Sociolinguistic biases and the automatic identification of discourse markers in dialogue

Popescu-Belis, Andréi; Zufferey, Sandrine

Sociolinguistic biases and the automatic identification of discourse markers in dialogue

Andrei Popescu-Belis & Sandrine Zufferey
University of Geneva

Contrast

1. He was *like* a son to me.
2. Nobody can sing that song *like* he did.

with

3. It took, *like*, twenty minutes.
4. He was *like*, yeah, I can make dogs raise their ears.

Questions

- How do these types differ?
- Can humans distinguish them reliably?
- Can a computer program distinguish them as well?
- Does knowledge of the speaker's background help?

Outline of the talk

- Discourse markers (DM) as pragmatic functions of lexical items
  - focus on *like* and *well*
- Corpus: multiparty meeting recordings
  - data includes sociolinguistic characteristics of the speakers
- Experiments in DM recognition by humans
- Speaker-related statistical preferences for DM use
- Automatic detection of DM use
- Relevance of speaker-related features for DM detection

Discourse markers: a definition

- General purpose definition (Andersen, 2001)
  - "A class of short, recurrent linguistic items that generally have little lexical import but serve significant pragmatic functions in conversation"
- Examples
  - *actually, and, but, I mean, kind of, like, now, really, so, therefore, well, you know*
- Notoriously ambiguous items
  - serve other functions such as verb, adjective, etc.

The discourse marker *like*

- Function of *like* as a DM
  - make explicit to the hearer that what follows the marker is a loose interpretation of the speaker's belief (Andersen, 2001).
- Examples of DM uses
  - It took, *like*, twenty minutes.
  - They had little carvings of, *like*, dead people on the walls or something.
  - He was *like*, yeah, I can make dogs raise their ears.
- Examples of non DM uses
  - He was *like* a son to me.
  - Nobody can sing that song *like* he did.
  - I *like* chocolate very much.

The discourse marker *well*

- Function of *well* as a DM
  - "signals that the context created by an utterance may not be the most relevant one for the interpretation of the next utterance" (Jucker, 1993: 450).
- Examples of DM uses
  - A: Is the rising pitch a feature, or is it gonna be in the same file? B: *Well*, the rising pitch will never be hand-annotated.
  - So they'll say, *well*, these are the things I want to do.
  - Oh, yes, but…*well*, uh, yes, but what I mean is that…
- Examples of non DM uses
  - It's as *well* not to offend her.
  - I do not feel very *well*.
  - He sings as *well* as he plays.

Data

- ICSI Meeting Recorder corpus (Janin et al., 2003)
- About 80 hours of staff meeting recordings
- 5-8 speakers
- Scientific and technical discussions, in English, among research groups in speech and language processing from ICSI, Berkeley
- Transcribed: 800,000 words
- Segmented into utterances : ca. 100'000
  - indications of interruptions and unfinished utterances
  - dialog act annotation available too (Shriberg et al., 2004)

Participants to the ICSI-MR meetings

- 52 different speakers
- Sociolinguistic information (collected using paper forms)
  - gender
  - age
  - education level: undergraduate, graduate, PhD, professor
  - proficiency in English: native or non-native
  - US region of origin → interpreted as US East, US West and US other
- Speakers cohort is well-balanced with respect to these features
- Independence of speaker-related features (χ^2 test)
  - gender and education level
  - gender and age
  - gender and origin
  - but not, of course, age and education level

Biases in the speakers' contributions

- Most of the data is produced by only a few speakers
  - 7 most frequent ones > 40,000 words each (64% of the data)
  - 10 least frequent ones < 1,000 words each (0.6% of the data)
- Unbalanced contributions to the corpus (in words)
  - female / male 22% / 78%
  - native / non-native 74% / 26%
  - US East / US West / US other / other 27% / 12% / 32% / 29%
  - undergrad. / graduate / PhD / professor 2% / 30% / 40% / 28%
- Sociolinguistic features not fully independent when weighted by the number of words produced by each speaker (χ^2 test)
  - correlated: age and origin, gender and origin
  - independent: gender and age

