

# Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation

MASSIMO POESIO, JON CHAMBERLAIN, and UDO KRUSCHWITZ, University of Essex  
LIVIO ROBALDO, University of Turin  
LUCA DUCCESCHI, University of Utrecht/Verona

We are witnessing a paradigm shift in Human Language Technology (HLT) that may well have an impact on the field comparable to the statistical revolution: acquiring large-scale resources by exploiting collective intelligence. An illustration of this new approach is *Phrase Detectives*, an interactive online *game with a purpose* for creating anaphorically annotated resources that makes use of a highly distributed population of contributors with different levels of expertise.

The purpose of this article is to first of all give an overview of all aspects of *Phrase Detectives*, from the design of the game and the HLT methods we used to the results we have obtained so far. It furthermore summarizes the lessons that we have learned in developing this game which should help other researchers to design and implement similar games.

Categories and Subject Descriptors: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Natural language

General Terms: Human Factors, Design

Additional Key Words and Phrases: Web cooperation, resource creation, human language technology, games with a purpose, corpus annotation, anaphora

## ACM Reference Format:

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.* 3, 1, Article 3 (April 2013), 44 pages.  
DOI: <http://dx.doi.org/10.1145/2448116.2448119>

## 1. INTRODUCTION

Ever since the shift towards statistical methods, research in Human Language Technology (HLT) has been driven by the availability of large-scale resources (corpora, lexica and, more recently, repositories of encyclopedic knowledge). The creation of such resources has traditionally been the task of dedicated experts who did their work manually. However, we may be now witnessing another significant change: Web collaboration has started to emerge as a viable alternative for obtaining the large resources that are needed to build and evaluate HLT systems.

Examples of collective intelligence such as Wikipedia demonstrated that a surprising number of individuals are willing to help with resource creation and scientific

---

The reviewing of this article was managed by special section Associate Editor David Robertson.

The initial funding for *Phrase Detectives* came from EPSRC project AnaWiki, EP/F00575X/1.

Authors' addresses: M. Poesio, J. Chamberlain, and U. Kruschwitz, School of Computer Science and Electronic Engineering, University of Essex, UK; L. Robaldo (corresponding author), Department of Computer Science, University of Turin, Italy; email: [robaldo@di.unito.it](mailto:robaldo@di.unito.it); L. Ducceschi, University of Utrecht/Verona, The Netherlands.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or [permissions@acm.org](mailto:permissions@acm.org).

© 2013 ACM 2160-6455/2013/04-ART3 \$15.00

DOI: <http://dx.doi.org/10.1145/2448116.2448119>

experiments, and raised a lot of interest in HLT. Wikipedia is now routinely used as a word sense repository, possibly even more than WordNet [Csomai and Mihalcea 2008] or as a source of encyclopedic knowledge [Ponzetto and Strube 2007]; and *crowdsourcing* through Amazon’s Mechanical Turk<sup>1</sup> or CrowdFlower<sup>2</sup> has quickly become the method of choice for the fast annotation of small-and-not-so-small corpora, and for some types of HLT system evaluation [Snow et al. 2008; Callison-Burch 2009]. Less used, so far, is a second approach to collaborative resource construction popularized by von Ahn and colleagues [von Ahn 2006]: incentivizing users to create resources by developing a so-called *Game-With-A-Purpose* (GWAP) which will produce the required resource as a byproduct of the users’ playing. The promise of this approach is that provided that the game is entertaining enough to attract sufficient players, it should be possible to carry out the annotation out at a smaller cost than with crowdsourcing, let alone with traditional annotation methods, thus potentially enabling the annotation of much greater amounts of data—the 100M words-plus corpora that are increasingly perceived as necessary to train high-performance HLT components. The challenges are to develop such a game and to maintain a high visibility for it.

In this article we discuss *Phrase Detectives*,<sup>3</sup> one of the first GWAP for corpus collection,<sup>4</sup> and one of the very few such games to result in the annotation of a substantial amount of data. *Phrase Detectives* was developed to annotate corpora for anaphora resolution [Kamp and Reyle 1993; Garnham 2001; Mitkov 2002; Poesio et al. 2011b], the semantic task concerned with recognizing that, for example, the pronoun *it* and the definite nominal *the town* in (1) refer to the same entity as the proper name *Wivenhoe*, and to a different entity from the mentions *Colchester* or *River Colne*.

- (1) *Wivenhoe* developed as a port and until the late 19th century was effectively a port for Colchester, as large ships were unable to navigate any further up the River Colne, and had two prosperous shipyards.  
*It* became an important port for trade for Colchester and developed shipbuilding, commerce and fishing industries.  
 The period of greatest prosperity for *the town* came with the arrival of the railway in 1863.<sup>5</sup>

Anaphora resolution is a key semantic task both from a linguistic perspective and for applications ranging from summarization to text mining, but one for which medium-sized corpora have only recently become available and our understanding of which is not such that linguists can produce a coding scheme with high reliability [Poesio and Vieira 1998; Zaenen 2006]. As we will see, it is one of the contentions of this article that the collaborative approach to resource creation can also result in a better understanding of the complexity of language interpretation.

This work is meant to be the definitive reference article on *Phrase Detectives*, collecting in a single publication material previously only found in separate papers such as Poesio et al. [2008], Chamberlain et al. [2008a, 2008b, 2009a, 2009b], and Kruschwitz et al. [2009] and additional material not presented before, including a cost comparison between games, traditional annotation, and crowdsourcing, and a discussion of recent developments such as the Facebook version of the game. Our objective is to provide an assessment of the methodology and to summarize the lessons we learned so that other researchers may decide whether this methodology is appropriate for other HLT tasks. In

<sup>1</sup><http://www.mturk.com>.

<sup>2</sup><http://crowdflower.com>.

<sup>3</sup><http://www.phrasedetectives.org>.

<sup>4</sup>The only earlier effort we are aware of is Chklovsky’s *1001 Paraphrases* [Chklovski 2005].

<sup>5</sup>Taken from Wikipedia’s page about *Wivenhoe*, the village next to the University of Essex where many of the authors live.

summary, we will argue that the game has already been moderately successful both in terms of quantity and quality of data, and furthermore that the data collection is still going strong after almost three years, which suggests that the GWAP approach can definitely be used to annotate at least medium-size (1M–10M words), high-quality corpora, provided, however, that a number of issues are paid attention to and that a continuous effort is made to maintain the game visible.

The structure of the article is as follows. Section 2 motivates the collective intelligence approach to annotation and surveys the main approaches to collective intelligence (Wikipedia-style “citizen science,” crowdsourcing, and games-with-a-purpose). We then discuss in Section 3 the issues that have to be tackled by the developers of GWAP in general and GWAP for annotation in particular, such as providing adequate incentives, quality control, and data selection. The approach to these issues taken in Phrase Detectives is discussed in Section 4, and the methods used to create a corpus in Section 5. The results obtained so far are analyzed in Section 6. We then summarize the lessons learned to serve as a guide for developers of future GWAP applications applied to HLT tasks, and briefly discuss ongoing and future developments.

## 2. COLLECTIVE INTELLIGENCE AND RESOURCE CREATION

### 2.1. Annotation of Language Corpora

The tremendous success of the statistical revolution in Human Language Technology has resulted in the first HLT components and applications truly usable on a large scale. It also created, however, a need for large amounts of annotated linguistic data for training and evaluating such systems.<sup>6</sup>

The first annotated corpora, such as as the 1-million-word Brown Corpus [Kucera and Francis 1967], were only concerned with low-level linguistic information such as lemmas and part-of-speech tags, and were created entirely by hand. This methodology is still used for the majority of annotation projects, in particular for projects concerned with the annotation of more complex types of linguistic information, and arguably still has a place to create resources of very high quality but the costs involved are considerable. Thanks to substantial investments in Germany and USA, such as the funding of the SALSA [Burchardt et al. 2009] and ONTONOTES [Hovy et al. 2006; Pradhan et al. 2007] projects, it has been possible in recent years to create Brown-Corpus-size annotated corpora for semantic tasks such as coreference, predicate argument structure and word-sense disambiguation. However, the costs required (in the order of over one million dollars per million words of annotated data for each level) make it clear that the traditional hand-annotation methods used in such projects are not feasible to annotate larger amounts of data. Yet work on training parsers using the Penn Treebank has made it very clear that 1M-word corpora have serious limitations in terms of coverage. The techniques used to create OntoNotes will not be applicable to the annotation of a 100-million-word corpus.

A faster and less expensive semi-automatic methodology has therefore become standard to annotate larger amounts of linguistic information for which relatively high-quality annotation systems existed. When this is the case, a preliminary annotation with automatic methods is followed by partial hand-correction. The methodology was pioneered in the annotation of the British National Corpus (BNC), the first 100M-word linguistically annotated corpus [Burnard 2000], thanks to the availability of relatively high-quality automatic part-of-speech taggers trained on smaller-scale data (in this case, the CLAWS system developed by the University of Lancaster). With the development of the first high-quality chunkers this methodology became applicable to the case of

---

<sup>6</sup>It also created a need for large-scale lexical and encyclopedic resources, although the use and creation of such resources will not concern this article.

syntactic annotation as well, and was used for the creation of the Penn Treebank, although more substantial hand-checking was required [Marcus et al. 1993]. This semi-automatic annotation methodology, however, cannot yet be used for semantic annotation as the quality of, for instance, anaphoric resolvers on general text is not yet high enough. Furthermore, the quality of the annotation might deteriorate unless every item is hand-checked.

A variety of alternative approaches to the problem created by the need for ever-larger annotated corpora have been proposed. In recent years many very large corpora have been annotated fully automatically (for an example, see Baroni et al. [2009]), but this approach clearly is only feasible when high-performance annotators for a given type of annotation already exist, and the resources thus created are not directly useful to train annotators for that type of task (although they can be very useful, e.g., for lexicographic purposes and/or to train annotators for other types of tasks). A more radical approach has been to develop unsupervised methods for performing a given HLT task that do not depend on annotated corpora (for anaphora, see, e.g., Ng [2008]) but so far these methods have not yet achieved a level of performance comparable to even the moderate level of supervised methods trained on current corpora. Weakly supervised techniques [Mintz et al. 2009] have proven effective for tasks such as named entity resolution, word sense disambiguation, and relation extraction, in which collaboratively created resources such as Wikipedia can be used to generate the training data. However, no such resources are available for a number of core HLT tasks, including coreference, predicate argument structure, and discourse structure. In the medium term, therefore, two approaches appear to hold the greatest promise: active annotation [Vlachos 2006; Settles 2009], in which the activity of (hand) annotation is guided by the needs of the system being trained; and Web collaboration, the approach followed in the work described here.

## 2.2. Collaborative Resource Creation on the Web

The idea of collaborative resource creation on the Web is motivated by the observation that a group of individuals can contribute to a collective solution which has a better performance and is more robust than an individual's solution (this was shown for example in simulations of collective behaviors in self-organizing systems [Johnson et al. 1998]).

The willingness of Web users to collaborate in the creation of multilingual resources is clearly illustrated by Wikipedia. English Wikipedia numbers (as of October 2011) 3,773,941 articles, written by over 15.5 million collaborators and 5559 reviewers<sup>7</sup>. By contrast the current edition of *Encyclopedia Britannica*, as of 2007, had 700 “macro” articles and 70,000 “micro” articles, created by around 4,000 experts coordinated by 100 editors. Wikipedia takes full advantage of the multilingualism of the Web and includes more than 8 million articles in French, German, Italian, Polish, Spanish, Dutch, Russian, Japanese, and Portuguese. Wikipedia also illustrates the effectiveness of “bottom-up” or “self-organizing” editorial control where the reviewers are themselves volunteers who are considered by the Wikipedia community to be competent (i.e., by having an approval rate of over 75%).

Wikipedia is perhaps the best known example of collaborative resource creation but it is not an isolated case. Open Mind Common Sense<sup>8</sup> demonstrated that Web collaboration can be relied on to create an AI resource [Singh 2002]. Specifically, 14,500 volunteers have contributed nearly 700,000 sentences to Open Mind Common Sense,

<sup>7</sup>[http://meta.wikimedia.org/wiki/List\\_of\\_Wikipedias](http://meta.wikimedia.org/wiki/List_of_Wikipedias).

<sup>8</sup><http://openmind.media.mit.edu>.

which has been turned into ConceptNet.<sup>9</sup> This is now one of the main sources of conceptual knowledge currently available. The Open Mind Common Sense project also led to the development of a “quasigame” for collecting common sense knowledge, the system *LEARNER* [Chklovski and Gil 2005]. Other efforts to acquire large-scale world knowledge from Web users include Freebase<sup>10</sup> and True Knowledge<sup>11</sup>. A slightly different approach to the creation of common sense knowledge has been pursued in the Semantic MediaWiki project [Krötzsch et al. 2007], an effort to develop a “Wikipedia way to the Semantic Web”, which aims to make Wikipedia more useful and to support improved search of Web pages using semantic annotation.

Numbers of volunteers such as those involved in Open Mind Common Sense are very promising for annotation as well. If 15,000 volunteers each annotated 7 texts of 1,000 words (an effort of about 3 hours) a 100M-words annotated corpus would result. However, it took almost ten years for Open Mind Common Sense to attract that many contributors and to collect that much data. Thus there have been attempts to find more powerful incentives to Web collaboration.

### 2.3. Crowdsourcing

The simplest way to provide an incentive is to pay the collaborators. Amazon Mechanical Turk (AMT) pioneered *crowdsourcing*: using the Web as a way of reaching very large numbers of collaborators (called *workers*) who get paid, although typically very little (in the order of 1 to 10 cents per item of work). AMT and CrowdFlower demonstrated that crowdsourcing is very competitive with traditional resource creation methods from a financial perspective, because even very little payment is enough to attract large number of collaborators (many of which are students or otherwise unemployed, or live in countries in which the cost of living is lower). Studies showed that the quality of resources created this way is comparable to that of resources created in the traditional way, provided that multiple judgments are collected in a sufficient number [Snow et al. 2008; Callison-Burch 2009]. A further advantage is that workers work very fast; it is not uncommon for a task (Human Intelligence Task or HIT) to be completed in minutes. These considerations resulted in crowdsourcing becoming a standard way of creating small-scale resources for HLT. But even using AMT becomes prohibitively expensive to create resources of the size that we have been discussing, that is, in the order of 100 million annotated words. Consider for example the case of anaphora resolution, where the items to annotate (the segments of text which mention entities, which we will refer to as *markables* using standard annotation terminology) are Noun Phrases (NPs) (see Section 4.1). On average, there is one NP every three words,<sup>12</sup> thus a 100-million-words corpus would contain around 30 million NPs. Annotating a corpus of this size by hand at the costs at which OntoNotes was created would cost 100 million dollars or more. Creating it with AMT would still cost in the order of 7.5 million dollars even if workers were only paid .05 US \$ per judgment, and only 5 workers were asked to annotate every markable.<sup>13</sup>

<sup>9</sup><http://conceptnet.media.mit.edu>.

<sup>10</sup><http://www.freebase.com>.

<sup>11</sup><http://www.trueknowledge.com>.

<sup>12</sup>For example the collection of texts currently being annotated in Phrase Detectives is 1.2 million words and contains 392,120 markables.

<sup>13</sup>Neither of these assumptions are realistic for anaphoric annotation; see Section 6.5 for a more realistic comparison.

## 2.4. Games-with-a-Purpose

Wikipedia, Open Mind Common Sense, and similar initiatives rely on people's altruism and interest in science. Luis von Ahn from Carnegie Mellon University, Timothy Chklovsky from the Open Mind Common Sense group, and others argue that the desire to be entertained is a much more powerful incentive. It is estimated that every year over 9 billion person-hours are spent by people playing games on the Web [von Ahn 2006]. If even a fraction of this effort could be redirected towards resource creation via the development of Web games that achieve resource creation as a side-effect of having people play entertaining games (von Ahn called such games *Games-With-A-Purpose* or *GWAP*) we would have enormous quantity of man-hours at our disposal.

von Ahn demonstrated his point through the development of several *GWAP*. The best known of these games is the ESP Game.<sup>14</sup> In the ESP Game two randomly chosen players are shown the same image. Their goal is to guess how their partner will describe the image (hence the reference to extrasensory perception or *ESP*) and type that description under strict time constraints. If any of the strings typed by one player matches the strings typed by the other player, they score both points. From the players' perspective that is all that matters. The descriptions of the images players provide are very useful information to train content-based image retrieval tools [von Ahn and Dabbish 2004]. von Ahn's intuition that the game would attract very large numbers of Web visitors proved correct. The game attracted 13,000 players between August and December 2003 and has attracted over 200,000 players since, who have produced over 50 million labels. The quality of the labels has also been shown to be as good as that produced through conventional image annotation methods. A crucial advantage of *GWAP* over crowdsourcing is that, once the game has been developed and made available, it can continue to generate annotations with very little maintenance and very little cost. Indeed, the game was so successful that a license to use it was bought by Google, which developed it into the Google Image Labeler which was online from 2006 to 2011. The story of the Google Image Labeler<sup>15</sup> illustrates many useful points about what is required to make a *GWAP* successful: from the need to provide incentives to players, to that of continuously revising the game's methods for controlling malicious behavior to stay one step ahead of the malicious players. We discuss these requirements in Section 3.

