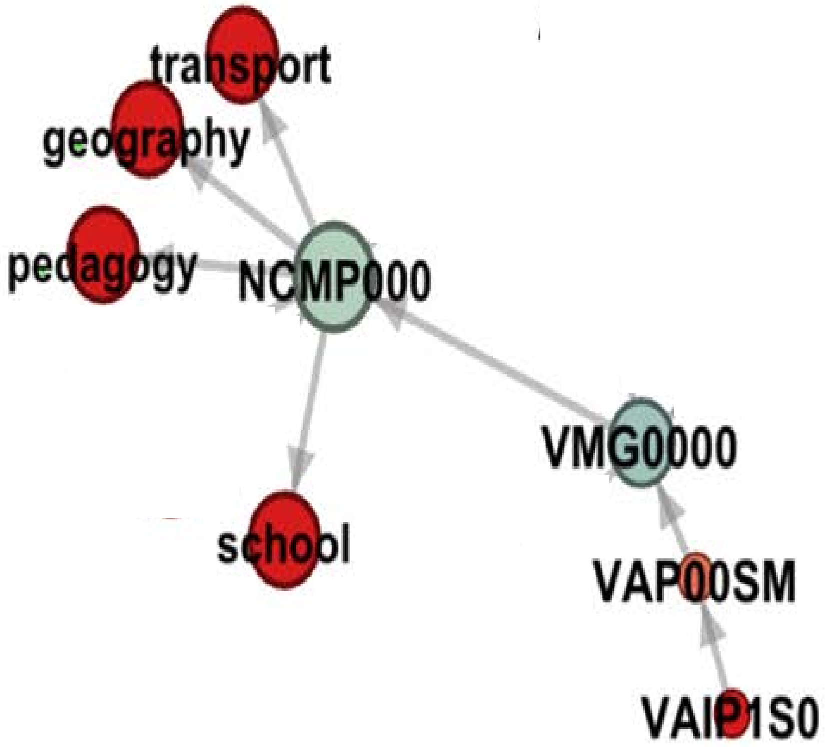
EmoGraph

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



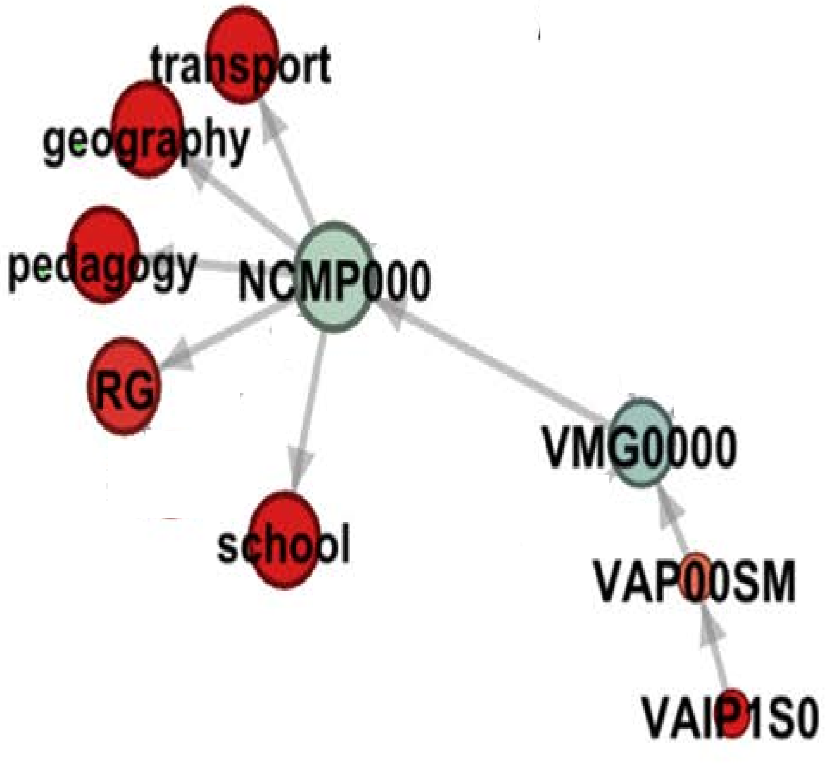
EmoGraph

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



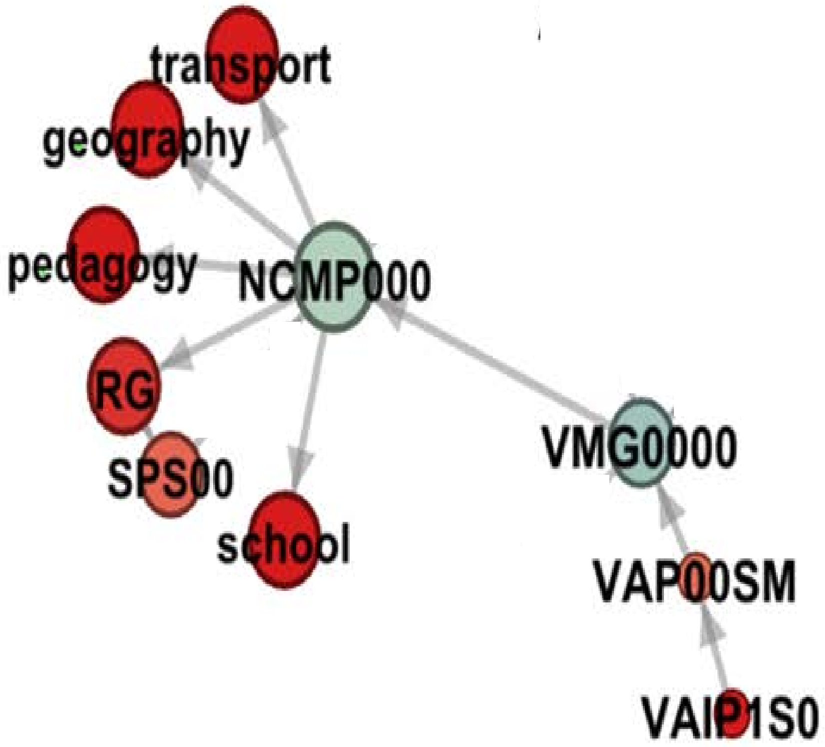
EmoGraph

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



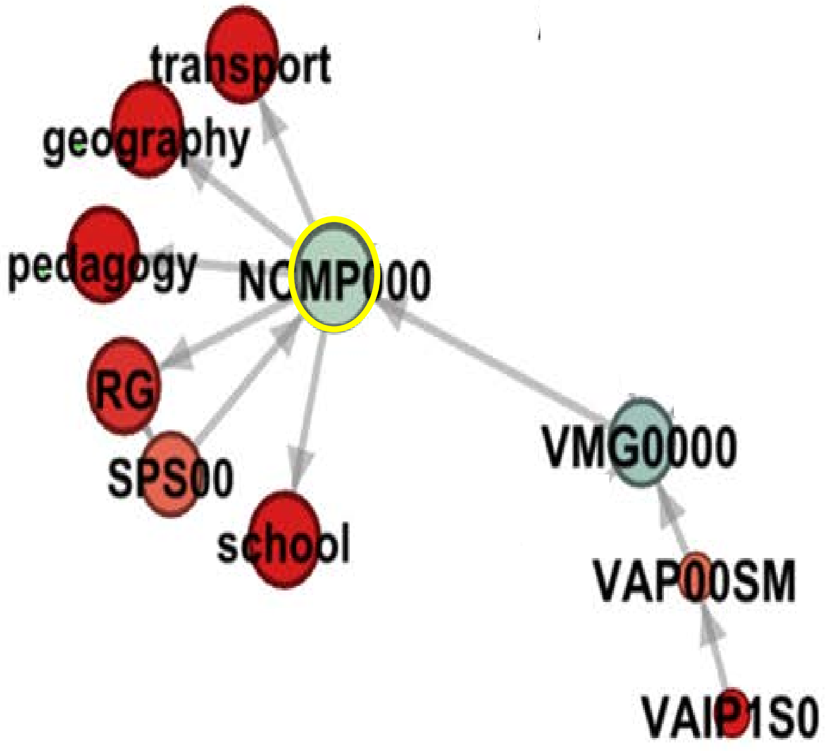
EmoGraph

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



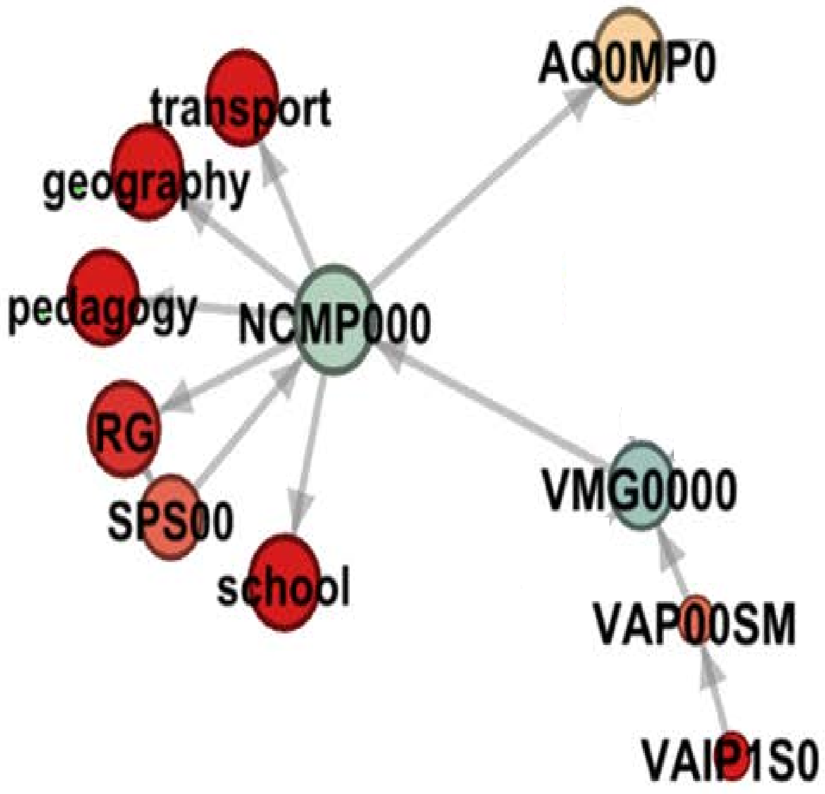
EmoGraph

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



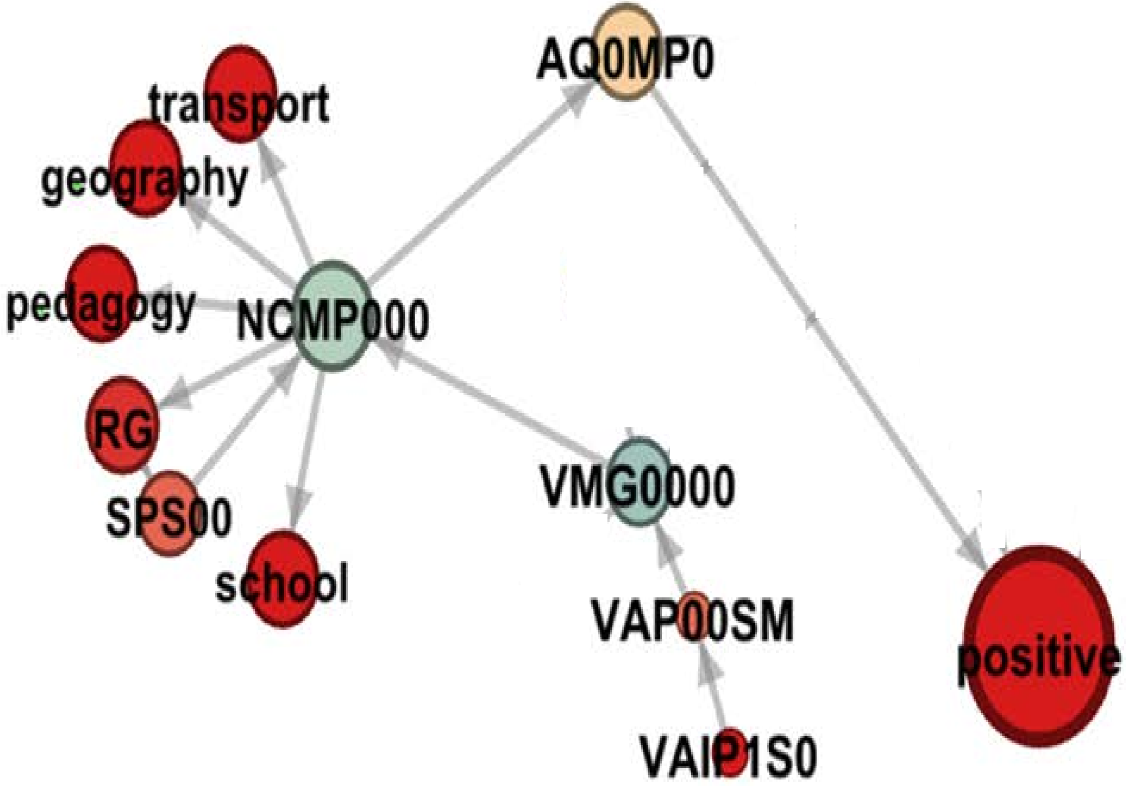
EmoGraph

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



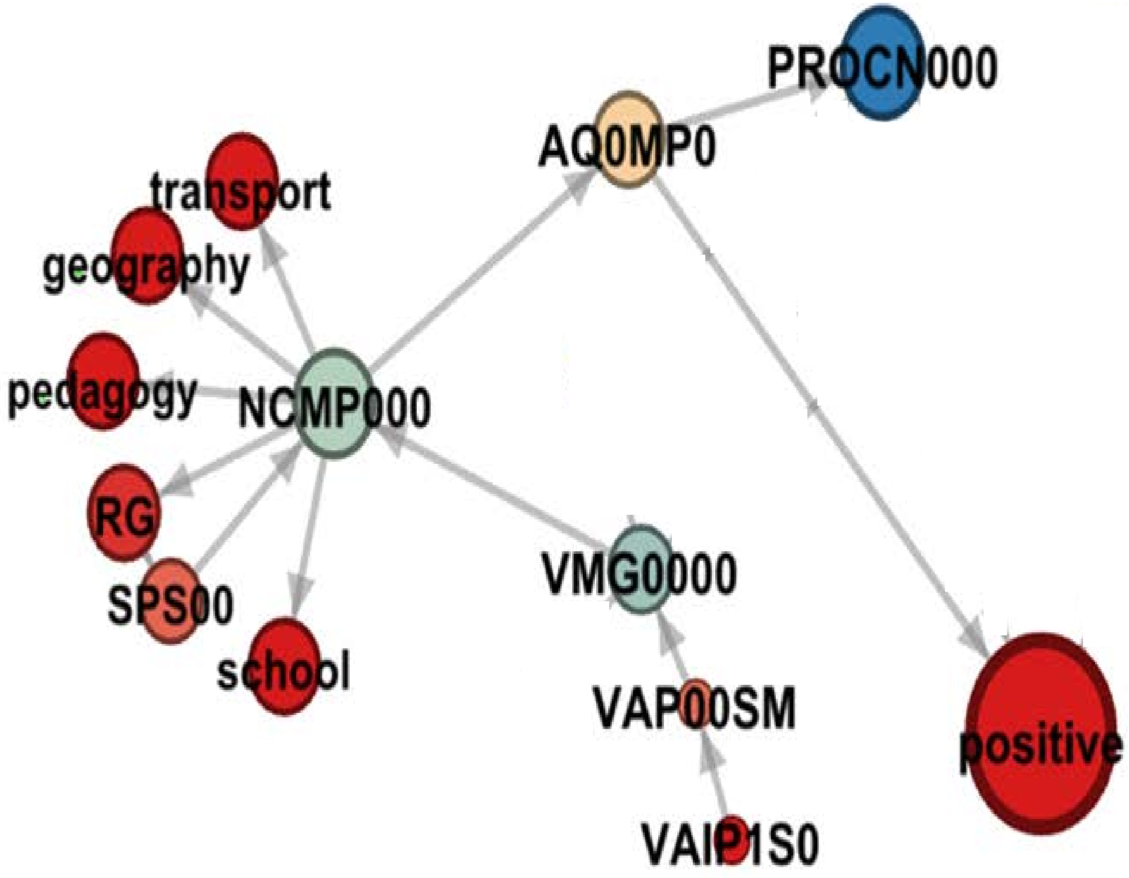
EmoGraph

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



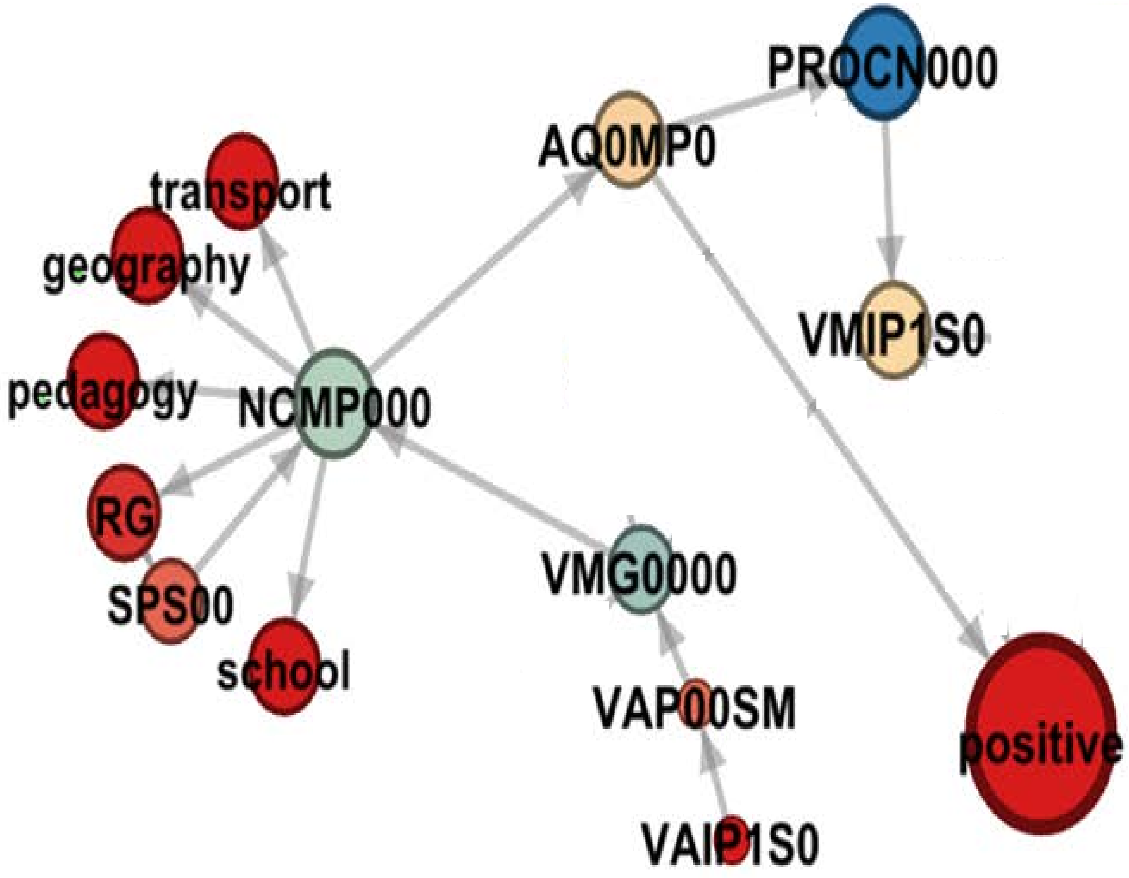
EmoGraph

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



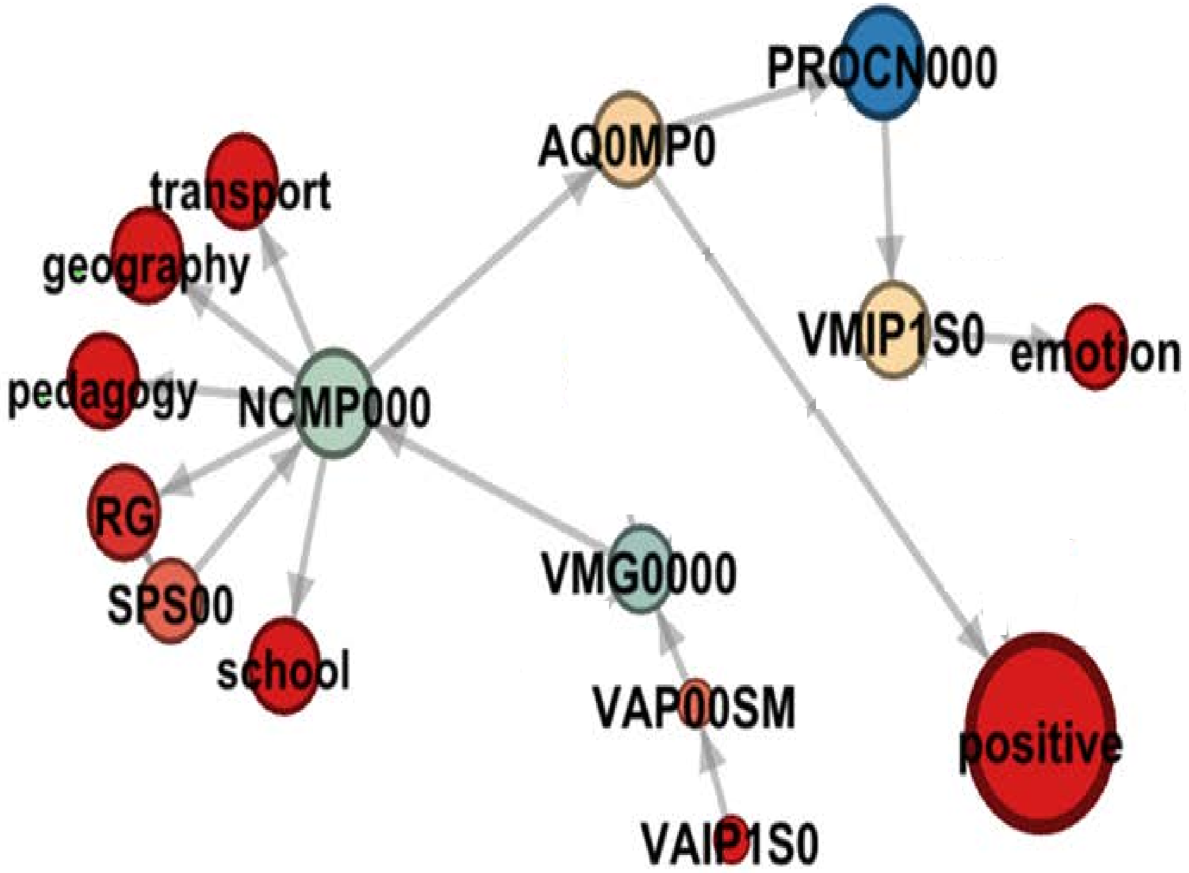
EmoGraph

