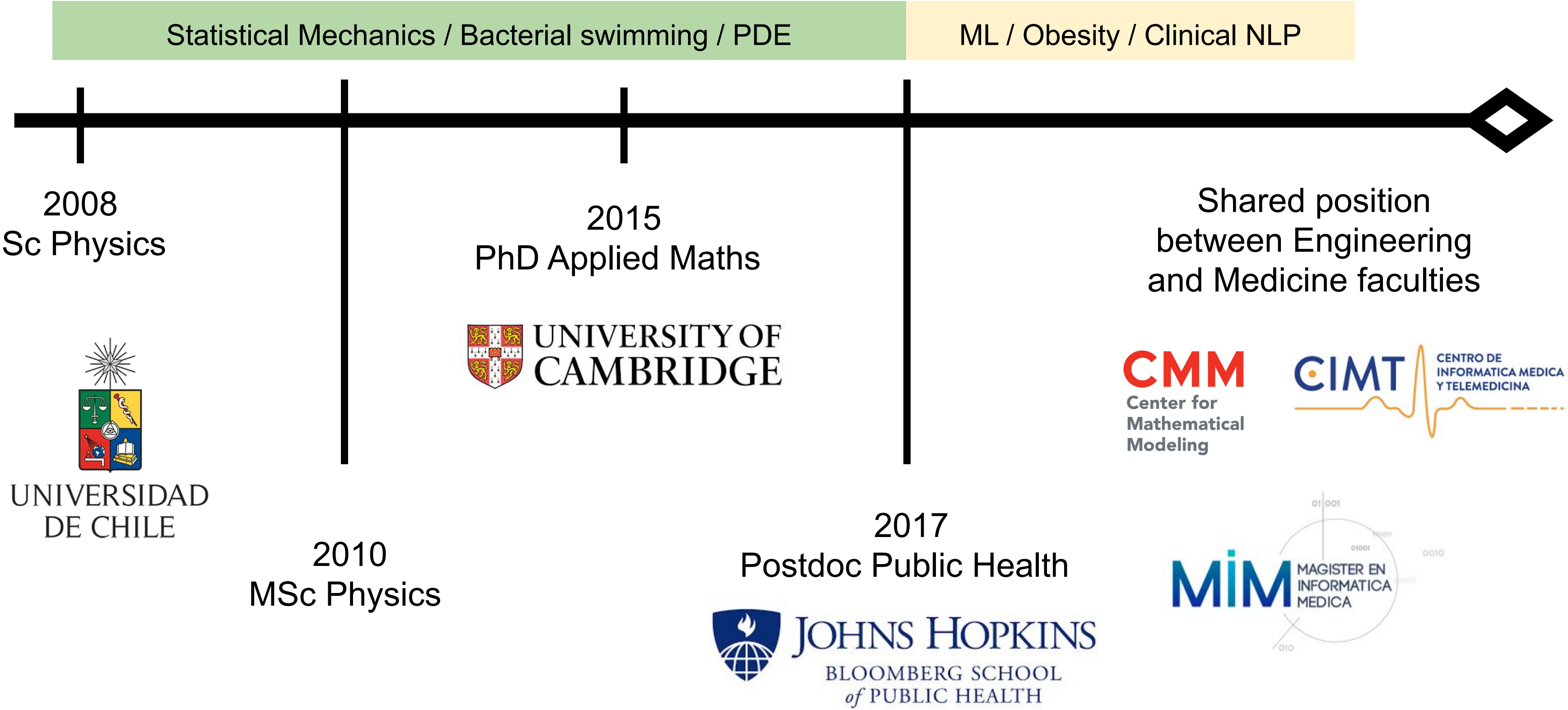
CIMT CENTRO DE INFORMATICA MEDICA Y TELEMEDICINA

CMM Centro de Modelamiento Matemático

Creating language resources to do clinical text mining in Chile

Jocelyn Dunstan, PhD MSc
University of Chile



Statistical Mechanics / Bacterial swimming / PDE

ML / Obesity / Clinical NLP

2008
BSc Physics

UNIVERSIDAD
DE CHILE

2010
MSc Physics

2015
PhD Applied Maths

UNIVERSITY OF
CAMBRIDGE

2017
Postdoc Public Health

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Shared position
between Engineering
and Medicine faculties

