PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 96 OCTOBER 2011 5–26

An attractive game with the document: (im)possible?

Barbora Hladká, Jiří Mírovský, Jan Kohout

Charles University in Prague, Institute of Formal and Applied Linguistics

Abstract

The annotation experience we have acquired while participating in the Prague treebanking projects provides us with a strong evidence to conclude that the linguistic data annotation by experts is a very intensive and expensive process. No surprise that we care whether we can get the annotated data in a less demanding process. We focus on an alternative way of annotation to generate the data for natural language processing tasks that either have not been implemented yet or have been implemented with a performance lower than human performance. To be more specific, we are interested in ways of annotation gathered mostly under the terms 'crowdsourcing' and 'human computation', i.e. we concentrate on activities that motivate as many non-experts as possible to devote whatever they prefer (effort, time, enthusiasm, responsibility, etc.) to carry out annotation.

In this paper, we review the notion of crowdsourcing, namely we turn our attention to crowdsourcing projects that manipulate textual data. As we are delighted with the games with a purpose, we carry out an implementation of the on-line games with texts. We introduce a game on coreference, *PlayCoref*, and games with words and white spaces in the sentence, *Shannon Game* and *Place the Space*, in great details. The game rules are designed to be language independent and the games are playable with both Czech and English texts by default. After a number of sessions played so far we revise our initial expectations and enthusiasm to design an attractive annotation game with a document.

1. Introduction

The Prague dependency treebanks[1] represent the annotation projects where both textual and spoken data have been annotated by experts. The annotation framework

[1] http://ufal.mff.cuni.cz/pdt.html

Corresponding author: hladka@ufal.mff.cuni.cz
Cite as: Barbora Hladká, Jiří Mírovský, Jan Kohout. An attractive game with the document: (im)possible?. The Prague Bulletin of Mathematical Linguistics No. 96, 2011, pp. 5–26. doi: 10.2478/v10108-011-0008-z.

has a solid theoretical background, namely the Functional Generative Description (FGD, Sgall et al. (1986)), thus the annotation guidelines are coordinated with this theory. Consequently, the annotators are trained according to the guidelines.

The FGD conceives a language as a system of layers, so the Prague treebanking annotation schemes respect this system in such a way that the data is annotated on three layers going from the simplest morphological one through the syntactic-analytical one to the most complex tectogrammatical one. The higher the layer, the higher requirements on the annotator's qualification are expected. While the annotation on the morphological layer can be performed by secondary school students, the annotation on the tectogrammatical layer can be performed by linguists and carefully trained students of the philological studies mainly. The quality of the annotated data must be pursued while formulating the annotation strategy, i.e. criteria to ensure a high quality annotation must be elaborated and a proper number of annotators must be selected. Most of the annotation projects are scheduled at least for five years and the number of people involved in them varies. In average, up to ten member teams are established including annotators and technical staff.

Summing up the annotation experience, we conclude that the linguistic data annotation by experts is a *labour* and *time* and *money* consuming process.[2] No surprise that we care whether we can get the annotated data in a less expensive process. At the same time, we ask *Do we really need (more) annotated data?* Considering data-driven approaches to address natural language processing tasks, the positive answer is replied every time the correlation between the performance and the volume of data needed is evaluated.

In this paper, we focus on an alternative way of annotation to provide the data for NLP tasks that either have not been implemented yet or have been implemented with a performance lower than human performance. To be more specific, we are interested in ways of annotation gathered mostly under the terms 'crowdsourcing' and 'human computation'. One can encounter many other synonyms but we will use these two terms throughout the paper.

The paper is organized as follows. In Section 2, we briefly introduce the notion of crowdsourcing. As we are immersed in the textual data annotation, we turn our special attention to the crowdsourcing projects with texts. At this point, we are very close to the topic of the paper (Wang et al., 2010) discussing the phenomenon of crowdsourcing in NLP for the first time, at least to our knowledge. We will summarize it and add our points of view. We are delighted with the games with a purpose so much that we conceive them as a possible way of textual data annotation. We have proposed and implemented a number of games that are described in detail in Section 3. We conclude with Section 4.

[2] This conclusion is valid for the data in general.

2. Crowdsourcing/Human computation

The online encyclopedia Wikipedia[3] is an exemplary crowdsourcing/human computation system so we list its definition of crowdsourcing and human computation:

- **Crowdsourcing** is the act of outsourcing tasks, traditionally performed by an employee or contractor, to an undefined, large group of people or community (a "crowd"), through an open call.
- **Human-based computation** is a computer science technique in which a computational process performs its function by outsourcing certain steps to humans. This approach uses differences in abilities and alternative costs between humans and computer agents to achieve symbiotic human-computer interaction.

We interpret the distinction between these two terms as follows: the human computation (HC) is the qualification of crowdsourcing to computer-based issues. It is not our intention to discuss the definitions in details. Instead, we refer to a number of more profound resources, like (Crowdsourcing.org, 2011), (Doan et al., 2011), (Ipeirotis and Paritosh, 2011).

The human computation systems can be classified along many dimensions, see e.g. (Quinn and Bederson, 2009), (Yuen et al., 2009). Here, we highlight two of them:

1. The *nature of collaboration*. We are mainly interested in the classes of Games With A Purpose (GWAP), Highly Intelligence Tasks (HITs) hosted by Amazon's Mechanical Turk and Wisdom of the Crowds (WotC) systems. The nature of collaboration closely relates to the motivation to collaborate. The three mentioned classes exemplify motivation by fun, profit and enthusiasm to share knowledge, respectively.
2. The *input data type*. The users absorb the information provided by the input data through different activities like observing the picture, watching the video, listening to music, reading the web page, etc. Each of these activities takes some time the amount of which strongly depends on the input data type. For example, image content understanding takes much less time than understanding of paragraph content.

As long as we search for an alternative way of textual data annotation, we review HC systems that manipulate with the textual data, i.e. either individual words, sentences, paragraphs or even whole documents. We list GWAPs first, then HITs and finally WotC.

