Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora

August 29th 2004

20th International Conference on Computational Linguistics
Editorial

Large linguistically interpreted corpora continue to play an important role for machine learning, evaluation, psycholinguistics, and (increasingly so) theoretical linguistics. Many research groups are engaged in the creation of corpus resources annotated with morphological, syntactic, semantic, and discourse information for an ever-increasing variety of languages. Since 1999, the LINC workshop series has brought together these activities in order to identify and disseminate best practice in the development and utilization of linguistically interpreted corpora. LINC workshops, thus, have traditionally served a dual purpose, viz. to (i) present an up-to-date survey of ongoing work on linguistically interpreted corpora and (ii) facilitate the exchange, propagation, and further development of research results with respect to the annotation, conversion, and exploitation of such corpora. We hope to continue this tradition at LINC-04, as a post-conference workshop at COLING 2004.

The call for papers this year encouraged, among others, experience reports on the use of linguistically interpreted corpora, e.g. for larger-scale, data-driven linguistic discovery, as training material for grammar acquisition and other machine learning approaches, and for the diagnostics and evaluation of NLP technology. In our view, all of the above reflect an emerging tendency in our field, viz. the growing availability of linguistically interpreted corpora combined with an increasing recognition of their utility for theoretical linguists and language engineers alike. At the same time, we observe new and continued initiatives on creating annotated resources for additional languages, as well as ongoing work to both enlarge available data sets and refine and augment existing linguistic annotations. The production, distribution, and maintenance of re-usable linguistically interpreted corpora is a process that requires sustained efforts (and funding) over time.

Finally, another emerging tendency in the field of corpus annotation appears to be in the growing connection of linguistically annotated corpora to independently developed resources, specifically broad-coverage computational grammars and parsing systems. Where only a few years ago, say, a majority of researchers may have perceived corpus annotation as an alternative paradigm to more traditional grammar writing approaches—i.e. shifting focus from linguistic introspection to the analysis of ‘naturally occurring’ data, thus having to confront a potentially much broader range of ‘representative’ phenomena—today, grammar building and corpus creation seem to not only co-exist but mutually benefit each other. Analysis (and, to a lesser extent still, generation) systems based on hand-constructed, symbolic grammars now standardly incorporate ambiguity management facilities that are derived from training stochastic models on linguistically annotated corpora; and more recently, such grammars themselves are put to use to aid annotation, both to reduce the cost of corpus construction but also to increase the granularity of available annotation and internal coherence of linguistically interpreted corpora. We have invited Prof. Rob Malouf of San Diego State University (CA) and formerly Rijksuniversiteit Groningen (The Netherlands) to reflect on this development in a topical presentation.

Like in the preceding years, broad interest in the LINC workshop was attested by a fair number of submissions: with the help of the programme committee and a number of added reviewers (see next page) the editors had to select one out of two papers for presentation during the workshop. We are indebted to everyone who submitted their work to the workshop, our numerous colleagues who assisted in the selection process (on a tight schedule), the contributors to this volume for their fine cooperation, and the COLING workshop programme committee—in particular Michael Hess and Fabio Rinaldi for their outstanding balance of precision and flexibility. In order to best achieve both of the aforementioned workshop goals in a one-day schedule, we opted for a mixed set-up, combining plenary presentations, discussions, and a slightly more informal poster session.

Oslo & Saarbrücken, June 2004
Silvia Hansen-Schirra,
Stephan Oepen, and
Hans Uszkoreit
Program Committee

- Francis Bond, Kyoto
- Pierrette Bouillon, Geneva
- Thorsten Brants, Mountain View
- John Carroll, Sussex
- Tomaz Erjavec, Ljubljana
- Dan Flickinger, Stanford
- Silvia Hansen-Schirra (co-chair), Saarbrücken
- Frank Keller, Edinburgh
- Brigitte Krenn, Vienna
- Joakim Nivre, Växjö
- Stephan Oepen (co-chair), Oslo & Stanford
- Laurent Romary, Nancy
- Geoffrey Sampson, Sussex
- Kiril Simov, Sofia
- Hans Uszkoreit (co-chair), Saarbrücken
- Jean Veronis, Aix-en-Provence
- Atro Voutilainen, Helsinki
- Jakub Zavrel, Antwerp

Additional Reviewing

- Stella Neumann, Saarbrücken
- Andreas Eisele, Saarbrücken
- Valia Kordoni, Saarbrücken
- Gregor Erbach, Saarbrücken
- Nuria Bertomeu, Saarbrücken
- Frederik Fouvry, Saarbrücken
- Michaela Mahlberg, Saarbrücken
### 5th International Workshop on Linguistically Interpreted Corpora (LINC-04)

