Transcription methods for consistency, volume and efficiency

Meghan Lammie Glenn, Stephanie M. Strassel, Haejoong Lee, Kazuaki Maeda, Ramez Zakhary, Xuansong Li

Linguistic Data Consortium, University of Pennsylvania

3600 Market Street, Suite 810, Philadelphia, PA 19104 USA

E-mail: {mlglenn,strassel,haejoong,maeda,rzakhary,xuangong}@ldc.upenn.edu

Abstract

This paper describes recent efforts at Linguistic Data Consortium at the University of Pennsylvania to create manual transcripts as a shared resource for human language technology research and evaluation. Speech recognition and related technologies in particular call for substantial volumes of transcribed speech for use in system development, and for human gold standard references for evaluating performance over time. Over the past several years LDC has developed a number of transcription approaches to support the varied goals of speech technology evaluation programs in multiple languages and genres. We describe each transcription method in detail, and report on the results of a comparative analysis of transcriber consistency and efficiency, for two transcription methods in three languages and five genres. Our findings suggest that transcripts for planned speech are generally more consistent than those for spontaneous speech, and that careful transcription methods result in higher rates of agreement when compared to quick transcription methods. We conclude with a general discussion of factors contributing to transcription quality, efficiency and consistency.

1. Introduction

This paper describes previous and ongoing efforts at Linguistic Data Consortium at the University of Pennsylvania to create manual transcripts as a shared resource to support human language technology research and evaluation. Research in speech recognition and related technologies in particular calls for large volumes of training data for system development, and for human gold standard references to evaluate system progress and development. LDC supports such efforts by providing a range of manual transcription approaches that are tailored to specific goals within a research program.

Recent efforts at LDC have targeted large-scale transcription of English, Arabic and Mandarin broadcasts, with smaller volumes in a wider range of languages in the conversational telephone speech, meeting and interview domains. The DARPA GALE program in particular has required LDC to create or commission hundreds of hours of Arabic and Chinese broadcast news and broadcast conversation transcripts to serve as training, development, and evaluation data for speech recognition. Creating manual transcripts on the scale demanded by programs like GALE can be costly and time-consuming. Data providers must strike a balance between cost and efficiency while still producing data that is useful for system development.

In this paper we give an overview of the different transcription guidelines LDC has created to promote efficiency and quality across transcription projects, languages, and domains, and report on real-time rates and inter-transcriber agreement observed for each of these categories. In designing inter-transcriber consistency experiments that would be representative of LDC's diverse transcription activities, we posited that the highest agreement rates would be achieved for carefully transcribed controlled speech with good audio quality, while agreement would decrease as the conversational nature of the recordings increased. We report preliminary consistency findings and discuss the impact of audio complexity on transcription agreement.

2. Transcription methodologies

All manual transcripts produced by LDC share the same core elements, which include time alignment at some level of granularity, speaker identification, and a transcript. Since there are often different requirements for system development versus system evaluation, LDC – with input from sponsors and researchers – has developed and published a set of transcription methodologies that target a range of data needs, from high volumes of approximate transcripts to small volumes of meticulously transcribed and annotated transcripts.

Each method strives to strike the appropriate balance among accuracy, efficiency, and cost while meeting program requirements. Each is also designed to apply with a unified approach to a variety of languages. Table 1 details the range of LDC's transcription methods, and includes required elements and approximate real-time rates for each.

2.1. Maximum efficiency

In 2002 a pilot experiment using 185 Switchboard calls showed that quick transcripts, which included automatic time alignment and a rough transcript, were of sufficiently high quality for system training purposes (Kimball, 2004). Even if the transcripts lacked some complexity of the recorded speech, the high volume of data made possible by this approach outweighed the possible disadvantages of less-precise

transcription. LDC's Quick Transcription (QTR) approach applies this principle and vastly accelerates real-time transcription rates, allowing a transcriber to complete one hour of data in approximately 5 hours for English (Strassel, et al., 2003). The QTR approach has since been adopted for the creation of training corpora, such as the 2003 Fisher English corpus and NIST Rich Transcription evaluations, among others. This transcription methodology optionally begins with automatic audio segmentation, which identifies speakers and divides the audio file into utterances. Transcribers listen to the automatically-produced segments and type what they hear, ignoring capitalization or punctuation rules, but marking a restricted set of non-lexemes (Cieri, et al., 2004).