- **Jinx**, a two player game, (Seemakurty et al., 2010), shows the players a context, usually a sentence, with an underlined word. The players enter synonyms for the underlined word and attempt to match each other. The output synonym sets are tested against the WordNet (Miller, 1995) and the game data presents a valuable data for a task of word sense disambiguation.

[3] `http://www.wikipedia.org/`

- **Onto Games**, (Siorpaes and Simperl, 2010), create a semantic content. Articles from Wikipedia are presented during the sessions and players answer the pre-generated questions concerning the onthology concepts.
- **PackPlay**, (Green et al., 2010), is a game framework consisting of the *Entity Discovery* game and *Named That Entity* game. The players are asked to annotate named entities in the sentences.
- **Page Hunt**, a single-player game, (Ma et al., 2009), shows the player a random web page (its contents, not its web address) and the player is supposed to ask such a query that brings a given page in the top N results on a search engine. The queries from the winning trials can be used as terms in a task of query alternation.
- **Phrase Detectives**, a single-player game (Chamberlain et al., 2008), traces a relationship between words and phrases in a short text, namely the relationship of coreference. We present details of the game description in Section 3.1.
- **Verbosity**, a two-player game, (von Ahn et al., 2007), generates common sense facts so that one player gets a secret word and provides the hints in a form of sentence templates to the second player that guesses the secret word.
- **Amazon's Mechanical Turk** is an online job market hosting so-called highly intelligence tasks (HITs). Browsing the HITs with textual data, we meet mostly machine translation tasks, tasks like 'write a sentence with a given phrase' or 'write a summary of an article'. (Snow et al., 2008) investigated HITs on affect recognition, word similarity, recognizing textual entailment, event temporal ordering, and word sense disambiguation. They showed high agreement between Mechanical Turk non-expert annotations and existing gold standard labels provided by expert labelers. Similarly, a study by (Kittur et al., 2008) compares the rating of Wikipedia's articles assessed by both Mechanical Turkers and Wikipedia admins. The two experiments they conducted differ in a feature that enables verification how much the Mechanical Turkers are familiar with the content of what they are rating, i.e. how carefully they are reading the articles. They conclude that the Mechanical Turk is a promising platform for conducting various tasks, but special care must be taken in the design of the tasks to avoid unfair processing, especially if the tasks are subjective or qualitative.
- **Wikipedia** is a freely accessible online encyclopedia that everyone can change. It represents the only HC system that works with whole documents.
- **reCAPTCHA** is a system enabling to improve the quality of digitalized books (von Ahn et al., 2008). It iss designed as an upgrade of CAPTCHA system that recognizes whether a person (not computer) is responsing. The recognition runs like a test to rewrite a distorted string of characters exactly. reCAPTCHA submits two character strings, one of them digitally recognized correctly and the other one unrecognized. A user has to rewrite both strings correctly. The system of reCAPTCHA can be classified as WotC system with the attribute 'no other

choice' since the users simply have to rewrite strings to proceed their further web activities.

At least to our knowledge, there is no HC system guiding the user to **carefully** read a document and do some annotation. This fact and our sympathy to the Games with A Purpose methodology strongly motivate us to design and implement such a system.

3. Play the Language Games

We have implemented three games with textual data and published them at `http://www.lgame.cz` portal. The subsection (3.1) describes the PlayCoref game – a game with coreference. It is the only game out of the three that is meant to produce linguistically valuable data. The subsequent subsections (3.2) and (3.3) describe two remaining games – Shannon game and Place the Space, respectively. Their primary purpose is to attract people to this game portal.

We use *A Study in Scarlet* by Sir Arthur Connan Doyle to present the input data into the sessions of all three games. The choice has been made for practical reasons, namely the novel is publicly available and has been translated into many languages; moreover, a free English audio book exists. The book is not difficult to read and it is enjoyable. The English version comes from the Guttenberg project[4] and the Czech translation comes from the portal Literární doupě[5]. The raw data undertook some processing that we specify in the *Game data preparation* sections below.

3.1. PlayCoref

The PlayCoref is a single-player and two-player game with text. During a 5 minute session, the players read a short text and connect words that co-refer. Their task is to connect all co-referring words in as many sentences as possible.

Notion of coreference Let us present the terminology we use: a *referent* is an object referred to in the given text. A *referring expression* is a lexical representation of a referent. *Coreference* occurs when several referring expressions in the text refer to the same entity (e.g. person, thing, fact). A *coreferential pair* (link) is marked between subsequent pairs of the referring expressions. A sequence of coreferential pairs referring to the same entity forms a *coreference chain*.

In the passage from (Doyle, 1887, 2005), one can read the following coreference chain: *I, I, me, I, me man*; another coreference chain *someone, Stamford, who, dresser* can be seen there: *On the very day that **I** had come to this conclusion **I** was standing at the*

[4] `http://www.gutenberg.org/ebooks/244`

[5] `http://ld.johanesville.net/doyle-06-studie-v-sarlatove`

*Criterion Bar, when someone tapped **me** on the shoulder, and turning round. **I** recognized young Stamford who had been a dresser under **me** at Barts. The sight of a friendly face in the great wilderness of London is a pleasant thing indeed to a lonely **man**.*

Simplicity Our primary goal was to design the game as enjoyable as possible, and thus to attract the greatest possible number of the Internet users. In order to make the game attractive, we have simplified the understanding of coreference so that we do not burden the players with linguistic definitions. Instead, the players are encouraged to follow their language instinct in deciding what corefers in the text.

The coreferential links are undirected and we restrict the part-of-speech classes of coreferential pair members only to coreference-relevant classes, for details see below 3.1 and (Hladká et al., 2009b); words of coreference-irrelevant part of speech classes are locked. A simple algorithm for the detection of a few types of the closest multi-word expressions is applied. Thus, for example, *Sherlock Holmes* is presented to the players as a single annotation unit.

