An Overview of NLP Crowdsourcing Systems

Sixth ENeL Action meeting in Budapest
25 Feb 2017

Federico Sangati

Terminology

Crowdsourcing Outsourcing

Citizen Science

Wisdom of the crowd

Human Computation

Amazon Mechanical Turk

Collective Intelligence

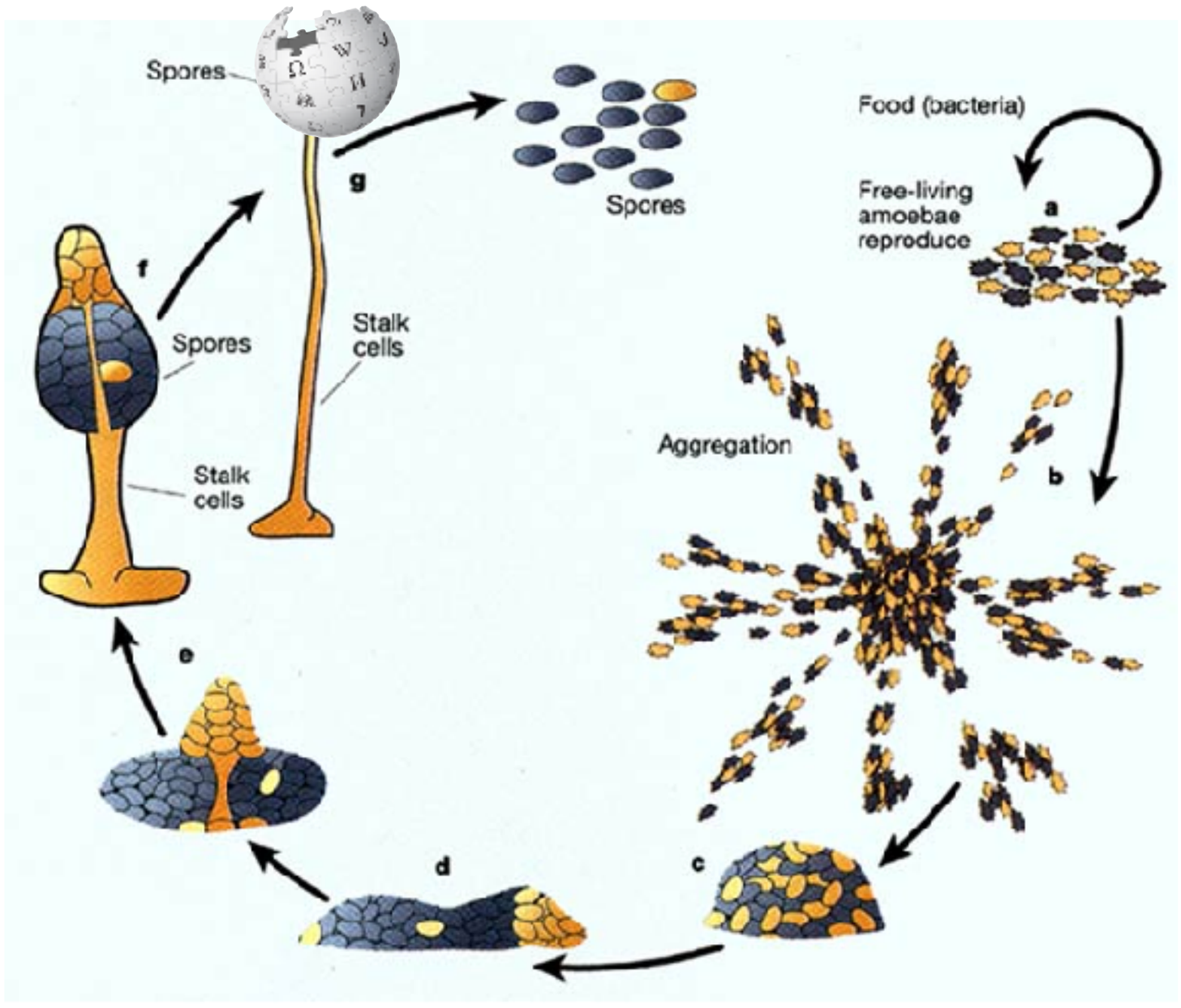
Game With a Purpose (GWAP)

Serious Games

Collaboratively Constructed Language Resources (CCLR)

Higher level of organization

Slime Mold



WIKIPEDIA (2001)