Many other *GWAP* have been developed by von Ahn and other labs to collect data for multimedia tagging (*OntoTube*,<sup>16</sup> *Tag a Tune*<sup>17</sup>) and for acquiring common sense knowledge (*Verbosity*,<sup>18</sup> *OntoGame*,<sup>19</sup> *Categorilla*<sup>20</sup>, *Free Association*<sup>21</sup>).<sup>22</sup> The *GWAP* concept has now also been adopted by the semantic Web community in an attempt to collect large-scale ontological knowledge because currently "the semantic Web lacks sufficient user involvement almost everywhere" [Siorpaes and Hepp 2008]. A number of *GWAP* have also been developed in other areas of computer science to support research in the biological sciences. The most famous of these games (and one of the most successful *GWAP* overall) is *Foldit*<sup>23</sup>, a *GWAP* about protein folding developed at the

<sup>14</sup><http://www.gwap.com/gwap/gamesPreview/espgame>.

<sup>15</sup>[http://en.wikipedia.org/wiki/Google\\_Image\\_Labeler](http://en.wikipedia.org/wiki/Google_Image_Labeler).

<sup>16</sup><http://ontogame.sti2.at/games>.

<sup>17</sup><http://www.gwap.com/gwap/gamesPreview/tagatune>.

<sup>18</sup><http://www.gwap.com/gwap/gamesPreview/verbosity>.

<sup>19</sup><http://ontogame.sti2.at/games>.

<sup>20</sup><http://www.doloreslabs.com/stanfordwordgame/categorilla.html>.

<sup>21</sup><http://www.doloreslabs.com/stanfordwordgame/freeAssociation.html>.

<sup>22</sup>The current *GWAP* from von Ahn's lab are playable from <http://www.gwap.com>.

<sup>23</sup><http://fold.it/portal>.

University of Washington. Other GWAP with a biomedical application include *Phylo*<sup>24</sup> and *EteRNA*.<sup>25</sup>

To our knowledge, however, prior to Phrase Detectives there had been only one GWAP aiming to exploit the effort of Web volunteers to annotate corpora, *1001 Paraphrases* [Chklovski 2005]. Other corpus annotation games have appeared after Phrase Detectives, the most successful being the GIVE family of games [Koller et al. 2010]. We discuss these other proposals in Section 7 on Related Work. We will, however, mention here another GWAP whose objective is not corpus annotation but transcription; transcription of ancient scrolls in fact: the University of Oxford's *Ancient Lives*.<sup>26</sup>

### 3. DESIGNING GAMES WITH A PURPOSE

Designing a game to annotate data is not easy. First of all, like for all other online games, a way has to be found to attract players, and then motivate them to keep playing either by making labeling items fun or by stimulating their competitive spirit. We discuss this requirement in Section 3.1.

A second requirement for a GWAP is that the interface should be easy to use and the task presented in a way that is simple to understand. A game deployed on the Web should observe all the normal guidelines regarding browser compatibility, download times, consistency of performance, spatial distance between click points, etc.<sup>27</sup> Game interfaces should be graphically rich, although not at the expense of usability, and aimed at engaging the target audience (i.e., a game aimed at children may include more cartoon or stylized imagery in brighter colors than a game aimed at adults). The game should also provide a consistent metaphor within the gaming environment.

Finally, there is one additional requirement for GWAP in comparison with normal games: quality control. Care has to be taken that malicious users or users who simply do not care or do not understand the underlying rules of the game do not end up making the collected data unusable. We discuss this requirement in Section 3.2.

We conclude this section by discussing von Ahn's proposals concerning how to address these issues in GWAP design [von Ahn and Dabbish 2008]. We discuss how these issues were addressed in Phrase Detectives in Section 4.

#### 3.1. Incentives

Three types of incentives can be used to encourage participation in a Web collaboration activity. An individual may be incentivized at a *personal* level—whether by making the activity entertaining, as in a game, or by giving him/her the chance to highlight his/her expertise, as in Wikipedia. Wikipedia also relies on a *social* incentive: giving individuals the sense that they are collaborating in a worthwhile enterprise and improve their standing in an online community. (More in general, a social incentive is one that strengthens individuals' membership in a community.) Finally, the incentive may be *financial*: the player is motivated by personal gain. One of the promising features of the GWAP method is that a well-designed game can simultaneously entertain, provide a sense of participating to a worthwhile enterprise, improve a player's standing among her/his peers, and—through prizes—offer a financial incentive, although this type of incentive should be applied with caution as rewards have been known to decrease annotation quality [Mrozinski et al. 2008]. We discuss each type of incentive in turn.

<sup>24</sup><http://phylo.cs.mcgill.ca>.

<sup>25</sup><http://eterna.cmu.edu>.

<sup>26</sup><http://ancientlives.org>.

<sup>27</sup><http://www.usability.gov/guidelines>.

*Personal incentives.* Simply participating in a fun online activity can be enough reward for some individuals. GWAP should therefore be fun, a point often overlooked.

There is substantial literature on what makes games fun [Koster 2005]. One of the simplest mechanisms is *scoring*: by getting a score the player gains a sense of achievement. A second common method to entertain players is to have them experience a *progression through the game*, whether by learning new types of tasks, becoming more proficient at current tasks, or gaining recognition for their effort (see the following).

A common form of progression is by assigning the player a named level, starting from novice and going up to expert [Koster 2005; von Ahn et al. 2006]. The level mechanism also provides one form of quality control, as we will see shortly. von Ahn developed a theory of entertainment in GWAP that we will discuss in Section 3.4.

But entertainment is not the only personal incentive GWAP can offer. The desire of Web users to contribute information to Wikipedia can also be considered as motivated by personal reasons such as the desire to make a particular page accurate, or the pride in one's knowledge in a certain subject matter. As we show in Section 4, GWAP such as Phrase Detectives can offer this type of personal incentive as well.

As we will see in the case of Phrase Detectives, GWAP may attract a considerable number of Web collaborators by giving players the sense that they are contributing to creating a resource from which a whole field may benefit (e.g., other computational linguists). This is indeed the key motivation for volunteers contributing to Wikipedia [Yang and Lai 2010].

*Social incentives.* A different sort of social incentive is provided by the scoring mechanism. *Public leaderboards* reward players by improving their standing amongst their peers (in this case their fellow players). Using leaderboards and assigning levels for points has been proven an effective motivator, with players often using these as targets [von Ahn and Dabbish 2008]. An interesting phenomenon has been reported with these reward mechanisms, namely that players gravitate towards the cutoff points (i.e., they keep playing to reach a level or high score before stopping) [von Ahn et al. 2006].

Both types of social incentives can be made even more effective when the game is embedded in a social networking platform like Facebook. In such a setting, the players motivated by the desire to contribute to a communal effort may share their efforts with their friends, whereas those motivated by a competitive spirit can compete against them. This was one of the motivations behind the Facebook version of Phrase Detectives, briefly discussed in Section 8.

*Financial incentives.* One of the most effective ways to incentivize players is to pay them money, or to induce them to believe that much money can be gained through betting, as in online gaming. However, apart from ethical considerations, offering substantial direct payments would make Web collaboration lose its cost effectiveness as a way for generating resources. The success of crowdsourcing demonstrates that it is possible to attract sufficient numbers of collaborators at relatively little cost and a similarly low-cost reward structure can be built into online games as well through the mechanism of prizes.

### 3.2. Quality Control

Whereas the designers of normal online games only need to worry about providing the right incentives to their players, the designers of GWAP also need to worry about obtaining good-quality data. Obtaining reliable results from non-experts is also a challenge for users of crowdsourcing, and in this context strategies for dealing with the issue have been discussed extensively [Kazai et al. 2009; Alonso and Mizzaro 2009; Alonso et al. 2008; Feng et al. 2009]. In the case of crowdsourcing the main strategy for achieving a good quality of labeling is to aggregate results from many users to approximate a single



expert's judgments [Snow et al. 2008]. Our work, however, is not motivated solely by the desire to label large amounts of data; as discussed in Section 4, we believe that Web collaboration could, in fact should, also be used to gather data about the interpretation of natural language expressions, which all too often is taken to be completely determined by context, often without much evidence [Poesio et al. 2006]. From this perspective it is important to attempt to avoid poor-quality individual judgments.

Controlling cheating may be one of the most important factors in game design. If a player is motivated to work hard and score points, he/she may become more motivated to find a way to cheat the system. But there are many ways to prevent or control cheating, including IP address checking, user profiling, validating submissions against known answers, and preventing resubmission of answers, as we will see in Section 4.4.

### 3.3. Attracting and Retaining Players

In order to attract the number of participants required to make a success of this methodology it is not enough to develop attractive games; it is also necessary to develop effective forms of *advertising*. The number of online games competing for attention is huge and without some effort to rise a game's profile, it will just never catch the attention of enough players. The importance of this strategy was demonstrated by von Ahn's lab. The ESP Game was constantly advertised in the press and also on tv. Other methods to reach players included blogs and being linked on gaming sites. As we will see discussing Phrase Detectives, not all advertising methods are equally successful and it is important to evaluate which works best for the game task, language, or country.

Retaining these players, once acquired, is not easy either: one of the biggest problems for any online game is *volunteer attrition*, where a player's contribution diminishes over time [Lieberman and Teeters 2007]. The level mechanism may help in this respect; other methods to reduce attrition include providing additional feedback on the player's efforts, besides scoring, and allowing the player to comment on the gaming conditions (perhaps to identify an error in the game, to skip a task, or to generate a new set of tasks). Through commenting the player feels more in control of the game (while providing essential quality control).

### 3.4. von Ahn's Theory of GWAP Design

von Ahn and Dabbish extracted out of the experience of designing a good number of games several proposals concerning the issues discussed in the earlier parts of this section [von Ahn and Dabbish 2008].

They focus on one type of incentive: *enjoyment*. The main mechanism exploited by von Ahn and colleagues to make players enjoy their GWAP is providing them a challenge. This is achieved through mechanisms such as requiring a timed response, keeping scores that ensure competition with other players, and having players of roughly similar skill levels play against each other.

With regards to quality control, von Ahn and Dabbish discuss two main concerns: ensuring correctness and variety of labels, and avoiding collusions between players. With respect to the first matter they propose two main mechanisms: player testing (assessing the quality of a player's output by occasionally matching it against already annotated data) and repetition, that is, redundancy (ensuring that each item is multiply labeled). Variety is achieved primarily through the mechanism of taboo words introduced with the ESP Game: once a particular label has been produced by a number of players, it becomes "taboo" and subsequent players are not allowed to use it. In order to avoid collusions, von Ahn and colleagues developed a number of methods to make sure that players do not play against themselves and do not know each other's identity. These include running IP address checks, and introducing random delays to

the moment each player starts so as to make it more difficult to synchronize starting points.

von Ahn and Dabbish also introduced a number of mechanisms for evaluating a game. Two main variables were proposed: *throughput* (the speed at which a particular player is labeling) and *average lifetime play* (a measure of enjoyability).

In the next section we discuss the solutions we adopted in Phrase Detectives, some of which were inspired by von Ahn's proposals, whereas others were novel.

#### 4. A GAME-WITH-A-PURPOSE FOR ANNOTATION: PHRASE DETECTIVES

Phrase Detectives is a single-player game-with-a-purpose developed to collect data about anaphora (this HLT task is briefly discussed in Section 4.1) and centered around the detective metaphor. The game architecture is articulated around a number of tasks and uses scoring, progression, and a variety of other mechanisms to make the activity enjoyable (Section 4.2). A mixture of incentives, from the personal (scoring, levels) to the social (competing for some players, participating in a worthwhile enterprise for others) to the financial (small prizes) are employed (Section 4.3). The GWAP approach to resource annotation was adopted not just to annotate large amounts of text, but also to collect a large number of judgments about each linguistic expression. This led to the deployment of a variety of mechanisms for quality control which try to reduce the amount of unusable data beyond those created by malicious users, from the level mechanism itself to validation to a number of tools for analyzing the behavior of players (Section 4.4). Last but not least, making a GWAP into a success requires a great deal of promotional activities to ensure the game achieves visibility; these matters are discussed in Section 4.6.

##### 4.1. Anaphora

Anaphora is the linguistic mechanism of referring back to an entity already introduced in a discourse, for example, Wivenhoe in (1), sometimes using the same expression again (as in the case of the two references to Colchester in the same example) but in many other cases using different expressions (as in, e.g., the two other references to Wivenhoe in the example using *it* and *the town*). Interpreting anaphoric reference therefore involves, first of all, keeping track of which entities have been mentioned (in linguistics this is called building a *discourse model* [Kamp and Reyle 1993]). Then, whenever a new linguistic expression of interest<sup>28</sup> is encountered—such expressions are usually called markables in an annotation context—the reader or system has to decide whether this markable introduces a new entity (in which case it is called *discourse new* [Prince 1992]) or whether instead it refers to an entity already introduced (this entity is called the *antecedent*; the term *discourse old* is used to indicate expressions which refer to a previously introduced antecedent)—and if so, which one. For example, in the second utterance in (1), pronoun *it* could refer to Wivenhoe, Colchester, or indeed the River Colne, whereas in the third utterance, the markable *the town* could be interpreted as having either Wivenhoe or Colchester as antecedent.

The problem of interpreting such markables is further complicated by the fact that not all nominal phrases in English (NPs) are *referential*, that is, either introduce a new entity or refer to one already introduced. First of all, expressions like *it*, that in texts like the one under discussion can be discourse old, in other contexts may have no semantic content at all: For example, in (2a), *It* is only used for syntactic reasons and is semantically empty. Second, many nominal phrases are used to express properties of

<sup>28</sup>In Phrase Detectives we focus on so-called *nominal anaphora*, that is, anaphoric relations involving nominal expressions. The linguistic expressions of interest are therefore Noun Phrases (NPs). Note that other types of linguistic expressions can be anaphoric, most notably verbal ellipses as in *John fell. Bob did too*.

entities, as opposed to referring to entities directly. Thus for instance the NP *a fireman* in (2b) is used to express a property of the entity referred to by the subject of the sentence, *Sam*. Third, certain nominal phrases, like *no town* in (2c), cannot be said to introduce or refer to any entity in particular; instead, they denote *quantifiers*, that is, relations between predicates; roughly speaking, (2c) asserts that the intersection of the denotations of the sets “towns in England” and “towns older than Colchester” is empty.

- (2) a. It is raining.  
 b. Sam is a fireman.  
 c. No town in England is older than Colchester.

Neither deciding the logical form content of a noun phrase (referring, empty, property) nor choosing an antecedent between the entities already introduced in discourse are easy tasks, and in many cases the text does not provide enough information to decide. Consider, for instance, the passage (3a) from *Alice in Wonderland*, one of the texts in the Gutenberg subset of the Phrase Detectives corpus. The four instances of *it* in the passage (underlined> are all ambiguous between being semantically vacuous and having a so-called *discourse deictic* reading, that is, referring to a proposition: for example *when she thought it over afterwards* could either simply mean that Alice was thinking about what happened (semantically vacuous interpretation), or that she was thinking about a specific episode, namely, the fact that the Rabbit was saying something to itself (discourse deictic interpretation).

- (3) a. There was nothing so VERY remarkable in that; nor did Alice think *it* so VERY much out of the way to hear the Rabbit say to itself, ‘Oh dear! Oh dear! I shall be late!’ (when she thought *it* over afterwards, *it* occurred to her that she ought to have wondered at this, but at the time *it* all seemed quite natural); . . .

- b.  
 3.1 M: can we .. kindly hook up  
 3.2 : uh  
 3.3 : engine E2 to the boxcar at .. Elmira  
 4.1 S: ok  
 5.1 M: +and+ send it to Corning  
 5.2 : as soon as possible please  
 6.1 S: okay

The identification of the antecedent of an anaphoric expression, as well, may also be problematic. Consider the instance of *it* in utterance 5.1 in (3b). In experiments reported in Poesio et al. [2006] subjects were asked about the interpretation of this and similar pronouns. About two-thirds of the subjects chose engine E2, whereas the other third chose the boxcar at Elmira.

These difficulties in interpretation suggest the need to collect multiple judgments for each expression—a task very well suited for Web collaboration of any type—and that in cases of disagreement it may be best to preserve such judgments rather than attempting to make a choice between them.