**Geneva, 29 August 2004**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 – 9:15</td>
<td>Welcome &amp; Introduction</td>
</tr>
<tr>
<td></td>
<td>Silvia Hansen-Schirra, Stephan Oepen, and Hans Uszkoreit</td>
</tr>
<tr>
<td>9:15 – 9:45</td>
<td>The HOLJ Corpus, Supporting Summarisation of Legal Texts</td>
</tr>
<tr>
<td></td>
<td>Claire Grover, Ben Hachey, and Ian Hughson</td>
</tr>
<tr>
<td>9:45 – 10:15</td>
<td>Towards User-Adaptive Annotation Guidelines</td>
</tr>
<tr>
<td></td>
<td>Stefanie Dipper, Michael Goetze, Stavros Skopeteas</td>
</tr>
<tr>
<td>10:15 – 10:45</td>
<td>The TIGER Dependency Bank</td>
</tr>
<tr>
<td></td>
<td>Martin Forst, Nuria Bertomeu, Berthold Crysmann, Frederik Fouvry, Silvia Hansen-Schirra, and Valia Kordoni</td>
</tr>
<tr>
<td>11:00 – 11:30</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:30 – 12:00</td>
<td>Word Order Variation in German Main Clauses</td>
</tr>
<tr>
<td></td>
<td>Andrea Weber and Karin Mueller</td>
</tr>
<tr>
<td>12:00 – 12:30</td>
<td>Inflectional Syncretism and Corpora</td>
</tr>
<tr>
<td></td>
<td>Dunstan Brown, Carole Tiberius, and Greville G. Corbett</td>
</tr>
<tr>
<td>12:30 – 13:00</td>
<td>Corpus-based Induction of an LFG Syntax-Semantics Interface for Frame Semantic Processing</td>
</tr>
<tr>
<td></td>
<td>Anette Frank and Jiří Semecky</td>
</tr>
<tr>
<td>13:00 – 14:00</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>14:00 – 15:00</td>
<td>Mining Corpora for Linguistic Insights (Invited Presentation)</td>
</tr>
<tr>
<td></td>
<td>Rob Maloul, San Diego State University (CA)</td>
</tr>
<tr>
<td>15:00 – 15:30</td>
<td>Discussion</td>
</tr>
<tr>
<td>15:30 – 16:00</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>16:00 – 17:00</td>
<td>The Hinoki Treebank, Working Toward Text Understanding</td>
</tr>
<tr>
<td></td>
<td>Francis Bond, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama, Eric Nichols, Akira Ohtani, Takaaki Tanaka, and Shigeaki Amano</td>
</tr>
<tr>
<td></td>
<td>The Szeged Corpus</td>
</tr>
<tr>
<td></td>
<td>A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus</td>
</tr>
<tr>
<td></td>
<td>Dora Csendes, Janos Csirik, and Tibor Gyimothy</td>
</tr>
<tr>
<td></td>
<td>Mining Linguistically Interpreted Texts</td>
</tr>
<tr>
<td></td>
<td>Cassiana Fagundes da Silva, Renata Vieira, Fernando Santos Osorio, and Paulo Quaresma</td>
</tr>
<tr>
<td></td>
<td>Automated Induction of Sense in Context</td>
</tr>
<tr>
<td></td>
<td>James Pustejovsky, Patrick Hanks, and Anna Rumshisky</td>
</tr>
<tr>
<td>17:00 – 17:30</td>
<td>Bootstrapping Parallel Treebanks</td>
</tr>
<tr>
<td></td>
<td>Martin Volk and Yvonne Samuelsson</td>
</tr>
<tr>
<td>17:30 – 18:00</td>
<td>Discussion</td>
</tr>
</tbody>
</table>
Table of Contents

Editorial
Silvia Hansen-Schirra, Stephan Oepen, and Hans Uszkoreit 3

The Hinoki Treebank. Working Toward Text Understanding
Francis Bond, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama,
Eric Nichols, Akira Ohtani, Takaaki Tanaka, and Shigeaki Amano 7

Inflectional Syncretism and Corpora
Dunstan Brown, Carole Tiberius, and Greville G. Corbett 11

The Szeged Corpus.
A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus
Dora Csendes, Janos Csirik, and Tibor Gyimothy 19

Towards User-Adaptive Annotation Guidelines
Stefanie Dipper, Michael Goetze, and Stavros Skopeteas 23

Towards a Dependency-Based Gold Standard for German Parsers.
The TIGER Dependency Bank
Martin Forst, Nuria Bertomeu, Berthold Crysmann, Frederik Fouvry,
Silvia Hansen-Schirra, and Valia Kordoni 31

Corpus-based Induction of an LFG Syntax–Semantics Interface for
Frame Semantic Processing
Anette Frank and Jiri Semecky 39

The HOLJ Corpus. Supporting Summarisation of Legal Texts
Claire Grover, Ben Hachey, and Ian Hughson 47

Automated Induction of Sense in Context
James Pustejovský, Patrick Hanks, and Anna Rumshisky 55

Mining Linguistically Interpreted Texts
Cassiana Fagundes da Silva, Renata Vieira, Fernando Santos Osorio, and Paulo Quaresma 59

Bootstrapping Parallel Treebanks
Martin Volk and Yvonne Samuelsson 63

Word Order Variation in German Main Clauses
Andrea Weber and Karin Mueller 71
Author Index

Shigeaki Amano 11
Nuria Bertomeu 35
Francis Bond 11
Dunstan Brown 15
Greville G. Corbett 15
Berthold Crysmann 35
Dora Csendes 23
Janos Csirik 23
Stefanie Dipper 27
Martin Forst 35
Frederik Fouvry 35
Anette Frank 43
Sanae Fujita 11
Michael Goetze 27
Claire Grover 51
Tibor Gyimothy 23
Ben Hachey 51
Patrick Hanks 59
Silvia Hansen-Schirra 7.35
Chikara Hashimoto 11
Ian Hughson 51
Kaname Kasahara 11
Valia Kordoni 35
Karin Mueller 75
Shigeko Nariyama 11
Eric Nichols 11
Stephan Oepen 7
Akira Ohtani 11
Fernando Santos Osorio 63
James Pustejovsky 59
Paulo Quaresma 63
Anna Rumshisky 59
Yvonne Samuelsson 67
Jiri Semecky 43
Cassiana Fagundes da Silva 63
Stavros Skopetelas 27
Takaaki Tamaka 11
Carole Tiberius 15
Hans Uszkoreit 7
Renata Vieira 63
Martin Volk 67
Andrea Weber 75