WWW.OEAW.AC.AT

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Crowdsourcing Linguistic Annotations

Tanja Wissik

Austrian Academy of Sciences

SLOVKO 2021, 14 October 2021

ACDH-CH – AUSTRIAN CENTRE FOR DIGITAL HUMANITIES AND CULTURAL HERITAGE

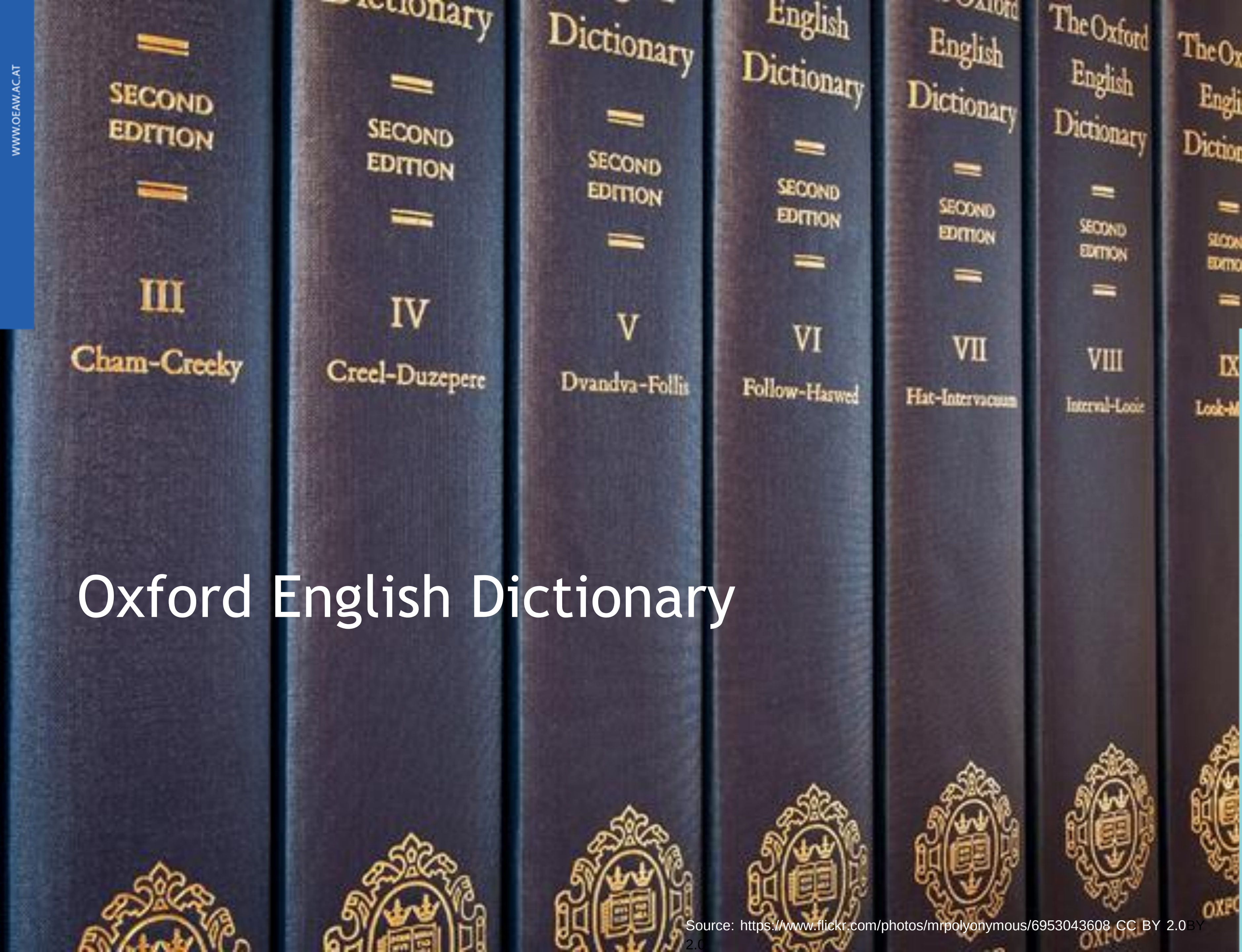
crowdsourcing
annotation
based
nlp
corpora
corpus
best
practice
guidelines
approach
platform
games
expert
sentiment
languages
create
approaches
problems
contributions
crowd
definition
practices
volunteers
crowdsourced
range
large
data
collecting
training
human
sense
support
used
automatically
annotations
domain
using
algorithms
set
key
ace
answers
community
creation
word
genres
volunteer
gate
forward
evaluation
collaborative
purpose
knowledge
acquisition
annotated

“Crowdsourcing is the process of leveraging public participation in or contributions to projects and activities.”

(Hedges & Dunn 2017)

“Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken.”

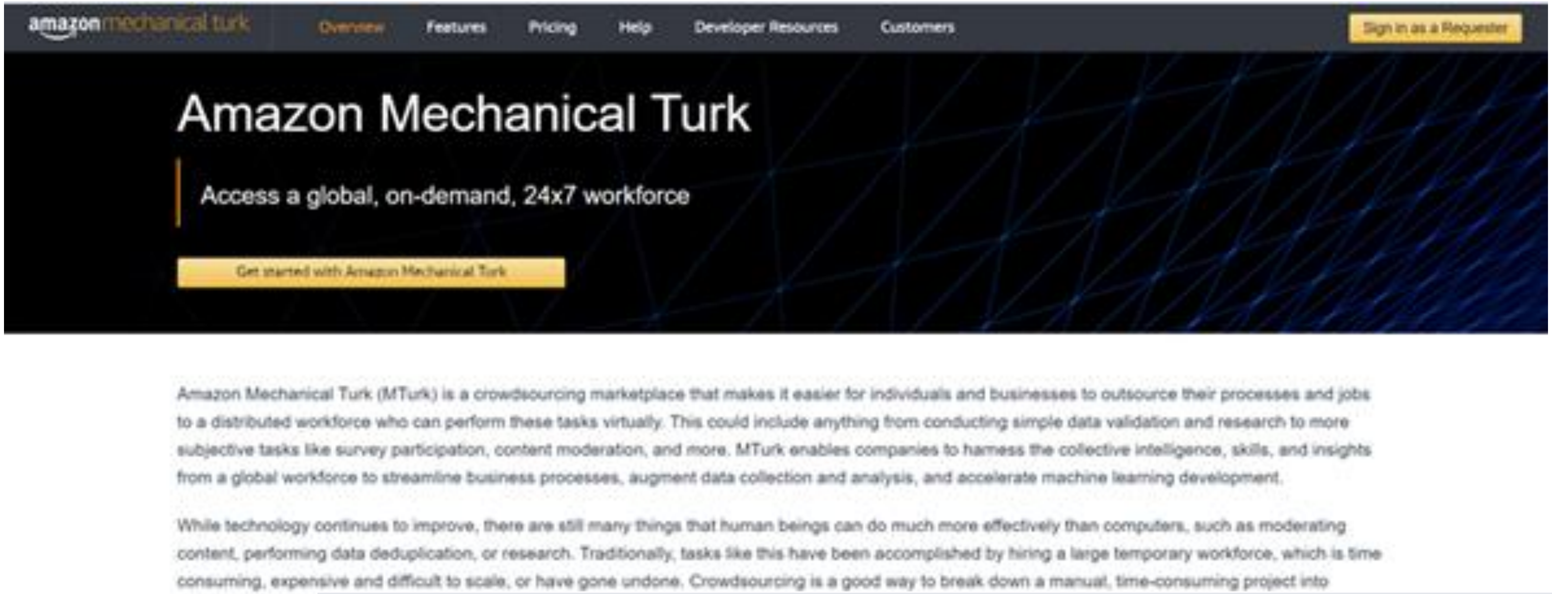
(Estellés-Arolas & González-Ladrón-de-Guevara 2012)



