Annotating for Precision and Recall in Speech Act Variation: The Case of Directives in the *Spoken Turkish Corpus*

Abstract

Summary

Speech act realizations pose special difficulties in search during annotation and pragmatics research based on corpora, in spite of the fact that their various forms may be relatively formulaic – hence amenable to (semi-)automatic annotation. Focusing on spoken corpora, this paper concerns the generation of discourse analytical annotation schemes that can address not only variation in speech act annotation but also variation in dialog and interaction structure coding. The major arguments in the paper are that (1) enriching the metadata features of corpus design can act as useful aids in speech act annotation; and that (2) sociopragmatic annotation and corpus-oriented pragmatics research can be enhanced by incorporating (semi-)automated linguistic annotations that rely both on bottom-up discovery procedures and the more relatively top-down, theoretical linguistic categorizations based on the literature in traditional approaches to pragmatics research. The paper illustrates implementations of enriched metadata and pragmatic annotation with examples drawn from directives in the demo version of the *Spoken Turkish Corpus*, and presents a qualitative assessment of the annotation procedures.

1. Speech acts as a challenge for corpus annotation

Speech act realizations are notorious for the special difficulties they pose in search both during annotation and pragmatics research based on corpora, in spite of the fact that their various forms may be relatively formulaic, hence amenable to (semi-)automatic annotation. While part-of-speech tagging has become semi-automatized, sociopragmatic annotation involves significant difficulties in the very process of identifying categories and units of pragmatic phenomena such as variation in manifestations of speech acts and the identification of conversational segments (Archer, Culpeper and Davies 2008: 635). As underscored by Schmidt and Wörner, this makes pragmatics research conducted on corpora "heuristic" in nature in that the relationship between theory and corpus analysis is bi-directional (2009: 4). This is all the more so in the identification of speech acts, as function only partially follows form.

To illustrate this with a short excerpt from a naturally occurring speech event, the utterance *iki çay* 'two teas' may be describing the number of cups of tea one has had. But followed by *tamam hocam* "okey deferential address term", the noun phrase would achieve the illocutionary force of a request when uttered to a service provider. It goes without saying that the initial utterance can occur with *please* as a politeness marker, which would certainly increase its chance of being identified as a request in a corpus. Communications, however, do not always exhibit such pre-fabricated forms. Moreover, there are a multitude of supportive acts in the performance of speech acts (see, e.g., Blum-Kulka, House and Kasper 1989). Thus their recall in corpora would require the analyst to increase the number of search expressions infinitely. Even so, that would not guarantee full recall; neither would it filter false cases. Jucker, Schneider, Taavitsainen and Breustedt (2008) remark, for instance, that, while the formulae identified in Manes and Wolfson's (1981) classic study on compliments account for a fair number tokens of the speech act in the BNC, they warn that there are likely to be hidden manifestations of non-formulaic compliments. This situation goes against the advantage of using corpora for the study of variation and largely limits the derivation of qualitative and quantitative conclusions from corpora.

In this paper we argue that annotation for studying variation in speech act realizations can be improved by (1) enriching metadata coding during the construction stage of a corpus; and (2) by implementing (semi-)automated annotation for sociopragmatic features of communications that rely both on bottom-up discovery procedures and top-down, linguistic categorizations based on traditional approaches to pragmatics research (e.g. annotation of socially and discursively significant verbal and non-verbal phenomena and non-phonological units such as multi-word expressions and changes in tone of voice.

The argumentation is based on insights from Multidimensional Analysis (Biber 1995) and vocabulary-based identification of discourse units (Csomay, Jones and Keck 20007), and the fact that pragmatic phenomena in conversational management (e.g., illocutionary force indicating devices, address terms, and politeness formulae) tend to form constellations of 'traces' in discourse. Annotating such traces can add "precision" and improve "recall" (Jucker et al. 2008) in searching for variation in speech acts. The main thrust of the paper is that speech events and discourse level units exhibit such verbal and non-verbal clusters, and that annotating such units can provide insights for further discursive coding (see, Carletta et al. 1997). Below, we explain the procedures for these two approaches to annotation with illustrations from the demo version of the *Spoken Turkish Corpus* (STC), which currently comprises 44,962 words from a selection of recordings in conversational settings, service encounters, and radio archives (STC employs EXMARaLDA corpus construction tools (Schmidt 2004), along with a web-based corpus management system.).