English
The Free Encyclopedia
3 907 000+ articles

日本語
フリー百科事典
799 000+ 記事

Español
La enciclopedia libre
879 000+ artículos

Deutsch
Die freie Enzyklopädie
1 383 000+ Artikel

Русский
Свободная энциклопедия
838 000+ статей

Français
L'encyclopédie libre
1 230 000+ articles

Italiano
L'enciclopedia libera
905 000+ voci

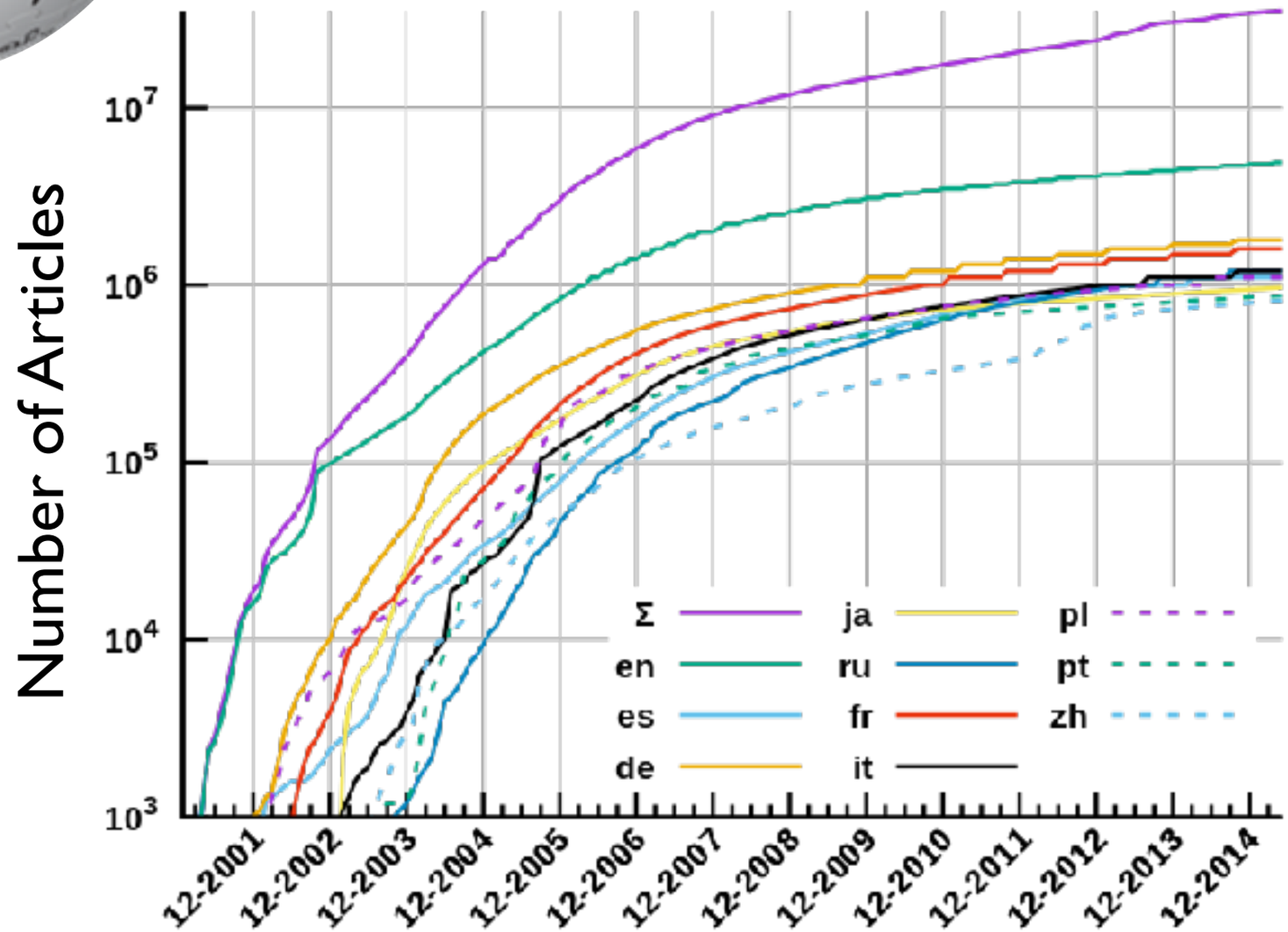
Polski
Wolna encyklopedia
887 000+ haseł

Português
A enciclopédia livre
718 000+ artigos

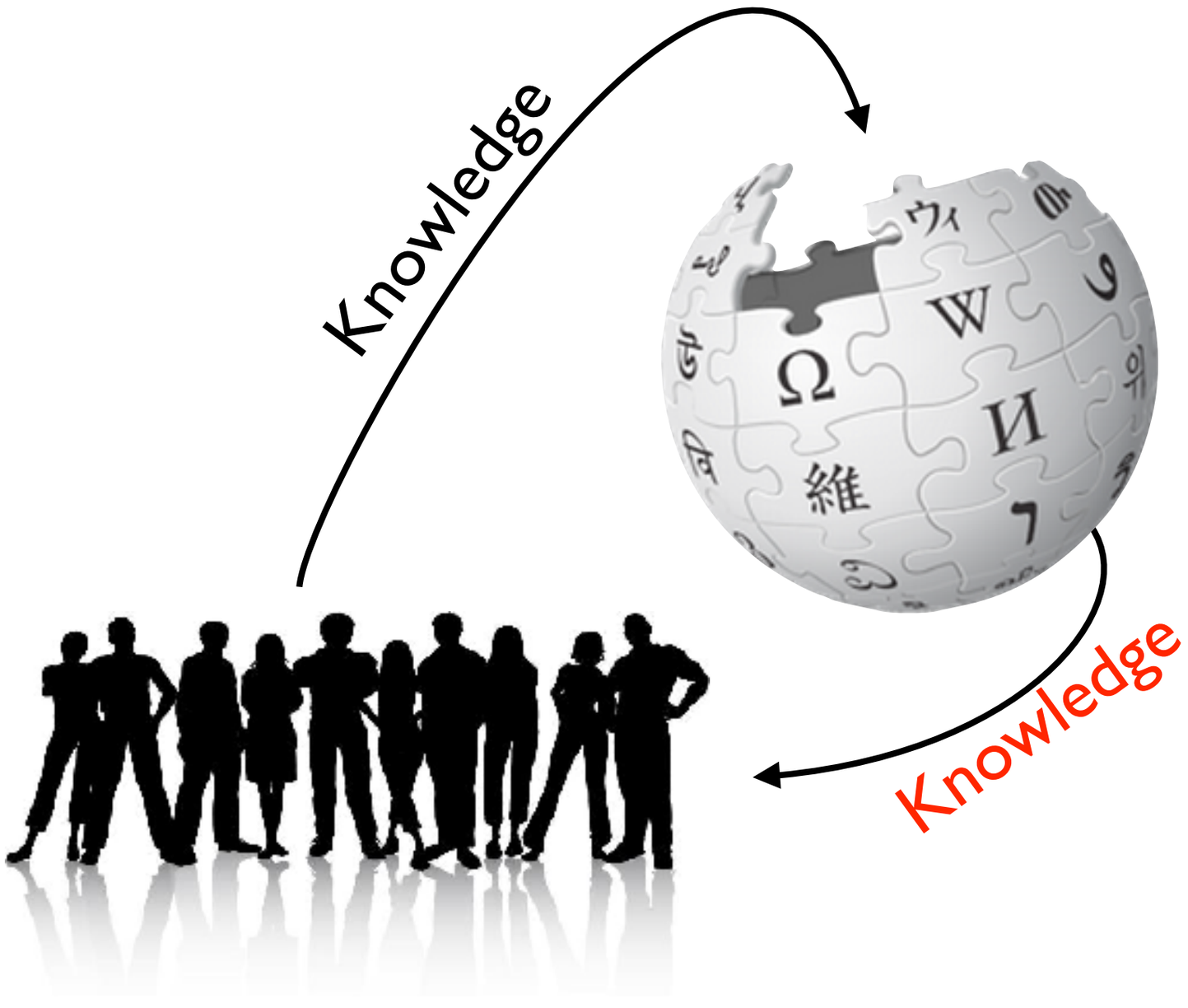
中文
自由的百科全書
429 000+ 條目

WIKIPEDIA

285 different languages



WIKIPEDIA



Amazon Mechanical Turk (2005)

