




Massimo Poesio

(Joint with R. Bartle, J. Chamberlain, C. Madge, U. Kruschwitz, S. Paun)

EXPLORING ANAPHORIC AMBIGUITY USING GAMES-WITH- A-PURPOSE: THE DALI PROJECT




Disagreements and Language Interpretation (DALI)

- A 5-year, €2.5M project on using games-with-a-purpose and Bayesian models of annotation to study ambiguity in anaphora
 - A collaboration between Essex, LDC, and Columbia
 - Funded by the European Research Council (ERC)
- 



Outline

- **Corpus creation and ambiguity**
 - Collective multiple judgments through crowdsourcing: Phrase Detectives
 - DALI: new games
 - DALI: analysis
- 

Anaphora (AKA coreference)

So she [Alice] was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly **a White Rabbit with pink eyes** ran close by her.

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); but when **the Rabbit** actually **TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET**, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after **it**, and fortunately was just in time to see **it** pop down a large rabbit-hole under the hedge.

Building NLP models from annotated corpora

- Use **TRADITIONAL CORPUS ANNOTATION / CROWDSOURCING** to create a **GOLD STANDARD** that can be used to train supervised models for various tasks
- This is done by collecting multiple annotations (typically 2-5) and going through **RECONCILIATION** whenever there are multiple interpretations
- **DISAGREEMENT** between coders (measured using coefficients of agreement such as κ or α) viewed as a serious problem, to be addressed by revising the coding scheme or training coders to death
- Yet there are very many types of NLP annotation where **DISAGREEMENT IS RIFE** (wordsense, sentiment, discourse)



Crowdsourcing in NLP

- Crowdsourcing in NLP has been used as a cheap alternative to the traditional approach to annotation
- The overwhelming concern has been to develop alternative quality control practices to obtain a gold standard comparable to those obtained with traditional high-quality annotation

The problem of ambiguity

15.12 M: we're gonna take the engine E3

15.13 : and shove it over to Corning

15.14 : hook [it] up to [the tanker car]

15.15 : _and_

15.16 : send **it** back to Elmira



(from the TRAINS-91 dialogues collected at the University of Rochester)

Ambiguity: What antecedent?

(Poesio & Vieira, 1998)

About 160 workers at *a factory* that made paper for the Kent filters were exposed to asbestos in the 1950s.

Areas of the factory were particularly dusty where the crocidolite was used.

Workers dumped large burlap sacks of the imported material into a huge bin, poured in cotton and acetate fibers and mechanically mixed the dry fibers in a process used to make filters.

Workers described "clouds of blue dust" that hung over *parts of the factory*,

even though exhaust fans ventilated the area.

Ambiguity: DISCOURSE NEW or DISCOURSE OLD?

(Poesio, 2004)

What is in **your cream**

Dermovate Cream is one of a group of medicines called topical steroids.

"Topical" means they are put on the skin. Topical steroids reduce the redness and itchiness of certain skin problems.

AMBIGUITY: EXPLETIVES

'I beg your pardon!' said the Mouse, frowning, but very politely: 'Did you speak?'

'Not I!' said the Lory hastily.


'I thought you did,' said the Mouse. '--I proceed. "Edwin and Morcar, the earls of Mercia and Northumbria, declared for him: and even Stigand, the patriotic archbishop of Canterbury, found **it** advisable--"

'Found **WHAT**?' said the Duck.

'Found **IT**,' the Mouse replied rather crossly: 'of course you know what "it" means.'



Ambiguity in Anaphora: the ARRAU project

- As part of the EPSRC-funded ARRAU project (2004-07), we carried out a number of studies in which we asked numerous annotators (~ 20) to annotate the interpretation of referring expressions, finding systematic ambiguities with all three types of decisions (Poesio & Artstein, 2005)
- 

Implicit and Explicit Ambiguity

- The coding scheme for ARRAU allows coders to mark an expression as ambiguous at multiple levels:
 - Between referential and non/referential
 - Between DN and DO
 - Between different types of antecedents
- BUT: most annotators can't see this ...

