Massimo Poesio

DISAGREEMENTS IN ANAPHORIC ANNOTATION

http://www.dali-ambiguity.org

Disagreements in anaphora (and other aspects of language interpretation)

Anaphora (AKA coreference)

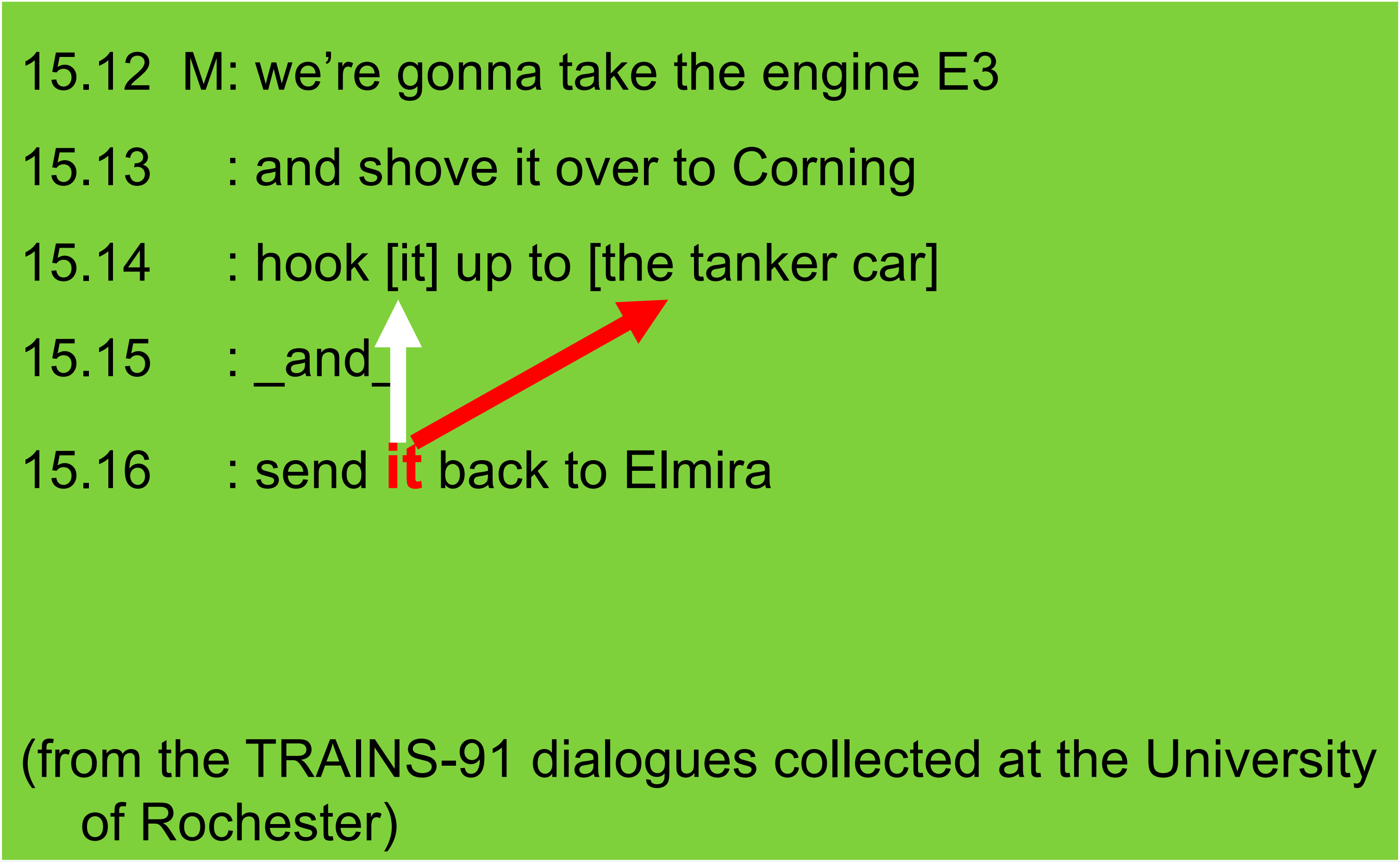
So she [Alice] was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly **a White Rabbit with pink eyes** ran close by her.

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)

Ambiguity in anaphora



15.12 M: we're gonna take the engine E3

15.13 : and shove it over to Corning

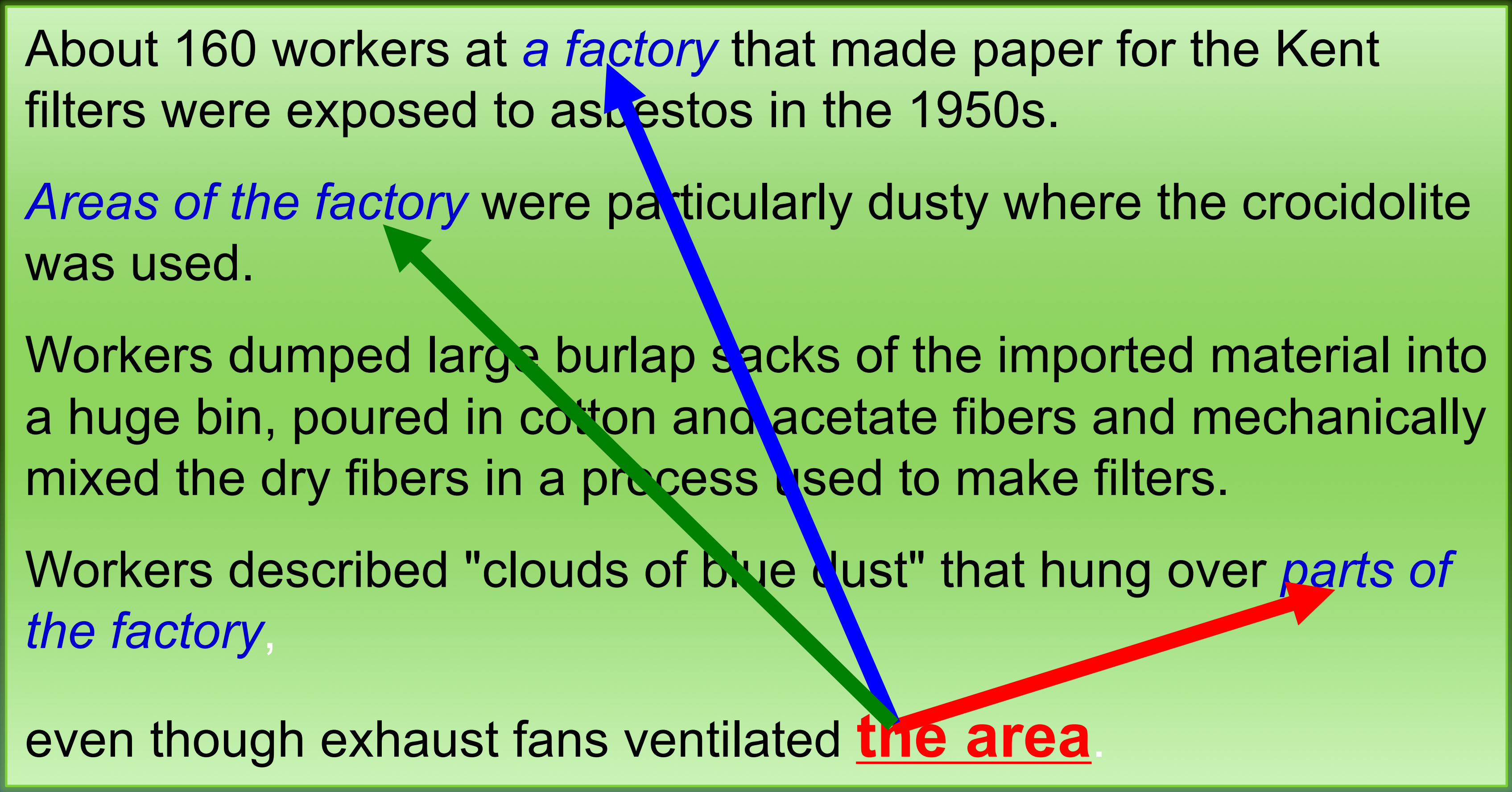
15.14 : hook [it] up to [the tanker car]

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.'

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

Facets (Versley, 2008)

As a lawyer in Boston, [1 John Travolta] sues two businesses that he holds responsible for eight children having died of leukemia.

At first, [2 the calculating career lawyer] only scents the high amount of compensation (. . .).

A court drama, environmental thriller and great actors' cinema, in which [3 Travolta] and his antagonist Robert Duvall reach top form.

