Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax

Bruno Guillaume, **Karën Fort** and Nicolas Lefèbvre

bruno.guillaume@inria.fr, karen.fort@paris-sorbonne.fr, nicolas.lefebvre@inria.fr

December 16th, 2016

1. Crowdsourcing: back to basics
2. Overview of the game
3. Behind the curtain
4. Results
5. Conclusion and future plans

1 Crowdsourcing: back to basics

2 Overview of the game

3 Behind the curtain

4 Results

5 Conclusion and future plans

Crowdsourcing

Crowdsourcing *is "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call." [Howe, 2006]*

- no *a priori* identification or selection of the participants ("open call")

Crowdsourcing

Crowdsourcing *is "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call." [Howe, 2006]*

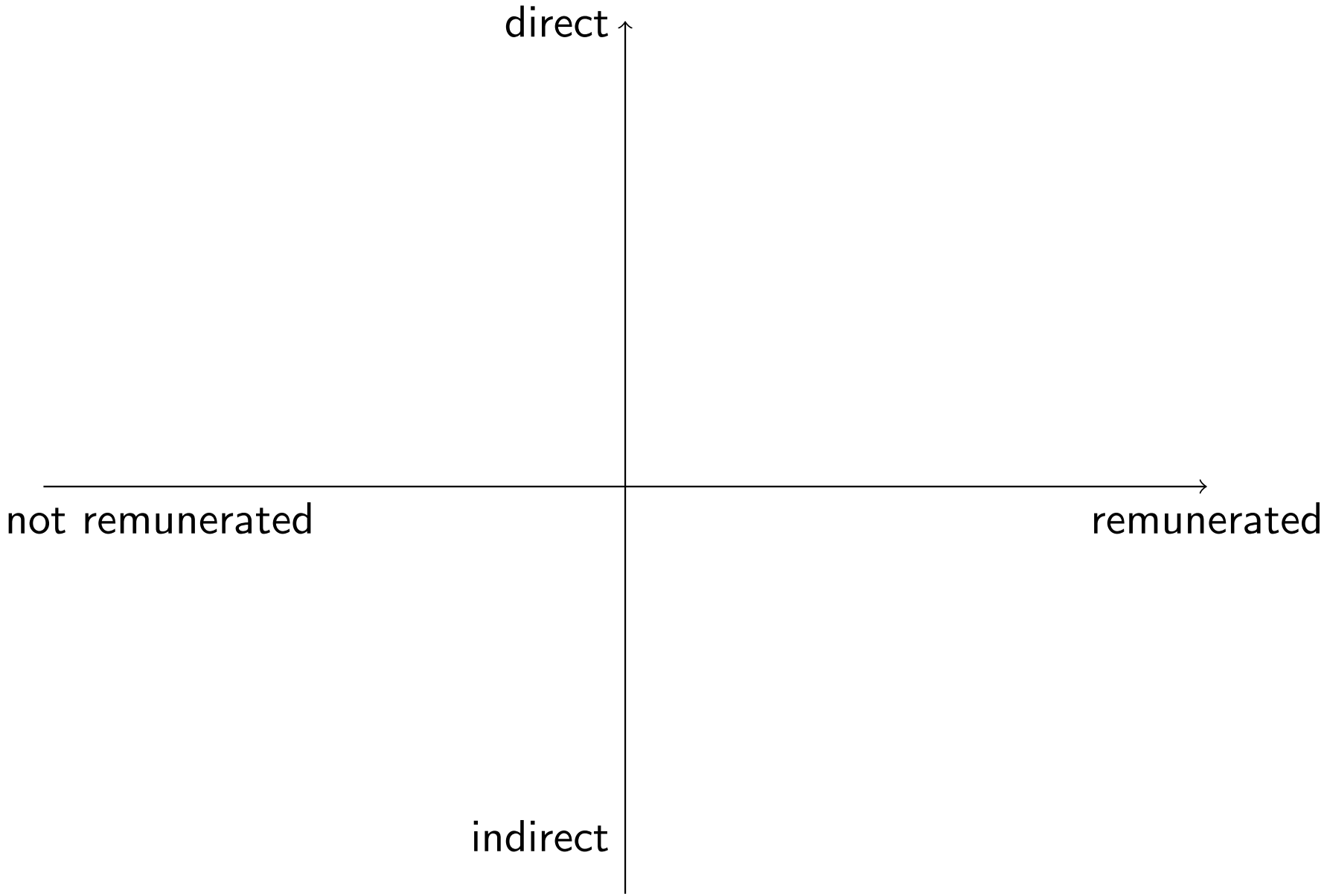
- no *a priori* identification or selection of the participants ("open call")
- massive (in production and participation)

Crowdsourcing

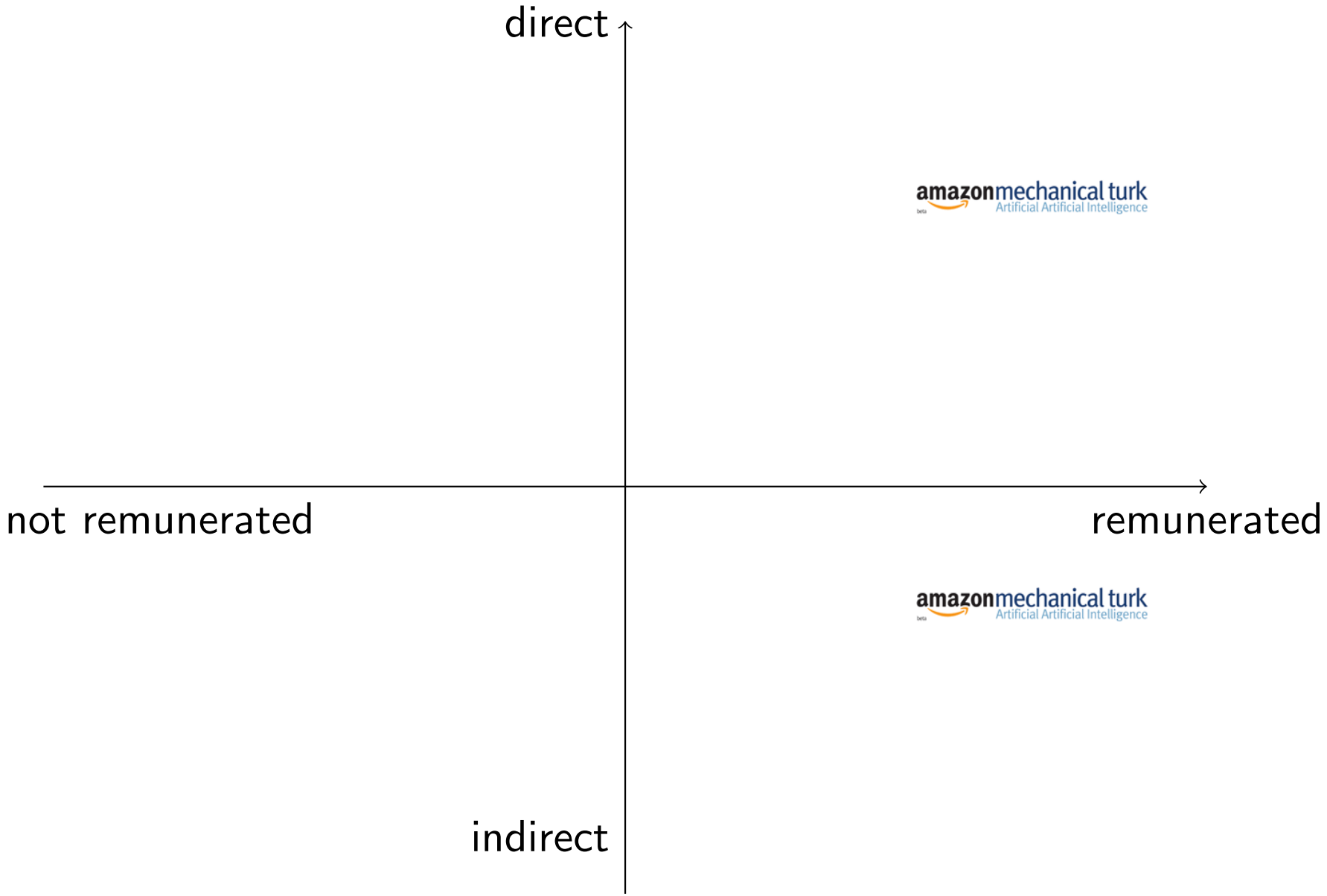
Crowdsourcing *is "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call." [Howe, 2006]*