Quickest	→			**Most Careful**
Segmentation	Automatic	Auto w/ verification	Manual	Manual w/ verification
Completeness	Content words	Add partial words, disfluencies	Add partial words, disfluencies	Add verification pass
Filled Pauses	Optional	Incomplete	Exhaustive	Exhaustive w/ verification
Disfluencies	None	Incomplete	Exhaustive	Exhaustive w/ verification
Transcriber Uncertainty	Flag and skip	Flag and best guess	Flag and best guess	Flagged best guess w/ verification
Feature Marking	None	Minimal	Full	Accurate, complete w/ correction
Speaker, Background Noise	None	Minimal	Exhaustive	Exhaustive w/ verification
Manual Passes	1	1-2	2-3	4+
Approx. Cost (x Real Time)	5 x	15 x	25 x	50 x

Table 1. Overview of transcription approaches, from quickest to most careful (Cieri and Strassel, 2009).

2.2. Efficiency and richness

The Quick-rich transcription (QRTR) approach was developed by LDC as an extension of QTR. The goal of QRTR is to add structural information like topic boundaries and manual SU annotation to the core components of a quick transcript. SUs are sentence-like units in spontaneous speech; they have semantic and syntactic cohesion and are critical for certain downstream tasks such as translation or part-of-speech annotation. QRTR also includes dialect identification for Arabic and Mandarin speech, where applicable. It is the prevailing transcription methodology for the DARPA GALE (Global Autonomous Language Exploitation) program, and has been used to produce thousands of hours of manual transcripts in Arabic and Chinese broadcast recordings for system training and development.

2.3. Maximum accuracy

Technology evaluations often require gold-standard references, which are produced with a Careful Transcription (CTR) method that involves multiple quality control passes and necessarily takes more time than a quick transcription approach. Elements of a careful transcript include a verbatim transcript; time-alignment to the level of sentences or breath-groups, speaker turns, and sections if required; consistent speaker identification; standard orthography and punctuation; markup of phenomena such as filled pauses, noises, and proper nouns; dialect annotation if applicable; and multiple manual and automatic quality control passes.

Method	**Transcript text**
QTR	well i don't know that i don't know that i'd score it as one better than the other i think every one of them to got a chance obama, edwards and senator clinton got a chance to provide a narrative of their own journey
QRTR	Well I don't know that uh I don't know that I would score it as one doing better than the other. I think that every one of them to got a chance Obama, Edwards and Senator Clinton got a chance to provide a narrative of their own faith journey.
CTR	Well I don't know that %uh – I don't know that I would score it as one doing better than the other. I think that every one of them to got a chance – %uh Obama, E- Edwards and Senator Clinton – got a chance to provide a narrative of their own faith journey.

Table 2. One excerpt, transcribed three ways.

3. Consistency analysis

3.1. Background

Scripted, measured speech by a single speaker will be less difficult for automatic processes and transcribers than spontaneous, multi-speaker conversations. An inter-transcriber consistency study conducted in 2004 as a part of the DARPA EARS (Effective, Affordable, Reusable Speech-to-Text) program illustrates this point. LDC and NIST examined careful transcripts of English broadcast news (BN) and conversational telephone speech (CTS) from the RT-03 test data. Broadcast news is primarily read speech, usually with minimal speaker overlap and good audio quality. CTS, on the other hand, is spontaneous conversation that comes with all of the challenges of unstructured speech – slang, disfluencies, and rapid speech, not to mention the acoustic variation in the telephone recordings.

The transcripts were carefully transcribed and scored with NIST's SCLITE toolkit (Fiscus, 2006).

