Structural Metadata Annotation: Moving Beyond English

Stephanie Strassel[1]*, Jáchym Kolář*[2]*, Zhiyi Song*[1]*, Leila Barclay*[1]*, Meghan Glenn*[1]

Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania, USA[1]
Department of Cybernetics, University of West Bohemia in Pilsen, Czech Republic[2]
{strassel,zhiyi,lbarclay,mlglenn}@ldc.upenn.edu, jachym@kky.zcu.cz

Abstract

The goal of metadata extraction (MDE) is to enable technology that can take raw speech-to-text output and refine it into forms that are more useful to humans and to downstream automatic processes. Starting in 2003, a structural metadata annotation task was defined for English as part of the DARPA EARS Program. A significant new challenge for MDE is the addition of new languages. This paper reports on work undertaken to apply MDE annotation to data from three very different languages: Mandarin Chinese, Levantine Arabic, and conversational Czech. Details of annotation task modifications are provided for each language; along with a general overview of data and annotation tools for non-English MDE.

1. Introduction

The goal of metadata extraction (MDE) is to enable technology that can take raw speech-to-text output and refine it into forms that are more useful to humans and to downstream automatic processes. In simple terms, this means the creation of automatic transcripts that are maximally readable. This readability might be achieved in a number of ways: creating boundaries between natural breakpoints in the flow of speech; flagging non-content words like filled pauses and discourse markers for optional removal; identifying sections of disfluent speech; and applying natural orthography and conventions for representing speaker turns and identity.

As part of the DARPA EARS (Efficient, Affordable, Reusable Speech-to-Text) Program, Linguistic Data Consortium at the University of Pennsylvania creates linguistic resources to support MDE technology evaluations. Initial work focused on annotation of broadcast news and conversational telephone speech data in English; recent efforts have extended the tasks to include Mandarin Chinese and Levantine Arabic conversational telephone data as well.

The research group at the Department of Cybernetics, University of West Bohemia (UWB) has further extended the MDE annotation task for Czech. This effort is primarily a front-end for NLP applications (speech summarization, information retrieval, machine translation, etc). Although Czech is not currently part of the EARS Program, the existing EARS MDE task lends itself well to these goals. To this end, a Czech spontaneous speech corpus of radio discussions has been annotated for MDE; and annotation of additional Czech data in new domains is planned, including broadcast news and sports commentaries. Czech is a good test bed for the Slavic MDE, because Czech is probably the most explored Slavic language for ASR research; and conclusions from Czech should be largely applicable to other Slavic languages.

2. The MDE Annotation Task

The earliest efforts to define an MDE annotation task relied heavily on previous work, in particular the Meteer manual for disfluency tagging of the Switchboard Corpus [1]. The early MDE task definitions within EARS were known as Full MDE; this task definition cycle culminated in the production of a set of pilot data labeled to the Full MDE Specification V2.6. Pilot annotation revealed a number of problems with the task specification; most importantly, many tasks could be performed only with very limited annotation consistency. In response, LDC developed a new task definition that eliminated some annotation tasks entirely and simplified others, with the goal of creating a task that could be performed by non-linguist annotators with reasonable consistency. This new Simple MDE task was the basis of the RT-03 EARS MDE evaluation; minor changes in 2004 resulted in SimpleMDE V6.2 [2]. This version supported the RT-04 EARS MDE evaluation and provided input for the development of non-English MDE annotation guidelines.

2.1. Fillers

In the context of MDE, fillers are defined as words that do not alter the propositional content of the material into which they are inserted, and their insertion does not depend on the word identities of the surrounding material. MDE annotation includes four types of fillers: filled pauses, discourse markers, asides/parentheticals and explicit editing terms.

Filled pauses (FP) are non-lexemes that speakers use to indicate hesitation or to maintain control of a conversation. Every language has a limited set of canonical FPs, though other non-words can occasionally be used as FPs. A discourse marker (DM) is a word or phrase that functions primarily as a structuring unit of spoken language. A DM signals the speaker's intention to mark a boundary in discourse, like a change in speaker or the beginning of a new topic. There is no exhaustive list of DMs for a given language due to their wide range of functions, colloquial variations, and the difficulty of defining them precisely. Asides and parentheticals (AP) occur when the speaker utters a short side comment either on a new topic (asides) or on the same topic of the larger utterance (parentheticals), then returns to the main topic. Both break up the stream of discourse and are often accompanied by noticeable prosodic features. Explicit editing terms (EET) occur during an edit disfluency, and consist of an overt statement (e.g., *I mean*) from the speaker recognizing the disfluency.

