Mineração de textos – revisão e exemplos de aplicações

Wagner de Lara Machado

PPGP PUCRS

Preparações pré-análise

- Arquivos em formato txt (outros formatos são suportados, porém o txt é o mas “fácil” e simples de trabalhar)
- Dividir o material em unidades que façam sentido (cada arquivo pode representar unidades diferentes de análise)
- Definir se o objeto da análise é a relação entre termos ou documentos (muito importante!)

- A função “Corpus” gera no RStudio seu espaço de análise
- A função “tm_map” (mappings) é importante para preparar o material textual para a análise
- Os usos mais comuns são:
  - Remover pontuações: tm_map(x,removePunctuation)
  - Converter as letras em minúsculas: tm_map(x,content_transformer(tolower))
  - Remover números: tm_map(x, removeNumbers)
  - Remover “stopwords”: tm_map(x, removeWords, stopwords("portuguese"))
  - “Stemming” (radical): tm_map(x,stemDocument,language = "portuguese")

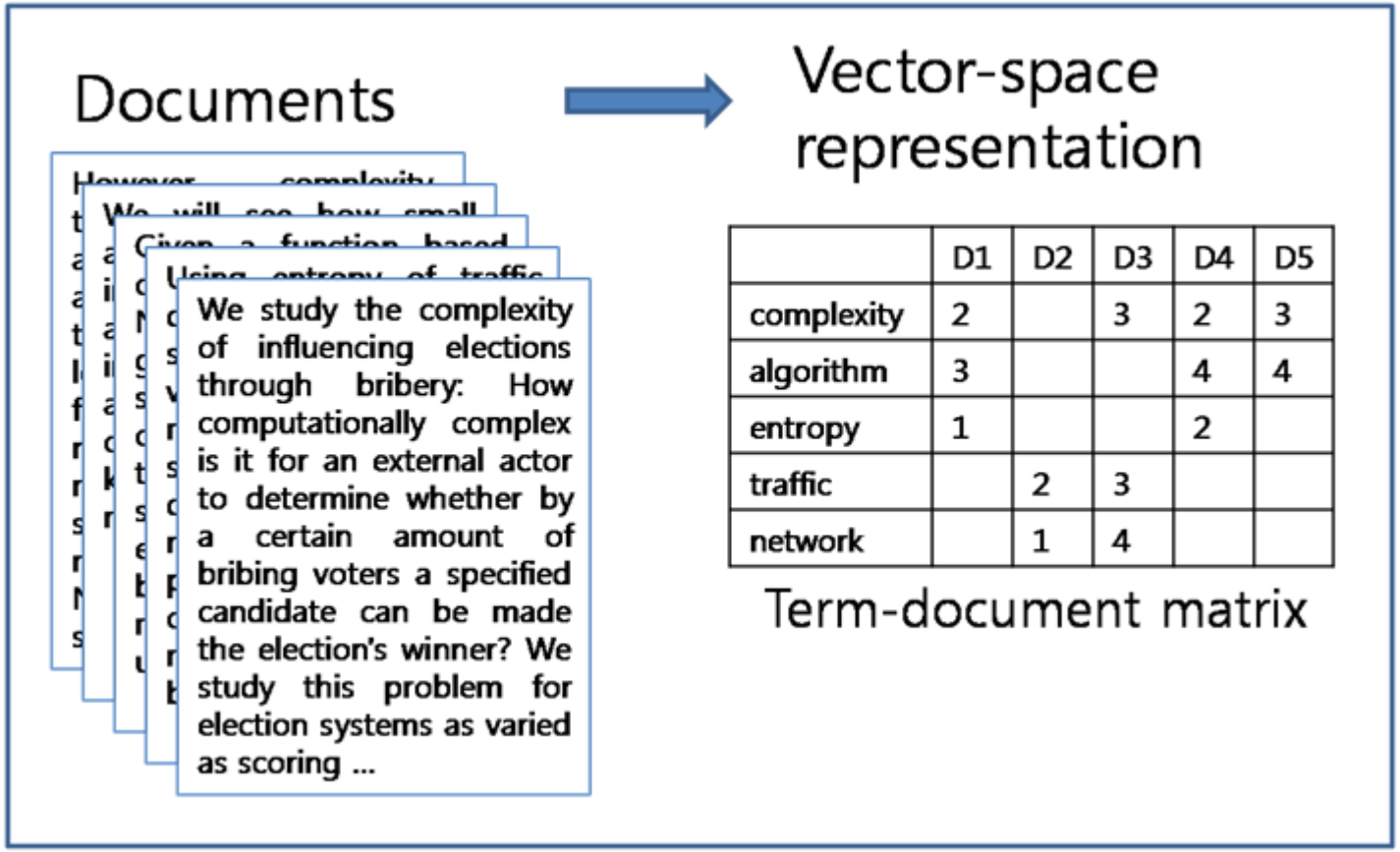
- Trocas eventuais podem ser feitas ainda:
  - tm_map(x, content_transformer(gsub), pattern = "á", replacement = "a")

Matriz de documentos/termos

- DocumentTermMatrix(x) ou TermDocumentMatrix (x)
- Pode representar:
  - Documentos/Termos

		Terms			
Documents		data	result	statistics	analysis
	Document1	0	1	0	1
	Document2	1	0	1	0
	Document3	0	0	1	0
	Document4	1	1	0	0

