Inter-annotator Agreement on a Multilingual Semantic Annotation Task

Rebecca Passonneau, Nizar Habash, Owen Rambow

Center for Computational Learning Systems, Columbia University
New York, NY USA
(becky|habash|rambow)@cs.columbia.edu

Abstract

Six sites participated in the Interlingual Annotation of Multilingual Text Corpora (IAMTC) project (Dorr et al., 2004; Farwell et al., 2004; Mitamura et al., 2004). Parsed versions of English translations of news articles in Arabic, French, Hindi, Japanese, Korean and Spanish were annotated by up to ten annotators. Their task was to match open-class lexical items (nouns, verbs, adjectives, adverbs) to one or more concepts taken from the Omega ontology (Philpot et al., 2003), and to identify theta roles for verb arguments. The annotated corpus is intended to be a resource for meaning-based approaches to machine translation. Here we discuss inter-annotator agreement for the corpus. The annotation task is characterized by annotators' freedom to select multiple concepts or roles per lexical item. As a result, the annotation categories are sets, the number of which is bounded only by the number of distinct annotator-lexical item pairs. We use a reliability metric designed to handle partial agreement between sets. The best results pertain to the part of the ontology derived from WordNet. We examine change over the course of the project, differences among annotators, and differences across parts of speech. Our results suggest a strong learning effect early in the project.

1. Introduction

Six sites participated in the Interlingual Annotation of Multilingual Text Corpora (IAMTC) (Dorr et al., 2004; Farwell et al., 2004; Mitamura et al., 2004).[1] The six sites are Carnegie Mellon University (CMU), Columbia University (CU), Information Sciences Institute (ISI), Mitre, New Mexico State University (NMSU), and University of Maryland (UMD). English translations of news articles from Arabic, French, Hindi, Japanese, Korean and Spanish were annotated by up to ten annotators. The annotation task was to match open-class lexical items (nouns, verbs, adjectives, adverbs) to one or more concepts taken from the Omega ontology (Philpot et al., 2003), and to identify theta roles for verb arguments. The annotated corpus is intended to be a resource for meaning-based approaches to machine translation and other natural language technologies.

Parameters of the IAMTC project that define the agreement issues include: sets of concepts or theta roles as the values that annotators select, an unbounded number of labels for concepts, and different sets of annotators on different datasets. In this paper, we discuss inter-annotator agreement with a focus on three questions: (a.) how to quantify agreement for set-valued items; (b.) how to rate individual annotators; and (c.) how to determine if the IAMTC annotations are reliably annotated.

In section 2, we describe the annotation tasks and datasets. In section 3, we discuss related work on inter-annotator agreement measures, and suggest that in pilot work such as this, agreement measures are best used to identify trends in the data rather than to adhere to an absolute agreement threshold. In section 4, we motivate the use of a metric for Measuring Agreement for Set-valued Items (MASI) (Passonneau, 2004). In Section 5, we present our results, showing an apparent learning effect as reflected in narrower differences between the two translations of the same documents later in the project. We also show how agreement can be used to prune less reliable annotators. The facts that agreement increases over time, and that some annotators are particularly proficient, are evidence that the IAMTC task can be performed reliably.

2. Annotation Tasks

The six sites collaborated on annotation procedures and instructions, but independently supervised annotators. There were six pairs of translations into English, or twelve document sets, as listed in column one of Figure 1. Although the original goal was to have twelve annotations with two contributed by each site, some annotators were unable to complete their assignments.

Annotators were given initial onsite training. They used a specialized graphical user interface (GUI) to access the dependency tree parses of the sentences to be annotated, the Omega ontology, and a list of possible

[1] http://aitc.aitcnet.org/nsf/iamtc/

theta roles. The Omega ontology combined two pre-existing ontologies, WordNet (Fellbaum, 1996), and Ontosem/Mikrokosmos (Nirenburg et al., 1996). The first task was to select one or more concepts for each lexical node in a dependency parse. The second task was to assign theta roles to those nodes that had grammatical roles associated with them.

