Difficulties with computational coreference tracking: How to achieve ‘coherence in mind’ without a mind?*

Peter Crosthwaite
(University of Hong Kong)

Crosthwaite, Peter. 2015. Difficulties with computational coreference tracking: How to achieve ‘coherence in mind’ without a mind? *Linguistic Research* 32(2), 451-468. The introduction and maintenance of reference in discourse is subject to a number of language-universal constraints on NP form and selection that allow speakers to maintain reference co-referentially (i.e. from mention to mention) with minimal processing effort across large spans of discourse. Automated approaches to co-reference resolution (i.e. the successful tracking of discourse referents across texts) use a variety of models to account for how a referent may be tracked co-referentially, yet automated co-reference resolution is still considered an incredibly difficult task for those working in natural language processing and generation, even when using ‘gold-standard’ manually-annotated discourse texts as a source of comparison. Less research has been conducted on the performance of automated co-reference resolution on narrative data, which is rich with multiple referents interacting with each other, with long sequences of continuous and non-continuous reference to be maintained. Five freely-available co-reference resolvers were trialled for accuracy on oral narrative data produced by English native speakers using a picture sequence as an elicitation device. However, accuracy levels of under 50% for each resolver tested suggest large differences between human and computational methods of co-reference resolution. In particular, zero anaphora, NPs with modifiers and errors in coding first-mentioned referents appear particularly problematic. This suggests that automated approaches to coreference resolution need a vast range of lexical knowledge, inferential capabilities based on situational and world knowledge, and the ability to track reference over extended discourse if they are to succeed in modelling human-like coreference resolution. (University of Hong Kong)

**Keywords** Co-reference resolution, anaphora, reference, discourse, narrative

* I would like to thank the two anonymous reviewers of this paper, whose insightful guidance helped greatly during the revision process.
1. Introduction

Introducing and maintaining reference to discourse entities is a complex linguistic task. All languages use a range of referring expressions that refer to ‘new’ or previously ‘given’ referents in discourse. The number and status of referents in discourse is determined pragmatically according to the needs of speaker, rather than following any fixed generic pattern. Therefore, once a character has been introduced into a discourse, continuing reference to that character needs to be maintained coherently. The following example from Marquez, Recasens & Sapena (2013) explains the task facing the speaker, where the terms in bold/italics refer to the same referent:

*Major League Baseball* sent *its* head of security to Chicago to review the second incident of an on-field fan attack in the last seven months. *The league* is reviewing security at all ballparks to crack down on spectator violence. (2013:662)

Early treatments of referential coherence fell under the umbrella of *cohesion* (Halliday & Hasan, 1976; Hasan, 1984), or ‘the relation between an element of the text and something else by reference to which it is interpreted in the given instance’ (Halliday & Hasan, 1976:308), and texts were judged to be coherent depending on the the number of cohesive *ties* or *chains* they contained, although it has since been shown that frequency of cohesive ties alone is not a significant predictor of coherence (Neuner, 1987; Reid, 1993; Hinkel, 2001). Modern accounts of referential coherence eschew a purely textual approach to coherence, considering the level of mutual knowledge shared in the ‘common ground’ (Stalnaker, 1974; Levinson, 2006; Clark, 2015) between language users in real time, where the coherent transfer of information between the speaker and their audience ‘emerges not in the text, but in two collaborating minds’ (Gernsbacher & Givon, 1995: viii). The choice of linguistic forms used to do this are determined by the discourse-pragmatic needs of the speaker (the ‘intentional’ structure of the discourse) as well as the ‘attentional state’ of the discourse in real time (Grosz and Sidner, 1986:175). Cognitive / discourse-pragmatic accounts of reference such as Gundel, Hedberg & Kacharski (1993), Givon (1995) and Ariel (1991, 1996, 2008, 2010) suggest that languages universally organise their
use of referential forms along scales of cognitive status that reflect the relative salience, identifiability, referential distance, or accessibility (in working memory) of the referent in question against the attentional state of the interlocutors at any given point in the discourse. Generally, this is evidenced through the typical use of shorter referential forms (pronouns, zeros) being used for more accessible, shorter distance (within and between adjacent clause) antecedents, with longer forms (definite NPs, full names) being used for less accessible, more distant antecedents, reflecting language-universal pragmatic principles of economy and relevance (e.g. Grice, 1975; Sperber & Wilson, 1995), and this is seen as ‘critical to discourse and language understanding in general’ (Zhou & Su, 2004:1). Diachronically, such concerns are claimed to have led to the grammaticalisation of reduced NP forms including pronouns and zero anaphora (i.e. the minimize forms principle of Hawkins, 2004).