- no *a priori* identification or selection of the participants ("open call")
- massive (in production and participation)
- (relatively) cheap

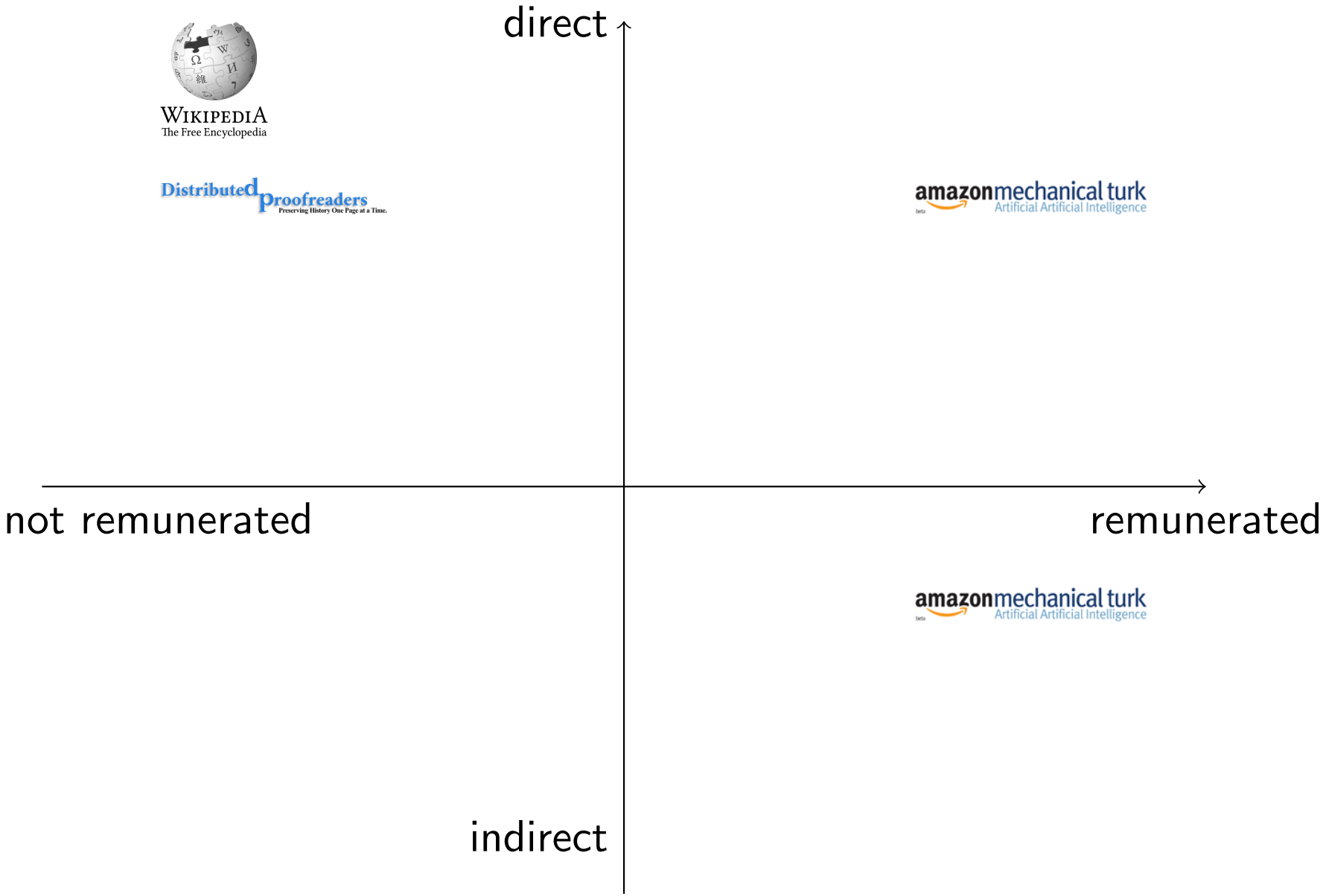
A simplified taxonomy (more in [Geiger et al., 2011])



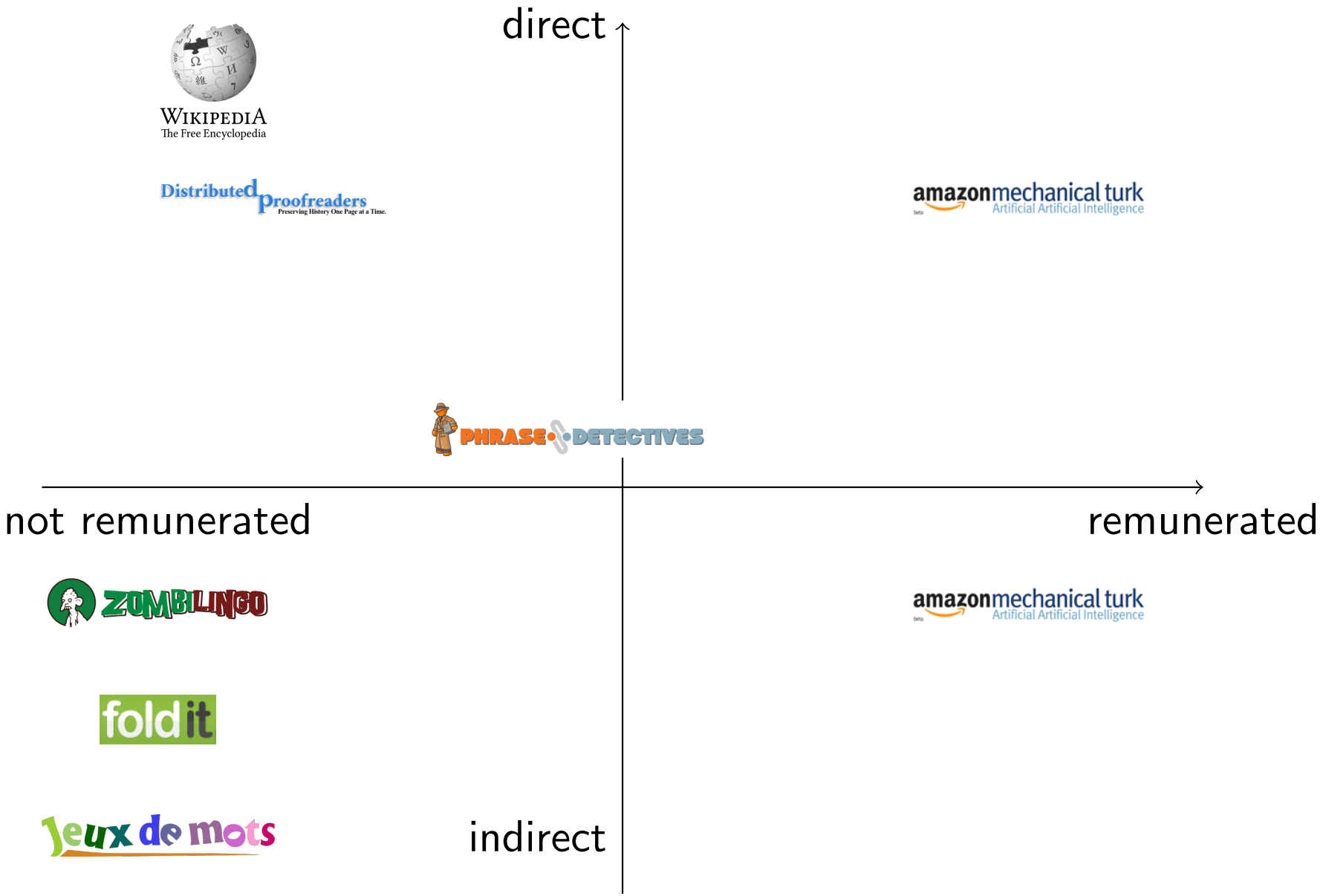
A simplified taxonomy (more in [Geiger et al., 2011])



