Techniques for transcribers: assessing and improving consistency in transcripts of spoken language

Peter Garrard
Stroke and Dementia Research Centre, St George's, University of London, London, UK

Anne-Marie Haigh and Celeste de Jager
Nuffield Department of Medicine, University of Oxford, Oxford, UK

Abstract

Spoken discourse is a uniquely valuable source of data in cognitive research. A natural way of representing spoken discourse is in the form of a transcript in standard orthography. However, since transcribing is, for neuroscientists at any rate, no more than a means to an end, many researchers give only cursory descriptions of the transcription process, including the assessment of agreement between transcribers. This article introduces a novel approach to the systematic assessment of agreement between transcripts. The method first involves the automated alignment of two texts, followed by the automatic identification and quantification of discrepancies. A similarity score is then computed, providing researchers with a tool to evaluate the accuracy of the pair of transcripts in question. Most importantly, the automated production of a set of comparison tables reveals and summarizes the actual mismatches found, making it possible to identify common causes of discrepancy. Through applying this approach to medical data collected for an investigation of dementia, the present study demonstrates its value in the amendment of transcripts and the improvement of transcription practices, which pave the way towards more reliable transcriptions for research purposes.

Correspondence:
Peter Garrard, Stroke and Dementia Research Centre, St George's, University of London, Cranmer Terrace, London SW17 0RE, UK
E-mail:
pgarrard@sgul.ac.uk

1 Introduction

Despite being a natural, seemingly effortless, everyday human activity verbal communication is a highly complex process, drawing on a wide range of cognitive abilities. Short-term 'working' memory, knowledge of phonological structure, grammatical convention, and word meaning are among the fundamental requirements (Garrard, 2008; Price, 2010). Understanding the language system in sufficient detail not only to enumerate its functional components, but also to explain how these map on to physical (i.e. neuronal) activity in the brain, has been a long-standing aim of cognitive neuroscience.

Literary and Linguistic Computing, Vol. 26, No. 4, 2011.

doi:10.1093/llc/fqr018 Advance Access published on 24 May 2011

Three types of methodological approach are currently used to test hypotheses arising from theoretical models: recordings of cognitive event-related potentials (ERPs), which reveal the distribution of changes in electrical potential (corresponding to simultaneous regional activity) over the brain's surface during specific cognitive tasks; functional neuroimaging, which is used to delineate brain regions where blood oxygen consumption (and hence neural activity) changes in response to a well-defined cognitive task; and finally, the study of the patterns of cognitive difficulty shown by individuals who have suffered some form of brain injury, showing how cognitive abilities (such as linguistic communication) fractionate into more basic subcomponents—an approach referred to as neuropsychology.

Although much of our understanding of the organization of the linguistic system has come from the study of patients with aphasic deficits due to focal brain damage (e.g. following a stroke), interesting neurolinguistic data have also been obtained from patients whose brain damage is secondary to progressive neurodegenerative dementia such as Alzheimer's disease (AD) and frontotemporal dementia (FTD) (a rarer, more focal, cause of brain degeneration). The cognitive profiles of individuals with these conditions suggest that knowledge of phonology, grammatical role, and meanings of individual words may all be independently disrupted. The experimental paradigms that are employed to demonstrate these problems are generally based on single words, whether as stimuli, responses (e.g. in picture naming tests), or both (Kertesz, 1986; Neary *et al.*, 2005). Such tasks, however, despite being both robust and straightforward to administer and score, are somewhat lacking in everyday validity. More recently it has been recognized that samples of connected discourse provide rich sources of insight into linguistic cognition in both the intact and damaged brain (Chi, 1997; Ash *et al.*, 2006; 2009; Meteyard and Patterson, 2009; Garrard and Forsyth, 2010).

There are a number of different approaches to the analysis of samples of connected language: quantification and classification of errors; comparisons of lexical variables such as frequency and concreteness; measures of syntactic complexity, narrative coherence, and semantic richness; latent semantic analysis (LSA); and principal components analysis of word type occurrences across groups of speakers. All of these methods require, as a first step, the production of a written version, in a standard orthography, of the text to be analysed. Although this might seem like a trivial procedure the reality is that transcription is fraught with problems relating to accuracy, consistency, and standardization. For English alone there are numerous schemes for transcribing episodes of individual and conversational connected speech (Atkinson and Heritage, 1984; Boden and Zimmermann, 1991; Schiffrin, 1994; Leech *et al.*, 1995; MacWhinney, 1995; Carter, 2004; Jefferson, 2004; MICASE, 2007; Forsyth *et al.*, 2008). Yet, even when a single set of conventions is adopted, it is common to see considerable variation between individual transcribers.

Why should this be so? One important factor is nicely summarized in the title of an article by Ochs (1979) 'Transcription as theory': in other words, there is simply no single objective method of transcribing speech, a point also emphasized by Cucchiarini (1996):

> . . . a transcription, whatever the type, is always the result of an analysis or classification of speech material. Far from being the reality itself, transcription is an abstraction from it. (p. 132).

In accepting that transcription is an 'act of interpretation' (Bucholtz, 2000, p. 1463), we acknowledge that multiple different interpretations are to be expected. An empirical study by Chiari (2007) has highlighted the extent to which a transcription is a representation of what is intended rather than what is actually said. This arises as a natural consequence of our normal mode of listening, in which we seek to understand what the speaker means rather than to record the form in which such meaning is expressed. As Chiari (2007) argues:

> . . . even with the best possible audio quality, when trying to concentrate attention on the reconstruction of linguistic form, we tend to shift and rely on our understanding strategies,

> that lead us to re-create text in a plausible way. (p. 10).

Despite widespread recognition that the process of transcription is problematic, transcribing is not a prime focus of concern for most researchers, and it is not surprising that the transcription process itself is seldom described in great depth or detail. The following extracts from the Methodology sections of a selection of papers that use transcripts as data amply illustrate this point:

> The interviews were tape recorded and transcribed for analysis. (Martinson *et al.*, 1993).
> The recordings of each participant were transcribed verbatim. (Hux *et al.*, 2008)
> Following transcription, each memory was segmented into informational bits or details. (McKinnon *et al.*, 2008)

It appears that practice among researchers varies even on important questions such as: (i) the number of listeners (one versus more than one) who produce independent transcriptions of a recording; (ii) how multiple transcripts are used—e.g. for cross-checking or for comparing content; (iii) how such cross-checks or comparisons are carried out; and (iv) how any differences or divergences that are identified between transcripts are resolved. The principal aim of the present paper is to bring these important issues into focus. We anticipate that our observations and methods will be of particular interest to all the researchers who use spoken language as research data, and hope that the guidelines we propose for mitigating the problems associated with transcribing, and the computational techniques we describe for measuring and maximizing adherence to those guidelines, will help to enhance the consistency and accuracy of their data.