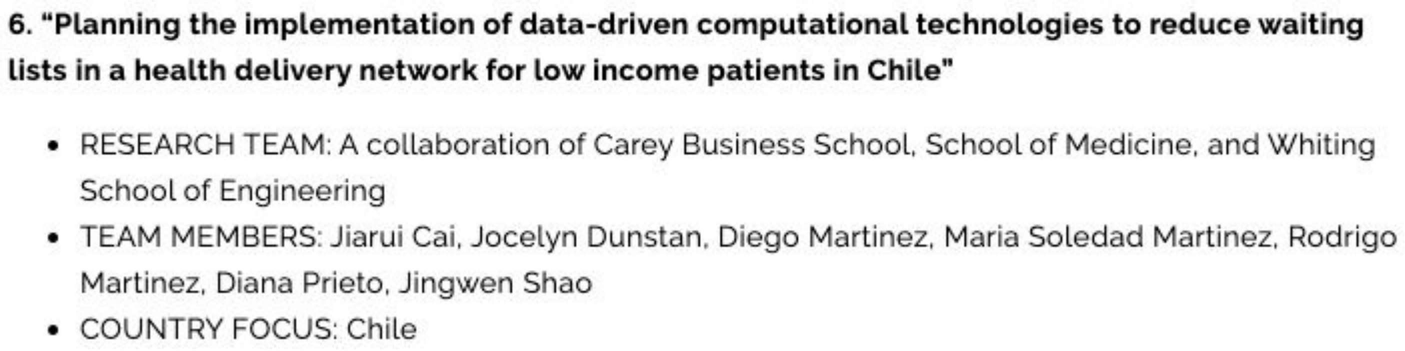
CMM
Center for
Mathematical
Modeling

CIMT
CENTRO DE
INFORMATICA MEDICA
Y TELEMEDICINA

MIM
MAGISTER EN
INFORMATICA
MEDICA

https://sites.google.com/view/jdunstan/home

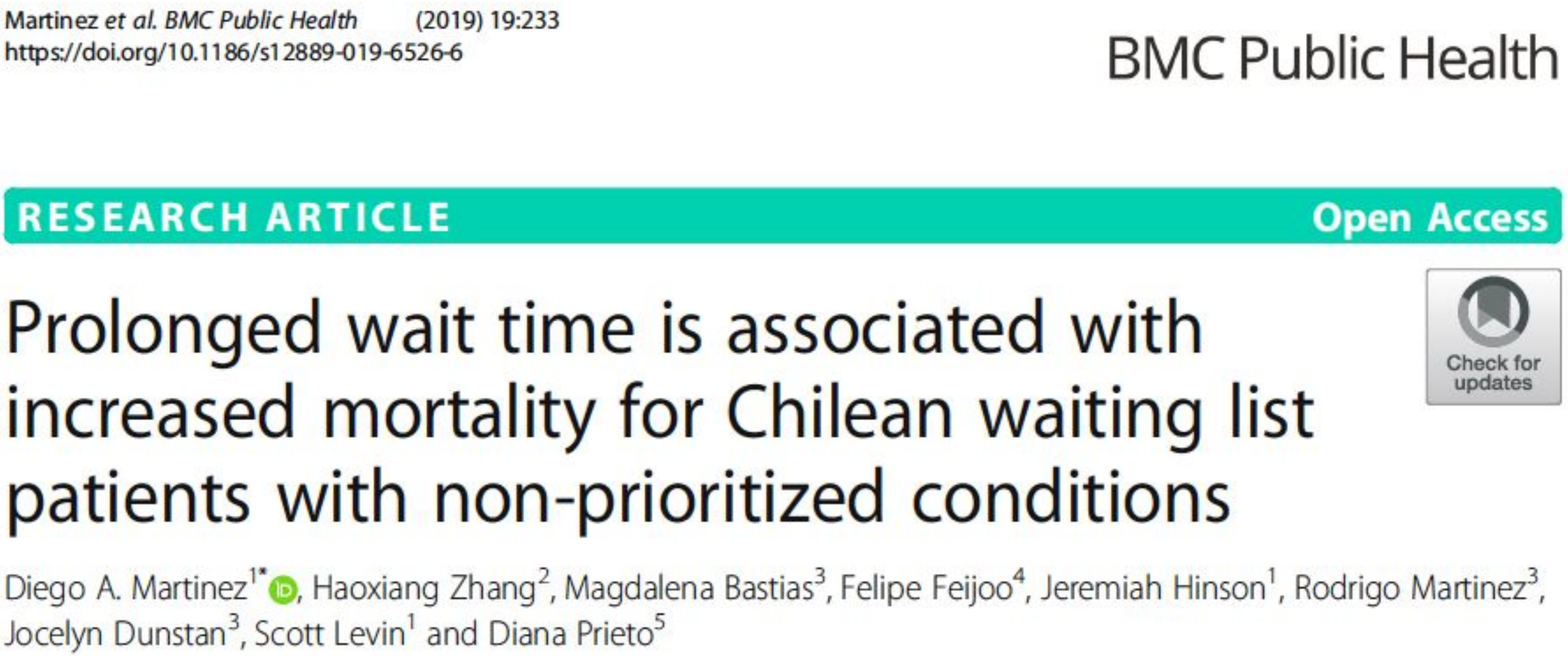
HEALTHIER WORLD CHALLENGE 2017
GRANT WINNERS
Innovative multidisciplinary research on health equity
Total of $318K across 14 Planning Grants
JOHNS HOPKINS ALLIANCE for a HEALTHIER WORLD
https://www.ahealthierworld.jhu.edu



6. “Planning the implementation of data-driven computational technologies to reduce waiting lists in a health delivery network for low income patients in Chile”

- RESEARCH TEAM: A collaboration of Carey Business School, School of Medicine, and Whiting School of Engineering
- TEAM MEMBERS: Jiarui Cai, Jocelyn Dunstan, Diego Martinez, Maria Soledad Martinez, Rodrigo Martinez, Diana Prieto, Jingwen Shao
- COUNTRY FOCUS: Chile

https://www.ahealthierworld.jhu.edu/ahw-updates/2018/2/5/six-seed-grants-totaling-148k-awarded-by-alliance-for-a-healthier-world



Martinez *et al. BMC Public Health* (2019) 19:233
https://doi.org/10.1186/s12889-019-6526-6

BMC Public Health

RESEARCH ARTICLE

Open Access

Prolonged wait time is associated with increased mortality for Chilean waiting list patients with non-prioritized conditions

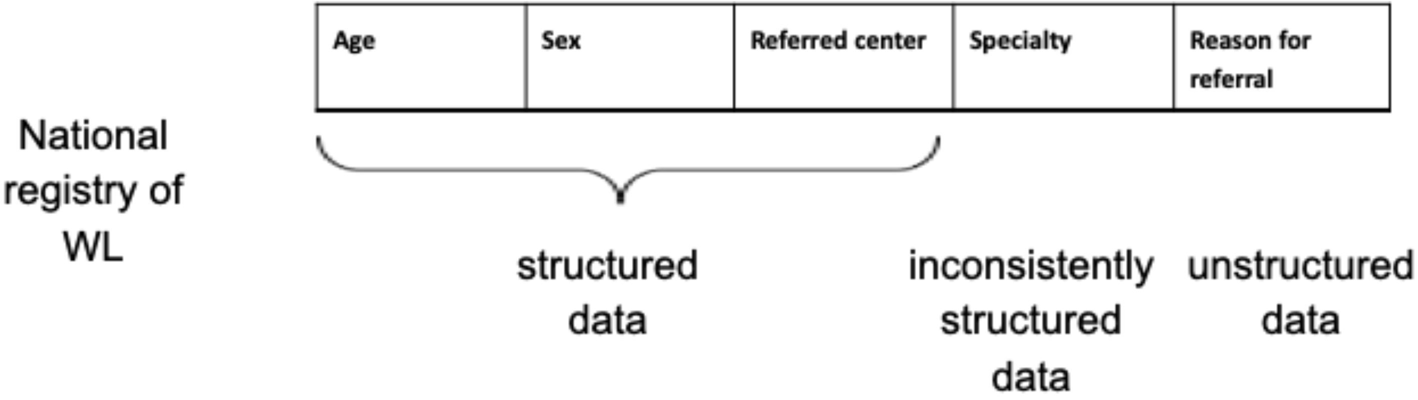
Check for updates

Diego A. Martinez[1*], Haoxiang Zhang[2], Magdalena Bastias[3], Felipe Feijoo[4], Jeremiah Hinson[1], Rodrigo Martinez[3], Jocelyn Dunstan[3], Scott Levin[1] and Diana Prieto[5]

Free text untouched in the analysis!

Classification of patients waiting in public hospitals

- 73% of the population in the in the public healthcare system
- To see an specialist you go in a waiting list (WL)



Age	Sex	Referred center	Specialty	Reason for referral

National registry of WL

structured data

inconsistently structured data

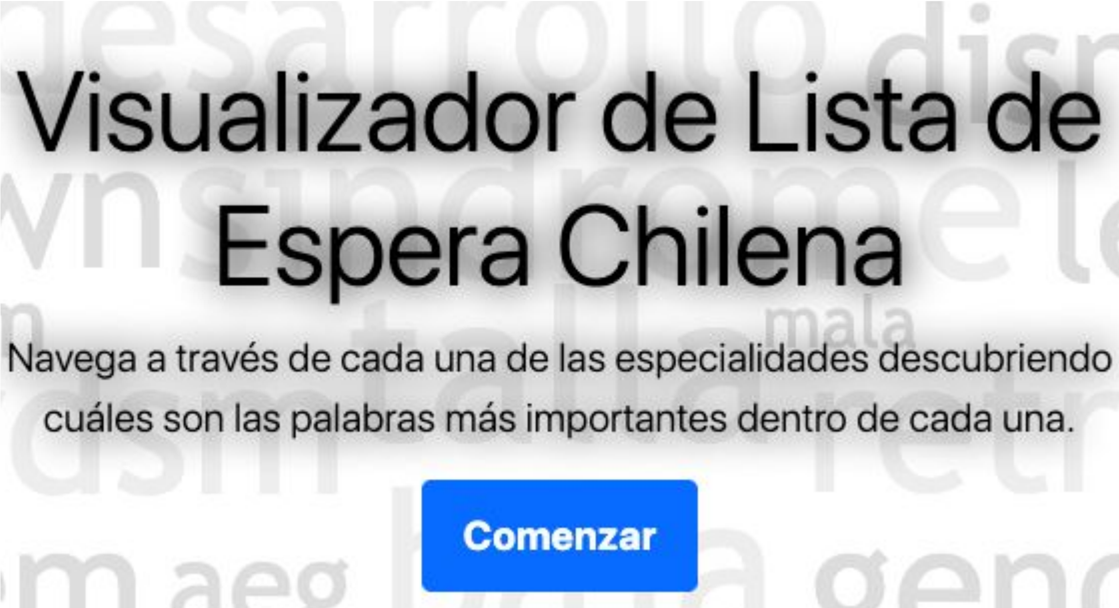
unstructured data

Can we improve the management of the Chilean waiting list?