## 4.2. The Game

A key decision in the design of Phrase Detectives was to follow the approach to data collection adopted by Chklovsky in *LEARNER* [Chklovski and Gil 2005]—namely, to have the Web collaborators perform both the task of providing the judgments (which we will call the *annotation* step) and the task of checking those judgments (that we will call *validation*); as we will see, the inclusion of the latter step plays a crucial role in our strategy for quality control. In Phrase Detectives the player is a detective that goes about resolving cases—expressing judgments about the interpretation of markables—in the so-called Name-the-Culprit activity, and providing opinions about

The screenshot shows the homepage of the game 'Phrase Detectives' for a player named 'jon'. The page is designed with a detective theme, featuring a blue and orange color scheme and cartoon detective characters. The central area is titled 'WELCOME BACK TO HEADQUARTERS' and includes a 'Start >>' button. To the left, the 'USERPROFILE' section shows 'jon' has 60 cases this week, 4 decisions, 31 agreements, and 25 extras. It also displays a 'Cunning Pirate' level and a 96% rating. The 'CASE OPEN' section indicates 89 tasks remaining. Below this are links for 'EDIT PROFILE | LOGOUT' and a 'Like' button. The 'Instructions' and 'FAQ' sections are also visible. The right side of the page features a 'TOPSCORES' section with 'THIS WEEK' and 'THIS MONTH' winners, a 'LEADERBOARD' table, and a 'MOST RECENT' section. The 'LEADERBOARD' table lists players and their scores for the week, month, and all-time. The 'MOST RECENT' section shows a recent submission by 'jon'. At the bottom, there is a 'Feedback' button and a footer with version information and copyright details.

WEEK	MONTH	ALLTIME
JRS	JRS	JRS
1100	1100	1100
stewart miller	1086	
MDKorpel	806	
papillon	610	
thomwd	543	
IveBrii_RijkvanBraak	504	
Folkert_Patrick_KI	402	
StefanenHein	390	
myrmidon	388	
MAJ	367	
VB	277	
s2011840	252	
AntonMulder	237	
michelleburghardt	233	
gully	229	
julie3164	194	
RB_NV_KI	180	
MarcoBosman	178	
MWithagen	166	
rikanna	155	

Fig. 1. Screenshot of the Phrase Detectives player homepage.

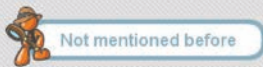
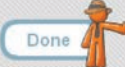
other detectives's judgments in the Detectives Conference activity. Both of these activities lead to point accumulation, which is the main objective of the players; in fact, as we will see shortly, validation (Detectives Conference) is the main scoring activity for players once they pass the training threshold. The graphical design of Phrase Detectives, centered around the detective metaphor, is exemplified in Figure 1.


*Name-the-Culprit.* Name-the-Culprit is the primary activity dedicated to the labeling of data by players. The players are shown a window of text in which a markable is highlighted in orange, as shown in Figure 2 (top).<sup>29</sup> They have to decide, first of all, whether the markable is referring, a property, or nonreferring. In case they decide the markable is referring, they then have to decide whether it introduces a new entity (i.e., whether it is discourse new), or whether it refers to an already mentioned entity, and in this case they have to locate the closest mention. Moving the cursor over the text reveals the markables within a bordered box; to select a markable, the player clicks on the bordered box and the markable becomes highlighted in blue. This process can be





<sup>29</sup>These markables are automatically extracted from the text using the pipeline(s) discussed in Section 5.

### Fairy Tales - Clever Elsie (The Brothers Grimm)

Then the woman said to the servant: 'Just go down into the cellar and see where Elsie is.' The maid went and found her sitting in front of the barrel, screaming loudly. 'Elsie why do you weep?' asked the maid. 'Ah,' she answered, 'have I not reason to weep? If I get Hans, and we have a child, and he grows big, and has to draw beer here, the pick-axe will perhaps fall on his head, and kill him.' Then said the maid: 'What a clever Elsie we have!' and sat down beside her and began loudly to weep over the misfortune. After a while, as the maid did not come back, and those upstairs were thirsty for the beer, the man said to the boy: 'Just go down into the cellar and see where Elsie and the girl are.' The boy went down, and there sat Clever Elsie and the girl both weeping together. Then he asked: 'Why are you weeping?' 'Ah,' said Elsie, 'have I not reason to weep? If I get Hans, and we have a child, and he grows big, and has to draw beer here, the pick-axe will fall on his head and kill him.' Then said the boy: 'What a clever Elsie we have!' and sat down by **her**, and likewise began to howl loudly.

 **Not mentioned before**  **Done**

 Comment on this phrase

-  Skip this one
-  Skip - closest phrase can't be selected
-  Skip - closest phrase is no longer visible
-  Skip - error in the text

### Shanghai Fugu Agreement (Wikipedia)

The 1985 Hesse coalition under Prime Minister Holger Börner was to be based on an official policy agreement negotiated by both parties.

During a final night session of the negotiations the Greens tabled a demand that Hesse join the "Shanghai Fugu Agreement". This was accepted by their tired Social Democratic counterparts and became official state policy.

The Greens argued that the fugu fish is well known to be a dangerous delicacy requiring specialized chefs who mostly come from Asia. Due to expanding restrictions on work permits restaurants have found it difficult to employ such specialists. The "Shanghai Fugu Agreement" provides special regulations for certified fugu chefs internationally.

(It should be noted that chefs in Japan require certification to handle the fish.)

The agreement was absolutely fictional but was neither discovered to be **a joke** by the Social Democrats during the nightly negotiations nor later by civil servants or the press who went through the coalition contracts. It took years before **the Agreement** was revealed to be a joke.

The phrase in blue is the **closest** phrase that refers to the phrase in orange.



 **Disagree**  **Agree**

Fig. 2. Screenshots of annotation mode (top) and validation mode (bottom).

repeated if there is more than one antecedent (e.g., for plural anaphors such as “they”). When the player has made his/her selection the annotation is submitted by clicking the Done! button.

The main issues to be considered while designing an activity of this type are the selection and presentation of the markable to annotate and the candidate antecedents. Name-the-Culprit is organized around *cases*: blocks of text in which a certain number of markables have been identified as *tasks*—items that the game has to get the player’s interpretation of. The tasks in a case are then presented for annotation to the player in order of appearance in the text. Players can set their profile to play in “all markables”

more, and the little girl sprang out, crying: 'Ah, how frightened I have been! How dark it was inside the wolf'; and after that the aged grandmother came out alive also, but scarcely able to breathe. Red-Cap, however, quickly fetched great stones with which they filled the wolf's belly, and when he awoke, he wanted to run away, but he collapsed at once, and fell dead.

Then all three were delighted. The huntsman drew off it; the grandmother ate the cake and drank the wine which Red-Cap had brought, and revived, but Red-Cap thought to herself: 'As long as I live, I will never by myself leave the path, to run into the wood, when my mother has forbidden me to do so.'

### HIDDEN PHRASES

... they ...

... the wolf's belly ...

... the wolf's ...

Fig. 3. The pop-up menu allowing selection of embedded markables.

mode in which every markable in a case is considered a task, but it was felt that most players would prefer to start detecting with a new text after a little while, so in general only a subset of the markables in a document segment become tasks, and a player can decide to initiate a new case at any time.<sup>30</sup> It is worth noting that our choice of an algorithm for generating new cases that aims at maximum variety (i.e., making sure that players rarely see twice the same text) rather than completion rate (i.e., maximizing the rate at which documents are completed) was one of the most consequential aspects of the design of Phrase Detectives from the point of view of resource creation, as discussed in Section 5.

The choice among candidate antecedents is carried out with respect to a *context window*—the portion of previous text displayed to the player. The presentation of this context was among the aspects of the game that required the most thought, as specifying the anaphoric interpretation of markables crucially depends on being able to point to the last mention of an entity in a context, yet players cannot be presented with too much context; in fact, in this version of Phrase Detectives, and the Facebook version discussed next, our goal was to avoid scrolling. To achieve this, we relied on results about the distance between entity mentions such as those in Vieira and Poesio [2000], which suggest that even for anaphoric expressions that can be used to refer to entities not mentioned in the current or previous sentence, such as definite descriptions (refer to *the town* in (1)), in the great majority of cases the distance between mentions is four sentences or less.<sup>31</sup> We chose therefore a context window of at least 1000 characters, rounded up to the nearest sentence, and at most four sentences, so as to fit comfortably within a single browser page at a standard 1024 × 768 resolution. The context ends with the sentence that contains the highlighted markable, that is, markables after the highlighted markable cannot be selected at present as at present we do not collect data regarding cataphors. (Some of these parameters, from the size of the context window to allowing for cataphors, can be reconfigured.) The context is recorded with every annotation.

A particularly tricky issue with respect to selection is *embedded* markables: markables that are syntactically embedded in other markables. Consider the example in Figure 3. The second reference to *the wolf* is embedded in the NP *great stones with which they filled the wolf's belly*. In order to select the most recent mention when specifying an interpretation for the subsequent mention of the wolf (the pronoun *he*

<sup>30</sup>The number of tasks within a case is one of the parameters of Phrase Detectives; current default is 50.

<sup>31</sup>For pronouns, it has long been known that between 90 and 95% of pronouns are used to refer to an entity last mentioned in the same or the previous sentence [Hobbs 1978; Hitzeman and Poesio 1998].

in the phrase *when he awoke*), the player needs to select this embedded NP. Pop-up menus were used to allow markable selection in cases like these. When a player hovers over the segment of screen in which the second reference to the wolf appears, a menu with several options will be presented, one for each markable that overlaps with that particular part of text. (See Figure 3.)

Each markable in a case is presented to several players in annotation mode (currently it is presented 8 times; this parameter can be configured). If every player chooses the same interpretation (for example, they all say the entity is discourse new, i.e., it has not been mentioned before) then that markable is classified as *complete*. Else, it is entered among the markables to be validated through the Detectives Conference activity, discussed next.

Given that players are only allowed to choose between a limited range of options (e.g., they are not allowed to mark bridging interpretations, or discourse deixis), and given also that there are restrictions on the context window, we quickly found that it is important that players are allowed to submit a *comment* about markables. We identified a number of standard problems that the players can choose from, and also allow the players to enter text. The standard comments include:

- text preprocessing was incorrect (e.g., the pipeline missed a markable, or assigned incorrect boundaries to a markable);
- the antecedent has been mentioned earlier in the text, but the latest markable is no longer visible (this can unfortunately happen given the limits on amount of text shown to players);
- the desired antecedent cannot be selected for some other reason;
- the markable has a discourse deictic interpretation (which cannot be specified with the current version of the game);
- the markable is ambiguous, a bridging reference, or a quantifier.

The commenting feature has been remarkably successful, the only problem being that the number of comments we receive resulted in a large backlog. Responding to comments is another aspect that we hope to turn into an activity for players.

In general, we feel that even with these limitations the implementation of the annotation task developed in Name-the-Culprit is general enough to be suitable for other types of language tasks that require either a section of text to be annotated or several sections of text to be linked together with a relationship.

*Detectives Conference.* Every markable for which multiple interpretations have been proposed (the great majority, as discussed in Section 6) must go through the validation process, *validation mode*, otherwise known as the Detectives Conference activity, displayed in the lowest screenshot in Figure 2. In Detectives Conference players have to say whether they agree or disagree with an interpretation entered in annotation mode. Both the candidate markable and the candidate antecedent markables are highlighted, in orange and blue, respectively. If the player disagrees with the proposed interpretation for the markable he/she enters annotation mode for that markable in order to specify an alternative interpretation. If the interpretation he/she specifies has not been entered before this will also be entered into the validation mode. Apart from making the game more interesting, it was assumed that validating annotations would be faster than creating annotations [Chklovski and Gil 2005]. This, however, proved not to be the case, with players taking almost twice as long to complete a validation task (although this does depend on the type of interpretation the player is validating).

*Scoring.* Scoring points is one of the most important incentives in Phrase Detectives. Through scores, players gain a sense of progress and achievement and compete with other players. Scoring also plays a key role in player training, and to motivate the

players to think carefully about their decision. Just as in the ESP Game and other GWAP, this is achieved by rewarding judgments that other players will agree with.

During training, the main function of scoring is to teach players about anaphora by comparing their judgments with those in a *gold standard* (previously annotated text). This goal can be achieved simply by having players score points by assigning to a given markable the same interpretation that can be found in the gold standard.

When players go past the training level, the way their points are counted in Phrase Detectives changes; the goal now is to motivate them to think carefully about what they do. In order to do this, the scoring mechanism was designed so that players can get more points when other players agree with them than they would by randomly choosing interpretations.

In annotation mode, players past training do get one point every time they produce a judgment, to encourage them to engage in this activity. In addition, however, players producing a judgment in annotation mode get an extra point for that judgment every time another player agrees with it in validation mode. If only one interpretation for a markable is chosen by all players being presented that particular markable in annotation mode, then all of these players get awarded an extra “agreement” point but that interpretation is not presented for validation, as discussed earlier.

Players in validation mode who agree with an interpretation get one point for every player who entered that interpretation in annotation mode. If they disagree with it, they get one point for every player who entered another interpretation while in annotation mode. (Note that these players will not get a point, however.) They are also asked to propose an alternative interpretation for that markable and if this interpretation is new it will go through validation. Only the initial annotating players gain points from agreement; further players gain their points from validation.

This scoring system is also designed to provide an incentive for players to return and inspect the scoreboard as they may gain points retrospectively. After scoring a certain number of points the player is promoted to the next level. Lower levels require fewer points to achieve in order to encourage new players to keep playing, but progressing to a higher level gets increasingly harder.

*Timing.* As discussed in Section 3.4, von Ahn and his colleagues view *timing constraints* as a key aspect of what makes games exciting [von Ahn and Dabbish 2008], and built them in all their games. This is true for games in general, where timing is usually considered essential both for excitement and for quality control. In Phrase Detectives, however, there are no timing constraints, although the time taken to perform a task is used to assess the quality of that particular annotation. There are two reasons for this.

First of all, it was considered important to allow players to read documents at a relatively normal speed while having the time to complete tasks. So the markables on which we are asking a player’s judgment are presented in the order in which they appear in the document (although by default not all markables are presented) and in a limited number (currently 50 markables are displayed from each document).<sup>32</sup>

But crucially, the decision was also based on the results of the first usability study of Phrase Detectives, discussed in Section 4.6. In the game prototype used in that study, players could see how long it had taken to do an annotation. But in contrast with suggestions that timing provides an incentive, the subjects complained that they felt under pressure and that they did not have enough time to check their answers, even though the time had no influence on the scoring. As a result, in all following versions

---

<sup>32</sup>Players are given bonus points if they change their profile settings to select every markable in each document, which makes reading slower, but only 5% of players chose to sacrifice readability for the extra points.



of Phrase Detectives the time it takes players to perform a task is recorded but not shown.

Given this, it may seem surprising that the throughput of Phrase Detectives is 450 annotations per human hour, that is, much higher than the throughput of 233 labels per human hour reported for the ESP Game in von Ahn and Dabbish [2008]. There is, however, a crucial difference between the two games: Phrase Detectives only requires clicks on preselected markables, whereas in the ESP Game the user is required to type in the labels. Designers of GWAP planning to make the task timed should therefore carefully consider the speed at which the player can process the input source (e.g., text, images) and deliver his/her response (e.g., a click, typing) in order to maximize throughput and hence the amount of data that is collected without making the game unplayable.

*Other aspects of the Phrase Detectives experience.* The realization of the detective metaphor in Phrase Detectives's graphical design is achieved in part through graphical devices (e.g., the buttons are stylized with a cartoon detective character), in part through the text on the pages, written as if the player was a detective solving cases (see Figure 1). The game task is integrated in such a way that task completion, scoring, and storyline form a seamless experience.

The detective metaphor is also reflected in the level system used in Phrase Detectives to foster the experience of progression through the game. Players begin at the rookie level and then achieve progressively higher detective-related levels.<sup>33</sup>

*Multilingualism.* From the very beginning it was intended that Phrase Detectives should support annotation in multiple languages, and users were able to choose in their profile the language of the texts they would see. The first version of Phrase Detectives only included English texts, but starting in 2009 work was begun to include documents in Italian as well by developing a second preprocessing pipeline, in collaboration with the Universities of Torino and of Utrecht. Italian documents were first made available to players in the summer of 2010. Both preprocessing pipelines are discussed in Section 5.

*Implementation.* The Phrase Detectives game was built primarily in PHP, HTML, CSS, and JavaScript. The overall design was created to conform to Internet usability, accessibility, and compatibility standards. The design incorporates licensed graphics from iStockphoto<sup>34</sup> and other sources with permission.<sup>35</sup>

In the initial plans, two types of Web collaboration would be supported: through a GWAP for casual users, and through an online annotation system developed by the University of Bielefeld called Serengeti [Stührenberg et al. 2007]. Both types of data would be stored in a single database. As a result, the Phrase Detectives data are stored in a MySQL database whose design is based on the Serengeti database, and new additions to the corpus are entered through the Serengeti interface, based on the SGF markup language [Stührenberg and Goecke 2008], discussed in more detail in Section 5 [Poesio et al. 2011a].<sup>36</sup>

### 4.3. Incentivizing and Retaining Players

It was our aim to ensure that Phrase Detectives would provide all sorts of incentives for players to play discussed in Section 3.1, so as to attract all types of players.