2.2. Edit Disfluencies

Edit disfluencies occur when a speaker corrects or alters his utterance, or abandons it entirely and starts over. Edit disfluencies have a more complex internal structure than fillers, consisting of the original utterance (reparandum), an interruption point, an editing phase and a correction. There are four types of disfluencies: repetitions; revisions; restarts; and complex disfluencies, which consist of multiple or nested edits. In Simple MDE, annotators label only the deletable region (DELREG) of the disfluency, which corresponds to the reparandum. In cases where the reparandum contains multiple disfluent utterances, annotators identify the maximal extent of the disfluent portion, starting with the left edge of the first disfluency and continuing to the right edge (IP) of the final disfluency.

2.3. SUs

One of the goals of MDE annotation is the identification of all units within the discourse that function to express a complete thought or idea on the part of the speaker. Within MDE these elements are called SUs (Syntactic, Semantic or Slash Units). As with disfluency annotation, the goal of SU labeling is to improve transcript readability by presenting information in small, structured, coherent chunks.

There are four sentence-level SUs. Statements are complete SUs that function as a declarative statement and are marked with **/.**; questions are complete SUs that function as an interrogative and are marked with **/?**. Backchannels are an open class of words uttered by the non-dominant speaker to indicate engagement in the conversation and are marked with **/@**. Incomplete SUs occur when an utterance does not constitute a grammatically complete sentence, phrase or continuer, and does not express a complete thought; these are marked with **/-**. To enhance inter-annotator consistency, there are also sentence-internal clausal and coordinating SUs (/, and /&).

3. Non-English MDE

A significant new challenge for MDE is the addition of new languages. As part of EARS, LDC began pilot work in 2004 to extend the English annotation task to Mandarin Chinese and Levantine Arabic. UWB has further developed the task for spontaneous Czech. In each case, the English task definition served as a starting point. Native speakers of each language were first trained to perform the English annotation task with consistency; they then began a multi-stage, cyclic annotation effort to produce both language-specific guidelines and annotated data in the new target languages.

With all three languages under discussion, the MDE annotation task has been applied only to spontaneous speech data; read speech has not been considered. The data annotated for Chinese and Arabic includes conversational telephone speech collected under LDC's Fisher telephone collection protocol [3]. The Chinese calls were drawn from a larger corpus of over 200 hours of transcribed Mandarin telephone speech collected by HKUST. The Arabic data is drawn from the Fisher Levantine Arabic corpus which consists of over 500 hours of speech from participants living in Jordan and Lebanon. The speech corpora and MDE annotations have already been distributed to EARS sites; future plans call for additional MDE annotation plus general publication of the data through LDC.

The Czech MDE corpus consists of recordings from the radio program Radioforum broadcast by Czech Radio 1. Radioforum is a live discussion show, where invited guests spontaneously answer topical questions asked by 1-2 interviewers. The material includes passages of interactive dialog, but longer stretches of monolog-like speech prevail. In all, the corpus contains 24 hours of transcribed speech. A more detailed description of the corpus is given in [4].

In addition to annotation guidelines, customized annotation software is required to handle the range of languages and variable task definitions. LDC created an MDE annotation tool using its Annotation Graph Toolkit (http://www.ldc.upenn.edu/Projects/MDE) [5]. The tool supports English, Chinese, Arabic and other languages and is highly customized for the MDE task, allowing users to highlight relevant spans of text, play the corresponding speech segments, and then record annotation decisions with a few mouse clicks or keystrokes. A new tool called Quick Annotator (QAn) was developed for Czech annotation (http://www.mde.zcu.cz). QAn has similar functionality as the AGTK MDE Toolkit, but utilizes a simpler linear XML-like format based on the Transcriber (.trs) format, with special MDE extensions. Conversion between the QAn format and LDC's AG format is possible.