Speakers are competent at organising their use of referring expressions in a way that allows for the unambiguous, cohesive and coherent tracking of entities in discourse. They are capable of doing this over long stretches of spoken or written discourse without difficulty by using the appropriate referential forms demanded of the intentional and attentional discourse structure. However, in terms of computational linguistics, one of the big challenges for natural language generation and processing (NLP) is the successful tracking of all mentions of an entity in discourse, a field of NLP known as co-reference resolution. Software designed for this purpose (co-reference resolvers) need to find the links (or ‘co-references’) between referring expressions and their antecedents in order to realize all the given expressions co-referential to that entity. However, this process represents ‘a significant engineering effort’ (Versley et. al. 2008) and ‘a highly challenging task […] The mere task of producing the same results as those produced by humans is difficult and largely unsolved’ (Marquez, Recasens & Sapena, 2013:662). The difficulty is sourced in the following ways:

1) Co-reference resolvers typically require a large semantic corpus to recognize the massive amount of possible NPs that can occur in discourse, and may not recognize a named entity that exists outside of this corpus. This problem is ‘acute’ in technical domains (Choi, Verspoor & Zobel, 2014) In addition, complex NP expressions (with many modifiers before the noun) present difficulties at the junction of
syntactic and semantic parsing output, resulting in errors of co-reference resolution.

2) Most co-reference resolvers use syntactic parsers during the pre-processing stage that produce ‘predicted’ accounts of referent number and type that might be different from the true number and type of referents present in a discourse text. Other resolvers are guided by ‘Centering Theory’ (Grosz, Joshi & Weinstein, 1995), the scope of which (in its basic form) is restricted to resolution within, but not between, discourse segments. This means that more ‘distant’ mentions across multiple sentences are most often processed as ‘new’ referents, rather than continuing ‘given’ reference by programs that rely on government and binding principles, and distant mentions across discourse segments (such as paragraphs or episodic shifts) struggle to be resolved through programs that use the principles of Centering theory. The importance of this ‘global inference’ is now increasingly apparent in the field coreference resolution (Cai, Mujdricza-Maydt, and Strube 2011; Lee et. al., 2013).

3) Often discourse texts contain multiple referents of the same gender or class, and while ambiguities can often easily be solved through pragmatic, real-world knowledge by native speakers (via plausibility, semantic roles, bridging descriptions etc.), computational approaches may lack the ability to solve ambiguities in this way (Rahman & Ng, 2011).

4) Certain languages (such as Chinese, Korean etc.) make regular use of topic/pro drop (zero anaphors) in discourse, where there is no surface mention of the referent, yet the referent can be pragmatically inferred from the discourse/deictic context (Crosthwaite, 2014). It is therefore difficult for parsers to resolve zero anaphors where there are no surface forms for a parser to detect.
In recent years, there has been an increasing number of demonstrations of these co-reference resolvers posted online, with accompanying reports quoting high levels of accuracy on manually annotated ‘gold standard’ training texts such as MUC-6, MUC-7 and ACE05 datasets (datasets used as benchmarks for co-reference resolution precision) (Hirschman & Chinchor, 1997; Doddington et. al. 2004; Orasan et. al. 2008). One example is found in Marquez, Recasens & Sapena (2013), where three programs (CISTELL by Recasens, 2010; RelaxCOR by Sapena et. al. 2010; and Reconcile by Stoyanov et. al. 2010) were compared on both ‘gold’ and ‘predicted’ (i.e. manual vs. automatically annotated) corpora, and found differences in performance between the two corpora types. However, less research has been conducted that tests the accuracy of these co-reference resolvers using natural, spontaneous production data instead of carefully-selected and well-trained written benchmark texts. Moreover, Marquez et. al. (2013) found that existing scoring of co-reference resolution can be problematic, citing Finkel & Manning (2008), Luo (2005) amongst others that have found weaknesses related to the ‘true’ number of entities in a text vs. the ‘predicted’ number of entities suggested by the software, or related to ‘mention-based’ or ‘link-based’ measures of scoring repeated reference. By comparing these programs’ performance against a human interpretation of co-reference, it is hoped that the underlying mechanism of the most accurate co-reference resolver may provide clues as to differences between how humans and computational approaches ‘compute’ reference tracking as the discourse unfolds.

