# RefRef: A Tool for Viewing and Exploring Coreference Space

**Hisami Suzuki and Gary Kacmarcik**

Microsoft Research
One Microsoft Way, Redmond WA 98052 USA
{hisamis,garykac}@microsoft.com

**Abstract**

We present RefRef, a tool for viewing and exploring coreference space, which is publicly available for research purposes. Unlike similar tools currently available whose main goal is to assist the annotation process of coreference links, RefRef is dedicated for viewing and exploring coreference-annotated data, whether manually tagged or automatically resolved. RefRef is also highly customizable, as the tool is being made available with the source code. In this paper we describe the main functionalities of RefRef as well as some possibilities for customization to meet the specific needs of the users of such coreference-annotated text.

## 1. Introduction

There have been many annotated resources created for the task of coreference resolution, such as the MUC and the ACE data sets, which have undoubtedly contributed to active research on the topic over the last decade. Tools that facilitate the creation of coreference-annotated corpora have also been developed in the past for many different languages. Some tools specialize in assisting the annotation process of coreference relations (e.g., Aone and Benett, 1994; Müller and Strube, 2003; Kawahara et al., 2005); others also support the comparison of coreference tags for evaluation purposes (e.g., Aone et al., 1998; Humphreys et al., 1998), or provide an environment for displaying a subset of tag sets (Day et al., 1997) as the annotated text typically consists of many layers of linguistic information, including the results of tokenization, chunking and named entity identification processes. These tools have played as important a role as the annotated data itself in advancing the state of the art of coreference resolution techniques.

Just as coreference resolution itself is a challenging task; it is equally challenging to find ways to utilize the resolved coreference relations for particular purposes. The tool presented in this paper, called RefRef, is a tool that is specifically intended to facilitate such data exploration. It does not offer tagging assistance; instead, it focuses on giving visually intuitive feedback on the already annotated data, whether manually tagged or automatically resolved, which is useful for performing error analyses or for simply viewing the annotated coreference relations in context. As the effectiveness of text exploration hinges on the ability to restrict the display to focus on the relevant part of the data, RefRef is designed to be highly customizable, and is made available with the source code (for research purposes).

## 2. Tool description

### 2.1. Displaying coreference relations from pre-annotated text

Figure 1 shows a basic display of an annotated English text from MUC-6. On the left side, we have the Text Pane where the original text is displayed. Those mentions that are annotated for coreference are indicated in blue. When mentions are nested within one another, a pair of brackets is used to indicate the outer mention. When any of these blue expressions are clicked, all mentions that are coreferential are highlighted in pink (in this case, the entity "the government"). This display is useful in getting a general overview of how coreferential nouns and pronouns are distributed in text.

On the top right, an automatically extracted list of entities in the corpus is displayed, using the name of the first (non-null) mention of the entity in the text. When one of these entities is selected, all the mentions associated with the entity are highlighted in pink in the text pane. The entities are sorted in the order they are first mentioned in text.

It is also possible to restrict the view to a specific subset of entities in a given text. For example, by choosing an option from a drop-down menu, one can only display the entities that are mentioned at least once by a pronoun. As pronouns play an important role in defining the information structure of text, such an option provides an additional utility in exploring the annotated data. The tool can also support other lists as a filter – for example, one can selectively display coreference relations that include only personal pronouns or zero pronouns, etc. The creation of such a list is driven by a specific need of the application which uses the coreference information.

RefRef has a native support for displaying commonly used annotations such as MUC-6 and 7; it also supports the Kyoto Corpus format for coreference annotation (Kawahara et al., 2005), to be discussed in the next subsection. Since source files are specified using Unicode (UTF-8), it accommodates a wide range of languages.

### 2.2. Displaying coreference and related information from the Kyoto Corpus

The Kyoto Corpus (version 4.0) has recently been made available with coreference annotations for 5,000 sentences. Displaying information annotated in the Kyoto Corpus poses new challenges for two reasons: first, as it is an annotated text of Japanese coreference relation, there is an extensive use of zero pronouns, i.e., mentions whose surface forms are null. These elements need a special attention so that they are displayed properly. Secondly, the Kyoto Corpus does not distinguish nominal and pronominal coreference relations from tagging surface case relations in general, whether the case elements (similar to arguments of a predicate except that they are defined by surface case relations rather than by grammatical functions) are present or absent in the surface structure. This is very reasonable in the context of

Japanese, where zero pronouns, which are functionally equivalent to English pronouns, are indistinguishable from other surface-null elements such as traces of movements (via relativization and topicalization) and the gap created by the deletion of shared elements in coordination, i.e., the type of information that are provided by a different set of annotations in English (e.g., the function tags of Penn Treebank, Bies et al., 1995). In that sense, the annotation provided by the Kyoto Corpus constitutes a superset of coreference annotation provided by such resources as MUC, and serves as a comprehensive set of annotation for the task of recovering complete predicate-argument structure for text understanding.