The upper table of Figure 1 shows the two English translations for one of the Hindi sentences. We show the dependency parse of the second sentence, but for simplicity of presentation, nodes from the parse are presented in tabular form rather than as a tree. The portion of the table bordered by a double line (columns 5, 6, and 7) shows the annotation choices made by annotator 2 (H2E2), one of the annotators who rated highest according to a ranking method we describe in Section 5. Note that symbols like "|", "<", "$" and "-" in the concept columns are part of the spelling of a single concept "name." Where there are entries on two lines within a concept column, the annotator chose two concepts, e.g., `|rate|` and `|pace>beat|` for the lexical item *rate*. The special concept `|DummyConcept|` is used to indicate that no proper concept was found in the ontology. The presence of "none" in the theta role column indicates the annotator decided none of the theta roles applied. The presence of "---" in any entry indicates the annotator made no selection, and this was an option for concepts or theta roles. All nodes were included in the computation of inter-annotator reliability, including those where annotators made no selection.

3. Background and Related Work

The use of datasets annotated by humans as training data for machine learning tools has accelerated the concern in computational linguistics for assessing the reliability of annotated data. As noted by Di Eugenio (2000) and Di Eugenio & Glass (2004) there has been an unfortunate tendency towards a cookbook approach in assessing reliability, meaning the assumption that there should be a single best metric, and an absolute threshold value that should be achieved, independent of the goals and constraints particular to each annotation project.

Inter-annotator reliability metrics can support other types of inference besides evidence for or against the reliability of the data for a particular purpose. In the data presented here, we use reliability metrics in a contrastive manner, to identify subsets of the data that are more or less reliable. We find, unsurprisingly, that some annotators are more consistently reliable than others, that some datasets were coded more reliably than others, and that different subtasks had greater reliability.

3.1. Choice of Metric

Typically, annotations are assessed by arranging the data in an **i** by **j** matrix of the observed annotation choices, where the **i** rows represent the **i** coders, the **j** columns represent the **j** units being coded, and each cell **(i,j)** contains the **kth** value (or category) that the ith coder chose for the **j**th unit. Thus each cell represents the decision made for a single item. In the case of the IAMTC data, as with many recent annotation efforts, the decision can have multiple parts.

The observed proportions of each of the **k** values are used to calculate the cell values that would be given by a chance distribution. A single probability distribution can be used for all coders, as in Siegel & Castellan's K (1988) or Krippendorff's Alpha (1980), or a separate one can be used for each coder, as in Cohen's Kappa (1960). As noted by Di Eugenio & Glass (2004), the choice of metric, and how to interpret the results, should depend on issues such as whether the annotations are skewed towards a small set of values (prevalence in the data), or whether coders make very different selections (bias in the data). Here we will use Krippendorff's Alpha because it allows for a distance metric to scale differences in a pair of values in an agreement matrix (cf. Passonneau 2004, 2006; Passonneau et al., 2005). As illustrated in the next section, whether coders agree or disagree is not always a binary question. This is particularly true when annotators are asked to make a decision with multiple components.

Artstein & Poesio (2005) indirectly suggest that since a reliability metric is simply a measure of association, there is no reason to be wedded to a single representation method, such as the type of matrix just described, or to a single metric. They propose a reliability metric that combines the probability estimation of (Cohen, 1960) with the type of distance metric proposed by (Passonneau 2004). In addition, they review entirely distinct approaches, such as the latent class analysis methods used by (Uebersax, 1988; Bruce & Weibe, 1998). The more the annotation task differs from the conventional model, the more necessary it may be to select or invent alternative metrics. Thus (Rosenberg & Binkowski, 2004) propose an augmented Kappa metric for a task in which coders were allowed to make a primary selection, and a secondary selection.

3.2. Interpreting Agreement

The agreement metrics discussed in the preceding section take on values ranging from one to very close to minus one. Values of zero represent no deviation from chance distribution. The closer the value is to one, the more support there is to conclude that similarities or differences across annotations are not due to accident.

Despite Krippendorff's advice to consider the cost of disagreements when interpreting results, advice which has been repeated and expanded upon by Di Eugenio & Glass (2004), there is still a tendency to rely on the 0.67 threshold suggested by Krippendorff. Ironically, Krippendorff offered the threshold of 0.67 only to exemplify the complexity of the issues, in the context of arguing against "ad hoc" standards, and against applying standards across the board. He was referring to a set of studies in which 0.67 was used to report phenomena supporting "cautious conclusions", and 0.8 for solid results. In significance tests of correlations using variables from the same data, significance was rarely achieved for variables with agreements of less than 0.70. However, he pointed out that this will not be the case in all datasets: “some content analyses are very robust in the sense that unreliabilities become hardly noticeable in the result.”