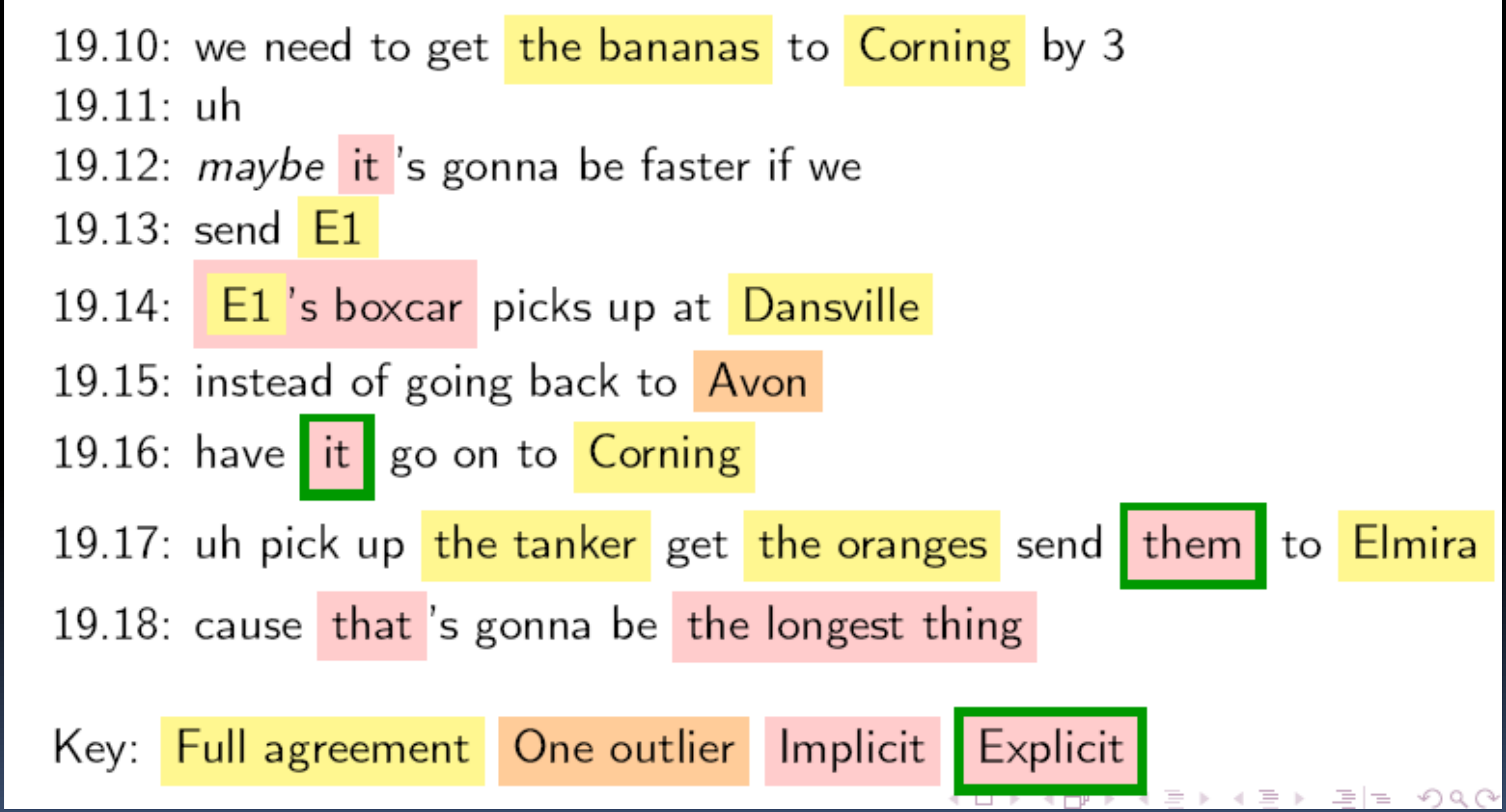
Near-identity cases (Recasens et al, 2011)

“[Your father]$_a$ was the greatest, but [he] was also one of us,” commented an anonymous old lady while she was shaking Alessandro’s hand—[Gassman]$_a$ ’s best known son.
“I will miss **[the actor]**$_{a1}$, but I will be lacking [my father]$_a$ especially,” he said.

“On homecoming night [Postville] feels like Hometown, USA . . . For those who prefer [the old Postville], Mayor John Hyman has a simple answer.

Collecting the data

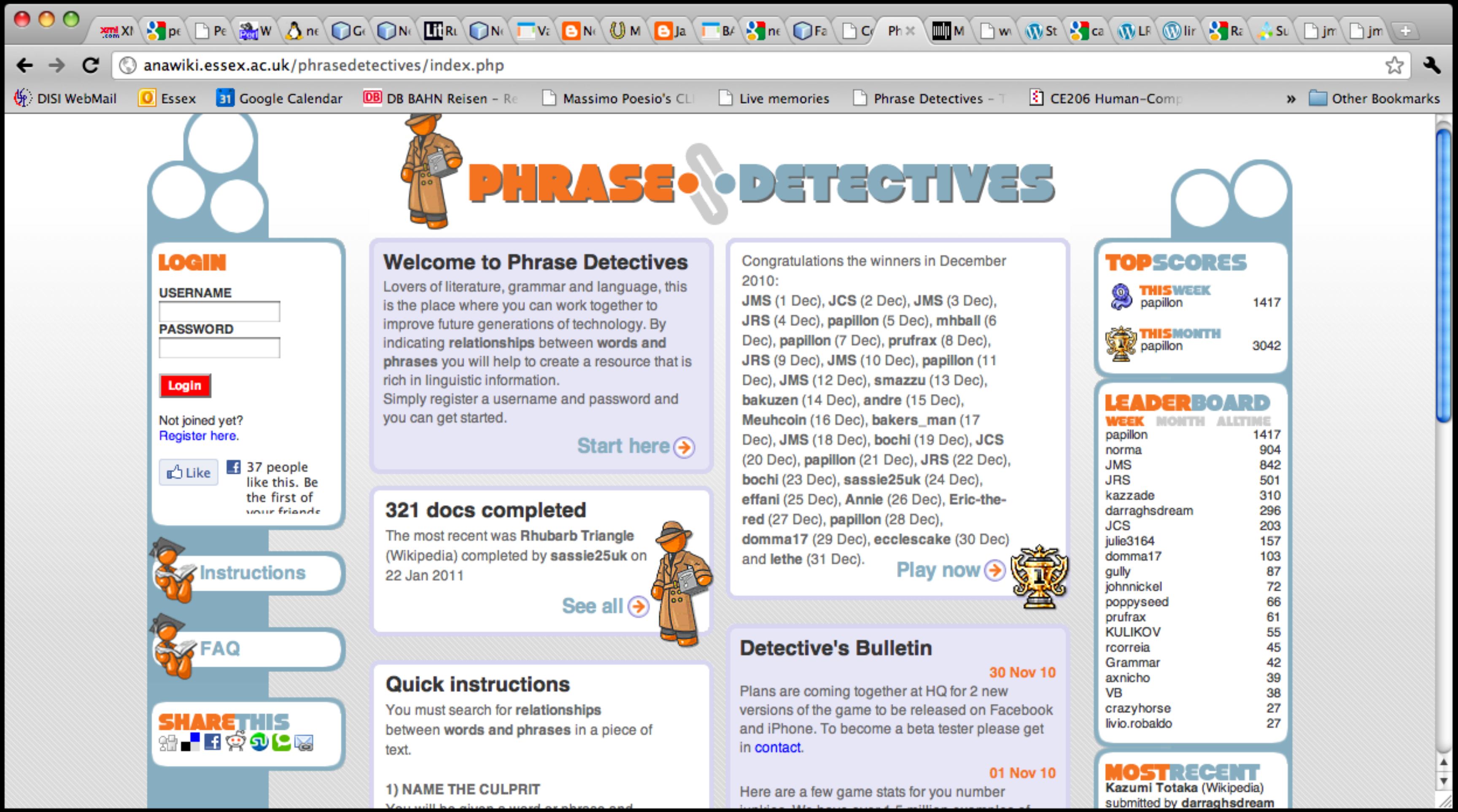
Explicit and implicit disagreements



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

Collecting disagreements online



LOGIN

USERNAME

PASSWORD

Login

Not joined yet?
Register here.

