LINGUISTIC GAMES WITH A PURPOSE

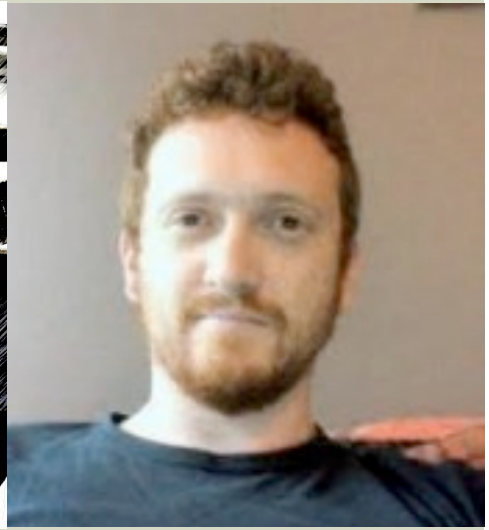
BUILDING A LARGE SEMANTICALLY ANNOTATED CORPUS

Johan Bos

14/09/2012

Linguistic Lunch





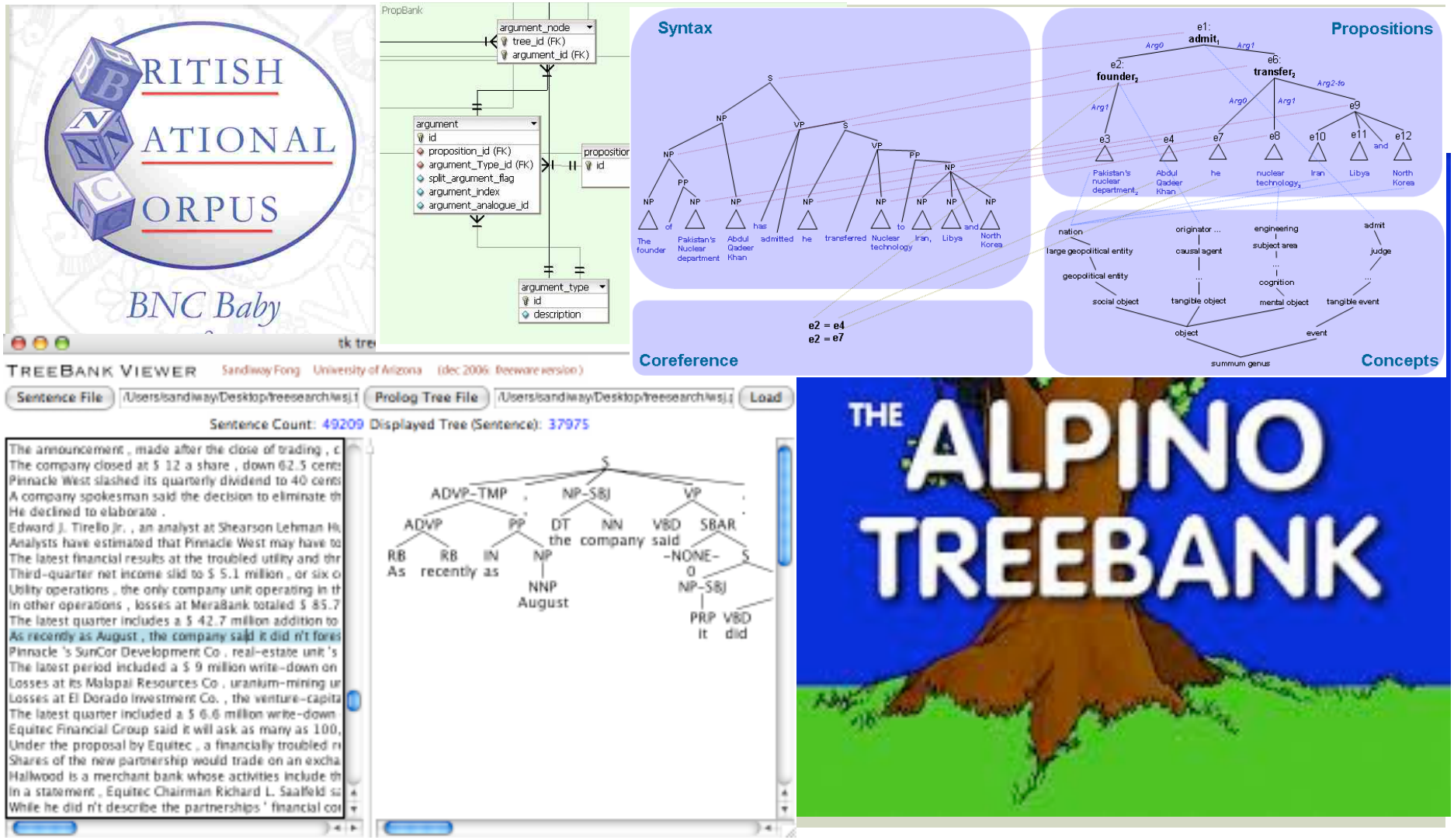
THIS IS JOINT WORK!