Can we have a second use of the information? (Incidence of diseases for example?)

Visualization of free-text in national waiting list

cimt.uchile.cl/lechile



Visualizador de Lista de Espera Chilena

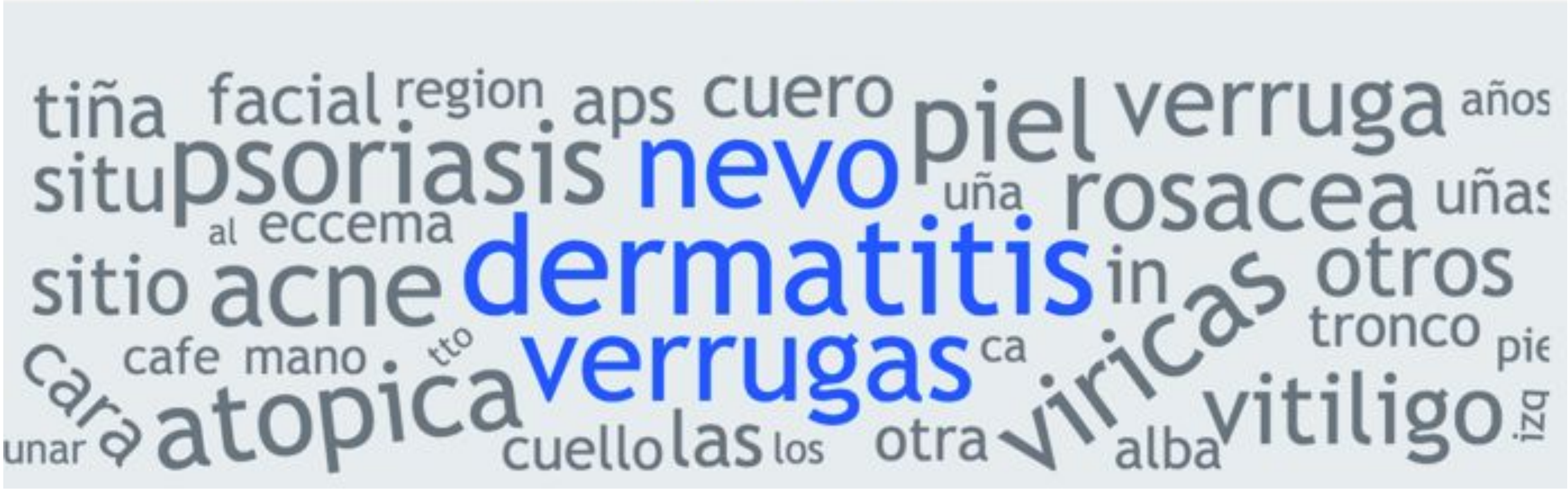
Navega a través de cada una de las especialidades descubriendo cuáles son las palabras más importantes dentro de cada una.

Fabián Villena
Odontologist, MSc (c) in Medical Informatics

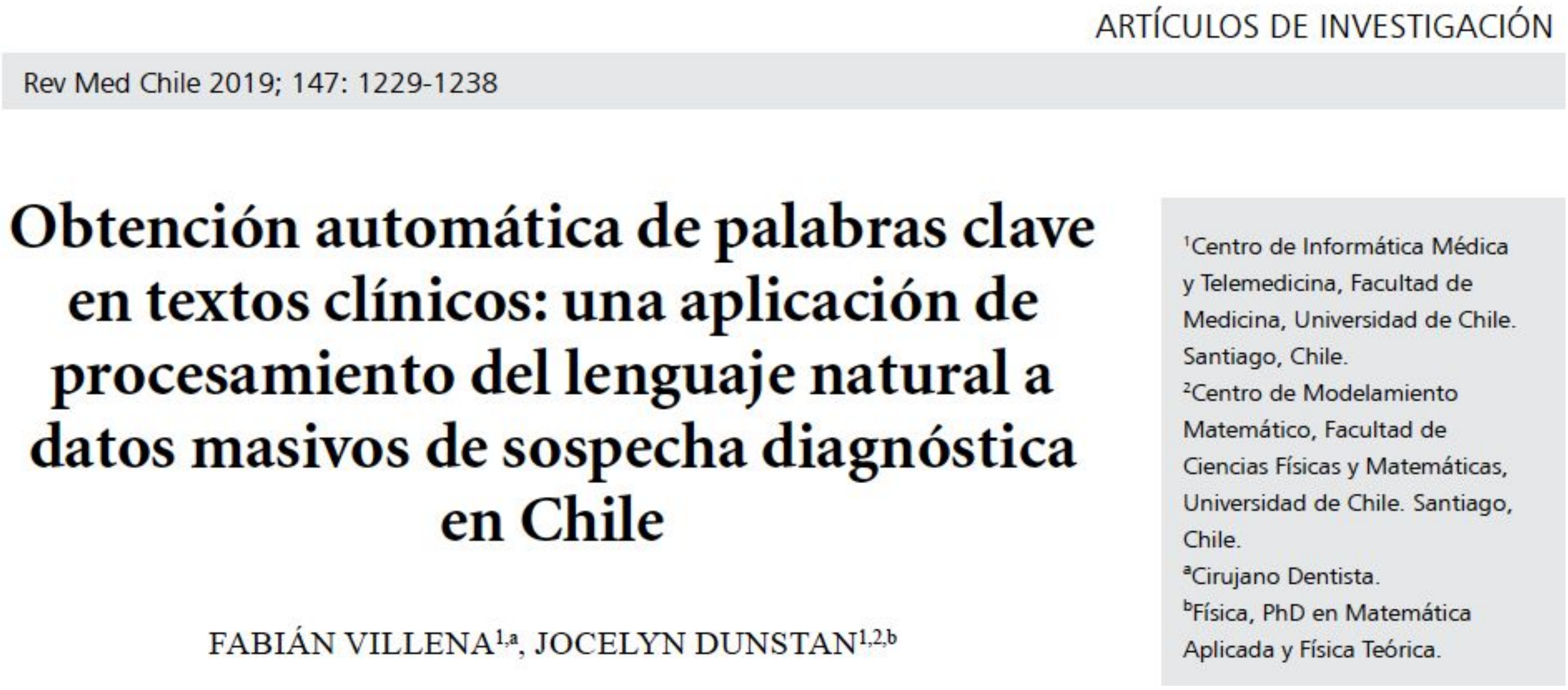
Part of the master in Medical Informatics of Fabián Villena

cimt.uchile.cl/lechile

DERMATOLOGIA



Master thesis of Fabián Villena (Medical Informatics)



Rev Med Chile 2019; 147: 1229-1238

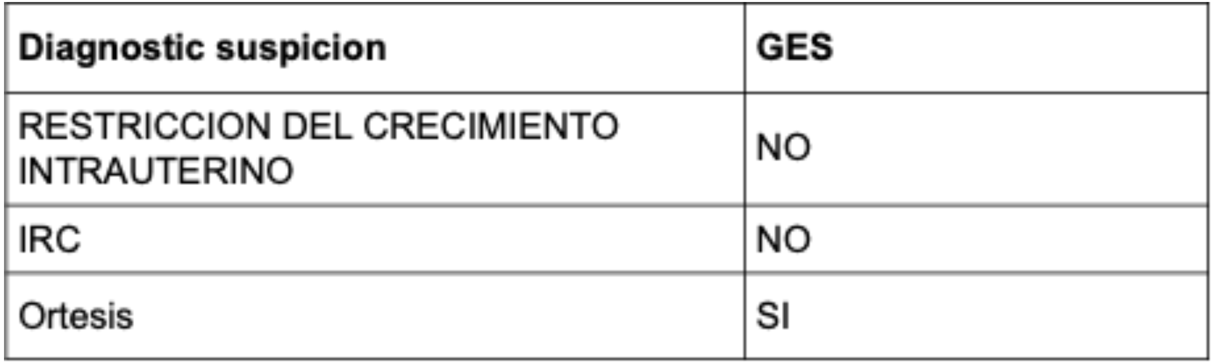
Obtención automática de palabras clave en textos clínicos: una aplicación de procesamiento del lenguaje natural a datos masivos de sospecha diagnóstica en Chile

