The Annotation of the Switchboard Corpus with the New ISO Standard for Dialogue Act Analysis

Alex C. Fang
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Hong Kong SAR
acfang@cityu.edu.hk

Jing Cao
College of Foreign Languages
Zhongnan University of Economics and Law
Wuhan, China
c_jinhk@yahoo.cn

Harry Bunt
Tilburg Center for Cognition and Communication
Tilburg University
The Netherland
harry.bunt@uvt.nl

Xiaoyue Liu
The Dialogue Systems Group
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Hong Kong SAR
xyliu0@cityu.edu.hk

Abstract

This paper is the description of a semantic annotation project that aims at the re-annotation of the Switchboard Corpus, previously annotated with the SWBD-DAMSL scheme, according to a new international standard for dialogue act analysis. A major objective is to evaluate, empirically, the applicability of the new ISO standard through the construction of an interoperable language resource that will eventually help evaluate the pros and cons of different annotation schemes. In this paper, we shall provide an account of the various aspects of the annotation project, especially in terms of the conversion between the two analytical systems, including those that can be fully automated and those that have to be manually inspected and validated. A second objective is to provide some basic descriptive statistics about the newly annotated corpus with a view to characterize the new annotation scheme in comparison with the SWBD-DAMSL scheme.

1 Introduction

The Switchboard Corpus is a valuable language resource for the study of telephone conversations. The Switchboard Dialogue Act Corpus, which is distributed by the Linguistic Data Consortium (LDC) and available online at http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2001T61, provides extensive added value because of its annotation of the component utterances according to an adapted DAMSL scheme for dialogue act (DA) analysis. More recently, the NXT-format Switchboard Corpus has been created (Calhoun et al. 2010). It combines orthographic transcriptions with annotations for dialogue act, syntax, focus/contrast, animacy, information status, and coreference in addition to prosodic and phonemic markings.

This paper describes a new development in the annotation of the Switchboard Dialogue Act Corpus. In this new version, each component utterance has been additionally annotated according to a new international standard, namely, ISO 64217-2:2012 (Bunt et al. 2010, 2012; ISO 2012). A major objective for the re-annotation of the corpus is to produce a new language resource where the same linguistic material is annotated according to two different schemes in order to facilitate a comparative study of different analytical frameworks. A second major objective is to verify the applicability of the new international standard through the practical annotation of authentic data and also to verify if the new scheme represents theoretical and practical advancement in real terms.

The basic principles for the project include the following:

1) The new DA scheme should be empirically applicable to a corpus of authentic conversations.
2) The re-annotation of the corpus should be realized by converting as much as possible

from its previous annotation in order to retain maximal data reliability.

3) The conversion should be optimized for automatic conversion, and manual mapping should be applied only when necessary.

A direct outcome for the project is a new language resource, which comprises transcribed real-life conversations and two different sets of DA annotations. As Figure 1 indicates, such a resource is especially well suited for comparative studies of DA annotation schemes and also in-depth investigation of the corpus through parallel annotations according to different schemes. As far as we know, such a resource is the first of its kind in the area of dialogue act analysis.

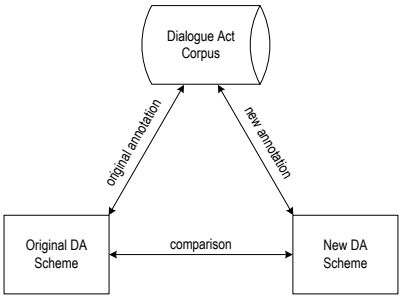

Figure 1: A new resource for DA research

The rest of this paper will describe how the original SWBD-DAMSL scheme has been converted to the new ISO DA standard. It will also describe the newly constructed corpus, including the ISO DA tags and their dimensions. The paper will then discuss some of the issues in the conversion and outline some future work.

2 SWBD-DAMSL

SWBD-DAMSL is a version of DAMSL (Dialogue Act Markup in Several Layers; Allen and Core 1997) that was specially adapted for the annotation of the Switchboard Corpus. The SWBD-DAMSL scheme consists of 220 DA types and has facilitated past studies such as Jurafsky et al. (1997) and Stolcke et al. (2000).[1]

To follow the practice of standoff markup, the original 1,155 annotated telephone conversations were re-processed, each slash-unit was coded and the utterance and its corresponding DA tag were separated and stored in individual files. Consider Example (1) below extracted from the file named `sw_0052_4378.utt`.