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



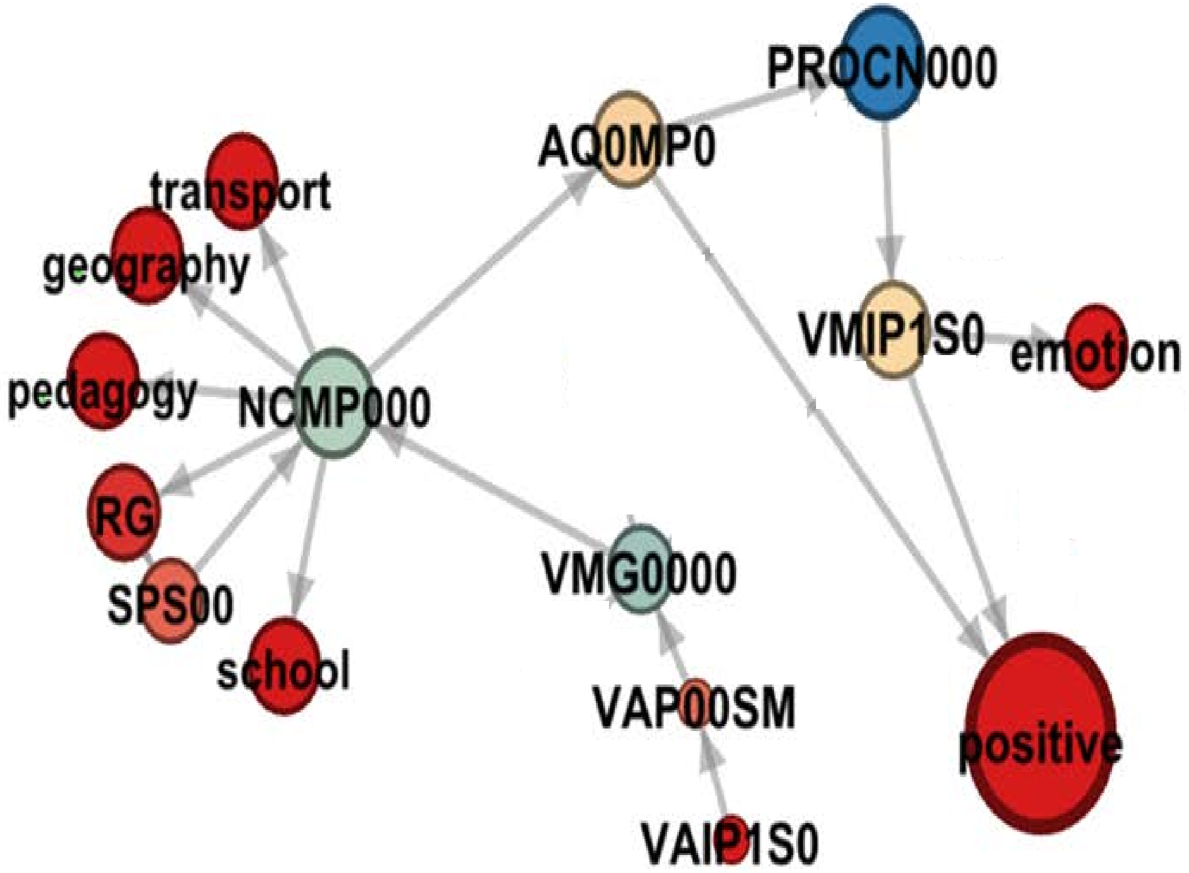
EmoGraph

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



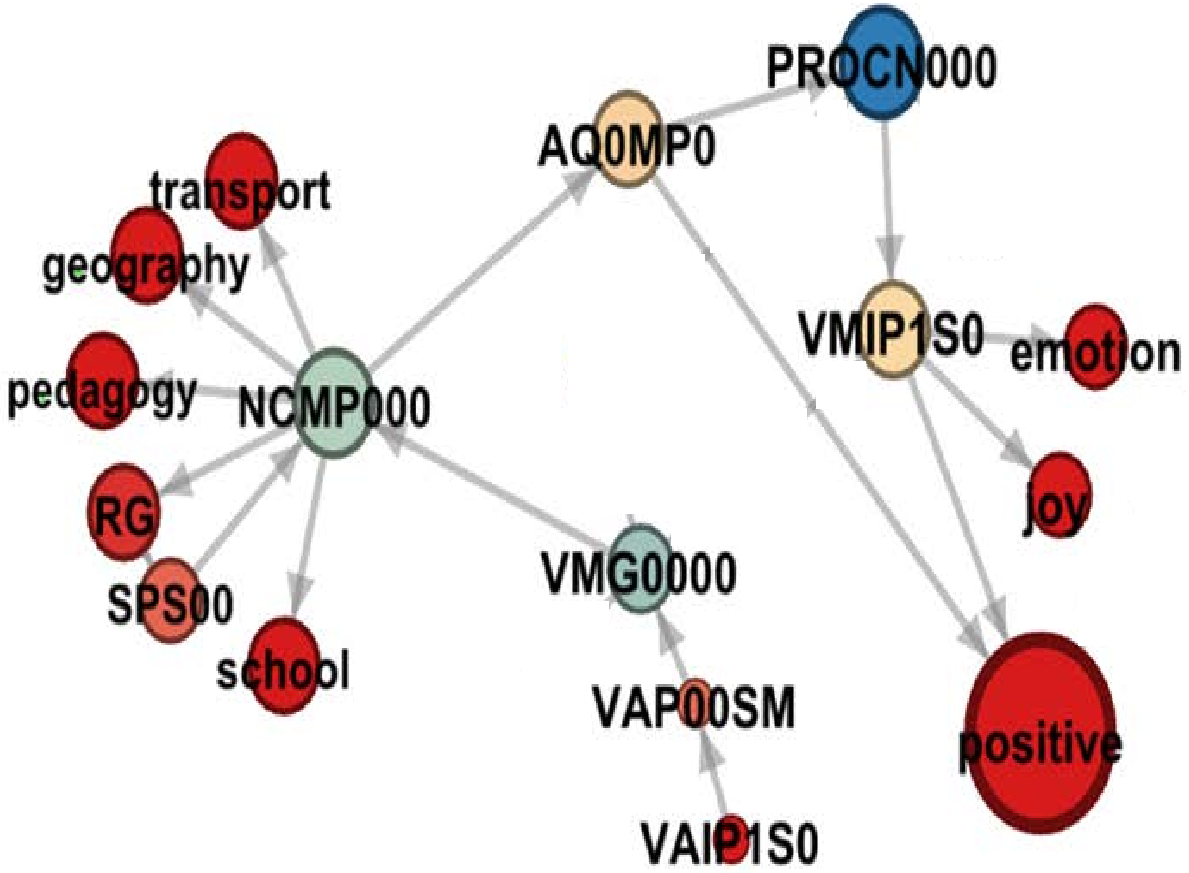
EmoGraph

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



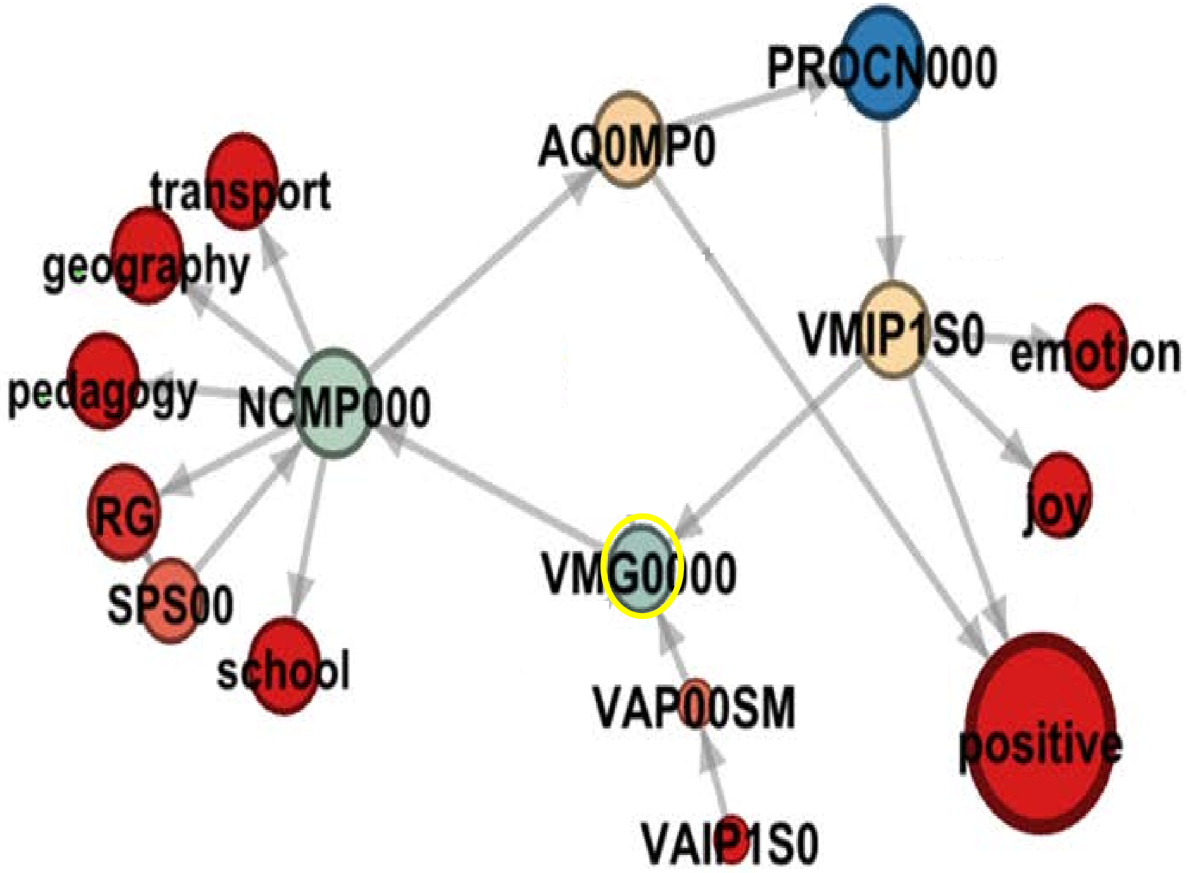
EmoGraph

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



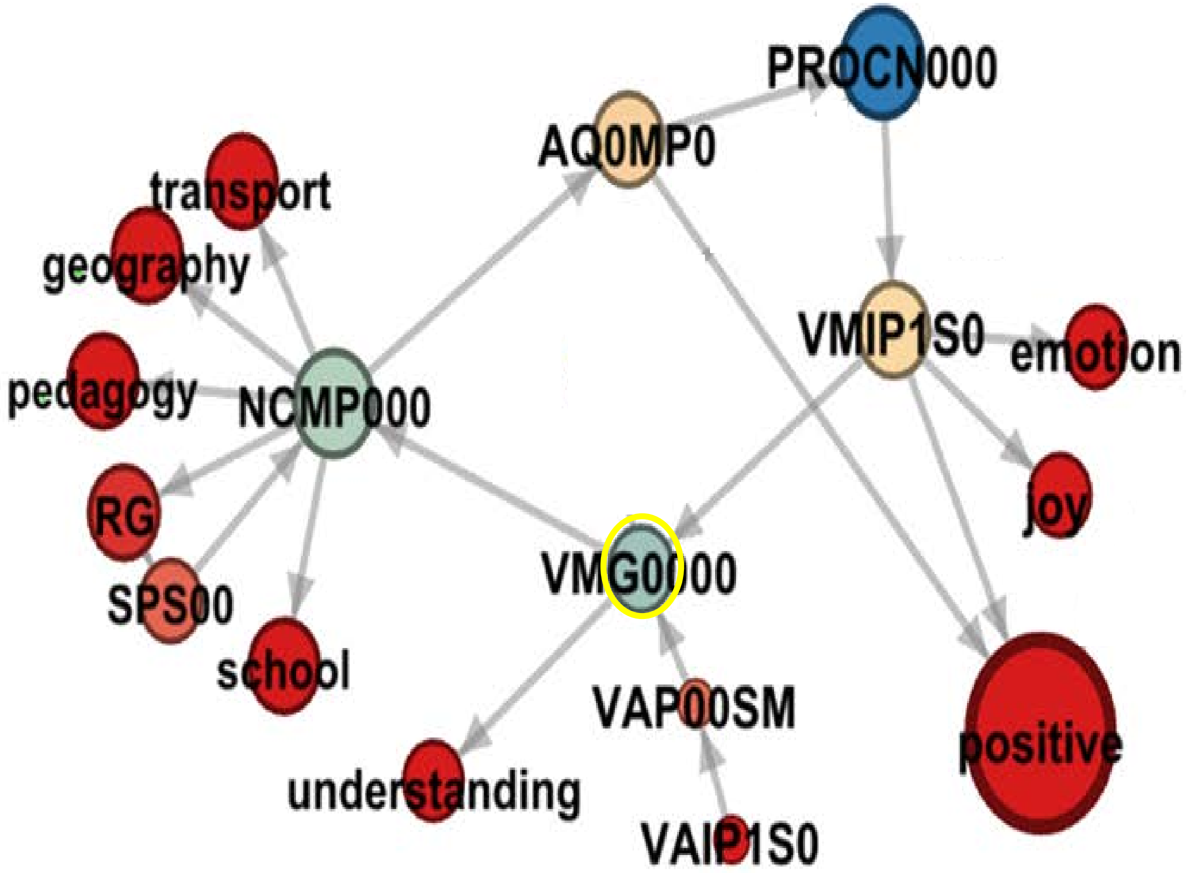
EmoGraph

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



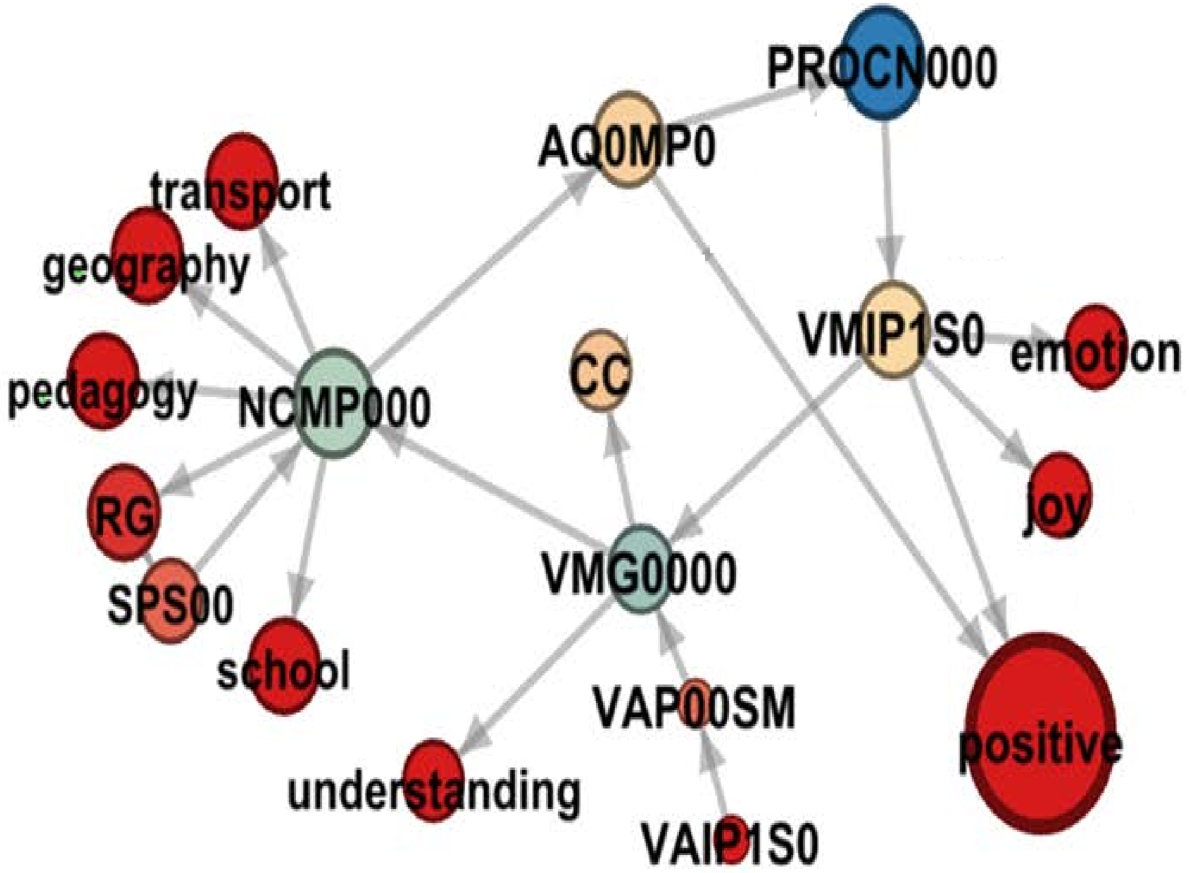
EmoGraph

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



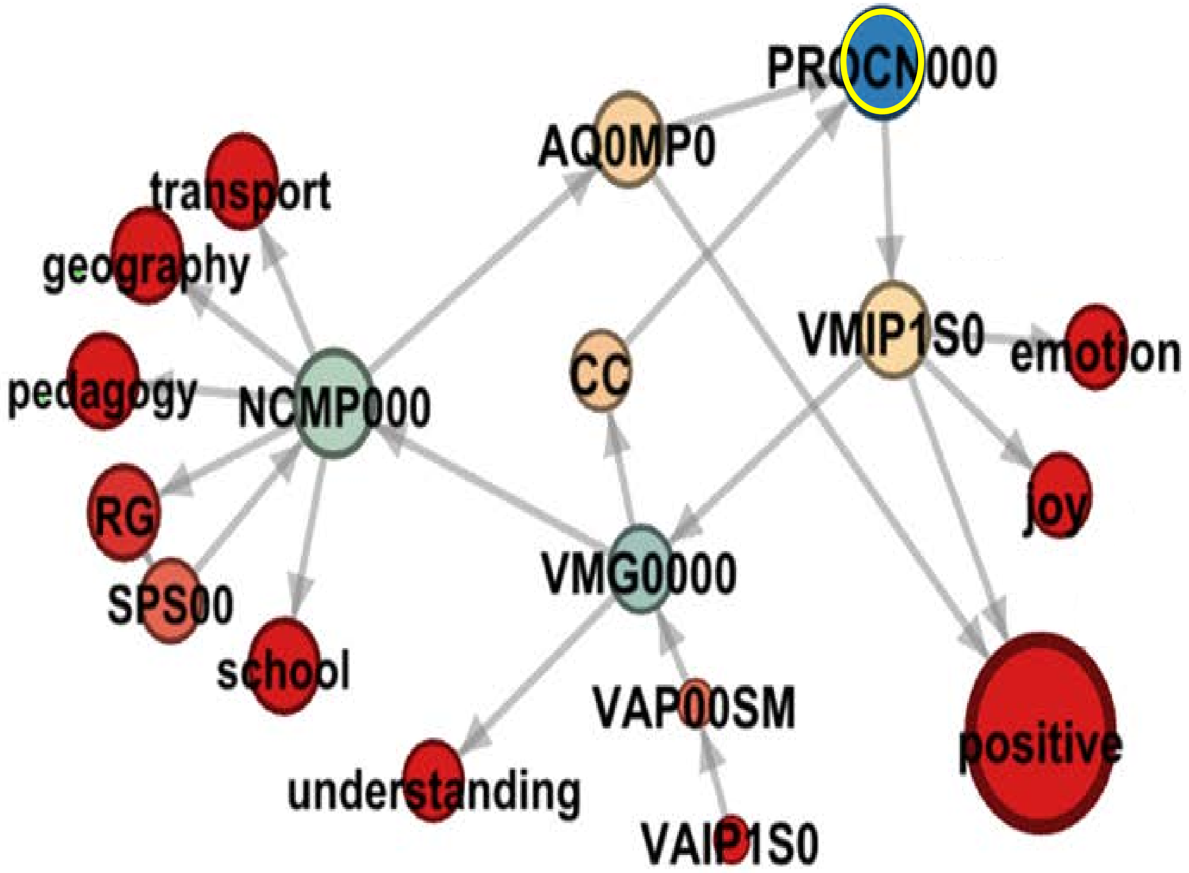
EmoGraph

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



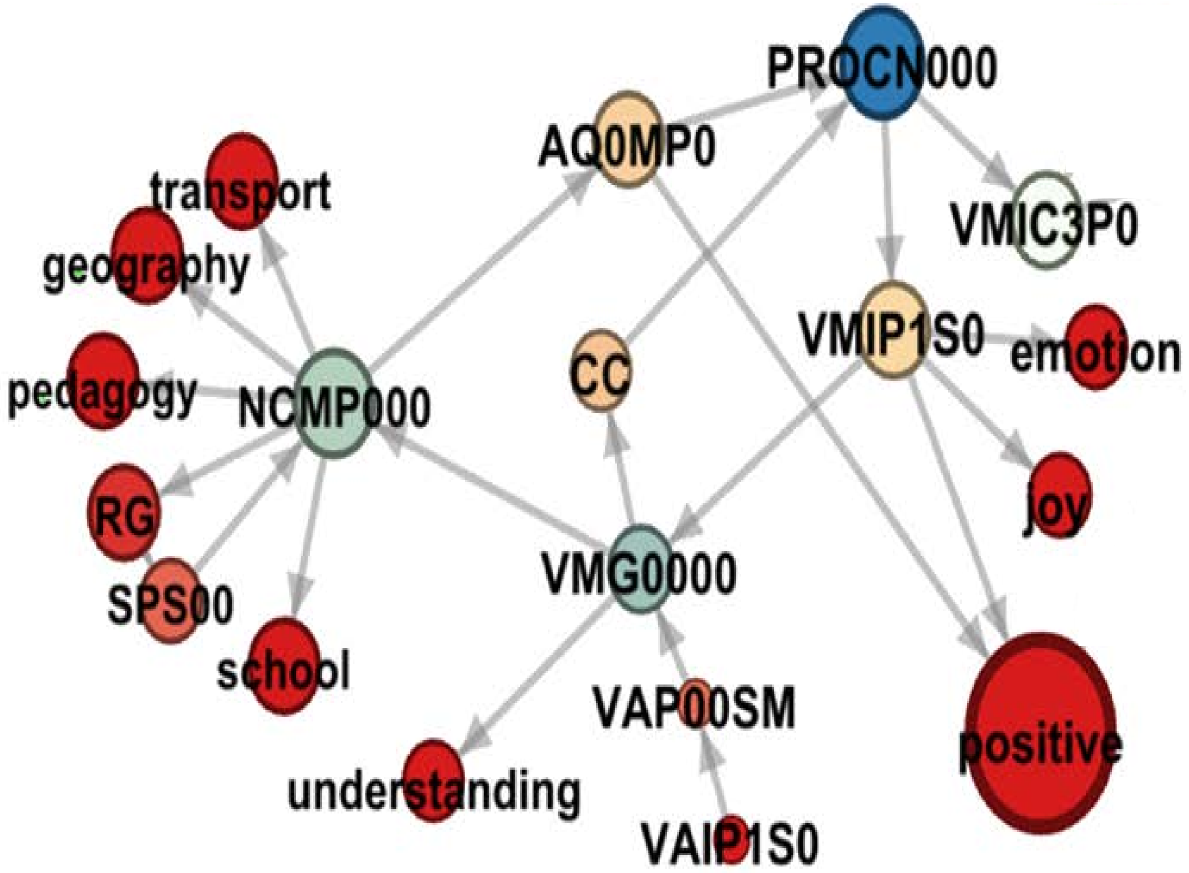
EmoGraph

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



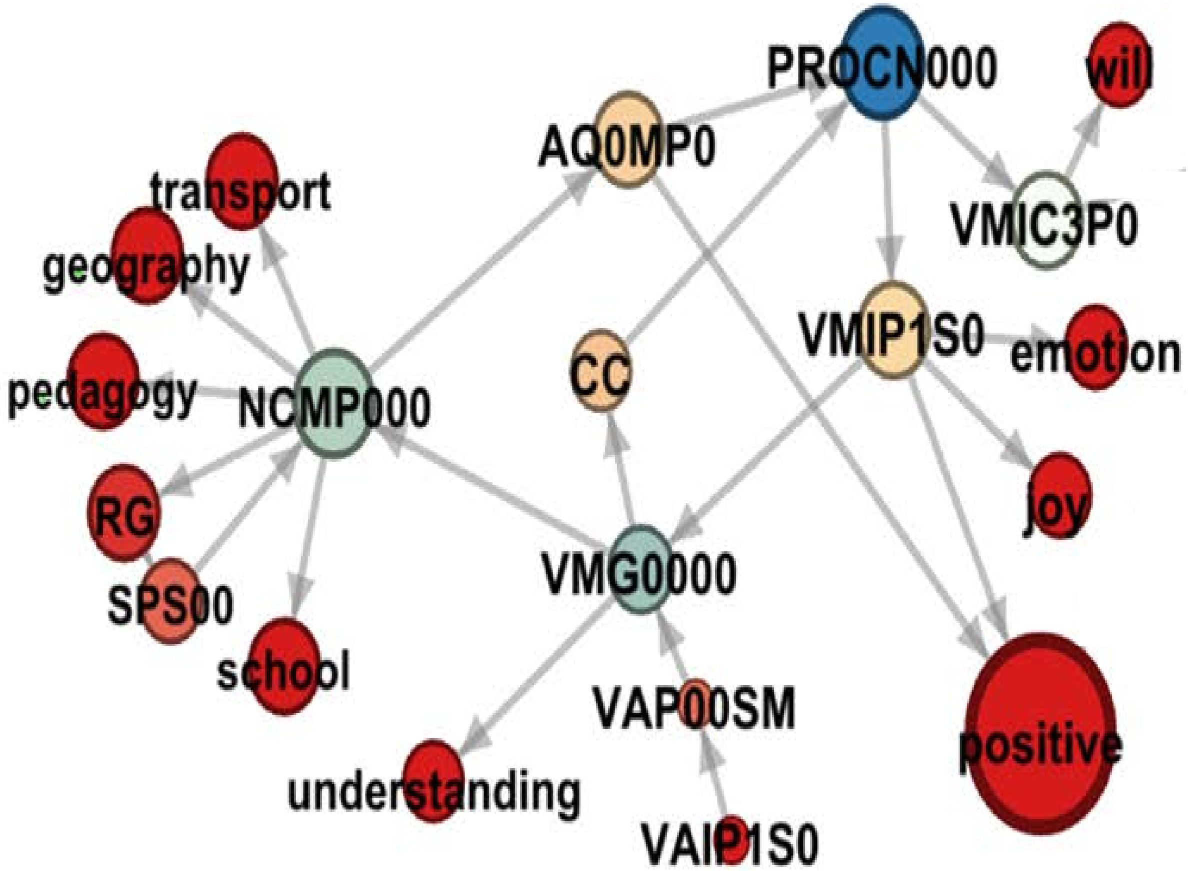
EmoGraph

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