FABIÁN VILLENA[1,a], JOCELYN DUNSTAN[1,2,b]

[1]Centro de Informática Médica y Telemedicina, Facultad de Medicina, Universidad de Chile. Santiago, Chile.
[2]Centro de Modelamiento Matemático, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile. Santiago, Chile.
[a]Cirujano Dentista.
[b]Física, PhD en Matemática Aplicada y Física Teórica.

Classification of referrals included or not in the Health Explicit Guarantees (GES)

GES and non-GES classification



Diagnostic suspicion	GES
RESTRICCION DEL CRECIMIENTO INTRAUTERINO	NO
IRC	NO
Ortesis	SI

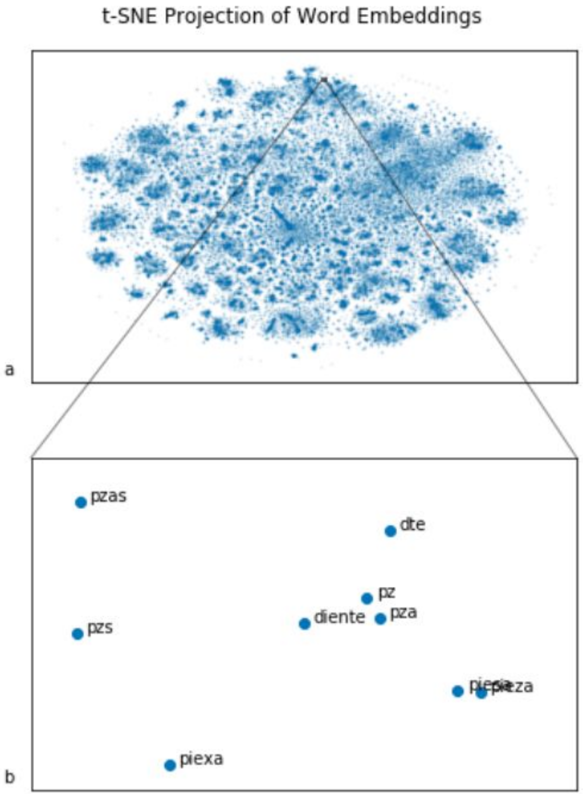
Free-text + age **Label**

Jorge Pérez
Assoc. Prof. DCC

Master thesis of Fabián Villena (Medical Informatics). In collaboration with Prof. Jorge Pérez (Computer Sciences)

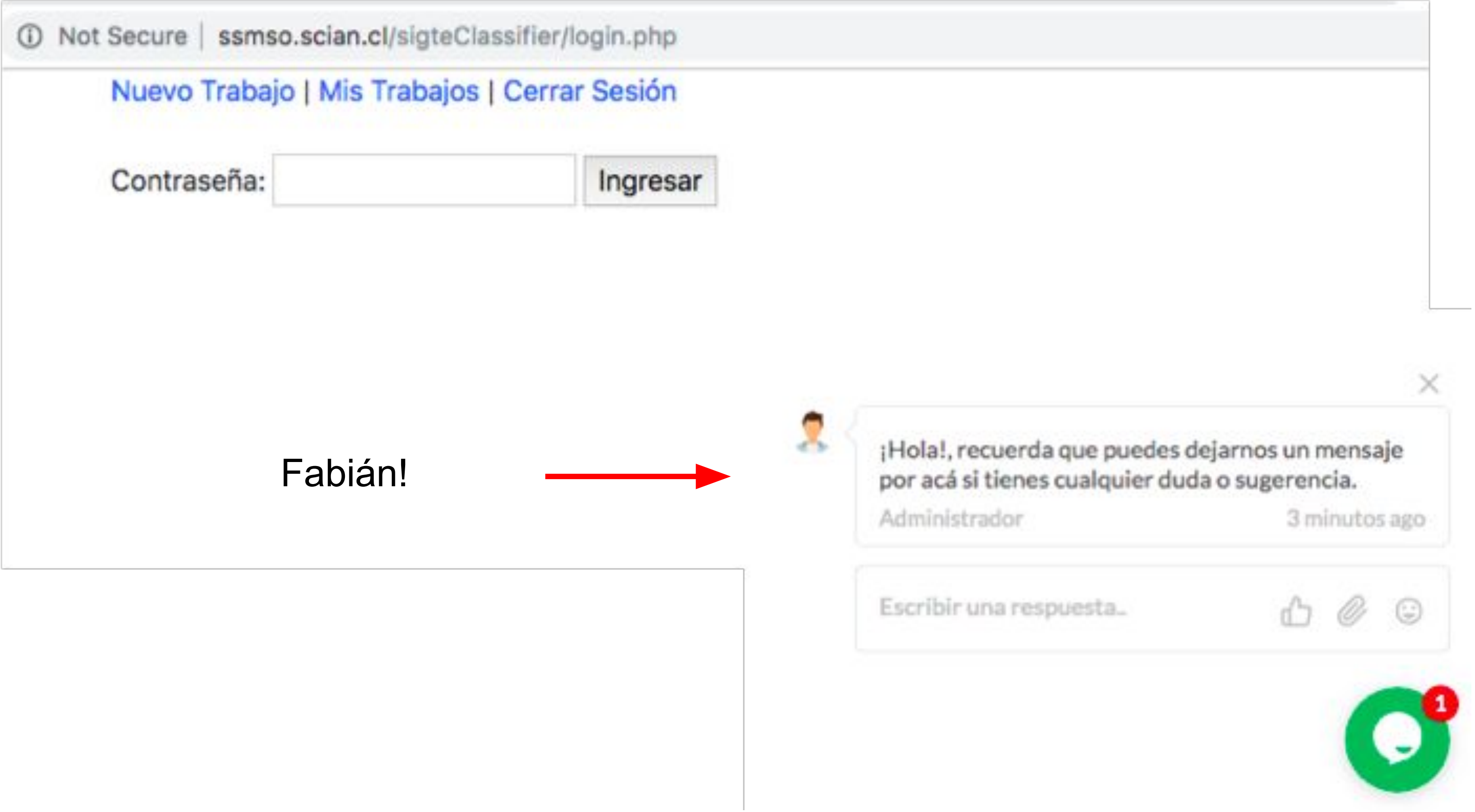
Word Embeddings

- Word embeddings in 300 dimensions were computed using 11 million referrals (56,079,828 word tokens and the vocabulary length is 252,513)
- Vector representation of referrals and age were the input of ML classifiers
- Our best performing algorithm reached an F1-score of 0.91 (RF)