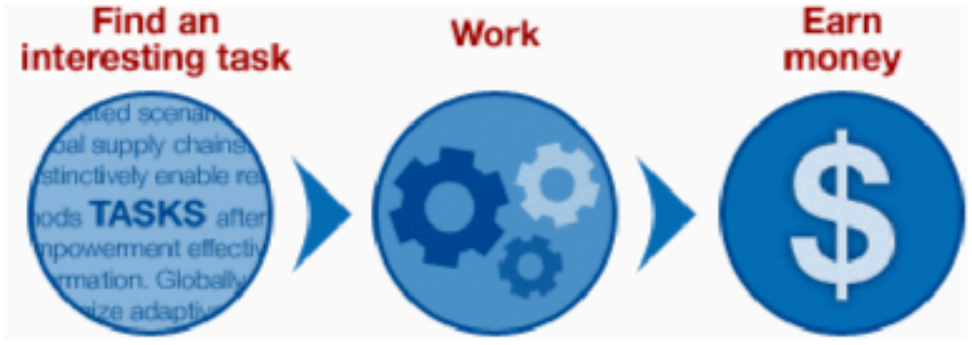
amazon mechanical turk™
Artificial Artificial Intelligence

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



Find HITs Now

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get Started.

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Get Started

Amazon Mechanical Turk (2005)