They were also compared using a transcript adjudication GUI (graphical user interface) developed by LDC that loads two transcripts and masks regions of agreement so that annotators may label discrepancies, shown in Figure 1.

Adjudication resulted in a 1.3% "word disagreement rate" (WDR)[1] between two transcribers for the broadcast news data. A careful analysis showed that 81% of these discrepancies were caused by insignificant differences in punctuation, while the remaining disparities arose from misspelled words, contractions, disfluent speech, or disagreement over the morphological status of a word. WDR for CTS data reached 4.1-4.5%; close examination of the discrepancies revealed that 95% were marked as "judgment calls" due to contractions, rapid or difficult speech, or disfluencies (Strassel, 2004). Each label is described in more detail in section 3.2.4.

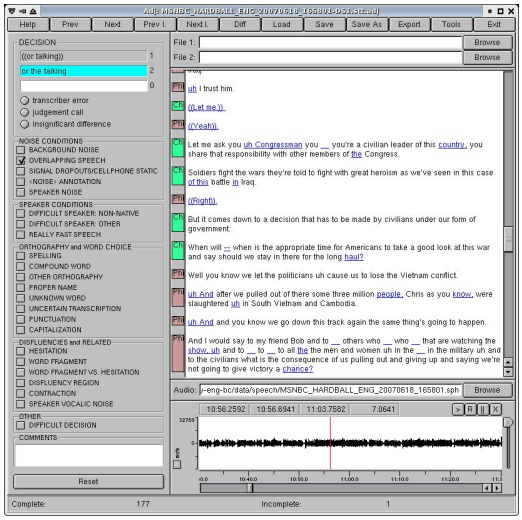

Figure 1. LDC Transcript Adjudication GUI.

In 2008 LDC conducted an impressionistic study of inter-transcriber consistency for highly conversational conference room data. For the NIST Rich Transcription 2007 conference room meeting test set, audio files were manually segmented and then assigned to two independent transcribers for a careful first pass transcript. Comparisons revealed that 64% of all dually-transcribed segments demonstrated some amount of disagreement, ranging from extreme disagreement where one transcriber understood the speaker completely differently from the other, to insignificant agreement such as punctuation variation (Glenn, 2008).

[1] *Word Disagreement Rate*. The number is calculated using SCLITE, which reports Word Error Rate. Since not all of the transcription "errors" are truly mistakes, we borrow this term from Strassel (2004) to refer to the percentage of disagreement between two transcribers.

3.2. Current analysis

The EARS study showed good inter-transcriber agreement on English BN and CTS data, with errors that are not detrimental to system development; however, these domains are not representative of the full spectrum of LDC's transcription approaches, audio genres, and languages. The current study targets a wider variety, in order to establish baseline human consistency rates for a broader range of local transcription efforts.

3.2.1. Data overview

In total, the current consistency study focused on 30 to 60 minutes for most of the following genres in English, Arabic and Mandarin: broadcast news, broadcast conversation, interviews, conversational telephone speech, and meetings. (Broadcast conversations include roundtable discussions, overlapping speakers, and rapid, highly disfluent speech.)

LDC selected English sociolinguistic-style interview and CTS transcripts that were produced for the Phonetic Annotation of Typicality in Conversational Speech (Phanotics) program, which supports forensic speaker recognition (Cieri, et al., 2008). English transcripts for conference-room meetings from the NIST Rich Transcription 2009 efforts were also selected. In addition to CTS, interview, and meeting recordings, LDC selected approximately 30 minutes of English broadcast news and conversation recordings, respectively, which were collected under DARPA GALE collection efforts. Also from GALE broadcast collection were approximately one hour of BC and BN transcripts for Arabic and Mandarin. In most cases, we also analyzed quick transcripts and careful transcripts for each language and genre combination, in order to evaluate the affect of transcription methodology on inter-transcriber agreement.