A simplified taxonomy (more in [Geiger et al., 2011])



A simplified taxonomy (more in [Geiger et al., 2011])

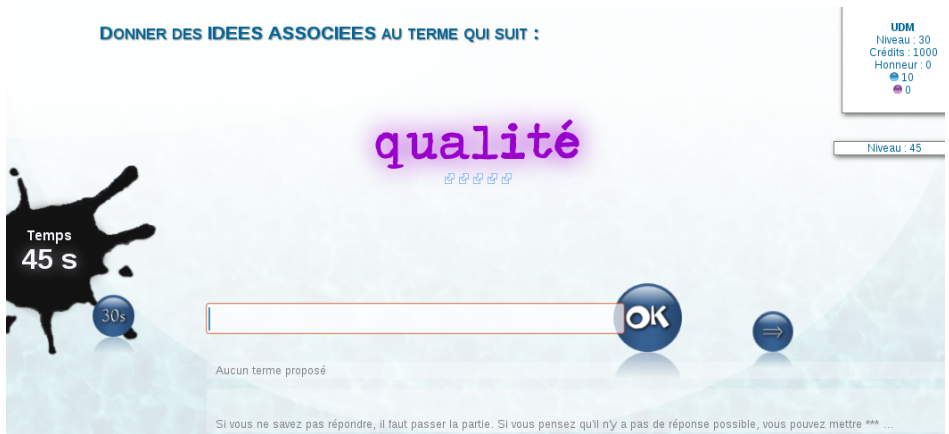


JeuxDeMots: playing association of ideas...

...to create a lexical network [Lafourcade and Joubert, 2008]

More than 60 million relations (created by 1,161 players), that are constantly updated

- play by pairs
- more and more complex, typed relations
- challenges
- lawsuits
- etc.



Phrase Detectives: playing detective...

...to annotate co-reference [Chamberlain et al., 2008]

3.5M decisions from 45k players

- pre-annotated corpus
- detailed instructions
- training
- 2 different playing modes
  - annotation
  - validation (correction of annotations)

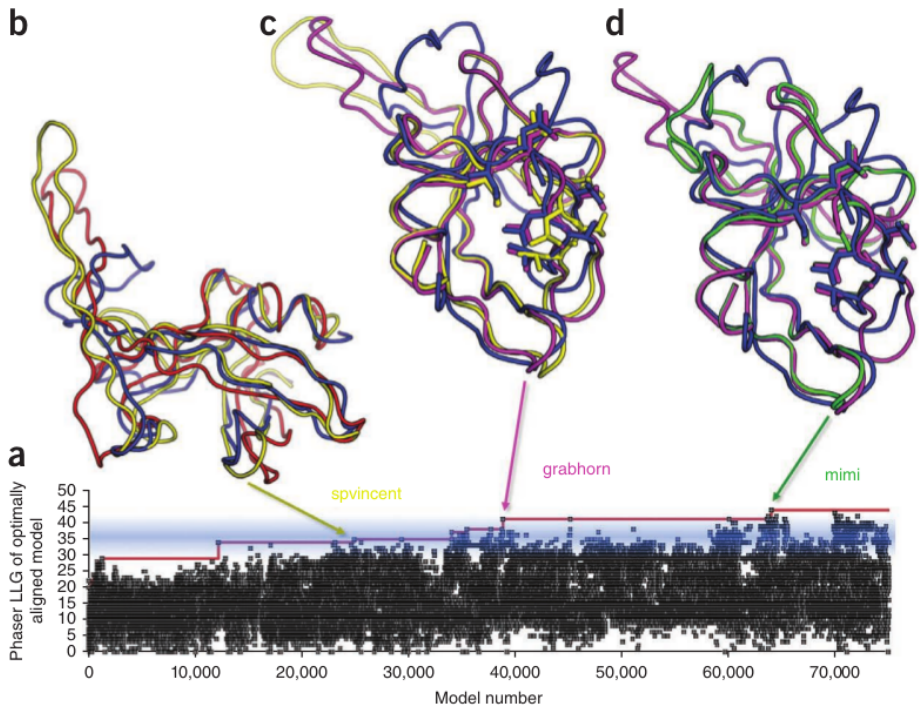
FoldIt: playing proteins folding...

...to solve scientific issues [Khatib et al., 2011]

Solution to the crystal structure of a monomeric retroviral protease (simian AIDS-causing monkey virus)

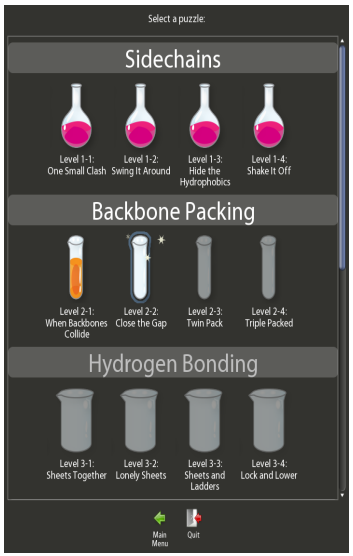
Solution to an issue unsolved for over a decade

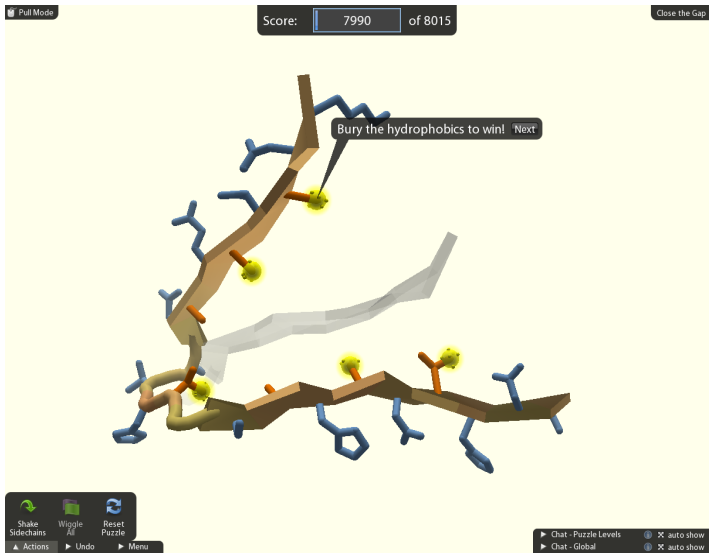
- found in a couple of weeks
- by a team of players
- that will allow for the creation of antiretroviral drugs



FoldIt: playing proteins folding...