<sup>33</sup>This is valid especially at the lower levels of progression; the names of the higher levels become increasingly more inventive, the main point being to stimulate the curiosity of the players.

<sup>34</sup><http://www.istockphoto.com>.

<sup>35</sup><http://www.pixeljoint.com/p/3794.htm>, <http://p.yusukekamiyamane.com>.

<sup>36</sup>In practice the Serengeti interface has not been used, primarily for lack of advertising.

*Competing with other players.* Phrase Detectives features the incentives usually found in online games for players motivated by a competitive spirit, such as weekly, monthly, and all-time leaderboards, cups for monthly top scores, and named levels for reaching a certain number of points.

In addition to leaderboards visible to all players, each player can also see a leaderboard of the players who agreed with them the most. Although this leaderboard provides no direct incentive (as one cannot influence one's own agreement leaderboard) this feature reinforces the social aspect of the scoring system. The success of games integrated into social networking sites like Sentiment Quiz<sup>37</sup> on Facebook indicates that visible social interaction within a game environment motivates the players to contribute more [Rafelsberger and Scharl 2009]. Indeed, this success was one of the motivations for developing the Facebook version of Phrase Detectives [Chamberlain et al. 2012], briefly discussed in Section 6.1.

*Collaborating within a community.* As discussed in Section 3.1 an important incentive for players of GWAP is the opportunity to participate in a project producing something of relevance to a (scientific) community. This type of incentive did play a role in attracting players to Phrase Detectives and retaining them: many of the players of the game are computational linguists who heard about the game through presentations and lectures, or thanks to the mention of Phrase Detectives in computational linguistics blogs with a substantial following such as Mark Liberman's<sup>38</sup> or Bob Carpenter's<sup>39</sup> during our first *recruitment drive*, a campaign we ran in January 2010 to celebrate the first year of Phrase Detectives activity (see the discussion on raising the visibility of a game in Section 4.6). The combination of increased advertising in particular among computational linguists and increased prizes (see next) was very effective: the number of players went from just over 1,200 to around 1,800 (an increase of over 50%) and the amount of completely annotated data doubled, from around 30,000 words to over 61,000 words. We observed similar effects in later drives, as discussed in Section 6.

*Financial incentives.* The whole point of using GWAP for resource creation, instead of crowdsourcing through Amazon Mechanical Turk or the like, is to have enjoyment of the game as the main incentive, instead of financial factors, thus hopefully lowering the overall cost. But as discussed in Section 3.1, a small financial incentive can still be provided in the form of prizes. Our experience suggests that prizes can have a substantial impact at a very low cost, but also that great care has to be paid to the type of prizes that are offered, and that frequent adjustments to the prize mechanism are required to ensure maximum effectiveness.

Monthly prizes for the highest-scoring players in the form of Amazon vouchers sent by email to the winners have been offered fairly regularly throughout the three years in which Phrase Detectives has been active. The monthly prize motivates the high-scoring players to compete with each other by doing more work, but also motivates some of the low-scoring players in the early parts of the month when the high score is low. Very quickly, however, we came to realize that given the dedication of some of our players, rewarding only the highest-scoring player would be discouraging to the other players, so we settled on offering prizes to the 3 highest scorers of the month. For the same reason, we introduced (typically, weekly) prizes awarded by randomly selecting an annotation. These prizes motivate low-scoring players because any annotation made during the prize time period has a chance of winning (much like a lottery) and the more annotations one makes, the higher one's chance of winning. These prizes are sometimes

<sup>37</sup><http://www.modul.ac.at/nmt/sentiment-quiz>.

<sup>38</sup><http://languagelog.ldc.upenn.edu/nll>.

<sup>39</sup><http://lingpipe-blog.com>.

awarded as an alternative to the highest-scoring prizes, sometimes in addition to those prizes.

The value range of the prizes is a third variable we experimented with. The prizes have ranged from £5–10 daily, £10–15 weekly, and from £30–£75 for the monthly high-scoring prizes.

Last but not least, we found it very effective, both as a recruitment tool and to increase productivity, to have occasional recruiting drives during which both promotion of the game (see the following) and prizes are stepped up, as already mentioned. During the January 2010 recruitment drive to celebrate the first year of Phrase Detectives activity, the advertising drive among computational linguists was paralleled with daily and weekly prizes and a grand prize of 500 euros for the top scorer of the month. As already said, this combination proved very effective. In Section 6 we present statistics about the success of Phrase Detectives in general including the overall effect of prizes (i.e., comparing prizes with no prizes). It has not been practical to study whether the specific details of the prizing mechanism (e.g., the frequency and level of prize) alters recruitment or performance as these were introduced with other methods of promotion.

It is, however, important to keep in mind that while financial incentives are important to recruit new players, a combination of all three types of incentives is essential for the long-term success of a site [Smadja 2009].

*Choice of text.* A final form of incentive provided to the players of Phrase Detectives is ensuring they read texts that they find interesting. From the very beginning the choice of documents was considered important in getting players to enjoy the game, to understand the tasks, and to keep playing. As discussed in Section 5, the texts to annotate consist for the most part of narrative texts from the Gutenberg collection and encyclopedic texts from Wikipedia. In both cases, the choice focused on texts we expected players to find most interesting to read (the Wikipedia texts, in particular, were chosen for their novelty or unusualness rather than their scholarly content).

While some of the chosen texts are straightforward, others can provide a serious challenge to readers, in particular when the task is resolving anaphors. Texts were therefore manually graded by administrators for complexity (on a scale of 1 to 4) after import. Players can choose the maximum level of document complexity they wish to read as they may be motivated to play the game to improve their English skills, or equally because they enjoy reading challenging texts.

Players can also specify a preference for particular topics in their profile; however, only 4% do so. This could be an indication that the corpus as a whole was interesting but it is more likely that they simply did not change their default options [Markey 2007].

We also allowed players to submit their own text to the system which would be processed and entered into the game. We anticipated that, much like in Wikipedia, this would motivate users to generate content and become much more involved in the game. Unfortunately this was not the case, with only one player submitting text. We have now stopped advertising this incentive, but the concept may still hold promise for games where the user-submitted content is more naturally created (e.g., collaborative story writing).

#### 4.4. Quality Control

The strategies for quality control in Phrase Detectives address four main issues:

- training and evaluating players;
- attention slips;
- malicious behavior;
- multiple judgments and genuine ambiguity.

We discuss each aspect in turn.

*Training and evaluating players.* One of the key differences between Phrase Detectives and the GWAP developed by von Ahn and his lab is the much greater complexity of judgments required of the players. Yet clearly we cannot expect players to be experts about anaphora, or to be willing to read a manual explaining how anaphora works, so all the training still has to be done while playing the game. Therefore, we developed a number of mechanisms that could help in this respect: giving suggestions and tips (global, contextual, and FAQ), comparing decisions with the gold standard, and sharing agreement with other players in validation mode.

*Help information* about the task is continuously presented to the players, using a variety of formats:

- very briefly on the homepage, covering the main aspects of the game;
- in a full Instructions page explaining in more detail the game, the scoring, the two gaming modes, and how a player should annotate the text;
- in a Frequently Asked Questions page where common email queries from players are added with explanations;
- in a small box on the player homepage where an instruction or hint is given about the game (chosen at random from over 20 such hints);
- during the game and when relevant to the markable text. For example, instructions specific to nonreferring markables appear whenever the markable is a variation of the pronoun *it* or *there*.

These instructions are constantly refined, with new examples and images added regularly in response to player feedback (in particular, examples of when to mark text as a property).

The second training mechanism is asking players to annotate text which has already been annotated (gold-standard text), so that their level of understanding and/or willingness to play correctly can also be assessed. Players always receive a training text when they first start the game, and may also need to complete one when being promoted to the next level (this is implemented in the Facebook version of the game). The training texts show the player whether their decision agrees with the gold standard (unambiguous markables are used in these cases, to avoid confusion). Once the player has completed all of the training tasks they are given a user rating (the percentage of correct decisions out of the total number of training tasks). The user rating is recorded with every future annotation because the user rating may change over time. Players are given training texts until the rating is sufficiently high to be given real text from the corpus (a minimum rating threshold of 50% is set for the game). This method is also used to eliminate noise, and is similar to the idea of “traps” [Tang and Sanderson 2010]. Last but not least, the training tasks prevent automated form completion software and malicious players from progressing far in the game.

Finally, players can learn about correct decisions by reinforcement, through validation mode. This builds on the assumption that the majority of players will agree with a good decision, which is not always the case, especially if the markable is complex or ambiguous. But by and large scoring high points in validation mode is an indication of a good interpretation.

*Attention slips.* Players may occasionally make a mistake and press the wrong button. We made a deliberate decision that there is no way that a player could go back and try again, else a player could try out all possible annotations and then select the one offering the highest score. Slips are identified and corrected by taking advantage of validation mode, where players can examine other players’ annotations and evaluate them. Through validation poor-quality interpretations should be voted down and

	AVERAGE	Good player	Bad player
<b>ANNOTATIONS</b>			
Total Annotations:	1423078	4587	11018
Average Annotation Time:	00:00:07	00:00:07	00:00:04
Total (Ratio) DN:	955520 (0.67)	1495 (0.33)	10935 (0.99)
Total (Ratio) DO:	378256 (0.27)	2696 (0.59)	58 (0.01)
Total (Ratio) PR:	79172 (0.06)	334 (0.07)	24 (0)
Total (Ratio) NR:	13395 (0.01)	64 (0.01)	2 (0)
<b>VALIDATIONS</b>			
Total Validations:	608982	3848	5256
Total (Ratio) Agree:	200174 (0.33)	1186 (0.31)	8 (0)
Ave Agree Time:	00:00:09	00:00:08	00:00:18
Total (Ratio) Disagree:	408808 (0.67)	2662 (0.69)	5248 (1)
Ave Disagree Time:	00:00:08	00:00:07	00:00:02
<b>OTHER</b>			
Total Skips:	51616	142	26
Skip per annotation:	0.04	0.03	0
Total Comments:	26593	229	0
Comment per annotation:	0.02	0.05	0

Fig. 4. Screenshot of the player profiling screen, showing the game totals and averages (left), a good player profile (center), and a bad player profile (right) taken from real game profiles. The bad player in this case was identified by the speed of annotations and the only responses were DN in annotation mode and disagree in validation mode. The player later confessed to using automated form completion software.

high-quality interpretations should be supported (in the cases of genuine ambiguity there may be more than one). Validation thus plays a key role as a second strategy for quality control.

*Malicious behavior.* Crowdsourcing systems attract spammers, which can be a real issue [Feng et al. 2009; Mason and Watts 2010; Kazai 2011]. However, in a game context we can expect spamming to be much less of an issue because there is less of an incentive when annotations are not conducted on a pay-per-annotation basis.

Nevertheless, several methods are used to identify players who are cheating or who are providing poor annotations. These include checking the player's IP address (to make sure that one player is not using multiple accounts), checking annotations against known answers (the player rating system), preventing players from resubmitting their decisions [Chklovski and Gil 2005], and keeping a blacklist of players to discard all their data [von Ahn 2006].

A new method of profiling players was developed for the game to detect unusual behavior. The profiling compares a player's decisions, validations, skips, comments, and response times against the average for the entire game; see Figure 4. It is very simple to detect players who should be considered outliers using this method (this may also be due to poor task comprehension as well as malicious input) and their data can be ignored to improve the overall quality.

However, the main method to filter out malicious input is again through validation.

*Multiple judgments and genuine ambiguity.* Collecting multiple judgments about every expression is a key aspect of Phrase Detectives, as in all other cases of using

**(67) Bristol Stool Scale - Wikipedia**

ID	Text	Skip			ReIs	Comments
9739	stool	0	<a href="#">6</a>	<a href="#">0</a>		
RelID	AnteID	RelType	Annotations	Agree	Disagree	Total
7551	9746	DO	13	3	1	15
12227	9749	DO	2	1	3	0
15658		PR	3	0	4	-1
19661		DN	5	1	3	3
88682	9745	DO	2	0	4	-2
91261	9761	DO	2	0	4	-2

Fig. 5. Screenshot of the administrative tool to view the annotations for a markable.

crowdsourcing for HLT [Snow et al. 2008; Feng et al. 2009; Albakour et al. 2010]. In the present version of Phrase Detectives we ask eight players to express their judgments on a markable. If they do not agree on a single interpretation, four more players are then asked to validate each interpretation<sup>40</sup>; see Figure 5.

Validation information has proven very effective at identifying interpretations produced by sloppy or malicious players: the value obtained by combining the player annotations with the validations for each interpretation

$$Ann + Agr - Disagr,$$

(where *Ann* is the number of players initially choosing the interpretation in annotation mode, *Agr* is the number of players agreeing with that interpretation in validation mode, and *Disagr* is the number of players disagreeing with it in validation mode) tends to be zero or negative for all spurious interpretations. This formula can also be used to calculate the “best” interpretation of each expression, which we will refer to in what follows as the *game interpretation*.

There is, however, one key difference between our judgment collection methods and the practice reported in other crowdsourcing work. As discussed in Section 4.1, anaphoric judgments can be difficult, and humans will not always agree with each other. For example, it is not always clear from a text whether a markable is referential or not; and in case it is clearly referential, it is not always clear whether it refers to a new discourse entity or an old one, and which one. In Phrase Detectives we are interested in identifying such problematic cases: if a markable is ambiguous, the annotated corpus should capture this information. We are therefore not aiming at selecting “the best,” or most common, annotation, but to preserve all interpretations in the corpus “exported” by the game (see Section 5); leaving it to subsequent interpretive processes to determine which interpretations are to be considered spurious and which instead reflect genuine ambiguity.

*Knowing more about the players.* Ultimately, our experience with Phrase Detectives suggests that the best way to filter out rogue players is to rely mostly or entirely on players picked from a social network of people that know each other. Although this would result in fewer players, our experience also suggests that most of the work is done by a minority of players, as discussed in Section 6. Such considerations are one

<sup>40</sup>It is possible for an interpretation to have more annotations and validations than required if a player enters an existing interpretation after disagreeing or if several players are working on the same markables simultaneously.

of the reasons for the development of the Facebook version of the game, discussed in Section 8.

#### 4.5. Administrative and Analysis Tools

We found it essential to invest time in the development of administrative tools to analyze the data produced by the players and to manage inputs, outputs, and users of the system. The tools we developed support the following.

- Analysis of Game Statistics.* These provide a selection of up-to-date statistics about the game that are useful for monitoring overall performance, such as total number of users, total words in the corpus, total annotations, average annotation times, throughput; as well as ways of monitoring how these numbers change over time.
- Analysis of Markable Statistics.* These allow us to visualize all annotations currently in the system broken down by document, then by markable, including annotations and validations for all interpretations, comments, skips, and markables excluded from the system by the administrators.
- Markable Administration.* All markables can be edited to correct mistakes created by the preprocessing pipeline or excluded from the game (but not deleted in order to maintain data integrity).
- Gold Standard Creation.* This is an interface for experts to annotate documents.
- Document Management.* All documents that have been imported to the game can have metadata attached, including complexity, language, and whether the theme is of an adult nature.
- Comment Management.* Users are allowed to provide comments, and such comments have proven invaluable to identify problems with the preprocessing or the annotation scheme. All user comments can be viewed for a given markable and dealt with, for example, to correct an error with a markable; see Figure 6.

#### 4.6. The Life of the Game: Testing, Deployment and Promotion

A great deal of the success of a game depends on testing it with potential players before going live, and after that, on ensuring it remains visible. We are convinced that the main reason why Phrase Detectives has attracted so many more players than other, equally well-designed GWAP for HLT is the effort we invested in raising its visibility. In this section we briefly discuss these aspects of the Phrase Detectives experience.

*Usability testing.* A first prototype of the game was built to test our initial ideas about game format and task design, using a small corpus of Aesop fables. This prototype was tested in February 2008 with a group of 16 players (staff and students at the University of Essex) who were paid a small amount to play the game for an hour while their actions and questions were recorded. This study led to significant interface refinements, in particular reducing task feedback (why the points were scored and how long it took to complete the task) and removing timing constraints. We also produced better instructions and examples of the tasks. The beta release to our friends and community took place in June 2008; this release was very important to fix bugs. The first full release took place in December 2008.

*Promotion.* In order to attract the number of participants required to make a success of the GWAP methodology it is not enough to develop attractive games; successful and continuous advertising is also required. Activities to raise the profile of Phrase Detectives were started from the very beginning, attempting to attract the attention not only of the computational linguistics community, but also of other scientists, and of the general public.