In pursuit of this overall objective, we propose a set of techniques that can assist in the process of arriving at a transcription that is adequate for the purposes in hand by quantifying the discrepancy between different transcripts of the same recording on an objective basis. We exemplify and test these techniques by applying them to a series of spoken interactions collected as part of an extensive longitudinal study of dementia and cognitive ageing: the Oxford Project to Investigate Memory and Ageing (OPTIMA) http://www.medsci.ox.ac.uk/optima.

2 Methods

2.1 Overall procedures

The techniques we propose are embedded within a set of overall procedures aimed at transforming digitized sound recordings into reliable written records of linguistic events. For the sake of clarity, these procedures are now summarized:

1 Capture discourse from study participants with the best affordable audio equipment. [We do not consider the additional complications introduced by video recording, for which see Carter and Adolphs (2008); Gorgos (2009).]
2 Establish and put in writing (or adapt from previously published guidelines) a preliminary set of transcription conventions.
3 Arrange for two transcribers independently to transcribe a sample of spoken interactions in the form of digital text files.
4 Align pairs of transcript files using a modified version of the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970).
5 Compute similarity scores using the *F*-measure (Weiss *et al.*, 2005) for benchmarking purposes.
6 Produce individual mismatch tables for amendment of transcripts.
7 Produce an aggregate discrepancy table to suggest improvements in transcription practices and/or amendment(s) of transcription conventions. The conventions developed by us using Steps 2–7 are reproduced in Appendix A1.
8 Transcribe the full data set using the practices and conventions agreed in Step 7.

According to this procedure, analysis of the linguistic attributes of the transcripts and their relationships with other information (about the speakers and the context of their interactions) only begins after completion of the eight steps outlined above, the purpose of which is to ensure, as far as possible, that the data derived from analysing the transcripts will bear the evidential weight that is likely to be placed upon it.

2.2 Alignment and comparison of transcripts

At the heart of our method is an automated technique for assessing the agreement between a pair of transcripts. Quantifying the similarity between two symbol sequences, however, is not a simple matter: some authors use phrases such as 'reliability at 80% or greater' (Williams *et al.*, 2003) without specifying how such figures are computed, despite the fact that a single mismatch at any point can throw both sequences out of alignment, resulting in close to 0% agreement between the subsequent symbols.

When authors quote a percentage agreement or an index such as kappa (Cohen, 1960), this should imply that the two symbol streams are aligned, whether by human inspection or using some form of automated procedure, such as the 'compare-and-merge-documents' facility in Microsoft Word, which can be used to obtain an optimal alignment of each document pair, prior to calculating a similarity score.

There are limits to the usefulness of this approach, however, because as soon as the number of mismatches rises above a certain level, the software's automatic highlighting becomes confusing. This difficulty is illustrated in Fig. 1, which shows an extract from a parliamentary interchange involving the Minister of State for the Cabinet Office on 1 July 2009. In this extract, word is comparing the official parliamentary report (Hansard, Vol. 495 Col. 283, 1 July 2009), with a transcription from audio recording following the conventions described in Appendix A1.

2.2.1 *Text alignment*

The problems associated with Word's built-in alignment software were overcome using a special-purpose text alignment programme, written in PYTHON 2.6. The programme implements a variant of the Needleman–Wunsch algorithm (1970), adapted to work on tokens rather than individual characters. The Needleman–Wunsch algorithm is commonly used to align DNA or amino acid sequences, though Sankoff and Kruskal (1983) describe several other applications of the family of dynamic-programming algorithms to which it belongs. A dynamic programme works in two phases: first it breaks a larger problem into a series of subproblems, saving the optimal solution found to each in a matrix (forward

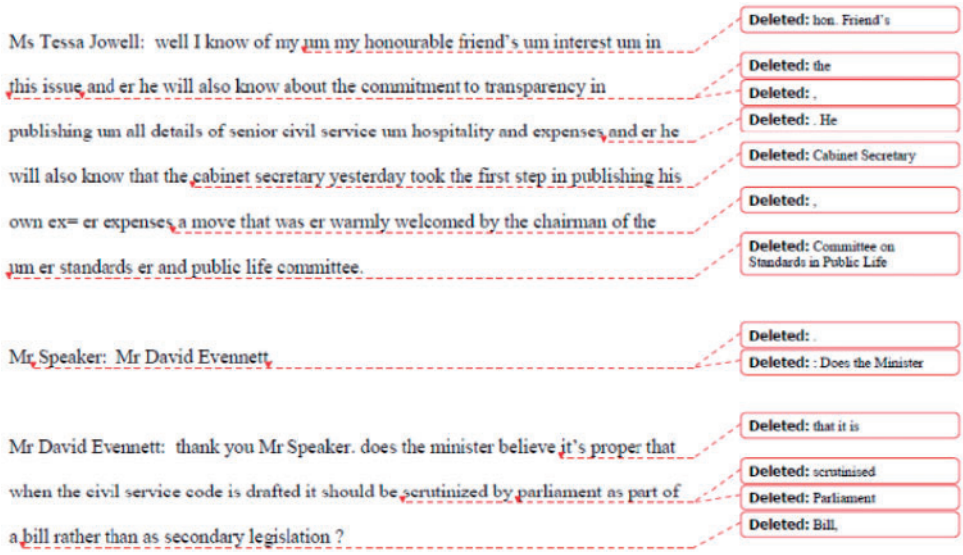

Fig. 1 Comparison of a verbatim transcript of a broadcast debate from the House of Commons with the official record of the debate published in Hansard, using the 'compare and merge documents' facility in Microsoft Word

Table 1 Types of match and mismatch recognized between pairs of aligned transcripts, together with defining criteria and illustrative examples

Token comparison result (code)	Definition	Score
Long match (LM)	An alignment of identical tokens of four or more characters	4
Short match (SM)	An alignment of identical tokens of less than four characters	3
Long partial match (LP)	An alignment between two tokens, one of which is a four-or-more- character-long substring of the other	2
Short partial match (SP)	An alignment between two tokens, one of which is a less-than-four-character-long substring of the other	0
Short mismatch (SX)	An alignment between two non-identical tokens, the shorter of which is less than four characters long	0
Short gap[a] (SG)	If a null token is aligned with a token consisting of four or more characters	0
Long gap (LG)	If a null token is aligned with a token consisting of less than four characters	−1
Long mismatch (LX)	An alignment between two non-identical tokens, the shorter of which is four or more characters long	−1

[a]A 'gap' occurs when a null token is introduced into either sequence to achieve an optimal alignment score.

phase); it then selects the solution that provided the optimal overall cost (backward phase). A technical description of the algorithm on which present programme is based can be found in Cannarozzi (2005).