...without any prior knowledge in biochemistry [Cooper et al., 2010]





Step-by-step training

- tutorial decomposed by concepts
- puzzles for each concept
- access to the following puzzles is given only if your level is sufficient

1 Crowdsourcing: back to basics

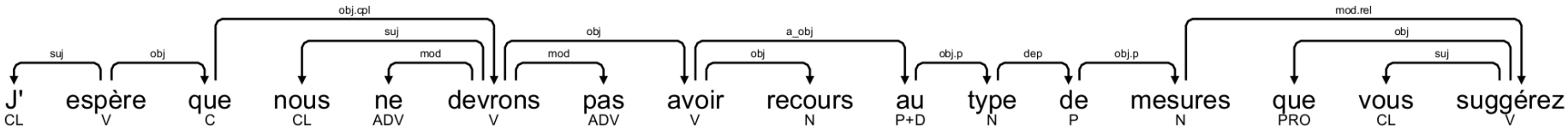
2 Overview of the game

3 Behind the curtain

4 Results

5 Conclusion and future plans

A complex annotation task



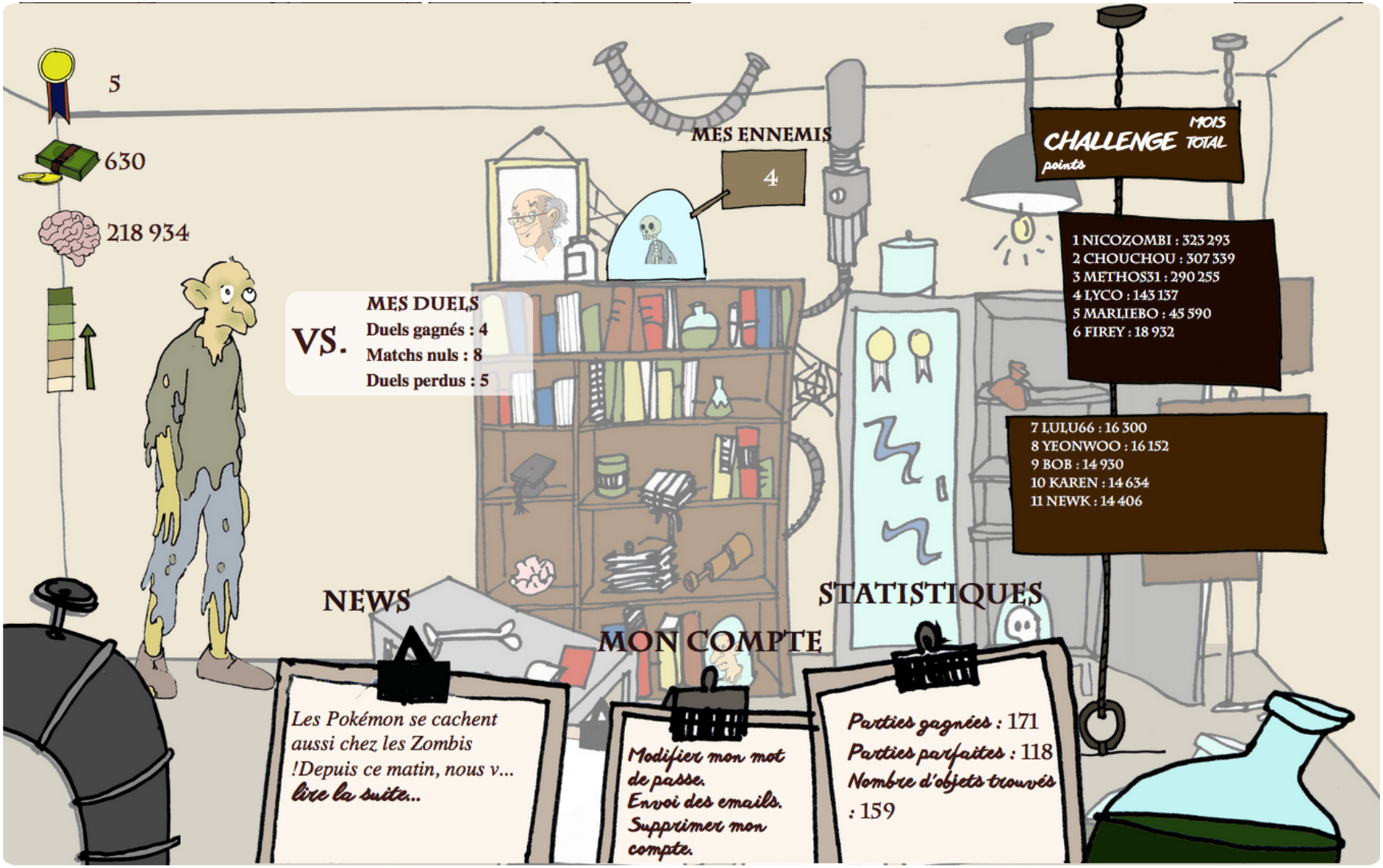
- annotation guidelines
  - 29 relation types
  - approx. 50 pages
- counter-intuitive decisions: `aobj` = *au*

  [...] avoir recours au type de mesures [...]

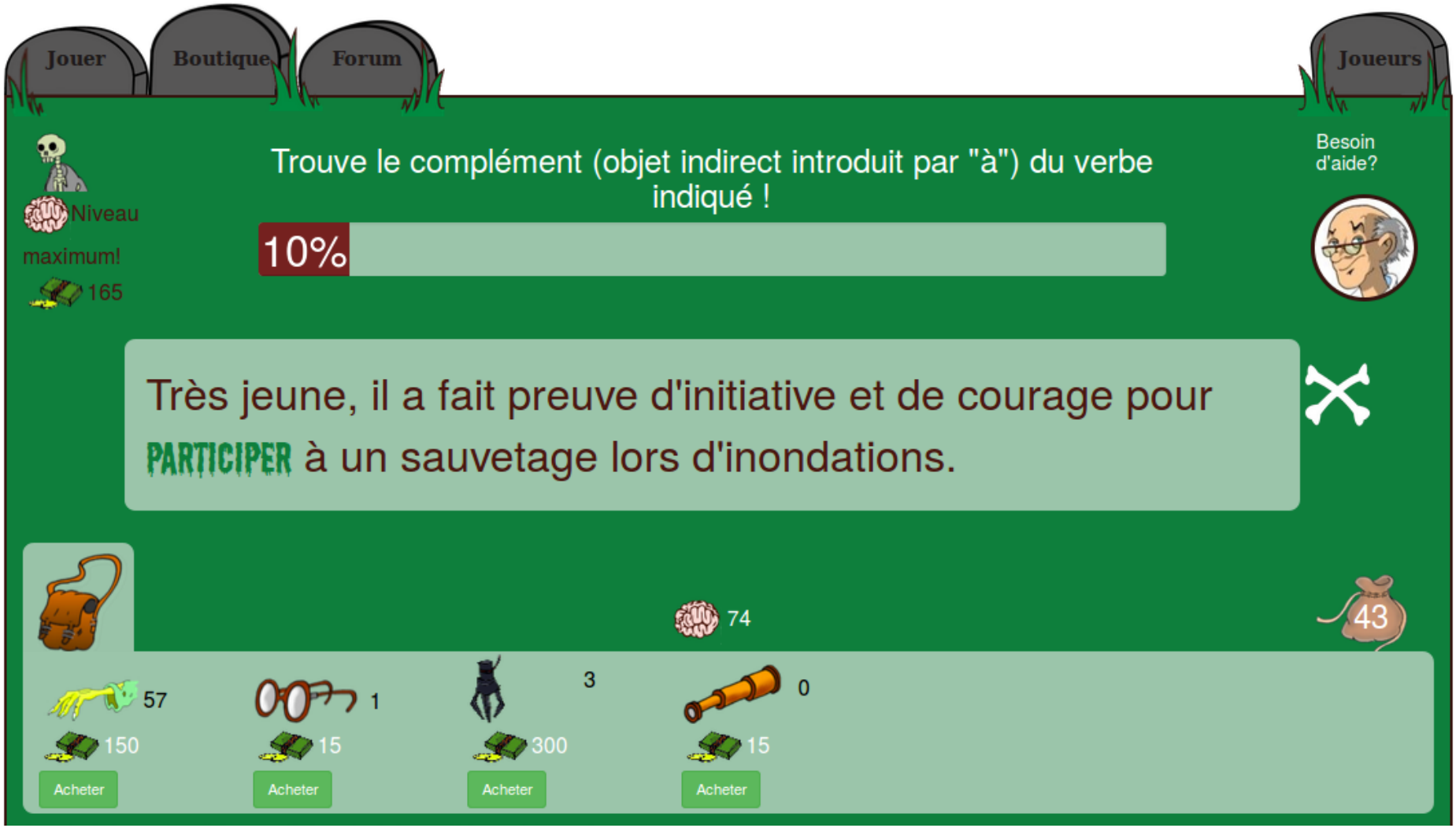
  i.e. head of the PP is the preposition