For the visual exploration of the Kyoto Corpus, we have chosen to display only the arguments of verbal and adjectival predicates in the main text window, suppressing the display of the arguments of nominal predicates. For example, in the third sentence displayed in Figure 2, we show the full list of arguments for 加えた "*added*", including zero pronouns indicated by φ followed by the case markers which indicate the case of the zero pronouns. As zero pronouns are by definition coreferential with another element, it always appears in a coreference chain. In Figure 2, the chain for 軍 "*troops*", which includes the *ga*-marked zero pronoun for 加えた "*added*", is highlighted in pink in Figure 2. Currently, we automatically place zero pronouns immediately preceding the predicate, as the location of zero pronouns is not specified in the annotation of the Kyoto Corpus.

Note that in Figure 2, we do not display the case relations for 制圧 "*control*" in the second sentence, which is tagged for two cases (*ga* and *wo* cases) in the Kyoto Corpus annotation. This is because 制圧 in this context is used as a noun. This decision was made simply for the sake of display so that the main text window displays the original text as is with minimal disruption. Predicate-argument structures including nominal arguments can be displayed in the right bottom pane, which is customizable to display additional information.

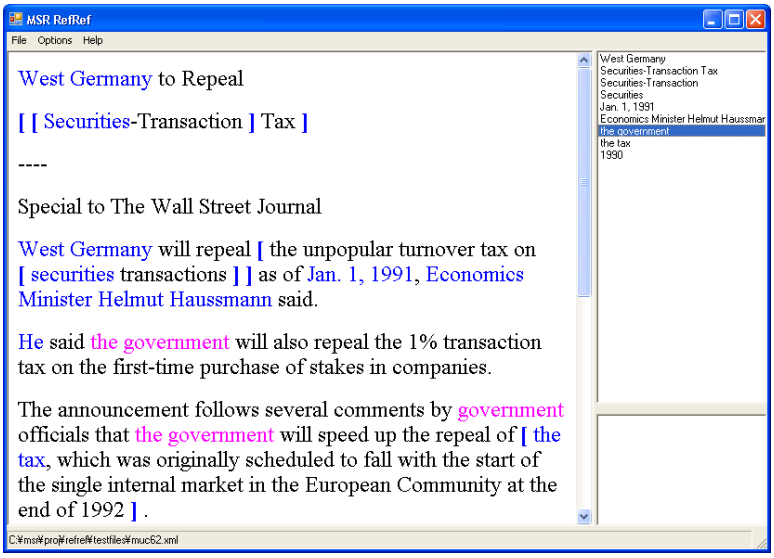


**Figure 1** : Displaying MUC-6 annotation. Mentions are displayed in blue and the currently selected mention is shown in pink. Square brackets are used to show nested mentions.

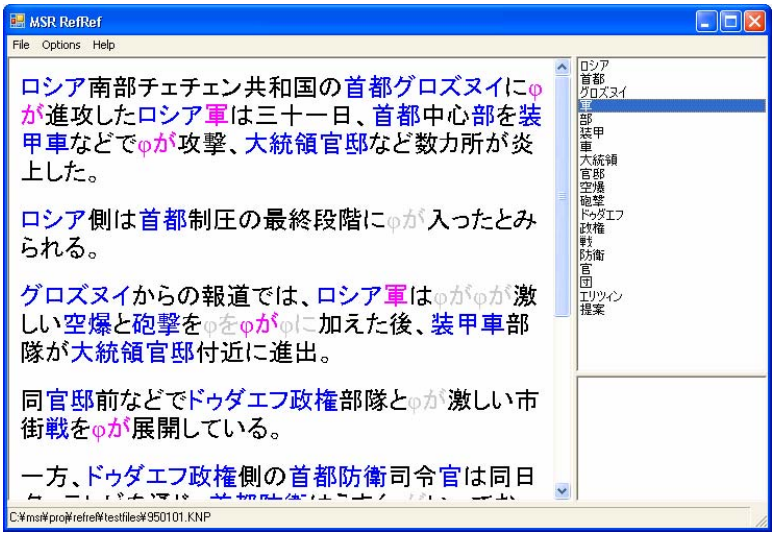


**Figure 2** : Visual exploration of a sample from the Kyoto Corpus. Zero pronouns are shown in gray / pink when highlighted.

