On Development of Consistently Punctuated Speech Corpora

Jáchym Kolář, Lori Lamel

Spoken Language Processing Group, LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
{jachym,lamel}@limsi.fr

Abstract

Punctuation of automatically recognized speech is important to enhance readability of transcripts and to aid downstream NLP processing. This paper is concerned with issues involved in developing training and test corpora for automatic punctuation systems. Punctuation annotation in speech transcripts is difficult since there are numerous cases for which no standard punctuation rules exist. Special punctuation annotation guidelines tailored to spoken language were developed. Using these guidelines, almost 100 hours of broadcast news and conversation data in English and French have been punctuated by trained annotators. Measures of inter-annotator agreement are provided for both languages and differences between languages and genre are analyzed and discussed, along with some of the most frequent disagreements between annotators. Overall, using the guidelines, the annotation consistency has been significantly improved.

Index Terms: punctuation, speech corpus, inter-annotator agreement, English, French

10.21437/Interspeech.2011-307

1. Introduction

The last two decades have witnessed significant progress in the area of automatic speech recognition (ASR). Large amounts of data can now be transcribed automatically. However, automatic transcripts typically do not have a form convenient for reading or downstream processing. The problem is that many ASR systems still output either only a raw stream of words, leaving out the structural information conveyed by punctuation in standard text completely, or text punctuated by using simple and very inaccurate punctuation models. To get an idea how punctuation may change meaning of an utterance, take a look at the following example — the sentences *"Woman without her man is nothing."* and *"Woman, without her, man is nothing."* have the same words but very different meanings. This is an extreme case, but various forms of ambiguity are quite frequent.

As shown by a number of studies, the absence of punctuation is confusing both for humans and computers. For example, Jones et al. [1] demonstrated that sentence breaks are critical for legibility of speech transcripts. Also many natural language processing (NLP) techniques typically trained on well-formatted text have problems when dealing with unstructured word streams. For instance, Furui et al. [2] reported that speech summarization improved when sentence boundaries were provided, and Matusov et al. [3] showed that the use of punctuation is beneficial for machine translation. To this end, we aim to produce an automatic punctuation system that would present the recognized text in a much more informative fashion. This paper deals with issues involved in development of reliable training and test data for such a system.

Several past publications have studied the problem of automatic detection of non-verbalized punctuation in speech [4, 5, 6, 7, 8, 9, 10]. However, these studies have focused on development of automatic punctuation methods rather than on the quality of reference punctuation in the speech corpora used. They have typically employed standard corpora distributed by LDC such as Hub-4, Switchboard or ACE, but none of these resources has been punctuated based on some punctuation guidelines tailored to speech. The only exception is [10] where the authors claim that their Portuguese data were revised by a linguist who corrected "many inconsistencies in punctuation". Nevertheless, the paper does not report whether some speech-specific punctuation guidelines were used or not.

This paper describes recent efforts towards improving punctuation of speech transcripts created for use in the Quaero project. The project focuses on the development of multimedia and multilingual indexing and management tools for professional and general public applications. Its speech-processing part uses broadcast data containing both news- and conversation-style speech. Although Quaero currently deals with nine European languages, these initial efforts have focused on two of them – English and French. Other languages are planned to be added later.

In addition to presenting the annotation approach and the data, the paper focuses on the analysis of inter-annotator agreement on speech punctuation. It is not only important to measure the quality of the guidelines and annotation, but also to indicate an upper-bound of performance when evaluating automatic punctuation system. To our best knowledge, this the first paper that publishes the inter-annotator measures for this task.

2. Punctuation annotation guidelines

The original Quaero transcription guidelines were oriented towards traditional ASR evaluation. Therefore, in terms of punctuation, they just defined that only periods, commas and question marks are allowed, but did not recommend any particular style for using them to punctuate spoken data. When taking a closer look at the transcripts, we found the punctuation there highly inconsistent, and thus inadequate for explicit work on automatic punctuation. To improve the consistency, we decided to define special guidelines for punctuating speech transcripts, and then re-punctuate part of the data.