The picture of ambiguity emerging from ARRAU

19.10: we need to get the bananas to Corning by 3

19.11: uh

19.12: *maybe* it's gonna be faster if we

19.13: send E1

19.14: E1's boxcar picks up at Dansville

19.15: instead of going back to Avon

19.16: have it go on to Corning

19.17: uh pick up the tanker get the oranges send them to Elmira

19.18: cause that's gonna be the longest thing

Key: Full agreement One outlier Implicit Explicit

More evidence of disagreement raising from ambiguity


- For anaphora
 - **Versley 2008**: Analysis of disagreements among annotators in the Tüba/DZ corpus
 - Formulation of the **DOT-OBJECT hypothesis**
 - **Recasens et al 2011**: Analysis of disagreements among annotators in (a subset of) the ANCORA and the ONTONOTES corpus
 - The **NEAR-IDENTITY hypothesis**
- **Wordsense: Passonneau et al, 2012**
 - Analysis of disagreements among annotators in the wordsense annotation of the MASC corpus
 - Up to 60% disagreement with verbs like *help*
- **POS tagging: Plank et al, 2014**

Exploring (anaphoric) ambiguity

- Empirically, the only way to see which expressions get multiple annotations is by having > 10 coders and maintain multiple annotations
- So, to investigate the phenomenon, one would need to collect many more judgments than one could through a traditional annotation experiment, as we did in ARRAU
- But how can one collect so many judgments about this much data?
- The solution: CROWDSOURCING



Outline

- **Corpus creation and ambiguity**
 - **Collective multiple judgments through crowdsourcing: Phrase Detectives**
 - DALI: new games
 - DALI: analysis
- 

Approaches to crowdsourcing

- Incentivized through money: microtask crowdsourcing
 - (As in Amazon Mechanical Turk)
- Scientifically / culturally motivated
 - As in Wikipedia / Galaxy Zoo
- Entertainment as the incentive: GAMES-WITH-A-PURPOSE (von Ahn, 2006)

Games-with-a-purpose: ESP

The screenshot shows the ESP Game interface in a Firefox browser. The browser's address bar displays the URL: <http://www.gwap.com/espgame-o/espgame.html?4341103&150041&guest150041&0>. The browser's menu bar includes File, Edit, View, History, Bookmarks, Tools, Window, and Help. The browser's status bar shows the time as 6:12 PM on Friday.

The game interface is displayed in a green-themed window. At the top, it shows the score (300) and time (2:16). The main question is "What do you see?" with a photo of sheep in a barn. The taboo words are "god" and "dog". The guesses section is empty. A "BONUS!" section offers 5,000 points. A leaderboard on the left shows the top 10 players:

Rank	Player	Points
1	heliobalione	32 K
2	ozcnobius	27 K
3	aks	14 K
4	scb	13 K
5	guest150020	12 K
6	byakugan333	12 K
7	fnyk	12 K
8	mdkairiys	7,000
9	guest150001	7,000
10	tgellot	6,500

The browser's search bar shows the search term "glove". The browser's status bar shows "Done". The browser's taskbar shows various application icons.

ESP results

- In the 4 months between August 9th 2003 and December 10th 2003
 - 13630 players
 - 1.2 million labels for 293,760 images
 - 80% of players played more than once
- By 2008:
 - 200,000 players
 - 50 million labels
- Number of labels x item is one of the parameters of the game, but on average, in the order of 20-30

Phrase Detectives

anawiki.essex.ac.uk/phrasedetectives/index.php

PHRASE DETECTIVES

LOGIN

USERNAME

PASSWORD

Login

Not joined yet?
[Register here.](#)

Like 37 people like this. Be the first of your friends

Instructions

FAQ

SHARE THIS

Welcome to Phrase Detectives

Lovers of literature, grammar and language, this is the place where you can work together to improve future generations of technology. By indicating **relationships** between **words** and **phrases** you will help to create a resource that is rich in linguistic information. Simply register a username and password and you can get started.

[Start here](#)

321 docs completed

The most recent was **Rhubarb Triangle** (Wikipedia) completed by **sassie25uk** on 22 Jan 2011

[See all](#)

Quick instructions

You must search for **relationships** between **words** and **phrases** in a piece of text.

1) NAME THE CULPRIT

You will be given a word, phrase and

Congratulations the winners in December 2010:

JMS (1 Dec), **JCS** (2 Dec), **JMS** (3 Dec), **JRS** (4 Dec), **papillon** (5 Dec), **mhbball** (6 Dec), **papillon** (7 Dec), **prufrax** (8 Dec), **JRS** (9 Dec), **JMS** (10 Dec), **papillon** (11 Dec), **JMS** (12 Dec), **smazzu** (13 Dec), **bakuzen** (14 Dec), **andre** (15 Dec), **Meuhcoin** (16 Dec), **bakers_man** (17 Dec), **JMS** (18 Dec), **bochi** (19 Dec), **JCS** (20 Dec), **papillon** (21 Dec), **JRS** (22 Dec), **bochi** (23 Dec), **sassie25uk** (24 Dec), **effani** (25 Dec), **Annie** (26 Dec), **Eric-the-red** (27 Dec), **papillon** (28 Dec), **domma17** (29 Dec), **ecclescake** (30 Dec) and **lethe** (31 Dec).