Oxford English Dictionary

Source: https://www.flickr.com/photos/mrpolyonymous/6953043608 CC BY 2.0

BY 2.0

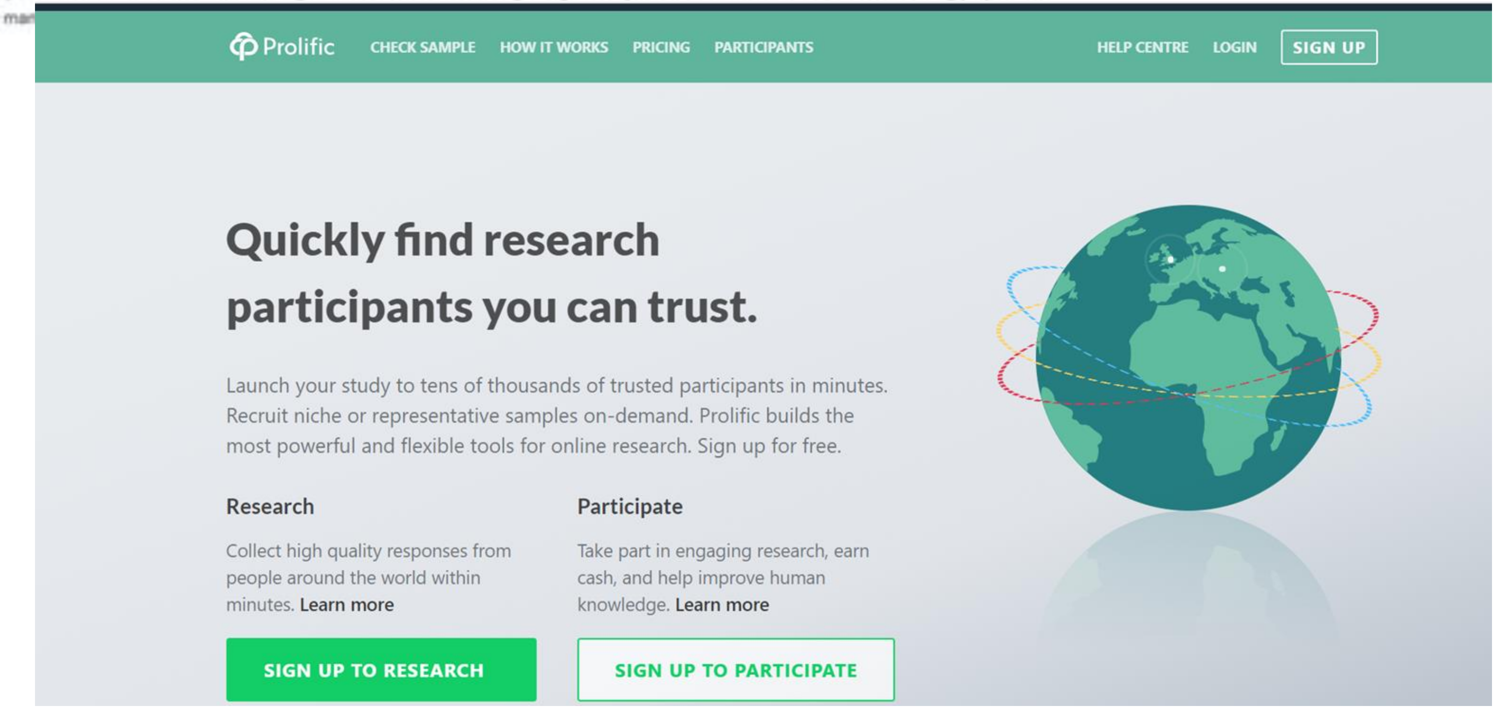


Amazon Mechanical Turk

Access a global, on-demand, 24x7 workforce

Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. This could include anything from conducting simple data validation and research to more subjective tasks like survey participation, content moderation, and more. MTurk enables companies to harness the collective intelligence, skills, and insights from a global workforce to streamline business processes, augment data collection and analysis, and accelerate machine learning development.

While technology continues to improve, there are still many things that human beings can do much more effectively than computers, such as moderating content, performing data deduplication, or research. Traditionally, tasks like this have been accomplished by hiring a large temporary workforce, which is time consuming, expensive and difficult to scale, or have gone undone. Crowdsourcing is a good way to break down a manual, time-consuming project into smaller, more man



Quickly find research participants you can trust.