Manuscript in preparation

7-months deployment in a hospital



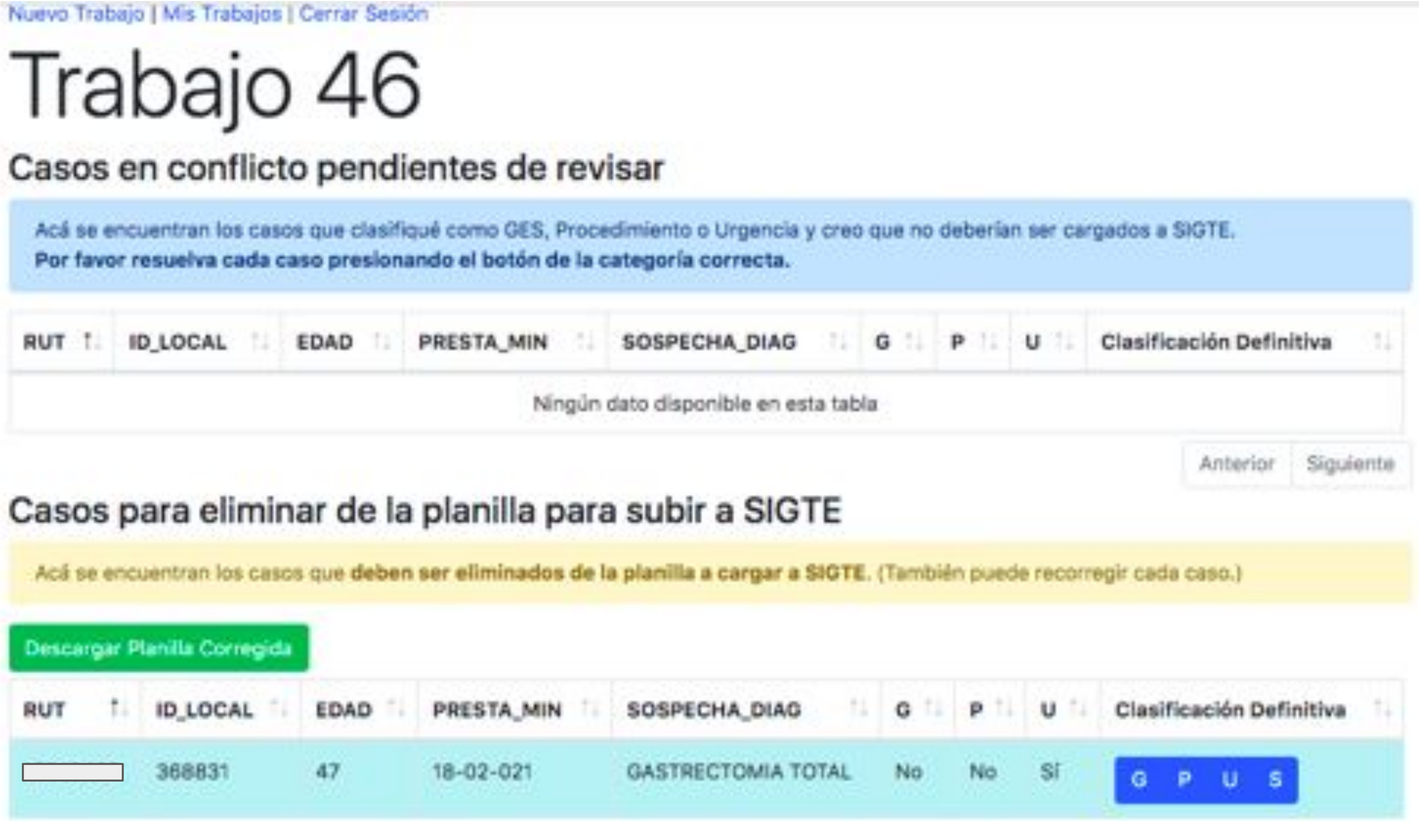
ⓘ Not Secure | ssmso.scian.cl/sigteClassifier/login.php
Nuevo Trabajo | Mis Trabajos | Cerrar Sesión
Ingresar
Escribir una respuesta..
1

Contraseña:

Fabián!

¡Hola!, recuerda que puedes dejarnos un mensaje por acá si tienes cualquier duda o sugerencia.

Administrador 3 minutos ago



Trabajo 46

Casos en conflicto pendientes de revisar

Acá se encuentran los casos que clasifiqué como GES, Procedimiento o Urgencia y creo que no deberían ser cargados a SIGTE.
Por favor resuelva cada caso presionando el botón de la categoría correcta.

RUT	ID_LOCAL	EDAD	PRESTA_MIN	SOSPECHA_DIAG	G	P	U	Clasificación Definitiva
Ningún dato disponible en esta tabla								

Casos para eliminar de la planilla para subir a SIGTE

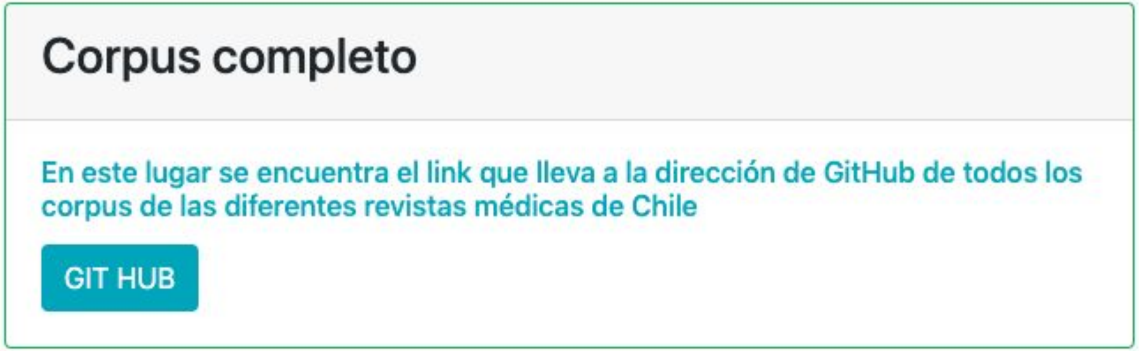
Acá se encuentran los casos que **deben ser eliminados de la planilla a cargar a SIGTE**. (También puede recorregir cada caso.)

RUT	ID_LOCAL	EDAD	PRESTA_MIN	SOSPECHA_DIAG	G	P	U	Clasificación Definitiva
[redacted]	368831	47	18-02-021	GASTRECTOMIA TOTAL	No	No	Sí	G P U S

Manuscript in preparation

Biomedical corpus from medical articles

Corpus médico en español de revistas médicas de Chile



Corpus completo

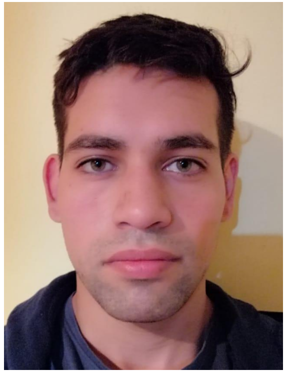
En este lugar se encuentra el link que lleva a la dirección de GitHub de todos los corpus de las diferentes revistas médicas de Chile

http://corpusmedico.cimt.cl/

Publicación corpus para el libre acceso de la comunidad

Se subió a GitHub y a una página web propia para que la comunidad pueda acceder (Link QR)