The game The game starts with several first sentences of the document displayed in the players' sentence window – see Figure 1. Unlocked words, i.e. potential members of coreferential pairs, are emphasized (here in black, e.g. *I*, *Sherlock Holmes*, *landlady*, *my*…), while the locked words (e.g. *good* or *usual*) are displayed in gray.

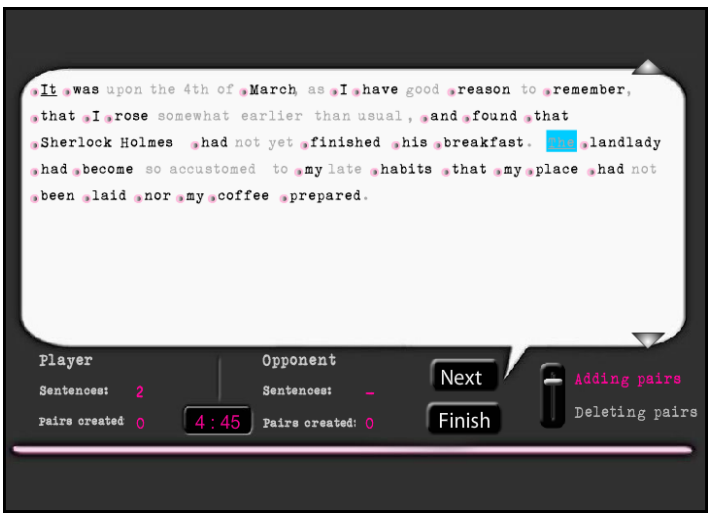

Figure 1. The PlayCoref game starts.

The players mark coreferential pairs as undirected links in the `Adding` mode simply by clicking on dots placed before the active words – see Figure 2 where the player has already created five links. Afterwards, he clicks the button `Next`, another sentence appears and the player adds more links. Links can be deleted in a similar way after switching to the `Deleting` mode.

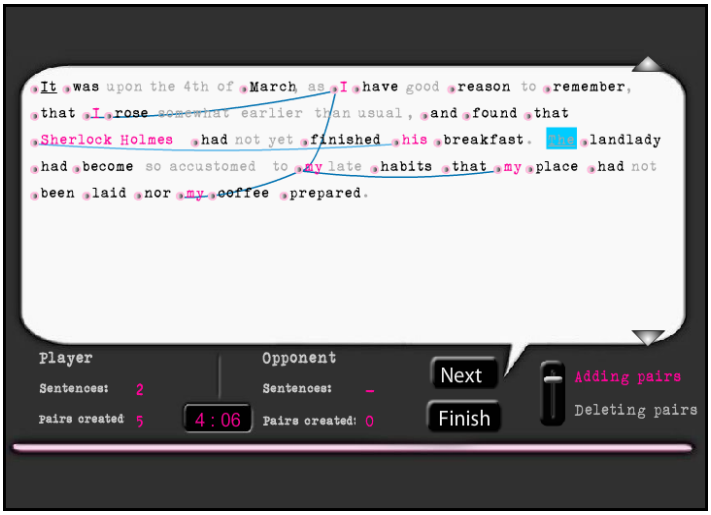

Figure 2. PlayCoref session: Adding links.

Whenever the player finishes pairing words in a visible part of the document (visible to him), he asks for the next sentence of the document by clicking the `Next` button. The sentence appears at the bottom of his sentence window, the first word of the added sentence is highlighted so that it can be recognized immediately. In this manner, the session goes on until the end of the session time (5 minutes) or until the player (both the players in the two-player version of the game) reaches the end of the document (no more sentences are offered and the button `Next` becomes inactive) and he decides to finish the session by clicking the `Finish` button.

During the session, the player has information about the remaining time, the number of his and the opponent's displayed sentences and the number of his and the opponent's created pairs. Revealing more information about the opponent's actions would affect the independency of the players' decisions. Especially, no running score is being presented during the game. Otherwise, the players might adjust their decisions according to the changes in the score, which is undesirable. See our elaboration on the

interactivity issues in (Hladká et al., 2009a). In the single-player version, naturally, no information about the opponent is available.

Figure 3 shows a possible situation of the game closely before its end. The player has already asked for several more sentences, so they do not fit into the window – the text can be scrolled up and down using the arrows on the right side or the mouse wheel. `Deleting` mode is active.

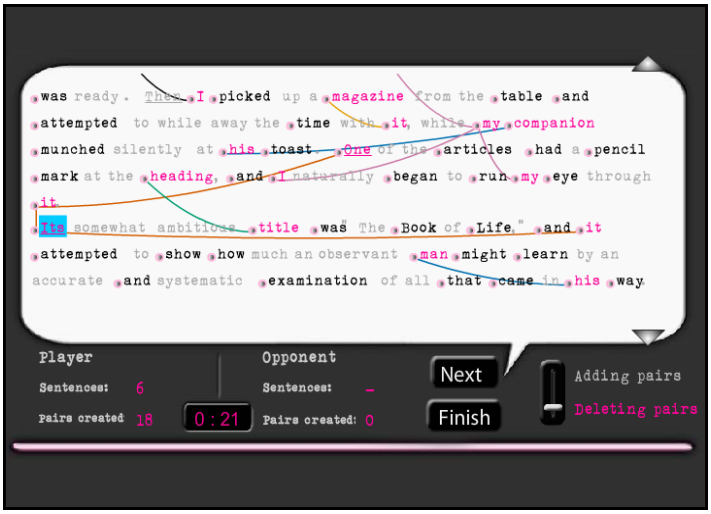

Figure 3. PlayCoref session closesly before its end. `Deleting` mode is active.

At the end of the session – see Figure 4, the result of the game is displayed. It contains information about the player: his final numbers of links, and, of course, his score (the scoring function is described below). In the two-player version, results for both the players are displayed.

Game data preparation In principle, any document can be used in the game, but the following processing steps are necessary.