Development of punctuation guidelines for the spoken language is not an easy task. Speech transcripts contain plethora of cases for which no standard punctuation rules exist. Standard punctuation was designed to indicate the structure and organization of *written* language, so its use in transcribing *spoken* language is often ambiguous. Frequent speech-specific phenomena include filled pauses, disfluencies (repetitions, self-corrections, false starts) or anacolutha.

The first problem is to decide what punctuation symbols to use. There are generally two possible approaches. One option is to use some special set of symbols designed to suite the struc-

ture of (spontaneous) speech including disfluencies, backchannels or incomplete utterances. This approach was used for the Structural Metadata (MDE) annotation [11] in the EARS project. Another option is to use a subset of punctuation for written language.

We have adopted the latter approach for a number of reasons. First, the special symbols would require some subsequent conversion for readability. Second, the annotation in terms of some structural metadata is more complex, and thus requires more annotator training and is more costly to produce. Third, automatic punctuation systems require large amounts of text data to train the language models, but standard text resources do not include any special symbols. Therefore, they could not be used for training directly.

The next step was to decide which subset of punctuation marks to choose. Using full punctuation would represent a too difficult task for an automatic system, mainly because of sparse training data for many punctuation marks. To this end, the three most frequent marks are used – comma, period, and question mark. We also contemplated using other marks, namely a dash to indicate a break in the flow of the sentence, and an ellipsis to mark incomplete sentences. Finally, we decided not to add them based on previous studies [8] which reported problems with their use because of the imprecise definition and data sparsity.

The guidelines proposed for both languages follow standard grammar and style books where possible, however, the standard conventions had to be extended to accommodate the phenomena specific to speech. The punctuation guidelines have two parts. The first deals with the general punctuation rules that are also valid for text, while the second focuses on punctuating speech. In particular, the first part reviews the most important grammar rules about punctuation in order to remind the annotators of them. It also provides illustrative examples of common errors. For some cases where multiple ways of punctuation are correct, it defines the preferred way. For example, we instruct the annotators not to use the Oxford comma (i.e., not to put comma after *grey* in *white, grey and black*). This first part is obviously very language-specific because English and French use quite different punctuation conventions. In contrast, the second part dealing with the specifics of speech transcripts is very similar for both languages. The most important rules of the second part are presented in the following paragraphs.

Since the punctuation is limited to question marks, periods and commas, it is necessary to substitute other standard marks (exclamation mark, ellipsis, dash, quotation mark, etc.) with a comma or leave them out without a substitution. For example, exclamation marks are simply replaced by periods. Direct (reported) speech is separated from the reporting clause by a comma. When a question is embedded in reported speech, the final punctuation of such a sentence is question mark (e.g., *Paul asked, who is that girl?*). On the other hand, a sentence with indirect speech is always terminated by a period, even if it reports a question (*Paul asked who the girl was.*). When reported speech contains more than one sentence, the sentences are separated by end-of-sentence punctuation. Furthermore, complete and incomplete sentences are not differentiated — both are marked by a period or a question mark at the end.

It is often difficult to decide whether to use a period or a comma at coordination breaks within "never-ending" compound sentences, which are common in spoken discourse. To support readability, the guidelines instruct annotators to avoid creating overly long sentences unless it is really the only option. On the other hand, if two independent clauses connected by a coordinating conjunction form a semantically and prosodically coherent and not overly long unit, they are separated just with a comma because such a presentation is easier to read (*Many people skip breakfast, but I need to eat something in the morning.*).

Our approach prefers not to put any comma that does not have syntactic or semantic motivation. Silent pauses that are not *very* long (say > 1 second, or significantly longer than other pauses) do not motivate any additional punctuation. When the pauses are *very* long, the utterance is split into two sentences at the pause. Filled pauses also do not themselves imply any punctuation marks. If there is a filled pause or a noise tag between the two words that should be separated by a comma, the comma is always put before the filled pause or the special tag (*Barbara voted for Saturday, uh but Thomas voted for Sunday.*). If the utterance contains disfluencies, or if it is ungrammatical, it is punctuated according to its imaginary fluent and grammatical version. As a consequence, repetitions are not separated with commas (*He's he's really out of out of line.*).