→ decompose the complexity of the task [Fort et al., 2012], not simplify it!

http://zombilingo.org/


5
630
218 934
MES ENNEMIS
4
CHALLENGE
MOIS
TOTAL
points
1 NICOZOMBI : 323 293
2 CHOUCHOU : 307 339
3 METHOS31 : 290 255
4 LYCO : 143 137
5 MARLIEBO : 45 590
6 FIREY : 18 932
7 LULU66 : 16 300
8 YEONWOO : 16 152
9 BOB : 14 930
10 KAREN : 14 634
11 NEWK : 14 406
VS.
MES DUELS
Duels gagnés : 4
Matchs nuls : 8
Duels perdus : 5
NEWS
MON COMPTE
STATISTIQUES
Les Pokémon se cachent aussi chez les Zombis !Depuis ce matin, nous v... lire la suite...
Modifier mon mot de passe.
Envoi des emails.
Supprimer mon compte.
Parties gagnées : 171
Parties parfaites : 118
Nombre d'objets trouvés : 159

AUXILIAIRE CAUSATIF
Points : 270
Annotations : 6/6
AUXILIAIRE DE TEMPS
Points : 7 986
Annotations : 34/34
CHOIX DU CORPUS
POKEMON
NIVEAU 1
SUJET
JEU
FORMATION
Points : 4 980
Annotations : 22/282
NIVEAU 1
AFFIXE
JEU
FORMATION
Points : 0
Annotations : 0/108
NIVEAU 1
DÉTERMINANT
JEU
FORMATION
Points : 0
Annotations : 0/148
NIVEAU 2
COMPLÉMENT EN « DE »
JEU
FORMATION
Points : 0
Annotations : 0/120
NIVEAU 2
COMPLÉMENT EN « À »
JEU
FORMATION
Points : 0
Annotations : 9/187
NIVEAU 2
OBJET DE LA PRÉPOSITION
JEU
FORMATION
Points : 0
Annotations : 0/213
NIVEAU 2
AUXILIAIRE PASSIF
JEU
FORMATION
Points : 598
Annotations : 9/56


Jouer
Boutique
Forum
Joueurs
Niveau maximum!
165
Trouve le complément (objet indirect introduit par "à") du verbe indiqué !
10%
Besoin d'aide?
Très jeune, il a fait preuve d'initiative et de courage pour PARTICIPER à un sauvetage lors d'inondations.
74
43
57
150
Acheter
1
15
Acheter
3
300
Acheter
0
15
Acheter