Tagging The morphological tagging, usually preceded by tokenization, is required to recognize part-of-speech classes and sub-part-of-speech categories (if needed),

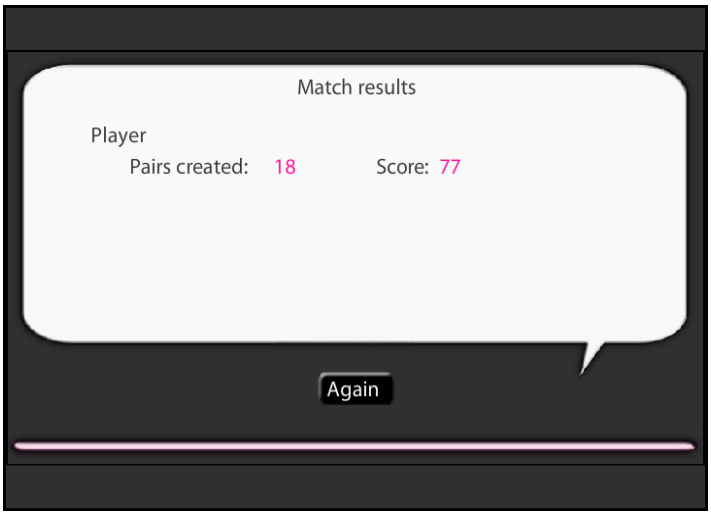

Figure 4. A result of the single-player version of PlayCoref.

in order to lock/unlock individual words for the game. For most languages, tagging is a well solved problem (e.g. for Czech the MORČE tagger[6], for English TnT tagger[7]).

Word locking Words of coreference-relevant POS classes are allowed to become parts of coreferential links marked between individual words or short named entities only. Coreferential pairs that link larger text parts (like several sentences) are disregarded since their marking would be too complex for the players.

We specify the coreference-irrelevant POS classes first. Then the particular words get locked and they are graphically distinguished, so that the players will not consider them at all during the sessions. For English, we work with the PennTreebank tagset (Marcus et al., 1993) and we lock words that are assigned with one of the following POS tags: DT (determiner), IN (preposition or subordinating conjunction), TO (*to*), RB (adverb), RP (particle), JJ (adjective). For Czech positional tag system (Zeman et al., 2005), Table 3.1 shows a list of locked sub-part-of-speech classes of pronouns. Some other POS classes get locked as well: A (adjective), C (numeral), D (adverb), I (interjection), R (preposition), T (particle), and Z (punctuation). So only nouns, selected pronouns, conjunctions and verbs are available for linking in the sessions with Czech texts.

[6] `http://ufal.mff.cuni.cz/morce`

[7] `http://www.coli.uni-saarland.de/~thorsten/tnt/`

Locked pronouns: subPOS and its description	
D	Demonstrative (”ten”, ”onen”, ..., lit. ”this”, ”that”, ”that”, ... ”over there”, ...)
E	Relative ”což” (corresponding to English which in subordinate clauses referring to a part of the preceding text)
L	Indefinite ”všechen”, ”sám” (lit. ”all”, ”alone”)
O	”svůj”, ”nesvůj”, ”tentam” alone (lit. ”own self”, ”not-in-mood”, ”gone”)
Q	Relative/interrogative ”co”, ”copak”, ”cožpak” (lit. ”what”, ”isn’t-it-true-that”)
W	Negative (”nic”, ”nikdo”, ”nijaký”, ”žádný”, ..., lit. ”nothing”, ”nobody”, ”not-worth-mentioning”, ”no”/”none”)
Y	Relative/interrogative ”co” as an enclitic (after a preposition) (”oč”, ”nač”, ”zač”, lit. ”about what”, ”on”/”onto” ”what”, ”after”/”for what”)
Z	Indefinite (”nějaký”, ”některý”, ”číkoli”, ”cosi”, ..., lit. ”some”, ”some”, ”anybody’s”, ”something”)

Table 1. List of pronoun sub-POS classes in the Czech positional tag system locked in PlayCoref.

Automatic and manual coreference annotation For calculating the players’ score (see below), some approximation of the correct solution is needed. If an automatic procedure for coreference resolution (ACR) is available for a language, it can be used. In our experience, however, all available ACR algorithms (both for English and Czech) perform very poorly[8] and cannot be used as a reasonable basis for the scoring function. Until another satisfactory way is found, we present to the game sessions data that is manually annotated, which is a sufficient solution for the initial experiments with the game.

A raw text format of Doyle’s novel was processed by a sequence of tools performing sentence segmentation, tokenization, morphological analysis, tagging, syntactical parsing and semantic parsing, using modules from the TectoMT system (Žabokrtský et al., 2008), and for Czech also the tool-chain from the CAC 2.0 CD-ROM (Hladká et al., 2008). Then two annotators trained for the coreference annotation [9] annotated

[8]For English, we tried Reconcile (Stoyanov et al., 2010), OpenNLP (`http://opennlp.sourceforge.net/models.html`), GuiTAR (`http://cswww.essex.ac.uk/Research/nle/GuiTAR/gtarNew.html`), and BART (Versley et al., 2008); some of the tools did not work at all, the others performed very poorly, especially on the text with dialogues. For Czech, there are almost no tools for ACR. The only one we know, (Novák, 2010) performs very poorly as well.

[9]Two students who participate in the project of coreference and bridging anaphora annotation in the Prague Dependency Treebank (Nedoluzhko et al., 2009)

coreferential links on the tectogrammatical layer. In English, this does not make much difference from the annotation on the surface layer, but in Czech, which is a pro-drop language, some post-processing had to be done. On the tectogrammatical layer in Czech, omitted pronouns are reconstructed and they naturally become parts of coreferential links/chains. As PlayCoref works on the surface layer, omitted pronouns have to be removed from the coreferential chains.