For the question of how reliable is reliable enough, Krippendorff says “there is no set answer” (p. 146). As with statistical inference in general, it depends on the uses the data will be put to. If there is no single use, which is the intended situation for the IAMTC data, there is no single answer. In a study of data from the 2005 Document Understanding Conference (Passonneau et al., 2005), we point out that datasets for computational linguistic applications are often assembled independent of a specific application, or are intended for multiple applications. As a consequence, it is necessary to resort to general criteria, such as those proposed by Krippendorff, to *begin* addressing the question of whether annotations are reliable. But the reliability analysis should not stop there.

In (Passonneau et al., 2005), we faced a situation more parallel to the one that gave rise to Krippendorff’s 0.67 threshold. We presented inter-annotator reliability results on pairs of annotators for a sample of semantic annotations of machine generated summaries that were evaluated against summarization models we refer to as pyramids. Because the annotated data was used to score the peer summaries, we had an independent *cost* measure, consisting of the correlation of scores of the same summaries from different peer annotations. While reliability measures met the 0.67 threshold, the major finding pertained to the relationship between the reliability scores and the statistical significance of the score correlations. Scores were very highly correlated, indicating that the reliability was more than sufficient to engender confidence in the scores. This is precisely the type of cost analysis Krippendorff was referring to, and that we cannot do for the IAMTC data.

4. Measuring Agreement on Set-valued Items

As explained above, annotators were allowed to select multiple concepts or roles if a single selection seemed insufficient. Table 2 shows an example of a token of the lexical item “*cost*” that was assigned WordNet concepts by nine annotators. Five annotators selected a singleton set, one selected a superset with two members, and three selected a larger superset with three members.

As discussed in (Passonneau, 2004; Passonneau et al., 2005), Krippendorff’s α (1980) allows a weighted comparison of values that can be adapted to count the three values in our example as partly alike, rather than wholly dissimilar. For very large samples, Alpha is quivalent to Scott’s (1955) pi; it corrects for small sample sizes; and generalizes to many scales.

The formula for Alpha, given m coders and r units, is:

$$\alpha = 1 - \frac{rm-1}{m} \frac{\sum_i \sum_b \sum_{c>b} n_{bi} n_{ci} \delta_{bc}}{\sum_b \sum_c n_b n_c \delta_{bc}}$$

The numerator is a summation over the product of counts of values b and c, for all pairs of values, times the distance metric δ, within rows. For categorical scales, because Alpha measures disagreements, δ is 0 when b=c, and 1 when b ≠ c. The denominator is a summation of agreements and disagreements within columns.

Number of annotators	WordNet Concepts selected
5	COST
1	COST, MONETARY_VALUE
3	COST, MONETARY_VALUE, TOLL<VALUE

Table 1. Partial annotation of a lexical item “cost”

For set-valued scales, we use MASI for the distance metric δ. It is equal to $1\text{-}J_{bc}*M_{bc.,}$ and ranges from 1 to 0. Briefly, J is the Jaccard (1908) metric for comparing two sets: a ratio of the cardinality of the intersection of two sets to their union. M is a four-point scale that takes on the value 1 when two sets are identical, 2/3 when one is a subset of the other, 1/3 when the intersection and both set differences are non-null, and 0 when the sets are disjoint. See the companion paper (Passonneau et al., 2006) in this proceedings for the motivation behind this scale. MASI becomes closer to 1 as two sets have more members in common and are more nearly equal in size.

WordNet			
DocSet	#annotators minus 1	# lexical nodes	Alpha/MASI (pair delta)
Arabic1	6	80	.66 (-0.15)
Arabic2	9	97	.51
Korean1	9	112	.50 (-0.15)
Korean2	9	92	.45
Japanese1	9	111	.66 (0.00)
Japanese2	9	117	.66
Spanish1	8	116	.60 (0.01)
Spanish2	8	124	.61
French1	9	130	.54 (-0.05)
French2	9	136	.49
Hindi1	8	77	.61 (-0.02)
Hindi2	8	76	.63
Mean			.58

Table 2. Alpha values using the MASI distance metric for WordNet concept annotations. For each dataset, the worst annotator was eliminated.