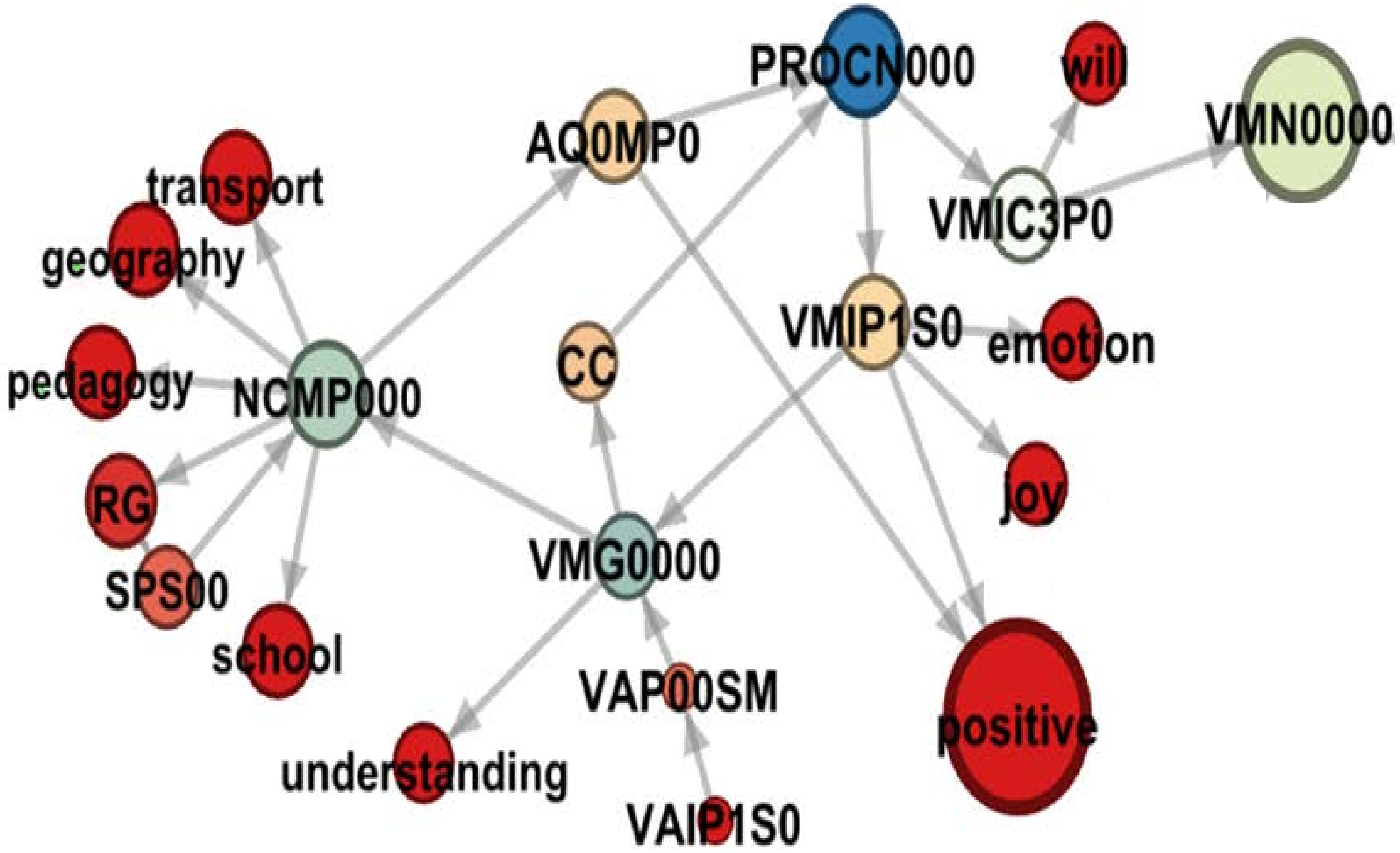
EmoGraph

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



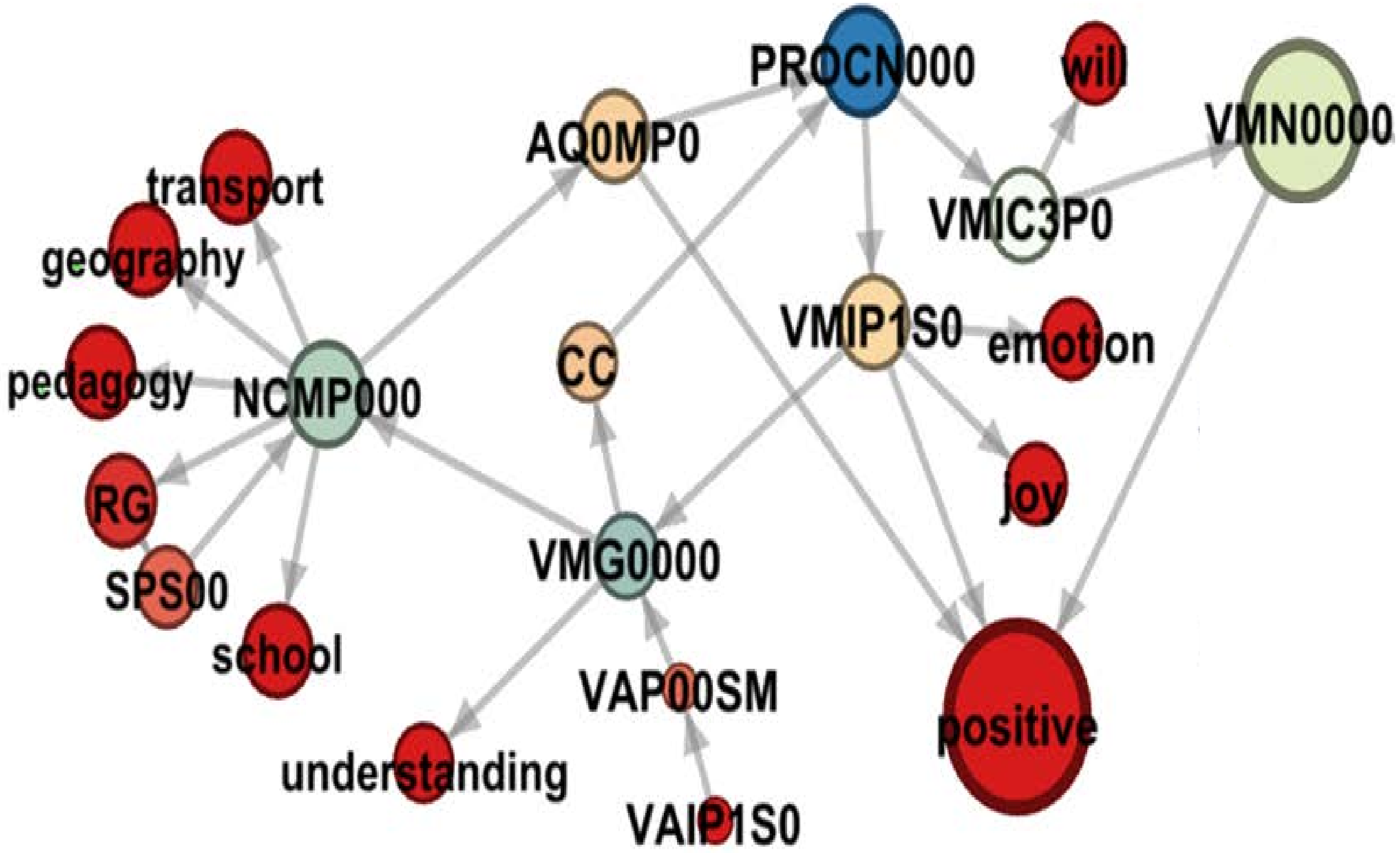
EmoGraph

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



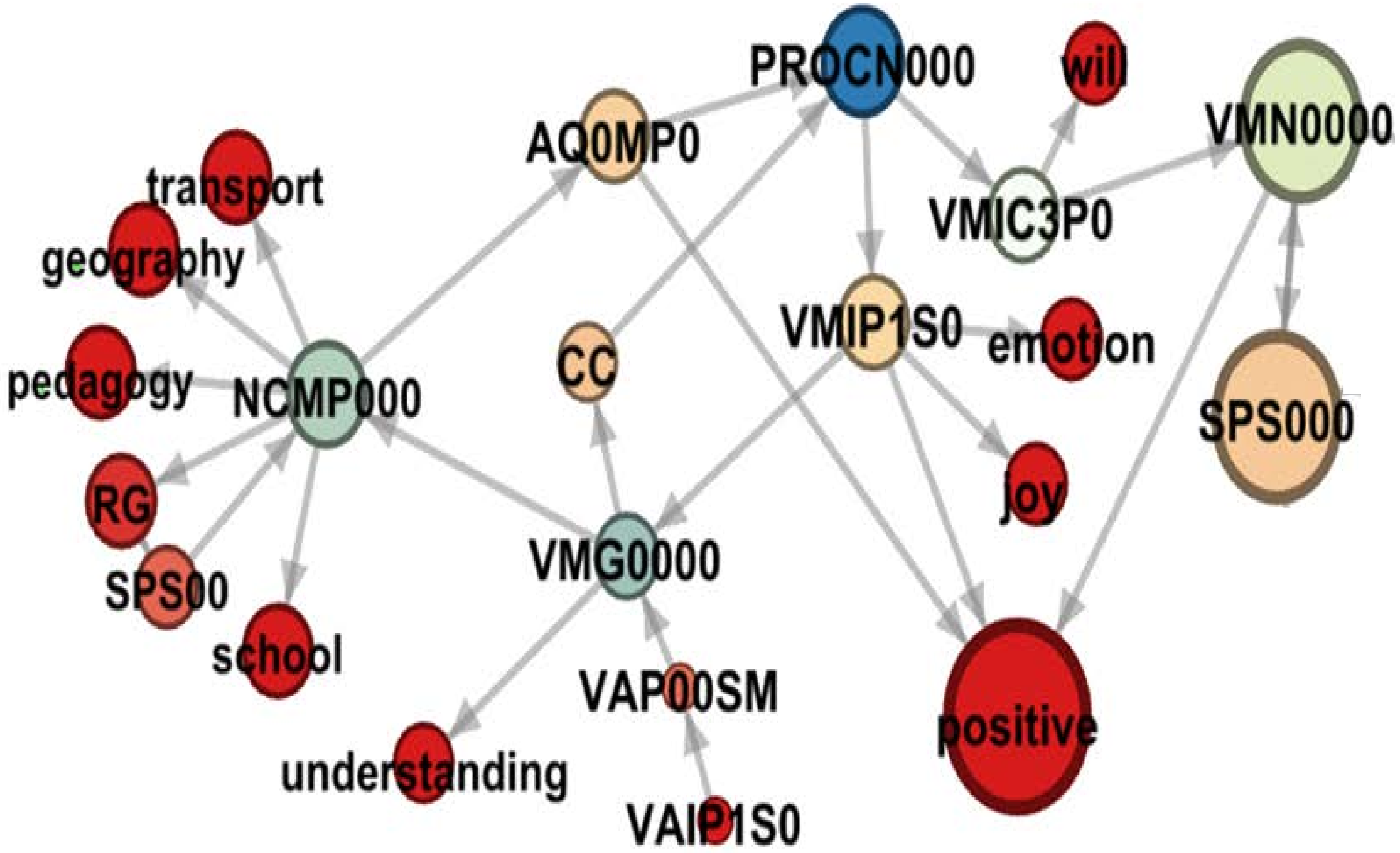
EmoGraph

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



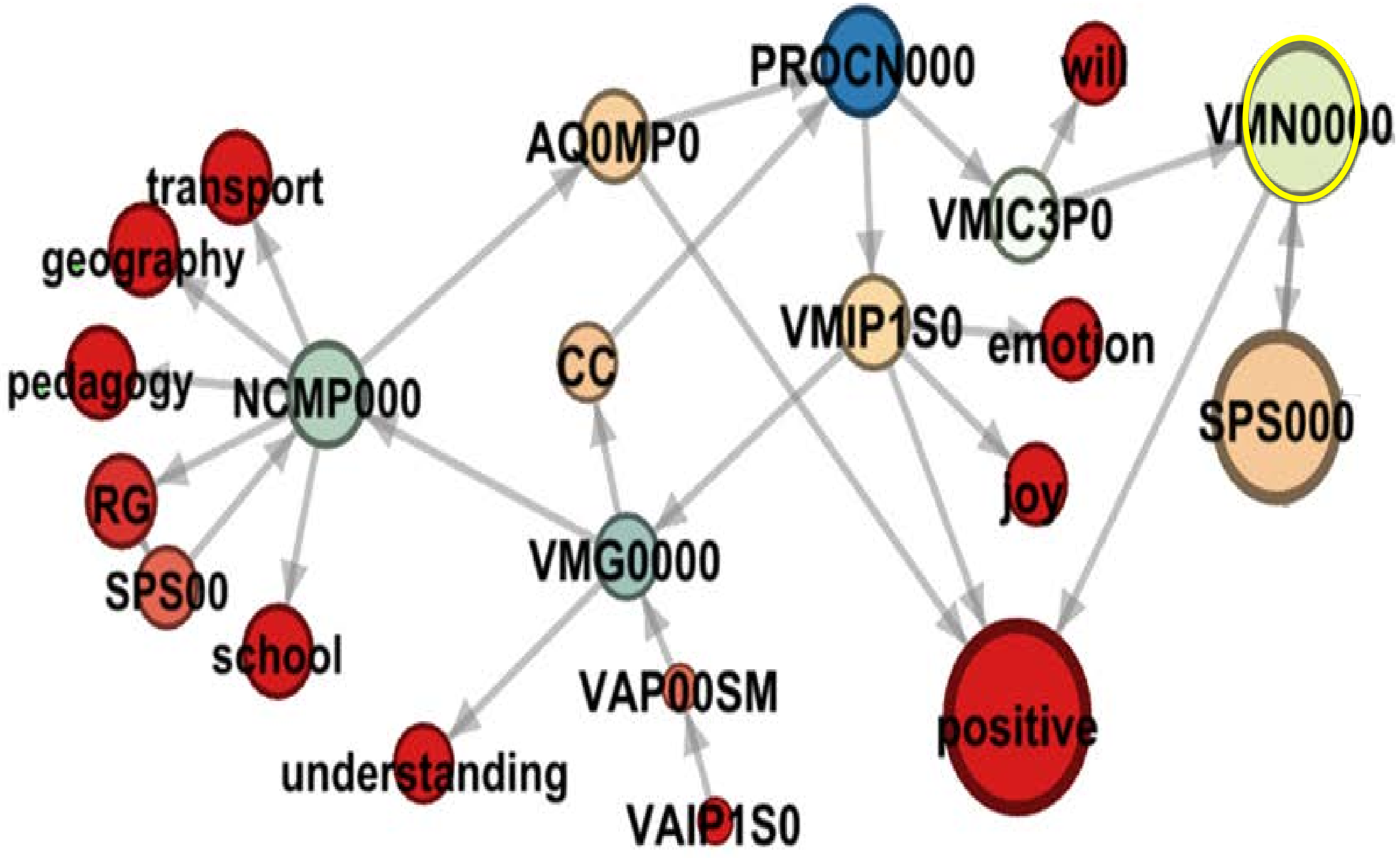
EmoGraph

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



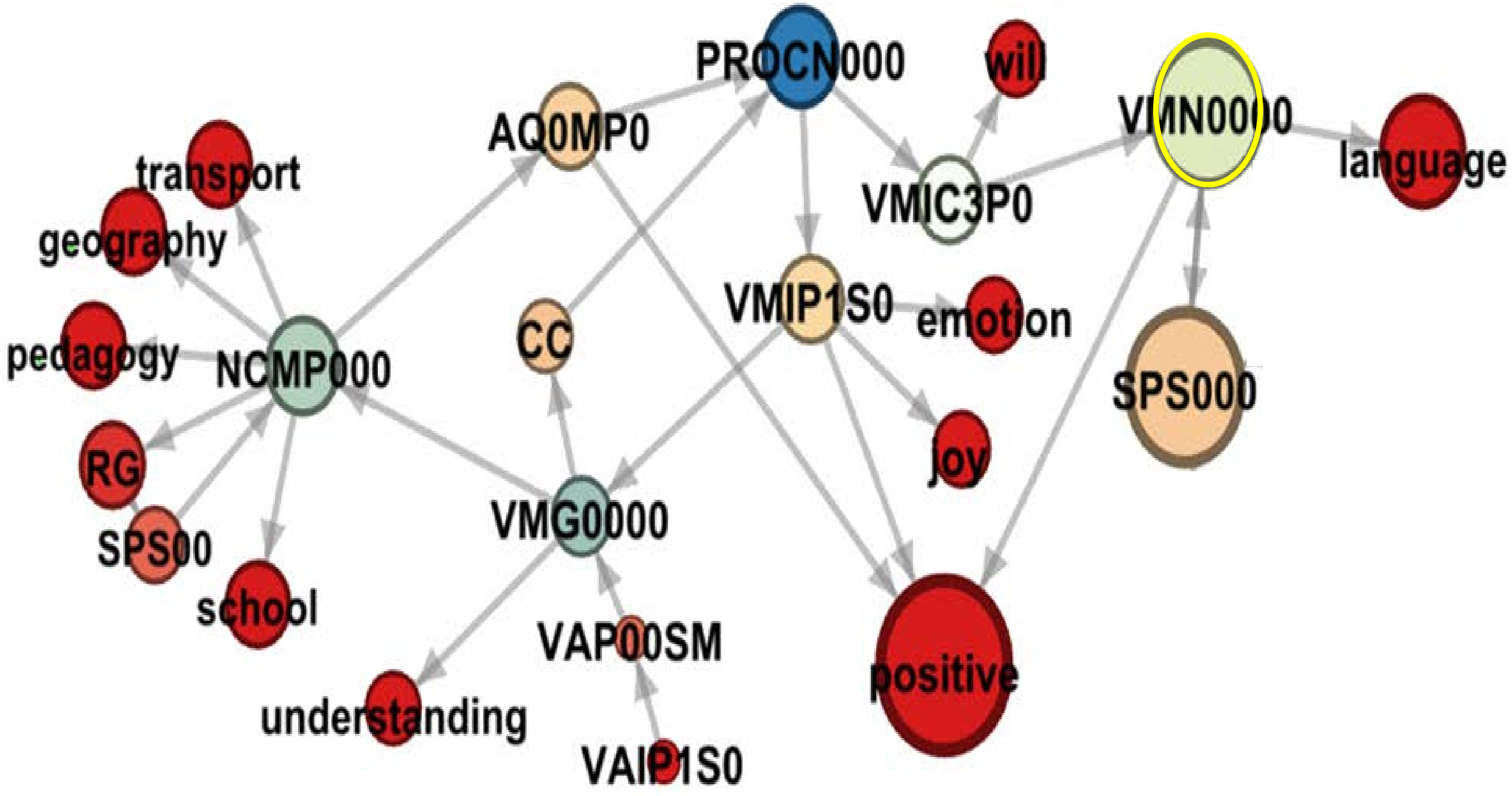
EmoGraph

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



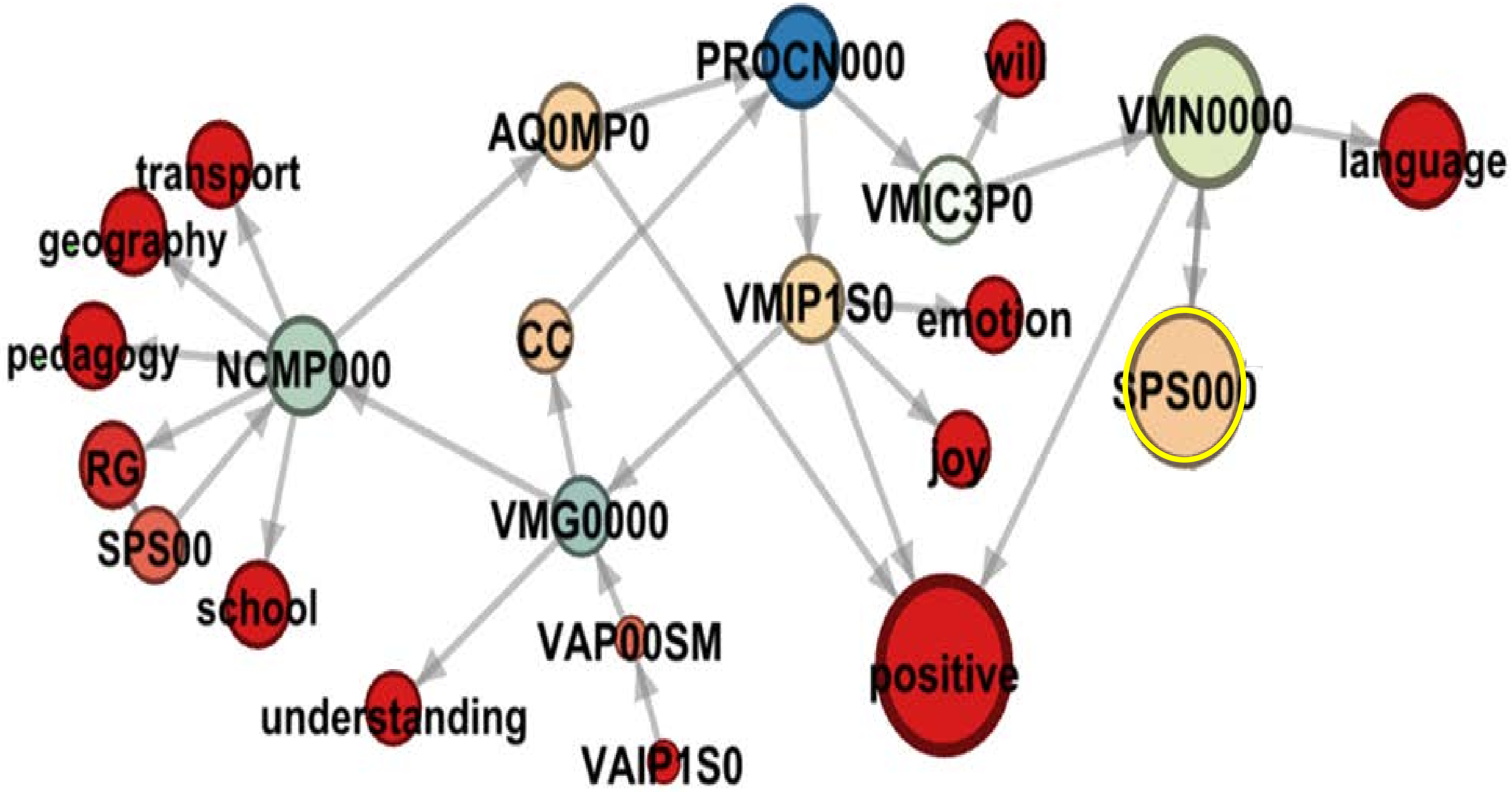
EmoGraph

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



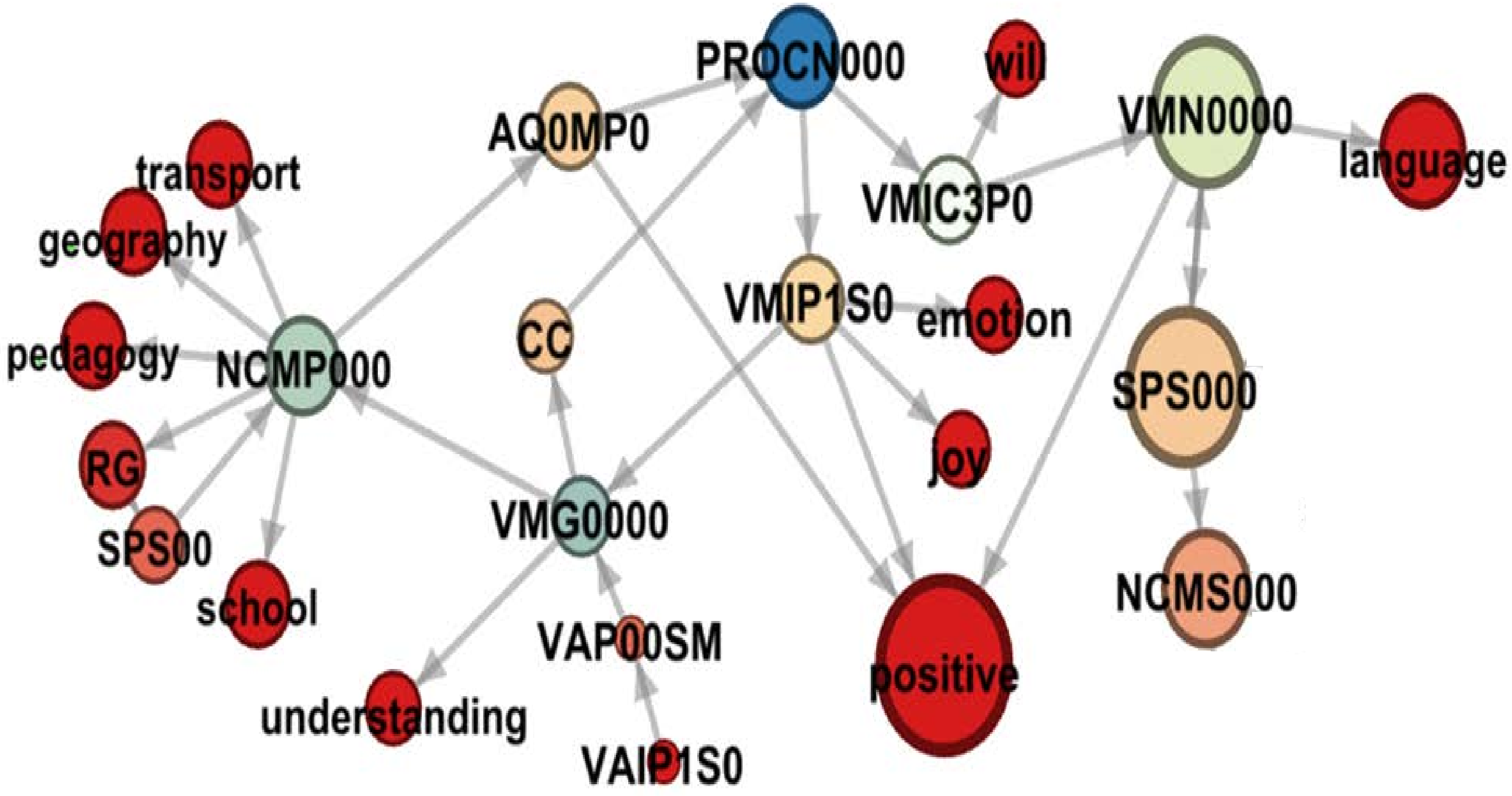
EmoGraph

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



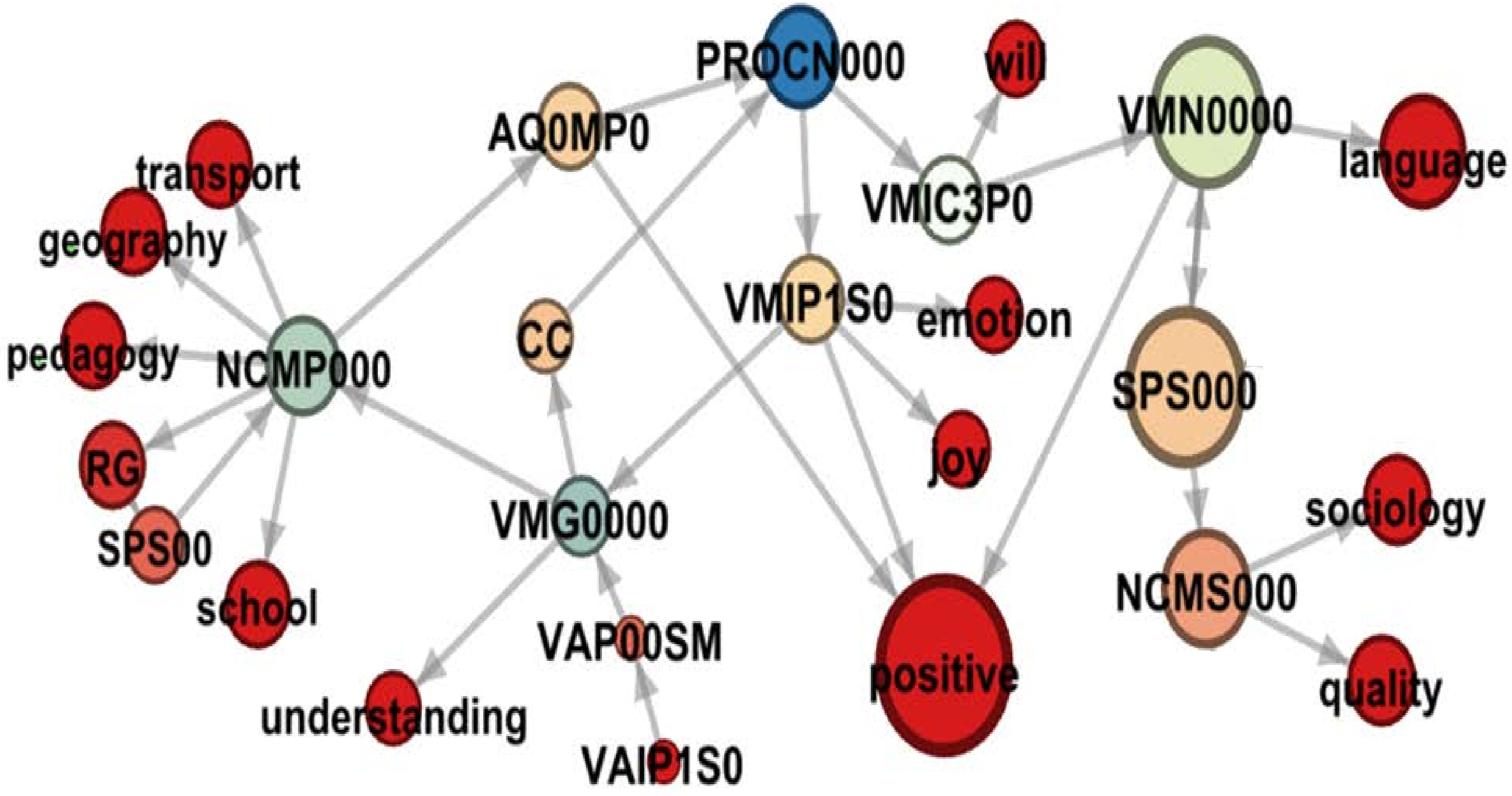
EmoGraph

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



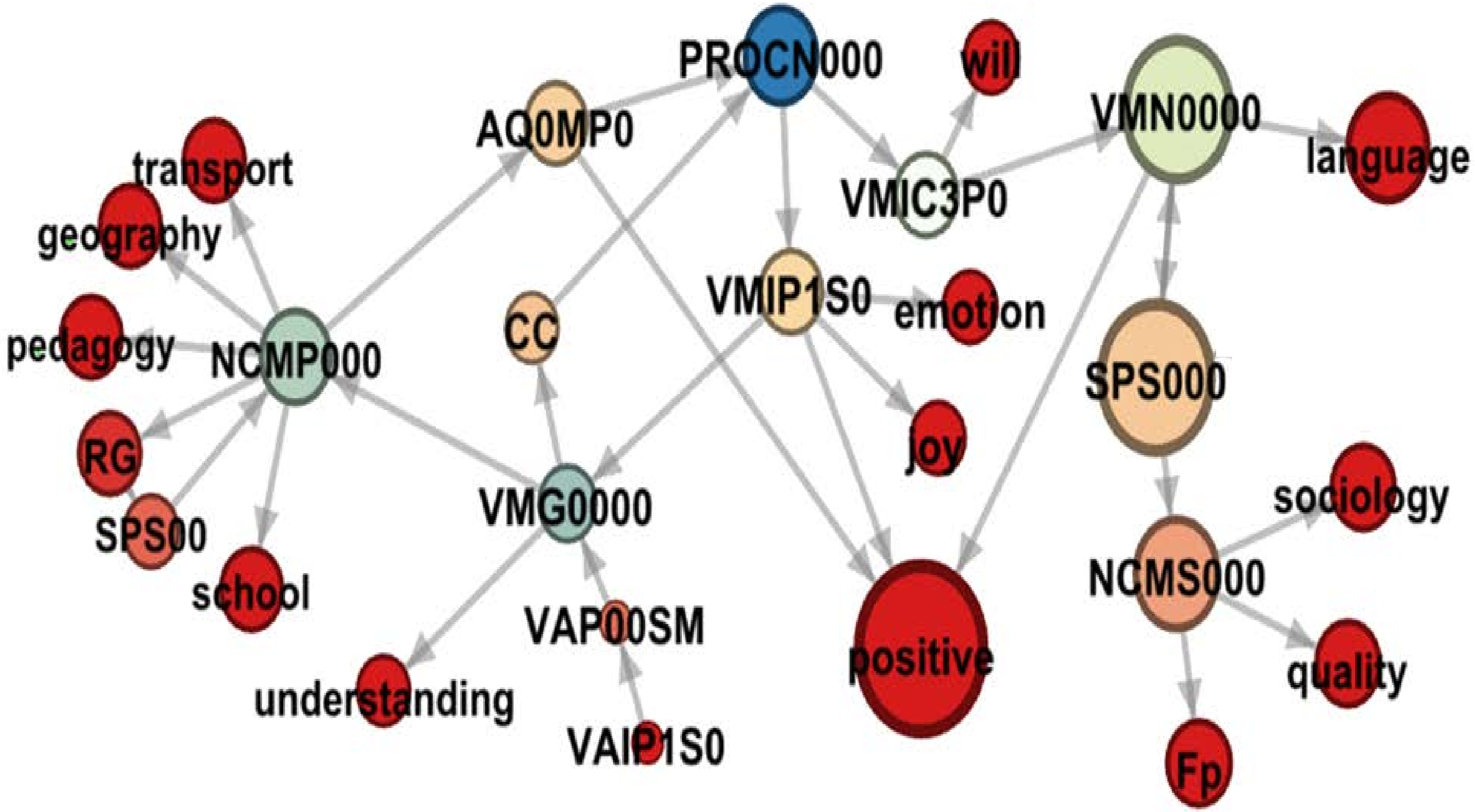
EmoGraph