Figure 5 shows an example of a coreferential chain from which a reconstructed pronoun (omitted on the surface) needs to be removed during the transformation of the coreference annotation to the surface form of the text. It is an automatically parsed sentence (actually, two sentences incorrectly parsed as one): *„Tak je to správné.“ „Ano, je, ale přehánět se to nesmí.“*, in English literally: *“It is right so.” “Yes, [it] is, but it must not be exaggerated.“* The three pronouns *it* form a coreferential chain, however the middle one is omitted in the surface form of the Czech sentence. It has to be removed from the coreferential chain. Thus, a new coreferenctial link is created between the two remaining pronouns.

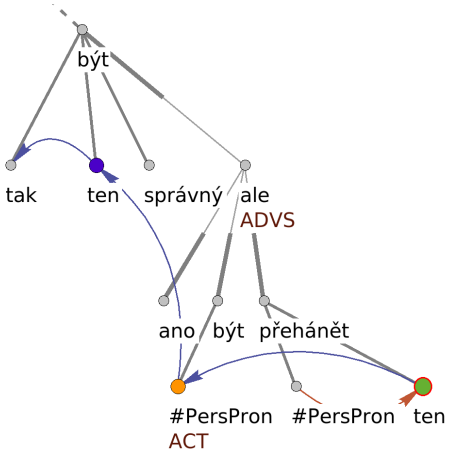

Figure 5. A reconstructed pronoun to be removed from a coreferential chain. In Czech, the two pronouns ”ten” are connected via the reconstructed pronoun (marked with #PersPron and ACT). During the transformation of the coreference annotation to the surface, a direct link between the two pronouns “ten“ is created.

Players' score We want to obtain a large volume of data and thus we must first attract the players and motivate them to play the game again and again. As a reward for their effort, we present scoring. We hope that the players' appetite to win, to confront with

their opponents and to place well in the long-term top scores tables correlate with our research aims and objectives.

Our goal is to ensure the highest quality of the annotation (see also (Hladká et al., 2009a)). The scoring function should reflect the data quality and thus motivate the players to produce correct data. The agreement with the manual expert annotation would be a perfect scoring function. However, the manual annotation is not available for all languages and above all, it is not our goal to annotate data already annotated.

An automatic coreference resolution procedure with a decent accuracy might serve as a first approximation for the scoring function (but as mentioned before, such procedures are not available). As the procedure makes errors, we need to add another component. We suppose that most of the players will mark the coreferential pairs reliably. Then an agreement between the players' pairs indicates correctness, even if the pair differs from the output of the automatic coreference resolution procedure. Therefore, the inter-player agreement becomes the second component of the scoring function. To motivate the players to ask for more parts of the text (and not only "tune" links in the initially displayed sentences), the third component of the scoring function awards the number of created coreferential links.

Scoring function After the game ends, coreference links are automatically checked for circles. If there are some, superfluous links are removed. Otherwise, the circles would harm the scoring function.

In the two-player version of the game, the players get scored (see also (Hladká et al., 2009b)) for their coreferential pairs according to the equation

$$\begin{aligned} score(Player_A) = \lambda_1 * F(Player_A, ACR \text{ or } Manual) \\ +\lambda_2 * F(Player_A, Player_B) \\ +\lambda_3 * min(12, sntncs)/12, \end{aligned}$$

where F stands for the $F - measure = \frac{2*Precision*Recall}{Precision+Recall}$. If the manual annotation is available, we check the player's annotation against it (i.e. we compute $F(Player_A, Manual)$). If a decent automatic coreference resolution method were available, we might check the player's solution against its output (i.e. we would compute $F(Player_A, ACR)$); $sntncs$ is the number of sentences used by the player in the game session. We include the ratio $min(12, sntncs)/12$ as a motivation parameter to inspire players to mark pairs in at least 12 sentences. We have selected the threshold of 12 sentences empirically, which is a reasonable number of sentences the players are able to read and process during the session time. Weights $0 \leq \lambda_1, \lambda_2, \lambda_3 \leq 1, \lambda_1, \lambda_2, \lambda_3 \in R$ (summing to 1) are set empirically.

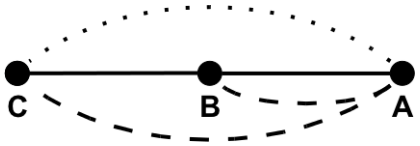

Figure 6. Player '1' pairs (A,C) – the dotted curve; player '2' pairs (A,B) and (B,C) – the solid lines; player '3' pairs (A,B) and (A,C) – the dashed curves. Although players '1' and '2' do not agree on the coreferential pairs at all, '1' and '3' agree only on (A,C) and '2' and '3' agree only on (A,B), for the purposes of the coreference chains reconstruction, the players' agreement is higher: players '1' and '2' agree on two members of the coreferential chain: A and C, players '1' and '3' agree on A and C as well, and players '2' and '3' achieved agreement even on all three members: A, B, and C.

In the single-player version of the game, the scoring function is similar – it only lacks the second member:

$$\begin{aligned} \mathrm{score}(\mathrm{Player_A}) = \lambda_1 * \mathrm{F}(\mathrm{Player_A}, \mathrm{ACR\ or\ Manual}) \\ +\lambda_2 * \mathrm{min}(12, \mathrm{sntncs})/12, \end{aligned}$$

During the calculation of the F-measure, the links of the "annotation" that we compare to are treated as a transitive relation. This solves the issues depicted and described in Figure 6. It does not matter whether the player connects the corefering words as a linear chain or as a star (all to one); also, omitting a word in the chain does not mean a complete disagreement.

Interactivity Issues The degree of the player-to-player interactivity contributes to the attractiveness of the game. From the player's point of view, the more interactivity, the better. For example, knowing both his and the opponent's running score would be very stimulating for the mutual competitiveness. From the linguistic point of view, once any kind of interaction is allowed, statistically pure independency between the players' decisions is lost. A reasonable trade-off between the interactivity and the independency must be achieved. Interactivity that would lead to cheating and decreasing the quality of the game data must be avoided.