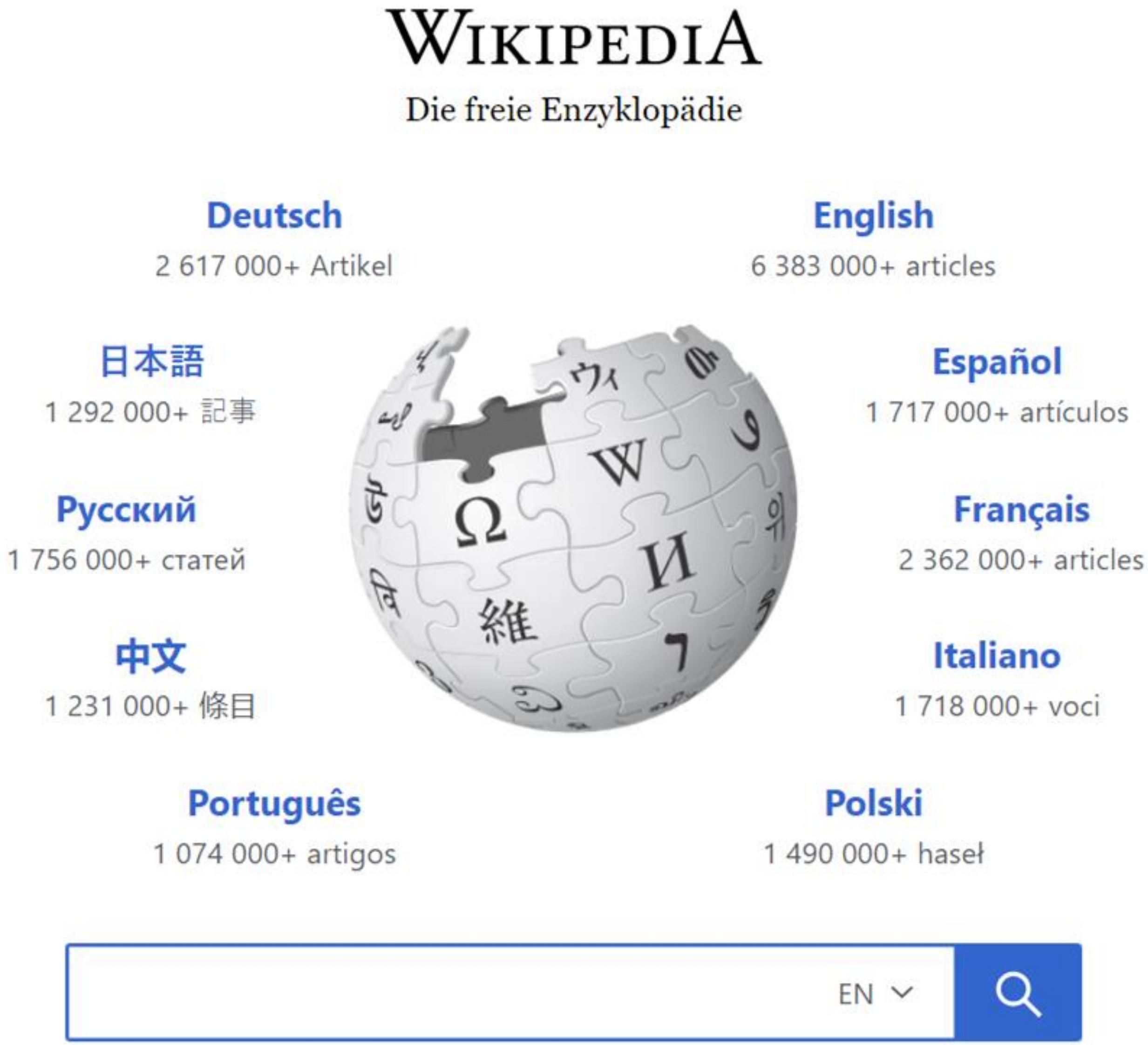
Launch your study to tens of thousands of trusted participants in minutes. Recruit niche or representative samples on-demand. Prolific builds the most powerful and flexible tools for online research. Sign up for free.

Research

Collect high quality responses from people around the world within minutes. Learn more

Participate

Take part in engaging research, earn cash, and help improve human knowledge. Learn more



Wikipedia

Die freie Enzyklopädie

Deutsch
2 617 000+ Artikel

English
6 383 000+ articles

日本語
1 292 000+ 記事

Español
1 717 000+ artículos

Русский
1 756 000+ статей

Français
2 362 000+ articles

中文
1 231 000+ 條目

Italiano
1 718 000+ voci

Português
1 074 000+ artigos

Polski
1 490 000+ haseł

EN

ZOONIVERSE
PLANET FOUR
GALAXY ZOO
THE PLASTIC TIDE
BASH THE BUG
SHAKESPEARE'S WORLD
MANATEE CHAT

Source: https://zooniverseblog.files.wordpress.com/2020/04/download.png?w=768&h=425&crop=1

Games with a purpose

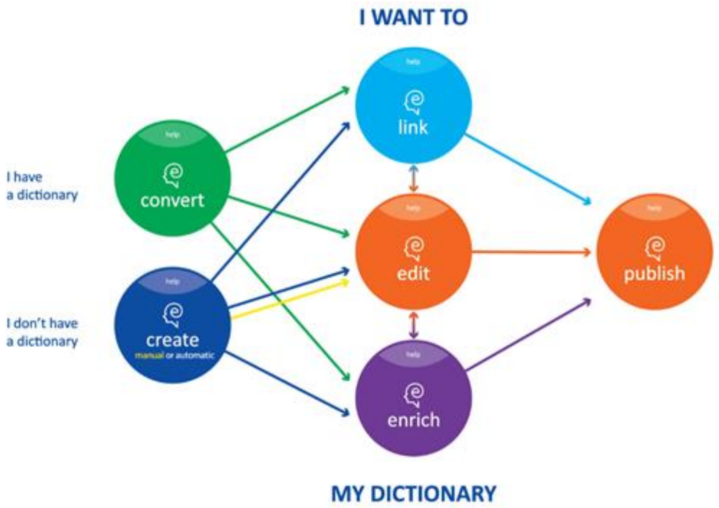
WWW.OEAW.AC.AT

ACDH-CH – AUSTRIAN CENTRE FOR DIGITAL HUMANITIES AND CULTURAL HERITAGE

Crowdsourcing Examples from my Recent Projects

ELEXIS - European Lexicographic Infrastructure

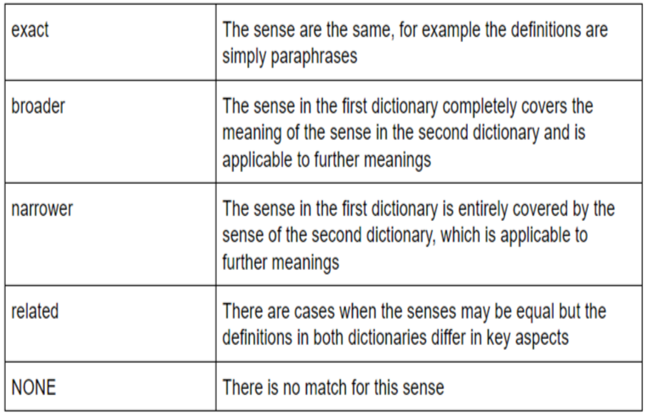
- H2020 Project
- February 2018 - July 2022
- 17 Partners, 52 Observers
- Integrate, extend and harmonise national and regional efforts in the field of lexicography, both modern and historical
- Create a sustainable infrastructure