2. Method

2.1 Software

The accuracy of 5 co-reference resolvers was compared. These were selected because they were freely available and could be either downloaded or run online. Some of the programs (e.g. Reconcile) allow for experimentation with a number of configurations for best performance, while others do not have this function. For this reason, the ‘default’ configuration was used in each case. A brief description of the programs is provided below:

ARKRef (Haghighi & Klein, 2009) – This approach modularizes syntactic,
semantic and discourse constraints on co-reference resolution to predict co-reference ‘from a deterministic function of a few rich features’ (2009:1152). A syntactic module calculates ‘paths’ from anaphor to antecedent mediated by deterministic constraints using a Treebank as training data. Semantic processing then occurs to determine the suitability of NPs. Finally, co-reference resolution is determined by the anaphor with the shortest tree distance to its antecedent, a technique which is said to outperform the accuracy of parsers that measure resolution through raw text distance (2009:1154). Accuracy rates are claimed to be in the high 70s-80s on the ACE and MUC-6 datasets.

**High Precision Co-reference Resolution in Newswire Domain (HPCRND) (Zhou & Su, 2004)** – This approach uses a ‘constraint-based multi-agent strategy’ (2004:1) for co-reference resolution, parsing the texts using a Hidden Markov Model of syntactic constraints, then using ‘general and special knowledge’ (2004:1) to select antecedents, followed by a final selection based on referential distance. ‘General’ knowledge appears to be defined as morphological and semantic agreements (2004:4) while ‘special’ knowledge appears to be defined as name alias co-reference (ex: Bill Gates & Mr. Gates), apposition, predicate nominal co-reference, and accessibility (defined as whether the anaphor is found in the same paragraph, rather than referring to the attentional state of the discourse as with Grosz & Sidner, 1986), amongst other constraints. The authors claim accuracy in the mid-eighties using the MUC-6 and MUC-7 datasets.

**Proxem Antelope (www.proxem.com, 2009)** – The Proxem Antelope software is a commercial solution for information retrieval and extraction that uses a Wordnet 3.0 lexicon for semantic analysis and the Stanford parser (Klein & Manning, 2003) for syntactic analysis. Antelope stands for Advanced Natural Language Object-oriented Processing Environment. The Stanford parser is used for shallow and deep syntactic processing (under a Chomskyan definition), while the Wordnet 3.0 semantic parser accounts for thematic roles and categorization frames.

**Reconcile (Stoyanov et. al. 2009)** – Reconcile is intended to be a test platform that allows simple reconfiguration according to the needs of any particular co-reference resolution algorithm to be tested. It uses a pre-processor that accounts for tokenizing and POS tagging (from OPENNLP, Baldridge, 2005) and uses the Stanford Parser (Klein & Manning, 2003). Once pre-processing is complete, reconcile produces feature vectors for linked NPs (from a bank of over 80 ‘features’ such as
agreement, number etc.). These features are then classified and clustered by class for final scoring. The authors claim accuracies in the mid-sixties for the MUC-6 and MUC-7 datasets.

**BART – (Versley et. al. 2008)** – Bart is described as a ‘highly modular toolkit’ for developing co-reference applications. It follows a pre-processing stage including Stanford POS tagging (Toutanova et al., 2003), chunking (Kudoh and Matsumoto, 2000), and Charniak & Johnson’s re-ranking parser (Charniak and Johnson, 2005). The program then uses feature extraction as with Reconcile, although the number and type of features are not given in the technical report. The authors claim an accuracy of 73.4% on the MUC-6 dataset.

Two other programs (JavaRAP, Qui et. al. 2004; and Mitkov’s Anaphora Resolution System [MARS], Mitkov et. al., 2010) were also initially tested, but as these only match pronouns to antecedents rather than tracking reference across continued pronoun and full NP reference, these programs will not be mentioned further.