Most algorithms of this sort use characters as the atomic units of comparison, but ours works with tokens as its base units, which entails a pre-processing stage in which texts are broken into sequences of tokens (mostly words, but also sometimes punctuation symbols). The programme also gives the option of folding all letters into lower case, an option that is used in all the examples quoted below.

Once the texts have been turned into token sequences, the alignment algorithm equalizes the lengths of the two sequences (if they are of unequal length) by inserting null strings into the shorter sequence. It then finds the maximum value for a payoff/penalty function (which assigns a cost or reward to the various kinds of match and mismatch that can occur). The types of match and mismatch that are recognized, the defining features of each, and the payoff/penalty scores assigned to them, are shown in Table 1. It will be noted that the programme recognizes eight distinct relationships and that it assigns scores at five levels, from minus one to plus four, which are used as heuristics to guide it towards an optimal alignment solution.

Table 2 Alignment between two sample texts showing the classification of the token match or mismatch at each point

Match or mismatch type	Token number	Text 1	Text 2
SG	0	[00]	in
SM	1	the	the
LG	2	ancient	[00]
LG	3	cretan	[00]
LX	4	palace	minoan
LM	5	ruins	ruins
SM	6	of	of
LM	7	knossos	knossos
SX	8	constitute	is
SP	9	a	an
LP	10	historical	historic
LM	11	treasure	treasure
LM	12	trove	trove
SM	13	.	.

To illustrate, the result of applying the algorithm to the text strings:

> The ancient Cretan palace ruins of Knossos constitute a historical treasure trove (Text 1)

and

> In the Minoan ruins of Knossos is an historic treasure trove (Text 2)

is shown in Table 2.

The output consists of the two texts aligned vertically with one another, together with the

classification of the match / mismatch between each token pair. All eight recognized match types described in Table 1 are represented. The string '[00]' is a place-holder used to indicate points at which the programme introduced a gap during the alignment process.

2.2.2 *Quanitfying similarity in aligned texts*

Once the two sequences have been aligned, the discrepancies (deletions, insertions, and substitutions) can be counted and used to obtain a variety of similarity scores. We have experimented with a number of agreement indices, of which the *F*1 measure (McCowan *et al.*, 2005), a metric based on notions of precision and recall as used in information retrieval (IR), seems most appropriate as a general index of agreement between symbol sequences.

The index ultimately derives from the work of van Rijsbergen (1979) and is defined as

$$F1 = \frac{(2 \times p \times r)}{(p + r)}.$$

*F*1 is thus the harmonic mean of p (precision) and r (recall), with precision and recall weighted equally, and yields a value in the range from 0 to 1, where 0 indicates complete difference and 1 means complete identity. Considered in the context of IR (for example, a search for documents in a repository such as the World Wide Web), precision is defined as the proportion of the retrieved entries that are relevant to the needs of the user, while recall is defined as the proportion of all the relevant entries that are retrieved. In most databases, the number of items in the archive will greatly exceed the number of relevant items. In the context of a comparison between two independent transcripts, however, this asymmetry disappears, since neither sequence can be regarded as a 'gold standard', so p and r become, in effect, interchangeable. Thus:

$$p = \frac{h}{n}$$

and

$$r = \frac{h}{m}$$

where h is the number of matching tokens (hits), and n and m are the lengths (number of tokens) in the first and second token sequences. In the small-scale example given in Table 2, n is 13, m is 12, and h (the number of matches) is 7. Using the above formulae:

$$\begin{aligned} p &= \frac{7}{13} \\ r &= \frac{7}{12} \\ F1 &= \left(2 \times \frac{7}{13} \times \frac{7}{12}\right) / \left(\frac{7}{13} + \frac{7}{12}\right) \\ &= \frac{98}{175} \\ &= 0.56. \end{aligned}$$

It should be noted that this measure is based on all-or-nothing matching: either a pair of tokens is the same or it is different. As noted above, the various levels of match shown in Table 1 are used to guide the alignment process; once aligned, the F1 measure is computed using a dichotomous distinction between matching and non-matching tokens.

2.3 Transcription data

Data used in the study came from a large set of recorded interviews with a group of volunteers in a long-term study of the effects of ageing on cognitive function and the incidence of AD—the OPTIMA project (see above). Participants undergo clinical and cognitive evaluations at roughly 6-monthly intervals, over a number of years. At entry, some of the participants have been diagnosed with AD or other types of dementia, while other study entrants enjoy normal cognitive function. Among the latter, serial evaluation will disclose either stability or later cognitive decline. Among the tasks participants are asked to perform is the cookie theft picture description test [a component of the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983)]. In this test, participants are shown a line drawing of a kitchen scene in which various events are occurring: a daydreaming housewife is allowing a sink to overflow, while a boy falls off a stool as, egged on by his sister, he reaches into a jar labelled 'cookies' behind her back. Participants are shown the picture and simply asked to 'describe what is going on'.

From the database of participants who had been interviewed at least five times we selected five

subjects. One was male (M1) and four female (F1–F4). The ages of the participants at the time of the first and fifth dialogue selected for the present analysis are given in Table 3. For each participant, the portions of their first five interviews dealing with the cookie theft description were transcribed independently by two different transcribers, giving twenty-five pairs of transcripts with which to investigate transcriber agreement.[1]

Although the cookie theft task is designed to evoke a monologue by the participant, in practice the interviewer often intervenes with comments, prompts or 'back-channel' signals (Yngve, 1970), producing a dialogue. As our interest extends to dialogue, we took the decision to take as the text to be processed all speech between the first and last recorded utterance by the volunteer, including the interviewer's utterances within that stretch of talk. A sample of this data—after processing by the alignment programme—is provided in Appendix A2. This is the shortest of the twenty-five transcripts referred to above.