1 Crowdsourcing: back to basics

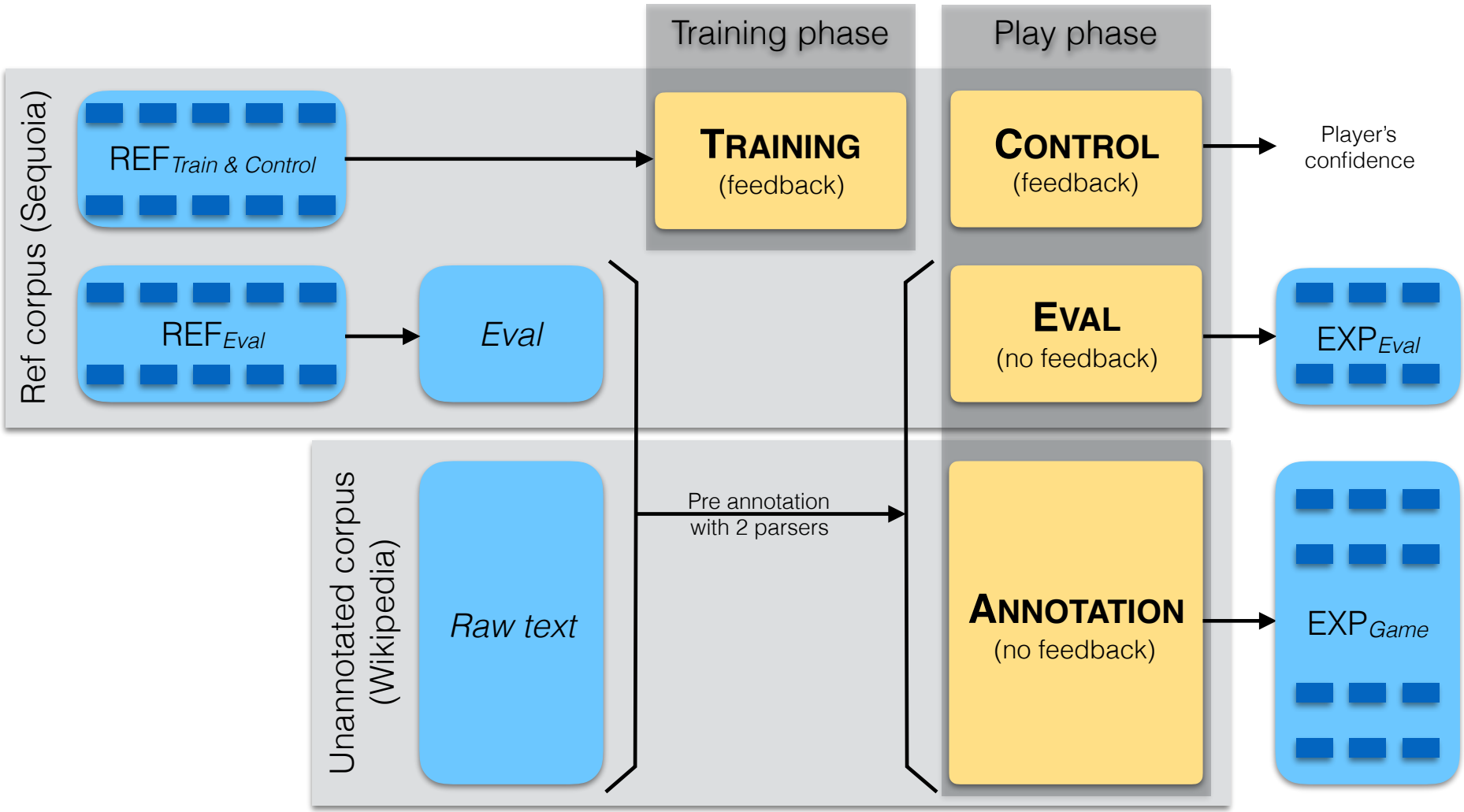
2 Overview of the game

3 Behind the curtain

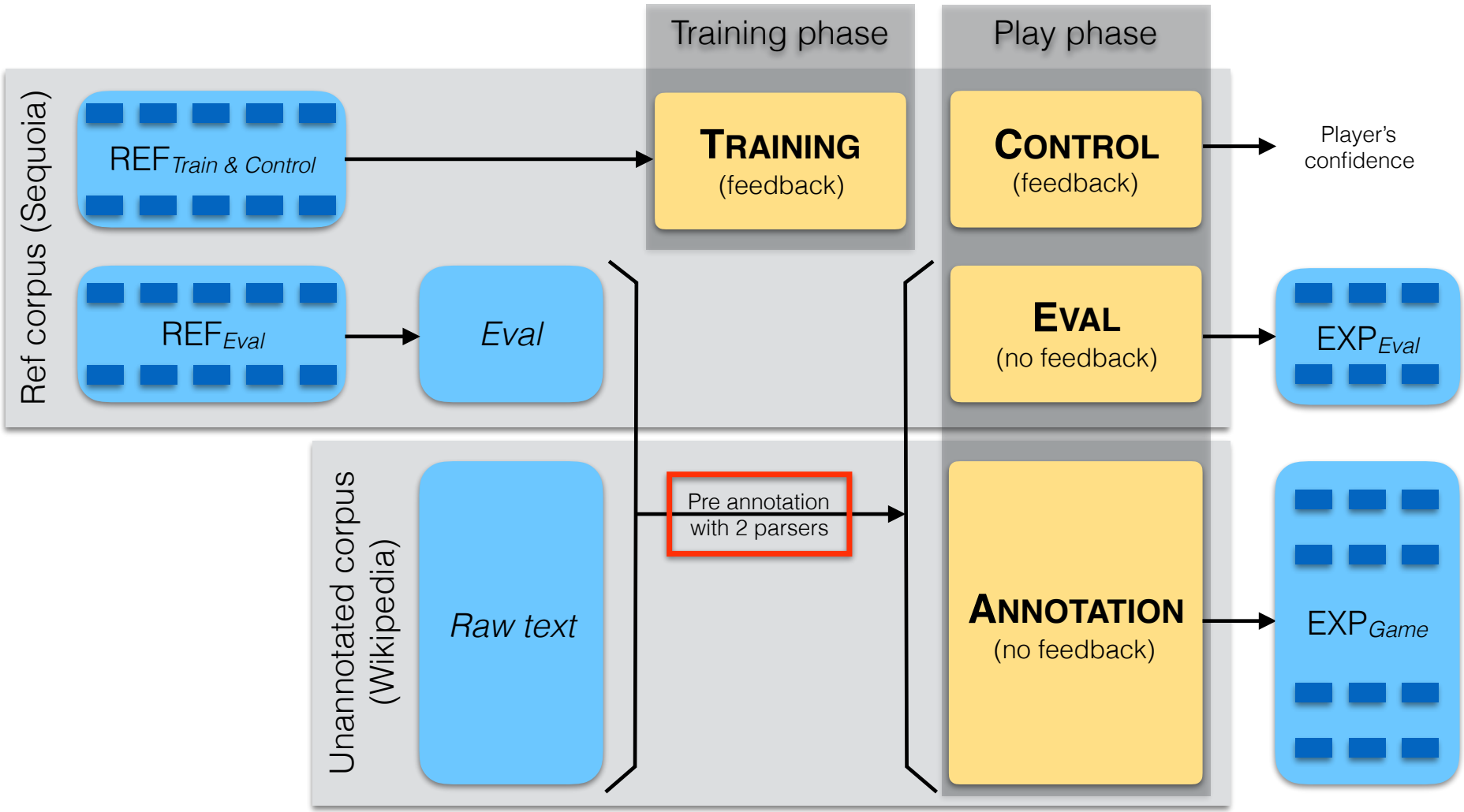
4 Results

5 Conclusion and future plans

Organizing quality assurance



Preprocessing data (freely available corpora)



Preprocessing data (freely available corpora)

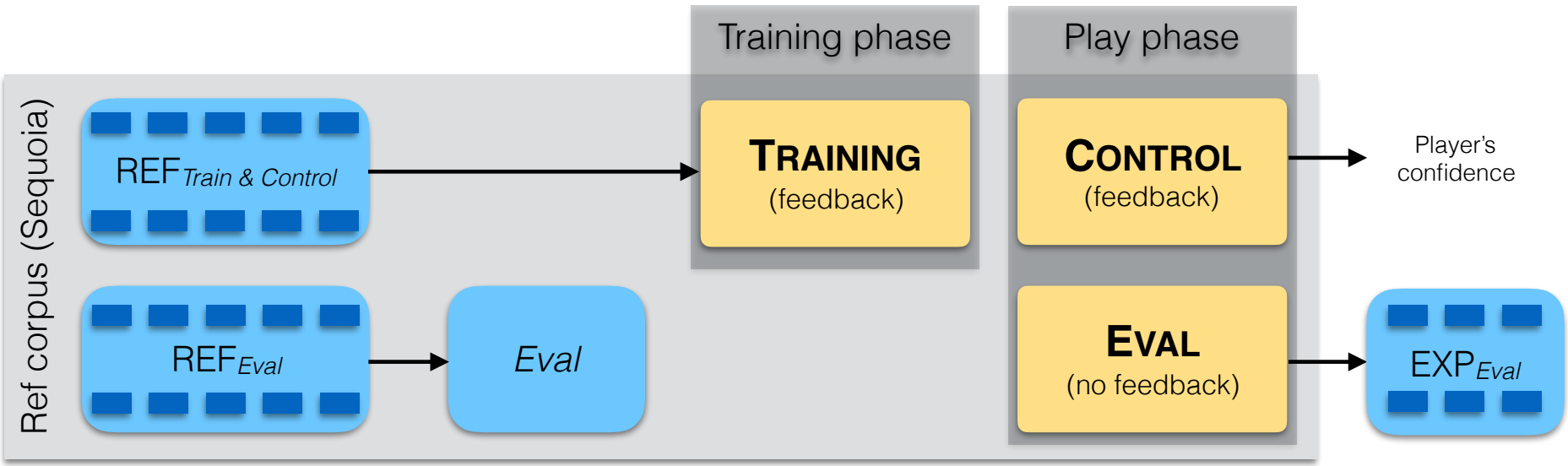
Pre-annotation with two parsers

1. a statistical parser: Talismane [Urieli, 2013]
2. a symbolic parser, based on graph rewriting: FrDep-Parse [Guillaume and Perrier, 2015]

→ play the items for which the two parsers give different annotations

Training, control and evaluation

Reference: 3,099 sentences of the Sequoia corpus [Candito and Seddah, 2012]

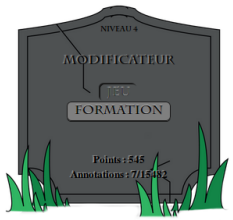


$REF_{Train\&Control}$	REF_{Eval}	Unused
50%	25%	25%
1,549 sentences	776 sentences	774 sentences