### 2.2 Corpus

4 sample texts were used to compare the performance of the co-reference software. These texts were oral narratives spontaneously produced by native English speakers, and were collected during a previous pilot study into reference introduction and reference maintenance. The stimulus for the narratives was the Edmonton Narrative Norms Instrument (ENNI) (Schneider et. al., 2005), which are sets of picture sequences that are controlled for complexity of story grammar as well as the number of new referents (4 per narrative), and are representative of the narrative genre (having an introduction, problem, and resolution). Texts 1 and 2 used story A3 of the ENNI (a story about a lost model airplane), while texts 3 and 4 used story B3 (a story about a lost balloon). Both stories have four referents to track (two main characters who appear in each picture, and two supporting characters who appear mid-story).

Each text was initially transcribed verbatim from the oral production data. However, as the software would be unlikely to process false-starts, audible pauses (‘ummm’) and self-repairs, these were later removed from the transcriptions, and punctuation was also added to allow the software to parse sentence structure. For
reference, a modified version of text 1 is supplied below:

An elephant and a giraffe meet at the pool. The giraffe takes his toy airplane to play with and shows it to the elephant. After seeing this, the elephant wants to play with the toy airplane and tries to snatch it off the giraffe, but it lands in the swimming pool. The giraffe is angry at the elephant for dropping his toy. The lifeguard comes along and asks what has been going on. The elephant explains what happened and how the toy ended up in the water. The giraffe and the elephant ask the lifeguard to rescue the airplane for them. The lifeguard leans over and tries to reach the toy, but the airplane is too far away. The giraffe starts to cry as it looks like his plane might be stuck in the pool. Another swimmer comes to help out and brings a long net to try and reach the toy. She reaches in to try and scoop the toy out and hands it back to the giraffe. He is happy to have his toy back safe and sound.

Data on the referring expressions used in each text is provided below and is separated into first mentions (‘a giraffe arrived’), within-clause references (W/C) (‘the giraffe threw his airplane), between adjacent clause references (B/C) (‘a giraffe arrived and Ø played’) and distant references (N/CR) (‘a giraffe arrived then an elephant arrived. The giraffe said hello’). These contexts represent the relative cognitive states or referent accessibility from new to highly accessible (W/C, B/C) to low accessibility (N/CR), following the scales of Ariel (1991, 1996, 2008, 2010).

Each text was between 170-190 words in length, with a total length of 725 words across the four texts, with a total of 118 NP target items for resolution across the four contexts mentioned above. The following table describes the NP forms and number of target references used in each co-reference context for each text.
Difficulties with computational coreference tracking: How to achieve...

Table 1. Forms used to introduce and maintain reference in corpus

<table>
<thead>
<tr>
<th>Text</th>
<th>First mentions</th>
<th>W/C Reference</th>
<th>B/C Reference</th>
<th>N/CR Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2 indefinite article</td>
<td>3 possessive pronouns</td>
<td>3 definite article</td>
<td>10 definite article</td>
</tr>
<tr>
<td></td>
<td>1 definite article</td>
<td></td>
<td>1 possessive pronoun</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 quantifier (another)</td>
<td></td>
<td>2 personal pronoun</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5 zero anaphor</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3 indefinite article</td>
<td>3 relative pronouns</td>
<td>5 definite article</td>
<td>8 definite article</td>
</tr>
<tr>
<td></td>
<td>1 quantifier (another)</td>
<td>1 possessive pronouns</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 personal pronoun</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 zero anaphor</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1 numeral NP (‘Two cartoon rabbits’) = 2 mentions</td>
<td>2 possessive pronouns</td>
<td>1 full proper name</td>
<td>5 definite article</td>
</tr>
<tr>
<td></td>
<td>1 indefinite article</td>
<td></td>
<td>3 definite article</td>
<td>1 quantifier</td>
</tr>
<tr>
<td></td>
<td>1 possessive + NP (her father)</td>
<td></td>
<td>2 possessive pronoun</td>
<td>(each)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 personal pronoun</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 quantifier NP (both, neither)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5 zero anaphor</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3 indefinite article</td>
<td>4 possessive pronouns</td>
<td>1 full proper name</td>
<td>9 definite article</td>
</tr>
<tr>
<td></td>
<td>1 possessive + NP (his mum)</td>
<td></td>
<td>9 definite article</td>
<td>1 personal pronoun</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5 personal pronoun</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 quantifier NP (both)</td>
<td></td>
</tr>
</tbody>
</table>