3.1. Chinese MDE

3.1.1. Fillers and Edits

The fundamental concept of fillers ports well to Mandarin Chinese. During the pilot annotation effort, filled pauses, discourse markers and EETs were labeled. The set of filled pauses in Chinese includes: 嗯（En), 唔（Um), 呃（Eh) and 啊（Ah). Some filled pause words can also act as backchannels:

***EN* as Filled Pause**: A: 作业*嗯*太多了 /.
There is too much uh homework /.
***EN* as Backchannel**: B: 嗯 /@
Oh /@

Like filled pauses, discourse markers in Chinese are functionally similar to English. One Chinese-specific challenge is the presence of sentence-final particles like 吗, 吧, 咯, 呢, 啊, 呀, and 么. These items can have an interrogative, modality, or discourse function in spoken Chinese; so in some cases they are quire similar to discourse markers. But because distinguishing among the three functions is quite difficult, and because particle usage is highly variable from one speaker to another, the current task definition does not label sentence-final particles at all. Interjections and emotives are also very common in spoken Chinese but are not labeled, although they can sometimes look like discourse markers. These items are primarily used to express emotion (admiration, surprise, sadness, blame) or to draw attention; they do not primarily serve to structure the flow of conversation, as is the case with true discourse markers as defined by MDE.

Revisions and restarts function identically in Chinese and English. Repetitions, however, pose a challenge to the current MDE task, largely due to the fact that in Chinese there is no clear concept of a word. A word in Chinese can contain only one character or multiple characters. Repetition occurs when the speaker repeats a single phoneme, a string of words or the full sound of a word more than once. In the following example, [*Chufei*] is a repetition of a whole word (two syllables) and [*ke*] is a repetition of a character that is part of the word of *kefu*:

[除非] 除非是自己[克] 克服不了的罗/.
[Chufei] chufei shi ziji [ke] kefu bu liao de LUO /.
Except except that it's a problem I can't sol- solve myself /.

3.1.2. SUs

Significant challenges arise in defining SUs for Chinese. Unlike English, semantic and syntactic boundaries do not often coincide in Chinese; this motivates a major reworking of the rules for labeling sentence-external as well as sentence-internal SUs. The following example illustrates one such challenge:

a: 李四这个家伙	lisi(1) this dude
b: 我因为救他	I (2) because save him
c: 受了伤	0(2) receive wound
d: 居然不来看我	0(1) even not come see me
e: 跑到纽约度假去了 /.	0(1) run to New York have vacation go LE /.

I hurt myself because of trying to save Lisi. Lisi that dude didn't even come to visit me. He went to New York for a vacation instead.

As the utterance is expressed in Chinese, there are two complete SUs, one embedded within the other. Such structures do not occur in English, and the existing SU rules do not specifically address such examples. By extension from English, one might be tempted to annotate three separate SUs (*a-c, d, e*). However, in Chinese this example is treated as a single SU because the whole utterance references the same subject/topic – "Lisi, that dude". Native speakers of Chinese interpret this passage as a single "sentence" and would only apply end-of-sentence punctuation after the final clause (*e*). Furthermore, creating separate SUs in Chinese would isolate clause *a* from *d* and *e* which rely on it for completion.

Clauses combine to form discourse units in Chinese. Several devices exist for building discourse units: prosodic elements; topic chains in which a set of clauses is linked by a topic in the form of zero anaphora [6]; and linking words like adverbs, conjunctions or subjunctives. When these devices are overtly expressed, clauses are annotated as a single SU. Otherwise, annotators are instructed to follow a rule of thumb: when the clauses share a subject, treat them as a single SU; if they have distinct subjects, treat them as multiple SUs. However, clauses with different subjects are sometimes closely linked to each other without an overtly expressed device. For example:

你不相信，我做给你看 /.
You not believe, I do to you see /.
If you don't believe it, I'll do it for you to see /.

These could be treated as two SUs, but the conditional subordinate relationship is missing. A more intuitive option is to treat such cases as a single SU.