[Play now](#)

Detective's Bulletin

30 Nov 10

Plans are coming together at HQ for 2 new versions of the game to be released on Facebook and iPhone. To become a beta tester please get in [contact](#).

01 Nov 10

Here are a few game stats for you number

TOPSCORES

THIS WEEK
 papillon 1417

THIS MONTH
 papillon 3042

LEADERBOARD

WEEK	MONTH	ALLTIME
papillon		1417
norma		904
JMS		842
JRS		501
kazzade		310
darraghsdream		296
JCS		203
julie3164		157
domma17		103
gully		87
johnnickel		72
poppyseed		66
prufrax		61
KULIKOV		55
rcorreia		45
Grammar		42
axnicho		39
VB		38
crazyhorse		27
lvio.robaldo		27

MOST RECENT
 Kazumi Totaka (Wikipedia)
 submitted by darraghsdream

The game

- **Find The Culprit** (Annotation)
User must identify the closest antecedent of a markable if it is anaphoric
- **Detectives Conference** (Validation)
User must agree/disagree with a coreference relation entered by another user



Find the Culprit (aka Annotation Mode)

The Count of Monte Cristo

Having arrived before the Pont du Gard, the horse stopped, but whether for his own pleasure or that of his rider would have been difficult to say. However that might have been, the priest, dismounting, led his steed by the bridle in search of some place to which he could secure him. Availing himself of a handle that projected from a half-fallen door, he tied the animal safely and having drawn a red cotton handkerchief, from his pocket, wiped away the perspiration that streamed from his brow, then, advancing to the door, struck thrice with the end of his iron-shod stick.



Not mentioned before!



Skip this one

Found it!



Find the Culprit (aka Annotation Mode)

The Count of Monte Cristo

Having arrived before the Pont du Gard, the horse stopped, but whether for his own pleasure or that of his rider would have been difficult to say. However that might have been, the priest, dismounting, led his steed by the bridle in search of some place to which he could secure him. Availing himself of a handle that projected from a half-fallen door, he tied the animal safely and having drawn a red cotton handkerchief, from **his** pocket, wiped away the perspiration that streamed from **his** brow, then, advancing to the door, struck thrice with the end of his iron-shod stick.



Not mentioned before!



Skip this one

Found it!



Detectives Conference (aka Validation Mode)

The Nurse and the Wolf - Aesop

"Be quiet now," said an old Nurse to a child sitting on her lap. "If you make that noise again I will throw you to the Wolf."

Now it chanced that a Wolf was passing close under **the window** as this was said. So he crouched down by the side of the house and waited. "I am in good luck to-day," thought he. "It is sure to cry soon, and a daintier morsel I haven't had for many a long day." So he waited, and he waited, and he waited, till at last the child began to cry, and the Wolf came forward before **the window**, and looked up to the Nurse, wagging his tail.

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



Disagree



Agree

Facebook Phrase Detectives (2013)

Facebook Phrase Detectives

Massimo Poesio
Learner
14391
Next level

YOUR CASE
Andean Condor
(Wikipedia)
445 tasks remaining
★ Perfect!

Home
Player HQ
Settings
Instructions
FAQ
About

YOUR TEAM
Post for friends

Andean Condor (Wikipedia)

In the male, the head is crowned with a dark red caruncle or comb, while the skin of his neck lies in folds, forming a wattle. The skin of the head and neck is capable of flushing noticeably in response to emotional state, which serves to communicate between individuals. Juveniles have a grayish-brown general coloration, blackish head and neck skin, and a brown ruff.

The middle toe is greatly elongated, and the hind one is only slightly developed, while the talons of all the toes are comparatively straight and blunt. The feet are thus more adapted to walking, and are of little use as weapons or organs of prehension as in birds of prey and Old World vultures. The beak is hooked, and adapted to tear rotting meat. The irises of the male are brown, while those of the female are deep red. The eyelids lack eyelashes. Contrary to the usual rule among birds of prey, the female is smaller than the male.

The Andean Condor is found in South America in the Andes. In the north, its range begins in Venezuela and Colombia, where **it** is extremely rare, then continues south along the Andes in Ecuador, Peru, and Chile, through Bolivia and western Argentina to the Tierra del Fuego.

NAME THE CULPRIT

Has the phrase shown in orange been mentioned before in this text or does it not refer to anything? Select the closest phrase(s) within the text if it has been mentioned before and click "Done".