Kilian Evang

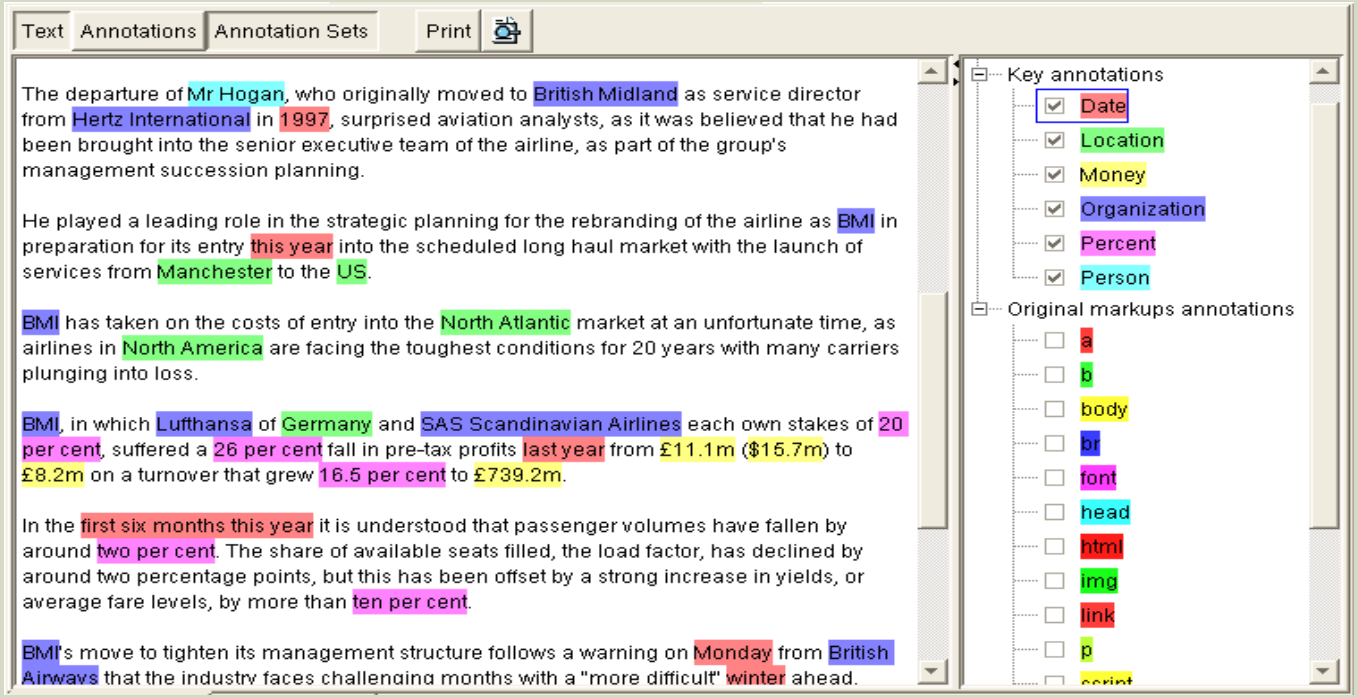
Noortje Venhuizen

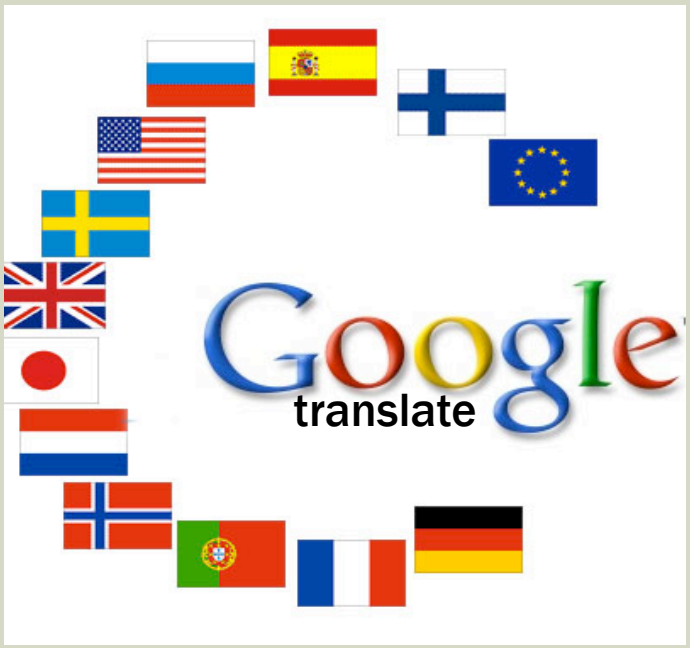
Valerio Basile

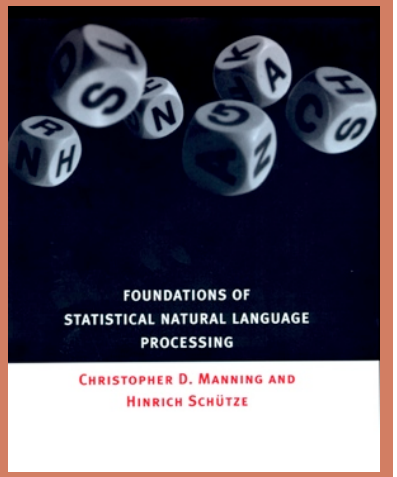
ANNOTATED CORPORA

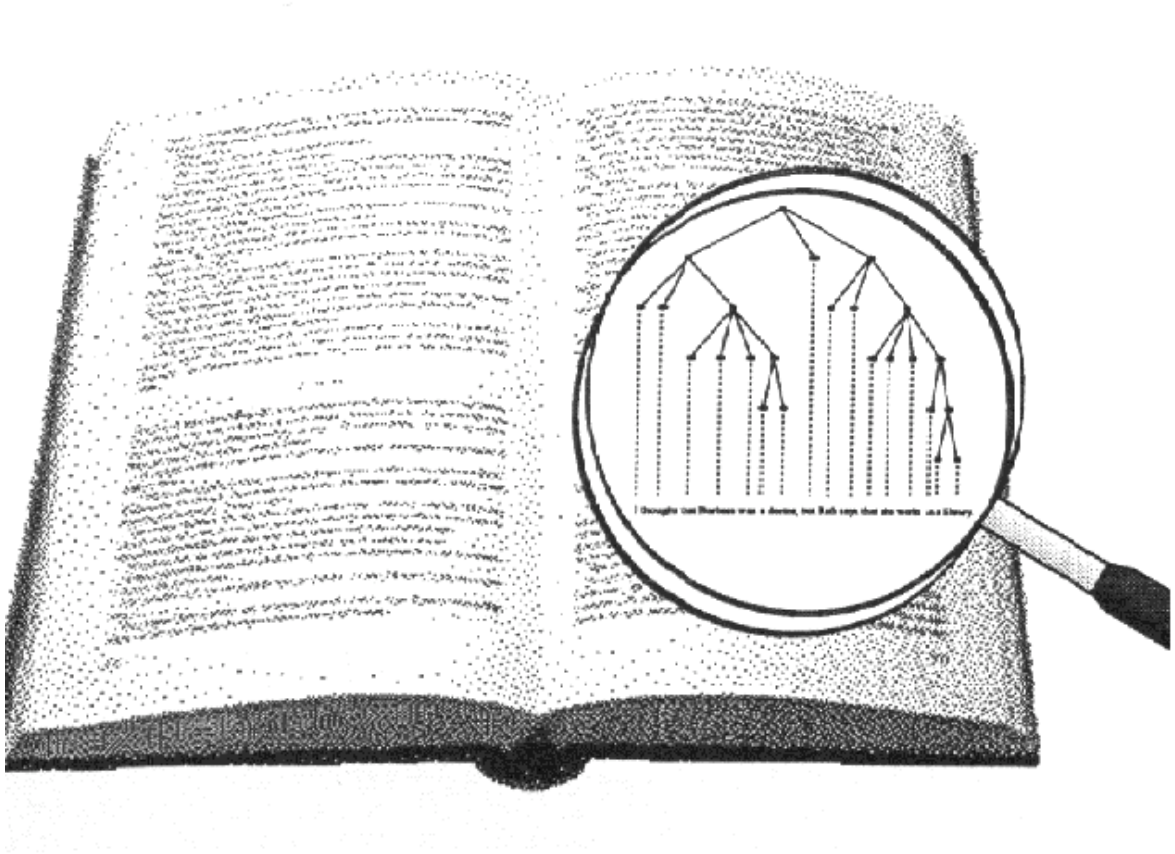