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

Instructions

FAQ

SHARETHIS

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

Congratulations the winners in December 2010:
JMS (1 Dec), **JCS** (2 Dec), **JMS** (3 Dec), **JRS** (4 Dec), **papillon** (5 Dec), **mhball** (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

THISWEEK	papillon	1417
THISMONTH	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
livio.robaldo	27

MOSTRECENT

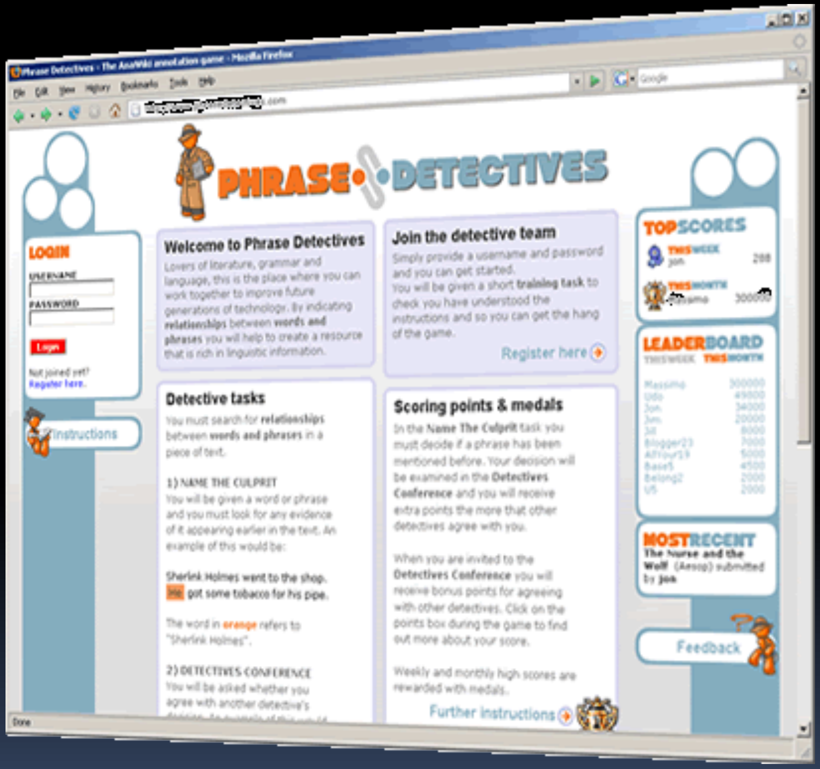
Kazumi Totaka (Wikipedia) submitted by **darraghsdream**

www.phrasedetectives.org

Gamifying annotation

- **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



Not joined yet? Register here.

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.

Join the detective team

Simply provide a username and password and you can get started. You will be given a short **training task** to check you have understood the instructions and so you can get the hang of the game.

Detective tasks

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

1) NAME THE CULPRIT
You will be given a word or phrase and you must look for any evidence of it appearing earlier in the text. An example of this would be:

Sherlink Holmes went to the shop. He got some tobacco for his pipe.

The word in orange refers to "Sherlink Holmes".

2) DETECTIVES CONFERENCE
You will be asked whether you agree with another detective's

Scoring points & medals

In the **Name The Culprit** task you must decide if a phrase has been mentioned before. Your decision will be examined in the **Detectives Conference** and you will receive extra points the more that other detectives agree with you.

When you are invited to the **Detectives Conference** you will receive bonus points for agreeing with other detectives. Click on the points box during the game to find out more about your score.

Weekly and monthly high scores are rewarded with medals.

TOP SCORES

THIS WEEK	Jon	288
THIS MONTH	Massimo	300000

LEADERBOARD
THIS WEEK THIS MONTH

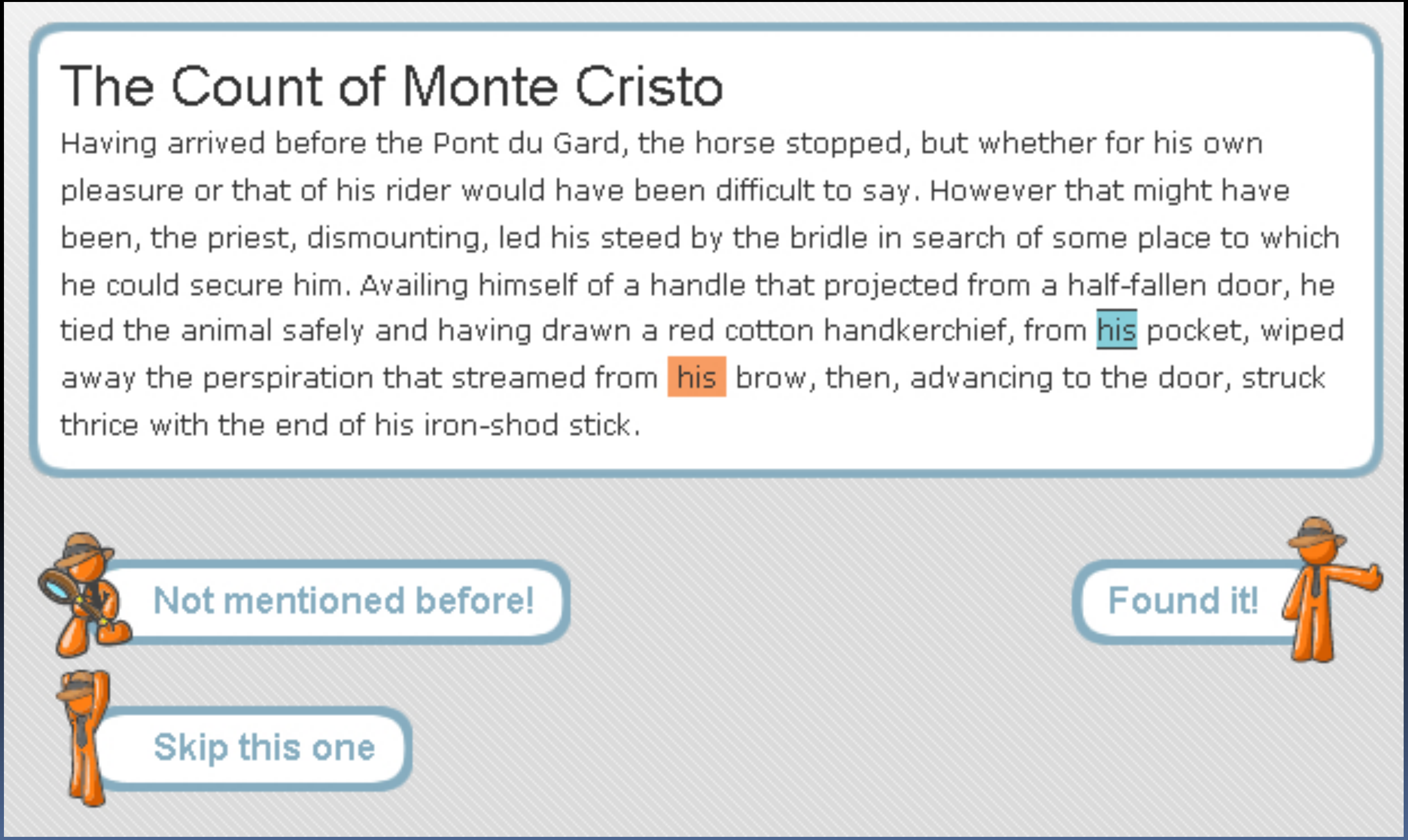
Massimo	300000
Udo	49000
Jon	34000
Jim	20000
Jill	8000
Blogger23	7000
Allyour19	5000
Base5	4500
Belong2	2000
US	2000

MOST RECENT
The Nurse and the Wolf (Aesop) submitted by **Jon**

www.phrasedetectives.com

Find the Culprit

(aka Annotation Mode)