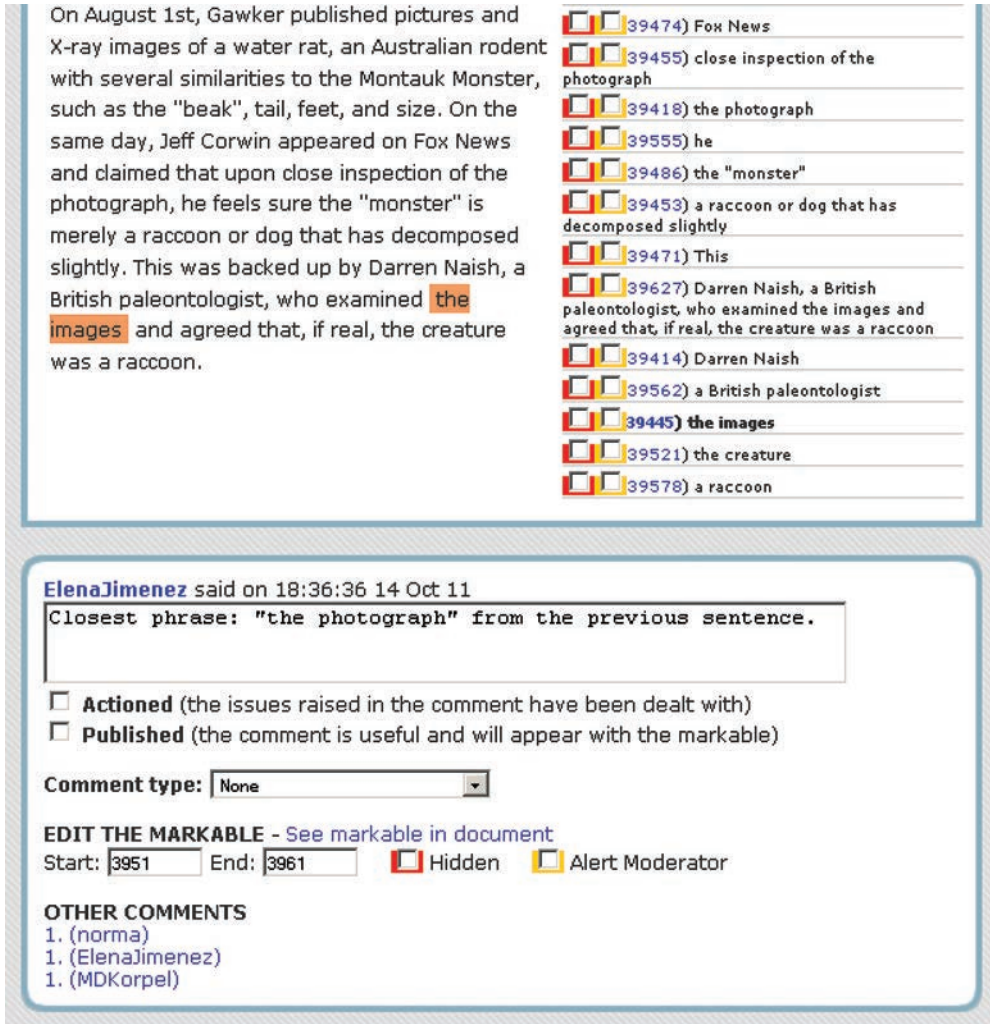


Fig. 6. Screenshot of the administrative tool to edit comments and markables.

The campaign to attract the general public began with press-releases in January 2009 that were picked up by *Science Daily* and *Innovations Report*, among other online publications, and by *Times Higher Education* among the regular academic journals; one of us (Chamberlain) was interviewed by the BBC. In addition the game was written about on blogs such as *Computer Science for Fun*<sup>41</sup> and was listed in bookmarking Web sites and gaming forums.<sup>42</sup>

At the same time, a pay-per-click advertising campaign was started on the social networking Web site Facebook. The analysis of the impact of any such type of advertising is quite difficult, but investigating the sources of traffic since live release using Google Analytics reveals that 46% of the incoming site traffic in February 2009 came from direct links, 29% from Web site links, 13% from the Facebook advert, 12% from

<sup>41</sup><http://www.cs4fn.org/linguistics/phrasedetectives.php>.

<sup>42</sup>For example, <http://www.gamescanteach.com/category/games/phrase-detectives>.



a search. This would suggest that the Facebook advert had some utility; however, the bounce rate (the percentage of single-page visits, where the user leaves on the page he/she entered on), which shows a relatively consistent figure for direct (33%), link (29%), and search (44%) traffic, is substantially higher for the Facebook advert (90%), meaning that 9 out of 10 users that came from this source did not play the game. This casts doubt over the usefulness of pay-per-click advertising as a way of attracting participants to a game.

Our efforts to reach out to the computational linguistics community in the first year involved announcements through mailing lists such as the Linguist List or Elsnet as well as presenting the game in a number of seminars, workshops, and conferences, also through postcard-size flyers. The efforts to reach out this community intensified during the first “recruitment campaign” of January 2010; in this period we also managed to get mentioned on blogs such as *Language Log*.<sup>43</sup> The success of these efforts cannot be measured in the same way (i.e., by tracking a link), but the figures on recruitment rate discussed in Section 6 suggest that they were effective.

Attracting large numbers of players to a game is only part of the problem. It is also necessary to attract players who will make significant contributions. We found that the top 5% highest-scoring players had 60% of the total points on the system and had made 73% of the annotations. This indicates that only a handful of users are doing the majority of the work, which is consistent with previous findings [Snow et al. 2008], however, the contribution of one-time users should not be ignored [Chamberlain et al. 2012].

## 5. PRODUCING A MULTILINGUAL CORPUS

The ultimate goal of Phrase Detectives is to obtain very large anaphorically annotated corpora for the languages covered (currently, English and Italian). In this section we discuss this aspect of the enterprise: what information is annotated; how data are imported and exported; how they are prepared for annotation; and the current composition of the corpus.

### 5.1. Coding Scheme

The Phrase Detectives corpus is annotated according to the linguistically-oriented approach to anaphoric annotation that is currently prevalent, having been adopted in OntoNotes [Pradhan et al. 2007], our own ARRAU corpus [Poesio and Artstein 2008], and in all the corpora used in the 2010 SEMEVAL anaphora evaluation [Recasens et al. 2010]. In this type of annotation, all NPs are considered markables, and anaphoric relations between all types of entities are annotated, unlike the practice in the MUC and ACE corpora.<sup>44</sup> (In the Phrase Detectives corpora, for instance, coordinated NPs like *John and Mary* are also considered markables.)

Players can assign four types of interpretation (labels) to markables:

- DN (discourse-new): this markable refers to a newly introduced entity;
- DO (discourse-old): this markable refers to an already mentioned entity (the player has to specify the latest mention);
- NR (nonreferring): this markable is nonreferring (e.g., pleonastic *it*);
- PR (property attribute): this markable represents a property of a previously mentioned entity (as in (2b)—e.g., *a teacher* in “He is a teacher”).

<sup>43</sup><http://languagelog.ldc.upenn.edu/nll/?p=2050>.

<sup>44</sup><http://projects.ldc.upenn.edu/ace/data>.

```

<sgf:corpusData xmlns:sgf="http://www.text-technology.de/sekimo" sgfVersion="1.1" xml:id="s_m1">
  <sgf:meta><!-- meta data goes in here --></sgf:meta>
  <sgf:primaryData start="0" end="24" xml:lang="en">
    <textualContent>The sun shines brighter.</textualContent>
  </sgf:primaryData>
  <sgf:segments>
    <sgf:segment xml:id="seg1" type="char" start="0" end="24"/>
    <sgf:segment xml:id="seg2" type="char" start="0" end="3"/>
    <sgf:segment xml:id="seg3" type="char" start="4" end="7"/>
    <sgf:segment xml:id="seg4" type="char" start="8" end="13"/>
    <sgf:segment xml:id="seg5" type="char" start="13" end="14"/>
    <sgf:segment xml:id="seg6" type="char" start="15" end="21"/>
    <sgf:segment xml:id="seg7" type="char" start="21" end="23"/>
  </sgf:segments>
  <sgf:annotation xml:id="a_morph">
    <sgf:level xml:id="a_morph_layer" priority="0">
      <sgf:meta><!-- meta data goes in here --></sgf:meta>
      <sgf:layer xmlns:morph="http://www.text-technology.de/sekimo/morphemes">
        <morph:morphemes sgf:segment="seg1">
          <morph:morphem sgf:segment="seg2"/>
          <morph:morphem sgf:segment="seg3"/>
          <morph:morphem sgf:segment="seg4"/>
          <morph:morphem sgf:segment="seg5"/>
          <morph:morphem sgf:segment="seg6"/>
          <morph:morphem sgf:segment="seg7"/>
        </morph:morphemes>
      </sgf:layer>
    </sgf:level>
  </sgf:annotation>
</sgf:corpusData>

```

Fig. 7. SGF representation of a morphological annotation.

Note that unlike the earlier coreference corpora, and following modern practice, in the Phrase Detectives corpora identity (annotated using the DO label) is sharply distinguished by predication, annotated using the PR label.

## 5.2. Input/Output

As discussed in Section 4.2, the data handled by Phrase Detectives are stored in a relational database whose design for the part concerned with storing texts and their annotations is based on that of the University of Bielefeld’s Serengeti system [Poesio et al. 2011a]. New texts are entered in the system through the Serengeti interface, that requires input in SGF format [Stührenberg and Goecke 2008]. The text must have been preprocessed to identify tokens, sentences, and noun phrases. The data are exported in an extended version of the MAS-XML format [Kabadjov 2007], designed to represent anaphoric information and to encode multiple interpretations. The extended version of MAS-XML, called PD-MAS-XML, can be used to export each interpretation assigned to each markable in the text. We briefly discuss SGF and PD-MAS-XML in turn.

*SGF.* The Sekimo Generic Format (SGF) [Stührenberg and Goecke 2008] was developed in the Sekimo project to support import and storage of multiple annotation layers, and as an exchange format for the Serengeti Web-based annotation tool (and other similar tools). The format uses a stand-off approach following the Annotation Graph’s model [Bird and Liberman 1999]. This makes it possible to use SGF for a great variety of linguistic annotations.

An example of SGF—the encoding in this format of the sentence *The sun shines brighter* and its morphological annotation—is shown in Figure 7. An SGF document includes, first of all, the declaration of a base layer which provides the primary data (i.e., the data that is annotated), inside a primaryData element. This is followed by the specification of the segments of the base layer that are annotated, that is, the markables,

using segment elements. (Note that SGF supports multiple levels of annotation; thus the segment elements specify the markables for all levels.) Segmentation of the base layer is usually character based. Finally, all annotations of are primary data are stored in annotation elements. For instance, the example in Figure 7 is an (automatically produced) annotation at the morph level in the University of Bielefeld's Sekimo annotation scheme identifying the segments as morph:morpheme elements.

In Phrase Detectives, the input SGF contains, in addition to the primary data, annotations indicating sentence and NP boundaries.

**MAS-XML.** The PD-MAS-XML format used to export Phrase Detectives data is a modified version of the Minimum Anaphoric Syntax (MAS-XML) format proposed in Kabadjov [2007]. MAS-XML is a form of inline XML in which the basic information required to carry out resolution is marked, including:

- sentences;
- words with their part-of-speech tags (for English, the Penn Treebank tagset is used);
- NPS (called Nominal Entities, ne), with their ID and the basic agreement features: gender (attribute gen for gold-standard info, AAgen for automatically extracted information), number (again two attributes are used, num and AAnum), and person (using the attributes per and AAPER);
- NP modifiers and heads, using the elements mod and nphead.

Note that the format does not require full syntactic information or named entity types. As an example, the representation in MAS-XML of NP *four little rabbits* is as follows.

```
<ne id="ne14819" AAcat="num-np"
  AAgen="neut" AAnum="plur" AAPER="per3">
  <mod id="AAm2" AAcat="AAPre">
    <W Lpos="CD">four</W>
    <W Lpos="JJ">little</W>
  </mod>
  <nphead id="AAh4">
    <W Lpos="NNS">rabbits</W>
  </nphead>
</ne>
```

Anaphoric information is marked using separate ante elements, a structured representation inspired by the Text Encoding Initiative link elements and that makes it possible to specify multiple anaphoric relations for each markable (identity and association) and to mark ambiguity using multiple anchor elements [Poesio 2004b], as in the following (made-up) example.

```
<ante current="ne3" rel="identity">
  <anchor antecedent="ne1"/>
  <anchor antecedent="ne2"/>
</ante>
```

The MAS-XML file for each document that is exported contains the original text and markup (sentences, NPS, and their features and constituents) automatically computed by the import pipeline, as well as the annotations produced by the players. To export the annotation information, the anchor mechanism from MAS-XML was replaced by a much more extensive format specifying for every player that expressed a judgment about a given markable the interpretation (DN for Discourse-New, DO for Discourse-Old, NR for NonReferring, or PR for Property), any antecedents selected for DO and PR interpretations, the user ID, the user rating, the time it took to make the annotation, whether the decision is an agreement, and in what mode the decision occurred (annotation or validation). Additionally players' comments are exported with the relevant markable and include the user ID, the type of comment, and the text that was submitted; and

so are skips. For instance, the (real-life) interpretation of markable ne14817, which all players interpreted as DN, is as follows.

```
<PDante id="ne14817">
  <interpretation>
    <anchor type="DN" user_id="281" user_rating="75" annotation_time="2" agree="y" mode="a"/>
    <anchor type="DN" user_id="728" user_rating="58" annotation_time="2" agree="y" mode="a"/>
    <anchor type="DN" user_id="779" user_rating="77" annotation_time="5" agree="y" mode="a"/>
    <anchor type="DN" user_id="281" user_rating="75" annotation_time="1" agree="y" mode="a"/>
    <anchor type="DN" user_id="18" user_rating="77" annotation_time="5" agree="y" mode="a"/>
    <anchor type="DN" user_id="1293" user_rating="64" annotation_time="15" agree="y" mode="a"/>
    <anchor type="DN" user_id="1364" user_rating="59" annotation_time="4" agree="y" mode="a"/>
    <anchor type="DN" user_id="163" user_rating="80" annotation_time="2" agree="y" mode="a"/>
    <anchor type="DN" user_id="1659" user_rating="92" annotation_time="9" agree="y" mode="a"/>
  </interpretation>
  <skip total="0"/>
</PDante>
```

Documents can be exported from Phrase Detectives in MAS-XML format either when they are complete (i.e., when all the markables have been annotated sufficiently according to the game configuration) or when they are partially complete. For the purposes of testing only complete documents have been exported.

### 5.3. Preprocessing

Adding texts in a new language to Phrase Detectives requires developing a pipeline to convert documents into SGF format importable in the database. Two such pipelines have been developed so far.

*The English pipeline.* The English Phrase Detectives pipeline converting raw text to SGF was developed by combining existing tools with ad hoc modules for correcting the output of such tools in the case of frequent errors, as follows.

- A preprocessing step normalizes the input, applies a sentence splitter, and runs a tokenizer over each sentence. The tokenizer and sentence splitter used to perform this process are from the popular *openNLP* toolkit.<sup>45</sup>
- A custom-developed postprocessing step is carried out to clean systematic errors by the tokenizer and sentence splitter.
- Each sentence is then analyzed by the Berkeley Parser [Petrov et al. 2006], often considered the best constituency parser for English.
- The parser output is then used to identify markables in the sentence. As a result a MAS-XML -like representation is created that preserves the syntactic structure of the markables (including nested markables, e.g., noun phrases within a larger noun phrase).
- A heuristic processor identifies additional features associated with markables such as person, case, number, etc. The output format is MAS-XML.
- MAS-XML is converted to SGF using XSL stylesheets and Saxon.<sup>46</sup>

*The Italian pipeline.* In order to use Phrase Detectives to annotate Italian data, a new pipeline [Robaldo et al. 2011] was developed using the TULE parser [Lesmo and Lombardo 2002]. The parser processed the raw text directly with Italian texts so no preprocessing is needed.

<sup>45</sup><http://incubator.apache.org/opennlp>.

<sup>46</sup><http://saxon.sourceforge.net>.

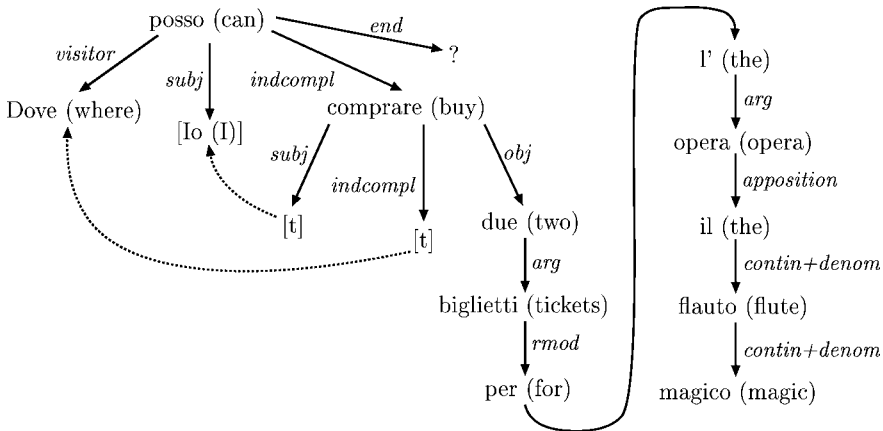


Fig. 8. TULE dependency tree for: “Dove posso comprare due biglietti per l’opera Il Flauto Magico?” (*Where can [I] buy two tickets for the opera The Magic Flute?*).