All 16 referents introduced 13 W/C targets 55 B/C targets 34 B/C targets

3. Evaluation criteria

Each sample text was analysed manually for reference introduction and reference maintenance. First-mentioned entities were coded with a number in ascending order as they were encountered in the text, and subsequent mentions of those entities were assigned the same number as their original antecedent. Only reference to animate discourse referents was tracked, as continued reference to inanimate referents is typically shorter in scope (Crosthwaite, 2014). Plural referring expressions (they, those children, both of them, etc.) were coded with the numbers of each referent. Complex definite article reference with relative pronoun use was coded as two
references in the original. Zero anaphors were also included in the count. An example original text with referent coding is provided below:

An elephant (1) and a giraffe (2) meet at the pool. The giraffe (2) takes his (2) toy airplane to play with and ø (2) shows it to the elephant (1). After seeing this, the elephant (1) wants to play with the toy airplane and ø (1) tries to snatch it off the giraffe (2), but it lands in the swimming pool. The giraffe (2) is angry at the elephant (1) for dropping his (2) toy. The lifeguard (3) comes along and ø (3) asks what has been going on. The elephant (1) explains what happened and how the toy ended up in the water. The giraffe (2) and the elephant (1) ask the lifeguard (3) to rescue the airplane for them (1+2). The lifeguard (3) leans over and ø (3) tries to reach the toy, but the airplane is too far away. The giraffe (2) starts to cry as it looks like his (2) plane might be stuck in the pool. Another swimmer (4) comes to help out and ø (4) brings a long net to try and reach the toy. She (4) reaches in to try and scoop the toy out and ø (4) hands it back to the giraffe (2). He (2) is happy to have his (2) toy back safe and sound.

The original text was then supplied as input to the pronoun/co-reference resolver, and the output was compared with the original to gauge the accuracy of the pronoun/co-reference resolution. The percentage accuracy by each program for each text was calculated by dividing the successful matching of NP to antecedent, by the total number of target NPs, multiplied by 100.

4. Results

The results for each text are provided below, followed by a summary across the four texts:
An important point to note is that zero anaphora were not processed by any of the programs used, despite their inclusion in the hand-coded original text. Zero anaphora were left in the hand-coded text analysis as they are considered to be important feature of reference, representing the highest possible accessibility for a given referent. It is a failure of these programs that zero anaphora are not taken into account, and this is reflected in the overall results. The best-performing program on these texts was ARKRef, yet this program’s accuracy was below 60% for either text. In most cases, pronouns and possessive forms were coded as separate (new) referents to their antecedents, so a pronoun such as ‘he’ was coded as a separate ‘first-mention’ referent, and subsequent use of ‘he’ was coded as co-referent to the ‘new’ first mention NP, rather than the original antecedent, e.g.:

‘Tony\textsubscript{1} had a ball. He\textsubscript{2} asked his\textsubscript{2} friend\textsubscript{3} to play with him\textsubscript{2}’.

On the other hand, often actual first mentions of new referents were instead coded as ‘given’ mentions for previous referents, and therefore each subsequent reference to the ‘new’ referent was linked co-referentially to the wrong referent e.g.:

‘Tony\textsubscript{1} had a ball. He\textsubscript{1} asked his\textsubscript{1} friend\textsubscript{1} to play with him\textsubscript{1}’.
On the texts collected using picture sequence B3 of the ENNI, the Reconcile program was the best performer. Text 3 contained a higher number of references to groups of referents than were found in text 4 (such as ‘neither of the rabbits’, ‘each of the children’) but these forms were not resolved well. Many of the first mentions of characters were coded as antecedents of other ‘given’ characters, forcing a high degree of inaccuracy as with texts 1 & 2. ARKRef, which performed well on texts 1-2, performed poorly compared to Reconcile and Bart on texts 3-4.

Reconcile was the best-performing resolver across the 4 texts sampled, but only produced 50% accuracy over all new and given mentions overall. While JavaRAP
and MARS analyzed only pronouns, the HPCRND program did not perform much better across other referential forms.

Reconcile was the best performer in all texts except text 1. What is interesting from the chart is that there does not appear to be a generally higher or lower accuracy found in any one particular text between the 5 programs trialled, with some programs performing better on some texts and worse on others. Each program experienced certain problems as the number and type of references shifted between texts.

