# Coling 2008 workshop on human judgements in Computational Linguistics

Ron Artstein
Gemma Boleda
Frank Keller
Sabine Schulte im Walde

August 23, 2008

1. Workshop Topic
2. Research Issues
    - Theoretical Issues
    - Reliability
    - Learning from Disagreements
    - Efficient Data Collection
3. Statistics

# Topic

Human judgments play a key role in computational linguistics:

- *category inventories and annotation schemes* are defined on the basis of judgments;
- *lexicon creation or corpus annotation* is conducted by experts via a sequences of linguistic judgments;
- *system evaluation* often involves judging the quality of system output or system performance.

# Topic

Questions concerning the *design of judgment experiments:*

- types of judgment experiments, design guidelines;
- lab-based vs. web-based experiments;
- methodologies for controversial tasks;
- role of ambiguity and polysemy in these tasks;
- appropriate level of granularity for judgment categories;
- type of participants (e.g., expert vs. naive);
- instructions and guidelines for participants.

# Topic

Questions concerning the *analysis and interpretation* of judgment data:

- importance of inter-annotator agreement;
- most suitable measures of agreement;
- other quantitative and qualitative methods for analyzing judgments;
- similarity/difference with practice in psycholinguistics;
- interaction of analysis procedures and annotation instructions.

# Theoretical Issues

Making a linguistics judgment is a *categorization task:*

- in the psychological literature, two main theories of categorization exist:
- the *property view* holds that the members of a category are defined by a unique set of features;
- the *prototype view* assumes that category membership is defined by in terms of similarity to a prototypical exemplar of that category;
- traditionally, generative linguistics have espoused a property view, and cognitive linguists a prototype view, of categories;
- it seems possible that some linguistic categories work through properties, while others through prototypes.

# Theoretical Issues

A recent theoretical issue is *gradience* in linguistic data:

- the literature on gradience has mainly focused on *gradient grammaticality* judgments;
- judgment techniques such as *magnitude estimation* have been developed to reliably elicit gradient judgments (Bard et al., 1996);
- however, there has been some work on *gradient linguistic categories* as well, e.g., Aarts' (2008) distinction between intersective and subsective gradience;
- these development are yet to be reflected in computational linguistics.

# Reliability

*Reliability of judgments* is an ongoing research issue:

- Cohen's $\kappa$ has been a widely used to measure agreement for linguistic annotation since Carletta (1996);
- but this has recently been criticized (e.g., Di Eugenio and Glass, 2004; Poesio and Artstein 2008);
- alternative measures exist in the form of Krippendorff's $\alpha$, Scott's $\pi$, Fleiss' $\kappa$, etc.
- Bhowmick et al. (this workshop) propose an extension of $\kappa$ to multi-category annotation.

# Learning from Disagreements

Another emerging issue is *disagreements in judgments:*

- disagreements can arise trivially due to *errors,* or due to genuine *subjectivity* in the judgment task;
- a key issue is the identification of the *source of disagreements* (Beigman Klebanov et al, this workshop);
- and how disagreements can be *exploited* for automatic classification (Reidsma et al., this workshop).

# Efficient Data Collection

In psychology, there has been a lot of interest in *data collection over the internet* (e.g., Birnbaum 2000):

- internet experimentation is very suitable for collecting linguistic judgments;
- offers access to a vast pool of *participants* and a wide range of *languages and demographics;*
- cost efficient, fast, experiments easy to set-up and analyze;
- but there are a number of *open issues:*
  - data integrity and participant authentication;
  - reliable presentation of instructions;
  - recruitment of expert partitions;
- software (e.g., WebExp) and infrastructure for recruiting subjects (e.g., Mechanical Turk) readily available.

# Statistics

Workshop organization:

- 22 submissions received
- reviewed by program committee of 30 reviewers
- 8 papers accepted as talks
- 33 registered participants

Sponsors:

- Spanish Education and Science Ministry via the KNOW project
- Sonderforschungsbereich 732, Universität Stuttgart