- The input is analyzed by TULE, which is a *dependency* parser.<sup>47</sup> An example of TULE output is shown in Figure 8. Note that TULE is able to identify morphologically unrealized components such as the subject of the verb *posso comprare*, so that such elements can be made explicit in the version of the text presented to the players and annotated.
- A custom Java module identifies markables on the basis of the dependency links among words. The Java module produces the MAS-XML format corresponding to the input text.
- MAS-XML is converted to SGF via Saxon, as for the English pipeline.

*An evaluation.* Developing a high-quality pipeline is one of the most important, yet most challenging, aspects of the development of GWAP for text, as the quality of the syntactic analysis greatly affects the experience of the players. The performance of the English and Italian pipelines was analyzed and compared by Robaldo et al. [2011]. In particular, using the markable administration administrative tool of Phrase Detectives, it was possible to analyze the number of markable identification errors in 10 English and 10 Italian documents, finding that the English pipeline produces on average 4.56 errors per text, whereas the Italian version only produces 0.67. It is not clear to us why there is such a difference in performance. The Italian parser is very good, regularly scoring first or joint first for parsing at the Italian evaluation campaigns EVALITA<sup>48</sup>, but so is the Berkeley parser. The only explanation we have at the moment is that a great deal more effort was invested in the development of the Italian pipeline, but a more in-depth analysis will have to be carried out before preprocessing further English text.

*Markable correction.* Our experience with Phrase Detectives suggests that the state-of-the-art in HLT is unfortunately not yet such that a pipeline composed of off-the-shelf modules can achieve adequate performance: the 4.56 error per text with the

<sup>47</sup>Two main types of parsers are used in HLT. The more traditional constituency parsers, like the Berkeley parser or the Charniak parser, analyze text according to traditional phrase structure theory, that is, they produce an output similar to that used in the Penn Treebank [Marcus et al. 1993]. Dependency parsers, by contrast, analyze text according to (some variant of) dependency grammar, a syntactic theory in which there are no phrasal nodes like NP or S, and the structure of a sentence expresses the dependencies between the lexical elements [Nivre 2005]. In recent years, dependency parsers have become increasingly dominant in HLT due to their higher accuracy (especially for languages other than English) and greater speed.

<sup>48</sup><http://www.evalita.it>.

English pipeline has proven too high. However, the results obtained with the Italian pipeline suggest that better results may be possible even for English if substantial effort is invested. In practice, at present the markable administration tool plays an important role in making the Phrase Detectives experience tolerable, at the expense of administrators having to spend a great deal of time to correct markables. This is clearly only a temporary solution as it is a substantial bottleneck. In the long run we would want, on the one hand, to improve the performance of the pipelines; on the other, to find effective ways to involve at least some experienced and trusted players in this aspect.

#### 5.4. The English and Italian Corpora

As our ultimate goal is to produce a freely distributable corpus, the texts of the English and Italian corpora are from collections not subject to copyright restrictions. We discuss each corpus in turn.

*English.* The English texts come from three main domains:

- Wikipedia articles selected from the “Featured Articles” page<sup>49</sup> and the page of “Unusual Articles”<sup>50</sup>;
- narrative text from Project Gutenberg<sup>51</sup> including in particular a number of tales (e.g., Aesop’s Fables, Grimm’s Fairy Tales, Beatrix Potter’s tales), and more advanced narratives such as several Sherlock Holmes short stories by A. Conan-Doyle, *Alice in Wonderland*, and several short stories by Charles Dickens;
- dialog texts from Textfile.<sup>52</sup>

The ultimate objective is to annotate over 100 million words, and several million words of text have already been converted, but in part because the accuracy of the present pipeline is not considered high enough, at present only around a million words have been actually uploaded in the English version of Phrase Detectives; to be precise, 1,206,597 words from 839 documents.

*Italian.* The same criteria concerning distribution were used for the texts in the Italian version of the game; an additional criterion has been the kind of linguistic phenomena that they are likely to include. The sources are the Italian version of Wikipedia and two novels by Wu Ming (CC licensed).

The texts from Wikipedia belong to two specific subgenres (plots and biographies) which are likely to contain a dense net of antecedents. The first kind displays a significant number of pronominal anaphors, while the second might display examples of lexical noun phrase anaphora (e.g., “the Queen” and “her Majesty”). In addition to the mentioned subgenres other uncategorized texts have been chosen in order to provide a comparison with the English version of the game (“Chess Boxing” and “Diet Coke and Mentos Explosion” are in both corpora).

The novels have been selected to test if the narrative style has an influence on the performance of the parser and of the players. This variety is more likely to display all the pronouns of the language, particularly 1st and 2nd person in reported speech, which are less likely to appear in Wikipedia articles.

<sup>49</sup>[http://en.wikipedia.org/wiki/Wikipedia:Featured\\_articles](http://en.wikipedia.org/wiki/Wikipedia:Featured_articles).

<sup>50</sup>[http://en.wikipedia.org/wiki/Wikipedia:Unusual\\_articles](http://en.wikipedia.org/wiki/Wikipedia:Unusual_articles).

<sup>51</sup><http://www.gutenberg.org>.

<sup>52</sup><http://www.textfiles.com>.

The Italian corpus for Phrase Detectives currently contains 30 texts, for a total of 11,373 words.

*Distribution.* Data from the game will be made available through the *Anaphoric Bank* [Poesio et al. 2011a].<sup>53</sup>

## 6. EVALUATION

In this section we report the results of several forms of evaluation of the results obtained with Phrase Detectives: from a quantitative perspective (how many players we recruited, how much labeling they did), as well as from the perspective of the quality of the results, evaluated using criteria including:

- agreement*: how the aggregated results obtained from the game compare to expert judgments;
- using the data to *train anaphoric resolvers*.

Last but not least, we evaluated the cost effectiveness of Phrase Detectives in comparison with other types of annotation methods we also use.

### 6.1. A Quantitative Assessment

Since the first release of the game in December 2008 to January 2012 just over 8,000 players have registered, 3,000 of which went beyond the initial training phase. These players did more than 5,000 hours of work, that is, 2.5 person-years. The average throughput of the game (labels per hour [von Ahn and Dabbish 2008]) is 450 annotations per hour. Average lifetime play (time in minutes spent on average by players in front of the game summing up all their interactions) is 2105 secs (35 mins and 5 secs), but our experience suggests that in the case of Phrase Detectives at least this statistic masks a massive difference between players that spend little or no time on the game and players that play continuously.

407 documents were fully annotated, for a total completed corpus of over 162,000 words, 13% of the total size of the collection currently uploaded for annotation in the game (1.2M words). This is about the size of the ACE2 corpus of anaphoric information, the standard for evaluation of anaphora resolution systems until 2007/08 and still widely used. The size of the completed corpus does not properly reflect, however, the amount of data we have collected, as the case allocation strategy adopted in the game privileges variety over completion rate. As a result, almost all the 800 documents in the corpus have already been partially annotated. This is reflected first of all in the fact that 84280 of the 392,120 markables in the active documents (21%) have already been annotated. This is already almost twice the total number of markables in the entire OntoNotes 3.0 corpus,<sup>54</sup> which contains 1 million tokens, but only 45,000 markables. But the number of partial annotations is even greater. Our players produced over 2.5 million anaphoric judgments between annotations and validations; this is way more than the number of judgments expressed to create any existing corpus. To put this in perspective, the GNOME corpus, of around 40K words, and regularly used to study anaphora until 2007/08, contained around 3,000 annotations of anaphoric relations [Poesio 2004a] whereas OntoNotes 3.0 only contains around 140,000 annotations.

It is also interesting to look at the rate at which players and data have been increasing in the last three years. It is illustrated in Figure 9 (for the players) and Figure 10 (for the amount of annotation). These charts show first of all that both the number of players and the amount of annotation are still growing; in fact, the rate of growth is

<sup>53</sup><http://anawiki.essex.ac.uk/anaphoricbank>.

<sup>54</sup><http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T24>.

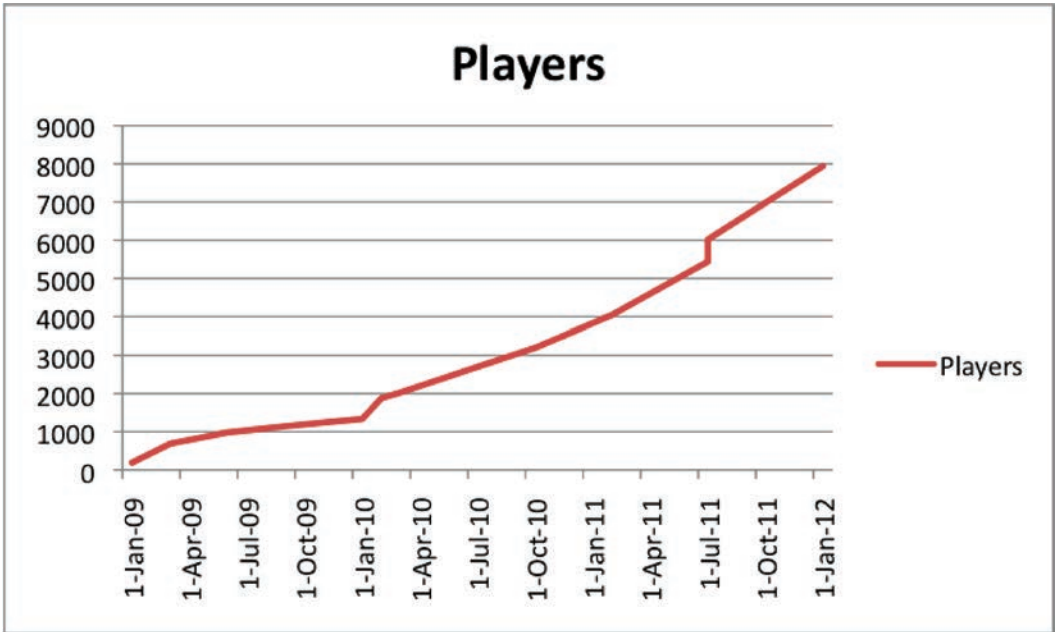


Fig. 9. Growth in number of players from January 2009 to January 2012.

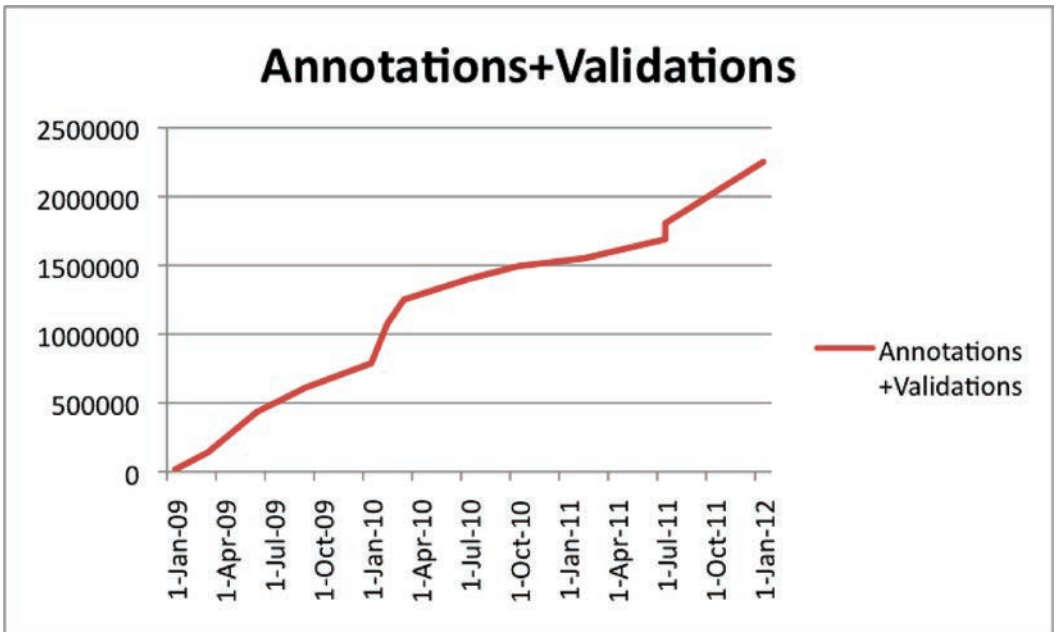


Fig. 10. Increase in total amount of judgments (annotations and validations).

increasing. They also highlight the impact of the two recruitment drives we carried out in January 2010 and in July 2011; in each case we see sudden steps upwards in the rate of growth. These charts give us reason to believe that we can expect Phrase Detectives to keep producing data for a few years more.



Table I. Agreement on Annotations

	Expert 1 vs. Expert 2	Expert 1 vs. Game	Expert 2 vs. Game
Overall agreement	94.1%	84.5%	83.9%
DN agreement	93.9%	96.0%	93.1%
DO agreement	93.3%	72.7%	70.0%
NR agreement	100.0%	100.0%	100.0%
PR agreement	100.0%	0.0%	0.0%

*Agreement figures:* DN = discourse-new, DO = discourse-old, NR = nonreferring, PR = property attribute.

## 6.2. Agreement on Annotations

One way to tell whether the game is indeed successful at obtaining good-quality anaphoric annotations is to check how the aggregated annotations produced by the game compare to those produced by an expert annotator. But because anaphoric annotation is much harder than, say, part-of-speech annotation, in which it is possible to reach very high agreement, we also looked at a second question, namely, what is the agreement between two experts annotating those texts?

In order to answer these questions, we randomly selected five completed documents from the Wikipedia corpus containing 154 markables. Each document was manually annotated by two experts (called Expert 1 and Expert 2 in the rest of this discussion) operating separately; we then compared the annotations produced by the experts with the most highly ranked interpretations produced by the players on the basis of the formula in Section 4.4 (henceforth, the *game interpretation*), and the experts' annotations with each other.

As discussed in Section 5.1, players can assign four types of interpretation (labels) to markables:

- DN (discourse-new): this markable refers to a newly introduced entity;
- DO (discourse-old): this markable refers to an already mentioned entity (the player has to specify the latest mention);
- NR (nonreferring): this markable is nonreferring (e.g., pleonastic *it*);
- PR (property attribute): this markable represents a property of a previously mentioned entity (e.g., as a *teacher* in “He is a teacher”).

Our experts judged DN to be the most common interpretation, with 70% of all markables falling in this category. 20% of markables are DO and form a coreference chain with previous markables. Less than 1% of markables are nonreferring. The remaining markables have been identified as expressing properties.

Overall, agreement between experts on the types is very high although not complete: 94% for a chance-adjusted  $\kappa$  value [Artstein and Poesio 2008] of  $\kappa = .87$ , which is extremely good. This value can be seen as an upper boundary on what we might get out of the game. Agreement between each of the experts and the game is also good: we found 84.5% percentage agreement between Expert 1 and the game ( $\kappa = .71$ ) and 83.9% agreement between Expert 2 and the game ( $\kappa = .7$ ). In other words, in about 84% of all cases the interpretation specified by the majority vote of non-experts was identical to the one assigned by an expert. These values are comparable to those obtained when comparing an expert with the “normally trained” annotators (usually students) that are typically used to create medium-quality resources (see Section 6.5). Table I gives a detailed breakdown of pairwise agreement values.

Looking separately at the agreement on each type of markables, we see that the figures for DN are very close for all three comparisons, and well over 90%. This seems

to be the easiest type of interpretation to identify. D0 interpretations are more difficult, with only 71.3% average agreement. If, however, we relax the notion of agreement for this type not comparing the antecedent specified by the players, we get agreement figures above 90% for this class as well: almost 97% between the two experts and between 91% and 93% when comparing an expert with the game. In other words, players agree to a considerable degree on a given markable being anaphoric, but much less on what the antecedent is. However, many of these disagreements are actually *spurious* ambiguities—cases in which players indicate different NPs as the last mention of an entity, but these NPs are actually mentions of the same entity. Also, due to the limited context presented by the game, players may not be able to select the last mention of a given entity that appeared earlier in a document (in many such cases the players indicate the problem by creating a comment). Analyzing these disagreements to identify spurious and real ambiguities and developing automatic methods for spotting them is one of our goals for the immediate future.

Of the other two types, the 0% agreement between experts and the game on property interpretations suggest that they are very hard to identify, or possibly our training for that type is not effective. Nonreferring markables on the other end, although rare, are correctly identified in every single case. We separately checked every completed markable identified as NR in the corpus and found that there was 100% precision in 54 cases.

Finally, looking at the disagreements between experts and the game (i.e., those cases where the experts' interpretation is different from the most highly ranked interpretation in the game) we find the following.