Table 3 Ages of participants whose descriptions were used in the study

Participant	Age at first transcribed episode	Age at fifth transcribed episode
M1	44	48
F1	53	63
F2	59	65
F3	64	69
F4	79	84

Since both transcribers were attempting to record the same speech events, full agreement is obviously desirable, but there is no established baseline that would specify how much discrepancy can be tolerated. To provide an empirical benchmark, we therefore collected a set of text pairs, of which each member was, in some sense, a replication of the other. These, along with the cookie theft descriptions, are detailed in Table 4.

The selection of texts is entirely exploratory: there are no strong theoretical reasons to predict what range of scores might be expected in each category. What can be said in advance of testing, however, is that the three categories of what might be termed semantic equivalence (Biblical versions, Poetry translation, and Sports press) would be expected to show to a high degree the phenomenon mentioned in the Introduction section—to wit, the tendency to replicate meaning rather than form. This tendency should be less evident in Parliamentary speeches, and less so still in the categories (Letters and Cookie theft transcriptions) where faithful reproduction of the form as well as the content of the discourse was a high priority. As

Table 4 Text pairs used for comparison, including the data of primary interest to this study

Category	Description
Biblical versions	Four extracts from the Bible (Genesis 1, Joshua 8, Psalm 101 and Romans 1) in the English Standard Version (ESV) and the New International Version (NIV).
Parliamentary speeches	Three short extracts from unscripted parliamentary exchanges in the House of Commons: one transcribed from audio and the other reproduced from the official record (Hansard).
Letters	Five letters handwritten by the novelist and philosopher Iris Murdoch to a friend, independently transcribed from manuscript by two transcribers.
Cookie theft descriptions	Twenty-five independently transcribed cookie theft description dialogues by OPTIMA participants (see text).
Plagiarism	Three student essays and the texts which they were later found to have plagiarized.
Poetry translation	Pairs of translations into English of poems—three from Hungarian and one from Chinese.
Sports press	Pairs of reports of three sporting events (Rugby League: England versus Australia, 1 November 2009; Boxing: Haye versus Valuev, 8 November 2009; Football: Brazil versus England, 15 November 2009) published in two different British Sunday newspapers (The Independent on Sunday and The Sunday Times)
OCR output	Three articles from The Times online archive (one from 1912 and two from 1917), one version transcribed manually from the page images and the other using optical character recognition (OCR) software.

for the examples of Plagiarism and OCR output, these can vary depending on numerous factors, such as the skill of the plagiarist in disguising his or her source, the font and quality of the texts subject to OCR, and the sophistication of the OCR procedure. This exercise thus provides a set of baselines against which to judge the performance of the similarity scoring algorithm as applied to English texts.

3 Results

3.1 Similarity scoring

The *F*1 scores of the twenty-five pairs of OPTIMA transcripts ranged from 0.6842 to 0.9787 with a mean of 0.87, a median of 0.87 and a standard deviation of 0.06. As it is rare for transcription agreement to be quoted in this manner, the same measures were computed for the other text groupings listed in Table 3, above, in order to provide a context for interpreting these similarity scores. Figure 2 displays a boxplot of the eight sets of results. Inspection of the figure confirms that the mean *F*1 measure for the cookie theft transcripts was the highest of the eight groups.

As expected, reporters working for different newspapers who write about the same sporting event have the lowest level of formal similarity, less even than free poetic translations, and much less than plagiarized material. Of the two groups that compared different translations of the same text, the biblical translations were noticeably more concordant than the poems. A plausible *post hoc* explanation for this is that translators of sacred scripture have to pay very close attention to the literal meaning of the original text and to the wording of earlier authoritative translations, while translators of secular poetry are free to rephrase the gist of the original in their own words, indeed are encouraged to use the original poem as a launch pad for their own creative efforts.

One of the most pertinent contrasts is between the cookie theft data and the House of Commons

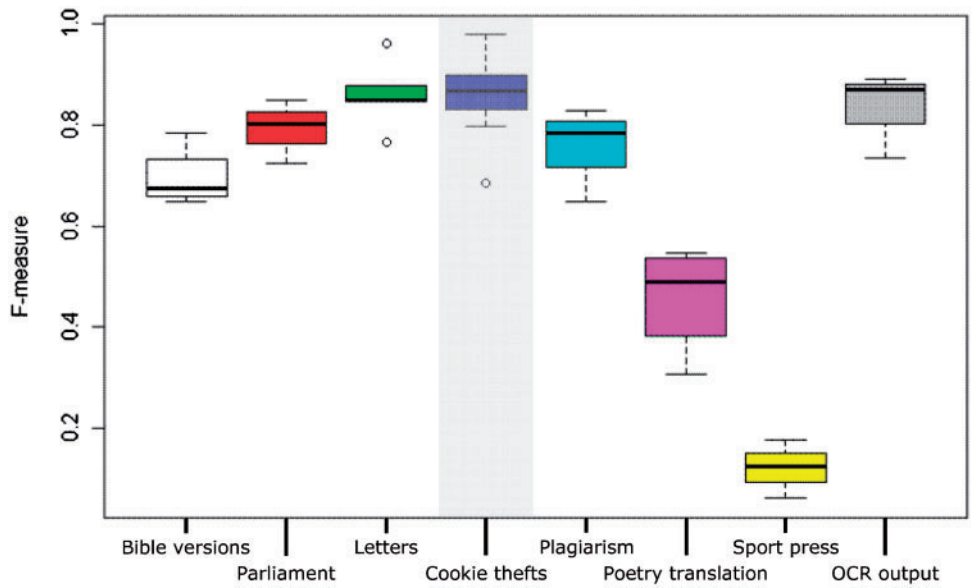

Fig. 2 Parallel boxplots (median, upper and lower quartiles, and range ± outliers) of the similarity measure (F1) between eight types of text pair, in which each member of a pair is some form of replication of the other. The highest similarity values are associated with independent transcriptions of verbal (cookie theft descriptions) and written (letters of Iris Murdoch) material; the lowest with descriptions of the same sporting event by two different sports journalists

transcripts. In both, two independent transcribers have produced written versions of the same dialogue and the *F*1 scores are similar, though not as close of those associated with handwritten letters and optically recognized news stories, which fall within the lower to upper quartile range of the OPTIMA transcripts. For all the other groups (including the House of Commons transcripts), the upper quartile has a lower score than the lower quartile of the OPTIMA transcripts. A Wilcoxon signed-rank test confirms that the OPTIMA and Commons subgroups differ significantly ($W = 65$, $P = 0.0409$, two-tailed) with respect to *F*1 scores. Whether this first-pass level of agreement is satisfactory depends on the uses to which the data will be put as well as the attributes of the transcribed language extracted for analysis, but it is encouraging that our transcripts, according to this measure, represent a more nearly verbatim record than official parliamentary proceedings.

