Journal of the Text Encoding Initiative

Issue 1 | June 2011
Selected Papers from the 2008 and 2009 TEI Conferences

A TEI-based Approach to Standardising Spoken Language Transcription

Thomas Schmidt

Electronic version
URL: http://journals.openedition.org/jtei/142
DOI: 10.4000/jtei.142
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Thomas Schmidt, « A TEI-based Approach to Standardising Spoken Language Transcription », *Journal of the Text Encoding Initiative* [Online], Issue 1 | June 2011, Online since 08 June 2011, connection on 20 April 2019. URL : http://journals.openedition.org/jtei/142 ; DOI : 10.4000/jtei.142

A TEI-based Approach to Standardising Spoken Language Transcription

Thomas Schmidt

AUTHOR'S NOTE

An earlier version of this paper, co-authored by Andreas Witt, was presented as "Transcription tools, transcription conventions and the TEI guidelines for transcriptions of speech" at the 2008 TEI Members Meeting in London. I am grateful to Peter M. Fischer and two anonymous reviewers for very helpful suggestions for improvement.

1. Introduction

1 Spoken language transcription is an important component of many types of humanities research. Among its central areas of application are linguistic disciplines like conversation and discourse analysis, dialectology and sociolinguistics, and phonetics and phonology. The methods and techniques employed for transcribing spoken language are at least as diverse as these areas of application. Different transcription conventions have been developed for different languages, research interests, and methodological traditions, and they are put into practice using a variety of computer tools, each of which comes with its own data model and formats. Consequently, there is, to date, no widely dominant method, let alone a real standard, for doing spoken language transcription. However, with the advent of digital research infrastructures, in which corpora from different sources can be combined and processed together, the need for such a standard becomes more and more obvious. Consider, for example, the following scenario: A researcher is interested in doing a cross-linguistic comparison of means of expressing

modality. He is going to base his study on transcribed spoken language data from different sources. Table 1 summarises these sources.

Table 1: File formats and transcription conventions for different spoken language corpora

Corpus (Language) [URL]	**File format**	**Transcription convention**
SBCSAE (American English) [http://projects.ldc.upenn.edu/SBCSAE/]	SBCSAE text format	DT1 (DuBois et al. 1993)
BNC spoken (British English) [http://www.natcorp.ox.ac.uk/]	BNC XML (TEI variant 1)	BNC Guidelines (Crowdy 1995)
CallFriend (American English) [http://talkbank.org/]	CHAT text format	CA-CHAT (MacWhinney 2000)
METU Spoken Turkish Corpus (Turkish) [http://std.metu.edu.tr/en]	EXMARaLDA (XML format)	HIAT (Rehbein et al. 2004)
Corpus Gesproken Nederlands (CGN, Dutch) [http://lands.let.kun.nl/cgn/ehome.htm]	Praat text format	CGN conventions (Goedertier et al. 2000)
Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, German) [http://agd.ids-mannheim.de/html/folk.shtml]	FOLKER (XML format)	cGAT (Selting et al. 2009)
Corpus de Langues Parlées en Interaction (CLAPI, French) [http://clapi.univ-lyon2.fr/]	CLAPI XML (TEI variant 2)	ICOR (Groupe Icor 2007)
Swedish Spoken Language Corpus (Swedish) [http://www.ling.gu.se/projekt/old_tal/SLcorpus.html]	Göteborg text format	GTS (Nivre et al. 1999)

2 Undoubtedly, the corpora have a lot in common as far as their designs, research backgrounds, and envisaged uses are concerned. Still, as the table illustrates, not a single one of them is compatible with any of the others, neither in terms of digital file formats nor transcription conventions used. In order to carry out his study, the researcher will thus have to familiarise himself with eight different file formats, eight different transcription conventions and, if he is not able or willing to do a lot of data conversion, eight different techniques or tools for querying the different corpora. Obviously, the world of spoken language corpora[1] is a fragmented one. The aim of this paper is to explore whether an approach based on the Guidelines of the TEI can help to overcome some of this fragmentation. In order for such an effort to be successful—that is, to really

reduce the variation—I think that it is necessary to take the following factors into account:

- Since spoken language transcription is a very time-consuming process, it is crucial for transcribers to have their work supported by adequate computer tools. Any standardisation effort should therefore be compatible with the more widely used tool formats. This compatibility should manifest itself in something that can be used in practice, such as a conversion tool for exchanging data between a tool and the standard.
- The reason for variation among transcription conventions and tool formats can be pure idiosyncrasy, but it can also be motivated by real differences in research interests or theoretical approaches. A standardisation effort should carefully distinguish between these two types of variation and suggest unifications only for the former type.
- Not least because the line between the two types of variation cannot always be easily drawn, any standardisation effort should leave room for negotiations between the stakeholders (that is, authors and users of transcription conventions, and developers and users of transcription tools) involved. This paper therefore does not intend to ultimately *define* a standard but rather to identify and order relevant input to it and, on that basis, suggest a general approach to standardisation the details of which are left to discussion.