- In 60% of all cases where the game proposed an interpretation different from the expert annotation, the expert marked this interpretation to be possible as well. In other words, the majority of disagreements are not incorrect annotations but alternatives such as ambiguous interpretations or references to other markables in the same coreference chain. If we counted these cases as correct, we get an agreement ratio of above 93%, close to pairwise expert agreement.
- In cases of disagreement, the expert-marked interpretation was typically the second or third highest-ranked interpretation in the game.
- The cumulative score of the expert interpretation (as calculated by the game) in cases of disagreement was 4.5, indicating strong player support for the expert interpretation. (A score around zero would be interpreted as one that has as many players supporting it as it has players disagreeing; a value above zero indicates a majority of supporters.)

### 6.3. A Linguistic Assessment: Ambiguity in the Corpus

We are in the process of analyzing the judgments accumulated so far in preparation for a paper on anaphora through the lens of Phrase Detectives, and some interesting results already came up, in particular about the notion of coreference (e.g., in many mysteries, the whole point of the story is that the identity of a character—the culprit, or some shady figure—is only discovered at the end). We will not enter into this discussion here, but one preliminary statistic is worth reporting given the motivating role that studying anaphoric ambiguity has had in the design of the game. In January 2011 there were 63009 completely annotated markables. Of these, 23479 (37.3%) had exactly one interpretation (i.e., the first eight players to be presented with that markable all chose the same interpretation). Of these, 23,138 were DN, 322 D0, and 19 NR. A further 13,772 markables (21%) had only 1 interpretation with a score greater than 0. Again, the majority of these (9,194) were DN; 4,391 were D0, and NR 175. These results, besides confirming what was said before about DN being easier to interpret than D0, and NR also

being relatively easy to identify, suggest that the percentage of NPs that are ambiguous could be even greater than we expected: 41.4% of markables have more than one interpretation supported by at least 3 players. (For an interpretation to have a score greater than 0 it must have been proposed by at least 1 player and be positively validated by at least half of validators, i.e., at least 2.)

#### 6.4. Using the Phrase Detectives Data to Train Anaphora Resolution Models

An alternative way to evaluate the quality of the annotated data is to use the Phrase Detectives data for training anaphora resolution algorithms. We carried out two studies of this type as sanity checks, one in 2009, and one in 2010. In both tests, the BART anaphora resolution toolkit [Broscheit et al. 2010] was used to train a Soon-et-al.-style model [Soon et al. 2001] by using the top interpretation from the game.

In the first study 21 documents from the Gutenberg data were used, all fairy tales, for a total of 12K words (i.e., about half the size of the commonly used MUC6 corpus). 16 documents were used for training and 5 for testing. The performance of the model was measured using the commonly used MUC score [Vilain et al. 1995] which measures precision, recall and F-value at finding anaphoric links. The model achieved  $F=.58$ , that is, on the higher end for systems doing the “all mentions”, “all modifier” task: for example, in the 2011 CONLL shared task, in which the OntoNotes 3.0 data were used which are annotated in a similar way, F-values of .58 were achieved by the top systems [Pradhan et al. 2011].

In the second study we used five times as much data as in the earlier experiment, and we took the documents from the Wikipedia subset instead. 190 documents were used (of which 130 used for training and 60 for testing), for a total of 60K words (i.e., about twice the size of the MUC6 corpus). This time the model achieved an F-value of .49, that is, on the lower end of the performance for this type of dataset but still in the general ballpark. We are analyzing the results to understand whether the lower results are due to increased difficulty in the types of anaphora or noisier data.

#### 6.5. Cost Effectiveness

In the Introduction we stated that one of the main reasons for using GWAP for annotation is the hope that this approach will result in much lower costs in comparison with traditional (high-quality) annotation—which, as discussed earlier, is at a cost of around 1 million US \$ per 1 million words, so is not a feasible approach to create a 100-million-word corpus, or even a 10-million-word one. In this final section we analyze our experience with Phrase Detectives from this perspective, also attempting a comparison with the costs of annotation using crowdsourcing.

Comparing annotation costs is always difficult as so much depends on local salary levels and factors that change from lab to lab, such as whether it is the principal investigators themselves who oversee the annotation, or whether postdocs are hired for this purpose, but we will try to estimate costs for what seems to us a fairly typical situation (the principal investigator develops the coding scheme, postdocs follow the actual annotation full-time) and using our own actual costs whenever possible. We distinguish between four annotation methods.

*Traditional, High Quality (THQ).* When people talk about annotation being expensive they usually think of this methodology, used, for example, in OntoNotes, but also to create the ACE and MUC corpora, the SALSA corpus in Germany, etc. In this approach, a formal coding scheme is developed, and often extensive agreement studies are carried out; then every document is doubly annotated according to the coding scheme by two professional annotators under the supervision of an expert, typically a linguist, and annotation is followed by merging of the annotations. These projects also generally

involve the development of suitable annotation tools or at least the adaptation of existing ones. It is this type of annotation which requires in the order of 1 million US \$ per 1 million tokens, that is, 1 dollar per token (several anonymous annotation experts, p.c.). On average our texts contain around 1 markable every 3 tokens, so we get a cost of 3 US \$ per markable.

*Traditional, Medium Quality (TMQ).* This type of annotation also involves the development of a formal coding scheme and training of annotators, but most items will be typically annotated only once, although around 10% of items are double-annotated to spot misunderstandings; also, in many cases annotation will have to be corrected, possibly extensively. Annotation is typically carried out by trained but not professional annotators, generally students, under the supervision of an expert annotator. Our own estimates (at UK/Italy costs) for this type of work are in the order of 330,000 euros/400,000 US \$ per 1,000,000 tokens, including expert annotator costs; that is, around .4 US \$ per token, or 1.2 US \$ per markable, that is, slightly more than two-fifths of the costs with THQ.

*Crowdsourcing.* Cost with Amazon Mechanical Turk depends on the amount paid per HIT and on the extent of reduplication. In our experience, .05 US \$ per HIT is the minimum required for nontrivial tasks, and for a task like anaphora, the cost is more like .1 US \$ per markable. Also, although many researchers only require five judgments per item, in practice we find that 10 is more like the number needed; this results in a cost of 1 US \$ per markable, that is, around 330,000 US \$ per million tokens. In addition, an expert annotator/Mechanical Turk user is typically required to set up the task and follow it up. On the other end, this work tends to be very fast, so we could imagine a very optimistic scenario in which the annotation of 1 million words is completed in 1 year, although we think 2 years is probably a more realistic estimate. Assuming a salary of 50,000 US \$ per year for the expert annotator, this would give a total cost in the range of 380,000–430,000 US \$ per million tokens/1.2–1.3 US\$ per markable, that is, about the same cost as with TMQ. Apart from the optimistic assumptions about speed, however, the real question is whether as complex a labeling task as anaphora resolution can be really turned into a HIT.

*GWAP.* The total cost for running Phrase Detectives so far has been around 60,000 £ in salary costs for setting up and running the game and around 6,000 £ in prizes for three years, that is, a total of around 100,000 US \$ per 162,000 complete tokens—but in fact over 84,000 markables have been completely annotated, at a cost of 1.2 US\$ per markable. But with GWAP, most of the expense takes place at the beginning, to set up the game: we spent 65,000 US \$ for the first two years at the end of which we had the game, but less than 60,000 words and 10,000 markables fully annotated. In the following 2 years, during which 74,000 markables were completely annotated, the cost has been 35,000 US \$, that is, .47 US \$ per markable, which is less than half the costs with Mechanical Turk. This figure gives a projected cost of 217,927 for 1 million words (65,000 US \$ for the first 10,000 markables, 152,927 US \$ for the other 323,333 markables that one can expect to find on average in 1,000,000 words). This is about half the cost one may expect with AMT. The one problem is that the rate of 34,000 completed markables a year is not fast enough: at this speed, it would take 9 years to complete the 307,480 markables remaining in the documents already active in Phrase Detectives. Many more players are needed to complete our goal of a 100-million-words corpus. (With 100,000 players, i.e., with half the players who played the ESP Game, this target could be achieved in 9 years.)

This comparison is summarized in Table II.

Table II. Comparison of Costs in US\$ Using Four Different Annotation Methods

Method	Cost/token	Cost/markable	Cost/million tokens
Traditional, High Quality	1	3	1,000,000
Medium, High Quality	.4	1.2	400,000
Amazon Mechanical Turk	.38	1.2-1.3	380,000-430,000
Games With A Purpose	.19	.47	217,927

## 7. OTHER GWAP FOR CORPUS CREATION AND ANNOTATION

In this section we discuss those GWAP whose aims are most closely related to those of Phrase Detectives, namely creating or annotating a corpus.

### 7.1. Creating a Corpus for Translation: 1001 Paraphrases

*1001 Paraphrases* [Chklovski 2005]—to our knowledge, the first GWAP whose aim was to collect corpora—was developed to collect training data for a machine translation system that needs to recognize paraphrase variants. In the game, players have to produce paraphrases of an expression shown at the top of the screen, like *this can help you*. If they guess one of the paraphrases already produced by another player, they get the number of points indicated on the window; otherwise the guess they produced is added to those already collected by the system, the number of points they can win is decreased, and they can try again. Chklovsky reports collecting 20,944 contributions.

From a methodological point of view, the main point to note is that the task in this game is crucially different from the task in Phrase Detectives: as in the ESP Game, players are required to enter text instead of choosing one interpretation. So the method could not be directly used for anaphoric annotation, but could be tried for other translation-related applications, or possibly other tasks such as summarization or natural language generation. However, many of the ideas developed by Chklovsky in *1001 Paraphrases* and the earlier *LEARNER* system (not really a game) are extremely useful, in particular the idea of validation. It is difficult, however, to assess how successful the game was as the paper mentioned only reports a small-scale pilot study.

### 7.2. Creating (and Annotating) a Corpus for Generation: GIVE

A family of GWAP which have been used to collect data actually used in computational linguistics are the GIVE games<sup>55</sup> developed in support of the the GIVE-2 challenge for generating instructions in virtual environments, initiated in the natural language generation community [Koller et al. 2010]. GIVE-2, for instance, is a treasure-hunt game in a 3D world. When starting the game, the player sees a 3D game window, which displays instructions and allows the players to move around and manipulate objects. In the first room players learn how to interact with the system; then they get in an evaluation world where they perform the treasure hunt, following instructions generated by one of the systems participating in the challenge. The players can succeed, lose, or cancel the game; this outcome is used to compute the *task success* metric, one of the metrics used to evaluate the systems participating in the challenge.

GIVE-2 was extremely successful as a way to collect data for HLT, collecting over 1825 game sessions in three months, which played a key role in determining the results of the challenge. No doubt this is due in part to the fact that it is an extremely attractive game to play. Again, this methodology would not be appropriate to annotate preexisting text; it may be possible, however, to learn about anaphora from the data produced this way.

<sup>55</sup><http://www.give-challenge.org>.

### 7.3. Other GWAP for Anaphora

Of the GWAP developed by the HLT community, the game more directly comparable with Phrase Detectives is *PlayCoref*, developed at Charles University in Prague [Hladká et al. 2009]. *PlayCoref* is a two-player game in which players can interact with each other. A number of empirical evaluations have been carried out showing that players find the game very attractive but to our knowledge the game has not yet been put online to collect data on a large scale.

### 7.4. GWAP for Other HLT Tasks

*PhraTris* [Attardi and the Galoap Team 2010] is a GWAP for syntactic annotation developed by Giuseppe Attardi's lab at the University of Pisa using a general-purpose GWAP development platform called GALOAP.<sup>56</sup> *PhraTris* is a very entertaining game and won the INSEMTIVES game challenge 2010 but has not yet been put online to collect data.

## 8. CONCLUSIONS

### 8.1. Developing GWAP for HLT: Top 10 Considerations

Phrase Detectives was one of the very first GWAP applied to resource creation for HLT and in quantitative terms has been the most successful, collecting over 2.5 million judgments from almost 8,000 players. In these years we learned a number of useful lessons about GWAP in general and GWAP for HLT in particular that we discussed throughout the article. In this section, we summarize our top tips in the style of a GWAP leaderboard.

- (1) Consider whether one's task is too difficult to be presented in a game format. If one only needs to annotate smaller amounts of data (in the order of hundreds of thousand words), and the task is fairly simple, crowdsourcing is probably the easiest solution.
- (2) Make use of the full range of player incentives: personal; social; and financial, if possible. A truly entertaining game will require few or no other incentives, but even such a game may benefit from prizes, and even small-scale prizes (in the order of US \$ 70 a month) are remarkably effective.
- (3) Ensure the interface is easy to use, intuitive to learn, and is designed to engage one's intended player demographic.
- (4) Promoting the game throughout is essential, and not all promotion methods are equally effective. There are thousands of online games out there; achieving visibility *and maintaining it* require constant effort. Balance the budget one has for promotions against the financial incentives one can offer. Both are effective ways of recruiting players, but one may get better value for money by offering prizes if one already has access to a large user group.
- (5) Consider how it might be possible to manipulate one's GWAP in ways one did not intend and how one will detect and control cheating players that do this. This may also apply to players who provide poor data and may be an indication that one's player training needs to be reviewed.
- (6) Validation is an extremely effective method for quality control.
- (7) Collecting multiple judgments for each expression is essential for quality control, and provides very useful linguistics data.
- (8) Investing time in ensuring that the preprocessing pipeline producing the input to the game is as robust, flexible, and as error free as possible will avoid many manual corrections later.
- (9) Player training and task instructions are worth spending time developing and refining in response to player feedback.

<sup>56</sup><http://galoap.codeplex.com>.

- (10) Social networks provide a perfect platform for GWAP delivery, with access to user demographic data and inherent social advertising. However, it comes with the cost of having to maintain the GWAP on a platform that may change its API protocols over time.

## 8.2. Further Developments

8,000 is a respectable number of players and much more than attracted by any other GWAP for HLT but it is not yet the number of players that could result in massive data creation over a short period of time (the 100,000 players of Google Image Labeler). The original Web game is still being played and new data accumulated, but the rate of new data creation and new players registering needs to increase. Therefore, we are continuously launching new initiatives aimed at reaching the numbers we are hoping to achieve. We briefly discuss two of these here.

*Phrase Detectives on Facebook.* Social networking now dominates the amount of time users spend online. It is reported to be as much as 22%, double the amount of time spent playing online games, three times the amount of time spent doing email, and seven times the amount of time spent doing searches.<sup>57</sup> It is therefore becoming vital for online games to be linked to social networking sites like Facebook, which are a very promising platform for online games to achieve visibility through word-of-mouth and to study new forms of collaborative playing. In addition, we felt that having (some of) our players come from a social network could make it possible to control better the quality of players, for example, by only using output from players part of the social network. For these reasons, a Facebook version of Phrase Detectives<sup>58</sup> was launched in February 2011. We briefly discuss the new platform here; a more detailed discussion can be found in Chamberlain et al. [2012].

Facebook Phrase Detectives maintains the overall game architecture while incorporating a number of new features developed specifically for the social network platform. The game was developed in PHP SDK (a Facebook API language allowing access to user data, friend lists, wall posting, etc.) and integrates seamlessly within the Facebook site. Data generated from this version of the game is compatible with previous versions and both current implementations of the game run simultaneously on the same corpus. The Facebook version of Phrase Detectives includes many refinements and bug fixes, including better instructions, cleaner imagery, and faster gameplay by removing the scoring feedback screen. For instance, score feedback is now presented in the left-hand menu as a phrase such as “Perfect!” or “Good agreement!” depending on how many other players the decision agrees with. Player levels now have more, well-defined criteria and the player must activate the new level once the conditions are met. Criteria includes:

- total points scored;
- total documents that have been completed;
- total number of Facebook posts made from the game;
- the player’s rating;
- total number of training documents completed.

The promotion criteria not only make the game more interesting but also prevent players who send automated, malicious, or poor decisions from getting very far in the game. A key element is that a training document must be completed at every level of promotion and, like its predecessor, the game asks the player to keep doing training

<sup>57</sup><http://mashable.com/2010/08/02/stats-time-spent-online>.

<sup>58</sup><http://apps.facebook.com/phrasedetectives>.

documents until a sufficiently high rating is achieved. The rating threshold is increased at higher levels and ensures higher-level players are capable of providing high-quality annotations.

The game makes full use of socially motivating factors inherent in the Facebook platform. Any of the player's friends who are playing the game form the player's team, which is visible in the left-hand menu. Whenever a player's decision agrees with a team member the player scores additional points.

Wall (or news) posting is integrated into the game. This allows a player to make an automatically generated post to his/her news feed (or wall) which will be seen by all of the player's friends. The wall post shows the document the player is working on, his/her position in the leaderboard, or directly asks friends to join his/her team. The posts include a link back to the game. Players are required to make a post from the game every time they are promoted to the next level (although the post can be deleted immediately). Posting is a very important factor in recruiting more players as studies have shown that a majority of social game players start to play because of a friend's recommendation.<sup>59,60</sup>

The Facebook game also incorporates new leaderboards including the highest-level player, highest rated player, and the player with the biggest team.