Human annotation of DM vs. non-DM

- Objective
  - identify each occurrence of *like* and *well* as either a DM or not
- **How reliable is this annotation?**
- Possible sources of disagreement
  - individual mistakes
  - different perceptions of what a DM is
  - intrinsic ambiguity of certain occurrences of *like* and *well*
- Measure of inter-annotator reliability
  - *kappa* (κ) score (Krippendorff, 1980; Carletta et al. 1997)
    - factors out the probability of agreement by chance
  - κ scale
    - $\kappa < 0.67$ → insufficient agreement
    - $0.67 \leq \kappa < 0.8$ → acceptable
    - $0.8 \leq \kappa$ → very good

Observed inter-annotator agreement

- Experiments with excerpts of the data
  - up to six annotators, native and non-native EN speakers
  - transcripts *and* recordings (→ prosody) are required
- Inter-annotator agreement: **$\kappa = 0.74$ → good**
  - details in (Zufferey & Popescu-Belis, 2004)
- DM annotation of the entire ICSI MR corpus
  - two annotators
  - ~0.5% of the tokens are ambiguous: not used in the study
  - annotations available online at: http://www.issco.unige.ch/projects/im2/mdm/data/discourse-markers

Observed frequencies of DMs

- Frequencies of *like* and *well* as DMs
  - 4,519 tokens of *like* → 2,052 are DMs (45%)
  - 4,136 tokens of *well* → 3,639 are DMs (88%)
- Comparative frequencies of other DMs
  - most frequent ones

but	*well*	*like*	*actually*
7,815	3,639	2,052	1,763
0.98%	0.46%	0.26%	0.22%

  - most infrequent ones (note: used by several speakers)

however	*furthermore*	*moreover*
59	16	0

Speaker-related preferences (1)

Proportions of		**DM *like***	&	**DM *well***	
Gender:	male	40%		88%	not significant
	female	55%		89%	
English:	native	44%		87%	
	non native	49%		90%	
Origin:	US East	31%		84%	
	other US	46%		89%	
	other c.	49%		90%	
	US West	55%		91%	
Education:	professor	22%		84%	
	PhD	48%		*90%*	
	graduate	50%		*88%*	
	undergraduate	67%		94%	

Speaker-related preferences (2)

- Individual preferences for DM *like* show much greater variability than those for DM *well*
  - "heavy DM *like* users"
- Significant preferences
  - **speakers from the US West favor DM *like***
  - **less educated speakers favor DM *like***
- But…
  - in the ICSI-MR corpus, speakers from US East are older and more educated than those from US West
  - not clear which of the factors is determinant

Automatic disambiguation of DMs

- Method to determine automatically if an occurrence of *like* or *well* is a DM or not
  - using features extracted from recording and transcript
- How well does such a method score?
- Does knowledge of speaker-related preferences increase the accuracy of the method?

Low-level features for the recognition of DMs *like* and *well*

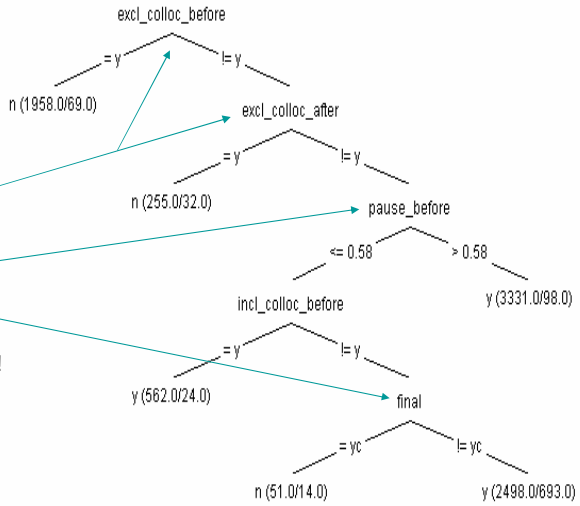
1. Collocations
   - the word immediately preceding or following a DM-candidate
   - examples for *like*
     - *like* that, *like* to, things *like*, seems *like*, would *like* → probably not a DM
     - of *like* → probably a DM
     - is *like*, was *like*, *like* a, *like* the, *like* you, it's *like* → uncertain
   - examples for *well*
     - as *well*, very *well* → probably not a DM
     - *well* I, *well* the, *well* it's, oh *well*, say *well* → probably a DM
     - *well* it → uncertain
2. Position and prosody
   - position in the utterance
     - initial, final, or middle
   - "prosody"
     - duration of the DM-candidate, duration of the pause before it and after it