ANNOTATED CORPORA HAVE CHANGED THE WORLD!









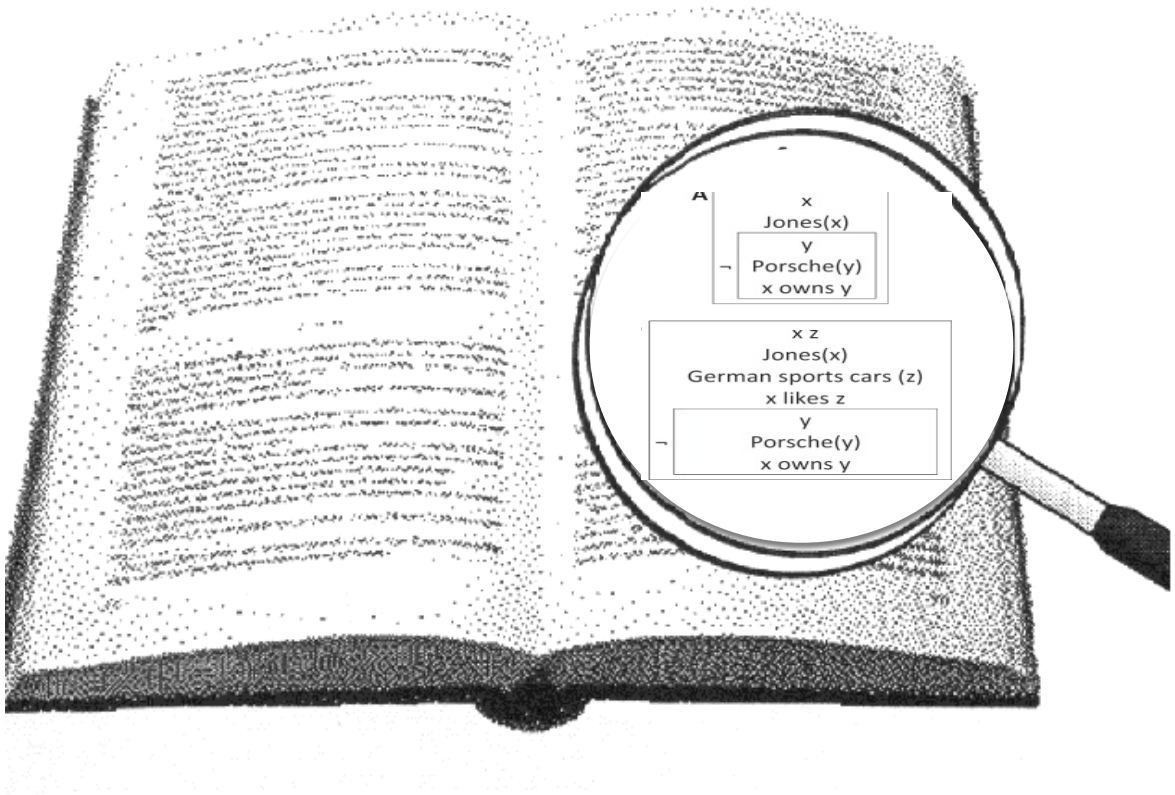
SYNTACTIC ANNOTATION

Examples:

- Penn Treebank
- Alpino
- Lingo Redwoods
- Negra

and many more!