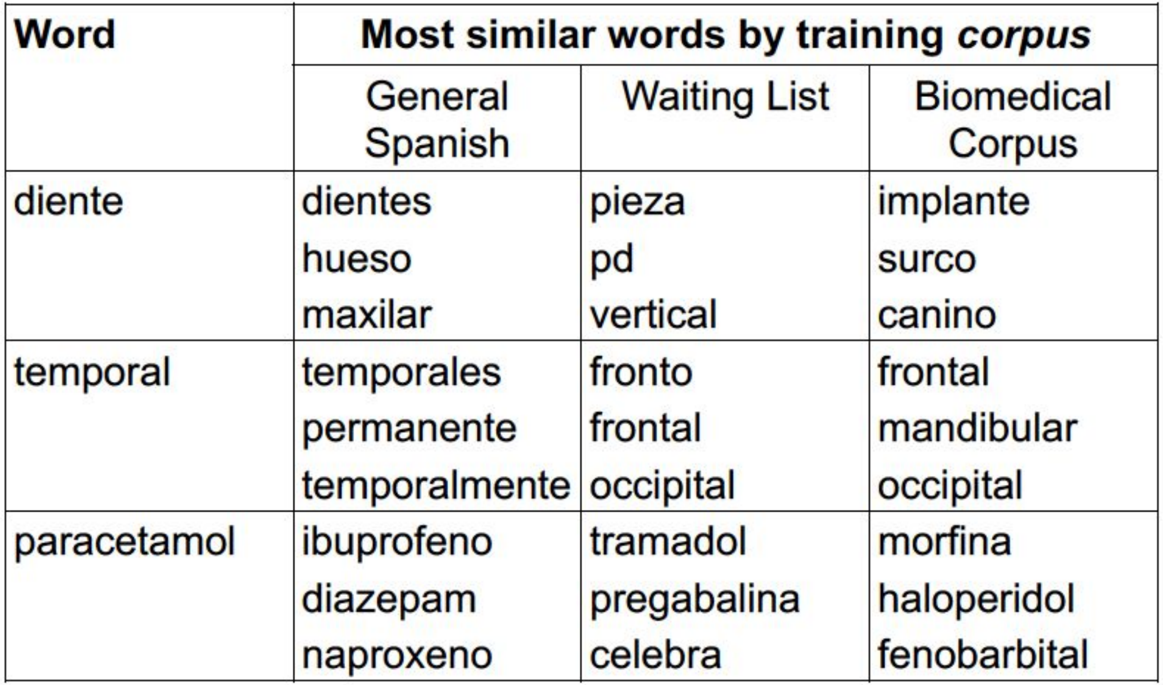
- 13 medical journals available in SciELO
- BeautifulSoup was used to preprocessed the text
- 13,000 text files equivalent to 1 GB
- 140 million tokens and 373,268 vocabulary.



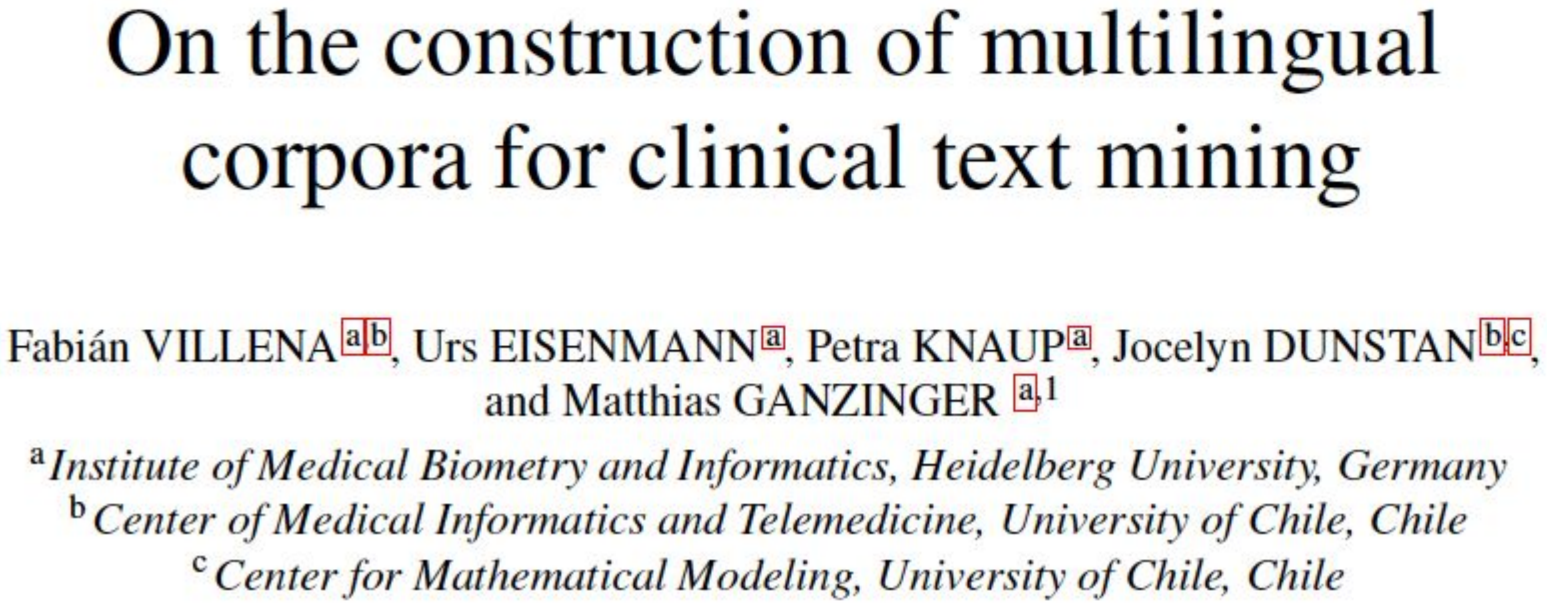
Manuel Durán
Graduated in Medicine,
MSc(c) in Medical
Informatics

Manuel Duran and Fabián Villena's work.

Word embedding computation with different training corpus



Word	Most similar words by training *corpus*		
	General Spanish	Waiting List	Biomedical Corpus
diente	dientes hueso maxilar	pieza pd vertical	implante surco canino
temporal	temporales permanente temporalmente	fronto frontal occipital	frontal mandibular occipital
paracetamol	ibuprofeno diazepam naproxeno	tramadol pregabalina celebra	morfina haloperidol fenobarbital



On the construction of multilingual corpora for clinical text mining

Fabián VILLENA[a,b], Urs EISENMANN[a], Petra KNAUP[a], Jocelyn DUNSTAN[b,c], and Matthias GANZINGER [a,1]

[a] *Institute of Medical Biometry and Informatics, Heidelberg University, Germany*

[b] *Center of Medical Informatics and Telemedicine, University of Chile, Chile*

[c] *Center for Mathematical Modeling, University of Chile, Chile*

Accepted in Medical Informatics Europe (MIE2020)

- German Medical Science Database (https://www.egms.de/)
- Chilean Scientific Electronic Library Online SciELO (https://scielo.conicyt.cl)

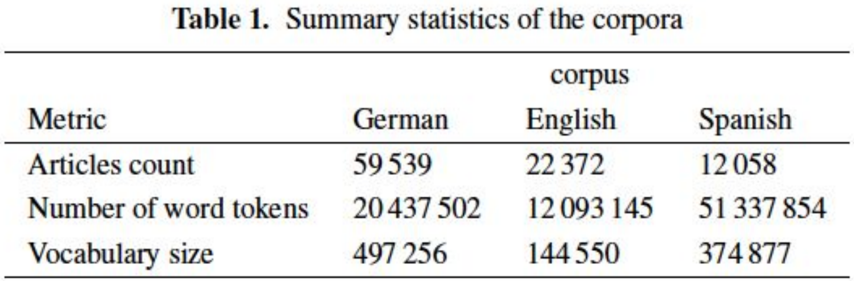


Table 1. Summary statistics of the corpora

	corpus		
Metric	German	English	Spanish
Articles count	59 539	22 372	12 058
Number of word tokens	20 437 502	12 093 145	51 337 854
Vocabulary size	497 256	144 550	374 877