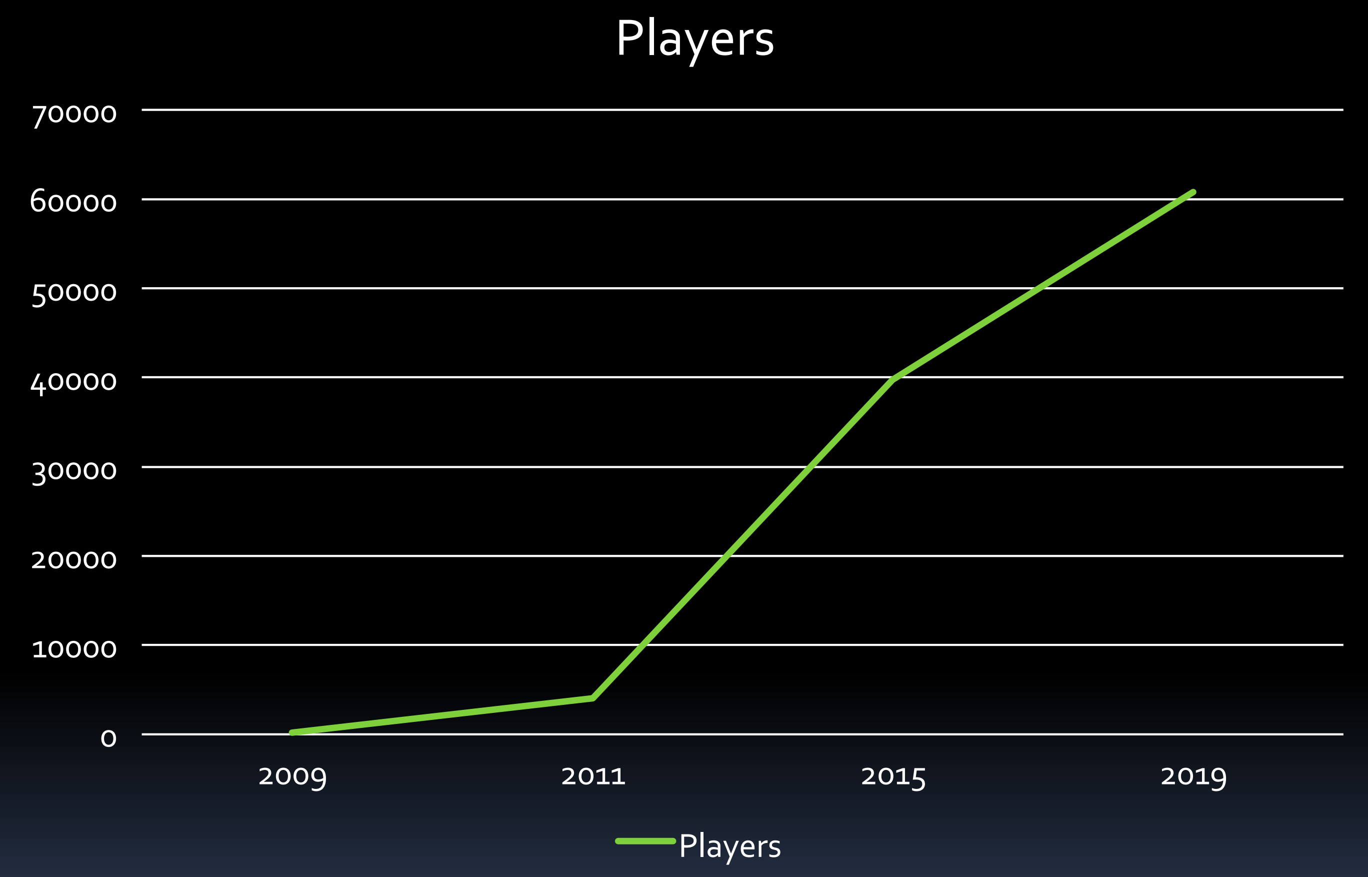


The Count of Monte Cristo

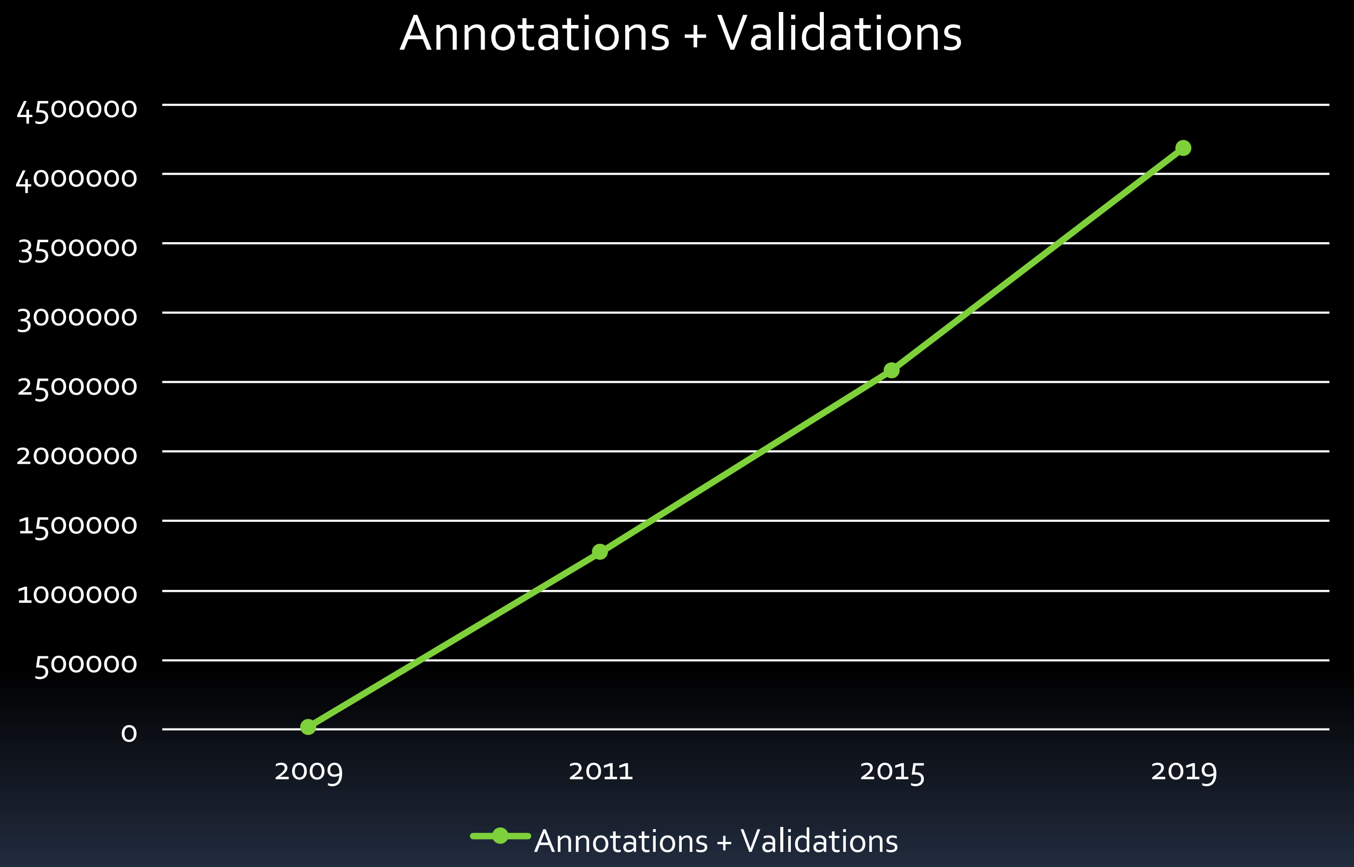
Having arrived before the Pont du Gard, the horse stopped, but whether for his own pleasure or that of his rider would have been difficult to say. However that might have been, the priest, dismounting, led his steed by the bridle in search of some place to which he could secure him. Availing himself 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.