DM recognition method

- Decision trees contain a set of tests
  - test = whether the features have particular values (nodes)
  - decision = whether the token (*like* or *well*) is a DM or not (leaf)
- Automatic learning of decision trees
  - based on training data
    - our set of hand-annotated examples (positive & negative)
  - algorithm
    - C4.5 / Weka, 10-fold cross-validation (Quinlan, 1993; Witten and Franck, 2000)
- C4.5 finds the best classifier for this data under certain constraints
- Score = *kappa* or the number of correctly classified occurrences

Nearly optimal binary decision tree

- Classification accuracy
  - **κ = 0.75, 89% CCI**
  - comparable to inter-annotator agreement
- Features used
  - collocations: esp. "excluding" ones
  - prosody: if pause before token > 580ms, then DM
  - position: final or not
- No sociolinguistic features!
  - but some "heavy DM-*like* users" are automatically identified in the *optimal* tree (which is more complex and scores slightly better)



Sociolinguistic features and decision trees

- Knowledge of speaker for C4.5 training:
  - does not help the recognition of DM *well*
  - slightly improves recognition of DM *like*
  - the linguistic features (collocation and prosody) are almost sufficient to reach the maximal score
- Proposed method for estimating the relevance of speaker-related features for DM recognition:
  - alternatively ignore the other features
  - look at the best classifier found and its score

Results/rules found automatically

- No significant correlation for 'gender' or 'native' – or for *well*
- **use of education only** ($\kappa = 0.39$)
  - undergraduate or graduate → all *like* are DMs
  - otherwise → all *like* are non-DMs
- **use of region of origin only** ($\kappa = 0.40$)
  - from the US West → all *like* are DMs
  - otherwise → all *like* are non-DMs
- **use of age only** ($\kappa = 0.44$ or 75% CCIs)
  - under 30 → all *like* are DMs
  - otherwise → all *like* are non-DMs
    - common view of DM *like* as a feature of adolescent speech (Andersen 2001)
- **Analysis**
  - a majority of speakers in ICSI-MR are graduates under 30 from the US West
  - it is not clear which of the three is the statistically relevant feature
  - but: correlation (κ) is better for **age**, **then** for **region**, **then** for **education**

Conclusion

- Two methods to study speaker-related effects on DM use
  - frequencies
  - role of features in automatic disambiguation
- Importance for sociolinguistic studies
  - distributional patterns of DMs in a meeting context
- Importance for computational linguistics/pragmatics
  - state-of-the-art method for automatic recognition
  - speaker-related features could be relevant for “regular users” of a system
- Future work: generalize the features to several markers
  - collocations, prosody
    - hand-crafted: precise, useful when not enough training data
    - could be extracted automatically if enough data is available
  - speaker-dependent features: must be “learned”

References

Andersen, G. 2001. *Pragmatic Markers of Sociolinguistic Variation: a Relevance-Theoretic Approach to the Language of Adolescents*. John Benjamins, Amsterdam.

Carletta, J., A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1).

Janin, A., D. Baron, J. A. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E., A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus. In *Proceedings of ICASSP 2003*, Hong Kong.

Jucker, A. 1993. The discourse marker well: a relevance-theoretical account. *Journal of Pragmatics*, 19.

Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufman, San Francisco, CA, USA.

Shriberg, E., R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of SIGdial 2004*, Cambridge, MA.

Witten, I. and E. Frank. 2000. *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco, CA.

Zufferey, S. and A. Popescu-Belis. 2004. Towards automatic identification of discourse markers in dialogs: The case of like. In *Proceedings of SIGdial 2004*, Cambridge, MA.

Acknowledgments

- Swiss National Science Foundation
  (IM)2 National Center of Competence in Research on **Interactive Multimodal Information Management**