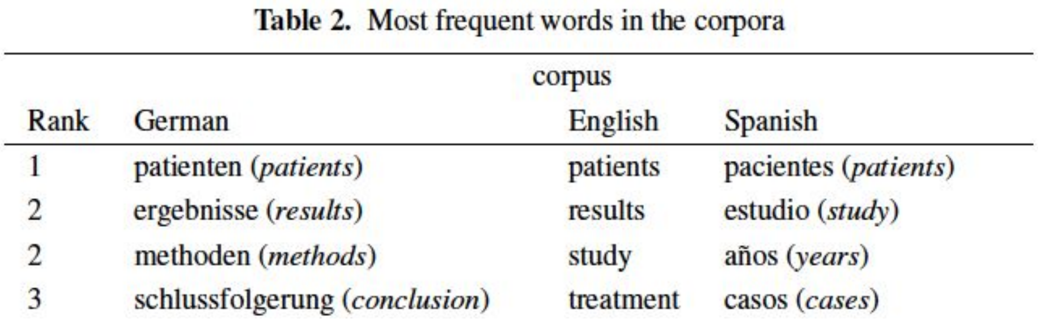
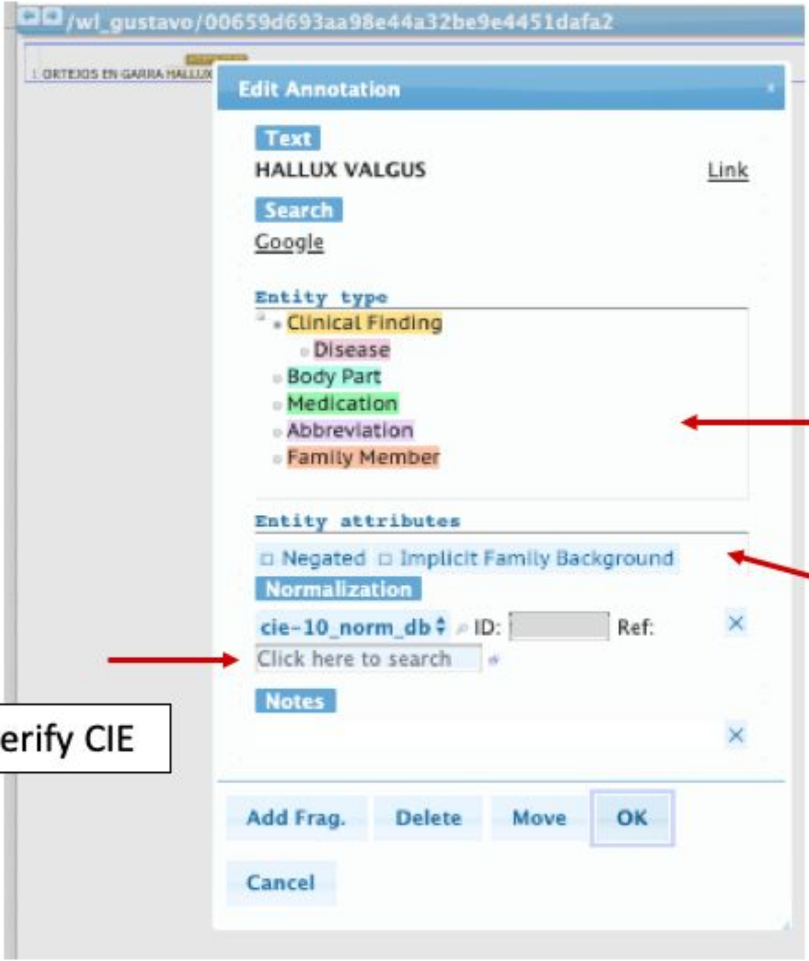


Table 2. Most frequent words in the corpora

		corpus	
Rank	German	English	Spanish
1	patienten (*patients*)	patients	pacientes (*patients*)
2	ergebnisse (*results*)	results	estudio (*study*)
2	methoden (*methods*)	study	años (*years*)
3	schlussfolgerung (*conclusion*)	treatment	casos (*cases*)

Annotating clinical text



/wl_gustavo/00659d693aa98e44a32be9e4451dafa2

1 ORTEJOS EN GARRA HALLUX

Edit Annotation

Text

HALLUX VALGUS — Link

Search

Google

Entity type

- Clinical Finding
  - Disease
- Body Part
- Medication
- Abbreviation
- Family Member

Entity attributes

□ Negated □ Implicit Family Background

Normalization

cie-10_norm_db ID: Ref:

Click here to search

Notes

Add Frag. | Delete | Move | OK

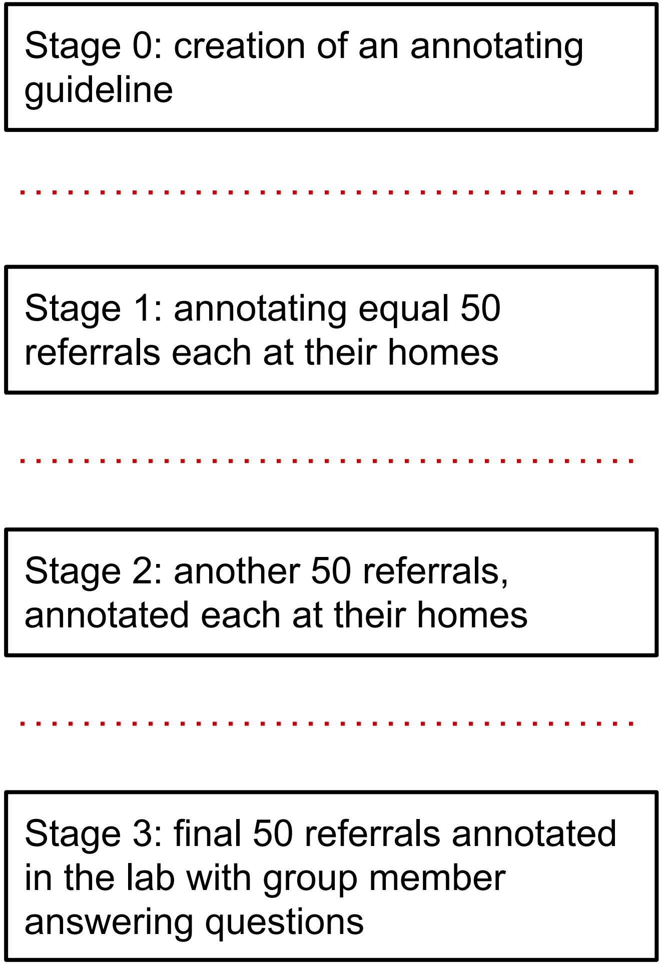
Cancel

Clinical entities

Additional attributes

Possiblity to verify CIE

Three independent labelers (medicine students)



Stage 0: creation of an annotating guideline

Meeting with the PI that introduced the idea and importance of annotating clinical text and the annotating guideline