3 Following these basic assumptions, the paper is structured as follows: Sections 2 and 3 look at two fundamentally different, but interrelated, things to standardise. Section 2 is concerned with the *macro structure* of transcriptions—that is, temporal information and information about classes of transcription and annotation entities (for example, verbal and non-verbal)—as defined in tool formats and data models. Section 3 is concerned with the *micro structure* of transcriptions—that is, names for, representations of, and relations between linguistic transcription entities like words, pauses, and semi-lexical entities. This is what a transcription convention usually defines. Both sections conclude with a suggestion of how to standardise commonalities between the different inputs with the help of the TEI. Section 4 then discusses some aspects of application—that is, ways of using the proposed standard format in practice.

2. Macro Structure and Tool Formats

4 Transcription tools support the user in connecting textual descriptions to selected parts of an audio or video recording. I will call the way in which such individual descriptions are organised into a single document the *macro structure* of a transcription. Transcription macro structures, and, consequently, the file formats used by the tools, usually remain on a relatively abstract, theory-neutral level. They are concerned with abstract categories for data organisation and with the temporal order of textual descriptions and their assignment to speakers, among other things, but they usually do *not* define any concrete entities derived from a theory of what should be transcribed (such as words and pauses). This latter task is delegated to transcription conventions (see the following section).[2]

2.1. Data Models: Commonalities and Differences

5 Disregarding word processors (like MS Word) and simple combinations of text editors and media players (like F4)[3], the following seven tools are among the most commonly used for spoken language transcription:[4]

- ANVIL (Kipp 2001), a tool originally developed for studies of multimodal behaviour

- CLAN/CHAT (MacWhinney 2000), the tool and data format belonging to the CHILDES database, originally developed for transcription and coding of child language data
- ELAN (Wittenburg et al. 2006), a multi-purpose tool used, among other things, for documentation of endangered languages and sign-language transcription
- EXMARaLDA Partitur-Editor (Schmidt and Wörner 2009), a multipurpose tool with a background in pragmatic discourse analysis, dialectology, and multilingualism research
- FOLKER (Schmidt and Schütte 2010), a transcription editor originally developed for the FOLK corpus for conversation analysis
- Praat (Boersma and Weenink 2010), software for doing phonetics by computer
- Transcriber (Barras et al. 2000), an editor originally developed for transcription of broadcast news

6 Although there are numerous differences in design and implementation of the tools, and although each tool reads and writes its own individual file format, their data models can all be understood as variants of the same base model. The basic entity of that data model is a time-aligned annotation—that is, a triple consisting of a start point, an end point, and a field containing the actual transcription or annotation.[5] Further structure is added by partitioning the set of basic entities into a number of tiers and assigning tiers to a speaker and/or to a type. As Schmidt et al. (2009) have shown, this simple structure can be viewed as a common denominator of all tools, and it can be used to establish a basic interoperability between them.