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

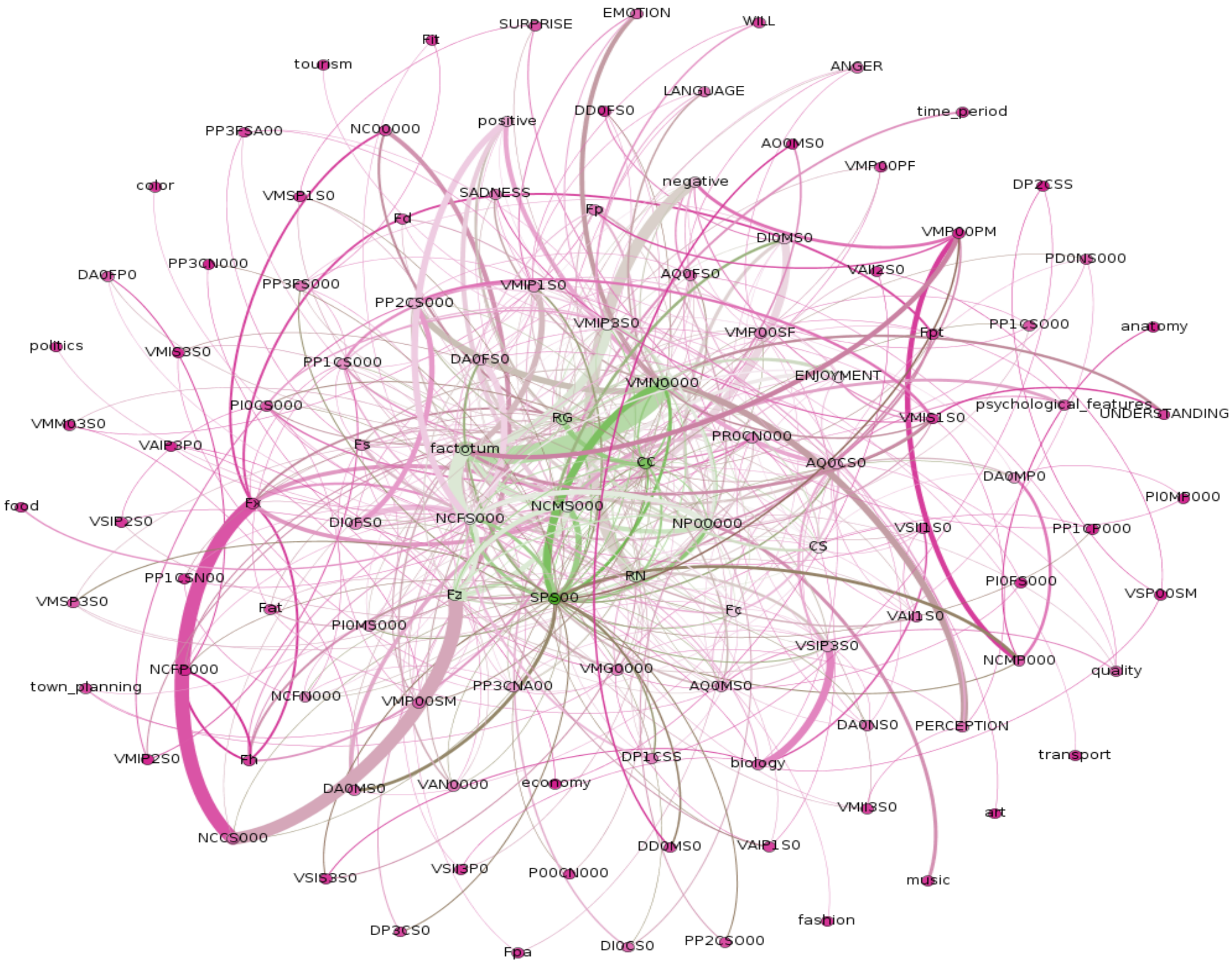


EmoGraph

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



EmoGraph: author’s sentences



Graph-based Features

Given a graph G={N,E} where:

- N is the set of nodes
- E is the set of edges

we obtain a set of:

- **structure-based** features from global measures of the graph
- **node-based** features from node specific measures

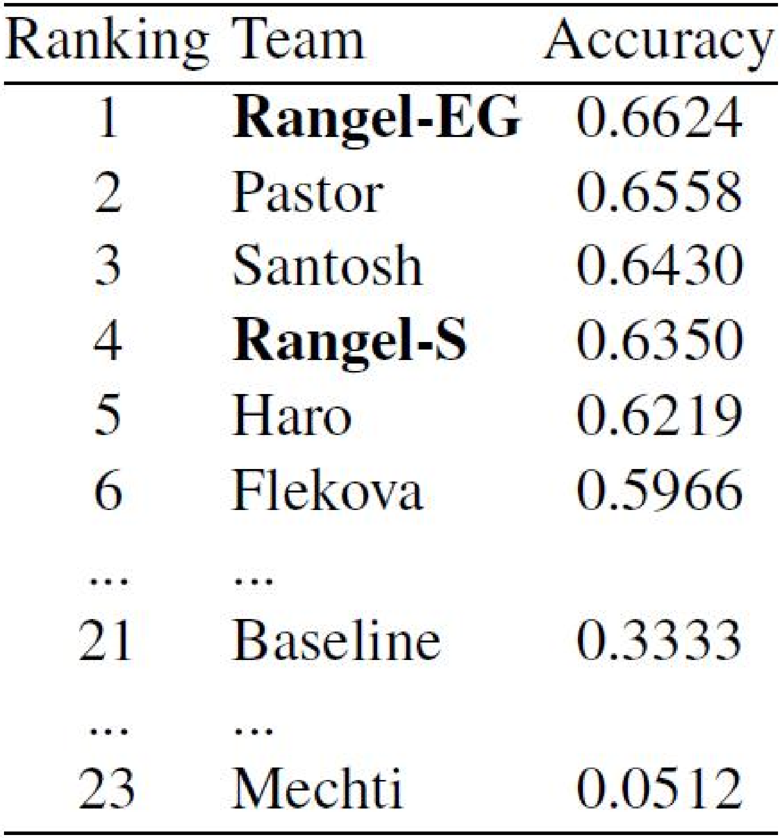
Structure-based Features

Nodes-edges ratio	Indicator of how connected the graph is, i.e., how complicated the discourse is	Theoretical maximum: $max(E) = N * (N - 1)$
Weighted average degree	Indicator of how much interconnected the graph is, i.e., how much interconnected the grammatical categories are	Averaging all nodes degrees. Scaling it to [0,1]
Diameter	Indicator of the greatest distance between any pair of nodes, i.e, how far a grammatical category is from others, or how far a topic is from an emotion	$d = max_{n \in N} \varepsilon(N)$ where E(N) is the eccentricity