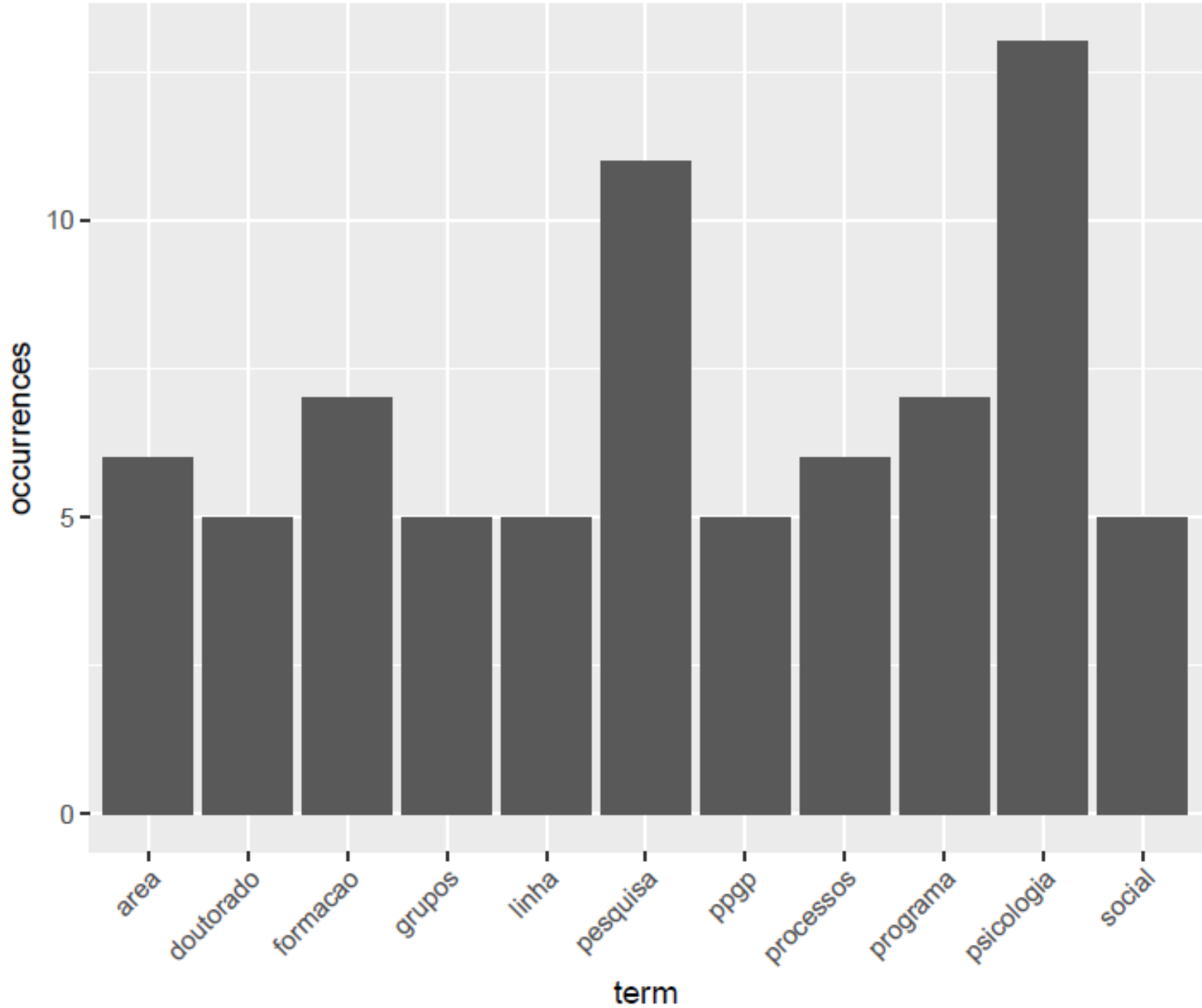
- Termos//Documentos

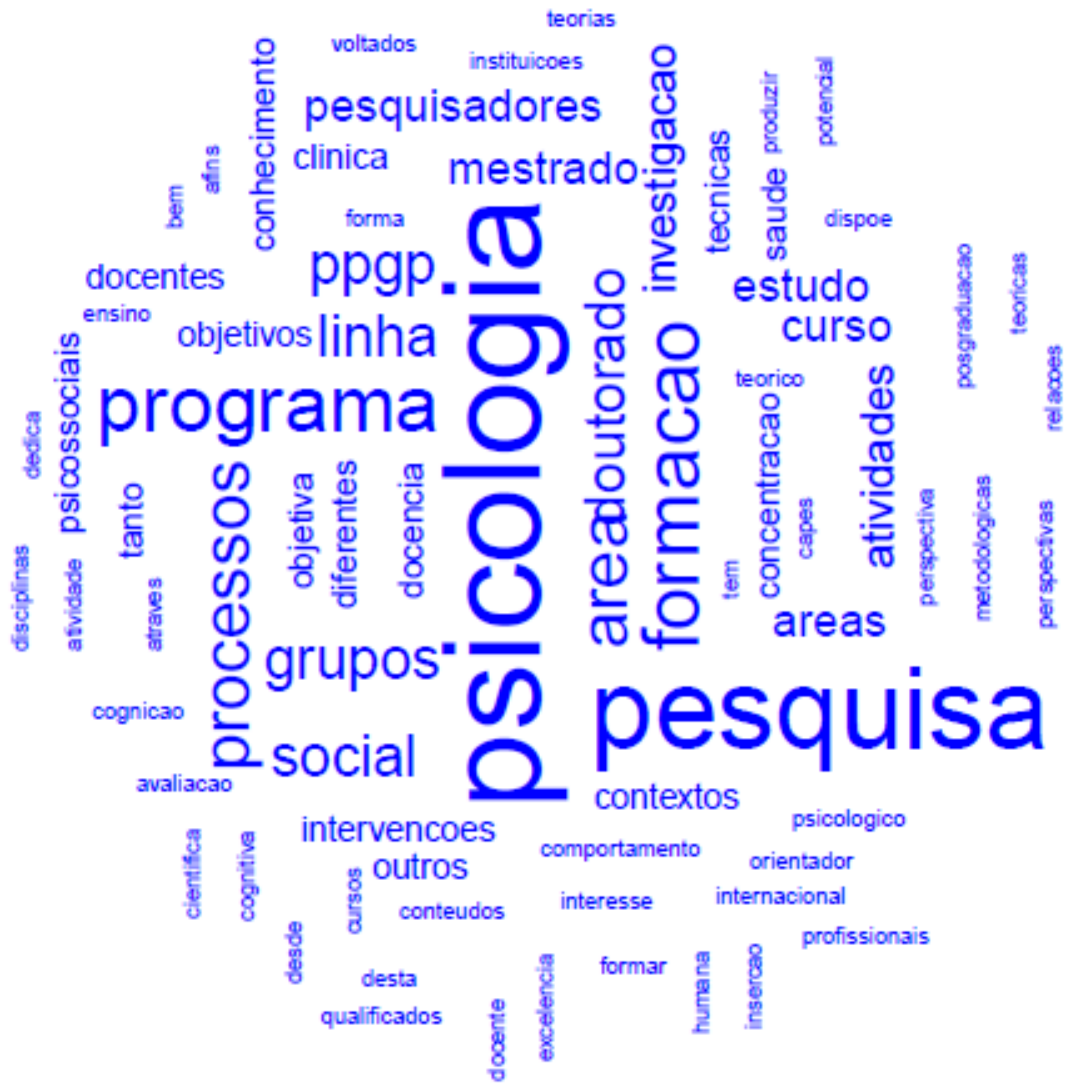


	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Análise descritiva de termos

- Frequências
  - Auxiliam a visualizar termos mais freq. e infreq. em documentos
  - Aproximação em termos de conteúdo
- Nuvens de palavras
  - Permitem a comparação e representação relativa da freq. de palavras em um ou mais documentos


occurrences
10
5
0
area
doutorado
formacao
grupos
linha
pesquisa
ppgp
processos
programa
psicologia
social
term


psicologia
pesquisa
programa
formacao
area
doutorado
processos
grupos
social
linha
ppgp
docentes
objetivos
pesquisadores
clinica
mestrado
investigacao
conhecimento
estudo
curso
atividades
areas
concentracao
contextos
intervencoes
outros
psicologico
comportamento
orientador
internacional
profissionais
interesse
conteudos
formar
desta
qualificados
docente
excelencia
humana
insercao
cursos
desde
cientifica
cognitiva
cognicao
avaliacao
psicossociais
tanto
objetiva
diferentes
docencia
tecnicas
saude
produzir
potencial
dispoe
forma
bem
afins
voltados
instituicoes
teorias
ensino
teorico
posgraduacao
teoricas
relacoes
perspectivas
metodologicas
perspectiva
capes
tem
atraves
atividade
disciplinas
dedica

Journal of Educational Evaluation for Health Professions

J Educ Eval Health Prof 2015, 12: 12 • http://dx.doi.org/10.3352/jeehp.2015.12.12

Open Access

eISSN: 1975-5937

TECHNICAL REPORT

Visualizing the qualitative: making sense of written comments from an evaluative satisfaction survey

Keith V. Bletzer*

National Community Health Partners, Tucson, AZ; School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA

- Data were collected by a community agency, National Community Health Partners, funded by the Centers for Disease Control and Prevention (CDC) to provide Capacity Building Assistance (CBA) to AIDS service organizations, and state/local health departments, across the United States and its territories.
- To improve evaluative rigor and complement quantitative satisfaction data, three open questions [5] asked participants what was most and least effective in the training, and what they would recommend to improve the learning experience. Question 16 (Q16) and Question 17 (Q17) focused on training elements that were considered by participants as *most* and *least* effective, respectively:
- Q16: What were the MOST effective parts of this CBA event for you?
- Q17: What were the LEAST effective parts of this CBA event for you?

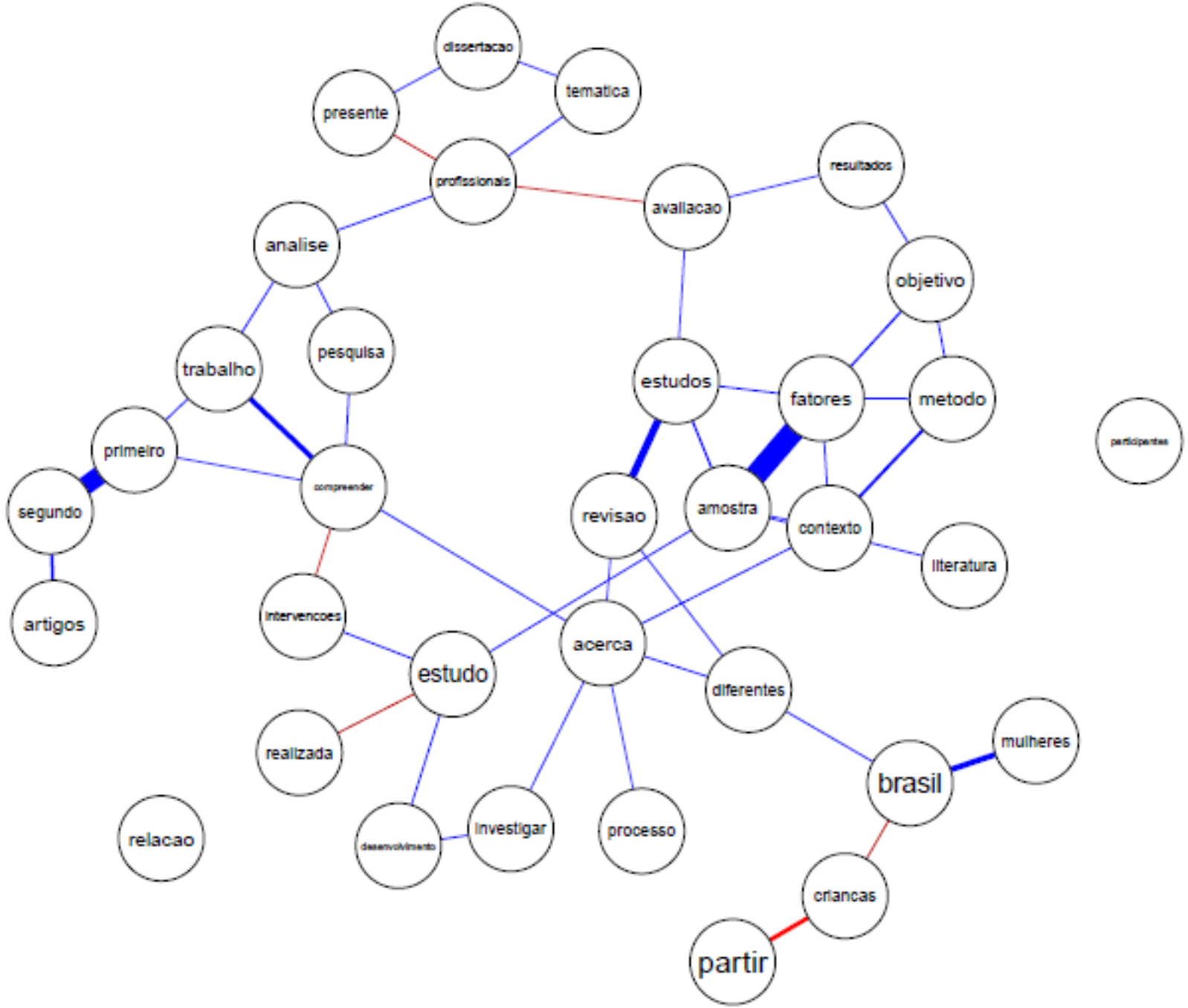
"What were the MOST effective parts of the CBA event?" (Question 16)