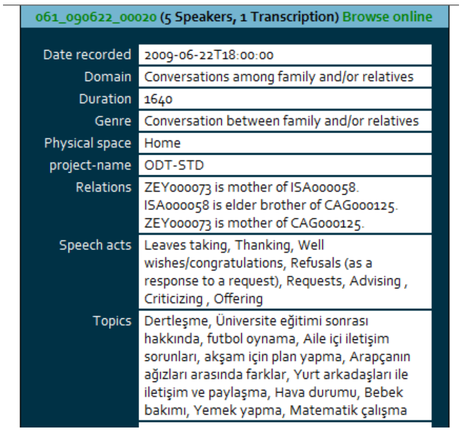
2. Metadata construction in the transcription and annotation workflow of STC

Besides constructing a metadata system for domain, interactional goal and speaker features, we maintain that the inclusion of speech acts along the lines of Searle (1973) and conversational topics as part of the metadata features of a corpus is a significant tool for tracing variation in speech acts in a systematic manner, as topical variation can impact their performance beyond the influence of domain and setting features. Viewed from another perspective, spoken texts are slippery resources of language in terms of domain and setting categorization such that they are often characterized by shifts in interactional goals. A service encounter in a shop, for example, can easily turn into a chat. Thus, if a communicative event were classified only for its domain of interaction, one would risk the chance of tracing subtle differences within the same domain along several dimensions. The simultaneous annotation of topics and speech acts during the compilation of the recordings and during their transcription can address the concern for achieving maximal retrieval of tokens of a speech act. It enables a bottom-up approach to search for variation through control for topic and speech acts, as manifestations of the act may not exhibit structures noted in the literature. It also allows for a corpus-driven categorization of speech acts that may not have been investigated at all in the particular language. The stages in this procedure in the construction of STC are outlined below:

1. Noting of local and global topics, and the communication related activities by recorders in the communication file (e.g. studying for an exam and having dinner)
2. Checking of topics and additions during the transfer of the recording to the corpus management system
3. Stages in transcription:
    a. Initial step: basic transcription of recording for verbal and non-verbal; editing of topics and addition of speech act metadata
    b. First check: Checking the transcription for verbal and non-verbal events; editing of topics and speech act metadata
    c. Second check: Checking the transcription for verbal and non-verbal events; editing of topics and speech act metadata

Stages (2) and (3a) involve expert and non-experts in linguistics, while steps (3b-c) are carried out by experts. To achieve a higher level of reliability in transcription, a different transcriber is responsible for the annotation in each step in (3), and differences in transcription are handled through consultation amongst the transcribers and the recorder. Stages (1) and (3a) ideally involve the same person so that the transcriber has an intuitive grasp of the topical content and the affective tone of the communication. This procedure has the added advantage of detecting regional variation with more precision. It also renders possible the construction of sub-corpora for initial pilot annotation not only through control for domain but also for topic and speech act, thus enhancing the likelihood of retrieval of a greater variety of tokens in a more economical manner. Naturally, this workflow taps into native speaker intuitions on speech act performance, but it is a viable methodological procedure in linguistics because it harnesses intuitions in a context-sensitive environment during text processing.

Figure 1 displays a select number of the metadata features of one communication in STC (Note that topics are written in Turkish, and that the term *requests* is used instead of *directives* because the former was a more transparent term for the transcribers in step (3a) above).



061_090622_00020 (5 Speakers, 1 Transcription) Browse online

Date recorded	2009-06-22T18:00:00
Domain	Conversations among family and/or relatives
Duration	1640
Genre	Conversation between family and/or relatives
Physical space	Home
project-name	ODT-STD
Relations	ZEY000073 is mother of ISA000058. ISA000058 is elder brother of CAG000125. ZEY000073 is mother of CAG000125.
Speech acts	Leaves taking, Thanking, Well wishes/congratulations, Refusals (as a response to a request), Requests, Advising , Criticizing , Offering
Topics	Dertleşme, Üniversite eğitimi sonrası hakkında, futbol oynama, Aile içi iletişim sorunları, akşam için plan yapma, Arapçanın ağızları arasında farklar, Yurt arkadaşları ile iletişim ve paylaşma, Hava durumu, Bebek bakımı, Yemek yapma, Matematik çalışma

Figure 1. Partial metadata for a communication in STC

3. Annotation procedure for speech acts in STC

Speech act annotation in STC is being implemented with Sextant (Wörner, n.d.), which also allows searches to be conducted with EXAKT. The search for tokens of directives employs a snowballing technique in developing regular expressions, and is similar to what Kohnen (2008: 21) describes as "structural eclecticism". In this paper, we focus only on the description of verbal cues for retrieval.