3.2.2. Methodology

Where possible, transcripts were generated by using identical time alignment. In some cases, the file was segmented manually and then assigned to two independent, trained transcribers for the first-pass transcript. In other cases, the time alignment from a completed transcript was extracted and assigned to a second transcriber. All transcripts were scored with NIST's SCLITE toolkit. (Fiscus, 2006) For the purposes of the current study, LDC ignored stylistic differences, such as capitalization or punctuation; however, since careful transcription requires consistent punctuation and capitalization, future analysis will include stylistic orthographic features in CTR comparisons. For a subset of English transcripts, further analysis was conducted using LDC's in-house adjudication GUI.

3.2.3. Scoring results

Preliminary results for all languages and genres, as shown in Table 3, support the findings of the RT-03 study: transcripts for controlled speech are generally more consistent than those for spontaneous data. We also observe that for most languages and domains – with the exception of Chinese BN – careful transcription methods result in higher rates of transcriber agreement when compared to quick transcription methods. Planned speech also generally produces better consistency between independent transcribers – regardless of the transcription methodology – than the more spontaneous genres.

Language	Genre	Careful Transcription WDR	Quick (Rich) Transcription WDR
English	CTS	4.1-4.5%	9.63% (5 pairs)
	Meeting	-	6.23% (4 pairs)
	Interview	n/a	3.84% (22 pairs)
	BN	1.3%	3.5% (6 pairs)
	BC	n/a	6.3% (6 pairs)
Chinese	BN	7.40% (23 pairs)	6.14% (18 pairs)
	BC	9.06% (24 pairs)	9.45% (4 pairs)
Arabic	BN	3.13% (14 pairs)	3.42% (16 pairs)
	BC	3.93% (12 pairs)	8.27% (18 pairs)

Table 3. Preliminary results with SCLITE scoring.

3.2.4. Results analysis

More detailed analysis was performed for most of the English-language quick transcripts, using LDC's customized transcription adjudication GUI. Annotators listened to and labeled each disparity as a "transcriber error," "insignificant difference," or "judgment call," just as in the EARS study.

Annotators label a discrepancy as a *transcriber error* when one transcriber omitted part of an utterance, transposed the order of words in an utterance, inserted words that were not originally spoken, or misunderstood the utterance. When both transcribers appeared to have made an error, the adjudicating annotator entered the correct transcription for that region. Approximately 15% of the differences across all the English quick transcripts were judged to be transcriber errors. The following CTS example shows a transcriber error:

Transcript	Decision	Analysis
A little bit? You sound like [you're not ready // you never going] to leave your friends.	transcriber error	"you're not ready" is correct

Insignificant differences, which are often caused by differences in capitalization or punctuation, speaker noise annotation variation, or spelling of hesitation sounds or partial words. Analysis showed that 65% of all discrepancies in the English quick transcripts belong to this category. For quick transcription approaches, omitting a disfluency is considered insignificant, since the goal of Q(R)TR is to produce content words for every utterance. The CTS example below shows an insignificant punctuation and capitalization difference:

Transcript	Decision	Analysis
the [scenes, but // scenes. But] to sit there and have a group and stuff like that and where you're actually	insignificant difference	both are correct

Judgment calls are cases where the adjudicator cannot deem on transcription for a particular utterance more correct than the other. Nearly 20% of all discrepancies in the English quick transcripts were labeled judgment calls. Such cases often occur in regions of disfluency or particularly fast or difficult speech.

Transcript	Decision	Analysis
Yeah [they would // then we] come [inside afterwards. // and sit afterwards.))]	judgment call	either option is plausible

Annotators optionally label each discrepancy in more detail, noting any audio conditions or speaker features that could have contributed to the disagreement. In the meeting domain example below, a single utterance contained three separate points of discrepancy: two judgment calls and one transcriber error, which was re-transcribed by the adjudicating annotator (marked in bold in the example). During adjudication, the utterance was also labeled as containing background noise and overlapping speech, which helps to explain the variation present in this transcript pair.