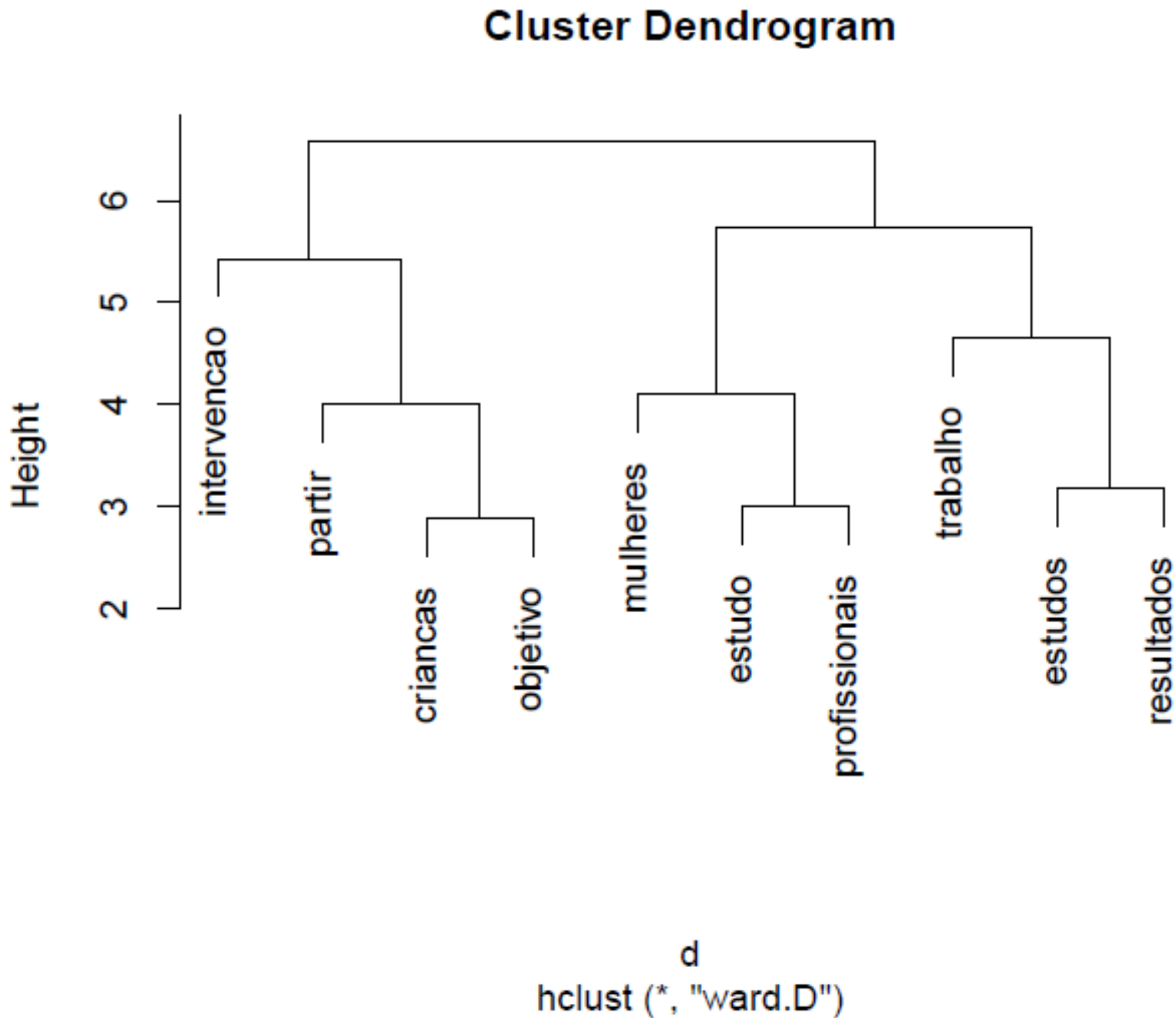
50 Terms Most Frequently Used, based on 1,058 responses from 1,286 Participants (YR5).

- **Results:** A three-tier display incorporated a Word Cloud at the top, followed by the corresponding frequency table, and a textual summary of the qualitative data represented by the Word Cloud imagery. This mixed format adheres to recognition that people vary in what format is most effective for assimilating new information
- **Conclusion:** The combination of visual representation through Word Clouds complemented by quantified qualitative materials is one means of increasing comprehensibility for a range of stakeholders, who might not be familiar with numerical tables or statistical analyses.

Correlações e agrupamentos de termos

- Permitem observar associações entre termos, indo além da mera descrição da freq.
- Análises exploratórias que possibilitam visualizar a associação e coocorrência de termos em uma coleção de documentos
- Permite descrever um único documento ou comparar dois ou mais documentos


dissertacao
tematica
presente
profissionais
avaliacao
resultados
analise
objetivo
trabalho
pesquisa
estudos
fatores
metodo
participantes
primeiro
compreender
segundo
revisao
amostra
contexto
literatura
artigos
intervencoes
acerca
estudo
diferentes
realizada
mulheres
brasil
relacao
desenvolvimento
investigar
processo
criancas
partir


Cluster Dendrogram
Height
2
3
4
5
6
intervencao
partir
criancas
objetivo
mulheres
estudo
profissionais
trabalho
estudos
resultados
d
hclust (*, "ward.D")

Revista de Psicología del Deporte. 2017, Vol 27, Suppl 1, pp. 59-65
Journal of Sport Psychology 2017, Vol 27, Suppl 1, pp. 59-65
ISSN: 1132-239X
ISSNe: 1988-5636

Universitat de les Illes Balears
Universitat Autònoma de Barcelona

Esporte é um contexto que possibilita emancipação ou colonização no processo de formação identitária?

João Ricardo Nickenig Vissoci*, Leonardo Pestillo de Oliveira, José Roberto Andrade do Nascimento Junior***, Fernanda Soares Nakashima**, Wagner de Lara Machado****, Antonio da Costa Ciampa*****, Renan Codonhato****** e Lenamar Fiorese Vieira******

A proposta Sintagma Identidade - Metamorfose - Emancipação foi adotada como suporte teórico para o processo de formação identitária. Esta trata identidade como um processo contínuo relacionado à atividade do ser humano de adotar personagens a partir de papéis estabelecidos (Ciampa, 1987). Tais personagens passam por transformações (metamorfoses) com as modificações do indivíduo e/ou do ambiente, direcionando-se como um movimento político de busca de autonomia (Habermas, 1990). Nessas metamorfoses o indivíduo pode integrar aspectos da cultura na sua expressão identitária (identidades políticas) em busca de emancipação da dominação ideológica.

Metamorfoses podem ser direcionadas para a manutenção da heteronomia e do *status quo* na expressão identitária (política de identidade) ao invés de conduzir o sujeito à busca de autonomia (Ciampa, 1987).

Método

Participantes

Participaram deste estudo 25 jogadores (25.49 ± 4.91 anos e 9.12 ± 3.59 anos de prática) das equipes participantes da Liga Nacional de Futsal. O estudo foi aprovado pelo comitê de ética da Universidade Estadual de Maringá (n°248.363/2013).