The annotation procedure starts off with the identification of forms that have been identified as being representative of directives in contemporary Turkish. Regular expressions based on these forms have been developed, and the coding of forms and the development of tag sets are done according to the syntactic and/or lexical structure of the head act. But instead of tagging only the head act, the full act is further coded by placing opening and closing tags for the relevant head act (see, Excerpts 1 and 2). This will allow further detailed tagging of the act in later stages of annotation.

The regular expressions are enriched based on tokens detected first in the sub-corpora of service encounters both by examining the larger context of the tokens recalled in initial searches and by manually investigating specific communications that are marked for directives in the corpus metadata. However, this procedure does not allow elliptical directives and hints to be recalled automatically. Based on the idea that a directive is ideally part of an adjacency pair – a directive and a response that may be a compliance or a refusal – the search for 'hidden' manifestations of the act is conducted through the presence of address terms and a select number of minimal responses, including lexical and non-lexical backchannels (e.g. *tamam* 'okey/enough/full', ha?, hm), which turned out to collate frequently with directives. Searches were thus conducted separately for these responses, and tokens that did not collocate with directives or form the head act itself (as is the case with *tamam*) were eliminated from the annotation.

Excerpt (1) shows the co-occurrence of *tamam* with an elliptical request (tag code: RNp), which could not be recalled with a regular expression (The head act is marked in bold):

Excerpt (1)

Speakers	*Interaction in Turkish*	*Translation*
XAM000066	şimdi ((0.3)) **T.C. kimlik numarası** ((0.2)) **ve öncelikli olarak** ((0.1)) **ev adresinizi**	$_{\textbf{RNp-open}}$now your Turkish ID number and first you home address ((XXX))$_{\textbf{RNp-close}}$
DIL000065	**tamam**. ((0.3)) ((filling in a form,10.8))	okey.

It is noteworthy that the sequence manifests the presence of the discourse marker *şimdi* ‘now’, and marks the speech act boundary, and illustrates how both minimal responses (*tamam*) and discourse markers collocate with the head act.

Excerpt (2) is an illustration from a service encounter in a travel agency. The head act (marked in bold) has a verb with the future in the past. In isolation, therefore, the utterance could be a manifestation of a representative, which the corpus actually revealed as being situationally-bound to requests in institutional service encounters. However, the collocation of the utterance with *buyrun* ‘lit. command’ (idiomatic equivalent: How can I help you?/Welcome) disambiguates it as a request:

Excerpt (2)

Speakers	*Interaction in Turkish*	*Translation*
MEH000222	((0.3)) **buyrun.**	welcome.
MED000112	iyi günler!	good day!
MEH000222	neresi olacak?	where is it to be? (idiomatic equivalent: where to?)
MED000112	**Dikili’ye bilet alacaktım**.	$_{\textbf{RImpFuI}\text{-open}}$I was going to get a ticket for Dikili$_{\textbf{RImpFuI-close}}$

Such collocations allow us to form a list of (semi-)formulaic conversational management units in the communications, which should be tagged as pragmatic markers for directives. In the current version of STC, *tamam* ‘okey’ is the item that exhibits the highest frequency. A search on the occurrence of the item was therefore conducted to check its collocation with directives. The search yielded 298 tokens, 20 of which were related to directives. In 8 instances, the item is a supportive move for the directive head act. In 2 recalls it was the head act itself to close off a conversational topic, while the remaining tokens were responses to a verbal or non-verbal request or part of the response to questions asking for advice/opinion. Amongst these we find the supportive function of *tamam* as a compliance gainer to be especially significant since the literature on directives in Turkish does not identify this function. Again, within these recalls, *tamam* collocates with 6 requests of the kind illustrated in Excerpt (2). This suggests that even with the limited size of the present corpus, *tamam* can function to disambiguate representatives from requests and can be used to retrieve elliptical directives and hints. Although the full classification of the pragmatics of *tamam* needs to be refined, we can safely say that in its semantically bleached use, it appears in topic closures, it functions as a backchannel to check comprehension, and is used as an agreement marker or as a pre-sequence to disagreement. In this regard, we can say that *tamam* is a pragmatic marker in its non-literal use and needs to be tagged accordingly.