Task development

Knowledge

amazon mechanical turk™
Artificial Artificial Intelligence

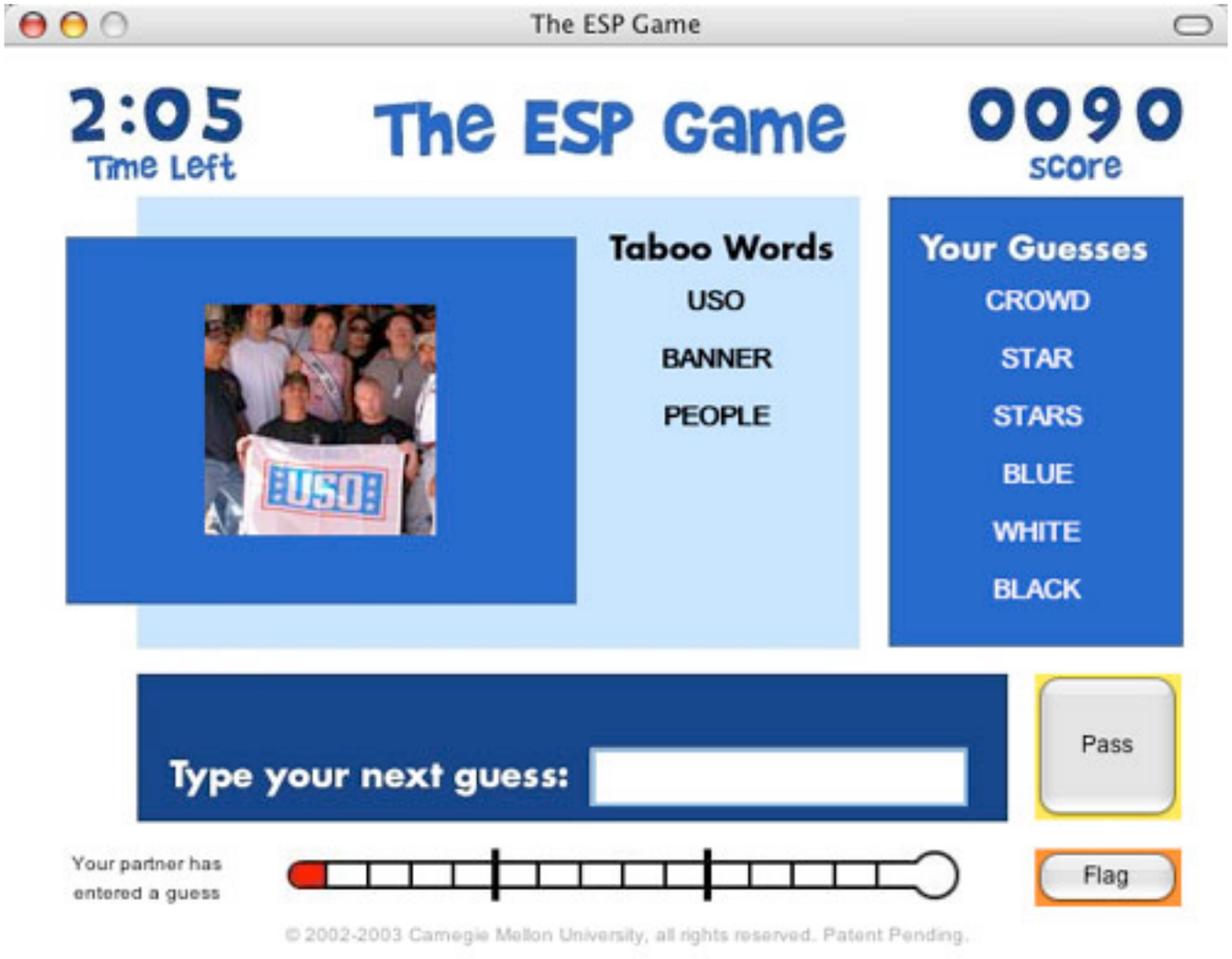
Data

$

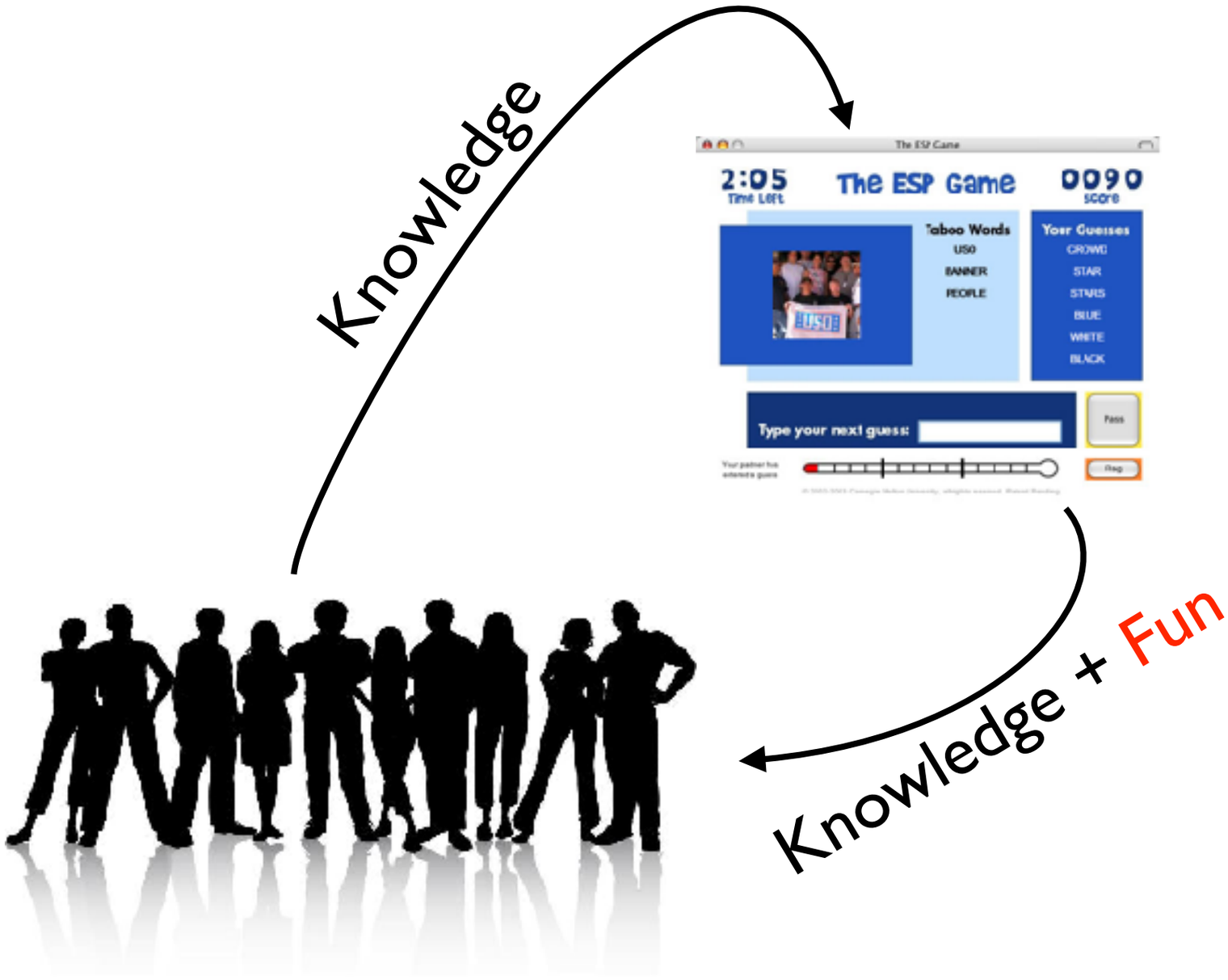
amazon mechanical turk™
Artificial Artificial Intelligence

ESP Game

Luis von Ahn (2004)



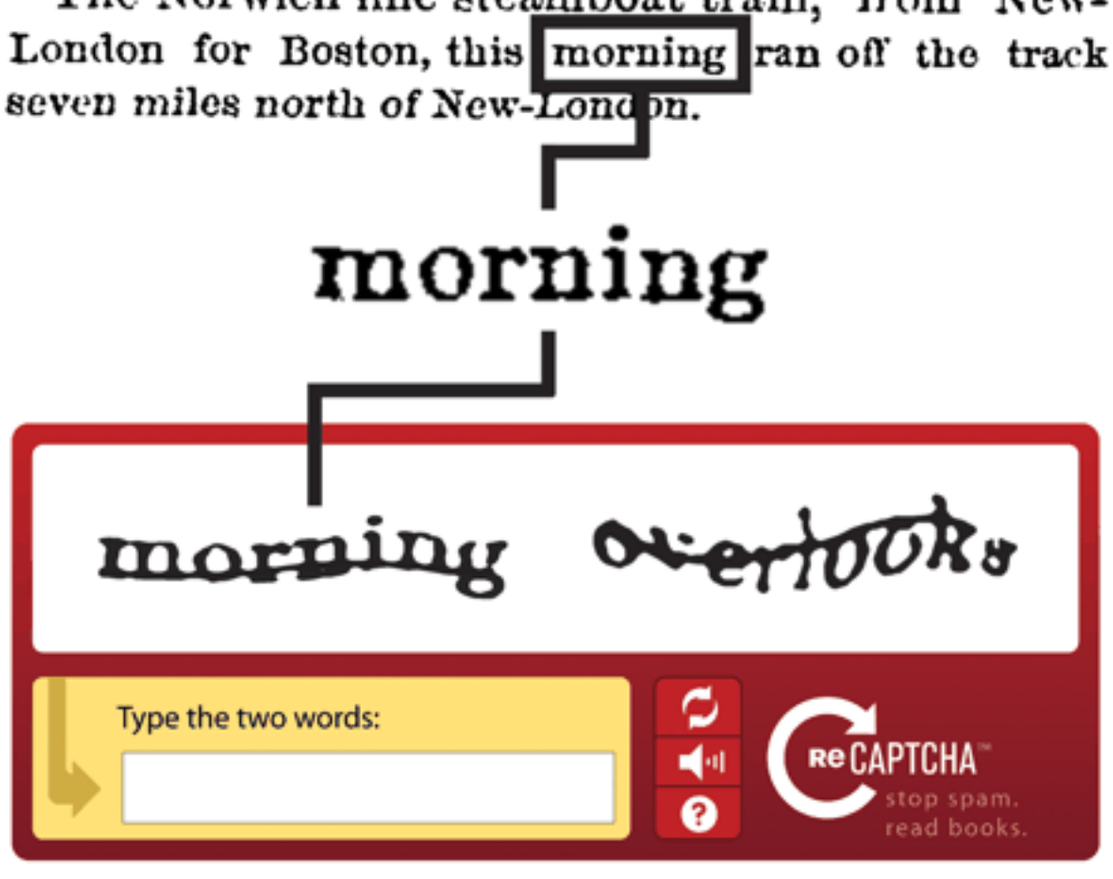
GWAP


Knowledge
The ESP Game
2:05
Time Left
The ESP Game
0090
Score
Taboo Words
USO
BANNER
PEOPLE
Your Guesses
CROWD
STAR
STARS
BLUE
WHITE
BLACK
Type your next guess:
Pass
Flag
Knowledge + Fun