https://elex.is

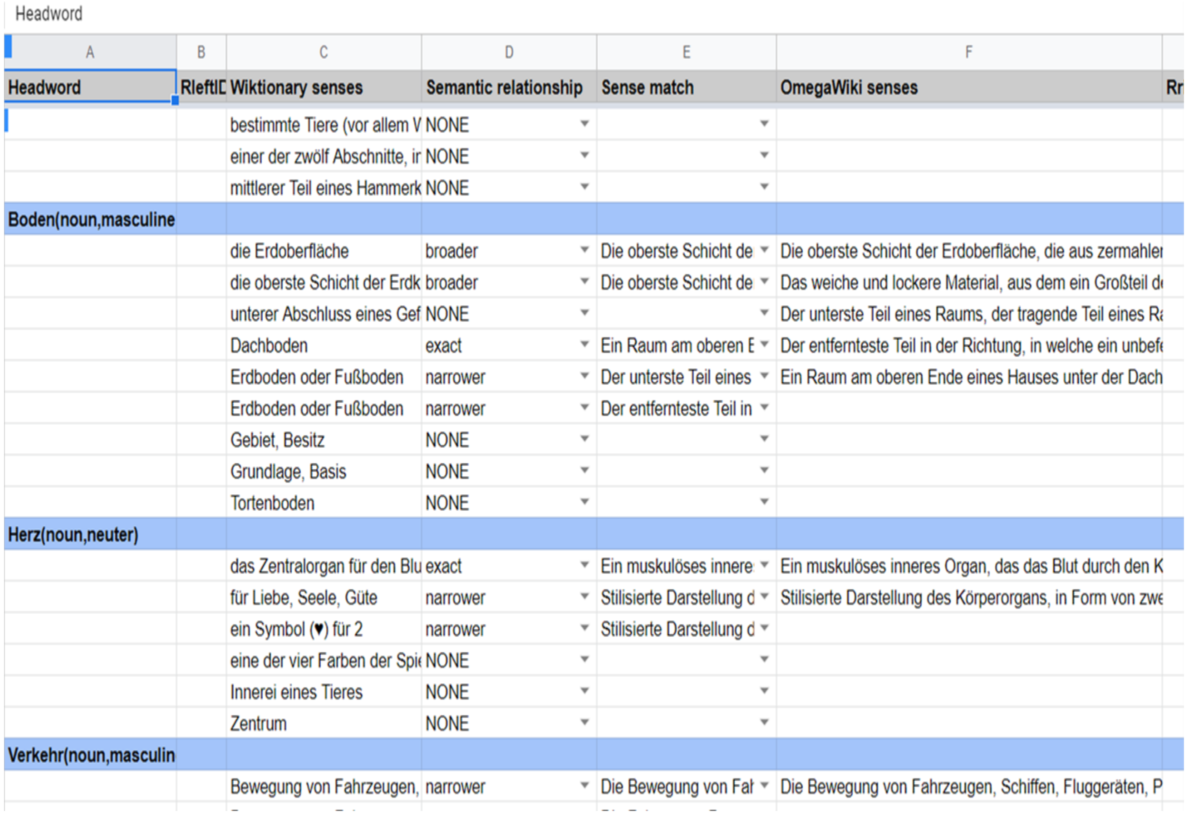
ELEXIS: Annotation of Semantic Relations

Semantic Relation



exact	The sense are the same, for example the definitions are simply paraphrases
broader	The sense in the first dictionary completely covers the meaning of the sense in the second dictionary and is applicable to further meanings
narrower	The sense in the first dictionary is entirely covered by the sense of the second dictionary, which is applicable to further meanings
related	There are cases when the senses may be equal but the definitions in both dictionaries differ in key aspects
NONE	There is no match for this sense