Density	Indicator of how close the graph is to be complete, i.e., how dense is the text in the sense of how each grammatical category is used in combination with others	$D = \frac{2 * \vert E \vert}{(\vert N \vert * (\vert N \vert - 1))}$
Modularity	Indicator of different divisions of the graph into modules (one node has dense connections within the module and sparse with nodes in other modules), i.e., how the discourse is modeled in different structural or stylistic units	Blondel,V.D.,Guillaume,J.L.,Lambiotte,R.,Lefebvre,E. Fast unfolding of communities in large networks. In: Journal of Statistical Mechanics: Theory and Experiment, vol. 2008 (10), pp. 10008 (2008)
Clustering coefficient	Indicator of the transitivity of the graph (if a is directly linked to b and b is directly linked to c, what's the probability that a node is directly linked to c), i.e., how different grammatical categories or semantic information are related to each other	Watts-Strogatzt: $cc1 = \frac{\sum_{i=1}^{n} C(i)}{n}$
Average path length	Indicator of how far some nodes are from others, i.e., how far some grammatical categories are from others, or some topics are from some emotions	Brandes, U. A Faster Algorithm for Betweenness Centrality. In: Journal of Mathematical Sociology 25(2), pp. 163-177 (2001)

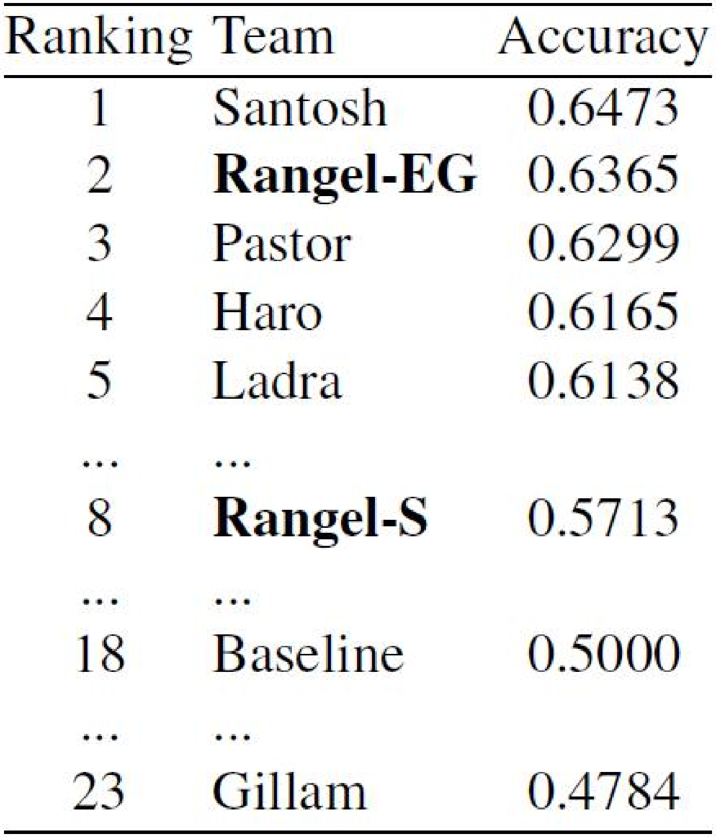
Node-based Features

EigenVector	It gives a measure of the influence of each node. In our case, it may give what are the grammatical categories with the most central use in the author’s discourse, e.g. which nouns, verbs or adjectives	Given a graph and its adjacency matrix $A = a_{n,t}$ where $a_{n,t}$ is 1 if a node n is linked to a node t, and 0 otherwise: $x_n = \frac{1}{\lambda}\sum_{t\in M(n)} x_t = \frac{1}{\lambda}\sum_{t\in G} a_{n,t}x_t$ where λ is a constant representing the greatest eigenvalue associated with the centrality measure.
Betweenness	It gives a measure of the importance of a each node depending on the number of shortest paths of which it is part of. In our case, if one node has a high betweenness centrality means that it is a common element used for link among parts-of-speech, e.g. prepositions, conjunctions or even verbs and nouns. Hence, this measure may give us an indicator of what the most common connectors in the linguistic structures used by authors	It is the ratio of all shortest paths from one node to another node in the graph that pass through x: $BC(x) = \sum_{i,j\in N-\{n\}} \frac{\sigma_{i,j}(n)}{\sigma_{i,j}}$ Where $\sigma_{i,j}$ is the total number of shortest paths from node i to j, and $\sigma_{i,j}(n)$ is the total number of those paths that pass through n.

