Inter rater agreement with CESAC

Erwin R. Komen
June 15, 2009

1. Introduction

The program CESAC is used to add coreference information to existing syntactically annotated corpora (Komen, 2009a). In the first half of 2009 several people[1] have been coding one and the same Old English text, in order to find out how well different coders agree on the information they add to the text. This paper describes the process of enriching the existing data, the causes of differences between coders, and the interrater agreement between three of the coders.

2. Enriching the data

The data used as input to the CESAC program consist of English texts from different time periods, which are all syntactically annotated according to the Penn Treebank format.

Cesac allows the user to establish coreference links between one kind of phrase (e.g. a noun phrase, a possessive pronoun) and another kind of phrase (e.g. another noun phrase, an IP etc.). Also, the user has to supply a category for the coreference link thus established. The list of categories is limited, and consists only of terms that are agreed upon by the coders.

3. Coding differences

The first difference between the coders was in the amount of coding done. Between two coders I have only evaluated the XPs where both coders had given positive input.

The second difference between coders lies in the phrasal categories actually chosen for coreferencing. These categories can be selected in Cesac under Tools/Settings. The agreed upon categories are shown in Table 1. The rules in this table give the following information. Any IP can serve as a destination for a coreference relation. Any NP or PRO$ (possessive pronoun) can serve as a source or a destination (target) for a coreference relation. Then there are two categories that *must* be supplied with coreference information: NPs containing a pronoun, and NPs containing a determiner or demonstrative.

Not all coders had their phrasal category rules set in the same way, however. One rater had allowed pronouns (i.e. phrases of type PRO*) to serve as source and destination of coreference relation. This poses a problem for calculating interrater agreement, since most pronouns have their own node in the syntactic tree, with its own absolute node ID number. However, unless a pronoun is a possessive pronoun (the PRO$ category), it is also part of an NP, which has a different absolute node ID number. So, while this rater could have the same coreference relation between a pronoun and some other target as one of the other raters had, their would actually be a difference in source and/or destination node ID numbers, resulting in a mismatch between raters. Unfortunately IRAT, the program used to do some preprocessing before calculating the interrater agreement, could only partly compensate for this difference (Komen, 2009b).

[1] There were 5 raters: Monique Tangelder, Gea Dreschler, Rosanne Hebing, Bettelou Los and Erwin Komen. The results of three of them have been taken into account in this paper.

Table 1 Agreed upon coding categories

Node	Child	Type	Target
IP*	*	Can	Dst
NP*	*	Can	Any
PRO$	*	Can	Any
NP*	PRO*	Must	Any
NP*	D^*	Must	Any

Finally, differences between coders can be found in the actual coreferencing. These differences could be subdivided as follows:

a) One of the coders did not establish a reference from a phrase, whereas the other did.
b) Both coders supplied a coreference link, but they chose a different destination.
c) Both coders supplied a coreference link, but chose a different coreference type.
d) Both coders supplied a coreference link, but the destination was different, and the coreference type was different too.

It would be profitable to differentiate between these four categories of inter rater disagreements.

4. Measuring agreement

Three coders made their data available to me, and I have subsequently calculated the interrater agreement between them. As measure for agreement Cohen's kappa has been taken. Table 2 shows the interpretation of this value. Based on this table I propose to accept a kappa of .81 or higher as sufficient for the coreference annotation work on English corpora.

Table 2 Interpretation of Cohen's kappa

Value	Interpretation
.00 - .20	slight
.21 - .40	fair
.41 - .60	moderate
.61 - .80	substantial
.81 - 1.0	almost perfect

The output provided by CESAC was not directly usable for the measurement of Cohen's kappa. I have used IRAT to make sure that the correct amount of phrases were compared between coders. Furthermore IRAT provided some harmonization to counter the effect of one of the coders using slightly different settings (see section 3). This effect could not be completely compensated for, however.