Annotated Data: Example for German



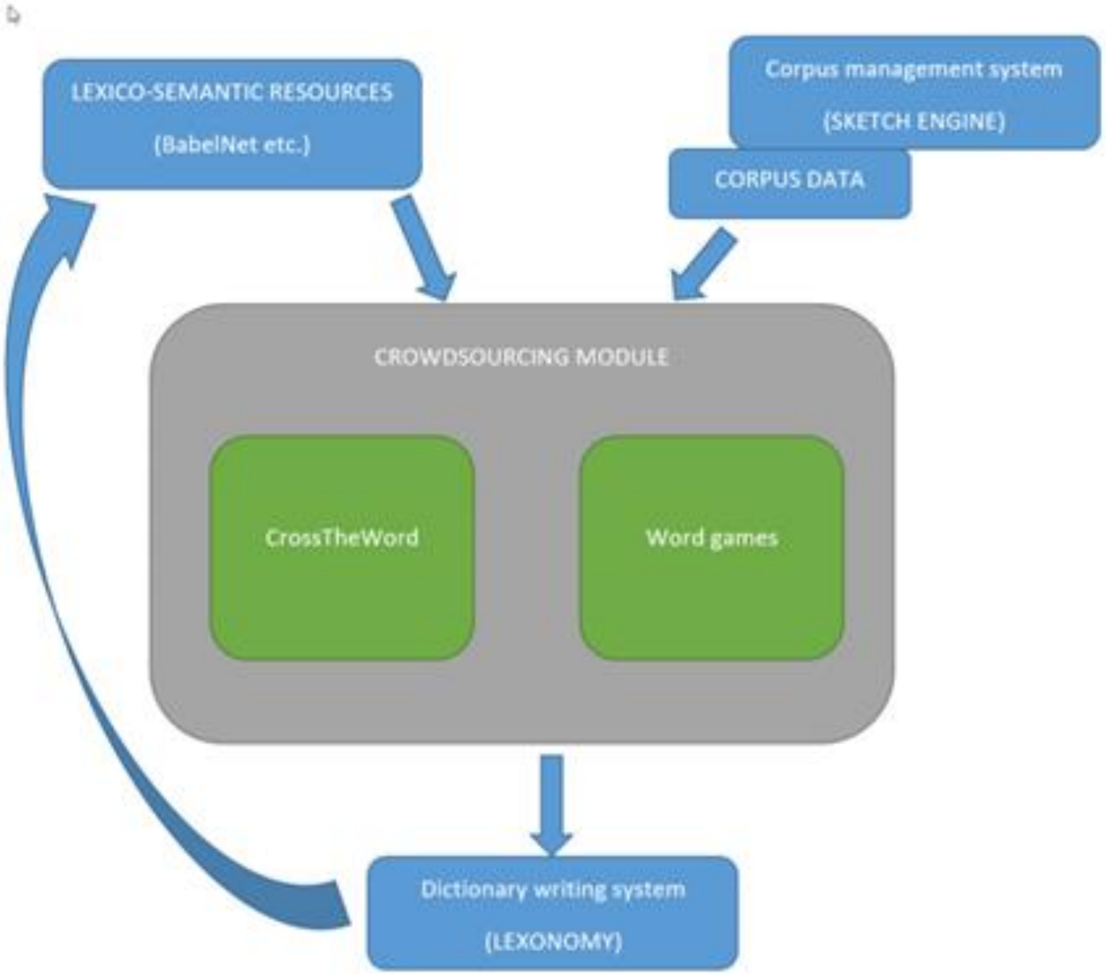
Headword

A	B	C	D	E	F	
Headword	RleftIL	Wiktionary senses	Semantic relationship	Sense match	OmegaWiki senses	Rr
		bestimmte Tiere (vor allem V	NONE			
		einer der zwölf Abschnitte, ir	NONE			
		mittlerer Teil eines Hammerk	NONE			
Boden(noun,masculine						
		die Erdoberfläche	broader	Die oberste Schicht de	Die oberste Schicht der Erdoberfläche, die aus zermahler	
		die oberste Schicht der Erdk	broader	Die oberste Schicht de	Das weiche und lockere Material, aus dem ein Großteil de	
		unterer Abschluss eines Gef	NONE		Der unterste Teil eines Raums, der tragende Teil eines Ra	
		Dachboden	exact	Ein Raum am oberen E	Der entfernteste Teil in der Richtung, in welche ein unbefe	
		Erdboden oder Fußboden	narrower	Der unterste Teil eines	Ein Raum am oberen Ende eines Hauses unter der Dach	
		Erdboden oder Fußboden	narrower	Der entfernteste Teil in		
		Gebiet, Besitz	NONE			
		Grundlage, Basis	NONE			
		Tortenboden	NONE			
Herz(noun,neuter)						
		das Zentralorgan für den Blu	exact	Ein muskulöses innere	Ein muskulöses inneres Organ, das das Blut durch den K	
		für Liebe, Seele, Güte	narrower	Stilisierte Darstellung d	Stilisierte Darstellung des Körperorgans, in Form von zwe	
		ein Symbol (♥) für 2	narrower	Stilisierte Darstellung d		
		eine der vier Farben der Spi	NONE			
		Innerei eines Tieres	NONE			
		Zentrum	NONE			
Verkehr(noun,masculin						
		Bewegung von Fahrzeugen,	narrower	Die Bewegung von Fal	Die Bewegung von Fahrzeugen, Schiffen, Fluggeräten, P	

(cf. Ahmadi et al. 2020)

european lexicographic infrastructure

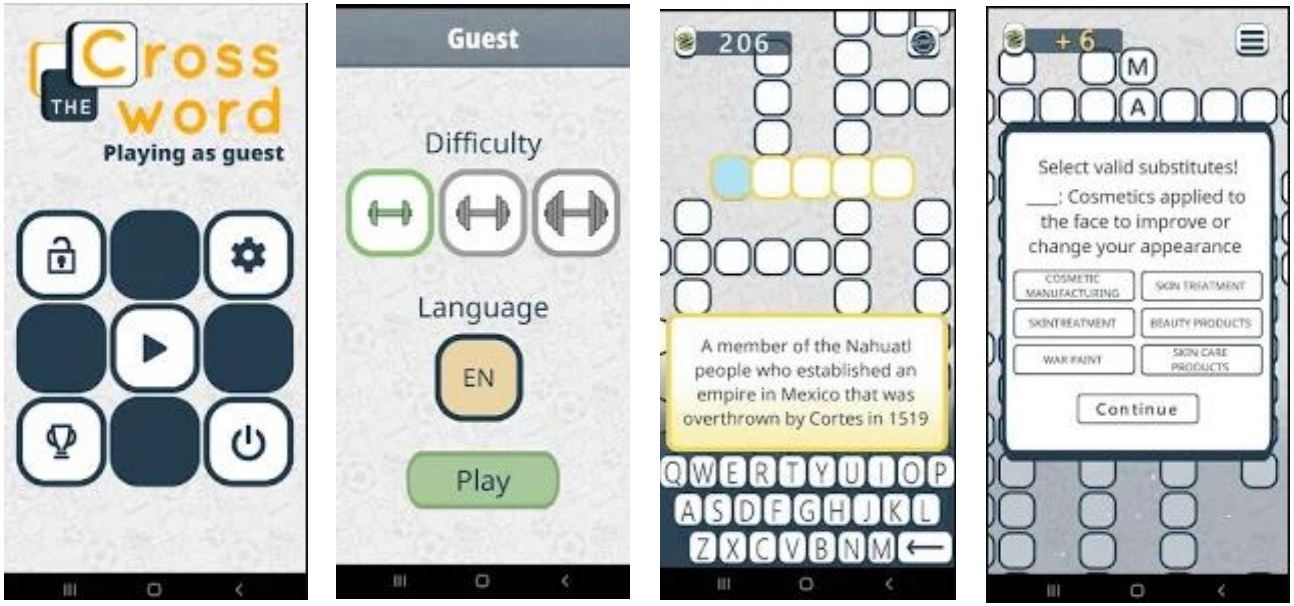
ELEXIS: Crowdsourcing Module



(ELEXIS D4.3 Crowdsourcing Module)

ELEXIS: CrossTheWord

CrossTheWord



A member of the Nahuatl people who established an empire in Mexico that was overthrown by Cortes in 1519

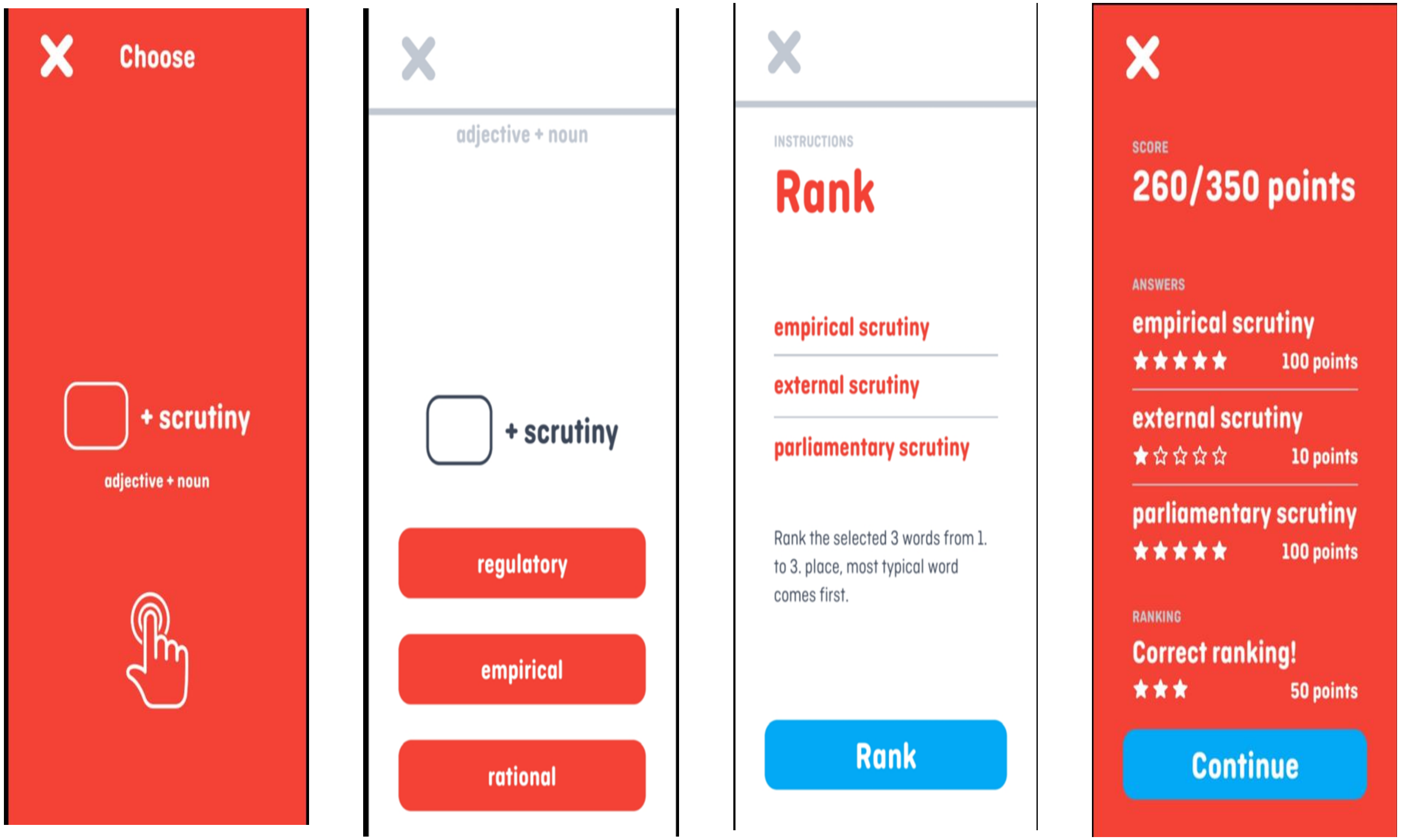
Select valid substitutes!