5. Results

5.1. Interannotator Agreement

We present overall inter-annotator results for the WordNet portion of the Omega ontology in Table 1, for Mikrokosmos in Table 2, and for theta roles in Table 3. Because different sets of annotators were involved in each dataset, the tables list results separately for each dataset (paired translations from Arabic, French, Hindi, Japanese, Korean, Spanish). The number of annotators is shown in column two, but not the identity. The nine annotators who coded French2, for example, are not the same as the nine coders for Arabic2. Column 3 gives the number of lexical nodes, which is a rough measure of the relative scope of the task across document sets. Finally, the double lines separate document sets that were annotated earlier versus later.

Among the eleven annotators across all sites, one was relatively uncooperative, often failing to complete the task, and another had much more difficulty than the remainder. Table 1 presents inter-annotator agreement for each of twelve datasets, and the mean of .58. In all cases, we dropped the coder with the lowest average pairwise reliability, largely because we found that this coincided with failure to complete a large portion of the document set. (The mean reliability for all coders on all document sets was .49.)

Mikrokosmos			
DocSet	#annotators minus 1	# lexical nodes	Alpha/MASI (pair delta)
Arabic1	6	80	.30 (-.01)
Arabic2	9	97	.29
Korean1	9	112	.40 (-.09)
Korean2	9	92	.31
Japanese1	9	111	.48 (0.00)
Japanese2	9	117	.48
Spanish1	8	116	.35 (0.04)
Spanish2	8	124	.39
French1	9	130	.28 (0.02)
French2	9	136	.20
Hindi1	8	77	.39 (0.00)
Hindi2	8	76	.39
Mean			.36

Table 3. Alpha values using the MASI distance metric for Mikrokosmos concept annotations on each document set. For each dataset, the worst annotator was eliminated.

Agreement values in Table 1 are almost always .5 or above, which means that annotator responses are halfway between chance and perfect agreement or better. The number in parentheses after the reliability score for the first of each translation pair indicates the difference in reliability between different translations of the same source documents. Given that translations are semantically very close, it is surprising to see such large deltas for the first phase of the project (Arabic, Korean). This delta almost disappears for all subsequent pairs of translations. We speculate that during the first phase of the project, annotators were still learning the task, the ontology, and the GUI.

Another difference one can see in Table 1 is that reliability scores seem to increase over time, apart from the French set. Again, we can only speculate as to the underlying reason, but note that the French set has the largest number of nodes to annotate. It is possible that annotators on these document sets had a lower rate of task completion, due to the greater size of the task, or a higher rate of inattention to the task.

Theta Roles			
DocSet	#annotators minus 1	# lexical nodes	Alpha/MASI (pair delta)
Arabic1	6	80	.43 (-.14)
Arabic2	9	97	.29
Korean1	9	112	.34 (-.06)
Korean2	9	92	.28
Japanese1	9	111	.39 (-.05)
Japanese2	9	117	.34
Spanish1	8	116	.23 (.05)
Spanish2	8	124	.28
French1	9	130	.25 (-.09)
French2	9	136	.16
Hindi1	8	77	.38 (.02)
Hindi2	8	76	.40
Mean			.31

Table 4. Alpha values using the MASI distance metric for theta role annotations on each document set. For each dataset, the worst annotator was eliminated.

Other possible sources of difference between the document set reliability scores would likely depend on differences in the semantic complexity of the concepts expressed, or to differences in the translation quality. In Figure 1, for example, we can see that the first translation is a less fluent sentence of English: the repetition of the NP “the growth rate” is somewhat awkward, and the word “less” would have been more correct instead of “lesser.” This could potentially affect the annotators certainty about the meaning. However, it is difficult to imagine how to control for either of these conditions, apart from conducting a very large scale study.

We computed separate reliability scores for the four parts of speech that were annotated: noun, verb, adj and adverb. In general, the reliability scores by part of speech were distributed very similarly to the full set, with nouns having somewhat higher reliability on average (mean=.60). Verbs, however, had much lower scores. For example, the mean reliability for verbs across the 12 document sets was .46. A t-test shows this is a significant difference from the group mean (p=.5).