Entrevista semiestruturada

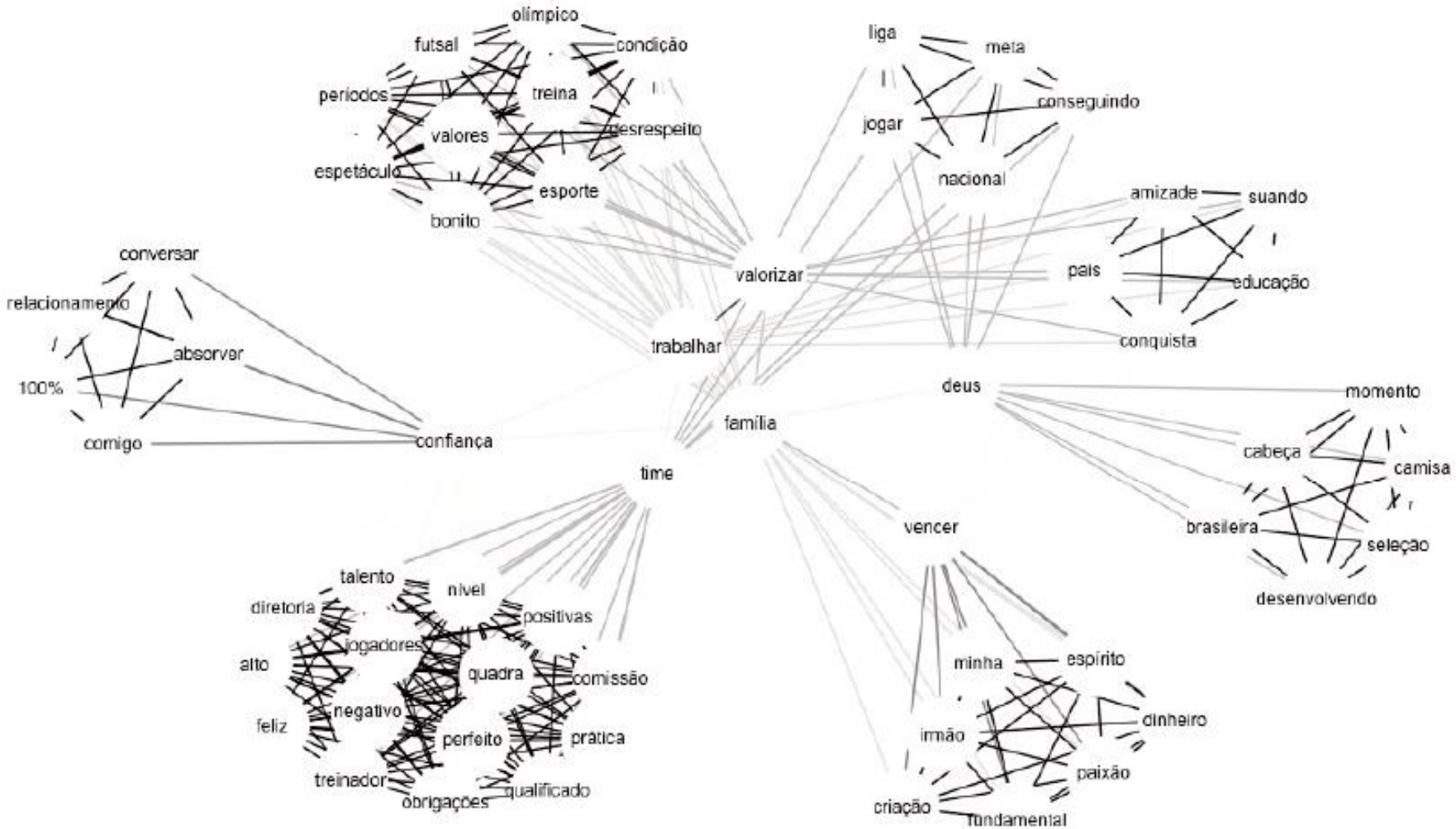
Foi conduzida abordando a trajetória de vida e esportiva dos atletas seguindo o método de história de vida. Os elementos

Análise dos dados

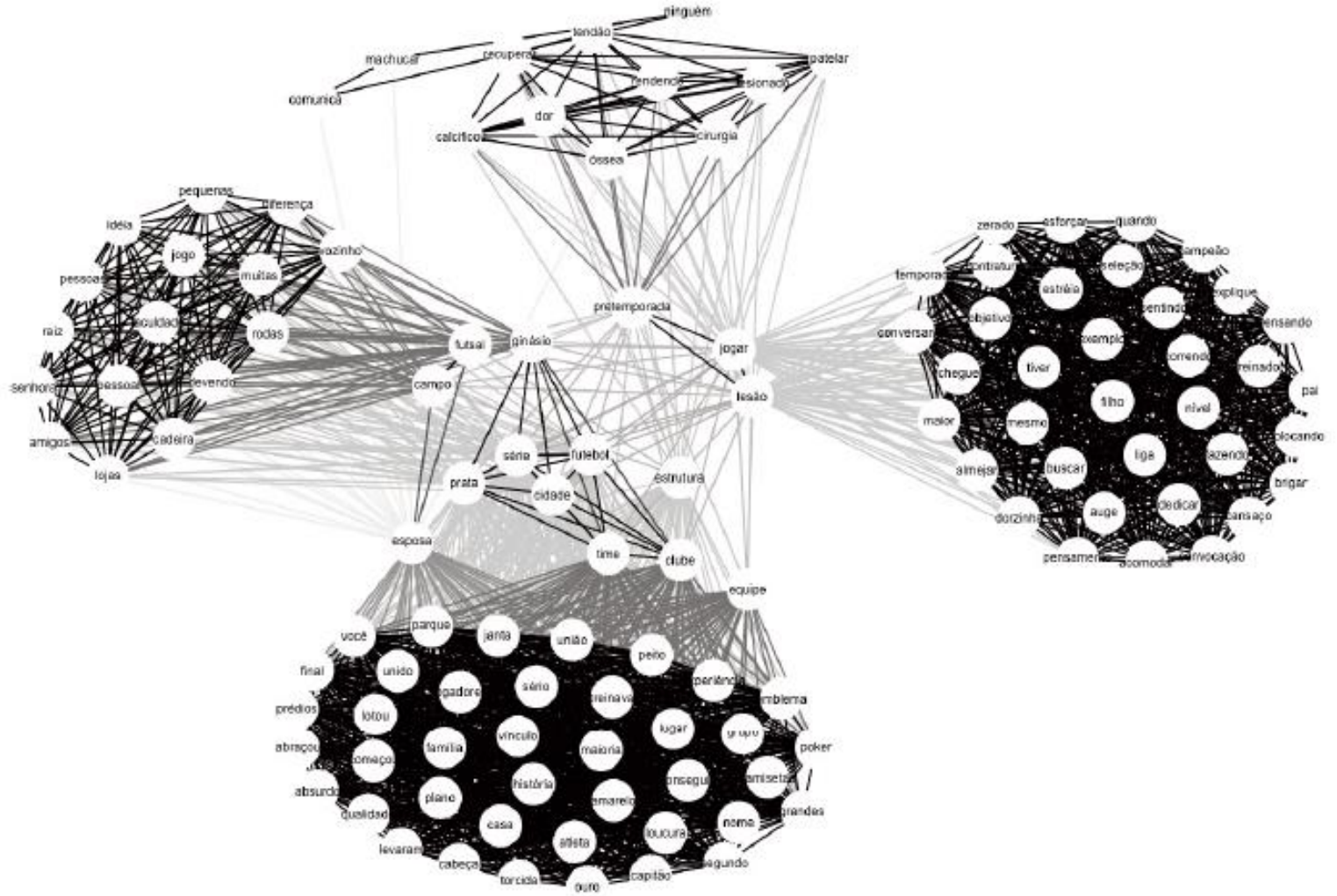
O método de análise foi do tipo dedutiva, partindo de categorias pré-definidas de classificação dos atletas que foram categorizados de acordo com o conteúdo direcionados para autonomia ou heteronomia e com os projetos de vida, fundamentados em políticas de identidade ou em identidades políticas (Ciampa 1987, Dantas 2013). Os grupos foram: (a) discurso voltado para heteronomia e que não evidenciam um projeto de vida; (b) discurso voltado para autonomia com um projeto de vida fundamentado em políticas de identidade; (c) discurso voltado para autonomia e com um projeto de vida fundamentado em identidades políticas. Cada grupo foi caracterizado dentro do processo de metamorfose-emancipação da formação identitária através de história de vida e da trajetória esportiva ilustradas com relatos de sujeitos emblemáticos e padrão semântico das palavras com análise em rede.

Cada palavra do corpus caracterizou um nodo na rede conectado pelas hastes, que representam a intensidade da sua correlação. Nessa rede bipartite, palavras que fossem expressas nos mesmos parágrafos teriam uma relação maior (direta), com relações indiretas estabelecidas pela presença de palavras em comum. Por exemplo, ao dizer "Eu gosto de futsal", os termos "gosto" e "futsal" teriam uma associação alta e direta. Contudo, se em outro parágrafo ao relatar "Eu gosto de vencer", "futsal" e "vencer" teriam uma associação indireta pelo compartilhamento da sentença com a palavra comum "gosto". Quanto mais próximos os nodos, maior é associação entre as palavras.