PAN 2013 - Style-based Approach (S) vs EmoGraph (EG): S+Emotion-graph



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

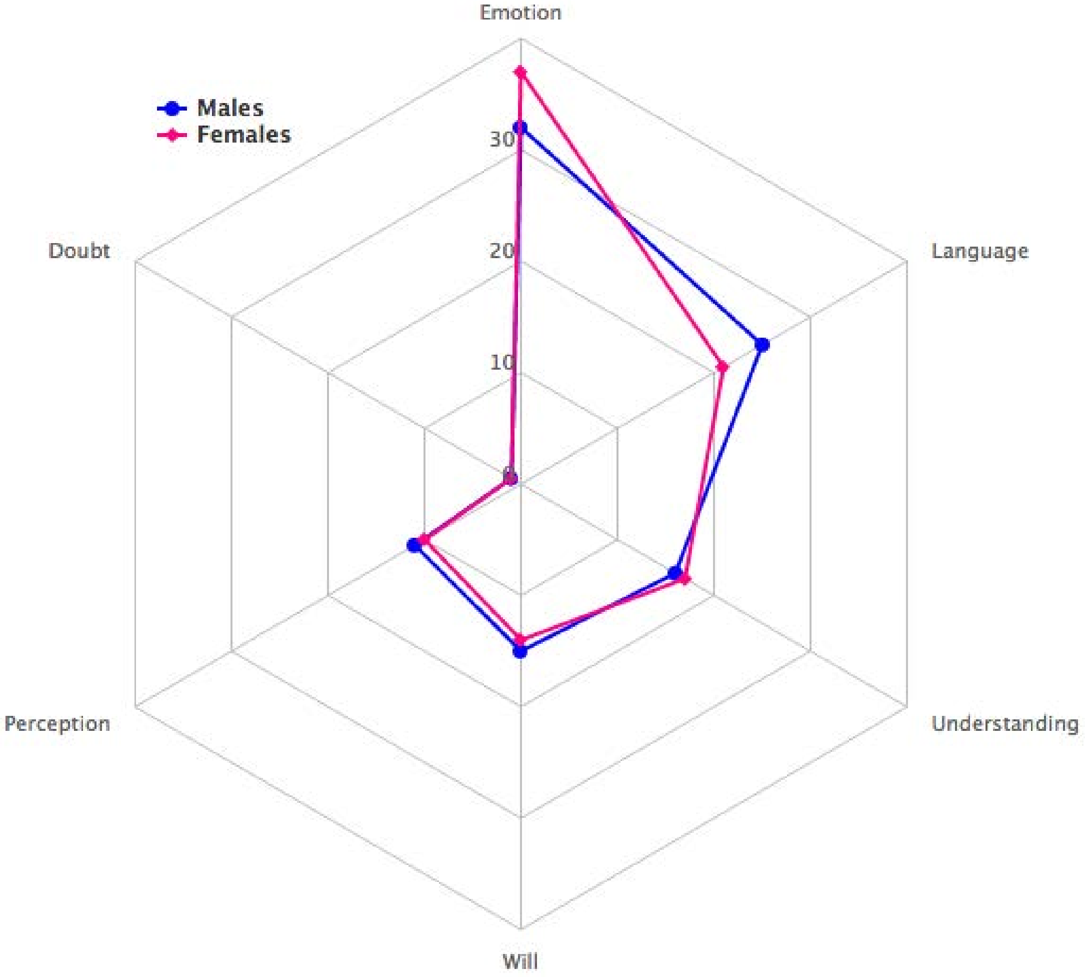


Ranking	Team	Accuracy
1	Santosh	0.6473
2	**Rangel-EG**	0.6365
3	Pastor	0.6299
4	Haro	0.6165
5	Ladra	0.6138
...	...	
8	**Rangel-S**	0.5713
...	...	
18	Baseline	0.5000
...	...	
23	Gillam	0.4784

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

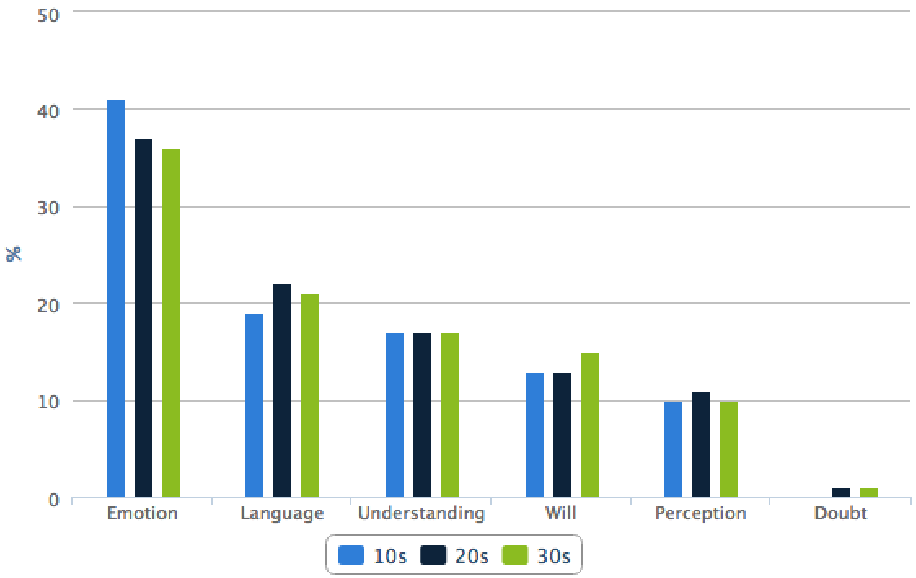
Verbs per Gender

- **Emotion**: feel, love, want…
- **Language**: say, tell, speak…
- **Understanding**: know, think, understand…
- **Perception**: see, listen…
- **Will**: must, forbid, allow…

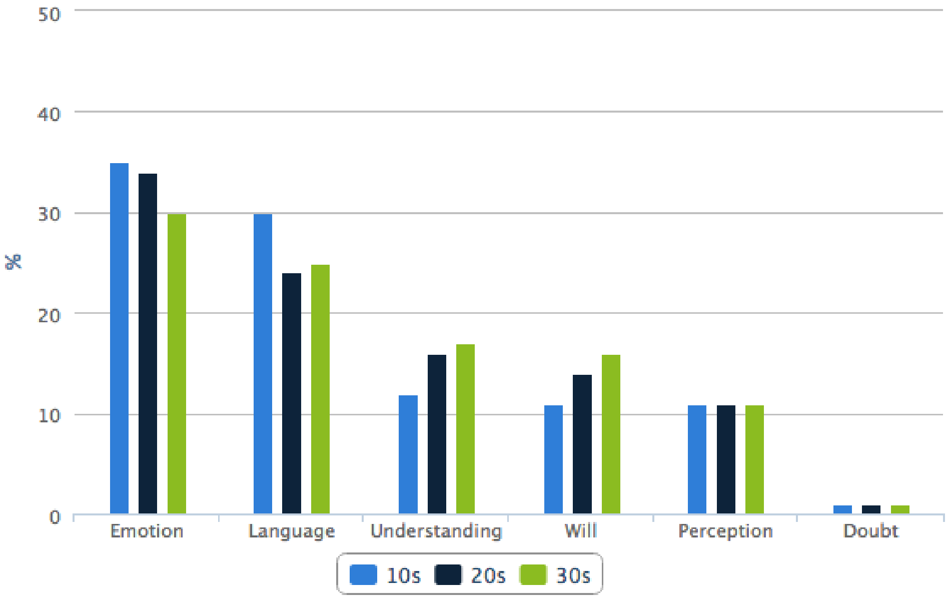


B. Levin. English Verb Classes and Alternations. University of Chicago Press, Chicago, 1993.

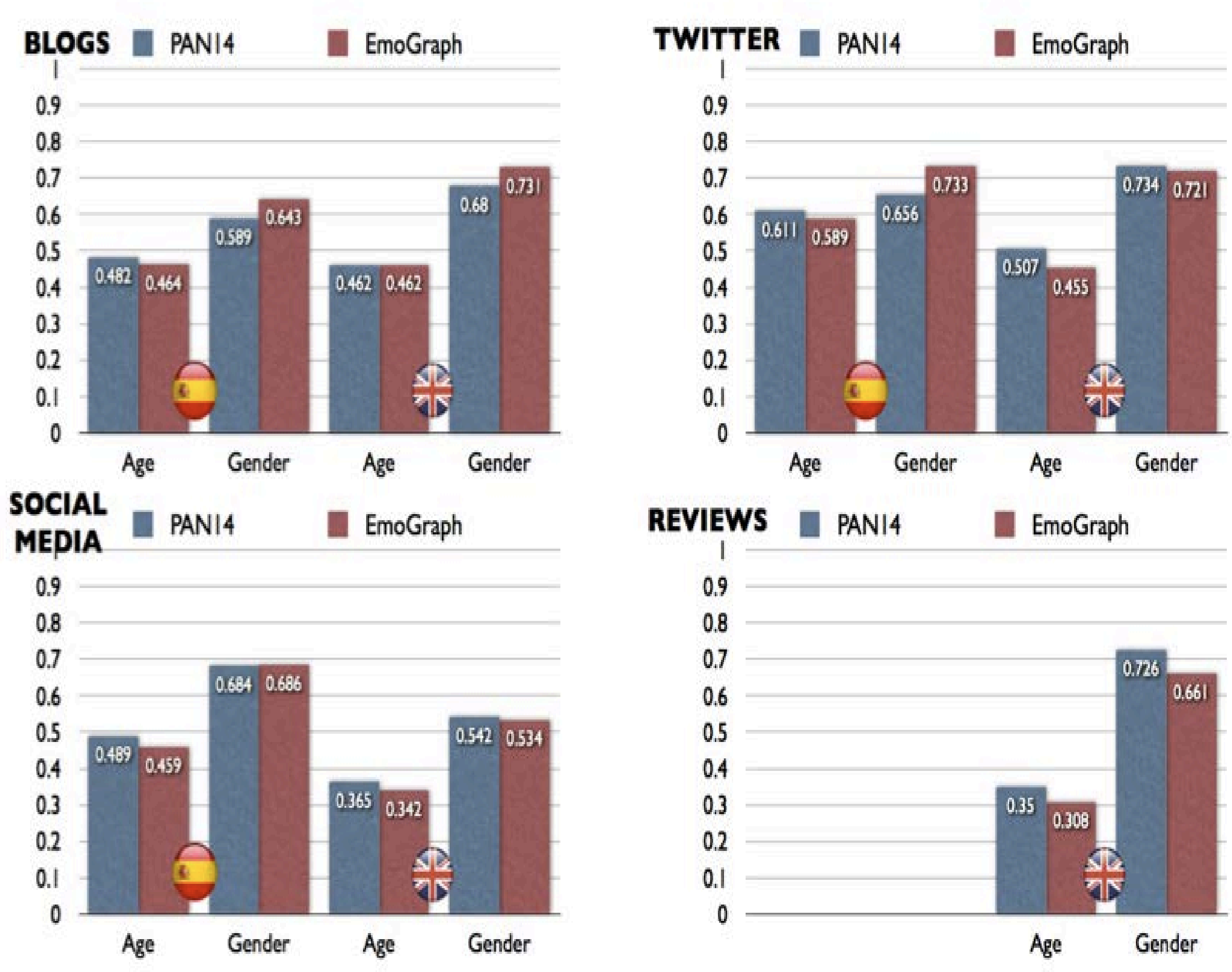
Verbs per Gender & Age



Females vs. Males



PAN 2014 - EmoGraph vs best approach



Accuracies of the best PAN14 team vs. EmoGraph on different languages and genres.

Conclusions and take-away message

Emotions may help profiling authors IF considered in the **discourse** structure

Danke / Merci / Grazie

Paolo Rosso: prosso@dsic.upv.es

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