www.phrasedetectives.com

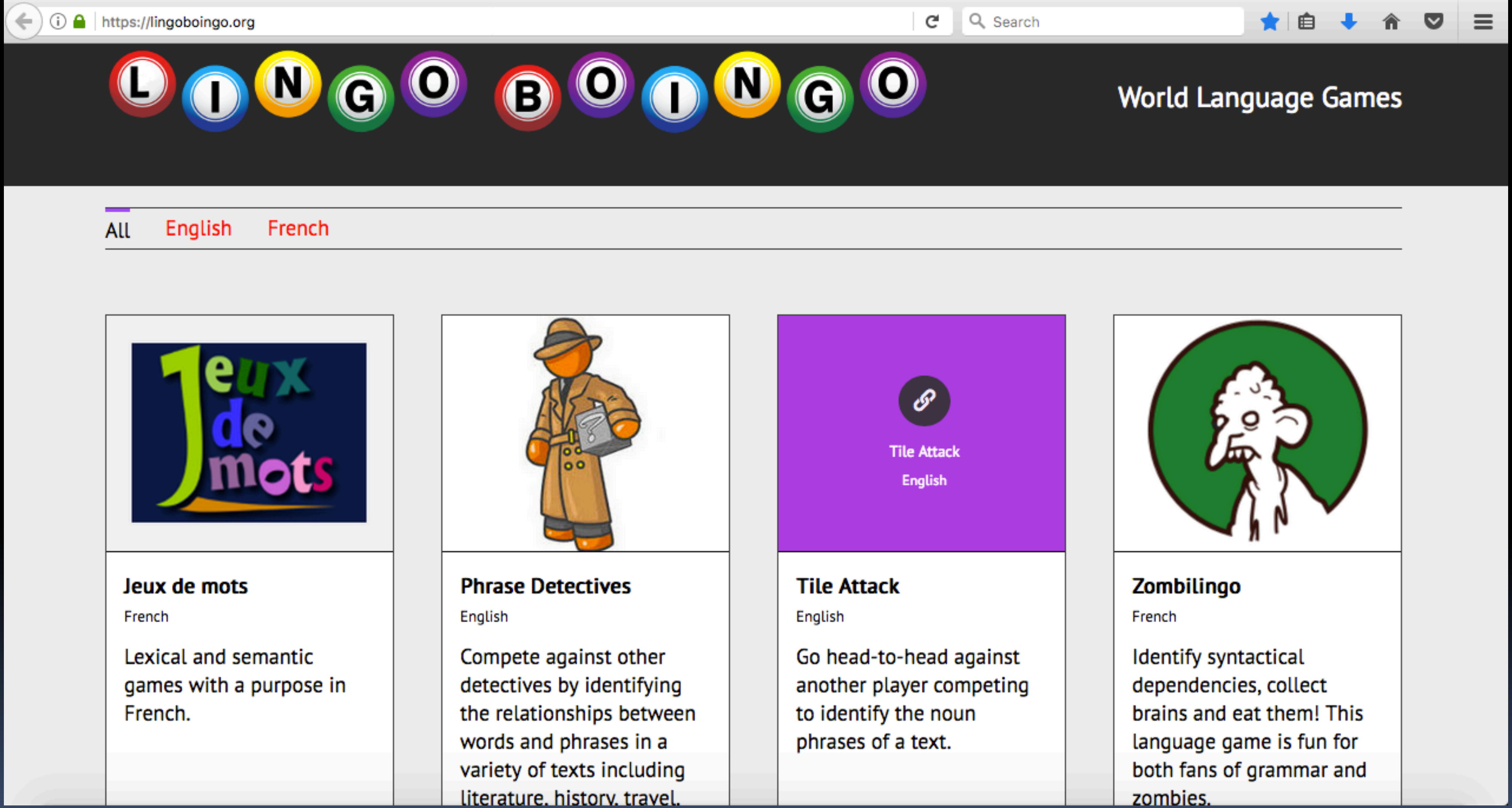
Number of players



Number of judgments



LingoBoingo



https://lingoboingo.org

LINGO BOINGO

World Language Games

All English French

Jeux de mots

French

Lexical and semantic games with a purpose in French.

Phrase Detectives

English

Compete against other detectives by identifying the relationships between words and phrases in a variety of texts including literature, history, travel,

Tile Attack

English

Go head-to-head against another player competing to identify the noun phrases of a text.

Zombilingo

French

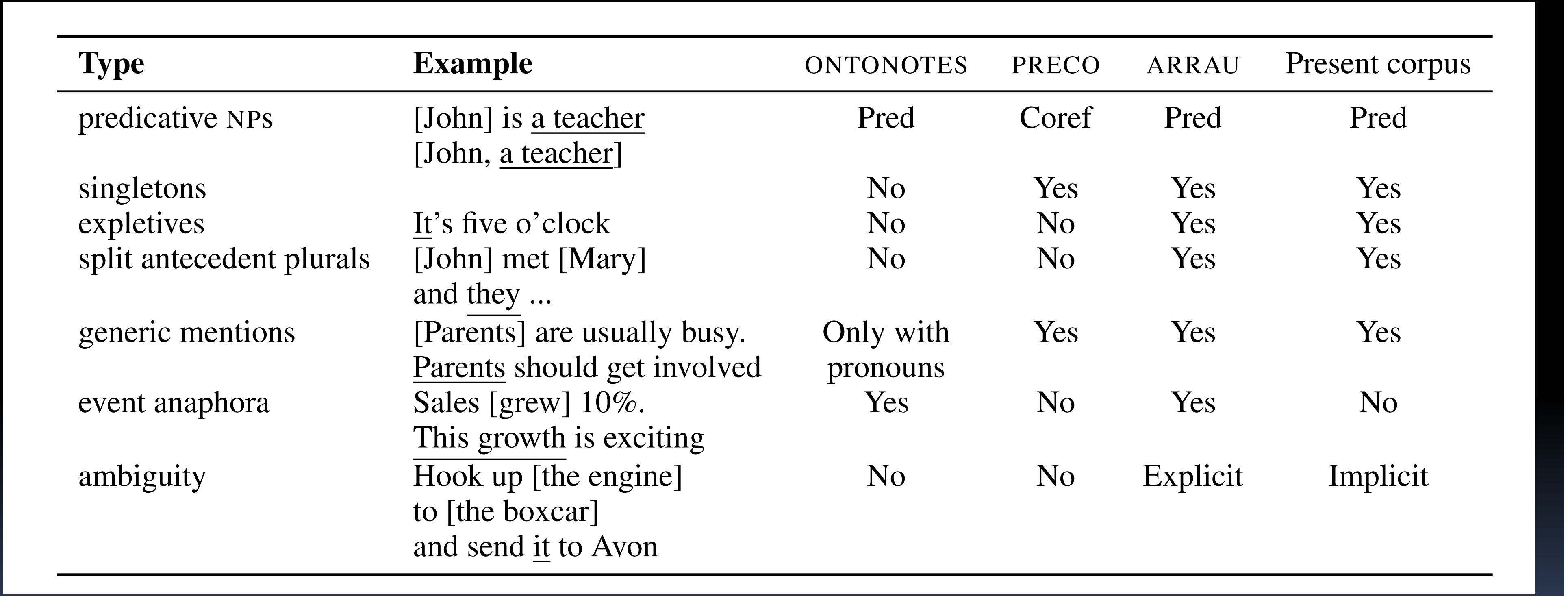
Identify syntactical dependencies, collect brains and eat them! This language game is fun for both fans of grammar and zombies.

The Phrase Detectives Corpus

The Phrase Detectives Corpus

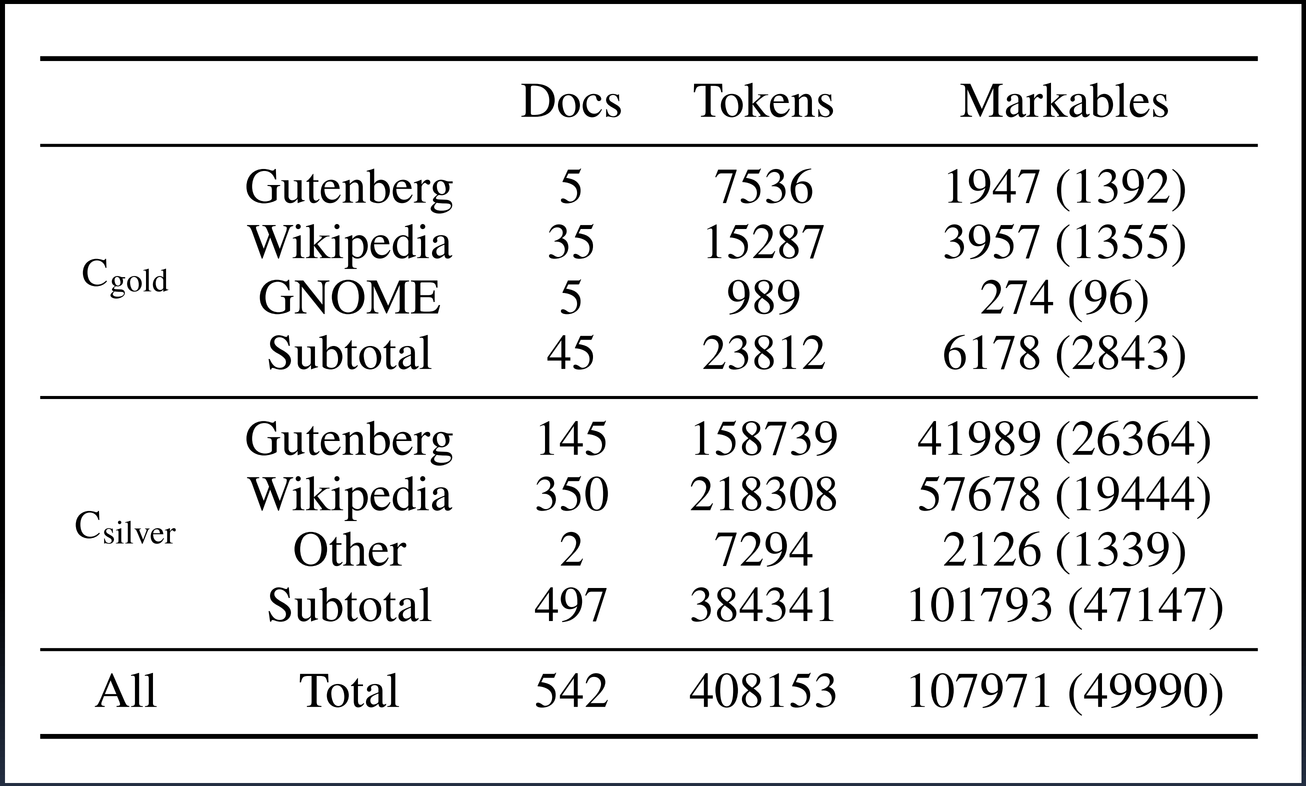
- Data:
  - 1.2M words total, of which around 330K totally annotated
  - About 50% Wikipedia pages, 50% fiction
- Markable scheme:
  - Around 25 judgments per markable on average
  - Judgments:
    - NR/DN/DO
    - For DO, antecedent
- Phrase Detectives 1 (with gold annotation) released via LDC in 2016
- Phrase Detectives 2 just released