Allowing the players to see their own running score would lead to cheating. The players might adjust their decisions according to the changes in the score. Another possible extension of interactivity that would lead to cheating is highlighting words that the opponent used in the coreferential pairs. The players might then wait for the opponent's choice and, again, adjust their decisions accordingly.

Such game data would be strongly biased. However, we still believe that a slight idea of what the opponent is doing can boost inter-coder agreement and yet avoid cheating. Revealing the information about the opponent's number of pairs and the

number of displayed sentences offers at least a little interactivity, yet it will not harm the quality of the data.

Comparison with Phrase Detectives At least to our knowledge, there are no other GWAPs dealing with the relationship among words in a text like PhraseDetectives and PlayCoref. Let us mention some important differences between these two games.

The main difference is in the basic principle of the games: PhraseDetectives game offers the player a whole paragraph and asks him one specific question at a time, e.g. "Are these two words coreferential?", or "Does this word co-refer with another word in the previous text? If so, with which one?". PlayCoref, on the other hand, presents the text to the player sentence by sentence and asks him to search for all coreferential relations in it. Table 2 offers a comparison of various features of the games.

PlayCoref	*PhraseDetectives*
detection of coreference chains	anaphora resolution
single or two-player game	single-player game
a document presented sentence by sentence	a paragraph presented at once
one text in several sessions	checking the pairs marked in the previous sessions
pairing not restricted to the position in the text	the closest antecedent
simple instructions	players training
scoring with respect to the automatic coreference resolution and to the opponent's pairs	scoring with respect to the players that played with the same document before
coreferential pairs correction	no corrections allowed

Table 2. PlayCoref vs. PhraseDetectives.

The very first sessions played We organized the very first PlayCoref competition as an associated event of the CLARA Course on Treebank Annotation.[10] The course participants were either computational linguistics graduates or research associates. In 10 days, 9 different players played 46 sessions that resulted in 945 coreferential pairs in 451 sentences.

We have measured the agreement between each player and the manual annotation and between the players. We use a very similar measure technique as in the scoring

[10] http://lgame.ms.mff.cuni.cz/lgame/sb/competition.php

	F_{chains} (%)
$Player_1$	75 | 57,9 | 69.5 | 72.1 | 75 | 62.6 | 56.3 | 56.1 | 32.1 | 38.6 | 55.8
$Player_2$	54.8 | 78.7 | 75 |81.6 | 79.9 | 72.7 | 55.3 | 68.6 | 68.5 | 56.2 | 58.3 | 46.5 | 75 | 69.1 | 68.2 | 72.5 | 71.9 | 57 | 64.6 | 65.7

Table 3. Most productive players and their performance

function. We calculate Recall, Precision and F-measure using their standard definitions, directly on the links and on the chains as well. Measuring the agreement between two players, only F-measure is interesting because it is symmetric. We propose two ways of calculation:

1. We assume individual links as annotation units for both players. We mark the agreement on links F_{links}.
2. We take links of one player and compare them with coreferential chains of the second player (or the manual annotation). I.e., if one player links nodes A—C, and the other player links nodes A—B and B—C, there is an agreement on the link A—C (see Figure 6). Using the first measure, it would be disagreement. This measure, marked F_{chains}, is the more important one. The same method is used in the scoring function in the game itself, as described above.

We observe first the game data for the players separately. We list statistics for two most productive players $Player_1$, $Player_2$ who played a great majority of the game sessions (11 and 20 out of 46) and we are also interested in whether their performance was improving with the increasing number of sessions. In Table 3, we list F_{chains} for successive sessions starting with the first session played. We can see that $Player_1$'s agreement was getting worse within his session series while $Player_2$'s agreement was more or less well balanced. In general, these numbers give a true picture how player concentration changes over playing time. Mainly, text comprehension in PlayCoref requires a relatively high concentration.

The average value of F_{chains} for all competitors is 60% and the minimal and maximal values are 13.5% and 81.6%, respectively. For illustration, we list the corresponding values of F_{links} – 55%, 10%, 81.6%.

We analyze the agreement between the manual annotation and the union and intersection of players' annotations (on the part of the data where the parallel annotations are available). The obvious expectation is that Precision of the intersection will be higher than Precision of the union, and on the other hand, Recall of the intersection will be lower than Recall of the union.

In total, 11 double player games were played, which resulted in 455 different coreferential pairs linked in 11 different documents: 123 pairs were linked by two players (i.e. they are elements of the intersection) and 332 links were marked by at least one

player (i.e. they are elements of the union). It is interesting to note that 252 pairs were linked in the manual annotation of 11 documents.

The average values of P^{union}_{chains} and R^{union}_{chains} are 65% and 71%, respectively, and the values of $P^{intersection}_{chains}$ and $R^{intersection}_{chains}$ are 88% and 43%, respectively. That confirms the theoretical expectation.

To set an upper bound of the players' annotation agreement, we measured the annotation agreement between the two annotators who manually marked the coreference chains during the preparation of the game data. On 110 sentences annotated in paralell we got $F_{links} = 94\%$ and $F_{chains} = 95\%$. The average agreeement of players who played 11 two-player games is $F_{links} = 57\%$ and $F_{chains} = 65\%$.

3.2. Shannon game

Shannon game is a game for one or two players with hidden words in the sentence. The players guess the words with the help of unhidden words in the sentence.

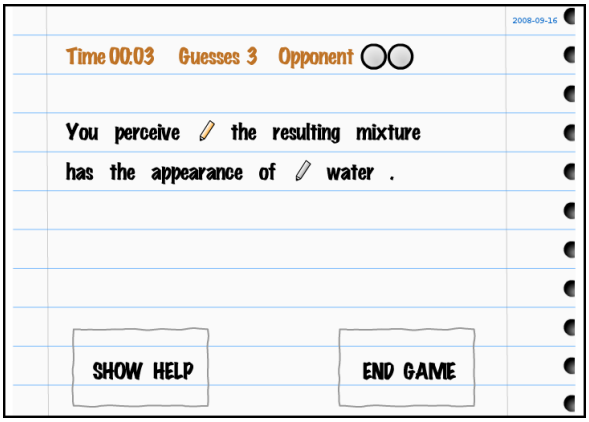

Figure 7. The Shannon game starts.