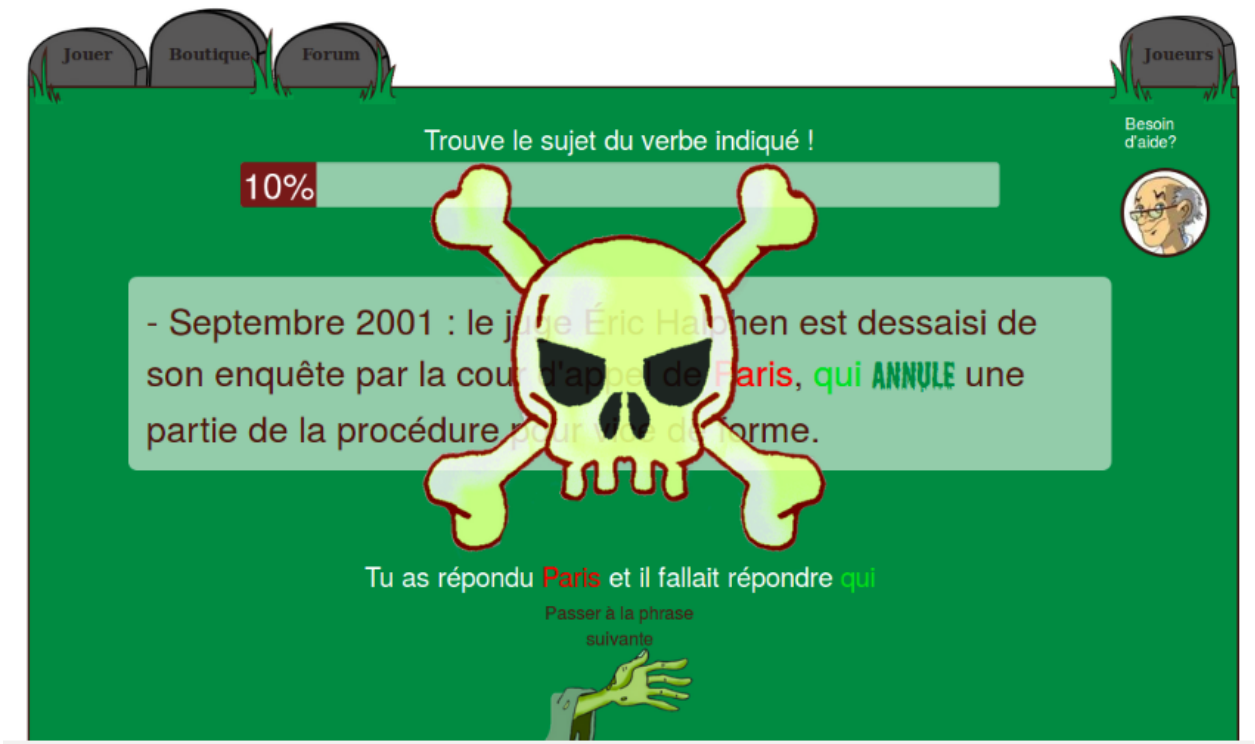
- $REF_{Train\&Control}$ is used to train the players
- REF_{Eval} is used like a raw corpus, to evaluate the produced annotations

Training the players

Compulsory for each dependency relation



- sentences are taken from the REF$_{Train\&Control}$ corpus
- a feedback is given in case of error



Dealing with cognitive fatigue and long-term players

Control mechanism

Sentences from the REF$_{Train\&Control}$ corpus are proposed regularly

1. if the player fails to find the right answer, a feedback with the solution is given

Dealing with cognitive fatigue and long-term players

Control mechanism

Sentences from the REF$_{Train\&Control}$ corpus are proposed regularly

1. if the player fails to find the right answer, a feedback with the solution is given
2. after a given number of failures on the same relation, the player cannot play anymore and has to redo the corresponding training

Dealing with cognitive fatigue and long-term players

Control mechanism

Sentences from the REF$_{Train\&Control}$ corpus are proposed regularly

1. if the player fails to find the right answer, a feedback with the solution is given
2. after a given number of failures on the same relation, the player cannot play anymore and has to redo the corresponding training

$\rightarrow$ we deduce a level of confidence for the player on this relation

1 Crowdsourcing: back to basics

2 Overview of the game

3 Behind the curtain

4 Results

5 Conclusion and future plans

Production: game corpus size

compared to other existing French dependency syntax corpora

As of July 10, 2016

- 647 players
- who produced 107,719 annotations

	Sequoia 7.0	UD-French 1.3	FTB-UC	FTB-SPMRL	Game
Sentences	3,099	16,448	12,351	18,535	5,221
Tokens	67,038	401,960	350,947	557,149	**128,046**
Tokens/sent.	21.6	24.4	28.4	30.1	24.5

Production: game corpus size

compared to other existing French dependency syntax corpora

As of July 10, 2016

- 647 players
- who produced 107,719 annotations

	Sequoia 7.0 *free*	UD-French 1.3 *free*	FTB-UC *not "free"*[1]	FTB-SPMRL *not "free"*	Game *free*
Sentences	3,099	16,448	12,351	18,535	5,221
Tokens	67,038	401,960	350,947	557,149	**128,046**
Tokens/sent.	21.6	24.4	28.4	30.1	24.5

[1] No redistribution allowed.

Production: game corpus size

compared to other existing French dependency syntax corpora

As of July 10, 2016

- 647 players (775 as of Dec. 13th)
- who produced 107,719 annotations (168,832 as of Dec. 13th)

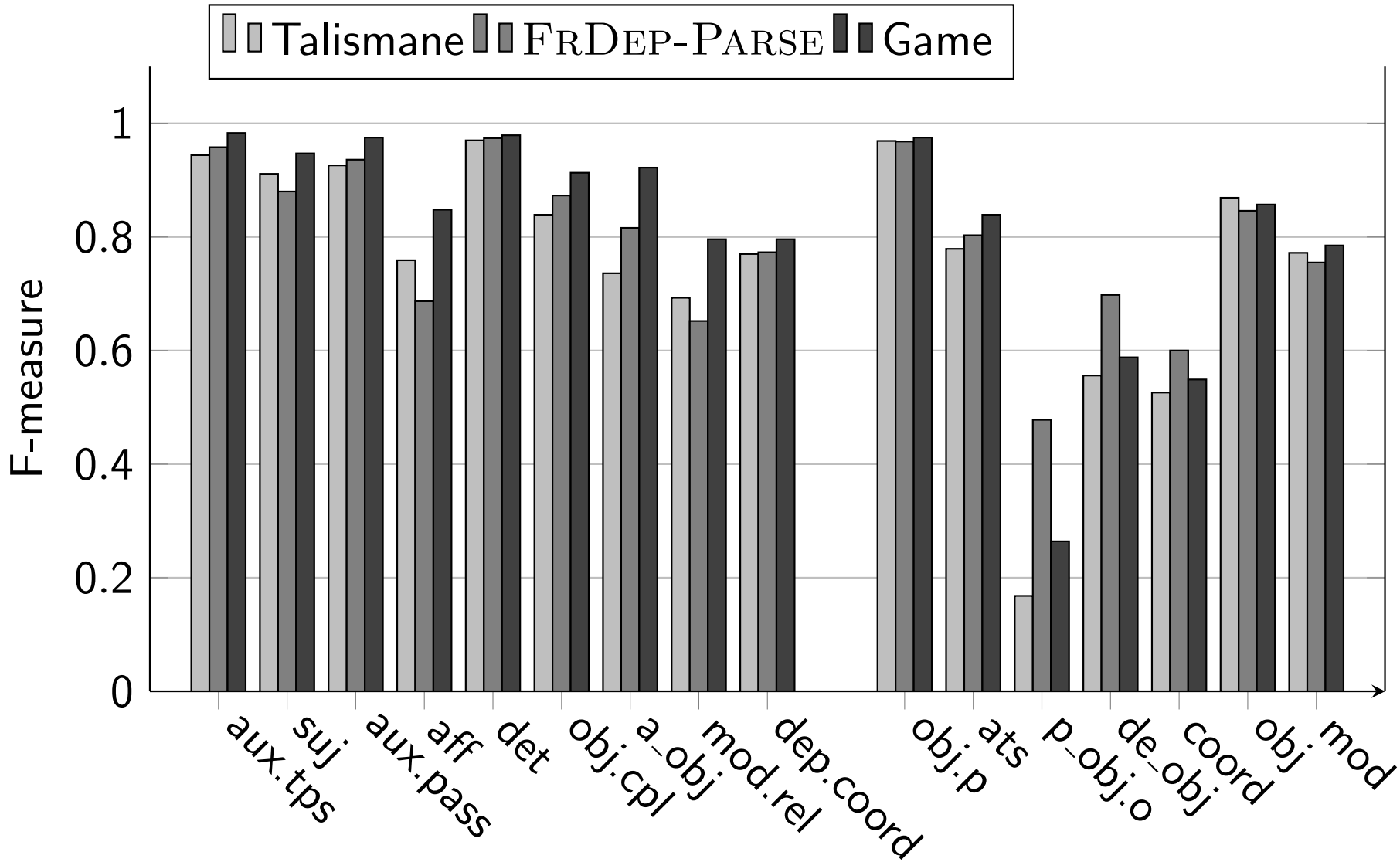
	Sequoia 7.0 *free validated*	UD-French 1.3 *free* after ZL[1] + *errors*	FTB-UC *not "free" validated*	FTB-SPMRL *not "free" validated*	Game *free validated*
Sent.	3,099	16,448	12,351	18,535	5,221
Tok.	67,038	401,960	350,947	557,149	**128,046**
Tok./sent.	21.6	24.4	28.4	30.1	24.5

+ (ever)growing resource!