(1) sd B.7 utt1: *{C And,} {F uh,} <inhaling> we've done <sigh> lots to it. /*

Such an utterance, which is annotated as `sd` (*statement-non-opinion*), resulted in two files, where `SBD` stands for SWBD-DAMSL:

File 1: `sw00-0052-0010-B007-01.utt`
Content: *{C And,} {F uh,} <inhaling> we've done <sigh> lots to it.*
File 2: `sw00-0052-0010-B007-01-SBD.da`
Content: `sd`

As a general rule, the transcribed utterance is stored in a file with the `.utt` suffix and its SWBD-DAMSL tag in `*-SBD.da`. Similarly, `*-ISO.da` represents the set of files containing the ISO DA tags and `*-ISO.di` their corresponding dimensions.

3 Conversion to the ISO DA Standard

The ISO scheme contains 56 core DA tags, representing a tagset size comparable to that of the SWBD-DAMSL scheme of 59 combined tags. The tags are grouped according to 9 core dimensions and additionally described by a number of qualifiers designed to provide additional information about subtleties of communication functions. To maximally facilitate the conversion from SWBD-DAMSL to SWBD-ISO, four types of relation between the SWBD-DAMSL scheme and the ISO scheme were identified, namely, exact matches, many-to-one matches, one-to-many matches and unique SWBD-DAMSL tags. In the project, we performed the first two types of conversions automatically, and the one-to-many conversion was mapped manually. The treatment of the last group of tags, i.e., those unique to SWBD-DAMSL, will be discussed in section 3.4.

3.1 Automatic Mapping

The automatic mapping was performed on exact matches and many-to-one matches between the two schemes. In this process, 46 SWBD-DAMSL tags were matched to 22 ISO DA types, with a

[1] It should be pointed out that for the sake of enough instances, some original SWBD-DAMSL DA types have been combined together, which resulted in 42 different DA types in Jurafsky et al. (1997). The current study uses the 59 DA tags in Fang et al. (2011).

total number of 187,768 utterances, or 83.97% of the corpus, which accounts for 94.29% of the whole corpus in terms of tokens.

3.2 Manual Mapping

Six SWBD-DAMSL DA types were observed to have multiple destinations in the ISO scheme. These include `accept`, `accept-part`, `reject`, `reject-part`, `action directive`, and `other answer`. A user-friendly GUI was specially constructed and all the utterances concerned manually inspected and assigned an ISO tag.

For this task, three postgraduate students majoring in linguistics were invited to perform the annotation. They were provided with the manual of SWBD-DAMSL and the ISO standard. The training session included three phases: First, the annotators got familiar with the two DA schemes through trail annotation of 2 files for each of the six DAs. During the second phase, supervised annotation was carried out with 10 additional files for each DA. Finally, unsupervised annotation was conducted with another set of 10 files for each DA, and the inter-annotator agreement test was calculated based on the unsupervised samples.

Results show that in most cases a predominant ISO DA type could be identified. In some cases, an annotator favoured just one particular ISO DA type, which creates the bias and prevalence problems for the calculation of the kappa value (e.g. Di Eugenio and Glass, 2004). To solve this problem, the prevalence-adjusted and bias-adjusted kappa (PABAK) was proposed by Byrt et al. (1993) and used in quite a few past studies such as Sim and Wright (2005), Chen et al. (2009), Cunningham (2009) and Hallgren (2012). The adjusted kappa is defined as:

$$PABAK = \frac{kP_{obs} - 1}{k - 1}$$

where k is the number of categories and P_{obs} the proportion of observed agreement. At the end of the training session, PABAK was calculated pair-wised and the mean was taken as the final result. The average PABAK value is 0.69. According to Landis and Koch (1977), the agreement between the three annotators is substantial and therefore judged acceptable for subsequent manual annotation.

The actual manual annotation saw six SWBD-DAMSL tags mapped to 26 different ISO DA tags, which involves 12,837 utterances (i.e. 5.74% of the corpus) and covers 2.03% of the corpus in terms of tokens.

Altogether, through both automatic and manual annotation, 200,605 utterances in the corpus (i.e. 89.71%) were treated with ISO DA tags. Table 1 presents the basic statistics of the SWBD corpus annotated with the ISO DA scheme, including the types of ISO DAs, the number of utterances and tokens, and their corresponding percentage and accumulative percentage. The ISO DA types are arranged according to the number of utterances in descending order.