PD corpus: annotation scheme



Type	Example	ONTONOTES	PRECO	ARRAU	Present corpus
predicative NPs	[John] is a teacher [John, a teacher]	Pred	Coref	Pred	Pred
singletons		No	Yes	Yes	Yes
expletives	It's five o'clock	No	No	Yes	Yes
split antecedent plurals	[John] met [Mary] and they ...	No	No	Yes	Yes
generic mentions	[Parents] are usually busy. Parents should get involved	Only with pronouns	Yes	Yes	Yes
event anaphora	Sales [grew] 10%. This growth is exciting	Yes	No	Yes	No
ambiguity	Hook up [the engine] to [the boxcar] and send it to Avon	No	No	Explicit	Implicit

PD2: Size



		Docs	Tokens	Markables
C_{gold}	Gutenberg	5	7536	1947 (1392)
	Wikipedia	35	15287	3957 (1355)
	GNOME	5	989	274 (96)
	Subtotal	45	23812	6178 (2843)
C_{silver}	Gutenberg	145	158739	41989 (26364)
	Wikipedia	350	218308	57678 (19444)
	Other	2	7294	2126 (1339)
	Subtotal	497	384341	101793 (47147)
All	Total	542	408153	107971 (49990)

PD2: Number of judgments

- 2,235,664 judgments from **425** 1958 players, of which
  - 1,358,559 annotations and
  - **426** 867,844 validations.
- On average, 20.6 judgments per markable
- Compare:
  - About 600K judgments for Ontonotes (~ 3 per markable)
  - About 10M judgments for PRECO (also ~ 3 per markable)

Assigning a probability to interpretations

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 aggregate these judgments
- A particularly popular approach is to use BAYESIAN MODELS OF ANNOTATION
  - Dawid and Skene 1979 (also used by Passonneau & Carpenter 2014)
  - Carpenter (2008)
  - Raykar et al 2010
  - Hovy et al, 2013

Bayesian Models of Annotation

- A Bayesian model of annotation 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

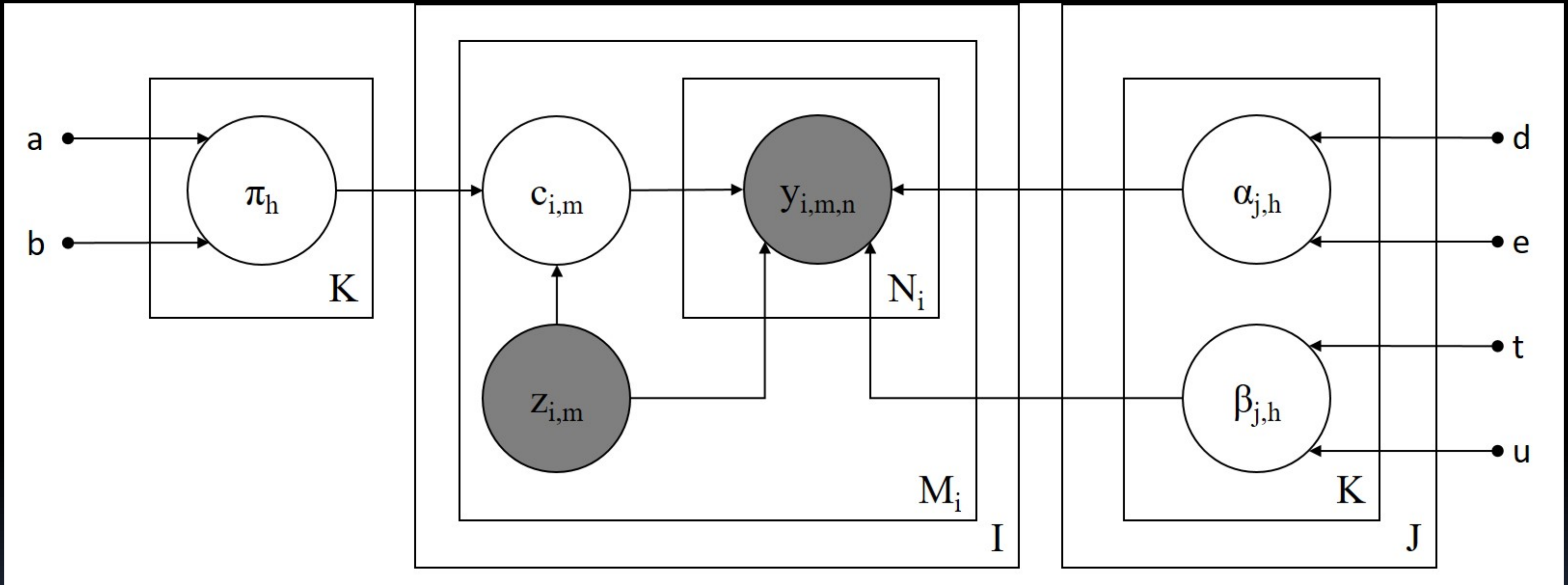
Comparing Bayesian Annotation Models

- Implemented in Stan (http://mc-stan.org/) some of the BAMs best-known in computational linguistics (Dawid & Skene, MACE, Carpenter's four models) & compared them on PD Gold data
- Evaluation metrics:
  - Accuracy
  - Annotator accuracy
  - Item difficulty
- The PD data are unique in a number of ways
  - Lots of judgments
  - Different types of noise from crowdsourcing
  - Gold info about spammers
- Paun et al, 2018. Comparing Bayesian Models of Annotation. *Transactions of the ACL*

Mention Pair Annotation (MPA)

- No existing BAM however can work with ANAPHORIC information, in which the 'labels' are not a discrete set, but coreference chains
- Our first model, called MPA, is a generative model of the process of linking mention pairs
- On the Phrase Detectives Data, it achieves an accuracy of 91.43% (as opposed to 84% for Majority Voting)
- Paun et al, 2018. A probabilistic annotation model for crowdsourcing coreference. *Proc. Of EMNLP*.

MPA

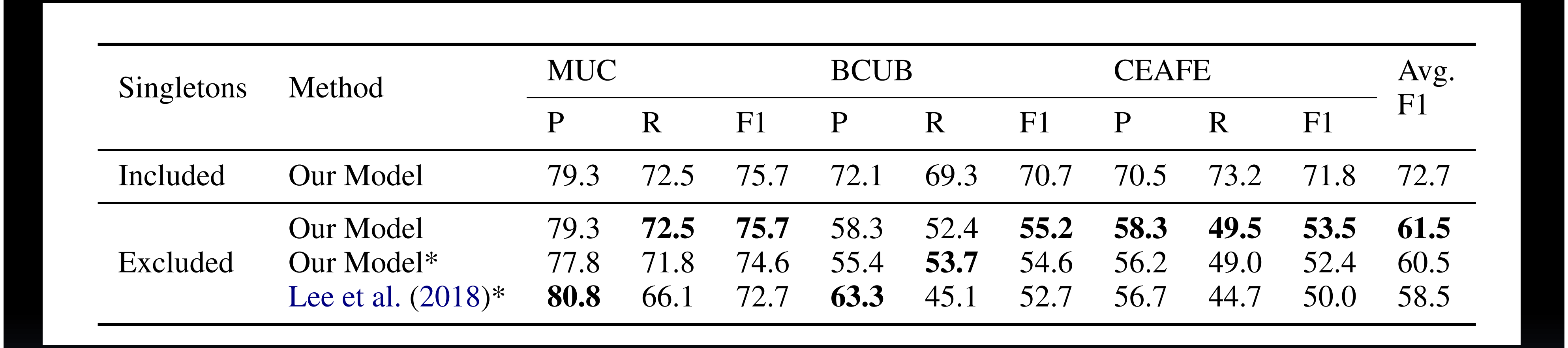


Anaphora resolution with PD 2

Methods

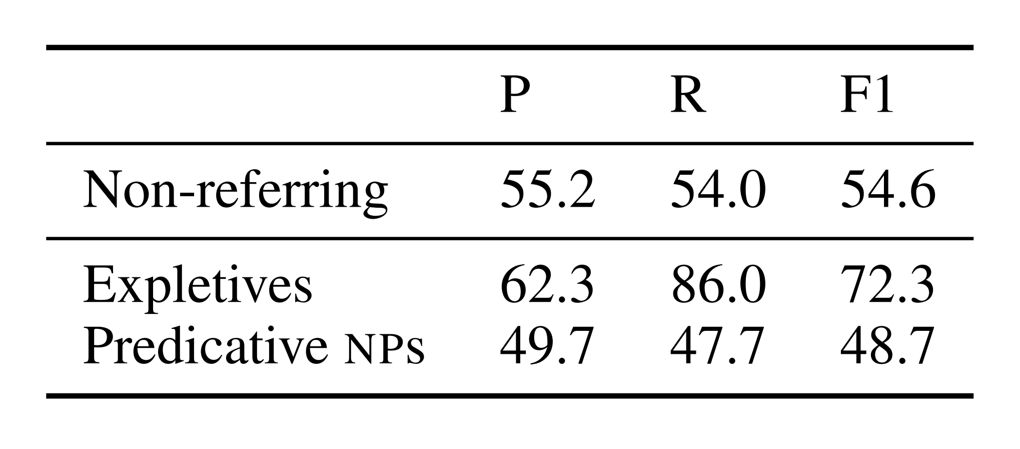
- The most likely (SILVER) labels extracted via MPA can be used to train CONLL-style coreference systems (if singletons and NR markables are ignored) or systems carrying out the full anaphora task
- For the second task, the Extended Coreference Score developed by Moosavi for the CRAC 2018 Shared Task can be used (Poesio et al, 2018)
- Two systems were trained and evaluated:
  - The state-of-the-art Lee et al 2018 system
  - Our own cluster ranking model (Yu et al, submitted)