5. Discussion

As the best performing program over the 4 texts only achieved 50% accuracy in resolving the range of referring expressions used in the text, it is apparent that the state-of-the-art is some way off in modelling the ability of humans to introduce and track reference to discourse entities across extended discourse. The main failures of these programs and suggestions to improve their performance are as follows:

**Zero anaphora**

Zero anaphora are common in English, and other languages make even more use of them. None of the programs trialled accounted for zero anaphors during
processing, presumably due to the inherent lack of surface information a zero anaphor provides to the parser. An alternative syntactic parser is required that recognizes zero anaphora in discourse and relates them to their antecedents. The same would apply to the use of relative pronouns, which were not recognized as separate references by these programs in their output, yet are still valid referential forms that signal a change (or boost) in referent accessibility.

Pronominal reference across clauses

Personal pronouns and possessive adjectives were very often realized as separate entities from their nominal antecedents, leaving many occasions when the antecedent of a pronoun or adjective was another pronoun realised as a ‘first mention’ (highlighted with 2 in the following example: ‘Tony1 had a ball. He2 asked his2 friend3 to play with him2’). This mostly occurred when the pronoun was found between adjacent clauses, as shown in the example above. However, as the majority of pronouns were found in this context, it is another failure of these systems not to account for this. The problem is likely to be caused by syntactic parsers that rely on government and binding theory due to the constraints on referential distance that such a theory places on co-reference resolution across clauses, i.e. the parsers were not often able to resolve pronouns under binding condition B in this context. The programs that followed a rank order of elements as their basis for pronoun resolution were slightly better at matching pronouns to their antecedents across clauses, but still ran into trouble when resolving reference across larger discourse segments. To solve this problem, it is possible that a resolver might need to analyse reference across the entire discourse until it is apparent that a particular discourse referent will not be referred to again in the discourse, yet this would require parsing for story grammar, which as yet has not been implemented into these systems.

Inappropriate coding of discourse newness

If a new referent is coded as a given referent, the entire chain of reference for that referent will be attributed to the wrong antecedent. This mainly occurred when modifiers were used to distinguish characters of a similar class (e.g. ‘the boy rabbit’, ‘the girl rabbit’, ‘the mother rabbit’). With some programs, this meant that when the full NP existed outside of the lexicon of the program, the program would take only the article and the head noun, code it as a referent, and apply that judgment to all
Difficulties with computational coreference tracking: How to achieve...

instances of that type, regardless of pre-head modifiers. In other cases, the definite article and first modifier (e.g. ‘the boy’) were coded as referents outside of the full NP (‘the boy rabbit’). These errors also lead to certain N/CR definite article forms taking the wrong antecedent after processing. Better recognition of complex NPs is needed at the syntactic and semantic levels of processing to avoid this kind of error.

6. Conclusion

From the results, all of the potential difficulties for these programs outlined in the rationale were found in the output of these programs. The best performing program approached 50% of what a competent native speaker does automatically as they read or listen to discourse texts. The native speaker is able to hold referents in working memory over discourse segments of much longer lengths than are processed by these programs. Native speakers are also able to recognize both long and complex forms of reference (accurately processing their syntactic and semantic content) as well as minimal and zero referential forms that allow the native speaker to realise the antecedents of these forms at a minimal cognitive cost. They thus have a lexical, pragmatic and cognitive advantage that automated approaches currently lack. Therefore, attempts to improve the state-of-the-art of co-reference resolution need to take these factors into account if they are to approach a human-like capability.

References


Difficulties with computational coreference tracking: How to achieve...


Recasens, Marta. 2010. Coreference: Theory, annotation, resolution and evaluation. Ph.D.
Dissertation, University of Barcelona, Barcelona, Spain


Schneider, Phyllis, Rita V. Dubé, and Denyse Hayward. 2003. The edmonton narrative norms instrument. Retrieved from University of Alberta Faculty of Rehabilitation Medicine. Website: http://www.rehabresearch.ualberta.ca/enni (27.2. 2015)


---

**Peter Crosthwaite**

Centre for Applied English Studies  
University of Hong Kong  
Room 6.38 Run Run Shaw Tower, Pokfulam, Hong Kong SAR  
E-mail: drprc80@hku.hk

Received: 2015. 01. 15.  
Accepted: 2015. 07. 30.