### 2.3. Displaying application-specific information

RefRef's primary functionality is to facilitate visual exploration of coreference-annotated text, regardless of whether the annotation comes from hand-annotated corpora (such as the MUC or Kyoto Corpus data), or from an automatic resolution module. An example of the latter is shown in Figure 3, which displays the results of the pronominal coreference resolution module within our natural language understanding system (Heidorn, 2002). The right bottom pane is used for displaying custom information provided by further processing the data based on coreference relations. In this example, predicate-argument structures for verbal predicates are displayed for the entities that are mentioned at least once by a pronoun, which provides a more succinct view of the text on the left pane, serving as a kind of summary of the original text. Note that RefRef is not extracting this information from the text, but is simply displaying information that was created externally and encoded in the annotation file. RefRef does, however, maintain the relationship between the custom data and the source text to facilitate navigation, for example, by highlighting the source text when a predicate-argument relation is selected, as shown in Figure 3.

This information is encoded by extending the MUC file format to include an extra `<info>` section. Each entry in the info section consists of the text to display and the sentence ID, from which RefRef manages the display to permit the aforementioned interaction.

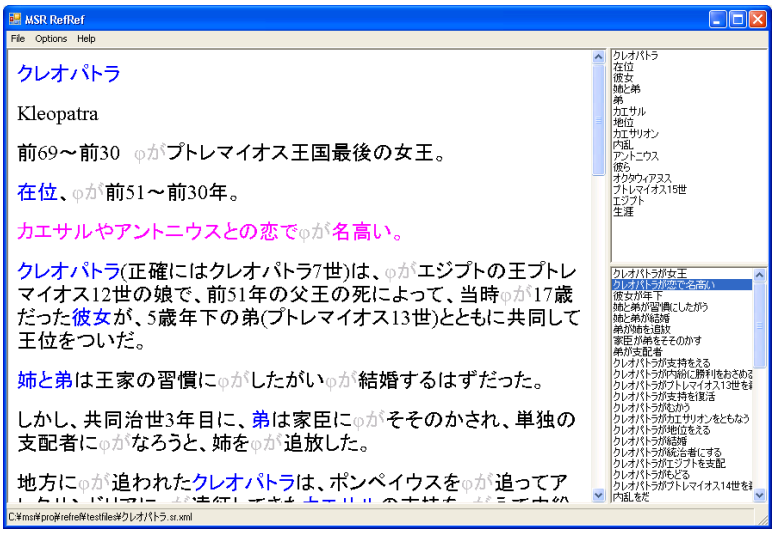


**Figure 3** : Displaying annotation created by an automatic coreference resolution module for Japanese. Zero pronouns have been added during annotation. The sentence corresponding to the selected predicate-argument relation (second item in lower right pane) is highlighted in pink.

## 3. Source code

As it is difficult to anticipate all possible features for this type of tool, we are providing this tool with source code. This will allow researchers to customize the tool to add support for tags and formats that are specific to their work.

The main display pane in the application hosts a web-browser control, so the full richness of HTML is available: tables, images, links and other HTML elements are all natively supported by this control.

The current version of RefRef uses only basic HTML formatting to support the selection of coreference chains. Each chain is assigned its own style in the document's stylesheet and all references in the document are enclosed in custom `<a href="ref:id=n"> <span class ="cn">` tags, where *n* is the ID of the coreference chain. The `ref:` URL format is custom to RefRef and is specially handled to dynamically update the stylesheet so that the entire reference chain containing the selected reference is highlighted.

The reliance on standard HTML-based formatting makes the program far more amenable to modification by other researchers. The code demonstrating the handling of MUC- and Kyoto Corpus-style files provides a useful starting point for those interested in supporting their own custom formats.

## References


Aone, Chinatu, Scott W. Bennett. 1994. Discourse tagging tool and discourse-tagged multilingual corpora. In *Proceedings of International Workshop on Sharable Natural Language Resources (SNLR)*, Nara, Japan.

Aone, Chinatsu, Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz. 1998. SRA: Description of the IE2 System Used for MUC-7. In *Proceedings for MUC-7*, Fairfax, Virginia.

Bies, Ann, Mark Ferguson, Karen Katz and Robert MacIntyre. 1995. *Bracketing Guidelines for Treebank II Style*, Penn Treebank Project.

Day, David, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson, and Marc Vilain. 1997. Mixed-initiative development of language processing

systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp.348-355.

Heidorn, George. 2000. Intelligent Writing Assistance. In R.H. Moisl and H. Somers (eds.), *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*, Chapter 8, Marcel Dekker, New York.

Humphreys, K, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. 1998. University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In *Proceedings for MUC-7*.

Kawahara, Daisuke, Ryohei Sasano, Sadao Kurohashi and Koichi Hashida. 2005. *Specification for annotating case, ellipsis and coreference. Kyoto Text Corpus Version 4.0* (In Japanese).

Müller, Christoph, and Michael Strube. 2003. Multi-level annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pp.198-107.