ISO DA Type	Utterance			Token		
	#	%	Cum%	#	%	Cum%
inform	120227	53.767	53.77	1266791	82.962	82.96
autoPositive	46382	20.743	74.51	66506	4.355	87.32
agreement	10934	4.890	79.40	20598	1.349	88.67
propositionalQuestion	5896	2.637	82.04	39604	2.594	91.26
confirm	3115	1.393	83.43	3698	0.242	91.50
initialGoodbye	2661	1.190	84.62	9442	0.618	92.12
setQuestion	2174	0.972	85.59	15841	1.037	93.16
disconfirm	1597	0.714	86.31	3392	0.222	93.38
answer	1522	0.681	86.99	8154	0.534	93.91
checkQuestion	1471	0.658	87.64	11053	0.724	94.64
completion	813	0.364	88.01	3188	0.209	94.85
question	680	0.304	88.31	5068	0.332	95.18
stalling	580	0.259	88.57	3004	0.197	95.37
choiceQuestion	506	0.226	88.80	4502	0.295	95.67
suggest	369	0.165	88.96	3320	0.217	95.89
autoNegative	307	0.137	89.10	798	0.052	95.94
request	278	0.124	89.22	1644	0.108	96.05
disagreement	258	0.115	89.34	689	0.045	96.09
acceptApology	112	0.050	89.39	366	0.024	96.12
instruct	106	0.047	89.44	961	0.063	96.18
acceptSuggest	99	0.044	89.48	195	0.013	96.19
apology	79	0.035	89.52	317	0.021	96.21
thanking	79	0.035	89.55	221	0.014	96.23
offer	71	0.032	89.58	590	0.039	96.27
acceptRequest	65	0.029	89.61	96	0.006	96.27
signalSpeakingError	56	0.025	89.64	75	0.005	96.28
promise	41	0.018	89.66	279	0.018	96.30
correction	29	0.013	89.67	210	0.014	96.31
acceptOffer	26	0.012	89.68	40	0.003	96.31
turnTake	18	0.008	89.69	28	0.002	96.31
alloPositive	17	0.008	89.70	21	0.001	96.31
correctMisspeaking	14	0.006	89.70	38	0.002	96.32
selfCorrection	8	0.004	89.71	41	0.003	96.32
acceptThanking	6	0.003	89.71	6	0.000	96.32
declineOffer	3	0.001	89.71	5	0.000	96.32
declineRequest	3	0.001	89.71	3	0.000	96.32
turnRelease	2	0.001	89.71	2	0.000	96.32
declineSuggest	1	0.000	89.71	1	0.000	96.32
other	23001	10.29	100.00	56175	3.679	100.00
Total	223606	100.00		1526962	100.00	

Table 1: Basic stats of the SWBD-ISO corpus

Other in Table 1 glosses together all the SWBD-DAMSL tags that cannot be matched to the ISO DA scheme. These represent 10.29% of the total

number of utterances in the corpus or 3.679% of all the tokens. They will be discussed in detail later in Section 3.4.

3.3 Dimensions

A feature of the ISO DA standard is that each utterance is also marked with dimension information. Consider Example (1) again. According to the ISO annotation scheme, it is annotated with the DA type *inform*, which belongs to the ISO dimension of *Task*. As a matter of fact, out of the nine ISO dimensions, eight are identified in the newly created SWBD-ISO corpus except for the dimension of *Discourse Structuring*.[2] Table 2 lists the eight dimensions and their corresponding ISO DA types, together with the percentage of the utterances they cover. Note that only those DA types observed in the corpus are listed in the table. The DA tag `alloNegative`, for instance, is missing from Table 2 since the corpus does not contain any utterance analysed as such. *Other** in Table 2 actually refers to the portion of utterances in the corpus that do not have an appropriate ISO DA tag and hence no dimension information. According to the table, those account for 10.29% of the total number of utterances in the corpus. The original SWBD-DAMSL analysis of the utterances is described in detail in Section 3.4 below and summarised in Table 3.