Not mentioned before Does not refer Done

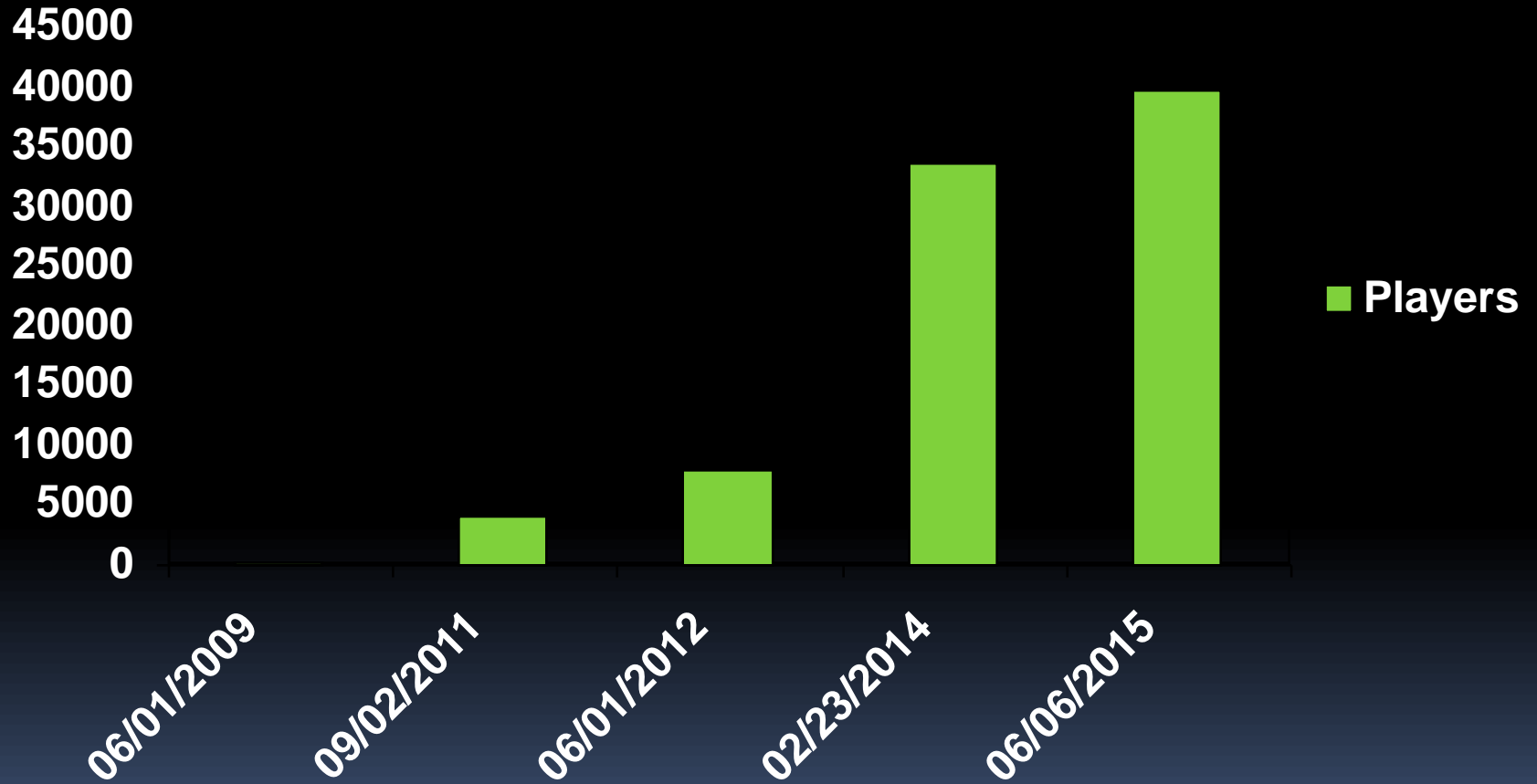
Comment on this phrase
Skip this one
Skip - closest phrase can't be selected
Skip - closest phrase is no longer visible
Skip - error in the text

