2006

COLING • ACL

# COLING • ACL 2006

## Linguistic Distances

## Proceedings of the Workshop

Chairs:
John Nerbonne and Erhard Hinrichs

23 July 2006
Sydney, Australia

Production and Manufacturing by
*BPA Digital*
*11 Evans St*
*Burwood VIC 3125*
*AUSTRALIA*

©2006 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 1-932432-83-3

# Table of Contents

Preface ..... v

Organizers ..... vii

Workshop Program ..... ix

*Linguistic Distances*
John Nerbonne and Erhard Hinrichs ..... 1

*Semantic Similarity: What for?*
Ido Dagan ..... 7

*Similarity Judgments: Philosophical, Psychological and Mathematical Investigations*
Claude St-Jacques and Caroline Barrière ..... 8

*Automatically Creating Datasets for Measures of Semantic Relatedness*
Torsten Zesch and Iryna Gurevych ..... 16

*Comparison of Similarity Models for the Relation Discovery Task*
Ben Hachey ..... 25

*Sentence Comparison Using Robust Minimal Recursion Semantics and an Ontology*
Rebecca Dridan and Francis Bond ..... 35

*Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification*
Grzegorz Kondrak and Tarek Sherif ..... 43

*Evaluation of String Distance Algorithms for Dialectology*
Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens and John Nerbonne ..... 51

*Study of Some Distance Measures for Language and Encoding Identification*
Anil Kumar Singh ..... 63

*Towards Case-Based Parsing: Are Chunks Reliable Indicators for Syntax Trees?*
Sandra Kübler ..... 73

*A Measure of Aggregate Syntactic Distance*
John Nerbonne and Wybo Wiersma ..... 82

*A Structural Similarity Measure*
Petr Homola and Vladislav Kuboň ..... 91

*Variants of Tree Similarity in a Question Answering Task*
Martin Emms ..... 100

*Total Rank Distance and Scaled Total Rank Distance: Two Alternative Metrics in Computational Linguistics*
Anca Dinu and Liviu P. Dinu ..... 109

Author Index ..... 117

# Preface

Welcome to the proceedings of ‘Linguistic Distances’ a workshop held in conjunction with ACL/COLING 2006 in Sydney. An introductory article explains our motivation for holding the workshop, which attracted 30 submissions, of which thirteen are included in these proceedings. We are gratified by this level of interest. In fact we restricted the remit of the workshop to exclude the use of distance (or its inverse, similarity) in evaluation because it was felt that evaluation was already regularly the subject of several focused workshops. So the topic of ‘Linguistic Distances’ seems to resonate within the computational linguistics community.

Perhaps we should add that we also hoped to attract computational interest in (non-applied) linguistic topics, and that this, too, emerged in the submissions, although it is not strongly reflected in the choice of articles. We’ll poll participants about interest in a possible follow-up, but we have no plans in that direction at this writing.

Our thanks are largely given in the acknowledgments section of the introductory article, but let’s add thanks to Suzanne Stevenson, the workshop chair of the conference, and to her review committee.

John Nerbonne and Erhard Hinrichs

# Organizers

**Chairs:**

John Nerbonne, Groningen
Erhard Hinrichs, Tübingen

**Program Committee:**

Harald Baayen, Nijmegen
Walter Daelemans, Antwerp
Ido Dagan, Technion, Haifa
Wilbert Heeringa, Groningen
Ed Hovy, ISI, Los Angeles
Grzegorz Kondrak, Alberta
Sandra Kübler, Tübingen
Rada Mihalcea, North Texas
Ted Pedersen, Minnesota
Dan Roth, Illinois
Hinrich Schütze, Stuttgart
Junichi Tsujii, Tokyo
Menno van Zaanen, Macquarie, Sydney

**Additional Reviewers:**

Gosse Bouma, Groningen
Gertjan van Noord, Groningen
and anonymous reviewers

**Invited Speaker:**

Ido Dagan, Bar Ilan University

**Sponsor:**

Netherlands Organisation for Scientific Research (NWO), Grant 200-02100,
for cooperation with the *Seminar für Sprachwissenschaft*, Tübingen.

# Workshop Program

**Sunday, 23 July 2006**

8:45–9:00 Opening Remarks:
*Linguistic Distances*
John Nerbonne and Erhard Hinrichs

**Session 1: Semantics I**

9:00–10:00 Invited Talk:
*Semantic Similarity: What for?*
Ido Dagan

10:00–10:30 *Similarity Judgments: Philosophical, Psychological and Mathematical Investigations*
Claude St-Jacques and Caroline Barrière

10:30–11:00 Break

**Session 2: Semantics II**

11:00–11:30 *Automatically Creating Datasets for Measures of Semantic Relatedness*
Torsten Zesch and Iryna Gurevych

11:30–12:00 *Comparison of Similarity Models for the Relation Discovery Task*
Ben Hachey

12:00–12:30 *Sentence Comparison Using Robust Minimal Recursion Semantics and an Ontology*
Rebecca Dridan and Francis Bond

12:30–14:00 Lunch

**Sunday, 23 July 2006 (continued)**

**Session 3: Pronunciation and Language Variation**

14:00–14:30 *Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification*
Grzegorz Kondrak and Tarek Sherif

14:30–15:00 *Evaluation of String Distance Algorithms for Dialectology*
Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens and John Nerbonne

15:00–15:30 *Study of Some Distance Measures for Language and Encoding Identification*
Anil Kumar Singh

15:30–16:00 Break

**Session 4: Syntax**

16:00–16:30 *Towards Case-Based Parsing: Are Chunks Reliable Indicators for Syntax Trees?*
Sandra Kübler

16:30–17:00 *A Measure of Aggregate Syntactic Distance*
John Nerbonne and Wybo Wiersma

17:00–17:30 *A Structural Similarity Measure*
Petr Homola and Vladislav Kuboň

17:30–18:00 *Variants of Tree Similarity in a Question Answering Task*
Martin Emms

**Reserve Paper**

*Total Rank Distance and Scaled Total Rank Distance: Two Alternative Metrics in Computational Linguistics*
Anca Dinu and Liviu P. Dinu