*Phrase Detectives on a smartphone.* Of the world's 4 billion mobile phones approximately 25% are smartphones.<sup>61</sup> It is expected that by 2014 Internet access via mobile phones will overtake Internet access via desktops or laptops.<sup>60</sup> Already 46% of the teenage US population play online games on their smartphones<sup>62</sup> and this is likely to increase. Developing GWP that can be downloaded for free on smartphones looks promising as another way of increasing the use of such games.

We are in the process of developing a cross-platform version of Phrase Detectives that will run on iPhones, Android phones and a number of other architectures. The new version will also feature a different graphical interface aimed at making the game more fun, in which players associate discourse entities with icons of their choice.

## ACKNOWLEDGMENTS

A great number of people in the community have helped us to develop the game and promote it. We would like to give particular thanks to the SEKIMO group at the University of Bielefeld (Daniela Goecke, Nils Diewald and Maik Stührenberg), to the audiences at a number of invited talks including the 2010 IK Winter School, Ans Alghamdi, Mark Schellhase, Richard Bartle, Bob Carpenter, Joshua Hartshorne, Mark Liberman, Kepa Rodriguez, Olga Uryupina and Peng Yan. We also extend our thanks to the anonymous reviewers who provided extensive feedback on an earlier version of this article.

## REFERENCES

- ALBAKOUR, M.-D., KRUSCHWITZ, U., AND LUCAS, S. 2010. Sentence-Level attachment prediction. In *Proceedings of the 1st Information Retrieval Facility Conference*. Lecture Notes in Computer Science, vol. 6107. Springer, 6–19.
- ALONSO, O. AND MIZZARO, S. 2009. Can we get rid of trec assessors? Using mechanical turk for relevance assessment. In *Proceedings of the Workshop on the Future of Information Retrieval Evaluation, Collocated at Special Interest Group on Information Retrieval Conference (SIGIR)*.
- ALONSO, O., ROSE, D. E., AND STEWART, B. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum* 42, 2, 9–15.

<sup>59</sup>[http://www.infosolutionsgroup.com/2010\\\_PopCap\\\_Social\\\_Gaming\\\_Research\\\_Results.pdf](http://www.infosolutionsgroup.com/2010\_PopCap\_Social\_Gaming\_Research\_Results.pdf).

<sup>60</sup><http://www.lightspeedresearch.com/press-releases/it/%E2%80%99s-game-on-for-facebook-users>.

<sup>61</sup><http://www.digitalbuzzblog.com/2011-mobile-statistics-stats-facts-marketing-infographic>.

<sup>62</sup><http://www.frankwbaker.com/mediause.htm>.



- ARTSTEIN, R. AND POESIO, M. 2008. Inter-Coder agreement for computational linguistics. *Comput. Linguist.* 34, 4, 555–596.
- ATTARDI, G. AND THE GALOAP TEAM. 2010. Phratris. In *Proceedings of the INSEMTIVES'10 (Demo)*.
- BARONI, M., BERNARDINI, S., FERRARESI, A., AND ZANCHETTA, E. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Lang. Resour. Eval.* 3, 209–226.
- BIRD, S. AND LIBERMAN, M. 1999. Annotation graphs as a framework for multidimensional linguistic data analysis. In *Proceedings of the Workshop "Towards Standards and Tools for Discourse Tagging"*. Association for Computational Linguistics, 1–10.
- BROSCHETT, S., POESIO, M., PONZETTO, S.-P., RODRIGUEZ, K. J., ROMANO, L., URYUPINA, O., VERSLEY, Y., AND ZANOLI, R. 2010. Bart: A multilingual anaphora resolution system. In *Proceedings of the Semantic Evaluation Workshop (SEMEVAL)*.
- BURCHARDT, A., ERK, K., FRANK, A., KOWALSKI, A., PADO, S., AND PINKAL, M. 2009. Framenet for the semantic analysis of German: Annotation, representation and automation. In *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, H. C. Boas, Ed., Mouton De Gruyter.
- BURNARD, L. 2000. The british national corpus reference guide. Tech. rep., Oxford University Computing Services, Oxford, UK.
- CALLISON-BURCH, C. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Vol. 1. Association for Computational Linguistics, 286–295.
- CHAMBERLAIN, J., KRUSCHWITZ, U., AND POESIO, M. 2009a. Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In *Proceedings of the Joint Conference of the 47<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and the 4<sup>th</sup> International Joint Conference on Natural Language Processing Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources*.
- CHAMBERLAIN, J., KRUSCHWITZ, U., AND POESIO, M. 2012. Motivations for participation in socially networked collective intelligence systems. In *Proceedings of the Conference on Collective Intelligence (CI12)*.
- CHAMBERLAIN, J., POESIO, M., AND KRUSCHWITZ, U. 2008a. Addressing the resource bottleneck to create large-scale annotated texts. In *Proceedings of the Symposium on Semantics in Systems for Text Processing (STEP)*.
- CHAMBERLAIN, J., POESIO, M., AND KRUSCHWITZ, U. 2008b. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (ISemantics'08)*.
- CHAMBERLAIN, J., POESIO, M., AND KRUSCHWITZ, U. 2009b. A new life for a dead parrot: Incentive structures in the phrase detectives game. In *Proceedings of the WWW Workshop on Web Incentives (WEBCENTIVES'09)*.
- CHKLOVSKI, T. AND GIL, Y. 2005. Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *Proceedings of the 3<sup>rd</sup> International Conference on Knowledge Capture*. 35–42.
- CHKLOVSKI, T. 2005. Collecting paraphrase corpora from volunteer contributors. In *Proceedings of the 3<sup>rd</sup> International Conference on Knowledge Capture (K-CAP'05)*. ACM Press, New York, 115–120.
- CSOMAI, A. AND MIHALCEA, R. 2008. Linking documents to encyclopedic knowledge. *IEEE Intell. Syst.* 23, 5, 34–41.
- FENG, D., BESANA, S., AND ZAJAC, R. 2009. Acquiring high quality non-expert knowledge from on-demand workforce. In *Proceedings of the Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources*. Association for Computational Linguistics, 51–56.
- GARNHAM, A. 2001. *Mental Models and the Interpretation of Anaphora*. Psychology Press.
- HITZEMAN, J. AND POESIO, M. 1998. Long-Distance pronominalisation and global focus. In *Proceedings of the 36<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and 17<sup>th</sup> International Conference on Computational Linguistics*. Vol. 1.
- HLADKA, B., MIROVSKY, J., AND SCHLESINGER, P. 2009. Play the language: Play coreference. In *Proceedings of the Joint Conference of the 47<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and the 4<sup>th</sup> International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 209–212.
- HOBBS, J. R. 1978. Resolving pronoun references. *Lingua* 44, 311–338.
- HOVY, E., MARCUS, M., PALMER, M., RAMSHAW, L., AND WEISCHADEL, R. 2006. Ontonotes: The 90% solution. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 57–60.
- JOHNSON, N. L., RASMUSSEN, S., JOSLYN, C., ROCHA, L., SMITH, S., AND KANTOR, M. 1998. Symbiotic intelligence: Self-Organizing knowledge on distributed networks driven by human interaction. In *Proceedings of the 6<sup>th</sup> International Conference on Artificial Life*. MIT Press.

- KABADJOV, M. A. 2007. Task-Oriented evaluation of anaphora resolution. Ph.D. thesis, University of Essex, Colchester, UK.
- KAMP, H. AND REYLE, U. 1993. *From Discourse to Logic*. D. Reidel, Dordrecht.
- KAZAI, G. 2011. In search of quality in crowdsourcing for search engine evaluation. In *Proceedings of the 33<sup>rd</sup> European Conference on Information Retrieval (ECIR'11)*. Lecture Notes in Computer Science, vol. 6611. Springer, 165–176.
- KAZAI, G., MILIC-FRAYLING, N., AND COSTELLO, J. 2009. Towards methods for the collective gathering and quality control of relevance assessments. In *Proceedings of the 32<sup>nd</sup> International Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval (SIGIR'09)*. ACM Press, New York, 452–459.
- KOLLER, A., STRIEGNITZ, K., GARGETT, A., BYRON, D., CASSELL, J., DALE, R., MOORE, J., AND OBERLANDER, J. 2010. Report on the second nlg challenge on generating instructions in virtual environments (give-2). In *Proceedings of the 6<sup>th</sup> International Natural Language Generation Conference*.
- KOSTER, R. 2005. *A Theory of Fun for Game Design*. Paraglyph.
- KROTZSCH, M., VRANDEIC, D., VOLKEL, M., HALLER, H., AND STUDER, R. 2007. Semantic wikipedia. *J. Web Semantics* 5, 251–261.
- KRUSCHWITZ, U., CHAMBERLAIN, J., AND POESIO, M. 2009. (Linguistic) science through web collaboration in the ANAWIKI project. In *Proceedings of the International Conference on Web Science (WebSci'09)*.
- KUCERA, H. AND FRANCIS, W. N. 1967. *Computational Analysis of Present-Day American English*. Brown University Press.
- LESMO, L. AND LOMBARDO, V. 2002. Transformed subcategorization frames in chunk parsing. In *Proceedings of the 3<sup>rd</sup> International Conference on Language Resources and Evaluation*. 512–519.
- LIEBERMAN, H., SMITH, A. D., AND TEETERS, A. 2007. Common consensus: A web-based game for collecting commonsense goals. In *Proceedings of the Workshop on Common Sense and Intelligent User Interfaces held in Conjunction with the International Conference on Intelligent User Interfaces (IUI'07)*.
- MARCUS, M. P., MARCINKIEWICZ, M. A., AND SANTORINI, B. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.* 19, 2, 313–330.
- MARKEY, K. 2007. Twenty-Five years of end-user searching, Part 1: Research findings. *J. Amer. Soc. Inf. Sci. Technol.* 58, 8, 1071–1081.
- MASON, W. AND WATTS, D. J. 2010. Financial incentives and the “performance of crowds”. *Special Interest Group Knowl. Discov. Data Min. Explorations Newslett.* 11, 100–108.
- MINTZ, M., BILLS, S., SNOW, R., AND JURAFSKY, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and the 4<sup>th</sup> International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language*. 1003–1011.
- MITKOV, R. 2002. *Anaphora Resolution*. Longman.
- MROZINSKI, J., WHITTAKER, E., AND FURUI, S. 2008. Collecting a why-question corpus for development and evaluation of an automatic QA-system. In *Proceedings of the 36<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 443–451.
- NG, V. 2008. Unsupervised models for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- NIVRE, J. 2005. Dependency grammar and dependency parsing. Tech. rep., Vaxjo University.
- PETROV, S., BARRETT, L., THIBAU, R., AND KLEIN, D. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21<sup>st</sup> International Conference on Computational Linguistics*. Association for Computational Linguistics, 433–440.
- POESIO, M. 2004a. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the Association for Computational Linguistics Workshop on Discourse Annotation*.
- POESIO, M. 2004b. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.
- POESIO, M. AND ARTSTEIN, R. 2008. Anaphoric annotation in the arrau corpus. In *Proceedings of the 6<sup>th</sup> International Conference on Language Resources and Evaluation*.
- POESIO, M., DIEWALD, N., STUHRENBERG, M., CHAMBERLAIN, J., JETTKA, D., GOECKE, D., AND KRUSCHWITZ, U. 2011a. Markup infrastructure for the anaphoric bank: Supporting web collaboration. In *Modeling, Learning, and Processing of Text Technological Data Structures*, A. Mehler, K.-U. Kuhnberger, H. Lobin, H. Lungen, A. Storrer, and A. Witt, Eds., Studies in Computational Intelligence, vol. 370, Springer, 175–195.

- POESIO, M., KRUSCHWITZ, U., AND CHAMBERLAIN, J. 2008. ANAWIKI: Creating anaphorically annotated resources through Web cooperation. In *Proceedings of the International Conference on Language Resources and Evaluation*.
- POESIO, M., STUCKARDT, R., AND VERSLEY, Y. 2011b. *Anaphora Resolution: Algorithms, Resources and Applications*. Springer.
- POESIO, M., STURT, P., ARSTEIN, R., AND FILIK, R. 2006. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes* 42, 2, 157–175.
- POESIO, M. AND VIEIRA, R. 1998. A corpus-based investigation of definite description use. *Comput. Linguist.* 24, 2, 183–216.
- PONZETTO, S. AND STRUBE, M. 2007. Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.* 30, 181–212.
- PRADHAN, S., RAMSHAW, L., MARCUS, M., PALMER, M., WEISCHEDEL, R., AND XUE, N. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the 15<sup>th</sup> Computational Natural Language Learning Conference (CoNLL)*. Association for Computational Linguistics, 1–27.
- PRADHAN, S. S., RAMSHAW, L., WEISCHEDEL, R., MACBRIDE, J., AND MICCIULLA, L. 2007. Unrestricted coreference: Identifying entities and events in ontonotes. In *Proceedings of the International Conference on Semantic Computing*.
- PRINCE, E. F. 1992. The ZPG letter: Subjects, definiteness, and information status. In *Discourse Description: Diverse Analyses of a Fund-Raising Text*, S. Thompson and W. Mann, Eds., John Benjamins, 295–325.
- RAFELSBERGER, W. AND SCHARL, A. 2009. Games with a purpose for social networking platforms. In *Proceedings of the 20th ACM Conference on Hypertext and hypermedia*. ACM Press, New York, 193–198.
- RECASENS, M., MARQUEZ, L., SAPENA, E., MARTI, M. A., TAULE, M., HOSTE, V., POESIO, M., AND VERSLEY, Y. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the Semantic Evaluation Workshop (SEMEVAL)*.
- ROBALDO, L., POESIO, M., DUCCESCHI, L., CHAMBERLAIN, J., AND KRUSCHWITZ, U. 2011. Italian anaphoric annotation with the phrase detectives game-with-a-purpose. In *Proceedings of the 12<sup>th</sup> Congress of the Italian Association for Artificial Intelligence*. Lecture Notes in Computer Science, vol. 6934. Springer, 407–412.
- SETTLES, B. 2009. Active learning literature survey. Tech. rep. 1648, Department of Computer Science, University of Wisconsin at Madison.
- SINGH, P. 2002. The public acquisition of commonsense knowledge. In *Proceedings of the AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*.
- SJORPAES, K. AND HEPP, M. 2008. Games with a purpose for the semantic web. *IEEE Intell. Syst.* 23, 3, 50–60.
- SMADJA, F. 2009. Mixing financial, social and fun incentives for social voting. In *World Wide Web Internet and Web Information Systems*.
- SNOW, R., O’CONNOR, B., JURAFSKY, D., AND NG, A. Y. 2008. Cheap and fast—But is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP’08)*. Association for Computational Linguistics, 254–263.
- SOON, W. M., LIM, D. C. Y., AND NG, H. T. 2001. A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.* 27, 4.
- STUHRENBERG, M. AND GOECKE, D. 2008. SGF—An integrated model for multiple annotations and its application in a linguistic domain. In *Balisage: The Markup Conference*.
- STUHRENBERG, M., GOECKE, D., DIEWALD, N., MEHLER, A., AND CRAMER, I. 2007. Web-Based annotation of anaphoric relations and lexical chains. In *Proceedings of the Association for Computational Linguistics, Linguistic Annotation Workshop*. 140–147.
- TANG, J. AND SANDERSON, M. 2010. Evaluation and user preference study on spatial diversity. In *Proceedings of the European Conference on IR Research (ECIR)*. Lecture Notes in Computer Science, vol. 5993, Springer, 179–190.
- VIEIRA, R. AND POESIO, M. 2000. An empirically based system for processing definite descriptions. *Comput. Linguist.* 26, 539–593.
- VILAIN, M., BURGER, J., ABERDEEN, J., CONNOLLY, D., AND HIRSCHMAN, L. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6<sup>th</sup> Message Understanding Conference*. 45–52.
- VLACHOS, A. 2006. Active annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining, Collocated at the European Chapter of the Association for Computational Linguistics*.
- VON AHN, L. 2006. Games with a purpose. *Comput.* 39, 6, 92–94.
- VON AHN, L. AND DABBISH, L. 2004. Labeling images with a computer game. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM Press, New York, 319–326.

- VON AHN, L. AND DABBISH, L. 2008. Designing games with a purpose. *Comm. ACM* 8, 58–67.
- VON AHN, L., LIU, R., AND BLUM, M. 2006. Peekaboom: A game for locating objects in images. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM Press, New York, 55–64.
- YANG, H. AND LAI, C. 2010. Motivations of wikipedia content contributors. *Comput. Hum. Behav.* 26, 6, 1377–1383.
- ZAENEN, A. 2006. Mark-Up barking up the wrong tree. *Comput. Linguist.* 32, 4, 577–580.

Received June 2011; revised January 2012; accepted April 2012