[1] ZL 1.0, July 2014 *vs* UD 1.0 January 2015.

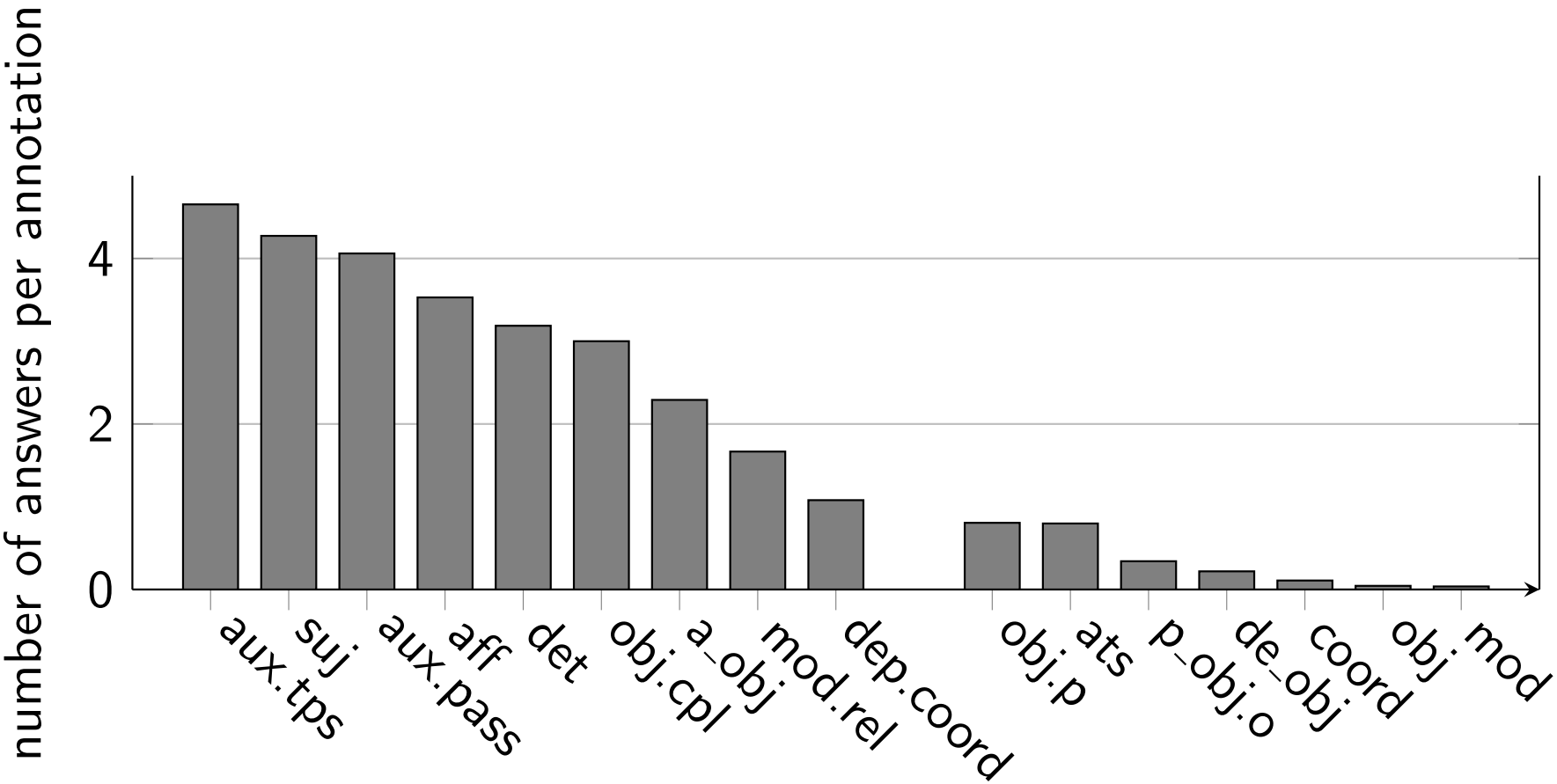
Evaluating quality

on the REF_{Eval} corpus



Annotation density

on the REF_{Eval} corpus



→ need **more** annotations on some relations

1 Crowdsourcing: back to basics

2 Overview of the game

3 Behind the curtain

4 Results

5 Conclusion and future plans

Games With(out) A Problem [Tuite, 2014]?

Achievements

- surprisingly good results in terms of quantity and quality
- we demonstrated that we can train people on a complex task

Yet to be validated

- relation types which are not played (too difficult or lack of players?)

Difficulties

- communication / advertisement
- community management

Improving gamification

- give more to explore and collect
- build a real story
- build a sense of community (how?)

Improving the exported resource

Test the influence of

- the pre-annotation score
- the level of the player in the game
- the confidence we have in the player for the relation type at hand

Expand to new languages and new annotation types

New languages

- English
- **less-resourced languages**

New annotation types

- **part-of-speech** (POS),
- corpus building,
- etc.

Alice Millour (PhD student)

Team and fundings

Bruno Guillaume (researcher)

Nicolas Lefèbvre (engineer)

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère
Culture
Communication

https://github.com/zombilingo

http://zombilingo.org/export

Candito, M. and Seddah, D. (2012).
Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical.
In Proceedings of Traitement Automatique des Langues Naturelles (TALN), Grenoble, France.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008).
Phrase Detectives: a web-based collaborative annotation game.
In Proceedings of the International Conference on Semantic Systems (I-Semantics'08).

Cooper, S., Treuille, A., Barbero, J., Leaver-Fay, A., Tuite, K., Khatib, F., Snyder, A. C., Beenen, M., Salesin, D., Baker, D., and Popović, Z. (2010).
The challenge of designing scientific discovery games.
In Proceedings of the Fifth International Conference on the Foundations of Digital Games, FDG '10, pages 40–47, New York, NY, USA. ACM.

Fort, K., Nazarenko, A., and Rosset, S. (2012).
Modeling the complexity of manual annotation tasks: a grid of analysis.
In Proceedings of the International Conference on Computational Linguistics (COLING), pages 895–910, Mumbai, India.

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. (2011).
Managing the crowd: Towards a taxonomy of crowdsourcing processes.
In AMCIS 2011 Proceedings.

Guillaume, B. and Perrier, G. (2015).
Dependency Parsing with Graph Rewriting.
In
Proceedings of IWPT 2015, 14th International Conference on Parsing Te
pages 30–39, Bilbao, Spain.

Howe, J. (2006).

The rise of crowdsourcing.
Wired Magazine, 14(6).

Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., et al. (2011).
Crystal structure of a monomeric retroviral protease solved by protein folding game players.
Nature structural & molecular biology, 18(10):1175–1177.

Lafourcade, M. and Joubert, A. (2008).
JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes.
In Journées internationales d'Analyse statistique des Données Textuelles (JADT), Lyon, France.

Tuite, K. (2014).
Gwaps: Games with a problem.
Foundations of Digital Games 2014.

Urieli, A. (2013).
Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit.
PhD thesis, Université de Toulouse II le Mirail, France.