reCAPTCHA

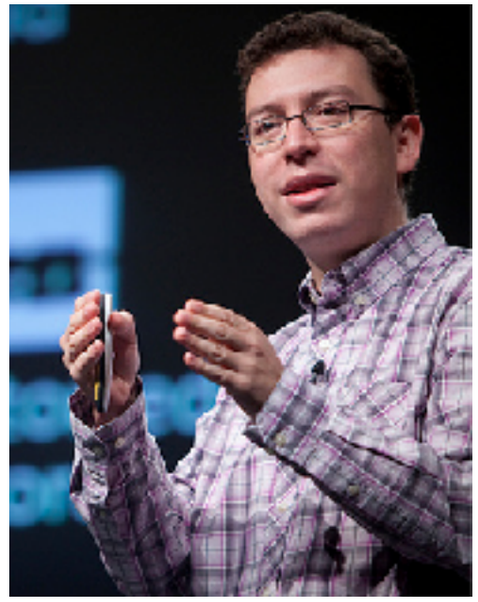
Luis von Ahn (2008)

The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.



Duolingo

Luis von Ahn (2012)



Edutainment

ZOONIVERSE

REAL SCIENCE ONLINE

CITIZEN SCIENCE ALLIANCE

(2007)

Nature

Arts

Biology

Climate

History

Literature

Language

Social Science

Space

Physics

Medicine

ZOONIVERSE

REAL SCIENCE ONLINE

CITIZEN SCIENCE ALLIANCE

Literature

SHAKESPEARE'S WORLD

MEASURING THE ANZACS

SCIENCE GOSSIP

OPERATION WAR DIARY

Why crowdsourcing in NLP

- Offset the high costs of language resource development and maintenance
- Seeking expertise outside the members of the project
- Create a public interest on linguistic research and synergies outside the academic environment (e.g., schools, elderly care taking infrastructures)

Main obstacles

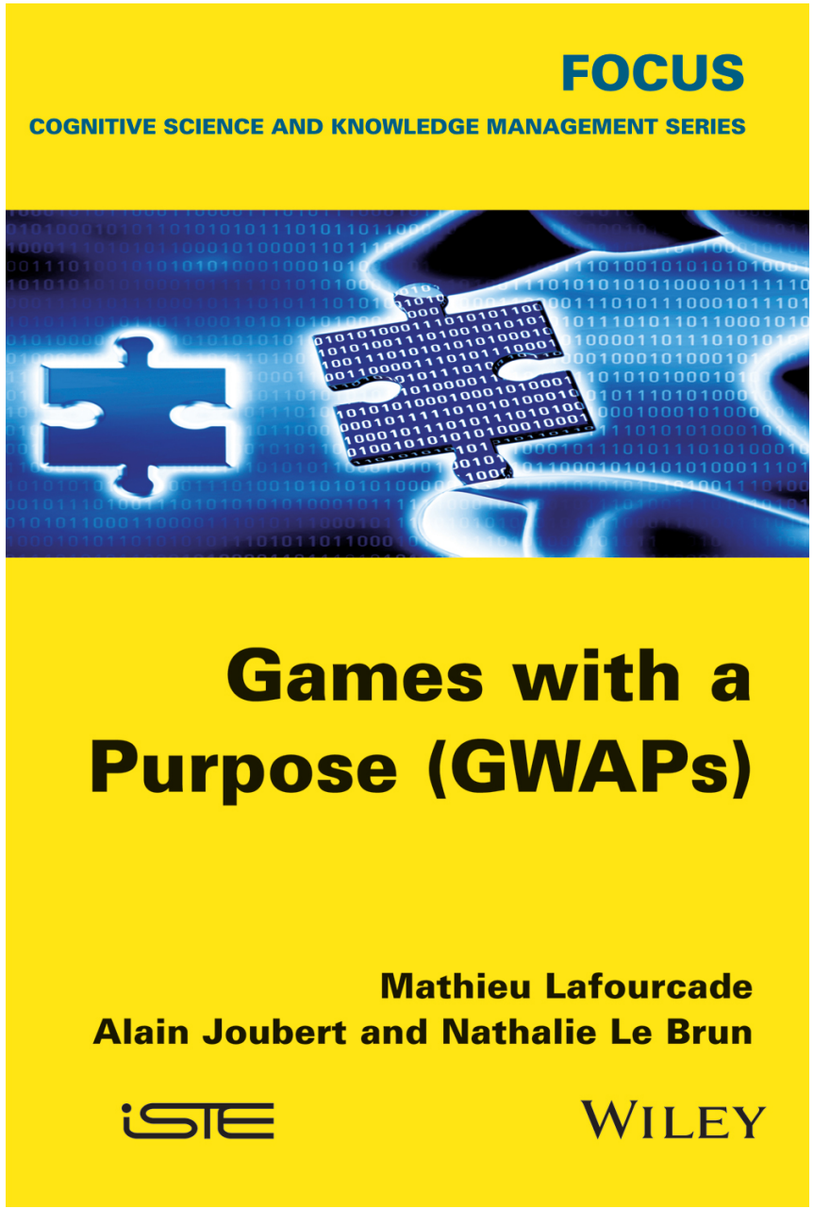
- Implementation: hard to program a successful system (paradigm, UX, robustness, scalability)
- Visibility: need to reach a critical mass of users in order for the project to succeed
- Dropouts: many people try the system just once and the abandon the project

Ingredients for success

- Implementation: Start simple and focus on game mechanics. Prototype the idea and test it with a small set of users before investing on interface and the rest.
- Visibility: enhance visibility of the project (social media) in order to attract new users.
- Dropouts: keep the community motivated and engaged.