____: Cosmetics applied to the face to improve or change your appearance

(ELEXIS D4.3 Crowdsourcing Module)

ELEXIS: Game of Words

Game of words

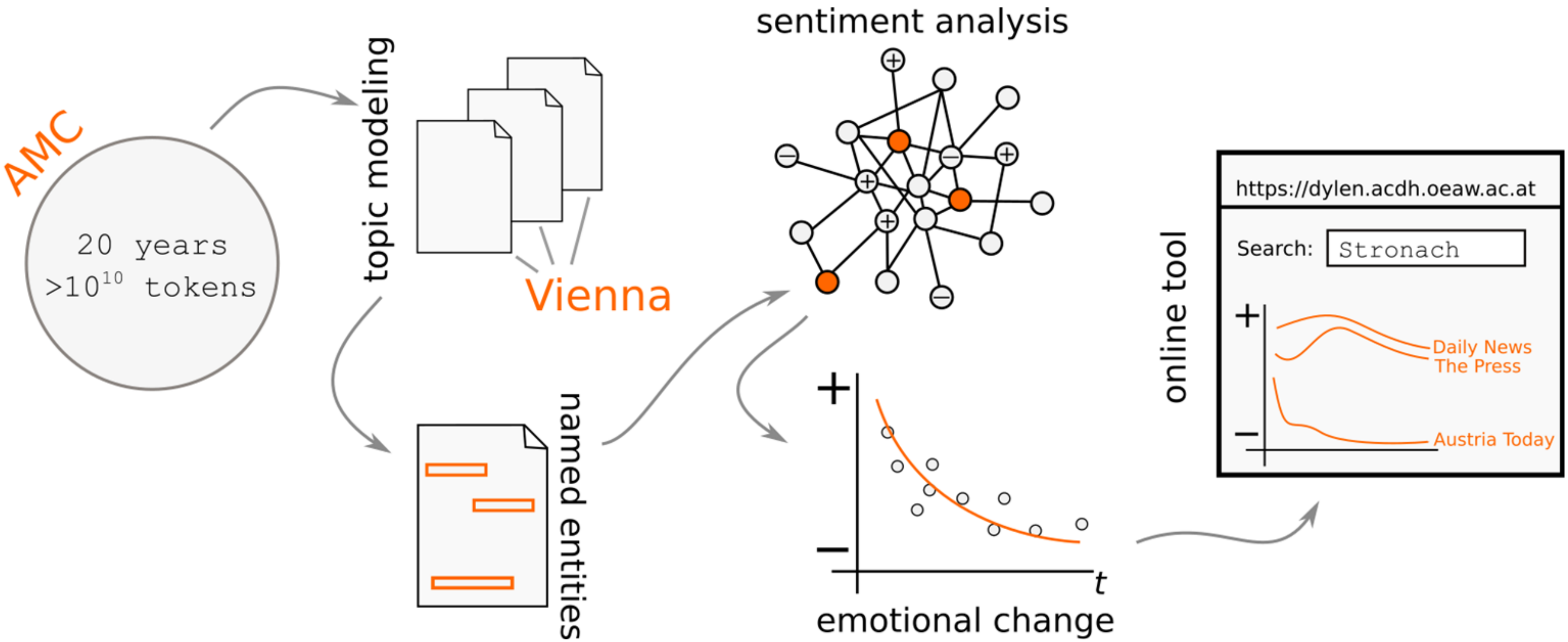


Rank the selected 3 words from 1. to 3. place, most typical word comes first.