The game For each missing word, the player has three attempts (guesses). The player simply writes a word and pushes the `Enter` button. If the word is correct, it becomes green and the player moves to the next missing word. If it is not correct, the player looses one guess and can start writing another word as his next attempt. If he guesses incorrectly for three times, the correct word is displayed in red and the player moves to the next missing word. If all missing words have been (correctly or three times incorrectly) guessed, the game ends. The player can also end the game sooner by clicking on the button "End game".

At the beginning of the game, the player chooses one of three difficulty levels. The higher the difficulty, the higher the number of missing words in the sentence: either 2 or 3 or 4 hidden words.

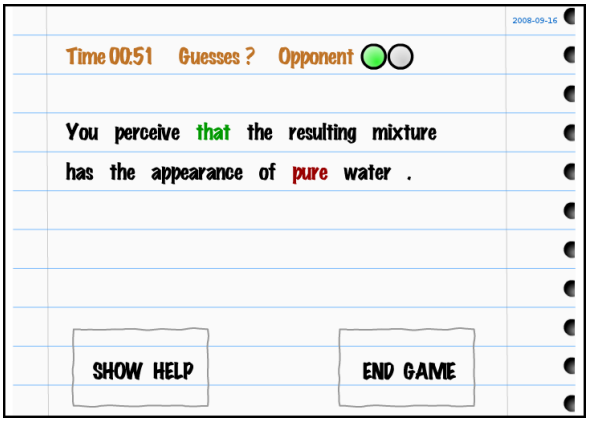

Figure 8. Shortly before the end of the Shannon game.

Game data preparation Any textual data can be used in the game. To pre-process the data, sentence segmentation and tokenization are needed. Only sentences of certain length and without the punctuation are selected. As proper names are almost impossible to guess without a broader context, they should not be used as missing words. Therefore, a procedure for proper names recognition is also needed.

Players' score We do not use any special formula to compute the players' score. Instead, a very straightforward point assignment is applied. For each word being guessed, the player gets points depending on the number of wrong guesses:

- 40 pts – if the 1st guess is correct
- 20 pts – if the 2nd guess is correct
- 10 pts – if the 3rd guess is correct
- -10 pts – if no guess is correct

For example, if two words are hidden, the total score of the player can range from -20 (no word guessed correctly) to 80 (both words guessed correctly at the first attempt) points.

Figure 9. The Shannon game – the players' score.

3.3. Place the Space

Place the Space is a single-player game of word segmentation. The player is presented with a sentence depicted without spaces between words. His task is to restore the spaces in a time-limit set up according to the length of the sentence.

The game To place the space, the player clicks on the character that should immediately follow the space. It can be later removed by clicking on the space.

Game data preparation Any textual data can be used in the game and only sentence segmentation is needed to process the data. To select sentences of a certain length (not too short and not too long), we also use tokenization and we select sentences according to the number of tokens. The number of characters would also be a sufficient measure.

Players' score The score ranges from 0 to 100 (both included) counted as the F-measure between the correct solution and the player's solution.

$$score(Player) = 100 * F(Player,\ Correct)$$

where F stands for the $F-measure = \frac{2*Precision*Recall}{Precision+Recall}$, and (in a standard way), Precision = (number of correctly guessed positions)/(number of guesses), and Recall = (number of correctly guessed positions)/(number of correct positions).

As spaces are naturally written in English and Czech texts, for these languages the game only serves to attract people to the game portal. However, it is a fast-paced and a

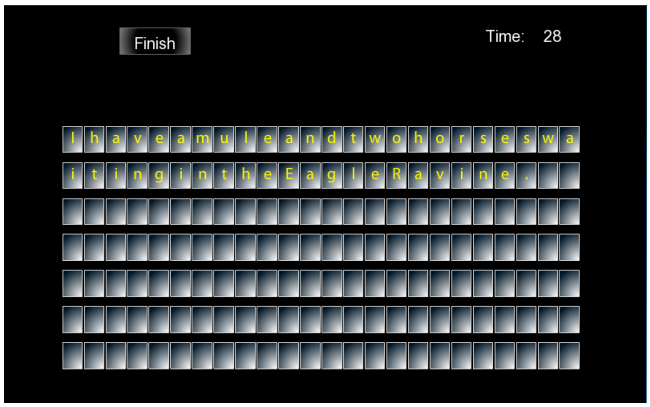

Figure 10. The Place the Space game starts.

very simple game that requires no training. For some languages like Thai or Chinese, where there are no spaces written between words, the game might produce some useful data. However, counting the score would require an automatic procedure for word segmentation; also a comparison to previous solutions by other players could be used.

4. Conclusion

We pay attention to crowdsourcing projects with the textual data, namely we concentrate on games with a purpose. These textual games present a minority of games simply because reading a text is not so enjoyable like for example observing pictures. Notwithstanding this fact, we have designed and implemented the PlayCoref game on marking coreference in the document. Even more, we have organized the very first PlayCoref competition where totally 46 sessions have been played. We are aware that such number of sessions is not large enough to make fundamental conclusions. On the other hand, the competition has encouraged our enthusiasm for PlayCoref game because the competition statistics make sense and the players enjoyed the game.

We implemented two more games, Shannon game and Place the Space. There is no specific natural language processing task to address through these games. We have designed them mainly to invite the Internet users to our language game portal.

To finish primary steps with the text games, a number of implementation actions will be done.