3.2. Arabic MDE

3.2.1. Fillers and Edits

As is the case with Chinese, fillers in Arabic are functionally similar to English. Filled pauses for all Arabic dialects include *ah, eh, ooh, mhm, uh, hmm*. As with Chinese, some filled pause words play other roles in spoken Arabic; this situation is further complicated by variation among different dialects of Arabic. For instance, the pause fillers *eh* and *ah* can mean "yes" in the Lebanese and Jordanian dialects, respectively. The English annotation practice of automatically pre-tagging common filled pauses must be revised for Arabic to include an additional manual check, and annotators must be instructed to be sensitive to the occurrence of dual-function words.

Dialect variation also complicates the situation for Arabic discourse markers. There exist several pan-dialectal DMs, for instance يعني (ya'3ni) "*I mean/it means*", بتعرف (bti'3rif) "*you know*", أوكي (okay) "*okay*", and حلو (Hilu) "*nice*". Conversely, there also exist dialect-specific forms such as Lebanese أو هيك (aw heik) "*or such*" and طب /طيب (Tab, Tayyib) "*good, okay*" in Jordanian Arabic. The use and choice of discourse markers varies widely depending on the geographic origin of the speaker. In mixed-dialect settings, speakers will attempt to be more precise in their speech, employing fewer discourse markers, but using a correspondingly higher rate of filled pauses.

One interesting case in spoken Arabic is the use of the name of God as a discourse marker or as a backchannel, in the form والله (wallah) which literally means "*by God.*" When used in its literal sense, wallah is not a discourse marker. The DM usage of ***wallah*** can take on many meanings, such as "*really*," "*actually*," "*okay*" or "*good enough*"; these are distinguished by variable prosody.

Edit disfluencies in Arabic are structurally identical to those in English, for all types; the English annotation rules seem to be adequate for handling Arabic as well.

3.2.2. SUs

SUs have been generally easy to recognize in the Levantine Arabic data. One source of complexity in the placement of SU breaks is caused by word order differences across varieties of Arabic. In Modern Standard Arabic (MSA), which is prevalent in anchored broadcast news data, the standard word order is Verb-Subject-Object. In colloquial Arabic like the Levantine dialect spoken in the current corpus, the word order is fluid. This fact along with other major differences between MSA and each Arabic colloquial dialect argues for separate annotation guidelines for each variety.

A notable feature of SUs in Arabic is their length. SUs are in general much shorter than in English; in fact, they can be as short as one verb, since subject dropping is prevalent and verbs carry full inflection. A sentence containing multiple dropped-subject clauses, each joined by coordinating conjunction, would be treated as a single statement SU with internal coordinating SU breaks in the English task definition. However, a more satisfying treatment for Arabic labels each clause as a separate sentence-level SU, because each clause can stand alone as an expression of a complete idea. For instance:

بده سيارة جديدة \. و يسافر \.
Biddu siyyara jdeedi /. Wysafir /.
He wants a new car **/.** *And to travel /.*

An exception is made when a semantic dependence exists between the two clauses, in which case the rule for English applies: the two clauses are joined as a single sentence-level SU but may also include an internal coordinating or subordinating break.

3.3. Czech MDE

Among the three languages discussed in this paper, Czech appears on the surface to be the most similar to English because of the romanized orthography, the common European origin and influence of Latin. However, despite some superficial similarities, Czech syntax differs significantly from English, and the MDE annotation rules, especially for SUs, required significant modifications.

3.3.1. Fillers and Edits

The concept of fillers ports very well to Czech, with a few exceptions. There is no generally agreed upon treatment for Czech FPs. In the Czech MDE corpus, FPs are treated as non-speech events rather than words. To aid annotation consistency, only two types of FPs are distinguished: the more frequent **EE** (similar to English *uh, er, eh*) and the rarer **MM** (sequence of consonant-like sounds, most often *mm* or *ww*). Likewise, the use of DMs in spontaneous Czech is similar to English. Short DMs like “no” (*well*) and “tak” (*so*) prevail over DMs containing a verb like "víte" (*you know*) and “podívejte se” (*you see*).