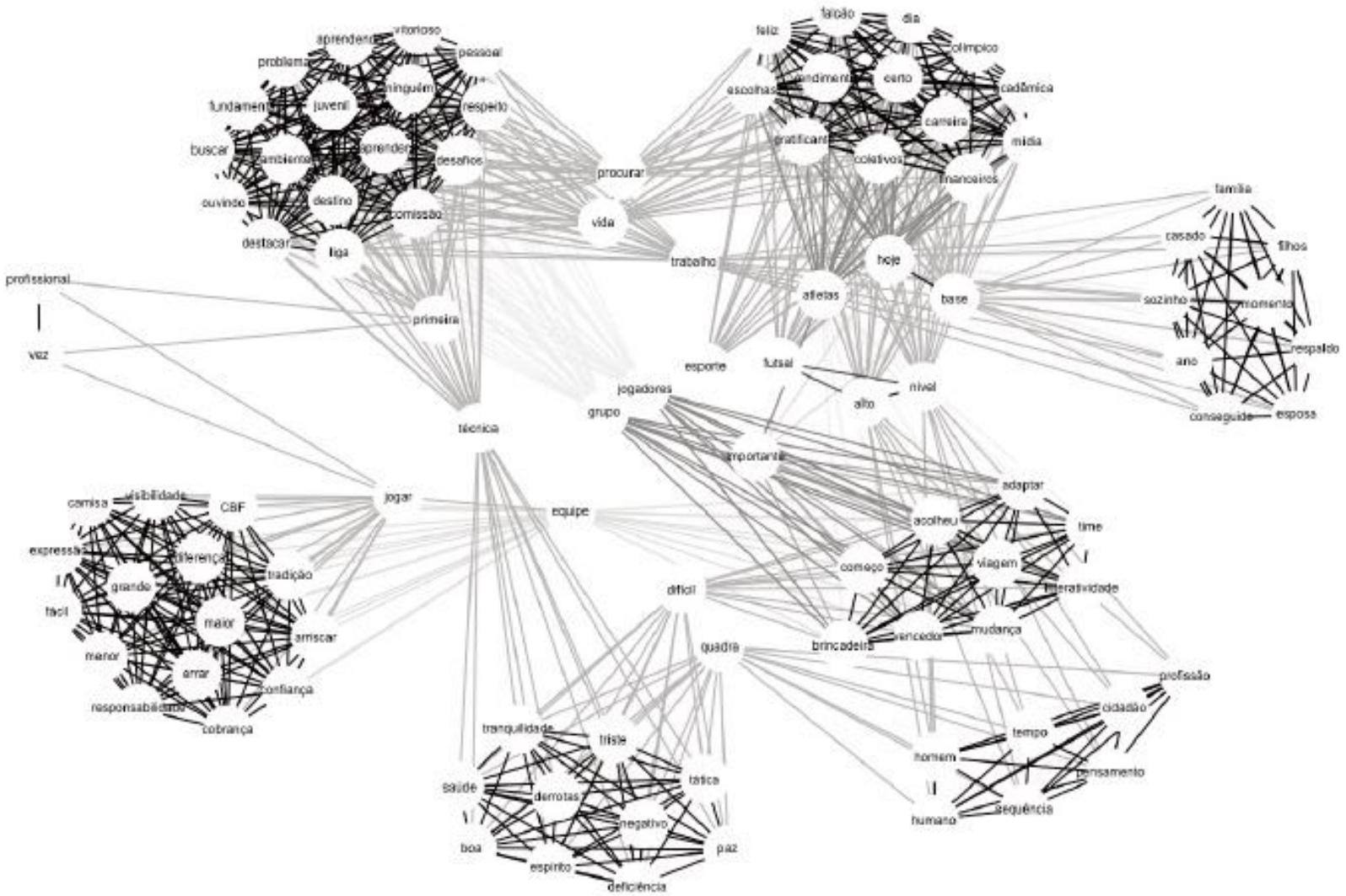
Heteronomia sem projeto



Autonomia sem projeto



Autonomia com projeto



Discussão

Este foi o primeiro estudo que analisou as possibilidades emancipatórias ou colonizadoras no processo de formação da identidade em atletas de Futsal. Identificamos três grupos caracterizados pela forma como expressavam suas metamorfoses. Atletas do primeiro grupo utilizaram com frequência palavras que representam uma forma de racionalidade instrumental ligados à vitória e sucesso, características do esporte espetáculo (Dumitriu, 2016). Os grupos mais autonômicos apresentavam conexões semânticas mais diversas, unindo experiências esportivas e não esportivas,

OPEN ACCESS Freely available online

PLOS | ONE

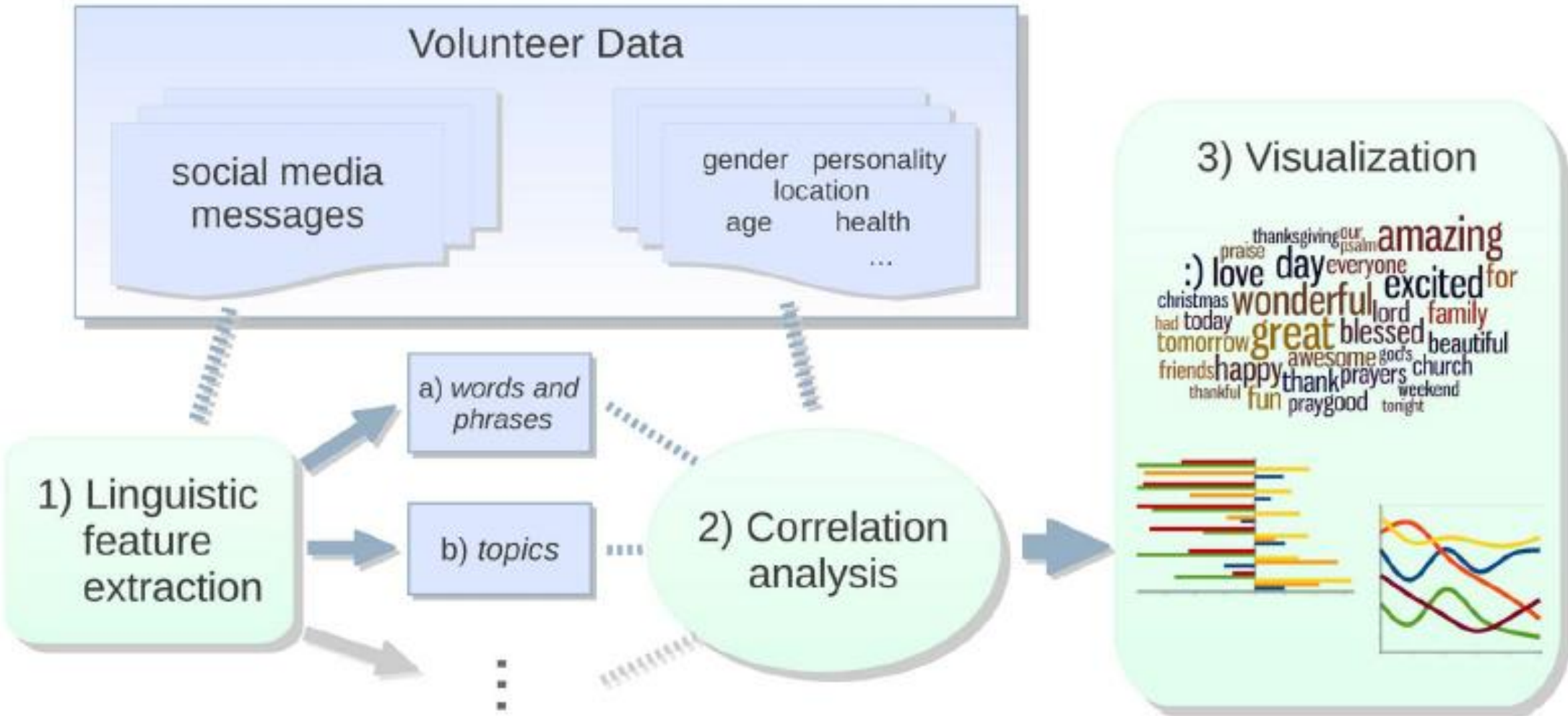
Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach

H. Andrew Schwartz[1,2]*, Johannes C. Eichstaedt[1], Margaret L. Kern[1], Lukasz Dziurzynski[1], Stephanie M. Ramones[1], Megha Agrawal[1,2], Achal Shah[2], Michal Kosinski[3], David Stillwell[3], Martin E. P. Seligman[1], Lyle H. Ungar[2]

1 Positive Psychology Center, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **2** Computer & Information Science, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **3** The Psychometrics Centre, University of Cambridge, Cambridge, United Kingdom

Abstract

We analyzed 700 million words, phrases, and topic instances collected from the Facebook messages of 75,000 volunteers, who also took standard personality tests, and found striking variations in language with personality, gender, and age. In our *open-vocabulary* technique, the data itself drives a comprehensive exploration of language that distinguishes people, finding connections that are not captured with traditional closed-vocabulary word-category analyses. Our analyses shed new light on psychosocial processes yielding results that are face valid (e.g., subjects living in high elevations talk about the mountains), tie in with other research (e.g., neurotic people disproportionately use the phrase 'sick of' and the word 'depressed'), suggest new hypotheses (e.g., an active life implies emotional stability), and give detailed insights (males use the possessive 'my' when mentioning their 'wife' or 'girlfriend' more often than females use 'my' with 'husband' or 'boyfriend'). To date, this represents the largest study, by an order of magnitude, of language and personality.