Still, a higher overall degree of similarity than the Hansard transcripts may not be a level to be content with, as it is well-known that verbal exchanges in parliament are subject to systematic 'rectifications' by the Hansard editorial staff, including but not limited to, the deletion of non-lexical fillers (such as 'er' and 'um'), repetitions (such as 'I I') and uncompleted phrases, and the replacement of 'have to' by 'must' and phrasal verbs by single lexical items (e.g. 'look at' by 'consider' or 'examine') (Mollin, 2007).

The purpose of the text-comparison software is not merely to assess similarity but also to indicate how it might be enhanced, and the value of a similarity scoring system such as this is that it can identify and quantify commonly occurring inconsistencies among transcribers. Studying such inconsistencies allows us to take a further step towards improving transcription practice.

3.2 Identifying mismatches

To facilitate correction of errors in individual cases, a programme was written (again in PYTHON 2.6) that took output from the alignment programme and produced, for each transcript-pair, a table of mismatches, to enable reference back to the episodes in which they occured. A total of 342 mismatches were found across the full set of twenty-five pairs of transcriptions. An example of the programme's output relating to a single pair of transcriptions is shown in Table 5. The two transcript sources are labelled 1 and 2 (though neither is considered to be the 'preferred' version at this stage). The aim of tabulation is to provide an initial focus for later attempts at improving consistency.

Table 5 Sample mismatch table for a single pair of cookie theft transcripts after optimal alignment

Transcriber 1	Transcriber 2
..	[00]
[?well]	Well
n:	[00]
mhhm	[00]
p:	[00]
n:	[00]
mhhm	Um
p:	[00]
er	[00]
[00]	Going
gonna	To
taking	[00]
out	Taken
of	From
..	[00]
and	[00]
..	Um
drying	Dyring
er	[00]
which	Um

As in Table 2, '[00]' indicates a point at which optimal alignment necessitated the introduction of a gap.

The programme can also be used to aggregate discrepancies of the same type and produce a table of such discrepancies in descending order of frequency. The combined table for the twenty-five transcripts examined here is displayed in Table 6 (with discrepancies that occur only once omitted to save space).

The first three columns of the table contain output from the programme. The last column, which is labelled 'origin of mismatch', contains a presumed interpretation of the mismatch. This means that a human reader can consider whether a common cause is likely to give rise to a number of mismatches of a certain type. Sorting the table by

Table 6 Table of discrepancy types occurring between all twenty-five cookie theft transcript pairs, ordered by number of occurrences (*n*)

Transcriber 1	Transcriber 2	*n*	Origin of mismatch
..	[00]	29	Pause recorded by T1 only
[00]	is	10	Contraction expanded, T2 only
and	[00]	7	Insertion by T1
n:	[00]	7	Speaker change, T1 only
the	[00]	7	Insertion by T1
a	the	6	Word mismatch
er	[00]	6	Filler, T1 only
p:	[00]	6	Speaker change, T1 only
um	[00]	6	Filler, T1 only
[00]	boy	5	Contraction expanded, T2 only
[00]	day	5	Insertion by T2
[00]	the	5	Insertion by T2
um	hm	5	Filler mismatch
[00]	n:	4	Speaker change, T2 only
boy's	is	4	Contraction expanded, T2 only
the	a	4	Word mismatch
um	and	4	Filler-word mismatch
[00]	a	3	Insertion by T2
[00]	p:	3	Speaker change, T2 only
falling	fallen	3	Word mismatch
i	[00]	3	Insertion by T1
?	[00]	2	Punctuation, T1 only
[#]	[00]	2	Unintelligible to T1 only
[00]	.	2	Punctuation, T2 only
[00]	..	2	Pause recorded, T2 only
[00]	[=laugh]	2	Non-verbal vocalization, T2 only
[00]	and	2	Insertion by T2
[00]	er	2	Filler, T2 only
[00]	for	2	Insertion by T2
[00]	me	2	Insertion by T2
all	[00]	2	Insertion by T1
and	um	2	Word-filler mismatch
daydreaming	dreaming	2	Word break/hyphenation
huh	[00]	2	Filler, T1 only
huh	ha	2	Interjection mismatch
in	[00]	2	Insertion by T1
kids	kid's	2	Word mismatch
mm	[00]	2	Filler, T1 only
on	[00]	2	Insertion by T1
stool's	stool	2	Contraction expanded, T2 only
that's	[00]	2	Insertion by T1
well	[00]	2	Insertion by T1
window's	window	2	Word mismatch
You	[00]	2	Insertion by T1

entries in this column allows the common causes of discrepancy to be shown together (Table 7). It will be seen that this table now includes a new column containing potential remedies that could remove some types of mismatch or mitigate their effects. By considering the frequencies associated with each potential remedy, 'low-hanging fruit' (i.e. simple changes with non-trivial effects) can be identified.

For example, if we were to ignore pauses, the table shows that we would avoid almost 10% of the mismatches [twenty-nine cases where Transcriber 1 has recorded a pause ('..') while Transcriber 2 has not, plus the two cases where the reverse is true]. It is also of passing interest to note that when the presence or absence of a token depends on a subjective judgement, it is inevitable that transcribers will have different thresholds; it is at least reassuring that Transcriber 1's lower threshold for recording pauses appears to be consistent! Fillers are another common source of discrepancy, and if unified (i.e. all treated as equivalent), then only five mismatches would be saved. In contrast, if fillers were ignored altogether, a further twenty-three mismatches would be eliminated. Importantly, both of these changes could be achieved by post-processing, without the need for re-transcription. Further reflections on the implications of such results are presented in the following section.

4 Discussion

In this article, we have outlined an approach to transcription, aimed at mitigating one of the main problems of transcribing, namely disagreement between transcribers. While acknowledging that agreement does not guarantee accuracy, we would argue that disagreement between transcribers is usually a sign that some aspect of the transcription process requires re-examination. We have therefore outlined an approach that enables classes of discrepancy to be identified automatically and systematically, and presented computational techniques for performing such automatic identifications. We have implemented these techniques in a suite of software modules and subjected them to an initial trial by applying them to real-world data from an important area of medical research.