Stage 1: annotating equal 50 referrals each at their homes

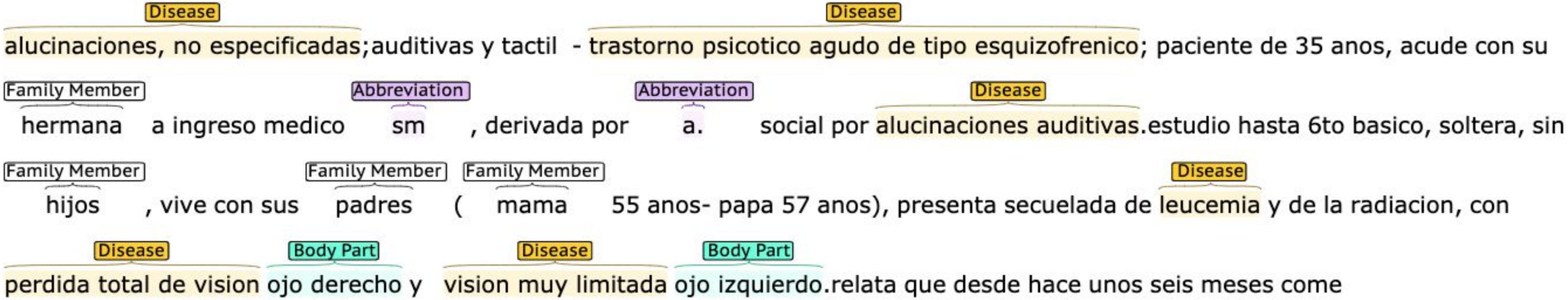
Discussion of the agrement results and analysing together some examples. Annotation guideline updated to clarify some points . WE DIDN’T FORCE AGREEMENT

Stage 2: another 50 referrals, annotated each at their homes

Percentage of agreement and Cohen’s kappa calculated

Stage 3: final 50 referrals annotated in the lab with group member answering questions

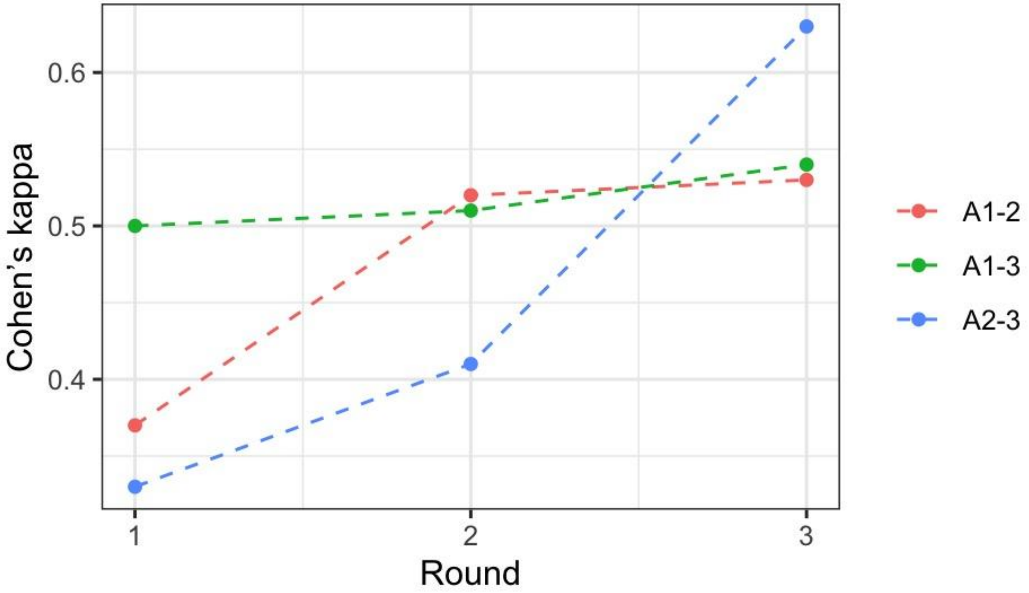
Annotation for named entity recognition



alucinaciones, no especificadas;auditivas y tactil - trastorno psicotico agudo de tipo esquizofrenico; paciente de 35 anos, acude con su
hermana a ingreso medico sm , derivada por a. social por alucinaciones auditivas.estudio hasta 6to basico, soltera, sin
hijos , vive con sus padres (mama 55 anos- papa 57 anos), presenta secuelada de leucemia y de la radiacion, con
perdida total de vision ojo derecho y vision muy limitada ojo izquierdo.relata que desde hace unos seis meses come

Done with Pablo Baez (PhD in biomedical sciences) and Fabián Villena

Inter-annotator agreement

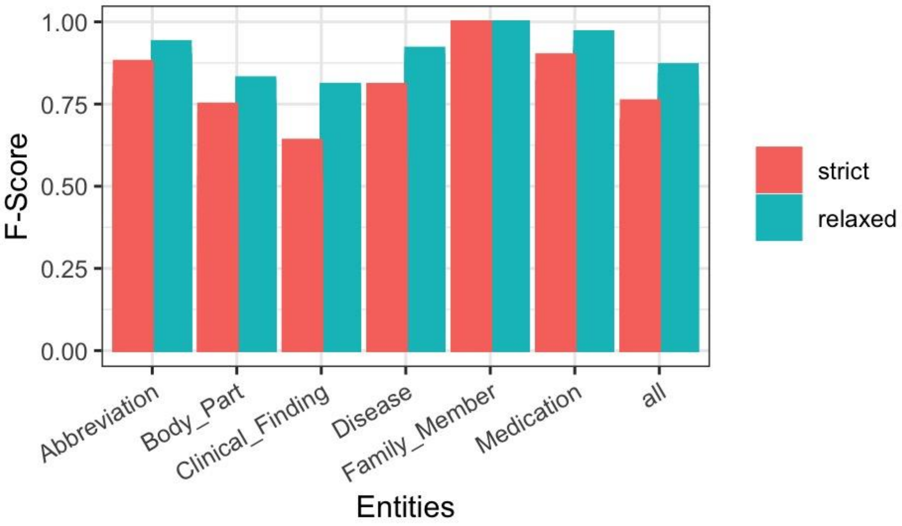


Pablo Báez
Microbiologist, PhD
in Biomedical
Sciences

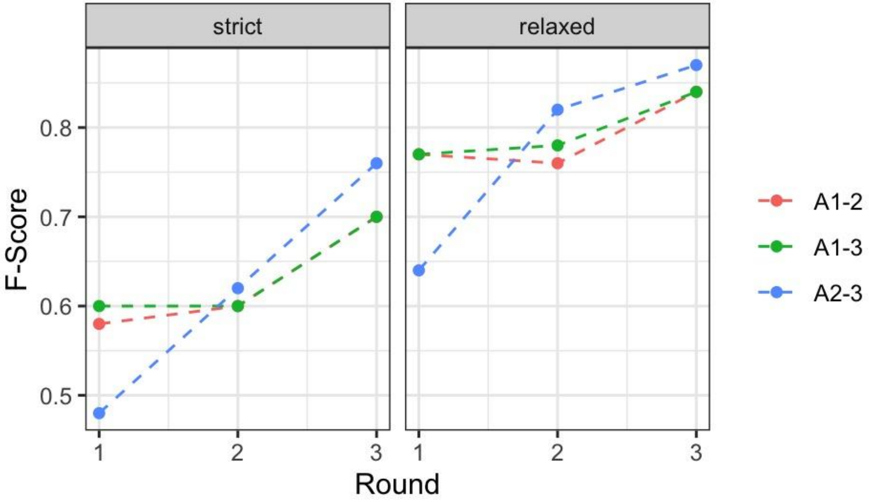
Done with Pablo Baez and Fabián Villena

Inter-annotator agreement

By entity type



Along the training period



Analysis done by Pablo Baez

Conclusions and future work

- We are trying to build a clinical text mining core in Chile incorporating human resources from Medicine and Engineering faculties
- Our aim is to create *corpora* and language resources and share them with the community. To secure funding we need to have a task in mind, and probably will be NER but I also like expansion of abbreviations and word disambiguation.
- There are challenges to overcome: get funding to expand the classifier, start coming to key conferences, get publications, secure more funding, annotate better and more text, have a task!
- To collaborate with Spain and other latin american countries to have pan-hispanic clinical text resources and a comparison between countries is a good big goal I think 😬

Thanks a lot for your attention!

jdunstan@uchile.cl
sites.google.com/view/jdunstan