GWAP Survey

- Mid January 2017: opened a survey on Corpora List of NLP-related Crowdsourcing Systems
- Selected answers at **tiny.cc/nlpcrowd**
- Survey is still open at tiny.cc/nlpcrowd_form



CHAPTER 3. GWAPS FOR NATURAL LANGUAGE PROCESSING 47

3.1. Why lexical resources? 47
3.2. GWAPs for natural language processing 48
3.2.1. The problem of lexical resource acquisition 49
3.2.2. Lexical resources currently available 50
3.2.3. Benefits of GWAPs in NLP 53
3.3. PhraseDetectives 54
3.4. PlayCoref 57
3.5. Verbosity 59
3.6. JeuxDeMots 61
3.7. Zombilingo 62
3.8. Infection 64
3.9. Wordrobe 66
3.10. Other GWAPs dedicated to NLP 68
3.10.1. Open Mind Word Expert 68
3.10.2. 1001 Paraphrases 69
3.10.3. Categorilla/Categodzilla 69
3.10.4. FreeAssociation 70
3.10.5. Entity Discovery 70
3.10.6. PhraTris 70

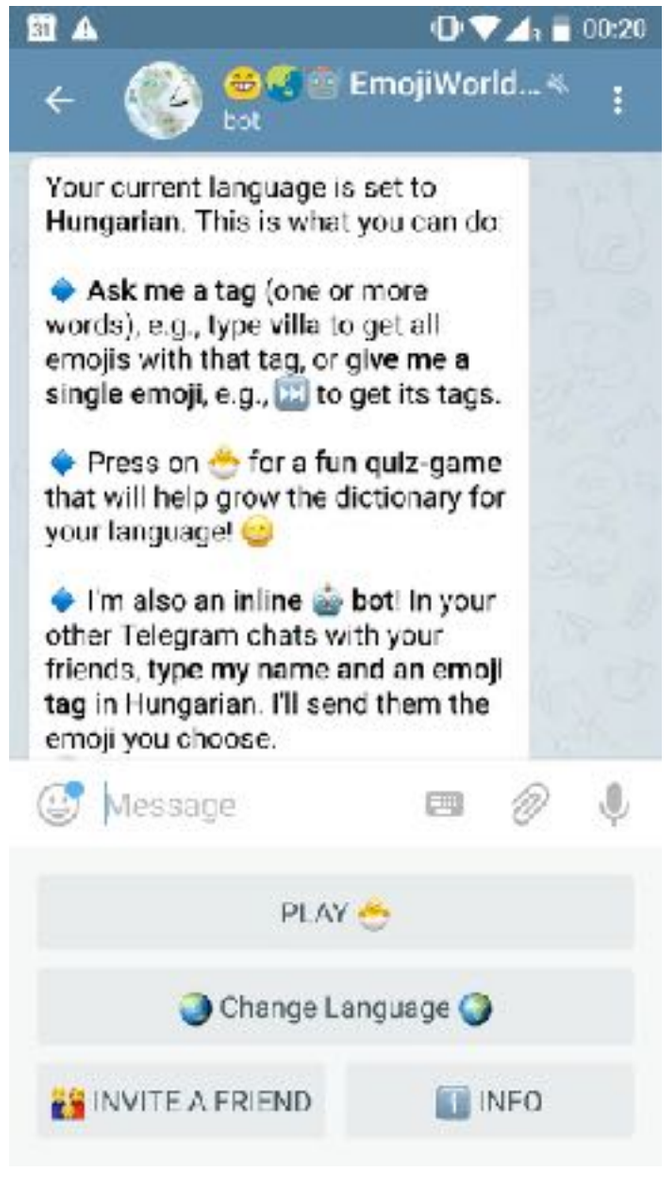
📖 M. Lafourcade, A. Joubert, and N. Brun. Games with a Purpose (GWAPS). Focus Series in Cognitive Science and Knowledge Management. Wiley, 2015.

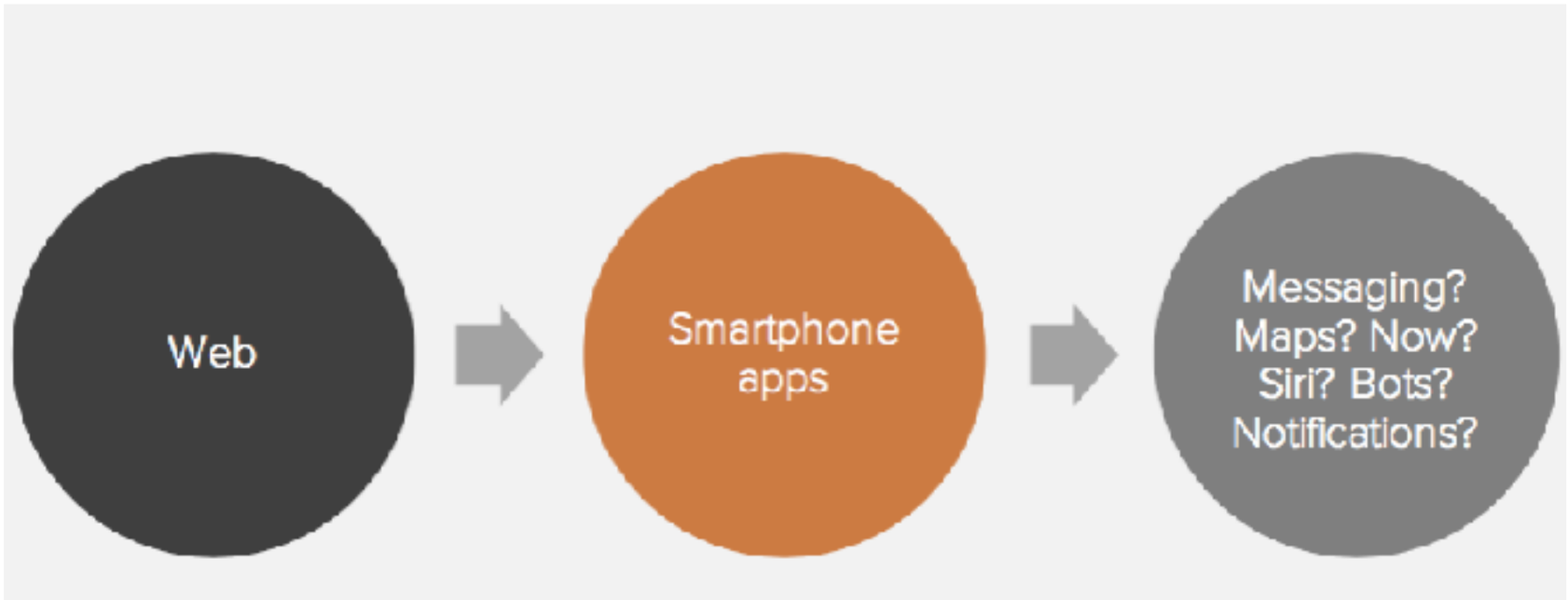
Name	Active	Topic	Launched
Open mind word expert	❌	Word Sense Tagging	2002
1001 Paraphrases	❌	Paraphrases	2005
Verbosity	❌	Word Common Sense Knowledge	2005
Jeuxdemots	✅	Lexico-Semantic Network	2007
Free Association / Categorilla / Categodzilla	❌	Word Associations	2008
OntoGames	❌	Word Ontologies	2008
Phrase Detective	✅	Anaphora Resolution	2008
Sentiment Quiz	❌	Sentence Sentiment Polarity	2009
PlayCoref	✅	Anaphora Resolution	2009
PhraTris	❌	Annotation of Syntactic Relations	2010
DuoLingo	✅	Foreign Language Learning	2012
Wordrobe	✅	Tagging (Part of Speech, Named Entity)	2012
Xtribe	✅	Writing Stories Collaboratively	2013
SmallWordlOfWords	✅	Collections of Words Associations	2013
ZombiLingo	✅	Annotation of Syntactic Relations	2014
Puzzle Racer Ka-boom!	❌	Concept to Picture Association	2014
Infection The Knowledge Towers	❌	Word Similarity, Antonymy, and Relations	2014
Clozemaster	✅	Language Learning	2014 (?)
Zoouniverse	✅	Literature Digitization and Tagging	2015 (?)
Bisame	✅	Part of Speech Tagging	2015 (?)
EmojiWorldBot	✅	Word Emoji Multilingual Dictionary	2016
Ingra-besed	✅	Word Collocations	2016