SEMANTIC ANNOTATION



Examples:

THE GRONINGEN MEANING BANK

- large (English) corpus of public domain texts
- texts are aligned with semantic representations (DRT)
- automatically produced, manually corrected

INGREDIENTS OF THE MEANING BANK

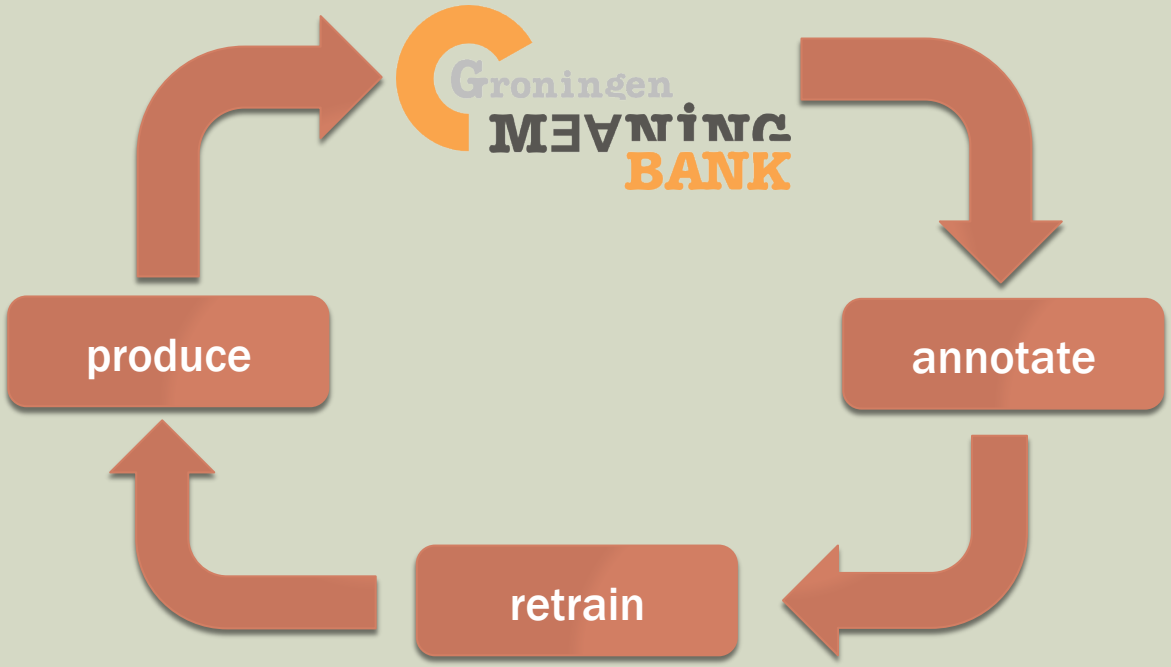
- PTB part of speech tags
- CCG supertags (lexical categories)
- Named entities
- WordNet senses
- VerbNet roles
- CCG derivations (syntax)
- DRT boxes (semantics)
- SDRT rhetorical relations (discourse)

THEORETICAL BACKBONE

- Discourse Representation Theory
- First-order logic as model theory
- Neo-Davidsonian event semantics
- Phenomena:
  - scope
  - quantification
  - pronouns
  - presupposition (Van der Sandt)
  - thematic roles (VerbNet)
  - rhetorical relations (SDRT)
  - tense and aspect (Kamp & Reyle)

ITERATIVE DEVELOPMENT

- The Meaning Bank Project is (ridiculously) ambitious!
- A lot of it is done by NLP tools
- But substantial human annotation cannot be avoided
- Note however: iterative development of tools