Volunteer Data
social media messages
gender personality location age health ...
1) Linguistic feature extraction
a) words and phrases
b) topics
...
2) Correlation analysis
3) Visualization
thanksgiving our psalm amazing
praise
:) love day everyone
christmas wonderful excited for
had today lord family
tomorrow great blessed beautiful
friends happy awesome god's church
thankful fun thank prayers
praygood weekend
tonight

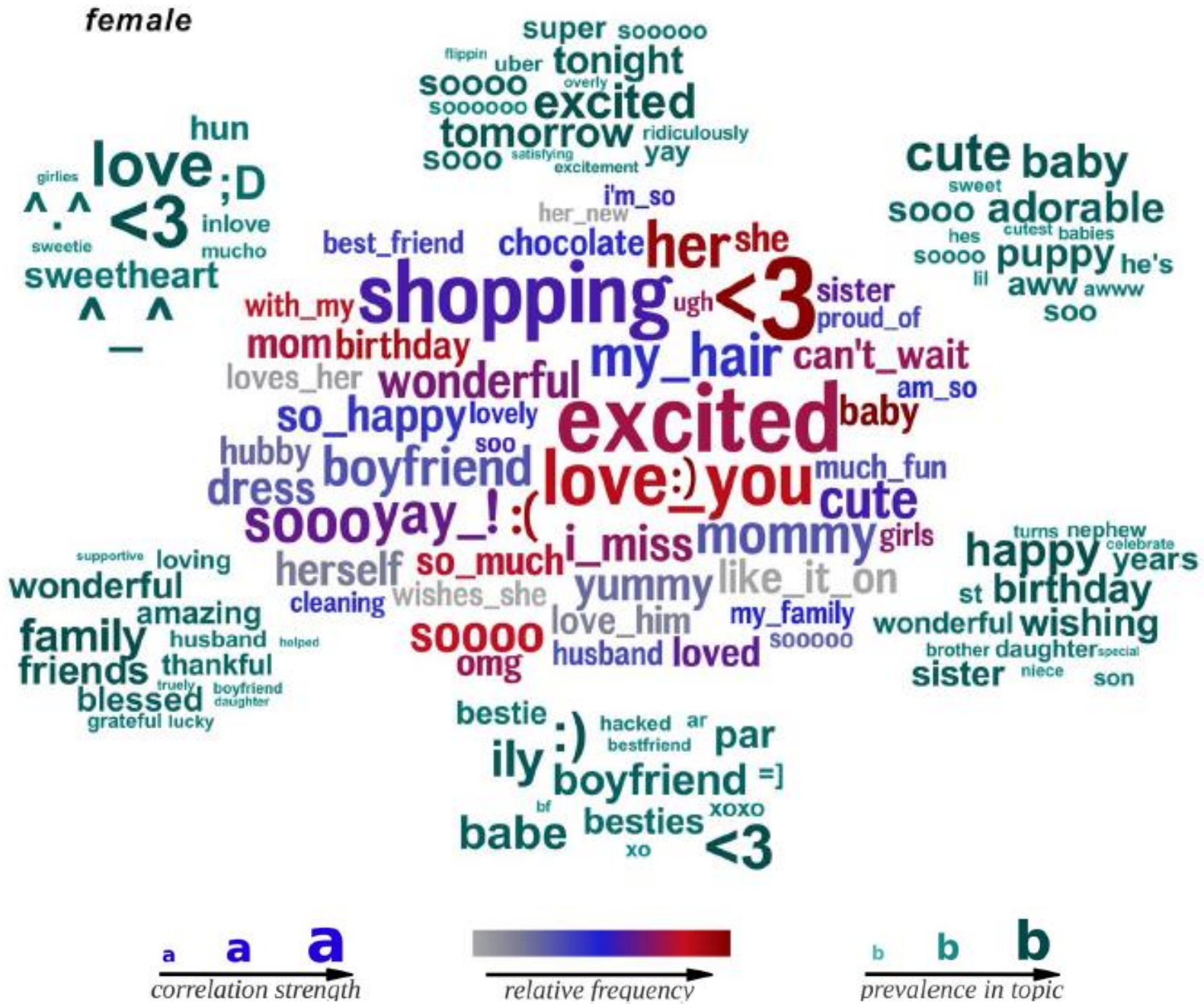
Closed Vocabulary: Word-Category Lexica

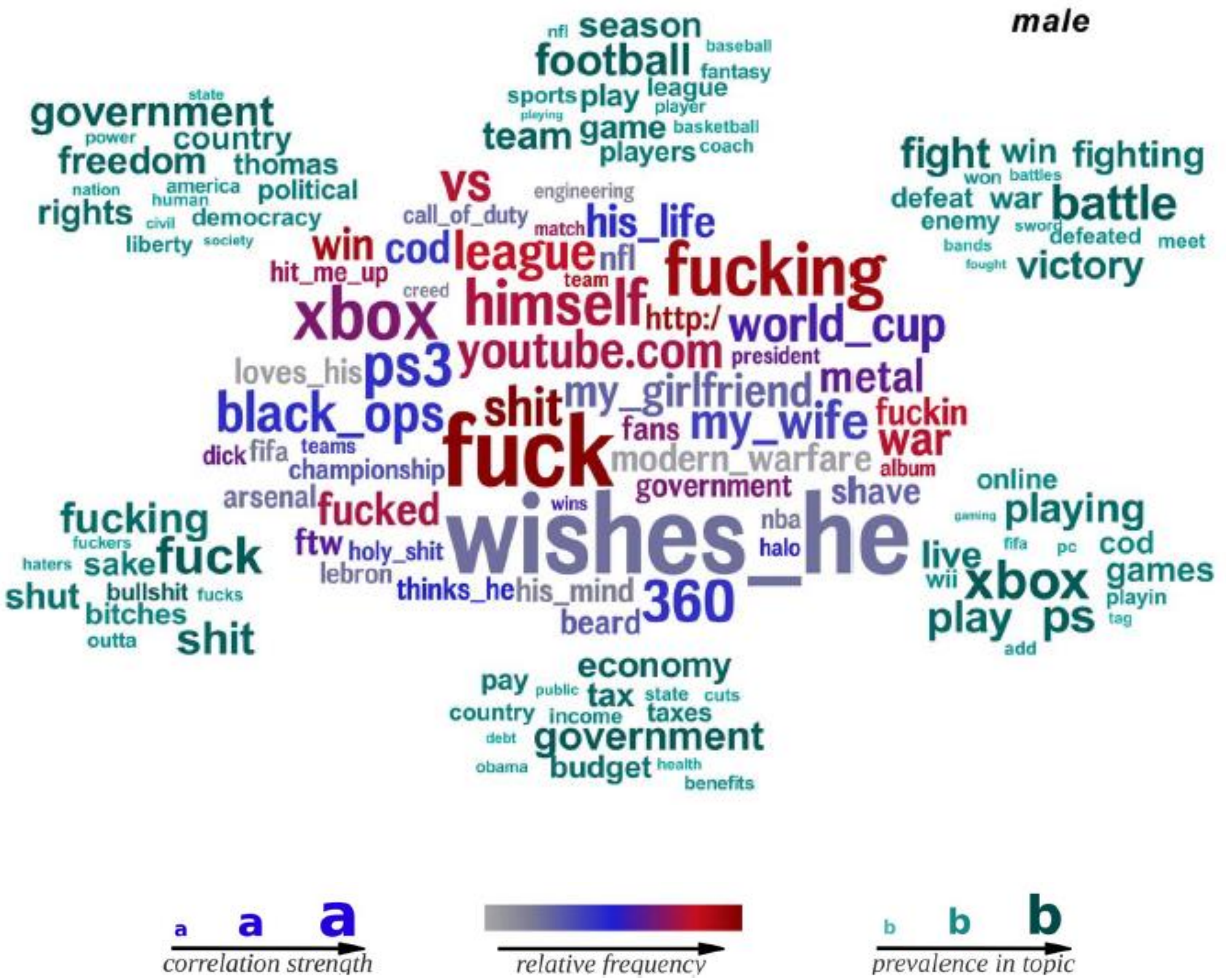
A common method for linking language with psychological variables involves counting words belonging to manually-created categories of language. Sometimes referred to as the *word-count* approach, one counts how often words in a given category are used by an individual, the percentage of the participants' words which are from the given category:

$$p\,(category \mid subject) = \frac{\sum_{word \in category} freq\,(word, subject)}{\sum_{word \in vocab\,(subject)} freq\,(word, subject)}$$

where $freq\,(word,subject)$ is the number of the times the participant mentions *word* and $vocab\,(subject)$ is the set of all words mentioned by the subject.