The interrater agreement was calculated using Cohen's kappa. However, it was not possible to use SPSS for the purpose of calculating this statistical measure. The problem was that each coder had some values for the coreference distance and/or the coreference type, which the other coders did not have. SPSS does not allow for this situation—it wants all coders to at least have used the same values once. So instead of SPSS the internet tool ReCal was used (Freelon, 2008).

Table 3 shows the results of the interrater agreement measurements of the coreference distance between the three coders. The agreement between coder RM and EK is 78 %, yet receives a kappa of .629, so can be described as "substantial". The 72% agreement between EK and MT receives a kappa of .267, while the 60% agreement between MT and RM results in a kappa of .198. These values are too low for a good agreement. But, as has been discussed above, the larger differences are mainly due to the different categories chosen to enrich with coreference information.

Table 3 Coreference distance agreement

	EK + RM	**EK + MT**	**RM + MT**
# phrases	1762	1614	1614
% agreement	77.6 %	71.7 %	59.5 %
Cohen's kappa	0.631	0.267	0.198

Table 4 shows the results of the interrater agreement measurements of the coreference *types* between the three coders. Both the agreement percentages and the kappa values are larger than those observed for the inter rater agreement of the coreference distance.

Table 4 Coreference type agreement

	EK + RM	**EK + MT**	**RM + MT**
# phrases	1762	1614	1614
% agreement	80.1 %	79.6 %	67 %
Cohen's kappa	0.613	0.425	0.238

After calculating the percentage agreements and the Cohen's kappa values, the program IRAT made a more detailed analysis of the kinds of disagreements (see section 4). Table 5 gives the results of this analysis. The analysis shows that there are relatively many instances where coder #1 has a coreference, but #2 has not and vice verse. The actual amount of disagreement between coders, in the sense that one coder choses a different target for the coreference relation or a different type of coreference relation, is much smaller.

Table 5 Types of disagreements between the coders

		EK-RM				EK-MT				RH-MT			
		Distance		Type		Distance		Type		Distance		Type	
Agree		77,7%	1369	80,1%	1412	75,5%	1218	79,6%	1284	64,2%	1036	68,5%	1105
	#2 empty	1,6%	28	1,6%	28	18,4%	297	18,4%	297	30,4%	490	30,3%	489
Disagree	#1 empty	14,9%	262	14,8%	261	1,4%	22	1,4%	22	0,5%	8	0,5%	8
	other	5,8%	103	3,5%	61	4,8%	77	0,7%	11	5,0%	80	0,7%	12

5. Conclusions

This paper describes how far different people agree on enriching a syntactically annotated text with coreference information using the program Cesac. One text, the story of Apollonius, was chosen as a testcase. The results of three coders have been compared for this text. The results show that the agreement is still lower than the target for this project (Cohen's kappa was not higher than 0.63, while an acceptable kappa would be at least 0.81).

There are several reasons for the interrater disagreements. First, one of the coders used different settings, which made comparisons more difficult, and not completely realistic. Second, the amount of phrases enriched with coreference information by one coder, but left untouched by another coder, was unacceptably high.

In view of the results I would like to make the following recommendations:

(a) An even more detailed analysis should be made to reveal the situations where and why there is disagreement between the three coders.

(b) As a result of this analysis a common coding practice should be developed with examples and recommendations, stating:
   i. When a coreference relation should be made
   ii. Under what circumstances what type of relation should be made
   iii. What our common set of coding categories should be.

(c) We should discuss this common coding practice, and then make another attempt at coding one text and determining the interrater agreement.

(d) Only then should we continue to code other texts.

6. References

Freelon, D. (2008). ReCal: reliability calculation for the masses [Electronic Version] from http://dfreelon.org/utils/recalfront/.

Komen, E. R. (2009a). *CESAC: Coreference Editor for Syntactically Annotated Corpora*. Nijmegen: Radboud University.

Komen, E. R. (2009b). IRAT: an inter rater agreement tool. Nijmegen: Radboud University.