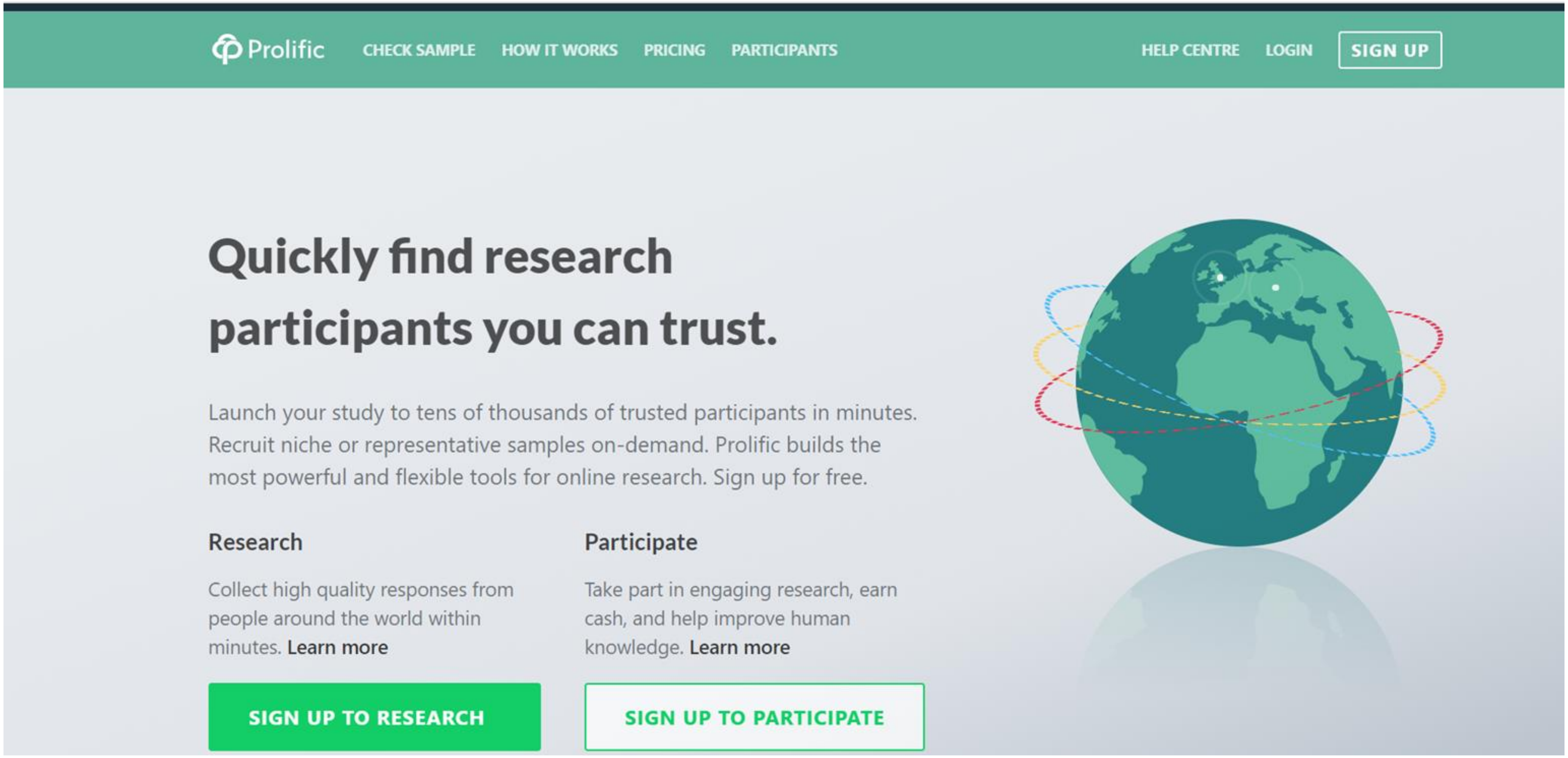
(ELEXIS D4.3 Crowdsourcing Module)

DYSEN: Sentiment Annotation of Media Text Snippets

- Dynamic Sentiment Analysis as Emotional Compass for the Digital Media Landscape (Original German Title: Dynamische Sentimentanalyse als emotionaler Kompass für die digitale Medienlandschaft)



Crowdsourcing Platform: Prolific



Quickly find research participants you can trust.

Launch your study to tens of thousands of trusted participants in minutes. Recruit niche or representative samples on-demand. Prolific builds the most powerful and flexible tools for online research. Sign up for free.

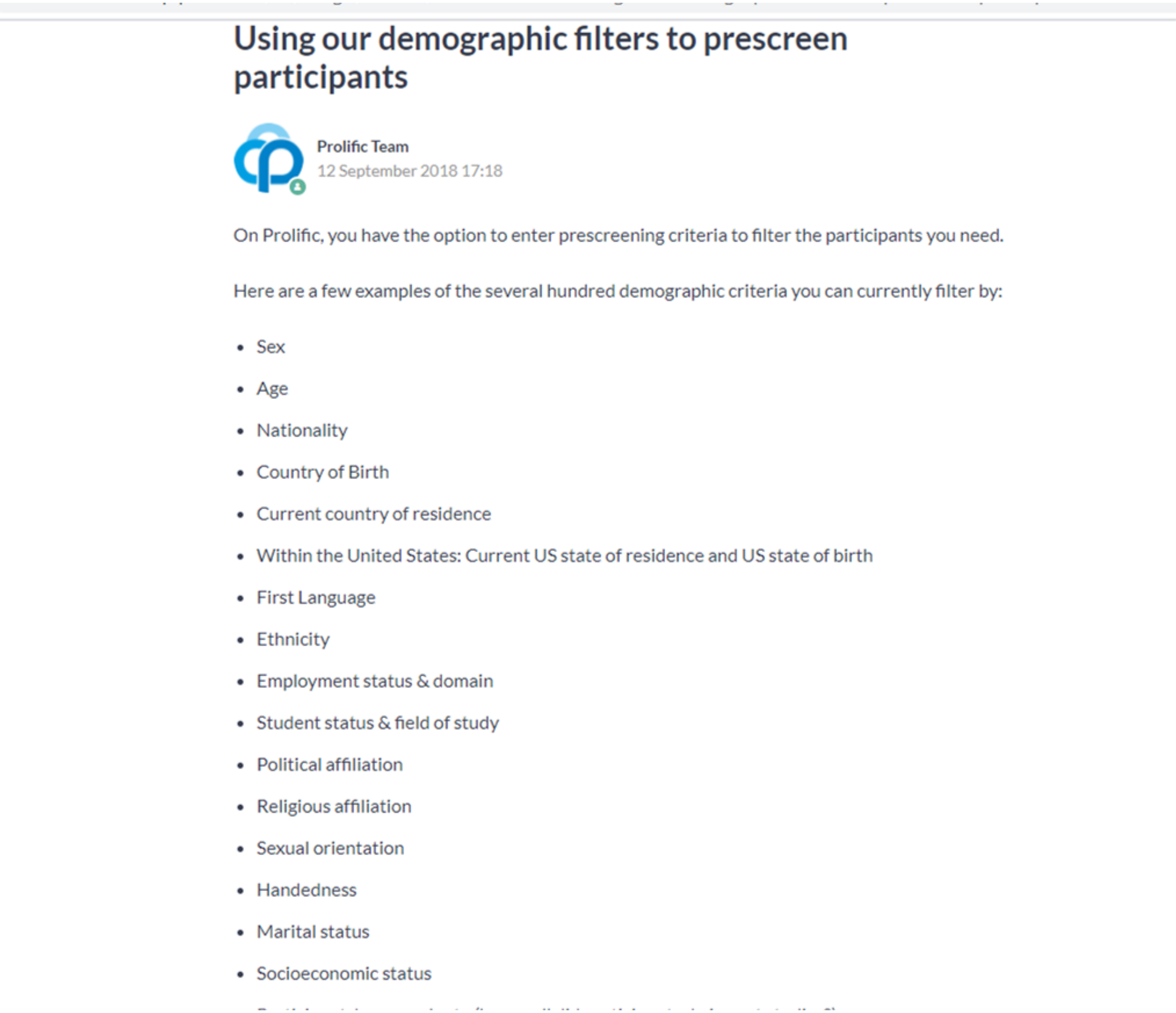
Research

Collect high quality responses from people around the world within minutes. Learn more

Participate

Take part in engaging research, earn cash, and help improve human knowledge. Learn more

Crowdsourcing Platform: Prolific



Using our demographic filters to prescreen participants

Prolific Team
12 September 2018 17:18

On Prolific, you have the option to enter prescreening criteria to filter the participants you need.