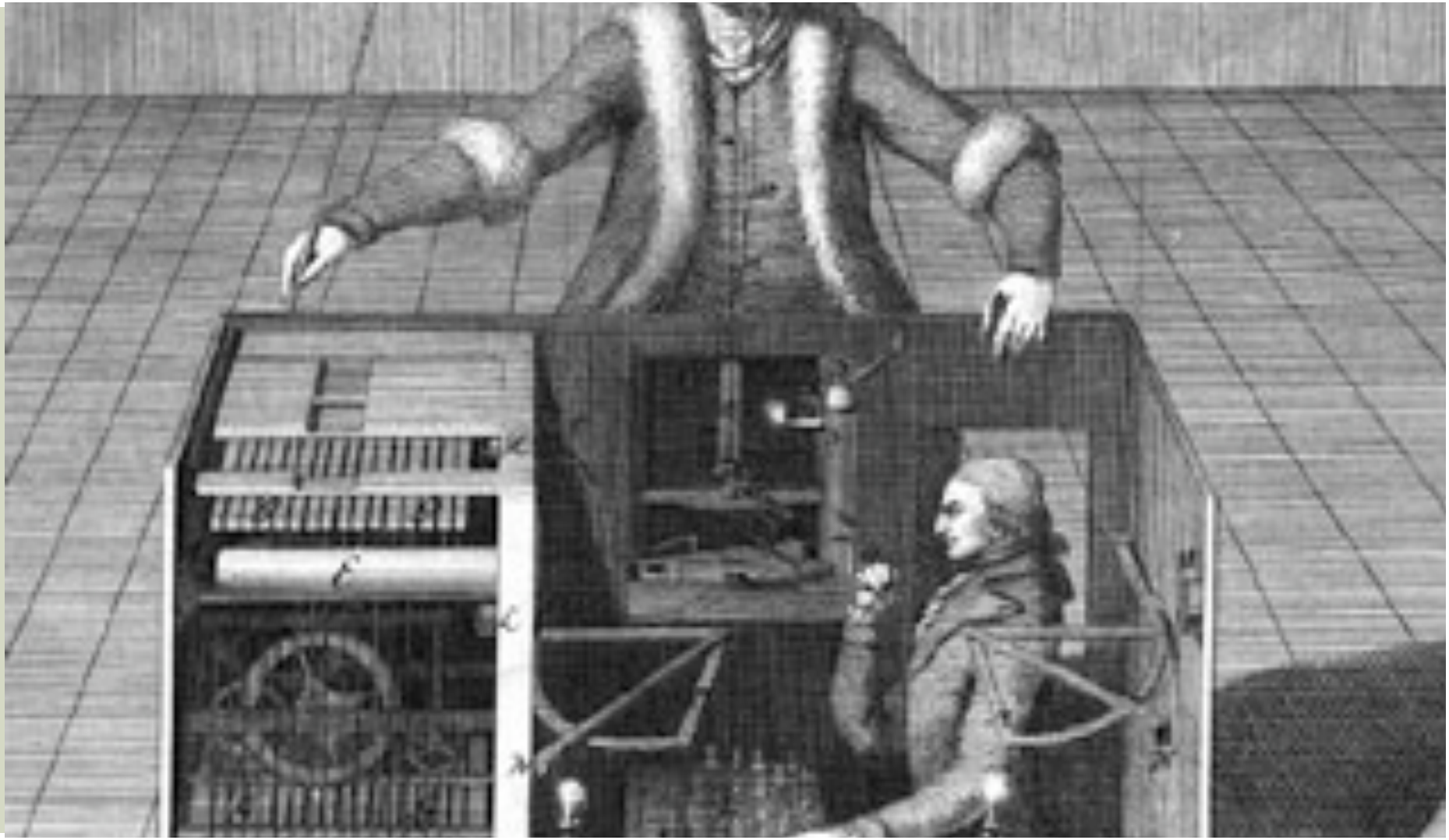


TRADITIONAL HUMAN ANNOTATION

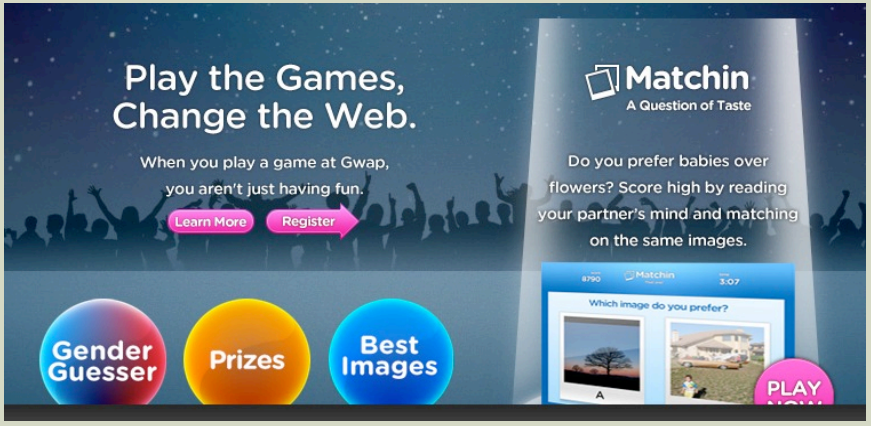
CROWDSOURCING

- outsourcing tasks to a distributed group of people
- the internet provides infrastructure
- the “crowd” is motivated: establishing social contacts, killing time, financial gain
- best example: wikipedia