4. Conclusion

Even though this paper touches only upon the disambiguating capacity of lexical pragmatic markers, the distribution of *tamam* also supports the claim that discourse segmentation and conversational structure annotation can use the clues provided by such ‘ traces’. The rough, functional description of *tamam* naturally raises the question as to coding principles for such items, including politeness formulae such as *please* and *welcome*. While non-lexical backchannels may not be too problematic, the classification and coding of pragmatic markers is a fuzzy area (Norrick 2009). At this stage, we propose that a semantic-based, broad categorization be made to distinguish lexical and non-lexical markers, interjections and discourse markers and discourse particles.

Our experience in testing the effect of pragmatic markers on recall of speech acts suggests that it might be possible to envision generic level schemes for speech act annotation. These would proceed first with a bottom-up approach, in which (multi-word) pragmatic markers, backchannels and non-verbal cues such as a classification of activity types (e.g., handing over money) are tagged. It is likely that such a venture will reveal commonalities between speech acts beyond what may be gleaned from the current pragmatics literature on speech act manifestations.

References

Archer, Dawn, Culpeper, Jonathan, Davies, Matthew, 2008. Pragmatic annotation. In: Lüdeling, A., Kytö, Merja (Eds.), *Corpus Linguistics: An International Handbook*, Vol. I, Berlin/New York: Walter de Gruyter, 613-642.

Biber, Douglas, 1995. *Dimensions of Register Variation*. New York: Cambridge University Press.

Blum-Kulka, Shoshana, House, Juliane, Kasper, Gabriele (Eds.), 1989. *Cross-Cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex.

Carletta, Jean, Isard, Amy, Isard, Stephen, Kowtko, Jacqueline C., Doherty-Sneddon, Gwyneth, Anderson, Anne H., 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics* 23, 13–32.

Csomay, Eniko, Jones, James K., Keck, Casey, 2007. Introduction to the identification and analysis of vocabulary-based discourse units. In: Biber, D., Connor, U., Upton, T. A. (Eds.) *Discourse on the Move. Using Corpus Analysis to Describe Discourse Structure*. Amsterdam / Philadelphia: John Benjamins, 155-173.

Huls, Erica (1987) Directives in Turkish. In H. E. Boeschoten and L. Verhoeven (eds), *Studies on Modern Turkish*, 242-258. Tilburg: Tilburg University Press.

Jucker, Andreas, Schneider, Gerold, Taavitsainen, Irma, Breustedt, Barb, 2008. “Fishing” for compliments. Precision and recall in corpus-linguistic compliment research. In: Jucker, A., Taavitsainen, Irma (Eds.). *Speech Acts in the History of English*. Amsterdam/Philadelphia: Benjamins, 273-294.

Kohnen, Thomas, 2008. Historical corpus pragmatics: Focus on speech acts and texts. In: Jucker, A., Taavitsainen, Irma (Eds.). *Speech Acts in the History of English*. Amsterdam/ Philadelphia: Benjamins, 13-36.

Manes, Joan, Wolfson, Nessa, 1981. The compliment formula. In: Coulmas, F. (Ed.), *Conversational Routines: Explorations in Standardized Communication Situations and Prepatterned Speech*. The Hague: Mouton Publishers, 115-132.

Norrick, Neal R., 2009. Interjections as pragmatic marker. *Journal of Pragmatics* 41, 5, 866-891.

Schmidt, Thomas, 2004. Transcribing and Annotating Spoken Language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*, Paris: ELRA. http://www1.uni-hamburg.de/exmaralda/Daten/4D-Literatur/Paper_LREC.pdf

Schmidt, Thomas, Wörner, Kai, 2009. EXMARALDA – creating, analysing and sharing spoken language corpora for pragmatic research**.** *Pragmatics* 19, 4, 565-582.

Searle, John, 1976. A classification of illocutionary acts. *Language and Society,* 5, 1-23.

Spoken Turkish Corpus. http://std.metu.edu.tr/en/

Wörner, Kai, n.d. Sextant tagger. http://www.exmaralda.org/sextant/sextanttagger.pdf