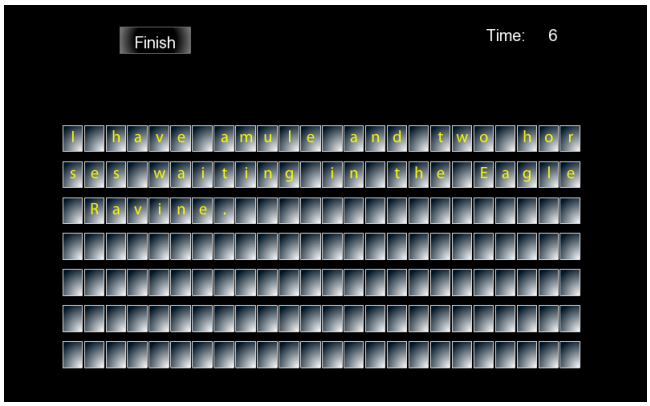

Figure 11. Place the Space: A player's solution with one error.

Figure 12. Place the Space: a player's score, along with the correct solution and an indication of the error.

Acknowledgements

We gratefully acknowledge the support of the Grant Agency of the Czech Republic (grants 405/09/0729 and P406/2010/0875) and the Czech Ministry of Education (grants MSM-0021620838 and LM2010013).

The authors would like to thank Eva Hajičová and Petr Sgall for their valuable comments that helped improve the paper. We really appreciate the patience of Martin Popel, the technical editor of PBML.

Bibliography

Chamberlain, Jon, Massimo Poesio, and Udo Kruschwitz. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of I-SEMANTICS '08 Graz, Austria, September 3-5*, 2008.

Crowdsourcing.org, 2011. URL `http://crowdsouring.org`.

Doan, Anhai, Raghu Ramakrishnan, and Alon Y. Halevy. Crowdsourcing Systems on the World-Wide Web. *Communication of the ACM*, 54(4):86–96, 2011.

Doyle, Arthur Conan. *A Study in Scarlet*. 1887, 2005.

Green, Nathan, Paul Breimyer, Vinay Kumar, and Nagiza F Samatova. PackPlay: Mining semantic data in collaborative games. In *Proceedings of the Fourth Linguistic Annotation Workshop, Uppsala, Sweden*, pages 227–234. 2010.

Hladká, Barbora, Jiří Mírovský, and Pavel Schlesinger. Designing a Language Game for Collecting Coreference Annotation. In *Proceedings of the Third Linguistic Annotation Workshop (ACL-LAW III), Singapore*, pages 52–55, 2009a.

Hladká, Barbora, Jiří Mírovský, and Pavel Schlesinger. Play the Language: Play Coreference. In *Proceedings of ACL-IJCNLP 2009, Singapore*, pages 209–212, 2009b.

Hladká, Barbora Vidová, Jan Hajič, Jiří Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Raab. Czech academic corpus 2.0, 2008.

Ipeirotis, Panagiotis G. and Praveen K. Paritosh. Tutorial: Managing Crowdsourced Human Computation, 2011.

Kittur, Aniket, Ed H. Chi, and Bongwon Suh. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the CHI 2008, Florence, Italy*, 2008.

Ma, Hao, Raman Chandrasekar, Chris Quirk, and Abhishek Gupta. Page Hunt: Improving Search Engines Using Human Computation Games. In *Proceedings of the SIGIR, Boston, MA, USA*, 2009.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Miller, George A. Wordnet: A lexical database for english. *Communication of the ACM*, 38(11): 39–41, 1995.

Nedoluzhko, Anna, Jiří Mírovský, Radek Ocelák, and Jiří Pergler. Extended coreferential relations and bridging anaphora in the prague dependency treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India*, pages 1–16. 2009.

Novák, Michal. Machine learning approach to anaphora resolution. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, 2010.

Quinn, Alexander J. and Benjamin B. Bederson. A taxonomy of distributed human computation, 2009.

Seemakurty, Nitin, Jonathan Chu, Luis von Ahn, and Anthony Tomasic. Word Sense Disambiguation via Human Computation. In *Proceedings of the KDD-HCOMP, Washington, DC, USA*, 2010.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht, 1986.

Siorpaes, Katharina and Elena Simperl. Incentives, Motivation, Participation, Games: Human Computation for Linked Data. In *CEUR Proceedings of the Workshop on Linked Data in the Future Internet at the Future Internet Assembly, Ghent, Belgium*, 2010.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of EMNLP 2008, Honolulu*, page 254–263. 2008.

Stoyanov, V., C. Cardie, N. Gilbert, E. Riloff, D. Buttler, and D. Hysom. Coreference resolution with reconcile. In *Proceedings of ACL 2010, Short Paper*, 2010.

Versley, Yannick, Simone Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. Bart: A modular toolkit for coreference resolution. In *Proceedings of LREC'08*, Marrakech, Morocco, May 2008. ELRA. ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

von Ahn, Luis, Shiry Ginosar, Mihir Kedia, and Manuel Blum. Verbosity: A Game for Collecting Common-Sense Knowledge. In *Proceedings of the SIGHI Conference on Human Factors in Computing Systems, ACM Press*, pages 75–78. 2007.

von Ahn, Luis, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321 (5895):1465–1468, 2008.

Wang, Aobo, Cong Duy Vu Hoang, and Min-Yen Kan. Perspectives on Crowdsourcing Annotations for Natural Language Processing, 2010. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.172.559`.

Yuen, Man-Ching, Ling-Jyh Chen, and Irwing King. A survey of human computation systems. In *Proceedings of the Computational Science and Engineering, IEEE International Conference*, page 723–728. 2009. URL `http://doi.ieeecomputersociety.org/10.1109/CSE.2009.395`.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *ACL 2008 WMT*, pages 167–170, Columbus, OH, USA, 2008. ACL. ISBN 978-1-932432-09-1.

Zeman, Dan, Jiří Hana, Hana Hanová, Jan Hajič, Barbora Hladká, and Emil Jeřábek. A manual for morphological annotation, 2nd edition. Technical Report 27, ÚFAL MFF UK, Prague, Czech Republic, 2005.

Address for correspondence:
Barbora Hladká
`hladka@ufal.mff.cuni.cz`
UK MFF, ÚFAL
Malostranské nám. 25, 118 00 Prague 1, Czech Republic