Here are a few examples of the several hundred demographic criteria you can currently filter by:

- Sex
- Age
- Nationality
- Country of Birth
- Current country of residence
- Within the United States: Current US state of residence and US state of birth
- First Language
- Ethnicity
- Employment status & domain
- Student status & field of study
- Political affiliation
- Religious affiliation
- Sexual orientation
- Handedness
- Marital status
- Socioeconomic status

DYSEN: Sentiment Annotation of Media Text Snippets

Sentiments: positiv, negativ, neutral

Examples of media text snippets

Der Sachverhalt der Disziplinaranzeige stütze sich lediglich auf die Erhebungen der Sonderkommission. In der Anzeige heißt es, dass in dem Verhalten Kreißls ein absoluter Vertrauensbruch gesehen werde. Deshalb würde die schwerste Disziplinarstrafe, die Entlassung aus dem Staatsdienst, angeregt.

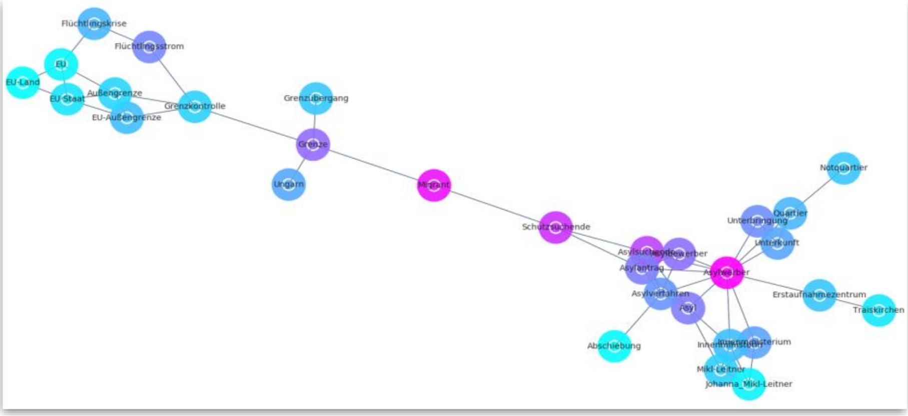
Riegler erklärte in Wien zum Antreten zweier Kandidaten, dies sei keine Zerreißprobe für die Partei, sondern könne ihr " viel Motivation bringen". Die Parteitagsdelegierten würden mit Busek " einen der fähigsten Politiker, den die VP aufzubieten hat", vorfinden. Mit Görg stehe " ein Mensch mit vielen Fähigkeiten" zur Verfügung.

WWW.OEAW.AC.AT

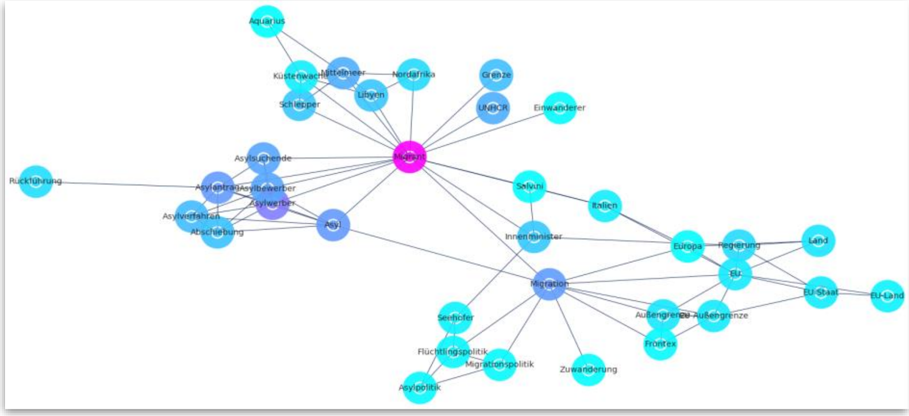
Diachronic Dynamics of Lexical Networks DYLEN

DYLEN: Sentiment Annotation of Austriacisms

- Diachronic Dynamics of Lexical Networks
- Detecting contextual shifts with networks: emergence/conflation of contexts/senses



‘Migrant’ in 2016



‘Migrant’ in 2018

Baumann et al. 2020

ACDH-CH – AUSTRIAN CENTRE FOR DIGITAL HUMANITIES AND CULTURAL HERITAGE

DYLEN: Sentiment Annotation of Austriacisms

Examples of Austriazisms

fesch A D Adj. [fe:S A, fES A D] <kurz für engl. fashionable ‚modisch'>: →schnieke D-nord/mittelwest (bes. Berlin) ‚hübsch, gut aussehend'

Exekution A die; –, -en <aus lat. ex(s)ecutio ‚Vollstreckung'> (Recht): →Betreibung CH, →Beitreibung D ‚Vollstreckung von [finanziellen] Ansprüchen, z. B. Pfändung, Zwangsräumung etc.'

(Variantenwörterbuch des Deutschen 2016)

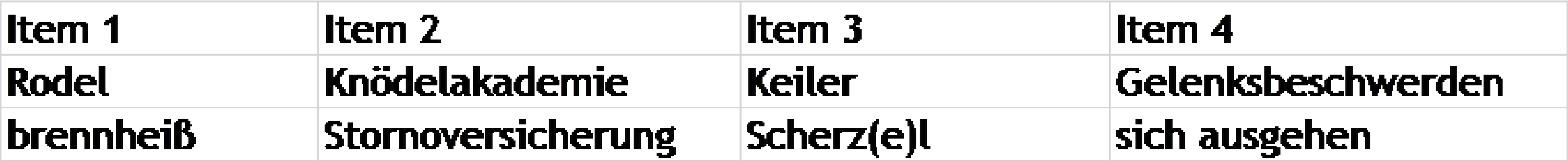
DYLEN: Sentiment Annotation of Austriacisms

- Multi-stage approach (cf. Rouces et al. 2018) with two rounds of crowdsourced human annotations

1. Direct score annotation (negative, neutral, positive)

2. Best-Worth Scaling (4-tuples, most negative, most positive)

Example of 4-tuples



Item 1	Item 2	Item 3	Item 4
Rodel	Knödelakademie	Keiler	Gelenksbeschwerden
brennheiß	Stornoversicherung	Scherz(e)l	sich ausgehen

Thank you for listening!

Project Websites:

ELEXIS https://elex.is/

DYLEN https://dylen.acdh.oeaw.ac.at/

DYSEN https://dylen.acdh.oeaw.ac.at/dysen/

Keep in touch: tanja.wissik@oeaw.ac.at

The research was partly supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015, by the go!digital Next Generation grant (GDNG 2018-020) and the Stadt Wien within the “Digital Humanism” call.