Results

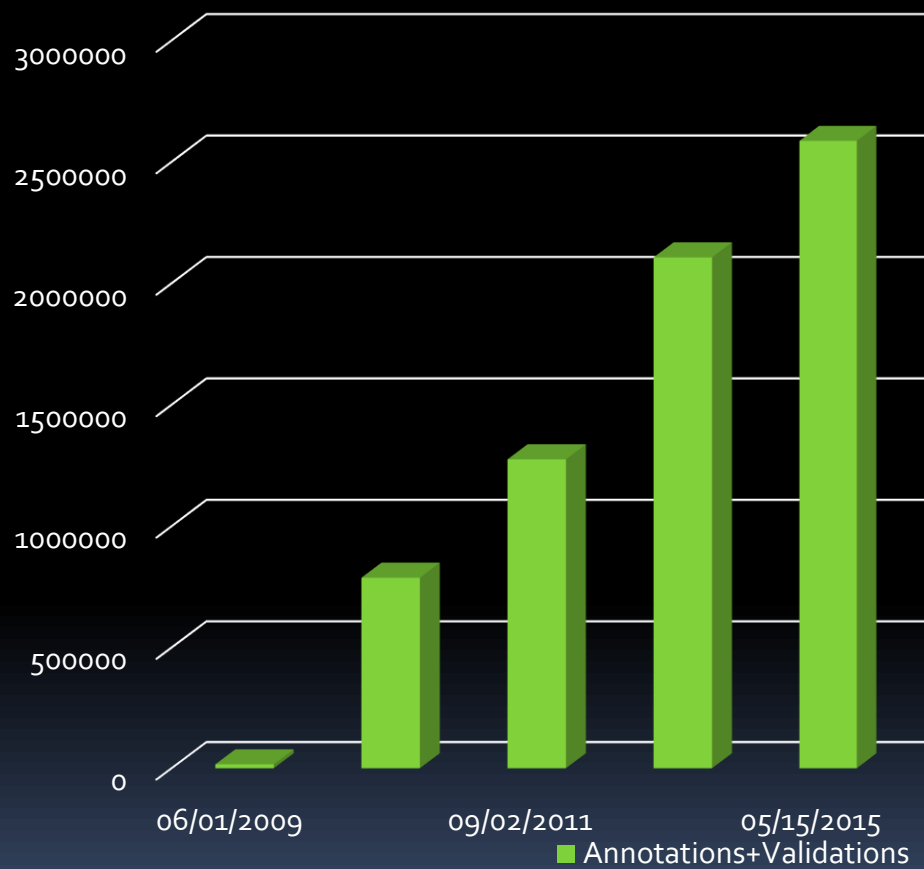
- Quantity
 - Number of users
 - Amount of annotated data
- The corpus
- Multiplicity of interpretations



Number of Players



Number of judgments



The Phrase Detectives Corpus

- Data:
 - 1.2M words total, of which around 330K totally annotated
 - About 50% Wikipedia pages, 50% fiction
- Markable scheme:
 - Around 25 judgments per markable on average
 - Judgments:
 - NR/DN/DO
 - For DO, antecedent
- Phrase Detective 1.0 just announced, to be distributed via LDC

Ambiguity in the Phrase Detectives Data

- In 2012: 63009 completely annotated markables
 - Exactly 1 interpretation: 23479
 - Discourse New (DN): 23138
 - Discourse Old (DO): 322
 - Non Referring (NR): 19
 - With only 1 relation with score > 0: 13772
 - DN: 9194
 - DO: 4391
 - NR: 175
 - In total, ~ 40% of markables have more than one interpretation with score > 0
 - Hand-analysis of a sample (Chamberlain, 2015)
 - 30% of the cases in that sample had more than one non-spurious interpretation

Ambiguity: REFERRING or NON REFERRING?

'I beg your pardon!' said the Mouse, frowning, but very politely: 'Did you speak?'

'Not I!' said the Lory hastily.

'I thought you did,' said the Mouse. '--I proceed. "Edwin and Morcar, the earls of Mercia and Northumbria, declared for him: and even Stigand, the patriotic archbishop of Canterbury, found **it** advisable--"'

'Found **WHAT?**' said the Duck.

'Found **IT,**' the Mouse replied rather crossly: 'of course you know what "it" means.'


Ambiguity: DN / DO

The rooms were carefully examined, and results all pointed to an abominable crime. The front room was plainly furnished as a sitting-room and led into a small bedroom, which looked out upon the back of one of the wharves. Between the wharf and **the bedroom window** is a narrow strip, which is dry at low tide but is covered at high tide with at least four and a half feet of water. The bedroom window was a broad one and opened from below. On examination traces of blood were to be seen upon the windowsill, and several scattered drops were visible upon the wooden floor of the bedroom. Thrust away behind a curtain in the front room were all the clothes of Mr. Neville St. Clair, with the exception of his coat. His boots, his socks, his hat, and his watch -- all were there. There were no signs of violence upon any of these garments, and there were no other traces of Mr. Neville St. Clair. Out of **the window** he must apparently have gone






Outline

- **Corpus creation and ambiguity**
 - Collective multiple judgments through crowdsourcing: Phrase Detectives
 - **DALI: new games**
 - DALI: analysis
- 