Table 7 Discrepancy table with suggested remedies

Transcriber 1	Transcriber 2	*n*	Origin of mismatch	Potential remedy
stool's	stool	2	Contraction expanded, T2 only	Expand contracted 'is'
boy's	is	4	Contraction expanded, T2 only	Expand contracted 'is'
[00]	boy	5	Contraction expanded, T2 only	Expand contracted 'is'
[00]	is	10	Contraction expanded, T2 only	Expand contracted 'is'
Um	hm	5	Filler mismatch	Ignore/unify fillers
Um	and	4	Filler-word mismatch	
Huh	[00]	2	Filler, T1 only	Ignore/unify fillers
Mm	[00]	2	Filler, T1 only	Ignore/unify fillers
Er	[00]	6	Filler, T1 only	Ignore/unify fillers
um	[00]	6	Filler, T1 only	Ignore/unify fillers
[00]	er	2	Filler, T2 only	Ignore/unify fillers
all	[00]	2	Insertion by T1	
in	[00]	2	Insertion by T1	
on	[00]	2	Insertion by T1	
that's	[00]	2	Insertion by T1	
well	[00]	2	Insertion by T1	
you	[00]	2	Insertion by T1	
i	[00]	3	Insertion by T1	
and	[00]	7	Insertion by T1	
the	[00]	7	Insertion by T1	
[00]	me	2	Insertion by T2	
[00]	for	2	Insertion by T2	
[00]	and	2	Insertion by T2	
[00]	a	3	Insertion by T2	
[00]	the	5	Insertion by T2	
[00]	day	5	Insertion by T2	
huh	ha	2	Interjection mismatch	Ignore/unify interjections
[00]	[=laugh]	2	Non-verbal vocalization, T2 only	
..	[00]	29	Pause recorded, T1 only	Ignore pauses
[00]	..	2	Pause recorded, T2 only	Ignore pauses
?	[00]	2	Punctuation, T1 only	Remove punctuation
[00]	.	2	Punctuation, T2 only	Remove punctuation
p:	[00]	6	Speaker change, T1 only	Analyse only p's text
n:	[00]	7	Speaker change, T1 only	Analyse only p's text
[00]	p:	3	Speaker change, T2 only	Analyse only p's text
[00]	n:	4	Speaker change, T2 only	Analyse only p's text
[#]	[00]	2	Unintelligible, T1 only	
daydreaming	dreaming	2	word break/hyphenation	
and	um	2	Word-filler mismatch	
kids	kid's	2	Word mismatch	
window's	window	2	Word mismatch	
falling	fallen	3	Word mismatch	
the	a	4	Word mismatch	
a	the	6	Word mismatch	

Using a novel method of text alignment and quantification of agreement, the level of similarity between cookie theft transcribers in the samples used in this study would appear to be acceptable: the degree of difference between transcripts is overall less than that between Hansard and a verbatim transcription. This suggests that the transcription conventions that we outline in Appendix A1 provide for an acceptable degree of reproducibility. It is also worth remarking that there were very few

misspellings or typographical errors ($n = 3$), so a spell-check would not significantly improve our level of agreement.

Nevertheless, there remain aspects of our transcription practices which are clearly less than ideal, particularly the recording of paralinguistic features. In particular: (i) pauses were not recorded consistently; (ii) coding of fillers (such as 'er' and 'um') was not consistent between transcribers; (iii) coding of interjections (such as 'ha' and 'huh') was also inconsistent between transcribers; (iv) there was disagreement about what constituted a backchannel signal (e.g. 'mhhm'); (v) other non-verbal vocalizations are not recorded consistently (e.g. '[=laugh]'); and (vi) some changes of speaker are indicated by only one transcriber.

It appears, therefore, that the recording and representation of various types of paralinguistic feature in transcription is somewhat idiosyncratic, and thus unreliable, suggesting that they should be removed in the interests of consistency. This is not to imply that such data are not useful markers; indeed Arciuli *et al.* (2010) showed that, in a laboratory setting, the frequency of the filler 'um' was the most important discriminator between statements made when subjects were instructed to tell the truth and those made when they were instructed to lie. If similar importance can be shown to attach to such items in the context of normal versus pathological age-associated language change, then clearly this decision would need to be revised, and more stringent guidelines for their representation included in the transcription guidelines instead.

We started out with the aim of recording the dialogue between a patient (p:) and nurse (n:), that almost always results from the performance of the cookie theft task. It is worth noting in this connection that in two of the three episodes with the lowest similarity scores, these low scores resulted from a final non-verbal utterance of the patient ('mm' and '[=laugh]', respectively) being omitted by one or other transcriber. Thus, in both cases an unmatched utterance by the nurse was included in one transcript but not the other. (Transcript 137BS_01/07/96 is reproduced in Appendix A2.) This side-effect of imperfect agreement on non-verbal items, leading to a disagreement about where the transcript ends, would have minimal effect on any analysis that confined itself only to the utterances of the patient, which in fact are our main subject of concern.

As far as the more lexical side of the data is concerned, our transcripts agree quite well, though two areas require specific attention: (i) contractions and genitives, such as ''boy's'' versus 'boy is', 'gonna' versus 'going to', ''sink's'' versus 'sinks' and ''you've'' versus 'you have'; and (ii) hyphenation and word-separation, such as 'all right' versus 'alright', 'overrunning' versus 'over-running' or 'thank you' versus 'thankyou'. There is no a simple practical way of standardizing these discrepancies *post hoc*, though one could do a guided search for places where they might be found. Point (ii) could also be addressed by expanding all enclitic versions of 'be' and 'have', such that ''boy's'' is rendered as 'boy is' or 'boy has' (except when context indicates a possessive). Such a restriction would, however, move us further towards the Hansard-style approach of recording what we think the speakers ought to have said and away from the ideal of verbatim transcription.

The central aim of these techniques was to provide information in a consistent fashion which will assist human reviewers to reflect on their transcription practices in a rational manner, ideally before making an expensive commitment to a large-scale transcription exercise. Such reviewing may lead to revisions that improve agreement and thereby increase the reliability of the transcripts. But revising the transcripts may not be the only outcome. Other reasonable responses might include adapting proposed transcription practices, amending the guidelines for transcribers, or avoiding analyses based on phenomena that have not been reliably recorded.

We do not claim to have solved the problems of transcription, nor do we offer a rigid prescription for all kinds of transcribing. We have, however, made explicit the stages required by a systematic approach to obtaining greater reliability in this richly informative source of language data.

Funding

This work was supported by a UK Medical Research Council research grant (grant number G0801370) to P.G. and C.deJ.