AMAZON'S MECHANICAL TURK

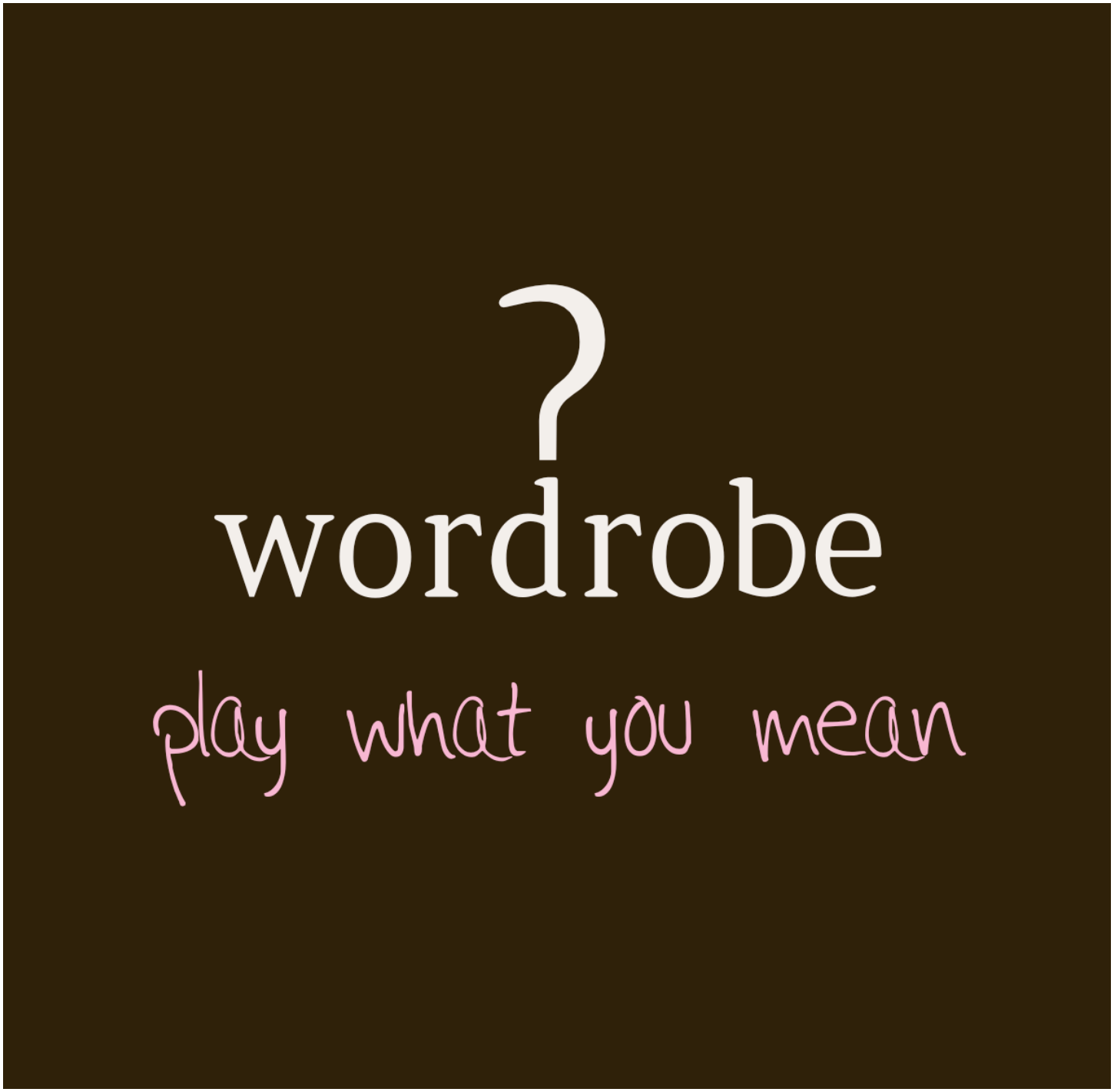


GAMES WITH A PURPOSE



With **Duolingo** you learn a language for free while helping to translate the web

PHRASE DETECTIVES



WWW.WORDROBE.ORG

Developed by the University of Groningen!

Operational since 12 Sept 2012

WORDROBE PHILOSOPHY

- not a single game, but a series of games that share same structure and scoring strategies
- each linguistic phenomenon that requires annotation corresponds to a different game
- every game consists of multiple-choice questions
- each question is presented by a text fragment plus a (small) number of possible answers
- these questions (and answers) are automatically generated from the corpus

HOMOGRAPHS

annotate part of speech of words that can both be noun and verb

difficult for machines (in some contexts)

PART OF SPEECH (NOUN OR VERB)

Twins Questions left until drawer is completed: 2

U.S.-led forces are hunting down remnants of Afghanistan's ousted Taleban regime who frequently carry out hit and run attacks on coalition and Afghan government **forces**, mainly in the country's eastern and southern regions.