The DALI project

1. Develop the GWAP approach to collecting data for anaphora
 2. Developing Bayesian annotation methods to analyze the data
 3. Develop models trained directly over multiple judgments data instead of producing a gold standard
 4. Develop an account of the interpretation of ambiguous anaphoric expressions building on Recasens et al 2011
- 



Beyond PD

- Phrase Detectives has been reasonably successful, and already allowed us to collect a large amount of data, but we're not going to be able to annotate 100M+ words through it
 - Not enough of a game
 - Humans still need to be involved in several behind-the-scenes activities
- We are also looking for new ways to gain visibility
 - We see the collaboration with LDC on NIEUW and being part of a 'GWAP-for-CL' portal as strategic

`New generation' GWAPS for CL

- Some more recent GWAPs have demonstrated that it is possible to design more entertaining games for CL, as well
- In particular, for collecting lexical resources
 - Jeux de Mots (Mathieu Lafourcade)
 - PuzzleRacer / Kaboom! (Jurgens & Navigli, TACL 2014)
- But also e.g., for Sentiment Analysis

Puzzle Racer



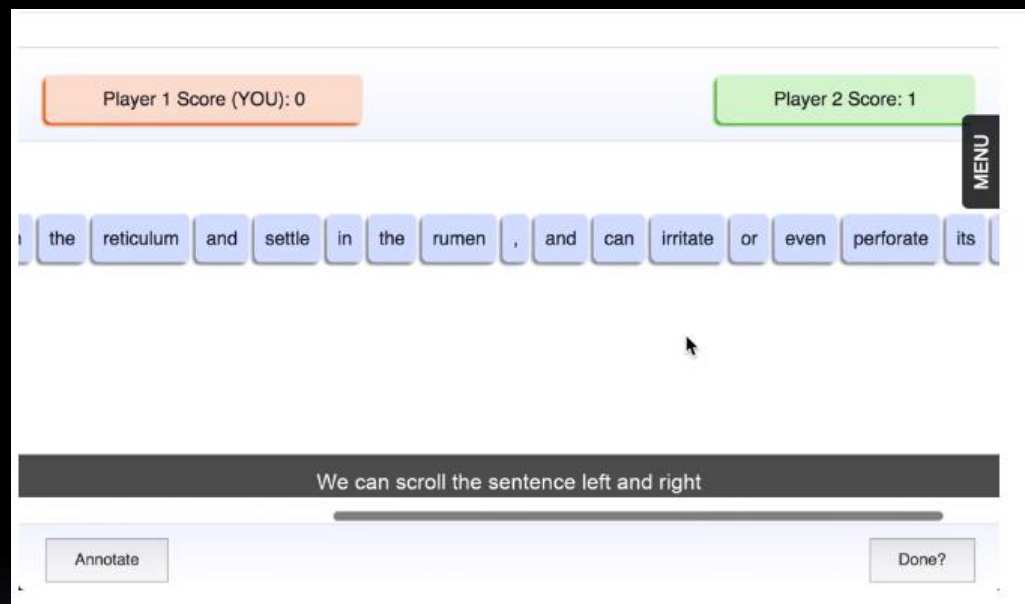
Gamify more aspects of the task

- Designer involvement is still required in PD to
 - Prepare the input to the game by correcting the output of the pipeline
 - Deal with comments
- We intend to develop games to remove these bottlenecks: a GAMIFIED PIPELINE

TileAttack! (Madge et al)

One such game is being developed to fix the input to the games

A first version has recently been tested



<http://tileattack.com/>

TileAttack: the game

The screenshot displays the TileAttack game interface. At the top, a timer shows 00:02:48. On the left, the score for Player 1 (YOU) is 3, and on the right, the score for Player 2 is 6. A vertical 'MENU' button is located on the right side. The main area contains a word search puzzle with the following words: 'The ray and disc petals are both yellow , and the fruits are brown .' The words 'The' and 'ray' are highlighted in orange, indicating they have been found. The words 'the' and 'fruits' are highlighted in green, indicating they have been found. At the bottom, there is a 'Done?' button and a notification that says 'Your opponent has finished, have you?'. An 'Annotate' button is also visible at the bottom left.

Player 1 Score (YOU): 3

00:02:48

Player 2 Score: 6

MENU

The ray and disc petals are both yellow , and the fruits are brown .

The ray

the fruits

Your opponent has finished, have you?

Annotate

Done?