Table 1: Basic characteristics of the re-punctuated data in English and French

	EN	**FR**
Duration of re-punctuated speech	45.1h	50.9h
Number of shows	101	103
Number of words	506.1k	512.3k
Avg. sentence length (#words)	17.0	16.9
Avg. #words between punc. marks	7.6	5.5

3. Data description

This work uses data selected and transcribed for the Quaero project. Both English and French data are split between Broadcast News (BN) and more varied data including talk shows, debates and web podcasts collectively called Broadcast Conversation (BC). The ratio between BN and BC is approximately 30% to 70%. The transcripts of development and evaluation data are produced using the detailed manual transcriptions (as used in NIST benchmark tests), while the transcripts meant to only be used for training are created in the quick transcription fashion. From Quaero resources for the two languages, recordings yielding in total approximately 500k words[1] for each language have been selected for punctuation re-annotation.

Some characteristics of the reannotated data are given in Table 1. Note that while the average sentence length is approximately the same for English and French, the average distance between two adjacent punctuation marks is lower in French. As shown in detail in Table 2, the frequency of periods and question marks is almost the same for both languages, but commas are much more frequent in French.

4. Inter-annotator agreement

An important part of this work was to evaluate the inter-annotator agreement (IAA) based on dually punctuated data. To this end, the Quaero 2010 development and test data created in the detailed transcription fashion were used. In total,

[1]The data were tokenized in the fashion we use for language modeling. Thus, English contracted forms (like *it's*) were kept concatenated, while most of the French were split at apostrophes (with the exception of *c'est* and *s'est*).

Table 2: Relative frequencies of punctuation marks [%]

Punctuation	EN [%]	FR [%]
Comma	7.2	12.3
Period	5.2	5.3
Quest. mark	0.7	0.7
None	86.9	81.7

Table 3: Inter-annotator agreement on English and French [K]

English	EN – Baseline	EN – Guidelines
Num. of words	91.0k	
Comma	0.51	0.70
Period	0.82	0.86
Quest. mark	0.88	0.87
Overall	0.70	0.81
Any vs. None	0.74	0.85
French	**FR – Baseline**	**FR – Guidelines**
Num. of words	89.0k	
Comma	0.65	0.70
Period	0.58	0.81
Quest. mark	0.81	0.80
Overall	0.68	0.79
Any vs. None	0.71	0.84

there are 91k words for English and 89k for French. Based on a show-level classification, 59 % of the English and 65 % of the French data are BC – the rest is BN. To perform this test, 4 native speakers with some linguistic background were employed for both languages. Not to be influenced by the original inconsistent punctuation, all punctuation marks were removed before passing the transcripts to the annotators. Using the words and corresponding audio, each show was punctuated by two of the annotators. The annotator pairs varied across the shows.

The K (kappa) statistic [12], which is considered to be a standard measure of agreement in many annotation tasks related to language processing, was used to measure IAA. It is defined as

$$K = \frac{A_o - A_e}{1 - A_e} \quad (1)$$

where A_o denotes the observed (i.e., measured in the test data) agreement and A_e stands for the expected agreement (i.e., based on label priors). If the annotators are in complete agreement, then $K = 1$, while agreement expected by chance corresponds to $K = 0$. The interpretation of values between 0 and 1 is not straightforward. In the original paper presenting this measure, Carletta claims that for tasks like content analysis, $K \geq 0.8$ is considered to be good reliability, and $0.66 < K < 0.8$ allows to draw tentative conclusions. However, note that these interpretations represent only a "rule of thumb" since they do not have profound theoretical background.