😂🌍🤖 EmojiWorldBot

Martin Benjamin, École polytechnique fédérale Lausanne, Switzerland
Francesca Chiusaroli, Macerata University, Italy
Johanna Monti, Napoli University, Italy

- Emoji ↔ Text in 130 different languages (and growing)
- Implemented as a chat-bot in the Telegram messaging platform

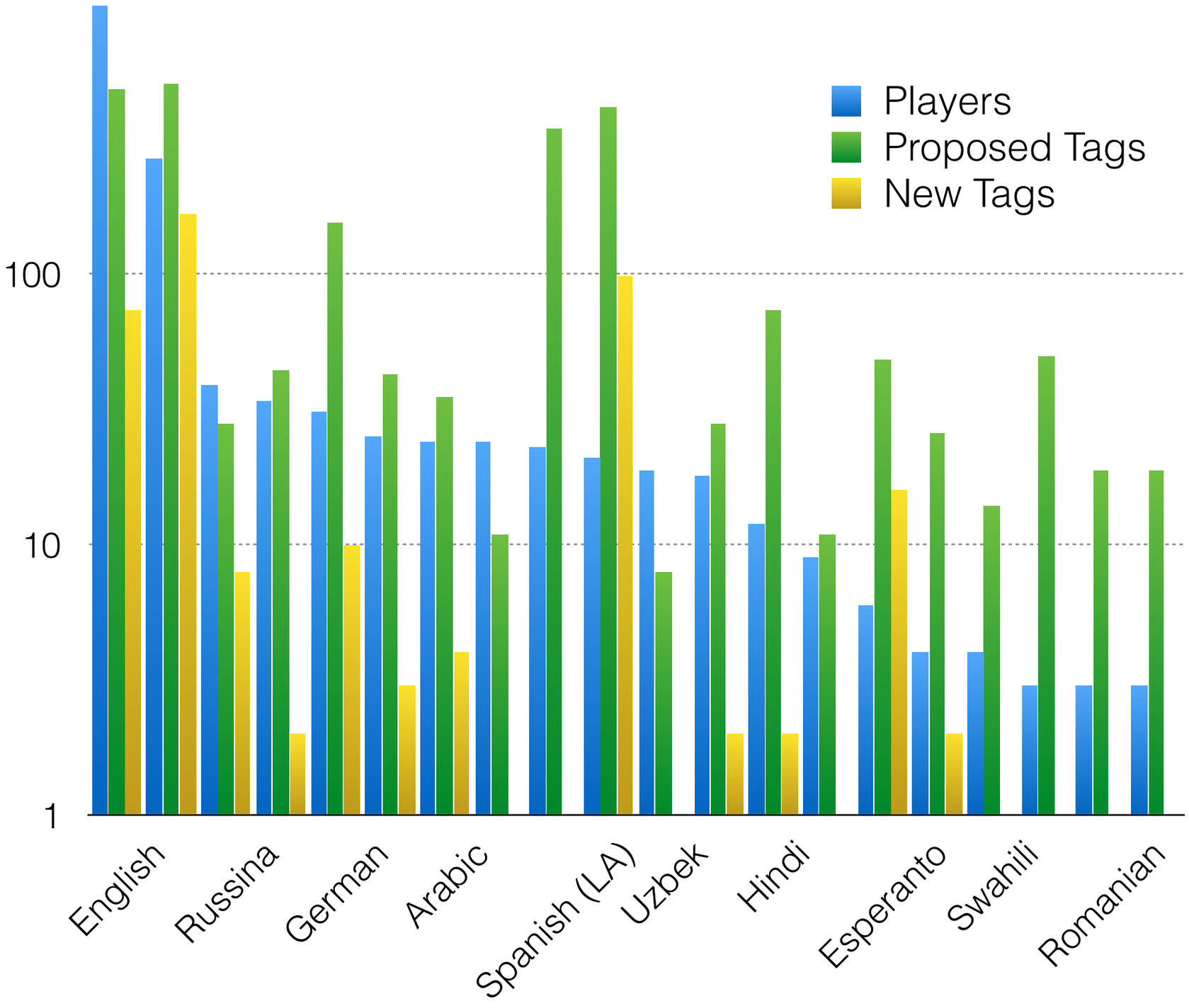



Web
Smartphone apps
Messaging? Maps? Now? Siri? Bots? Notifications?


www
HTML5
Js
Cloud
XML
CSS3
flat
PHP
header
.com
footer
.net
.org

Collected Annotations (since Sept. 2016)

- 61 languages with at least one annotation
- ~1700 players, ~2500 proposed tags, ~500 new tags in total



Final Remarks

- Get inspired by non-NLP crowdsourcing systems.
- Create single platform for NLP based crowdsourcing projects (boost visibility, code sharing)?
- Seek synergies with other types of institutions (e.g., school, elderly care taking infrastructures).
- New platforms chat-bots platforms.