Tables 2 and 3 present reliability scores for Mikrokosmos concepts and theta roles. As shown, they are much lower than for the WordNet concepts. This seems to be due to a much higher rate where no selection was made. Overall, annotators made no selection for a node at twice the rate for theta roles as for Mikrokosmos concepts, and even chose no selection even more frequently in the case of the theta role annotations.

5.2. Rating annotator performance

Table 5 presents the key results from an analysis in which we computed inter-annotator reliability for all combinations of annotators from 2 to N, where N is the total number of annotators. In this way, we were able to identify groups of individual annotators with relatively higher inter-annotator agreement, as well as determine which selection of annotators would yield the most consistent annotations.

Column M of Table 5 indicates the maximum number of coders to achieve an agreement of .70 or higher. Column WHO indicates the identity of the best subset of coders that achieves this threshold, while AVG$_{\mathrm{M}}$ gives the average reliability over all combinations of coders of the same cardinality M. Clearly, reliability depends on which annotators are used. In addition, with the exception of the French set, Table 4 illustrates that later in the project, very good reliability can be achieved by dropping relatively fewer coders. Note that annotator 2, used as an example in Figure 1, appears in ten rows of Table 4.

DocSet	M	WHO	Alpha MASI	AVG$_{\mathrm{M}}$
Arabic1	4	5,9,10,11	.73	.51
Arabic2	2	2,6	.75	.40
Korean1	3	2,6,8	.71	.39
Korean2	3	5,9,10	.71	.34
Japanese1	7	2,5,6,8,9,10,11	.71	.56
Japanese2	7	1,2,5,6,9,10,11	.70	.52
Spanish1	4	2,6,9,10	.70	.57
Spanish2	6	2,5,6,9,10,11	.70	.56
French1	4	2,6,9,11	.72	.47
French2	3	2,6,11	.71	.40
Hindi1	5	2,6,9,10,11	.70	.55
Hindi2	6	2,5,6,9,10,11	.70	.56

Table 5. Maximum number (M) of coders to achieve agreement of .70 or above, versus average across all combinations of M coders (AVG$_{\mathrm{M}}$).

A similar analysis of the Mikrokosmos and theta role reliability results indicates that relatively more coders would need to be dropped to reach the same threshold of .70. In addition, the values for AVG$_{\mathrm{M}}$ are much lower.

6. Discussion and Conclusion

In the NLP literature, inter-annotator agreement measures are usually presented in order to make the claim that an annotation is reliable. We have repeated arguments from previous literature that it is difficult to infer much about the absolute values of reliability measures without a context, such as an independent assessment of the significance of variables derived from the annotation. We have illustrated another use of reliability measures, namely to examine variations in reliability along different dimensions. As a consequence, we have been able to demonstrate overall improvement over time, to identify subsets of annotators that are more reliable, and to show that verbs are more difficult to assign conceptual labels to than the other parts of speech.

Presumably, if the goal is to measure variations in reliability along different dimensions, then the absolute values of the measurements are less important, as is the choice of metric. However, we have also argued that when annotators make complex decisions for each coding unit, it is important to choose, or if necessary, to design, an appropriate metric. We have illustrated the application of a method for measuring reliability that was originally designed for co-reference annotation. In future work, it would be useful to compare this metric with other reliability measures, such as Arstein & Poesio's Beta^3 (2005), or Rosenberg & Binkowski's Augmented kappa (2004), on the same datasets, or with alternative approaches to reliability measurement.

Acknowledgements

This work has been supported by NSF ITR Grant IIS-0325887. We would like to thank all the members of the IAMTC project for providing data and for helpful discussions.

References

Artstein, R. and M. Poesio. (2005). Kappa^3=Alpha (or Beta). Technical Report NLE Technote 2005-01, University of Essex. Essex.

Bruce, R. and J. Wiebe. (1998). Word-sense distinguishability an inter-coder agreement. In *Proceedings of Empirical Methods in Natural Language Processing*.

Di Eugenio, B. and M. Glass. (2004). The kappa statistic: A second look. *Computational Linguistics*, 30(1):95-101.

Di Eugenio, B. (2000). On the usage of Kappa to evaluate agreement on coding tasks, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece.