ISO Dimension	%	ISO DA Type
Task	66.85	inform; agreement; propositionalQuestion; confirm; setQuestion; disconfirm; answer; checkQuestion; question; choiceQuestion; suggest; request; disagreement; instruct; acceptSuggest; offer; acceptRequest; promise; correction; acceptOffer; declineOffer; declineRequest; declineSuggest
Auto-Feedback	20.88	autoPositive; autoNegative
Social Obligations Management	1.31	initialGoodbye; acceptApology; apology; thanking; acceptThanking
Time Management	1.19	stalling
Partner Communication Management	0.37	completion; correctMisspeaking
Own Communication Management	0.03	signalSpeakingError; selfCorrection
Allo-Feedback	0.01	alloPositive
Turn Management	0.01	turnTake; turnRelease
*Other**	*10.29*	**See Table 3 for a detailed breakdown*
Total	100.00	

Table 2: Basic stats for ISO dimensions

[2] In the current project, the dimension of *Discourse Structuring* is not explicitly treated since it most often overlaps with the more general *Task* dimension.

In addition, a particular feature of the ISO standard for DA annotation is that an utterance can be associated with more than one dimension, known as multi-dimensionality of DA. Example (1) has two dimensions, namely, *Task* and *Time Management*, for which the following files would be created:

File 3: `sw00-0052-0010-B007-01-ISO-21.da`
Content: `inform`
File 4: `sw00-0052-0010-B007-01-ISO-21.di`
Content: `task`
File 5: `sw00-0052-0010-B007-01-ISO-22.da`
Content: `stalling`
File 6: `sw00-0052-0010-B007-01-ISO-22.di`
Content: `timeManagement`

In our annotation scheme, `.da` files contain the name of the ISO DA types, while `.di` the name of the ISO dimensions. The first digit following `ISO-` (i.e. 2 in file names above) indicates the number of dimensions that a certain utterance is contextually associated with, while the second digit indicates the current number in the series.

Of the 200,605 mapped utterances, 144,909 utterances are annotated with one dimension, 44,749 with 2 dimensions and 10,947 with 3 dimensions.

3.4 Unmatched SWBD-DAMSL Tags

The conversion process left 13 SWBD-DAMSL tags unmatched to the ISO scheme. They account for 23,001 utterances and 56,175 tokens, representing respectively 10.29% and 3.68% of the corpus. See Table 3 for the basic statistics.

SWBD-DAMSL Tag	Utterance			Token		
	#	%	Cum%	#	%	Cum%
abandoned	12986	5.81	5.81	35363	2.32	2.32
non-verbal	3730	1.67	7.48	77	0.01	2.33
uninterpretable	3131	1.40	8.88	5729	0.38	2.71
quoted material	1058	0.47	9.35	8114	0.53	3.24
other	820	0.37	9.72	1603	0.10	3.34
transcription errors	649	0.29	10.01	3028	0.20	3.54
conventional opening	225	0.10	10.11	529	0.03	3.57
exclamation	136	0.06	10.17	282	0.02	3.59
3rd party talk	118	0.05	10.22	508	0.03	3.62
self talk	106	0.05	10.27	630	0.04	3.66
double quoted	27	0.01	10.28	189	0.01	3.67
explicit performative	9	0.00	10.28	81	0.01	3.68
other forward function	6	0.00	10.29	42	0.00	3.68
Total	23001	10.29		56175	3.68	

Table 3: Basic stats of unique SWBD tags

It is noticeable that a majority of these tags (e.g. *abandoned*, *non-verbal*, *uninterpretable*, and *quoted material*) are not defined on the basis of the communicative function of the utterance. Only two tags, i.e., *exclamation* and *explicit performative*, are clearly defined in functional terms and yet could not be matched to any of the DA types in the ISO standard.

3.5 Unmatched ISO Tags

An examination of the converted corpus has revealed that some ISO DA tags cannot be empirically observed in the corpus. See Table 4 for the specific ISO DA tags along with their corresponding dimensions.