In addition to reporting IAA figures themselves, we compared the IAA achieved with the guidelines with the original punctuation consistency. Since the original data did not contain any dually punctuated transcripts, we had to designate the baseline in a different way. Instead of the original IAA, we consider the baseline as the average IAA between the original and the re-punctuated versions of the transcripts. The results of the agreement experiment for both English and French are displayed in Table 3. The table shows results for the three punctuation marks (period, question mark, comma) as well as the overall agreement. The row "Any vs. None" corresponds to the condition in which all punctuation marks are grouped together. Thus, the metric only takes into account whether there is a punctuation after the word or not. This figure gives insight into the impact of substitution errors.

Table 4: Differences in inter-annotator agreement between English BN and BC [K]

	EN – BN	EN – BC
Num. of words	37.2k	53.8k
Comma	0.72	0.69
Period	0.89	0.84
Quest. mark	0.85	0.88
Overall	0.83	0.79
Any vs. None	0.87	0.83

It can be seen that much higher agreement was achieved for periods and question marks than for commas. The magnitude of agreement on commas is the same for both languages, while the agreement on periods and question marks was higher for English. The overall IAA numbers indicate that the agreement is slightly higher on English than on French (0.81 vs. 0.79). Bear in mind that K is normalized for the chance agreement, which is, due to the higher proportion of punctuation, lower for French. This makes the total value of K for French close to the value for English despite the higher absolute number of disagreements. For "Any vs. None", the Ks are higher than the overall Ks by not very different proportions – 4.9 % relative for English and 6.3 % relative for French.

The comparison of IAA achieved in this test with the baseline IAA also shows similar improvements for both languages – 15.7% relative for English and 16.2% relative for French. However, the improvements come from different sources. In English, we have largely improved IAA for commas, whereas in French, we have mainly improved IAA for periods. The IAA for question marks did not improve indicating that transcribers are able to agree on what is a question without guidelines. Overall, K about 0.8 has been achieved for both languages, which is considered to be on the edge of a good reliability according to [12]. However, the partial Ks for commas are in the interval $0.66 < K < 0.8$, allowing only tentative conclusions to be drawn according to the same paper. Although special punctuation guidelines were used, annotation of commas was still influenced by subjective decisions of the annotators.

We also analyzed the influence of genre on IAA. Table 4 compares IAA for English BN and BC. We do not use the French data for this analysis because French annotators punctuated BN data earlier than BC, and thus were more experienced when doing the latter. Also, French data were more difficult to categorize because some of the TV news contained a lot of conversational material, which was not the case for English. The numbers indicate that IAA was higher on BN but the difference was not very large (4.8 % relative). Among individual marks, BN had higher IAA for commas and periods, while IAA on question marks was higher in BC.

5. Analysis of annotator disagreement

An analysis was carried out to gain insights into the disagreement between annotators. Table 5 shows the distribution of disagreement types. The order of the disagreement types is the

Table 5: Relative frequencies [%] of disagreement types

Disagreement Type	EN [%]	FR [%]
Comma – None	71.3	66.6
Comma – Period	19.7	26.8
Period – None	6.5	2.8
Period – Quest. mark	1.6	1.2
Comma – Quest. mark	0.6	1.5
Quest. mark – None	0.4	0.3

same, but their proportions differ between the two languages. The by far most frequent disagreement pair is a comma and no punctuation mark. The second place belongs to the comma-period confusions, which are typical for the "never-ending" sentences consisting of many coordinated clauses. These confusions are more common in French. In English, where commas are less frequent, part of these difficult sentence boundaries is projected to the period-none disagreements. Confusions involving question marks are infrequent, but this can be mostly attributed to the low frequency of questions in the data.