End of game

Round Over - Sorry, you lost

Matches

the

fruits

Continue

Scoreboard

mp


Position: 6 Score: 275 Wins/Losses: 3/8

Position	Player	Score	Wins/Losses
1	chris	1006	28/30
2	TheUdo	796	8/27
3	Lex	505	16/7
4	tom	320	8/13
5	jon	288	3/8
6	mp	275	3/8
7	Richard	222	1/13
8	john m	212	5/10
9	liujish931	173	7/4
10	CatLinux	162	2/3
11	john_b	150	5/2
12	SteveH	120	4/1




TileAttack! In action

<https://www.youtube.com/watch?v=fcmrsPkiMvA&feature=youtu.be>





Outline

- **Corpus creation and ambiguity**
 - Collective multiple judgments through crowdsourcing: Phrase Detectives
 - DALI: new games
 - **DALI: analysis**
- 



Analyzing multiple judgments on a large scale

- Poesio et al 2006, Versley 2008, Recasens et al 2011, ourselves all analyzed a small sample of the annotations by hand
- Next challenge: analyze this multiplicity of judgments to distinguish real readings from noise on a large scale
- This requires using AUTOMATIC methods

Bayesian models of annotation


- The problem of reaching a conclusion on the basis of judgments by separate experts that may often be in disagreement is a longstanding one in epidemiology
- A number of techniques developed to analyze these data
- More recently, BAYESIAN MODELS OF ANNOTATION have been proposed:
 - Dawid and Skene 1979 (also used by Passonneau & Carpenter)
 - Latent Annotation model (Uebersax 1994)
 - Carpenter (2008)
 - Raykar et al 2010
 - Hovy et al, 2013

Bayesian Models of Annotation

- The probabilistic model specifies the probability of a particular label on the basis of PARAMETERS specifying the behavior of the annotators, the prevalence of the labels, etc
- In Bayesian models, these parameters are specified in terms of PROBABILITY DISTRIBUTIONS



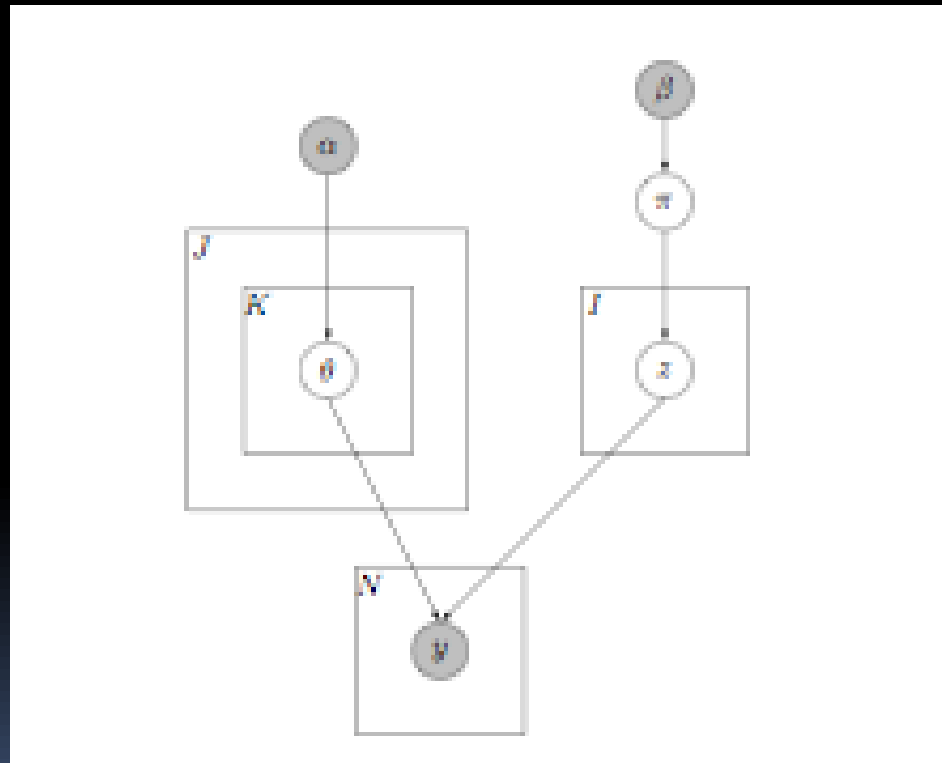
A GENERATIVE MODEL OF THE ANNOTATION TASK

- What all of these models do is to provide an EXPLICIT PROBABILISTIC MODEL of the observations in terms of annotators, labels, and items
- 

DAWID AND SKENE 1979

- Model consists of likelihood for
 1. annotations (labels from annotators)
 2. categories (true labels) for items given
 3. annotator accuracies and biases
 4. prevalence of labels
- Frequentists estimate 2–4 given 1
- Optional regularization of estimates (for 3 and 4)