- noun
- verb

Place your bet: low high

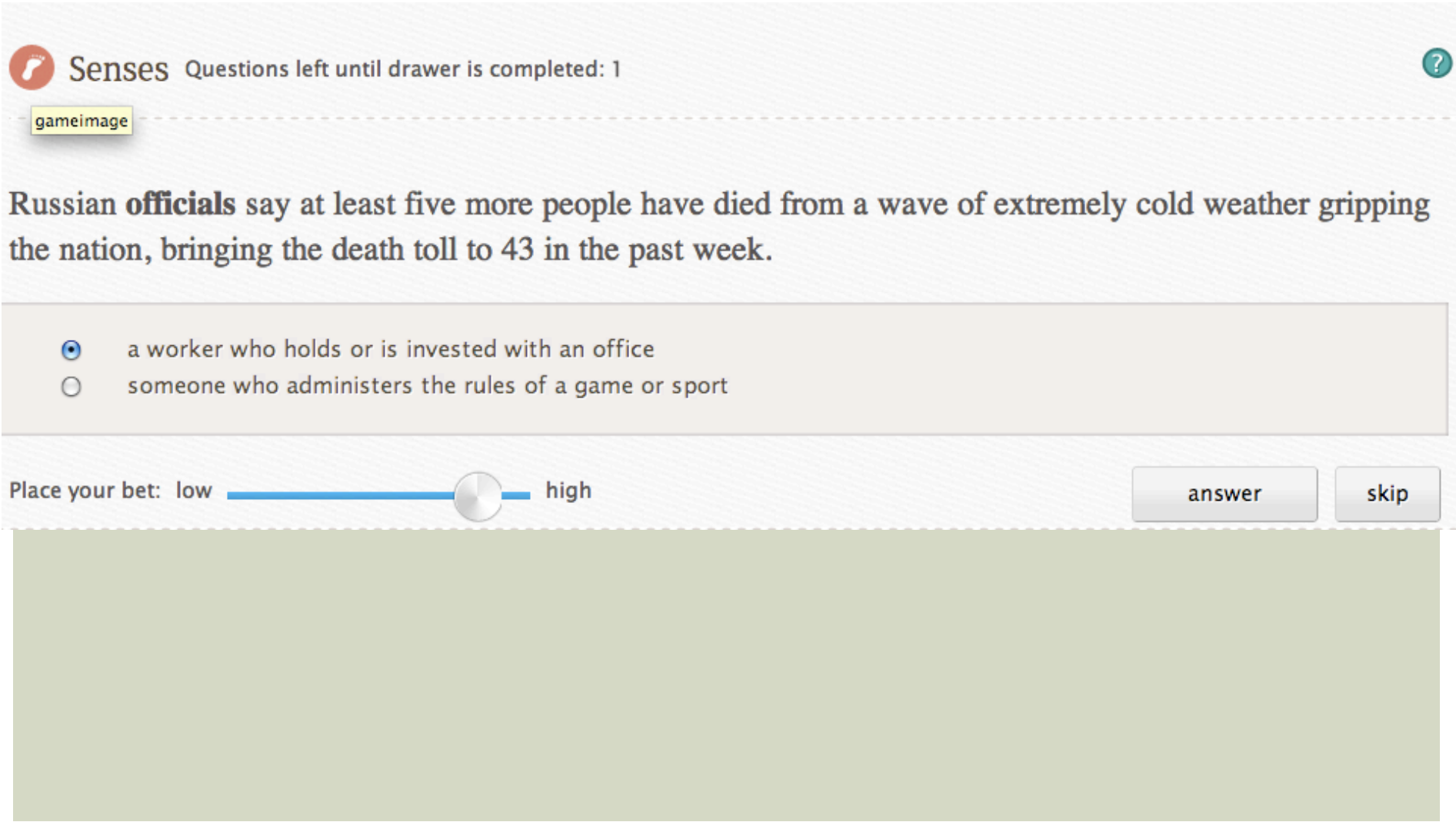
answer skip

WORD SENSES

annotate sense of word (noun or verb) using the WordNet sense inventory

difficult for machines (and humans too)

SENSES (WORDNET SENSES)