Acknowledgements

Thanks are due to Dr R. Forsyth for writing the PYTHON scripts, and to the University of Southampton Medical School, where much of this research was carried out.

References

Arciuli, J., Mallard, D., and Villar, G. (2010). Um, I can tell you're lying: Linguistic markers of deception versus truth-telling in speech. *Applied Psycholinguistics*, **31**: 397–411.

Ash, S., Moore, P., Antani, S., McCawley, G., Work, M., and Grossman, M. (2006). Trying to tell a tale: Discourse impairments in progressive aphasia and frontotemporal dementia. *Neurology*, **66**: 1405–13.

Ash, S., Moore, P., Vesely, L. *et al.* (2009). Non-fluent speech in frontotemporal lobar degeneration. *Journal of Neurolinguistics*, **22**: 370–83.

Atkinson, J. M. and Heritage, J. (1984). *Structures of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press.

Boden, D. and Zimmerman, D. H. (1991). *Talk and Social Structure*. Cambridge: Polity Press.

Bucholtz, M. (2000). The politics of transcription. *Journal of Pragmatics*, **32**: 1439–65.

Cannarozzi, G. M. (2005). String alignment using dynamic programming. http://www.biorecipes.com/DynProgBasic/code.html (accessed 19 January 2010).

Carter, R. (2004). *Language and Creativity*. London: Routledge.

Carter, R. and Adolphs, S. (2008). Linking the verbal and visual: new directions for Corpus Linguistics. *Language and Computers*, **64**: 275–91.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: a practical guide. *The Journal of the Learning Sciences*, **6**: 271–315.

Chiari, I. (2007). Transcribing speech: errors in corpora and experimental settings. *Proceedings of Corpus Linguistics 2007, Birmingham University, July 2007*. Full text available at: http://www.corpus.bham.ac.uk/corplingproceedings07/paper/248_Paper.pdf (accessed 3 February 2008).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**: 37–46.

Cucchiarini, C. (1996). Assessing transcription agreement: methodological aspects. *Clinical Linguistics and Phonetics*, **10**: 131–55.

Forsyth, R. S., Clarke, D. D., and Lam, P. (2008). Timelines, talk and transcription: A chronometric approach to simultaneous speech. *International Journal of Corpus Linguistics*, **13**: 225–50.

Garrard, P. (2008). Cognitive Archaeology: uses, methods and results. *Journal of Neurolinguistics*, **22**: 250–65.

Garrard, P. and Forsyth, R. S. (2010). Abnormal discourse in semantic dementia: a data-driven approach. *Neurocase*, **16**: 520–8.

Goodglass, H. and Kaplan, E. (1983). *The Assessment of Aphasia and Related Disorders*. Philadelphia: Lea and Febiger.

Gorgos, K. A. (2009). Lost in transcription: Why the video record is actually verbatim. *Buffalo Law Review*, **57**: 1057–27.

Hux, K., Wallace, S., Evans, K., and Snell, J. (2008). Performing Cookie Theft content analyses to delineate cognitive-communication impairments. *Journal of Medical Speech-Language Pathology*, **16**: 83–102.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In Lerner, G.H. (ed.), *Conversation Analysis: Studies From the First Generation*. Amsterdam/Philadelphia: John Benjamins.

Kertesz, A., Appell, J., and Fisman, M. (1986). The dissolution of language in Alzheimer's disease. *Canadian Journal of Neurological Sciences*, **13**: 415–8.

Leech, G., Myers, G., and Thomas, J. (1995). *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman.

MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, NJ: Erlbaum.

Martinson, I. M., Chesla, C., and Muwaswes, M. (1993). Caregiving demands of patients with Alzheimer's Disease. *Journal of Community Health Nursing*, **10**: 225–32.

McCowan, I., Moore, D., Dines, J. *et al.* (2005). *On the Use of Information Retrieval Measures for Speech*

Recognition Evaluation. IDIAP Research Report, IDIAL-RR 04-73.

McKinnon, M. C., Nica, E. I., Sengdy, P. *et al.* (2008). Autobiographical memory and patterns of brain atrophy in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, **20**: 1839–53.

Meteyard, L. and Patterson, K. (2009). The relation between content and structure in language production: an analysis of speech errors in semantic dementia. *Brain and Language*, **110**: 121–34.

MICASE. (2007). The Michigan Corpus of Academic Spoken English. http://legacyweb.lsa.umich.edu/eli/micase/MICASE_MANUAL.pdf (accessed 27 December 2007).

Mollin, S. (2007). The Hansard hazard: gauging the accuracy of British parliamentary transcripts. *Corpora*, **2**: 187–210.

Neary, D., Snowden, J., and Mann, D. (2005). Frontotemporal dementia. *Lancet Neurology*, **4**: 771–80.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**: 443–53.

Ochs, E. (1979). Transcription as theory. In Ochs, E. and Schieffelin, B. B. (eds), *Developmental Pragmatics*. New York: Academic Press, pp. 43–72.

Price, C. J. (2010). The anatomy of language: A review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, **1191**: 62–88.

Sankoff, D. and Kruskal, J. B. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley Publishing Company.

Schiffrin, D. (1994). *Approaches to Discourse*. Oxford: Blackwell.

van Rijsbergen, C. J. (1979). *Information Retrieval*. London: Butterworths.

Weiss, S. M., Indurkhya, N., Zhang, T., and Damerau, F. J. (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer.

Williams, K., Holmes, F., Kemper, S., and Marquis, J. (2003). Written language clues to cognitive changes of aging: an analysis of the letters of King James VI/I. *Journal of Gerontology*, **58**: 42–4.

Yngve, V. H. (1970). On getting a word in edgewise. *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, Chicago*. Chicago: University of Chicago, pp. 567–78.

Note

1 Although a sample size of twenty-five transcript pairs is adequate to an exploratory analysis of the methods described here, we acknowledge that they originate from a small number of individuals, and that this limits the range of production styles and variety of challenges against which the methods are actually tested. Although our contention is that these methods would lead to better transcriptions, however challenging an individual's speech, confirmation will depend on their successful application in other samples.

Appendix A1: Transcription conventions (Version 1.1, November 2009)

The aim of this document is to clarify the conventions we would like to be used in transcribing the data. Please try to stick with the same convention consistently throughout the transcript, and between different transcripts. (Please contact the authors if you have any questions.)

In general, standard British spellings, as in a reputable dictionary, are to be used, even when words are pronounced unclearly or in a non-standard accent. We do not require indications of loudness or tone. Thanks for your help!