Dorr, B.; Green, R.; Levin, L.; Rambow, O.; Farwell, D.; Habash, N.; Helmreich, S.; Hovy, E.; Miller, K.J.; Mitamura, T.; Reeder, F.; Siddharthan, A. (2004).. Semantic annotation and lexico-syntactic paraphrase. In *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora*, (LREC) Portugal,.

Farwell, D.; Helmreich, S.; Dorr, B. J.; Habash, N.; Reeder, F.; Miller, K.; Levin, L.; Mitamura, T.; Hovy, E.; Rambow, O.; Siddharthan, A. (2004). Interlingual annotation of multilingual text corpora. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Workshop on Frontiers in Corpus Annotation*, Boston, MA, pp. 55-62, 2004.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles* 44:223-270.

Krippendorff, Klaus. (1980.) *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA.

Mitamura, T.; Miller, K.; Dorr, B. J.; Farwell, D.; Habash, N.; Levin, L.; Helmreich, S.; Hovy, E.; Levin, L.; Rambow, O.; Reeder, F.; Siddharthan, A. (2004). Semantic annotation of multilingual text corpora. In *Proceedings of the Workshop on Beyond Named Entity Recognition: Semantic Labeling for NLP Tasks*, LREC, Portugal.

Nirenburg, S., S. Beale, K. Mahesh, B. Onyshkevych, V. Raskin, E. Viegas, Y. Wilks and R. Zajac. (1996). Lexicons in the Mikrokosmos project. *Proceedings of the Society for Artificial Intelligence and Simulated Behavior Workshop on Multilinguality in the Lexicon*, Brighton, UK.

Passonneau, R.; Nenkova, A.; McKeown, K.; Sigelman, S. (2005). Applying the pyramid method in DUC 2005. In *Proceedings of the Workshop of the Document Understanding Conference*. Vancouver, B.C.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation* (LREC). Genoa, Italy.

Passonneau, R. (2004). Computing reliability for co-reference annotation. *Proceedings of the International Conference on Language Resources and Evaluation* (LREC). Portugal.

Philpot, A.; Fleischman, M.; Hovy, E.H. (2003). Semi-automatic construction of a general purpose ontology. *Proceedings of the International Lisp Conference*. New York, NY. Invited.

Rosenberg, A. and Binkowski, E. (2004). Augmenting the kappa statistic to determine inter-annotator reliability for multiply labeled data points. In *Proceedings of the Human Language Technology Conference and Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL).*

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* 19:321-325.

Siegel, S. and N. John Castellan, Jr. (1988) *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition. McGraw-Hill, New York.

Teufel, S. and H. van Halteren. (2004). Evaluating information content by factoid analysis: human annotation and stability. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 419-426.

Uebersax, J. (1988). Validity inferences from interobserver agreement. Psychological Bulletin, 104(3):405-416.

H2E1-6	Last year, due to a famine, the growth rate had been 1 per cent lesser than the estimated growth rate.
H2E2-6	Last year due to drought conditions, India's economy grew at a rate 1% less then estimated.

H2E2-6 Annotation, by Annotator 2

Node Id	**Lexical Item**	**Parent Node Id**	**Gram. Role**	**Theta Role**	**WordNet Concept**	**Mikrokosmos Concept**
20	year	90	Mod	TIME	|yr|	|YEAR$NOUN|
40	drought	50	Mod	none	|drought|	|PRECIPITATION$NOUN|
50	condition	30	Obj	---	|condition<way|	|INFORMATION$NOUN|
80	economy	90	Subj	THEME	|economy<system|	|GOVERNMENTAL-IDEOLOGY$NOUN|
90	grow	0	Root	---	|grow<boom|	|GROW$VERB|
120	rate	100	Mod	---	|rate| |pace>beat|	|DummyConcept|
140	less	120	Mod	none	|less_than|	|DummyConcept|
150	than	140	Mod	none	|less_than|	|DummyConcept|
155	<pro>	160	Subj	AGENT	|DummyConcept|	|INCREASE$NOUN|
160	estimate	150	Mod	none	|estimate>set|	|ESTIMATE$VERB|
165	<pro>	160	Obj	THEME	|DummyConcept|	|DummyConcept|

Figure 1. Two translations of the same sentence, and a sample annotation of the lexical items of the second sentence