Results on the CONLL task and with singletons



Singletons	Method	MUC			BCUB			CEAFE			Avg. F1
		P	R	F1	P	R	F1	P	R	F1	
Included	Our Model	79.3	72.5	75.7	72.1	69.3	70.7	70.5	73.2	71.8	72.7
Excluded	Our Model	79.3	**72.5**	**75.7**	58.3	52.4	**55.2**	**58.3**	**49.5**	**53.5**	**61.5**
	Our Model*	77.8	71.8	74.6	55.4	**53.7**	54.6	56.2	49.0	52.4	60.5
	Lee et al. (2018)*	**80.8**	66.1	72.7	**63.3**	45.1	52.7	56.7	44.7	50.0	58.5

Results with NR markables



	P	R	F1
Non-referring	55.2	54.0	54.6
Expletives	62.3	86.0	72.3
Predicative NPs	49.7	47.7	48.7

Ambiguity in the PD corpus

Raw disagreements

	1	2	3	4	5	6	7
PD_G	38.8%	30.6%	18.5%	7.3%	2.5%	1%	0.6%
PD_C	36%	30%	19%	8.8%	3.8%	1.8%	0.8%

Total number of markables in PD_C: 108,013
Total number of markables with no disagreements: 38579

61.2% of markables in PD_G, and 64% in PD_C, has more than 1 interpretation

An example of disagreement

RB ne75965

The day came that had been fixed for the marriage. The bridegroom arrived and also a large company of guests, for the miller had taken care to invite all his friends and relations. As [they] sat at the feast, each guest in turn was asked to tell a tale; the bride sat still and did not say a word.

DO ne75948 {for the miller had taken care to invite [all his friends and relations]} (11,3,1,13),
DO ne75945 {a large company of [guests]} (2,2,2,2),
DN (10,3,1,12),
DO ne75936 {the girl}, ne75942 {[the bridegroom]}, ne75945, ne75948 (1,1,3,-1),
DO ne75942, ne75946 {[the miller]}, ne75948 (2,2,2,2),
DO ne759370001 {[her]}, ne75942 {[the bridegroom]}, ne75945 {[the large company of guests]}, ne759490001 {his (the miller)} (1,0,4,-3,e2,e18),
DO ne75942 ne75948 ne759370001 ne75946 (1,3,1,3),
DO ne75942 ne75948 ne759370001 (1,2,2,1),
DO ne75942 ne759370001 ne75945 (1,0,4,-3),
DO ne75948 ne75946 (2,1,3,0),
DO ne75942 ne75948 ne759370001 ne75945 ne75946 (2,1,3,0),
DO ne75936 ne75937 {her father aka the miller} ne75942 ne75948 (1,0,4,-3)

+ 2 not_selectable, 3 skips

81 A+V, 5 comments skips, Total: 86 judgments

Not all disagreements are due to ambiguity

- Pradhan et al (2012): The analysis of the around 20,000 mentions on which there was disagreement in the ONTONOTES coreference annotation suggests that reasons include
  - Ambiguity proper ('unclear interpretation' or `disagreements on reference') (30% of disagreements, 7% of all mentions)
  - Annotator error (25% of the cases of disagreement)
  - Limitations of the coding scheme (36.5% of all disagreements)
  - Interface limitations (7.5% of all disagreements)

Interface limitations in PD

Note in this case we also have the type of ambiguity with DDEIX discussed in Poesio et al 2003, 2006

- Interface limitations
  - DDEIX: ne75896
    The old woman then mixed a sleeping draught with their wine, and before long they were all lying on the floor of the cellar, fast asleep and snoring. As soon as the girl was assured of this,

    DN (15, 2, 2, 15),
    DO ne75894 {[fast asleep] and snoring} ?? (2, 1, 3, 0),
    DO ne75895 {[the girl]} ?? (1, 0, 4, -3),
    DO ne75908 ?? {they were all lying on [the floor] in the cellar} (3, 0, 4, -1),
    DO ne75889 {they} ne75890 ?? (1, 2, 2, 1)

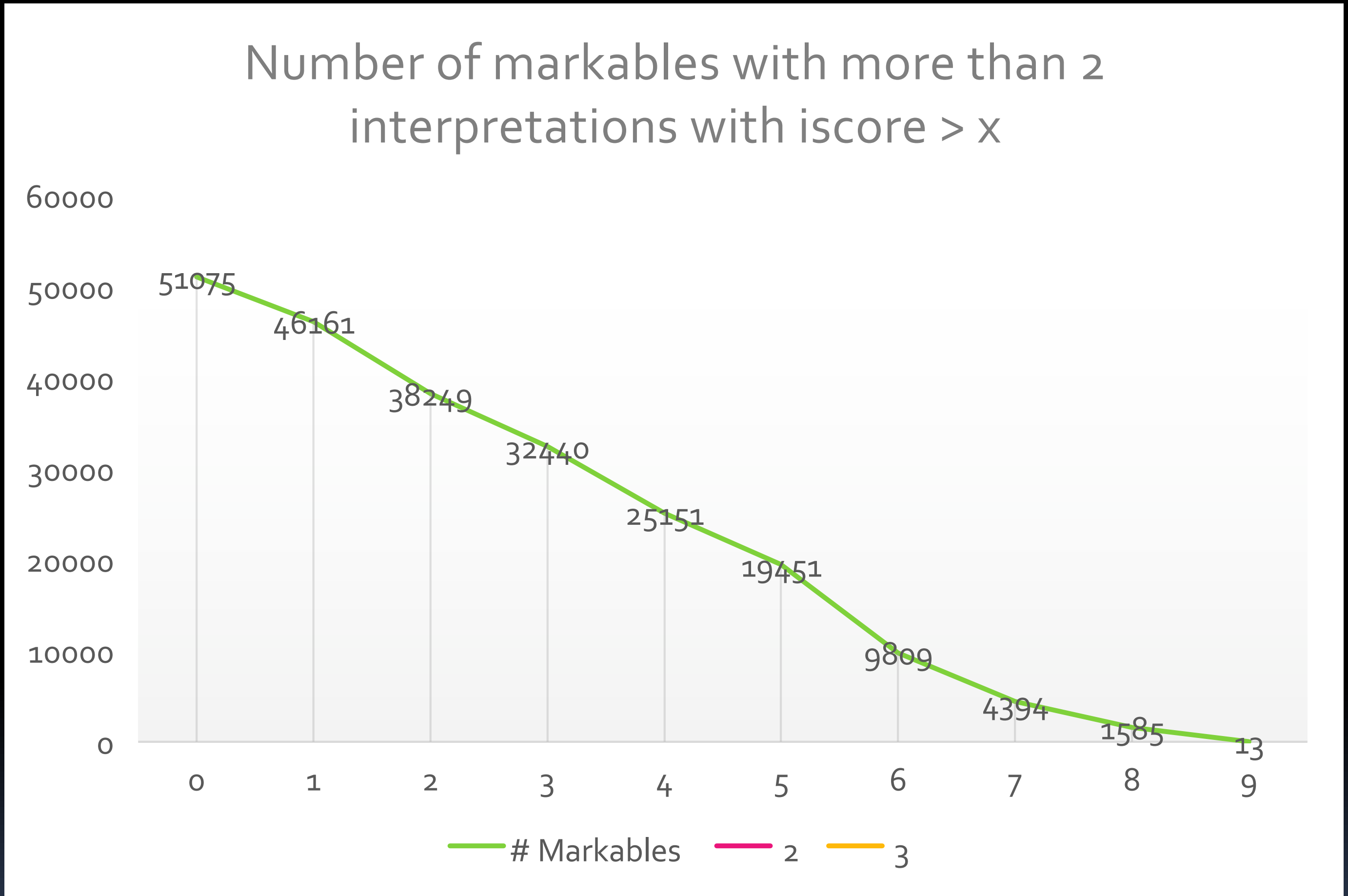
    3 skips

The validation filter

An interpretation can be `scored’ by counting the number of players who produce / agree with it, and subtracting the number of players who disagree with it

$$ISCORE\_i = ANN\_i + AGR\_i – DISAGR\_i$$

Filtering using validation



A second filter: MPA

	.5	.3	.1
#markables	104194	106042	106857
#mentions with more than one int.	2587	5263	10283
Highest number of int.	6	5	6

A second filter: MPA

	0	1	2	0 or 2
PD_G	2.3%	93.4%	4.3%	6.6%
PD_C	3.5%	94%	2.4%	5.9%

MPA and ambiguity

- The questions:
  - What types of ambiguity are there?
  - Which cases of ambiguity are correctly predicted by MPA?
  - Which cases of ambiguity are not caught by MPA, if any?