An analysis was also made of the most frequent contexts in which annotators disagree, comparing the bigrams across the inter-word boundary in question and the unigrams right before and after this boundary. For English, the most frequent bigrams of disagreements are (sorted by their relative disagreement rate): *now to, so they, so what, that in, so if*. The most frequent of these *now to* is typical for announcement of topic change in news, and the anchors often make a short pause after *now*, which causes confusions. Other frequent bigrams include *so* which is a difficult word because the annotator must judge whether it serves as a conjunction (no comma) or a discourse marker (followed by a comma). The most frequent unigrams before the disagreements are *so, today, fact, ok, now*, and the most frequent unigrams after the disagreements are *which, because, but, actually, who*. The relative pronoun *which* is difficult because one has to judge whether the following clause is restrictive (no comma) or non-restrictive (comma). In French, the most frequent disagreement bigrams are *donc c'est, oui oui, que si, et moi, que dans*. The case of French *donc c'est* is similar to the English bigrams with *so* – there is an ambiguity about the function of *donc* in this particular context. An interesting example is *oui oui*. We did not have a specific rule for the double use of the agreement word in direct answers so the annotators did not know how to punctuate it. The most frequent unigrams before the boundary are *puis, hier, ah, vraiment, mais* and the most frequent unigrams after *aussi, parce, mais, quand, non*.

The punctuation disagreements come from several different sources. After manually revising the data, we found 4 main categories: (1) Particular examples not covered by the guidelines; (2) The rule for the example depend on subjective assessment of prosody; (3) Completely ungrammatical structure; and (4) One of the annotators not used the rule correctly.

6. Summary and Conclusions

Punctuating automatically recognized speech is important to enhance readability of transcripts and to aid downstream NLP processing. In this paper, we have discussed some issues involved in the development of training and test corpora for automatic punctuation systems. We have developed speech punctuation guidelines and punctuated almost 100 hours of English and French data. Furthermore, we performed IAA tests on dually annotated transcripts. For both languages, a K about 0.8 was achieved, which is on the edge of what is considered to be "good" IAA. Among the individual punctuation marks, the IAA was higher on periods and question marks than on commas. In a genre-based comparison, slightly higher agreement was observed on BN than on BC.

Given the agreement achieved, we propose to use two or more references for future automatic punctuation evaluation in the Quaero program. The overall system performance might be evaluated either as the average of results on more references, or in a more lenient form, one might count a decision as an error only if the mark does not match any of the references. The quality of automatic punctuation may also be assessed in terms of the influence on the target downstream NLP module, or by measuring transcript readability in a perceptual test.

7. Acknowledgments

This work was achieved as part of the Quaero Program, funded by OSEO, French State Agency for Innovation. The authors also thank Ioana Vasilescu and Natalie Snoeren at LIMSI for their help with the organization of the annotation project, and Sandrine Courcinous and Julien Despres at Vocapia Research for providing us with the text normalization for both English and French.

8. References

[1] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, "Measuring the readability of automatic speech-to-text transcripts," in *Proc. EUROSPEECH*, Geneva, Switzerland, 2003.

[2] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

[3] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, M. Ostendorf, and H. Ney, "Improving speech translation with automatic boundary prediction," in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007.

[4] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in *Proc. of ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, USA, 2001.

[5] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. of ICSLP 2002*, Denver, CO, USA, 2002.

[6] J. H. Kim and P. Woodland, "A combined punctuation generation and speech recognition system and its performance enhancement using prosody," *Speech Communication*, vol. 41, no. 4, pp. 563–577, 2003.

[7] B. Favre, R. Grishman, D. Hillard, H. Ji, D. Hakkani-Tür, and M. Ostendorf, "Punctuating speech for information extraction," in *Proc. ICASSP*, Las Vegas, NV, USA, 2008.

[8] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *Proc. ICASSP*, Taipei, Taiwan, 2009.

[9] J. Kolář and Y. Liu, "Automatic sentence boundary detection in conversational speech: A cross-lingual evaluation on English and Czech," in *Proc. ICASSP*, Dallas, TX, USA, 2010.

[10] F. Batista, H. Moniz, I. Trancoso, H. Meinedo, A. I. Mata, and N. Mamede, "Extending the punctuation module for European Portuguese," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010.

[11] S. Strassel, "Simple metadata annotation specification V6.2," http://www.ldc.upenn.edu/Projects/MDE/Guidelines/SimpleMDE_V6.2.pdf, 2004.

[12] J. Carletta, "Assessing agreement on annotation tasks: The kappa statistic," *Computational Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.