7 Beyond the common denominator, the tool models also differ in several details:

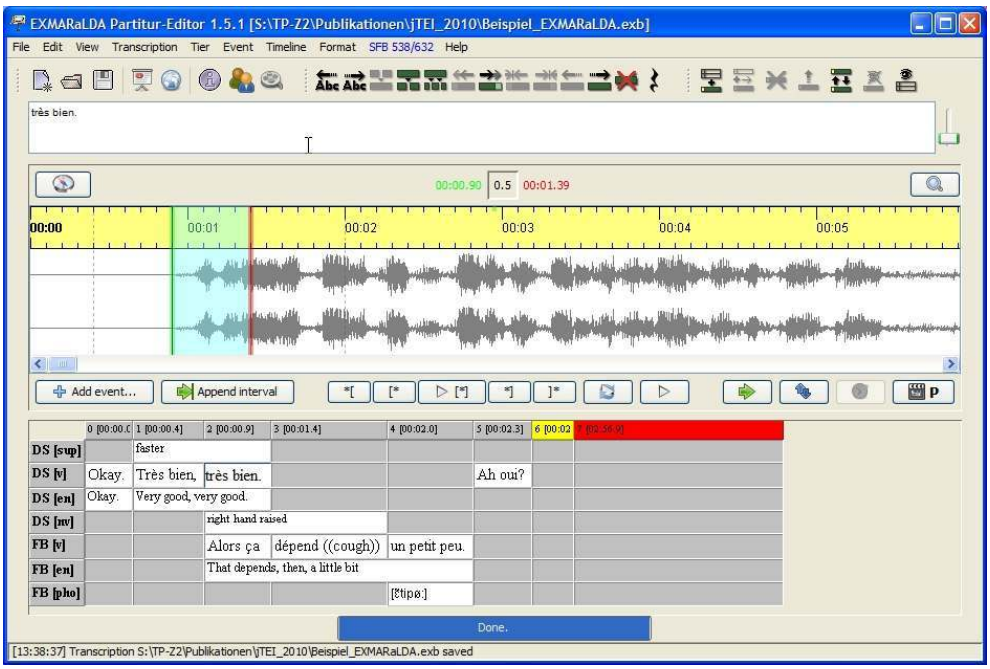
- Implicit vs. explicit timeline: In some models (like ANVIL and Praat), start and end points of the basic entities point directly to a time point in the recording. In other models (like EXMARaLDA and ELAN), they point to an external timeline— an ordered set of time points, which, in turn, can (but need not) have timestamps pointing into the recording.
- Speaker assignment of tiers: Some models (like EXMARaLDA and ELAN) allow (and sometimes require) tiers to be explicitly assigned to a speaker entity. Other models (like ANVIL and Praat), although they allow tiers to be characterised by a name and other features, do not have an explicit concept for speakers.
- Simple and structured annotations: In some models (like ANVIL and ELAN), the basic entities can have an internal structure, while in others (like EXMARaLDA and Praat), they always consist of simple text strings.
- Single layer and multi-layer: Some models (like FOLKER and Transcriber) provide a single tier for each speaker in which all annotation for that speaker has to be integrated. Other models allow multiple tiers for each speaker onto which annotations of different kinds (such as verbal vs. non-verbal or segmental vs. supra-segmental) can be distributed. In most models of the latter type, tier categories and semantics can be freely defined on the basis of a few abstract tier types (as in ANVIL, ELAN, EXMARaLDA, but see next point), whereas CLAN/CHAT predefines an extensive set of tier categories and a semantics for them.
- Tier types and dependencies: All multi-layer tools provide a system for classifying tiers according to their structure and semantics. The tier types can be associated with certain structural constraints on annotations within the respective tier or in relation to annotations in another tier. This often results in a tier hierarchy where one tier is regarded as primary and other tiers as subordinate to (or dependent on) the primary tier. No two tools use the same system of tier types, but there are some obvious commonalities and interrelations between the systems.

8 Schmidt et al. (2009) conclude that, "given that the diversity in tool formats is to a great part motivated by the different specializations of the respective tools", a full assimilation of the different data models is neither theoretically desirable nor practically possible. However, the similarities between the data models clearly outweigh the differences. I would therefore like to argue that, at least for the purposes of this paper, it will be sufficient to declare one of the formats as a typical exponent of a class containing all the others, and use this typical exponent as the basis for a transformation to TEI. The fact that EXMARaLDA has conversion filters for importing the formats of all the other tools shows that this assumption is not only true in theory, but can also be put to use in practice. In what follows, I will therefore use EXMARaLDA's data model as a representative of all the other tools.

2.2. EXMARaLDA's Data Model and Format

9 Concerning the above parameters, EXMARaLDA's data model has an explicit timeline, allows speaker assignment of tiers, uses only simple annotations, allows multi-layer annotations, and distinguishes three tier types which I will illustrate with the help of the following example. Figure 1 shows a transcription as displayed by the EXMARaLDA Partitur-Editor.

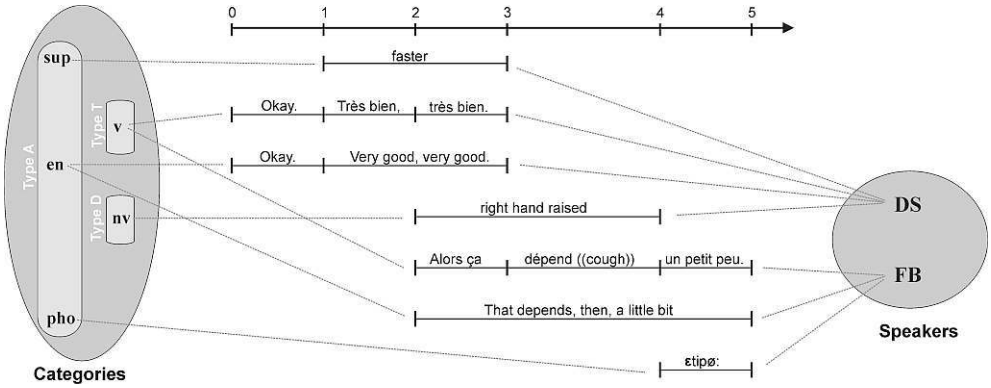
Figure 1: Example transcription as displayed in the EXMARaLDA Partitur-Editor with a waveform representation of the recording (top) and a musical score representation of the transcription (bottom). Annotations (white fields in the musical score) are assigned to tiers ("rows" of the score) and intervals of the timeline ("columns" of the score). The tiers are labelled with abbreviations for the corresponding speakers ("DS" and "FB") and with a category ("sup", "v", etc.).