Topic	Guidelines	Example(s)
Abbreviations, acronyms, etc.	Acronyms are written in upper case, without dots. So are standard abbreviations such as CIA and MI5. (But see also: Spoken letters, below.) The following six abbreviations are exceptions (mixed case, with no hyphens or dots): Dr Miss Mr Mrs Ms PhD	BBC2 CANCODE NATO USSR
Backchannels	Any affirmations, agreements with statements etc such as 'uh huh', 'mhm' are called backchannels. As far as possible, for ease of analysis, all positive (agreeing, supportive) backchannel vocalizations should be coded: mhhm, uhha and all negative or (disagreeing or doubting) vocalizations should be coded: uhoh. Hums, ums, and other pauses in normal speech are NOT backchannels. (See: Fillers.)	mhhm yeah uhoh
Capitalization	Only proper names and acronyms are capitalized. (The first-person pronoun 'I' may be capitalized if doing otherwise causes delay, though this is optional.) When a name mentioned by a speaker has to be anonymized to preserve confidentiality, it should be written as capital initial followed by an underscore, e.g. X_, Y_ Z_. Optionally, the first word of a speaker's turn may be capitalized. (See also: Speaker Turns.)	Argentina Beijing Dr Hu The London Eye
Contractions	All standard contractions of the words is, am, are, have, had, would, not are represented with apostrophes as in conventional writing. Only the following six non-standard forms are accepted: gonna, gotta, innit, wanna, yeah, yep. Other similar contractions should be regularized, e.g. 'coulda' becomes 'could've', 'kinda' becomes 'kind of', 'salt n pepper' becomes 'salt and pepper', 'y'know' becomes 'you know'.	Can't, don't, i'll, it's, we're I'm gonna get there in a minute It's sort of, you know, round at the edges She'd agree if you asked nicely
Exclamations and interjections	Try to stick only to the following forms: ah, aha, ha, hi, eh, hey, huh, oh, ooh, oops, ouch, oy, tsch.	huh [=laugh]
Fillers	There are no recognized standards for filled pauses. Try to stick to the following forms only: er (vowel only) um (vowel plus nasal) hm (aspiration plus nasal) mm (nasal sound only). (See also: Backchannels, above.)	
Foreign words	If the foreign language uses the Roman alphabet, spell the word as in the language of origin; otherwise try to make a phonetic transliteration.	Ciao sang froid sushi tabu
Hyphens	Standard hyphenation rules apply. (Consult your dictionary.)	nonlinear pre-christian
Interruptions	Interruptions are dealt with by starting a new line to show the new speaker's contribution. If the first speaker continues, that continuation follows on a new line.	A: It's down below the B: the mine, right A: diamond mine
OK	OK should be written as 'okay'.	okay?
Overlaps	See Interruptions (above).	
Non-verbal vocalizations	Any non-words included in the transcript should be enclosed in square brackets and prefixed with an equal sign, e.g. [=laugh]. For consistency use the noun form, e.g. [=cough] rather than [=coughs] and the shorter option, e.g. [=laugh] rather than [=laughter] or [=laughing].	[=cough] [=laugh] [=sigh] [=sniff]

(continued)

Topic	Guidelines	Example(s)
Numbers	All numbers are written as words. If the speaker says 'zero' write it, if 'O' write o.	nineteen seventy-five zero point nine nine o seven o five one thousand and twenty-four
Pauses	Short pauses aren't marked. Silences that are subjectively noticeable are indicated by double dots (..). If in doubt don't mark pauses: only conspicuous pauses are marked.	B: I didn't .. I never even saw her
Punctuation of sentences („.!?)	Exclamation marks are not used. Punctuation in general is optional. Question marks may be placed at the end of an utterance when it is clear, by intonation and/or syntax, that a question has been asked (separated by a space from the preceding word). Commas and full-stops are only included (separated by a space from the preceding word) if the effort of omitting them seems unnatural and would slow down the transcription process.	Should we write this down?You're sure?
Quotations	Do NOT use 'smart' quotes! When a speaker is clearly reading or reciting another person's words, enclose the passage spoken in double quotation marks.	It says 'stop valve' on the top
Speaker turns	Each new speaker turn begins on a new line. For clarity we recommend double newlines between utterances. For OPTIMA transcripts we normally use N: for Nurse (or Doctor) and P: for Patient (or Participant) to identify speakers at the start of each utterance. The essential consideration is that each line consists only of words produced by a single speaker.*	P: it looks like there's water. N: where? P: er on the um ground er the floor
Spoken letters	When a speaker spells out a word, the letters should be in lower case with hyphens between letters.	no e on the end on the *x*-axis it's time in seconds that's k-e-y-n-s-h-a-m with a k w-e-t-h-e-r-a-l-l
Time stamps	Not recorded.	
Transcriber's comments	Enclose transcriber's comments within double parentheses. These should be used sparingly.	[[drilling in next room]] [[knock at door]] [[phone rings and speaker gets up]] [[seems to be a break in recording here]]
Unfinished words	Incomplete words should be terminated by an equal sign.	individ = It's fu = bloody impossible
Unintelligible speech	Words that simply can't be interpreted should be coded as [#]. Even if the transcriber is sure that a stretch of more than one uninterpretable word has been uttered this should still be coded as [#]. We only recognize two states, unintelligible or intelligible, so if the transcriber is uncertain he/she has to make a definite decision, either inserting [#] or the presumed word.	[#]

*Philosophical note: In general, we prefer splitting to lumping when it comes to utterance division; so for instance if one speaker inserts a word or short phrase or backchannel communication without breaking another speaker's flow, that would still be coded as three turns or utterances on three lines. This helps to ensure that each speaker's contributions can be separately identified. (Example(s): P: it looks like a pool of; N: yeah; P: water leaking from the er sink)

Appendix A2: Sample transcript after alignment

0. p: p:
1. um hm
2. the the
3. sink's sink
4. [00] is
5. overflowing overflowing
6. um and
7. the the
8. little little
9. [00] boy
10. boy's is
11. coming coming
12. off off
13. the the
14. stool stool
15. um and
16. and and
17. the the
18. little little
19. [00] boy
20. boy's is
21. trying trying
22. to to
23. take take
24. some some
25. cakes cakes
26. out out
27. of of
28. the the
29. tin tin
30. .. ..
31. n: n:
32. that's that's
33. fine fine
34. p: p:
35. [#] [00]
36. n: [00]
37. thank [00]
38. you [00]
39. very [00]
40. much okay
41. p: [00]
42. mm [00]