Transcript	Decision	Details
[Right so the // So ((it would be))] little things like wires and stuff we should just check on ~E bay and order them up.	judgment call	background noise
Right so the little things like wires and stuff [we should just check on ~E bay and // **we should just look up on E-bay and** // (()) in the] order them up.	transcriber error	background noise
Right so the little things like wires and stuff we should just check on ~E bay and order [them up. // of the –]	judgment call	background noise, overlapping speech

3.2.5. Transcription challenges

Different domains present unique challenges. The

conference room meeting domain, for example, poses difficulties to human transcribers and automatic processes alike by way of massively multi-channel sessions containing overlapping speech, whispered asides, non-native speakers, and "insider" language and content. English sociolinguistic interviews score slightly worse than broadcast data of either genre; many of the interviews contained idiosyncratic or rapid speech that could have contributed to lower inter-rater agreement overall.

Broadcast conversations present similar obstacles to consistent transcription: massively overlapping speech, multiple speakers, non-native speakers, and dialectal speech. Dialect poses a particular challenge in transcribing Arabic conversations. For the Arabic broadcast genres, Modern Standard Arabic (MSA) is the targeted language, but real data contains significant volumes of dialectal Arabic, especially in the broadcast conversation domain. In QRTR, transcribers mark all dialectal speech as "non-MSA," but do not identify individual dialects at a finer granularity. Broadcast conversations may contain multiple dialects in a single recording. The transcriber's personal knowledge or background will impact his or her ability to transcribe multiple Arabic dialects, which contributes to lower agreement in the conversational domain. The following BC example shows several instances of non-MSA terms that were transcribed differently.

Arabic transcript	مو مشكلة يعني إحنا المواضيع [متاعنا // إمتاعنا] [ما فيهاش // مافيش] مشكل ثلاثة محرمات [هاذي // هذه] موجودة الجنس والسياسة والدين ما [تتقربولهمش // تتقربلهومش]
English translation	No problem, we don't have problems in our subjects, three restrictions, sex, politics, and religion, you shouldn't approach.
Analysis	Non-MSA terms spelled differently

Transcription of Mandarin BN and BC recordings is less often complicated by dialect than Arabic, but it too becomes increasingly difficult as the data grows more complex. In particular, strongly-accented speech affects transcription quality; transcribers often struggle to decide if particular terms are mispronounced or merely accented speech.

Chinese transcript A	*最近网上许多人都说 [many people on the internet recently said]*，我们今天我们是汶川人
Chinese transcript B	*这些 [these] ((涌向 [flow to] 体温 [body temperature])) 不说*，我们，今天我们是汶川人
Analysis	unclear pronunciation, conversational style produce different interpretations

In QRTR, transcribers create time-aligned SUs while producing a near-verbatim transcript of very spontaneous speech, and sometimes experience uncertainty in identifying sentence boundaries consistently and efficiently. The following example illustrates SU annotation variability.

Chinese transcript A	我首先在这里面我要道歉，我要跟姜岩道歉，但是可能没有用。
Chinese transcript B	我首先在这里面我要，道歉，我要跟姜岩道歉，当然可能没有用，*但是我也－我还要就是跟她父母，道歉，因为就说不管怎么样我是她的丈夫，这件事情肯定跟我有很大的关系*
Translation of region of difference	*However, I also -- I will still say sorry to her parents because I'm her husband anyway, and this definitely affects me greatly*
Analysis	Segment length reveals transcriber variability in identifying sentence boundaries based on conjunctions

As we found with the other languages, regions of disfluency are by far the most prevalent contributors to transcriber disagreement in quick-rich Mandarin transcripts: the number of filled pauses such as "呃 [er]", "呵 [ah]", and "哎 [eh]" often varies; backchannels such as "嗯[eh]" and "对[yes]" may be missed; and partial words are often left out. Conversational data tends to contain many articles or determiners such as "这个 [this]" or "那个 [that]", which are frequently omitted. The BC excerpt below shows higher disagreement around hesitation sounds and other disfluencies.