ISO DA Type	ISO Dimension
addressRequest; addressSuggest; addressOffer	Task
alloNegative	Allo-Feedback
turnAccept; turnAssign; turnGrab; turnKeep	Turn Management
pausing	Time Management
interactionStructuring; opening	Discourse Structuring
initialGreeting; returnGreeting; initialSelfIntroduction; returnSelfIntroduction; returnGoodbye	Social Obligations Management
retraction	Own Communication Management

Table 4: DA Tags unique to ISO scheme

As is worth noting here, Table 4 should not be taken to suggest that the corpus does not contain any utterance that performs those communicative functions specified in the new ISO standard. Bear in mind that the ISO annotation of the corpus is achieved through mapping the original SWBD-DAMSL tags. Hence, the non-observation of the ISO tags listed in Table 4 only suggests that there is no direct mapping between the SWBD-DAMSL and ISO tagsets as far as these particular ones are concerned. Annotation of these unique ISO tags can be realized by considering the actual content of the utterances. Secondly, it should also be noted that the unmatched tags in the *Task* dimension include the mother nodes (e.g. *addressRequest*) of some more specific DAs (e.g. *acceptRequest* and *declineRequest*) and the utterances concerned have been annotated with the more specific daughter nodes as requested by the manual of annotation.

4 Conclusion

This paper described a project to re-annotate the SWBD DA corpus with the new ISO standard for DA analysis and reported some of the basic statistics concerning the conversion between SWBD-DAMSL and SWBD-ISO. A significant contribution of the current work is the creation of an interoperable language resource which can serve as the test-bed for the evaluation of different DA annotation schemes. The same resource can also be used for the exploration and verification of the contribution of different DA taxonomies to the automatic identification and classification of DAs. Our immediate future work will include a comparative study of the SWBD-DAMSL and ISO DA schemes. It is also expected that attempts will be made to address the treatment of the unmatched DA tags with a view how best to accommodate empirically encountered dialogue phenomena that were not considered in the drafting process of the standard. At the same time, we are performing some preliminary research to assess the performance of an automatic classifier of ISO dialogue acts with the specific intent to construct a DA model from the SWBD-ISO Corpus to be applied to other linguistic resources for dialogue studies. An issue that is of particular interest at this stage is the prospect of applying the ISO DA standard to dialogue resources in the Chinese language.

Acknowledgement

The project described in this article was supported in part by grants received from the General Research Fund of the Research Grants Council of the Hong Kong Special Administrative Region, China (RGC Project No. 142711) and City University of Hong Kong (Project Nos 7008002, 7008062, 9041694, 9610188, and 9610226). The authors would like to acknowledge the academic input and the technical support received from the members of the Dialogue Systems Group (http://dsg.ctl.cityu.edu.hk) based at the Department of Chinese, Translation and Linguistics, City University of Hong Kong.

References

Alex C. Fang, Harry Bunt, Jing Cao and Xiaoyue Liu. 2011. Relating the Semantics of Dialogue Acts to Linguistic Properties: A machine learning perspective through lexical cues. In Proceedings of the 5th IEEE International Conference on Semantic Computing, September 18-21, 2011, Stanford University, Palo Alto, CA, USA.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul

Taylor, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3): 339–371.

Barbara Di Eugenio and Michael Glass. 2004. The Kappa Statistic: A Second Look. Journal of Computational Linguistics, 30(1): 95-101.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-discourse-function Annotation Coders Manual, Draft 13. University of Colorado, Boulder Institute of Cognitive Science Technical Report 97-02.

Guanmin Chen, Peter Faris, Brenda Hemmelgarn, Robin L. Walker, and Hude Quan. 2009. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. BMC Medical Research Methodology. 9: 5.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kyong Lee, Volha Petukhova, Andrel Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO Standard for Dialogue Act Annotation. In Proceedings of the Seventh International Conference on Language Resources and Evaluation. Valletta, MALTA, 17-23 May 2010.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrel Popescu-Belis, and David Traum. 2012. A Semantically-based Standard for Dialogue Annotation. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). Istanbul.

ISO Standard 24617-2. 2012. Language resource management – Semantic annotation framework (SemAF), Part 2: Dialogue acts. ISO, Geneva, 2012.

James Allen and Mark Core. 1997. DAMSL: Dialogue Act Markup in Several Layers (Draft 2.1*)*. Technical Report, Multiparty Discourse Group. Discourse Resource Initiative, September/ October 1997.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1): 159-174.

Julius Sim and Chris C. Wright. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy, 85 (3): 257-268.

Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in Quantitative Methods for Psychology. 8(1): 23-34.

Michael Cunningham. 2009. More than Just the Kappa Coefficient: A Program to Fully Characterize Inter-Rater Reliability between Two Raters. SAS Global Forum 2009.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language Resources and Evaluation. 44(4): 387-419.

Ted Byrt, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and Kappa. Journal of Clinical Epidemiology. 46(5): 423-429.