A/Ps are very frequent in the Czech radio corpus. As in English, some very common words or short phrases like “řekněme" (*say*) and “například" (*for example*) are not annotated as A/Ps; these "lexicalized parentheticals" were specified for annotators; a short list of common phrases that are treated as A/Ps was also prepared, including “řeknu příklad" (*(I) will say an example*) and “cituji” (*(I) quote*). EETs function as in English and are quite rare. The most frequent Czech EET is *“*nebo*”,* (*or)*. Edit disfluencies in Czech are very similar to English, but also appear surrounding A/Ps. In longer A/Ps speakers often repeat or revise the last word(s) uttered before the A/P, immediately after it.

3.3.2. SUs

The SU annotation task proves most challenging for Czech. As with Chinese and Arabic, subject dropping is an issue. In Czech and almost all Slavic languages, the subject (pronoun) can be dropped every time it is understood, from either the context or the form of the conjugated verb (predicate). Coordinated clauses are separated with an SU-external break, even if the subject is present in the first clause and dropped in the second clause; for instance “Pokusil se posadit /. ale nepodařilo se mu to /." *(He tried to sit up /. but didn't succeed /.*). If both predicates share an auxiliary verb (i.e. it is dropped in the second clause), the clauses cannot stand alone and a coordination break is used, as in "Dnes večer budeme studovat /& a potom odpočívat /." (*Tonight we will study /& and then have a rest /.).*

Czech syntax discriminates between compound sentences sharing a single common subject, and simple sentences with compound (multiple) predicates. Unfortunately, there is not absolute agreement in the literature on the borderline between them. For our purposes, the compound predicate is recognized if 1) The predicate verbs share a common constituent (e.g., object), as in "Nacpal /& a zapálil si dýmku /.” *(He filled /& and lit up his pipe /.);* or 2) The predicate verbs joined by a copulative conjunction have the same or very similar meaning, as in *“*Naši hosté často slaví /& a radují se/". *(Our guests often rejoice /& and celebrate /.)* In order to support annotation consistency, parts of compound predicates are separated by a coordination break.

The other rules for recognizing coordination breaks remain the same as for English. For clausal SU breaks, some minor modifications were applied, such as separating relative clauses with clausal breaks.

4. Conclusion

In this paper, we have described the structural metadata annotation task as defined for the EARS English MDE evaluation, and its extension to non-English for the first time. We have provided description of numerous language-specific modifications to the English MDE annotation task that were required to support annotation in spontaneous Mandarin Chinese, Levantine Arabic and Czech. We have also described the data and annotation tools developed to support non-English MDE.

Future plans for non-English MDE include additional annotation and guidelines development, including more coordination among the divergent task definitions. For all three languages, annotation of data in new genres, in particular broadcast news, is planned. In addition, several extensions adopted for Czech, including limited prosodic labeling at SU boundaries that distinguishes 2-3 categories, will be considered for all languages including English [4]. Finally, as task definitions are stabilized and additional data becomes available, we hope to distribute these linguistic resources – data, annotations, tools and guidelines – more broadly to the larger HLT community.

5. Acknowledgements

Support for this work was provided by the DARPA EARS (Efficient, Affordable, Reusable Speech-to-Text) Program and by the Ministry of Education of the Czech Republic, projects No. LC536 and MSM235200004.

6. References

[1] Meteer, M., et. al., "Dysfluency annotation stylebook for the Switchboard corpus," ftp://ftp.cis.upenn.edu/pub/treebank-/swbd/doc/DFL-book.ps, 1995.

[2] Strassel, S., “Simple Metadata Annotation Specification Versions 2.6, 5.0, 6.2”, http://www.ldc.upenn.edu/Projects/MDE, 2004

[3] Cieri C., Miller D., Walker K., "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text." *LREC 2004, Lisbon, Portugal,* 2004

[4] Kolar, J., Svec, J., Strassel, S., Walker, C., Kozlikova, D., Psutka, J.,“Czech Spontaneous Speech Corpus with Structural Metadata”, *Interspeech*, *Lisbon*, 2005

[5] Maeda, K and Strassel, S. "Annotation Tools for Large-Scale Corpus Development: Using AGTK at the Linguistic Data Consortium", *LREC 2004, Lisbon, Portugal,* 2004

[6] Chu, Chauncey C., "A discourse grammar of Mandarin Chinese." Peter Lang Publishing, New York, 1998