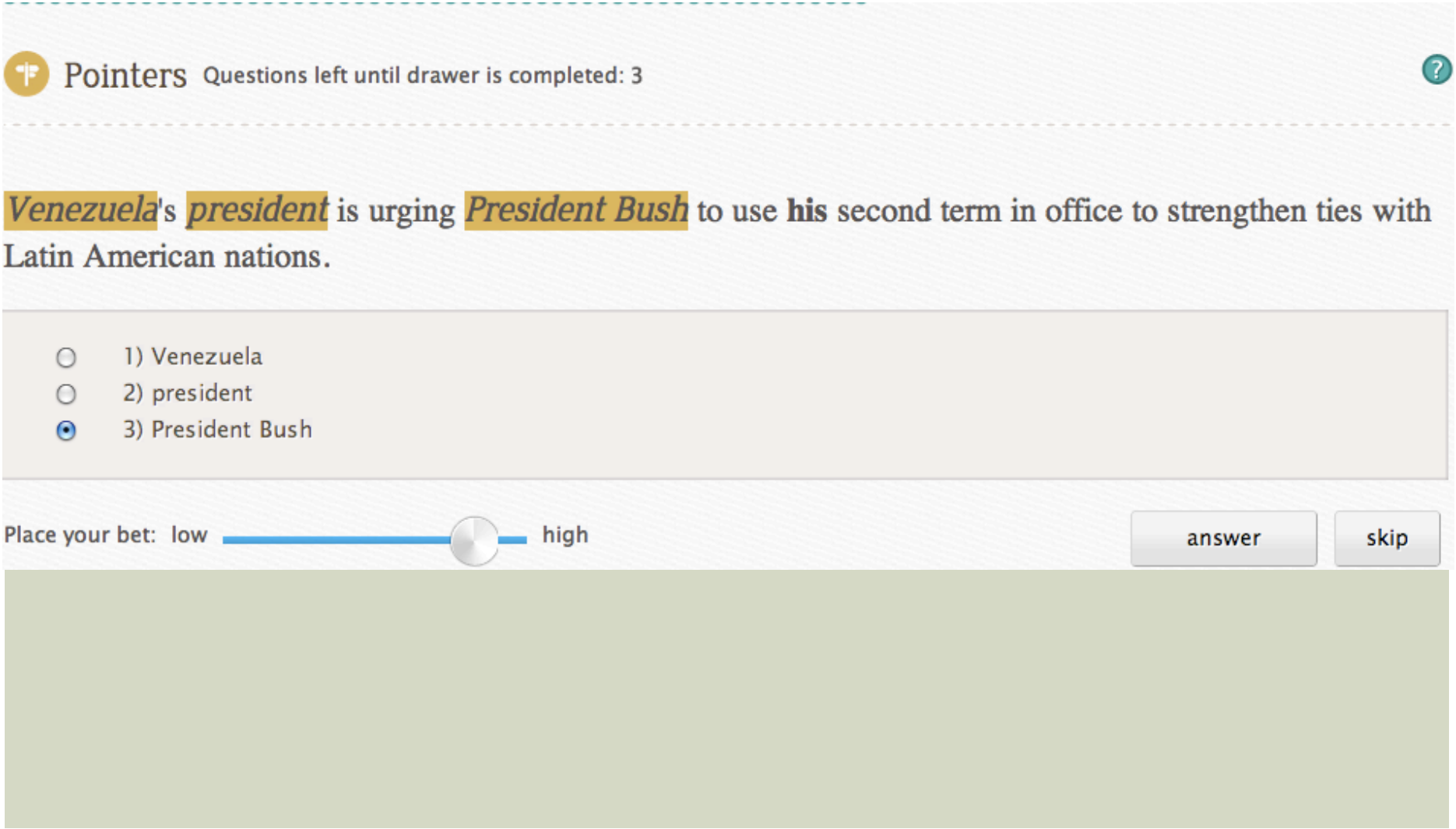


PRONOUNS

annotate antecedent of a pronoun (singular third-person)

very difficult for machines!

POINTERS (PRONOUNS)

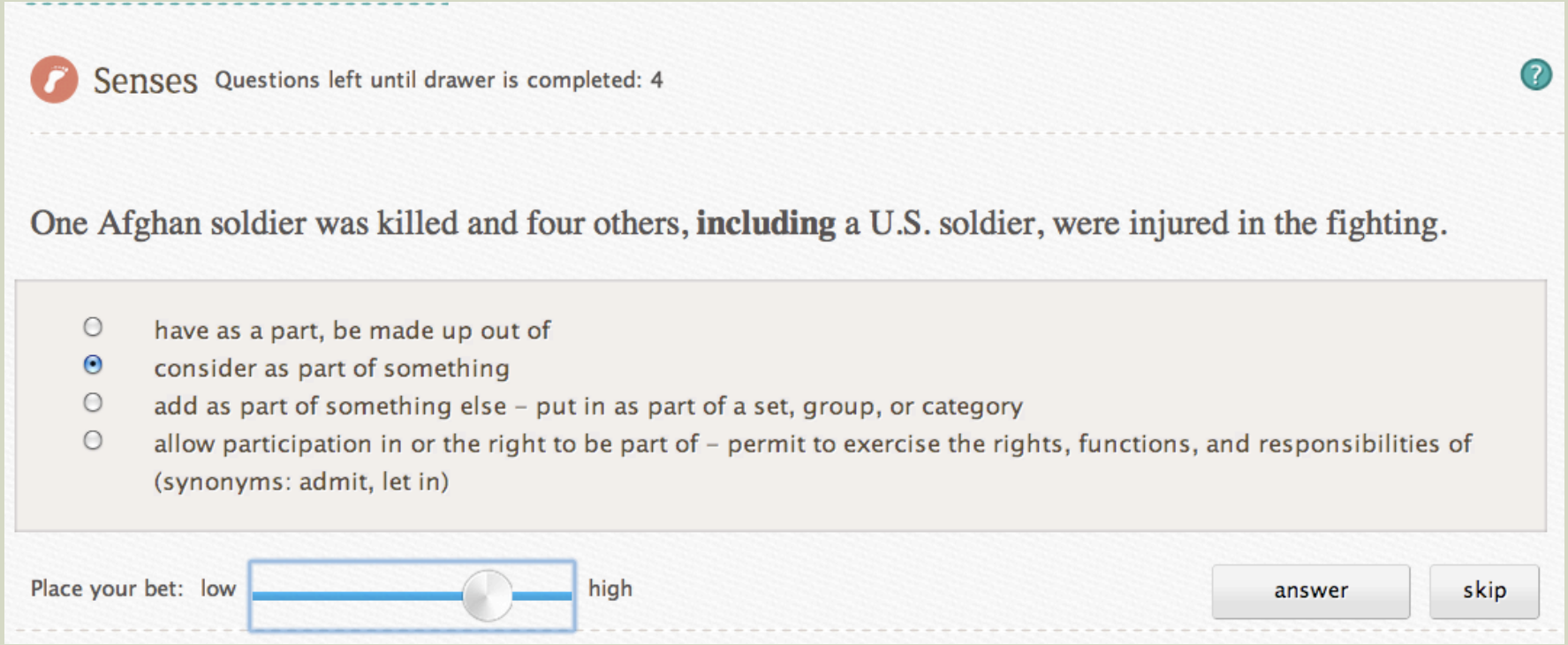


SCORING IN WORDROBE

- every answer increases the score of a player
- the more overlap of a player's answer with other players, the higher the score
- total scores for a game are calculated over answers given in the last N days (currently N=50)

A SCORE WITH A TWIST

- In addition, Wordrobe players can take risks and bet on the correctness of an answer
- The higher the bet, the more points you can win (or loose)



Kilian

Facebook page

Personal information

Name: Kilian

Completed drawers: 33

Senses: 10

Pointers: 13

Twins: 10

WHY WILL PEOPLE PLAY WORDROBE?

- unlocking achievements
- outperforming other players
- learning about language
- to win amazon vouchers!

wordrobe

play what you mean

WWW.WORDROBE.ORG