DH
DIGITAL HUMANITIES

FONDAZIONE BRUNO KESSLER

SCHOOL-TAGGING

Classroom Exercises for Language Research

Web platform for **collaborative language exercises** in classrooms with the help of the **teacher**.

School Classrooms

Computational Linguistics Research

Exercises result will be used to **create linguistic resources**.

http://dh.fbk.eu
https://twitter.com/DH_FBK

DH
DIGITAL HUMANITIES

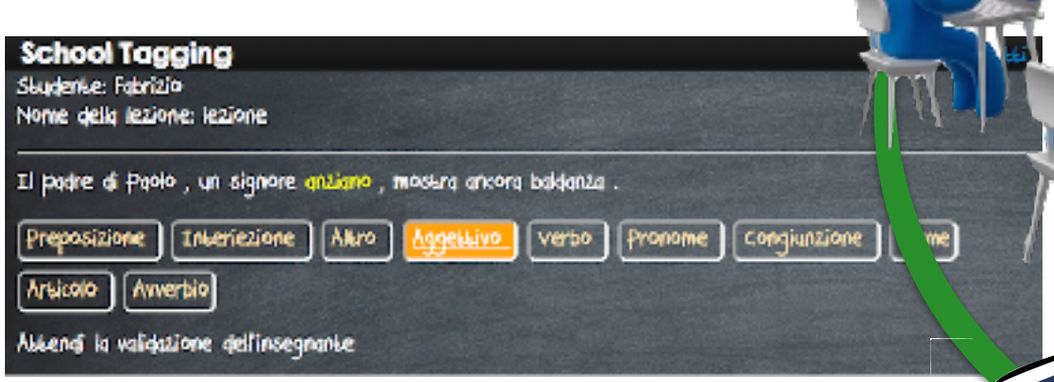
FONDAZIONE BRUNO KESSLER

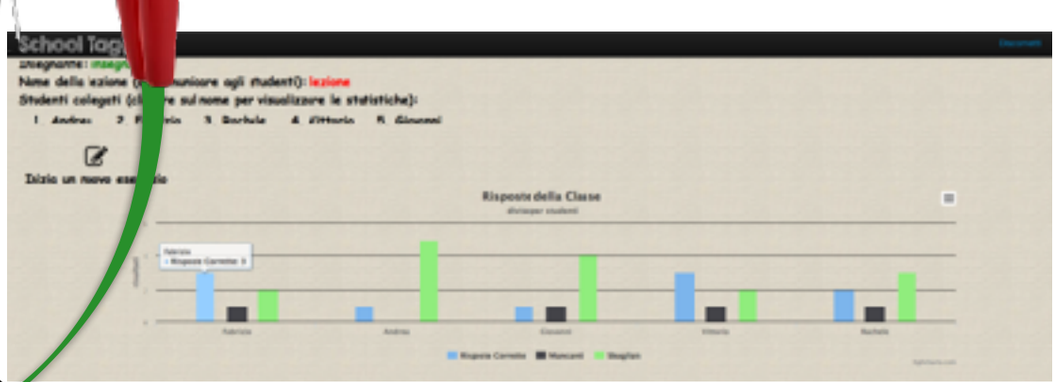
SCHOOL-TAGGING

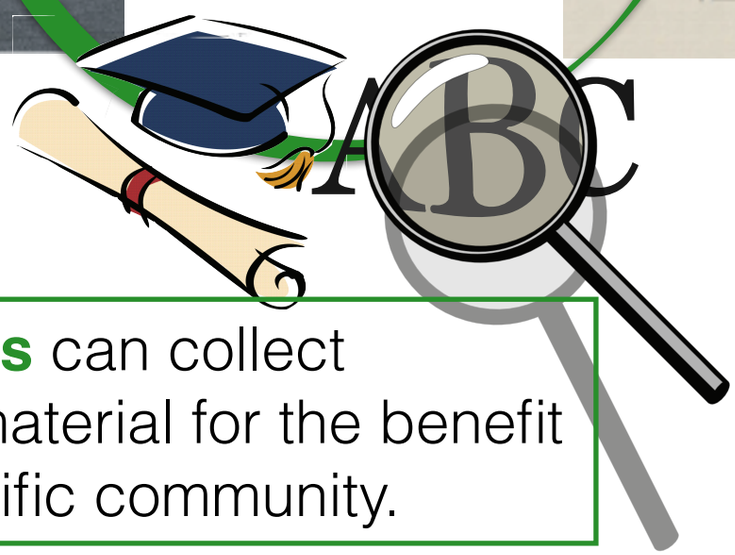
Classroom Exercises for Language Research

Students engage in game-like exercises with immediate feedback.

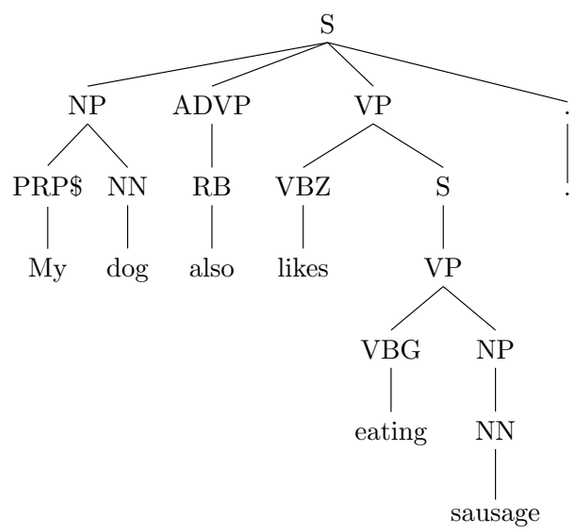
Teachers can monitor individual and aggregated answers in real time and validate results.







Researchers can collect annotated material for the benefit of the scientific community.



http://dh.fbk.eu
https://twitter.com/DH_FBK