A GRAPHICAL VIEW OF THE MODEL



THE PROBABILISTIC MODEL OF A GIVEN LABEL

$$\begin{aligned} p(z_i|y_i, \theta, \pi) &\propto p(z_i|\pi) p(y_i|z_i, \theta) \\ &= \pi_{z_i} \prod_{j=1}^J \theta_{j, z_i, y_{i,j}} \end{aligned}$$

DALI WP 3/4: Raykar et al 2010

- Propose a Bayesian model that simultaneously ESTIMATES THE GROUND TRUTH from noisy labels, produces an ASSESSMENT OF THE ANNOTATORS, and **LEARNS A CLASSIFIER**
 - Based on logistic regression

Conclusions

- Phrase Detectives shows that GWAPs are a promising approach to collect data for Computational Linguistics
 - In particular when multiple interpretations are of interest
- But much is still to be done in terms of
 - Developing more entertaining games
 - Analyzing the data
- We view the collaboration with LDC as strategic to attract players / deliver the data widely

The DALI Team (so far)



Richard Bartle



Jon Chamberlain



Udo Kruschwitz

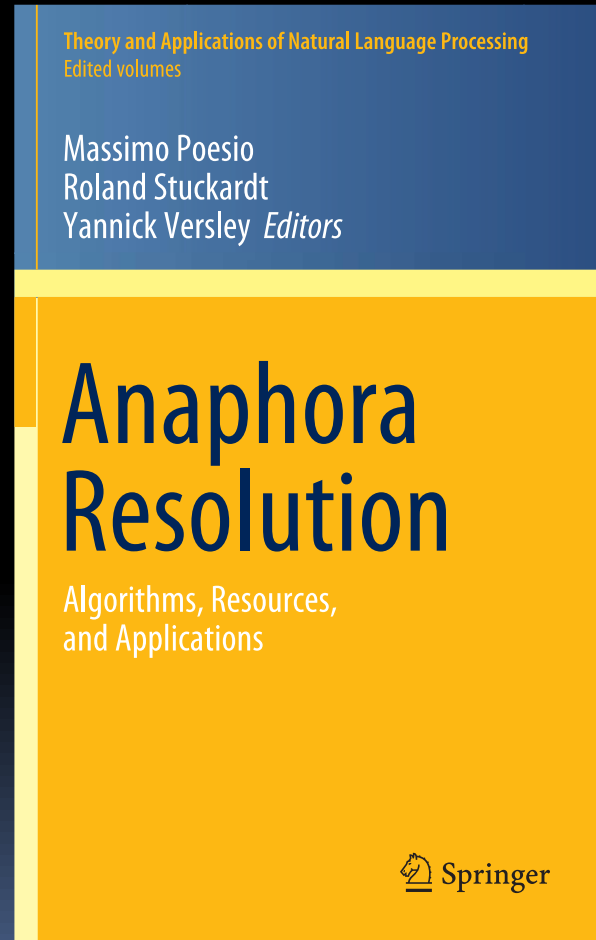


Chris Madge



Silviu Paun

Shameless plug #147




References

- M. Poesio, R. Stuckardt and Y. Versley (eds), 2016. *Anaphora Resolution*, Springer.
- M. Poesio, J. Chamberlain, U. Kruschwitz, 2013. Phrase Detectives, *ACM Transactions on Intelligent Interactive Systems (TIIS)*
- J Chamberlain, 2016. *Using a Validation Approach for Harnessing Collective Intelligence on Social Networks*, Uni Essex PhD



AGREEMENT STUDIES

- The aspects of anaphoric information that can be reliably annotated have been identified through a series of agreement studies with different degrees of formality (Hirschman et al., 1995; Poesio & Vieira, 1998; Poesio & Arstein, 2005; Mueller, 2007)
- 

Agreement on annotation

- Crucial requirement for the corpus to be of any use, is to make sure that annotation is RELIABLE (i.e., two different annotators are likely to mark in the same way)
- A number of COEFFICIENTS OF AGREEMENT developed to study reliability (Krippendorff, 2004; Artstein & Poesio, 2008)
- METHODOLOGY now well established*
- Agreement more difficult the more complex the judgments asked of the annotators
 - E.g., on givenness status
- The development of the annotation likely to follow a develop / test / redesign test
 - Task may have to be simplified

* Except that coefficients of agreement difficult to interpret

FOOD FOR THOUGHT: NO ANTECEDENTS

'Well!' thought Alice to herself, 'after such a fall as this, I shall think nothing of tumbling down stairs! How brave **they**'ll all think me at home! Why, I wouldn't say anything about it, even if I fell off the top of the house!' (Which was very likely true.)

Extremely prevalent: 30% of zero anaphors in Japanese of this type (Iida and Poesio, 2011)