An analysis of disagreements in the PD2 corpus

- Chosen a few docs from PD $_G$
  - So far completely analyzed two Gutenberg docs:
    - Little Red Riding Cap (Grimm)
    - The Robber Bridegroom (Grimm)

- Labelled the disagreements as indicating
  - Ambiguity (definitely, possibly)
  - Cheating/Misunderstanding
  - Spurious ambiguity
  - Interface Problems (with attempt at classification)

Plurals 2: bare plurals

LRC ne7546

'Little Red-Cap raised her eyes, and when she saw the sunbeams dancing here and there through the trees, and pretty flowers growing everywhere, she thought: 'Suppose I take grandmother a fresh nosegay; that would please her too. It is so early in the day that I shall still get there in good time'; and so she ran from the path into the wood to look for **[flowers]**

DN (5,1,3,3,e18),
DO ne7532 {and [pretty flowers] growing everywhere ... } (8,2,2,8,e2),
DO ne7536 {Suppose I take grandmother [a fresh nosegay]} (6,2,2,6),
PR ne7536 ?? (1,1,3,-1),
PR ?? (1,1,3,-1),
DO ne7537 {[that] would please her too} ?? (1,3,1,3)

Plurals 3: `we’ and `you’

RB ne75698 (MPA: none)

And you, my love,' said the bridegroom, turning to her, 'is there no tale you know? Tell us something.' 'I will tell **[you]** a dream, then,' said the bride.

DO ne75965 (9,3,1,11,e2,e18),
DO ne75960 (4,2,2,4),
DN (2,0,4,-2)

Additional sources of ambiguity: paths

Her betrothed only replied, 'You must come and see me next Sunday; I have already invited guests for that day, and that you may not mistake the way, I will strew ashes along the path.'
When Sunday came, and it was time for the girl to start, a feeling of dread came over her which she could not explain, and that she might be able to find [her path] again,

DN (6,1,3,4,e18),
DO ne75663 (8,1,3,6,e2)

Ambiguity: REFERRING or NON-REFERRING?

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.

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

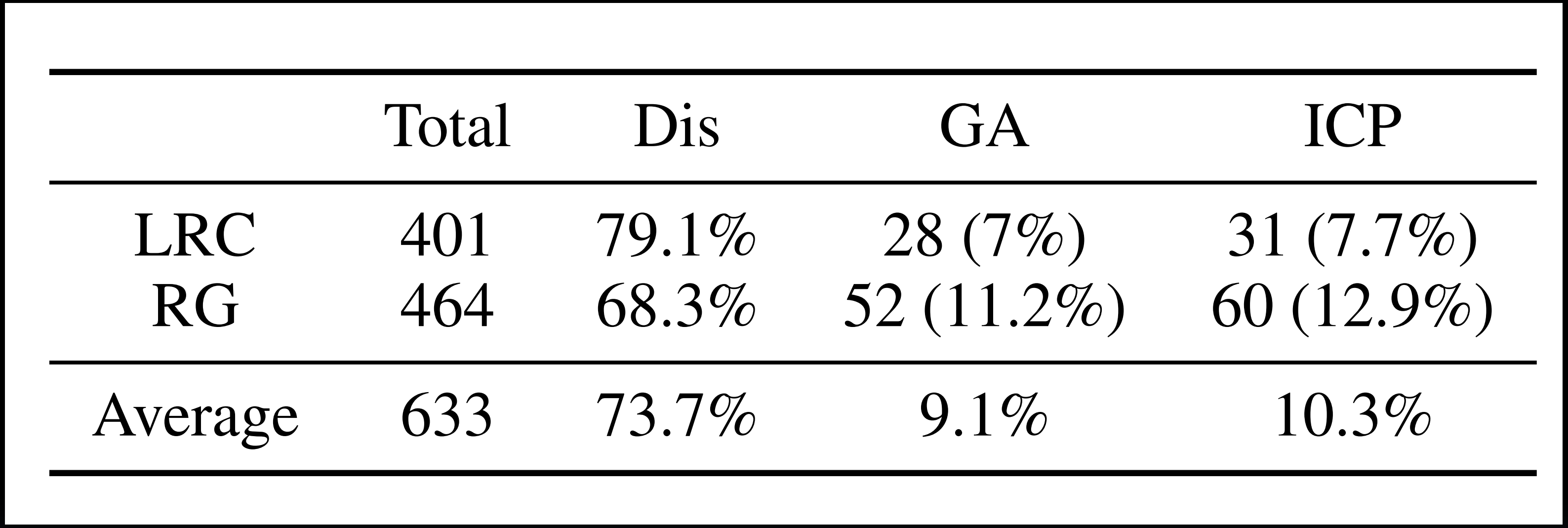
`DN’ when retelling a story as a dream

I went alone through [ne75972 a forest] and came at last to [ne75974 a house] ….

DN (9,0,0,9,e2)

3 out_of_context_window, 3 skips

Preliminary figures



	Total	Dis	GA	ICP
LRC	401	79.1%	28 (7%)	31 (7.7%)
RG	464	68.3%	52 (11.2%)	60 (12.9%)
Average	633	73.7%	9.1%	10.3%

MPA as ambiguity detector

- MPA is good at
  - Catching misunderstandings
  - Catching spurious ambiguity
- But not as good as ambiguity detector:
  - R: ~ 20%
  - P: ~ 50%

An hypothesis about justified and unjustified ambiguity

Previous theories of ´unproblematic’ ambiguity

- Poesio et al (1999, 2003, 2006)
  - JUSTIFIED SLOPPINESS: ’ambiguous’ references considered felicitous when candidate antecedents form a MEREOLOGICAL STRUCTURE
- Versley (2008)
  - GENERALIZED JUSTIFIED SLOPPINESS: ambiguous references felicitous when antecedents part of a DOT OBJECT in the sense of Pustejovsky and Asher
- Recasens et al (2010, 2012, 2014)
  - QUASI-COREFERENCE: coreference relation is a CONTINUUM between IDENTITY and NON-IDENTITY

Some additional evidence

- Frazier and Rayner (1990) and subsequent work on LEXICAL POLYSEMY: interpretation of polysemy different from interpretation of homonymy in that initial interpretation is not completely resolved (today we would say: UNDERSPECIFIED)
- The mereological cases cannot really viewed as dot-objects
- Recasens et al 2014: identity, near-identity and not-identity NOT A CONTINUUM

Preliminary new theory

- UNDERSPECIFICATION HYPOTHESIS:
  - Ambiguity is not problematic if the interpretations are part of an UNDERSPECIFIED STRUCTURE
    - But: mereological structure / dot objects are DISTINCT types of underspecified interpretation
- There are cases of UNJUSTIFIED SLOPPINESS
  - E.g., references to plans, areas
  - More similar to GOOD-ENOUGH cases (Ferreira et al)

Using information about disagreement in anaphora resolution

Previous work: using disagreement to filter

- Reidsma & Carletta (2008) and Beigman-Klebanov & Beigman (2009): use NOISE MODELS to exclude ‘hard cases’ from training
- The CrowdTruth project (Arroyo & Welty, 2014): DISAGREEMENT IS SIGNAL
  - Aroyo & Welty, 2013: use disagreement information to filter workers / sentences for relation extraction
  - See also Inel et al, 2014, 2017; Dumitrache et al, 2017, 2018
  - http://www.crowdtruth.org

Previous work: train with a probabilistic model

- Plank et al (2014): develop a loss function such that weight update is discounted by a factor depending on disagreement on an item

Conclusions

- Between 10% (written text, not considering discourse deixis) and 30% (spoken language, with deixis) of nominal expressions in language could be anaphorically interpreted in different ways
- This suggests that the assumption that each such expression has a unique `gold' interpretation is only a convenient idealization
- We are developing (freely available) resources that will allow ourselves NLP researchers to train models that do not make that assumption

The rest of the DALI Team

Richard Bartle

Jon Chamberlain

Udo Kruschwitz

Derya Cokal

Doruk Kicikoglu

Silviu Paun

Juntao Yu

The rest of the DALI Team (2)

Janosch Haber

Chris Madge

Alexandra Uma