Chinese transcript	[// 对，] 所以啊，你看那么迅速的期间 [啊 // 呢]，其实对我们来讲，从时间来看呢，我们的 [这-这个 // 这个]，所有的这个解放军也好，武警也好，他们把 [他 // 它] 当作一个战争的任务，是分秒必争
Translation	So, ah, you see how fast they are. Actually, to us, in terms of time, eh, uh, all our, uh, uh, PLA, as well as the PAP troops, they treated it as a war mission and fought for every second.
Analysis	filled pauses and repeated partial words missing or transcribed differently; non-standard pronoun employed by one transcriber

Conversational data may also contain speakers who use dialect words instead of standard Mandarin. This introduces transcription irregularities since the character set may not support consistent transcription of dialectal words or phrases. Another unique challenge encountered in Mandarin BC transcription is onomatopoeic terms, for which there are often no characters in the character bank. Transcribers use their best judgment or mark such terms as "uncertain."

4. Conclusions and future work

This paper has given an overview of LDC's manual transcription approaches, and has shown that humans demonstrate a high level of agreement on carefully transcribed, read speech in English, and that agreement rates for quick transcription for conversational telephone speech are also good. We

show that as the complexity of the speech increases, so does the disagreement between two or more independent transcribers.

The preliminary results presented in this paper offer opportunities for future work, including deeper analysis of the discrepancies among transcribers for efforts in Arabic and Mandarin – particularly for the careful Chinese transcripts – and further exploration of English meeting recording transcription consistency.

The resources described in this paper will be made available to the broader research community over time. Many resources have already been distributed to LDC members and non-member licensees through the usual methods, including publication in LDC's catalog. Other resources including transcription specifications and tools are freely distributed via LDC's website. Transcription specifications are available at http://projects.ldc.upenn.edu/gale/Transcription/, and LDC's in-house transcription tool, XTrans, which was used to create all of the transcripts discussed in this paper, is freely available at http://www.ldc.upenn.edu/tools/XTrans.

5. Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

We gratefully acknowledge the LDC transcription team for their work and analysis, and extend our thanks to Jonathan Fiscus for his guidance in using SCLITE.

6. References

Cieri, C. (2006) "What is Quality? Workshop on Quality Assurance and Quality Measurement for Language and Speech Resources." LREC2006: The 5th Language Resource and Evaluation Conference, Genoa.

Cieri, C., S. M. Strassel. (2009) "Closer Still to a Robust, All Digital, Empirical, Reproducible Sociolinguistic Methodology." NWAV 38: New Ways of Analyzing Variation, University of Ottawa, Ottawa, Canada.

Cieri, C., S. M. Strassel, M. Glenn, R. Schwartz, W. Shen, J. Campbell. (2008) "Bridging the Gap between Linguists and Technology Developers: Large-Scale, Sociolinguistic Annotation for Dialect and Speaker Recognition." LREC2008: The 6th Language Resource and Evaluation Conference, Marrakech, Morocco.

Fiscus, J., J. Ajot, N. Radde, C. Laprun. (2006) "Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech." LREC2006: The 5th Language Resource and Evaluation Conference, Genoa.

Glenn, M., S. M. Strassel. (2008) "Shared Linguistic Resources for the Meeting Domain," in Stiefelhagen, R., R. Bowers, and J. Fiscus, (Eds.), in Lecture Notes in Computer Science, vol. 4625, Multimodal Technologies for Perception of Humans, Heidelberg: Springer, pp. 401-413.

Kimball, O., C. Kao, R. Iyer, T. Arvizo, and J. Makhoul. (2004) "Using quick transcriptions to improve conversational speech models," in Proc. Int. Conf. Spoken Language Process. Jeju Island, Korea, pp. 2265–2268.

Strassel, S. M., D. Miller, K. Walker, and C. Cieri. (2003) "Shared resources for robust speech-to-text technology," in Proc. EUROSPEECH 2003, Geneva, Switzerland.

Strassel, S. M. (2004) "Evaluation resources for STT: Careful Transcription and Transcriber Consistency." Unpublished Manuscript.