- - NEO-PI-R
- Questionário sociodemográfico
- Facebook updates
- Gênero e sexo


female
super
sooooo
flippin
uber
tonight
soooo
overly
excited
soooooo
tomorrow
ridiculously
sooo
satisfying
excitement
yay
hun
girlies
love
;D
^.^
<3
inlove
sweetie
mucho
sweetheart
^ ^
—
i'm_so
her_new
best_friend
chocolate
her
she
shopping
with_my
ugh
<3
sister
proud_of
mom
birthday
my_hair
can't_wait
loves_her
wonderful
am_so
so_happy
lovely
excited
baby
soo
hubby
boyfriend
love:)_you
much_fun
dress
cute
soooyay_!:(
i_miss
mommy
girls
supportive
loving
herself
so_much
yummy
like_it_on
wonderful
cleaning
wishes_she
amazing
family
husband
helped
love_him
my_family
soooo
sooooo
friends
thankful
omg
husband
loved
blessed
truely
boyfriend
daughter
grateful
lucky
bestie
:)
hacked
ar
bestfriend
par
ily
boyfriend
=]
bf
xoxo
babe
besties
xo
<3
cute
baby
sweet
sooo
adorable
hes
cutest
babies
soooo
puppy
he's
lil
aww
awww
soo
turns
nephew
celebrate
happy
years
st
birthday
wonderful
wishing
brother
daughter
special
sister
niece
son
a
a
a
correlation strength
relative frequency
b
b
b
prevalence in topic


male
nfl season baseball
football fantasy
sports play league player playing
team game basketball
players coach
state
government
power country
freedom thomas
nation america human political
rights civil democracy
liberty society
fight win fighting
won battles
defeat war battle
enemy sword defeated meet
bands
fought victory
vs engineering
call_of_duty match his_life
win cod league nfl fucking
hit_me_up team
creed
xbox himself http:/ world_cup
youtube.com president
loves_his ps3
black_ops shit my_girlfriend metal
fans my_wife fuckin
dick fifa teams fuck war
championship modern_warfare album
arsenal government shave
fucked wins
wishes_he nba halo
ftw holy_shit
lebron
thinks_he his_mind 360
beard
fucking
fuckers
haters sake fuck
shut bullshit fucks
bitches
outta shit
online
gaming playing
live fifa pc cod
wii xbox games
playin
play ps tag
add
economy
pay public tax state cuts
country income taxes
debt government
obama budget health
benefits
a a a
correlation strength
relative frequency
b b b
prevalence in topic

Análise Semântica Latente

- Tem por objetivo classificar observações de acordo com padrões de similaridade
- Em geral envolvem um sistema de ponderação, do uso de um termo em relação ao uso geral para cada observação
- Após calcular os pesos em uma matriz de termos e documentos, é possível aplicar diversas técnicas como Escalonamento Multidimensional, Cluster Hierárquico e Análise de Componentes Principais/Fatoriais

npj | Schizophrenia

www.nature.com/npjschz

ARTICLE OPEN

Automated analysis of free speech predicts psychosis onset in high-risk youths

Gillinder Bedi[1,2,9], Facundo Carrillo[3,9], Guillermo A Cecchi[4], Diego Fernández Slezak[3], Mariano Sigman[5], Natália B Mota[6], Sidarta Ribeiro[6], Daniel C Javitt[1,7], Mauro Copelli[8] and Cheryl M Corcoran[1,7]

BACKGROUND/OBJECTIVES: Psychiatry lacks the objective clinical tests routinely used in other specializations. Novel computerized methods to characterize complex behaviors such as speech could be used to identify and predict psychiatric illness in individuals.
AIMS: In this proof-of-principle study, our aim was to test automated speech analyses combined with Machine Learning to predict later psychosis onset in youths at clinical high-risk (CHR) for psychosis.
METHODS: Thirty-four CHR youths (11 females) had baseline interviews and were assessed quarterly for up to 2.5 years; five transitioned to psychosis. Using automated analysis, transcripts of interviews were evaluated for semantic and syntactic features predicting later psychosis onset. Speech features were fed into a convex hull classification algorithm with leave-one-subject-out cross-validation to assess their predictive value for psychosis outcome. The canonical correlation between the speech features and prodromal symptom ratings was computed.
RESULTS: Derived speech features included a Latent Semantic Analysis measure of semantic coherence and two syntactic markers of speech complexity: maximum phrase length and use of determiners (e.g., *which*). These speech features predicted later psychosis development with 100% accuracy, outperforming classification from clinical interviews. Speech features were significantly correlated with prodromal symptoms.
CONCLUSIONS: Findings support the utility of automated speech analysis to measure subtle, clinically relevant mental state changes in emergent psychosis. Recent developments in computer science, including natural language processing, could provide the foundation for future development of objective clinical tests for psychiatry.

npj Schizophrenia (2015) **1,** Article number: 15030; doi:10.1038/npjschz.2015.30; published online 26 August 2015

Table 1. Demographics

	CHR+ *(N = 5)*	*CHR −* *(N = 29)*
Age (in years)	22.2 (3.4)	21.2 (3.6)
Gender (% male)	80%	66%
Race (% Caucasian)	40%	38%
Medications prescribed (antipsychotics and/or antidepressants)	20%	21%

Abbreviations: CHR+, clinical high-risk participants who transitioned to psychosis during follow-up; CHR −, clinical high-risk participants who did not transition to psychosis during follow-up.

