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 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 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 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])

- Direct
- Indirect
- Remunerated
- Not remunerated

---

Crowdsourcing: back to basics

**Definition**

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])

- Direct
  - Remunerated
    - Wikipedia
    - Distributed Proofreaders
    - Phrase Detectives
  - Not Remunerated
    - Zombilingo
    - Foldit
    - Jeux de Mots

- Indirect
  - Remunerated
    - Amazon Mechanical Turk
  - Not Remunerated
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
Crowdsourcing: back to basics

Overview of the game

Behind the curtain

Results

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!
Overview of the game ZombiLingo

http://zombilingo.org/

**ZOMBI LINGO**

BIENVENUE À TOI, JEUNE ZOMBIE!

LE MONDE EST CONDAMNÉ, TA TRANSFORMATION EN ZOMBIE A COMMENCE.

POUR SURVIVRE, SUIS MES RÈGLES, IDENTIFIE LES TÊTES ET MANGE-LES.

ATTENTION AUX PIÈGES, ILS SONT NOMBREUX!

**Jouer**

Pas de limite pour toi! Tu accèdes à toutes les options, bonus cachés!

**Karen**

Retrouve ici tes statistiques, et compare ton score avec celui de tes amis!
Overview of the game

ZombiLingo

Mes Ennemis

Challenge

Points

Total

1. NICOZOMBI : 323,293
2. CHOUCHOU : 307,339
3. METHOSSI : 290,255
4. LYCO : 143,137
5. MARLIEBO : 45,590
6. FIREY : 18,932
7. LULU66 : 16,300
8. YEONWOO : 16,182
9. ROB : 14,930
10. KAREN : 14,634
11. NEWK : 14,406

Mes Duels

Duels gagnés : 4
Matchs nuls : 8
Duels perdus : 5

Mes News

Les Pokémon se cachent aussi chez les Zombis ! Depuis ce matin, nous v... lise la suite...

Mon Compte

Modifier mon mot de passe.
Envoi des emails.
Supprimer mon compte.

Statistiques

Parties gagnées : 171
Parties parfaites : 118
Nombre d'objets trouvés : 159
Trouve le complément (objet indirect introduit par "à") du verbe indiqué !

10%

Très jeune, il a fait preuve d'initiative et de courage pour participer à un sauvetage lors d'inondations.
1. Crowdsourcing: back to basics

2. Overview of the game

3. Behind the curtain

4. Results

5. Conclusion and future plans
Organizing quality assurance

- Unannotated corpus (Wikipedia)
- Ref corpus (Sequoia)
  - REF_{Train & Control}
  - REF_{Eval}
  - Raw text
  - Pre annotation with 2 parsers
- Training phase:
  - TRAINING (feedback)
- Play phase:
  - CONTROL (feedback)
  - EVAL (no feedback)
- Annotation (no feedback)

Player's confidence

EXP_{Eval} → EXP_{Game}
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]

![Diagram showing training, control, and evaluation phases.]

<table>
<thead>
<tr>
<th>REF_{Train &amp; Control}</th>
<th>REF_{Eval}</th>
<th>Unused</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>1,549 sentences</td>
<td>776 sentences</td>
<td>774 sentences</td>
</tr>
</tbody>
</table>

- 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\textsubscript{Train&Control} corpus are proposed regularly. 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.

- 1er FÉVRIER 1995 : Jean-Paul Schimpf, un ami intime de Didier Schuller, est arrêté sur un parking, alors que la dirigeante d'une entreprise d'assainissement disait vouloir lui remettre une somme d'argent en liquide.

Tu as répondu une et il fallait répondre arrêté

Tu as un peu oublié comment jouer ce phénomène. Pour continuer à jouer sur celui-ci, tu vas devoir refaire le tutoriel correspondant.
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

→ 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
As of July 10, 2016

- 647 players
- who produced 107,719 annotations

<table>
<thead>
<tr>
<th></th>
<th>Sequoia 7.0</th>
<th>UD-French 1.3</th>
<th>FTB-UC</th>
<th>FTB-SPMRL</th>
<th>Game</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>3,099</td>
<td>16,448</td>
<td>12,351</td>
<td>18,535</td>
<td>5,221</td>
</tr>
<tr>
<td>Tokens</td>
<td>67,038</td>
<td>401,960</td>
<td>350,947</td>
<td>557,149</td>
<td><strong>128,046</strong></td>
</tr>
<tr>
<td>Tokens/sent.</td>
<td>21.6</td>
<td>24.4</td>
<td>28.4</td>
<td>30.1</td>
<td>24.5</td>
</tr>
</tbody>
</table>
Production: game corpus size
compared to other existing French dependency syntax corpora

As of July 10, 2016
- 647 players
- who produced 107,719 annotations

<table>
<thead>
<tr>
<th></th>
<th>Sequoia 7.0 free</th>
<th>UD-French 1.3 free</th>
<th>FTB-UC not &quot;free&quot;</th>
<th>FTB-SPMRL not &quot;free&quot;</th>
<th>Game free</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>3,099</td>
<td>16,448</td>
<td>12,351</td>
<td>18,535</td>
<td>5,221</td>
</tr>
<tr>
<td>Tokens</td>
<td>67,038</td>
<td>401,960</td>
<td>350,947</td>
<td>557,149</td>
<td>128,046</td>
</tr>
<tr>
<td>Tokens/sent.</td>
<td>21.6</td>
<td>24.4</td>
<td>28.4</td>
<td>30.1</td>
<td>24.5</td>
</tr>
</tbody>
</table>

¹No 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)

<table>
<thead>
<tr>
<th></th>
<th>Sequoia 7.0</th>
<th>UD-French 1.3</th>
<th>FTB-UC</th>
<th>FTB-SPMRL</th>
<th>Game</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>free</td>
<td>free</td>
<td>not ”free”</td>
<td>not ”free”</td>
<td>free</td>
</tr>
<tr>
<td></td>
<td>validated</td>
<td>after ZL$^1$+errors</td>
<td>validated</td>
<td>validated</td>
<td>validated</td>
</tr>
<tr>
<td>Sent.</td>
<td>3,099</td>
<td>16,448</td>
<td>12,351</td>
<td>18,535</td>
<td>5,221</td>
</tr>
<tr>
<td>Tok.</td>
<td>67,038</td>
<td>401,960</td>
<td>350,947</td>
<td>557,149</td>
<td>128,046</td>
</tr>
<tr>
<td>Tok./sent.</td>
<td>21.6</td>
<td>24.4</td>
<td>28.4</td>
<td>30.1</td>
<td>24.5</td>
</tr>
</tbody>
</table>

+ (ever)growing resource!

---

$^1$ZL 1.0, July 2014 vs UD 1.0 January 2015.
Evaluating quality
on the $\text{REF}_{\text{Eval}}$ corpus
Annotation density
on the $\text{REF}_{\text{Eval}}$ 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)
Acknowledgements

https://github.com/zombilingo

http://zombilingo.org/export


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

Crystal structure of a monomeric retroviral protease solved by protein folding game players.

JeuxDeMots : un prototype ludique pour l’émergence de relations entre termes.
In Journées internationales d’Analyse statistique des Données Textuelles (JADT), Lyon, France.

Gwaps: Games with a problem.