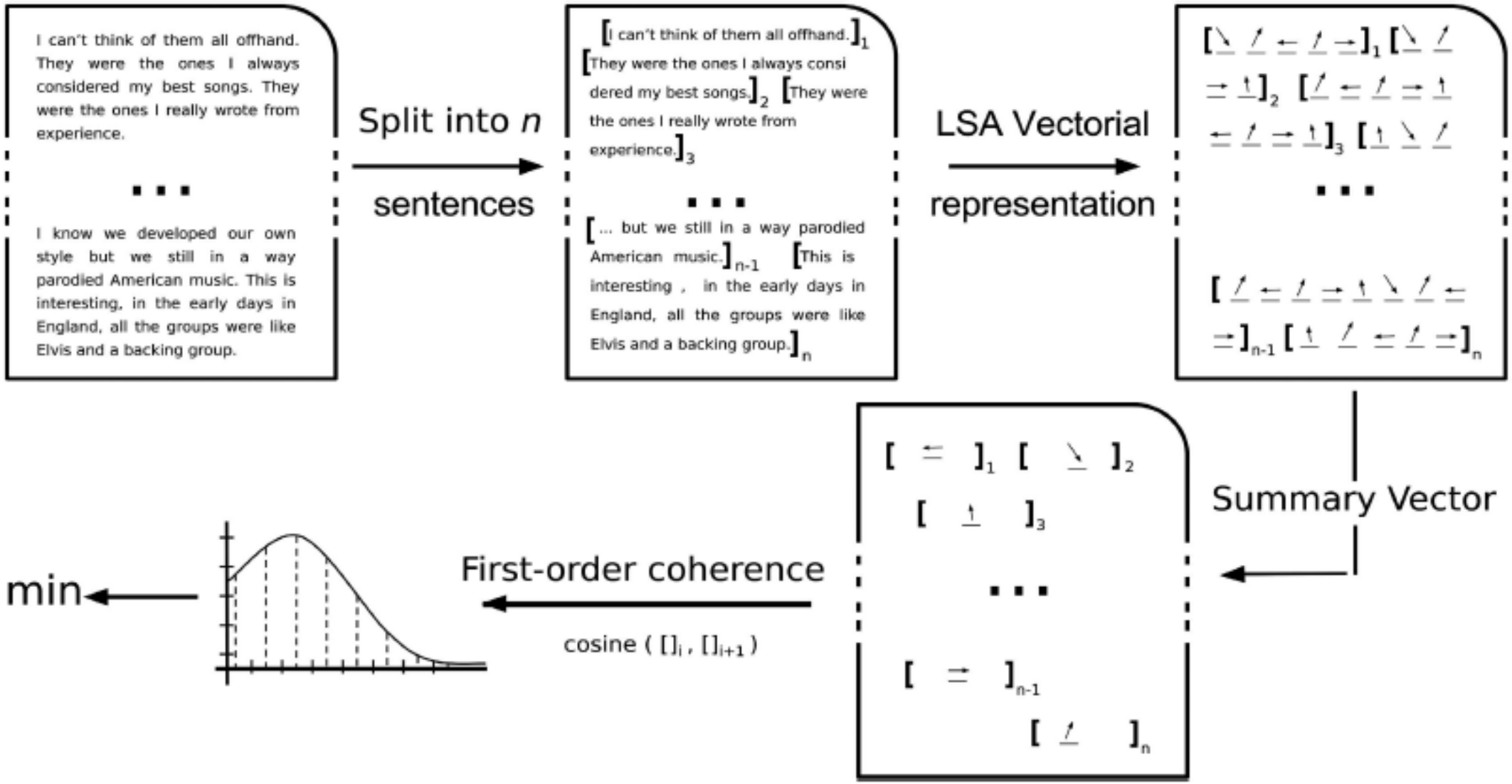


Figure 1. Pipeline for automated extraction of the semantic coherence features. Texts were initially split into sentences/phrases. Each word was represented as a vector in high-dimensional semantic space using Latent Semantic Analysis (LSA). Summary vectors were calculated as the mean of each vector in each phrase. Coherence was determined based on the semantic similarity between adjacent phrases, calculated as the cosine of their respective vectors. The semantic coherence feature that best discriminated those who transitioned to psychosis from those who did not was the minimum semantic coherence value (i.e., the coherence at the point of maximal discontinuity) within each transcribed text.

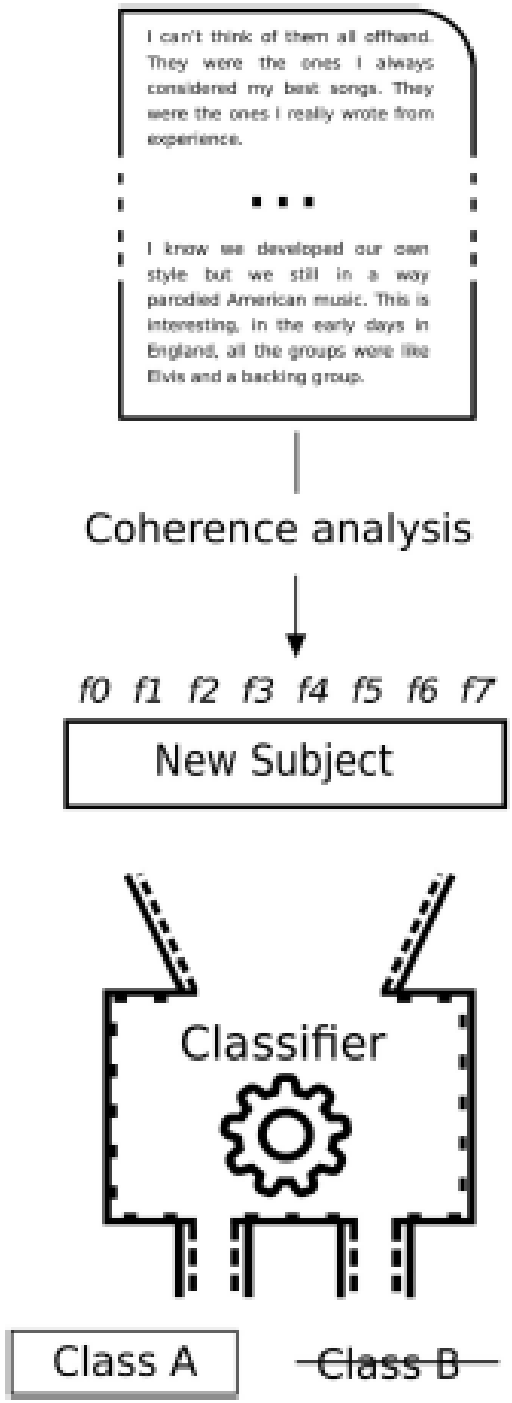
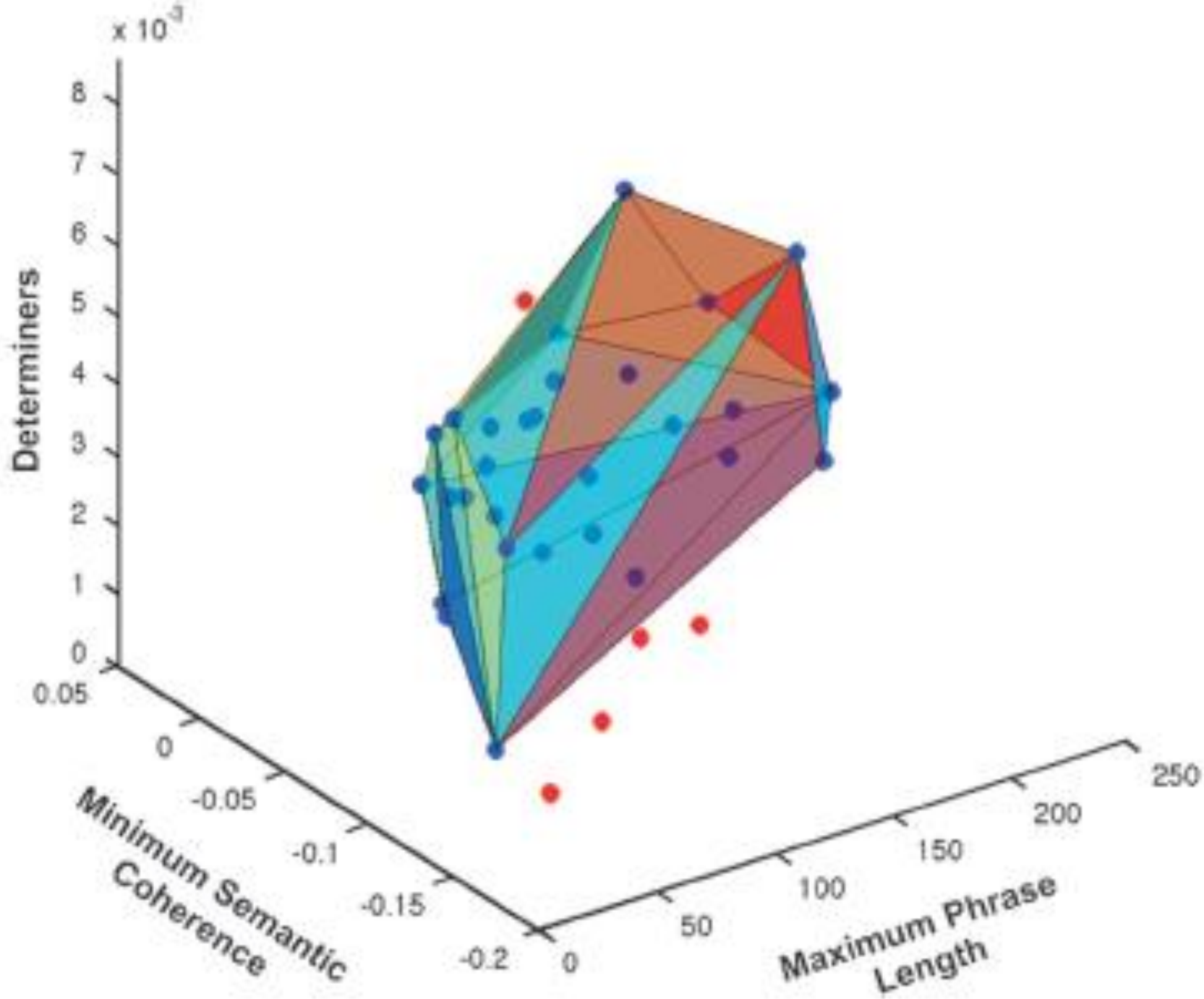

I can't think of them all offhand. They were the ones I always considered my best songs. They were the ones I really wrote from experience.
. . .
I know we developed our own style but we still in a way parodied American music. This is interesting, in the early days in England, all the groups were like Elvis and a backing group.
Coherence analysis
f0 f1 f2 f3 f4 f5 f6 f7
New Subject
Classifier
Class A
Class B

Table 2. Classification performance metrics

Classification	*PPV*	*NPV*	*Sens.*	*Spec.*	*ROC*
Convex Hull 3-feature	100	100	100	100	1.00
SIPS/SOPS	33	89	40	86	0.47

Abbreviations: NPV, negative predictive value; PPV, positive predictive value; ROC, receiver operating characteristic area under the curve; Sens, sensitivity; SIPS/SOPS, classification based on baseline scores on the Structured Interview for Prodromal Syndromes/Scale for Prodromal Symptoms; Spec, specificity.


x 10^-3
Determiners
8
7
6
5
4
3
2
1
0
0.05
0
-0.05
-0.1
-0.15
-0.2
Minimum Semantic Coherence
0
50
100
150
200
250
Maximum Phrase Length

Questões e dúvidas...