10 The transcription consists of twelve annotation triples, organised into seven tiers, each of which is attributed to one of two distinct speakers (`DS` and `FB`), one of five distinct (freely definable) categories (`sup`, `v`, `en`, `nv` and `pho`) and one of three (predefined) tier types. Note that the same mechanism—assigning identical start and end points to the respective annotations—is used to represent both temporal simultaneity (as in the speaker overlap

between “très bien” and “Alors ça”) and semantic equivalence (as between the orthographic transcription “un petit peu” and its phonetic counterpart “ [ɛ̃tipø:] ”). Figure 2 gives a schematic representation of the underlying data model.

Figure 2: Schematic representation of the EXMARaLDA data model



11 Tiers of type **T(RANSCRIPTION)** contain the primary information—that is, the transcription of words uttered by the respective speaker alongside with descriptions of non-phonological phenomena (such as coughing and pauses) which are alternative (rather than simultaneous) to the actual speech. Tiers of type **A(NNOTATION)** contain information which is dependent on the primary tiers. For instance, the tiers of category `en` contain English translations of the speakers' French utterances, whereas the tier of type `sup` contains annotations which describe suprasegmental features of transcribed words. Finally, in tiers of type **D(ESCRIPTION)** secondary information, which is independent of the transcribed words etc., can be entered. In the example, the tier of category `nv` contains an annotation for a non-verbal action by speaker `DS`. The data model has the following simple constraints with respect to tier types:

1. Tiers of type **T** and **A** must be attributed to a speaker (if a tier of type **A** and a tier of type **T** are attributed to the same speaker, the latter is the *parent tier* of the former).
2. There has to be exactly only one tier of type **T** for each speaker, but there can be any number of tiers of type **A** and **D**.
3. For each annotation in a tier of type **A**, there must be an annotation or a connected sequence of annotations in the parent tier with the same start and end point.

12 As illustrated in figure 3, EXMARaLDA represents this data model in an XML file which hierarchically organises individual annotations (`<event>` elements) into tiers (`<tier>` elements). All other structural relations, in particular the assignment of annotations to points in the timeline and the assignment of tiers to speakers, are not expressed in the document hierarchy, but with the help of pointers to `@id` attributes.

Figure 3: XML representation of an EXMARaLDA transcription (simplified)

```
<basic-transcription>
  <head>
    <speakertable>
      <speaker id="SPK0" abbreviation="DS"/>
      <speaker id="SPK1" abbreviation="FB"/>
    </speakertable>
  </head>
  <body>
    <common-timeline>
      <tli id="T1" time="0.0"/>
      <tli id="T2" time="0.4"/>
      <tli id="T3" time="0.9"/>
      <tli id="T4" time="1.4"/>
      <tli id="T5" time="2.0"/>
      <tli id="T6" time="2.3"/>
      <tli id="T7" time="2.6"/>
    </common-timeline>
    <tier id="TIE1" speaker="SPK0" category="sup" type="a">
      <event start="T2" end="T4">faster</event>
    </tier>
    <tier id="TIE2" speaker="SPK0" category="v" type="t">
      <event start="T1" end="T2">Okay.</event>
      <event start="T2" end="T3">Très bien,</event>
      <event start="T3" end="T4">très bien.</event>
      <event start="T6" end="T7">Ah oui ?</event>
    </tier>
    <tier id="TIE3" speaker="SPK0" category="en" type="a">
      <event start="T1" end="T2">Okay.</event>
      <event start="T2" end="T4">Very good, very good.</event>
    </tier>
    <tier id="TIE4" speaker="SPK0" category="nv" type="d">
      <event start="T3" end="T5">right hand raised</event>
    </tier>
    <tier id="TIE5" speaker="SPK1" category="v" type="t">
      <event start="T3" end="T4">Alors ça</event>
      <event start="T4" end="T5">dépend ((cough))</event>
      <event start="T5" end="T6">un petit peu.</event>
    </tier>
    <tier id="TIE6" speaker="SPK1" category="en" type="a">
      <event start="T3" end="T6">That depends, then, a little bit</event>
    </tier>
    <tier id="TIE7" speaker="SPK1" category="pho" type="a">
      <event start="T5" end="T6">[ɛ̃tipø:]</event>
    </tier>
  </body>
</basic-transcription>
```

13 While this has proven a practically adequate representation of the data model for the purposes of the EXMARaLDA editor (and similar XML formats are used, for instance, by ANVIL and ELAN), it ha s some obvious drawbacks from the point of view of XML based data modelling and processing:

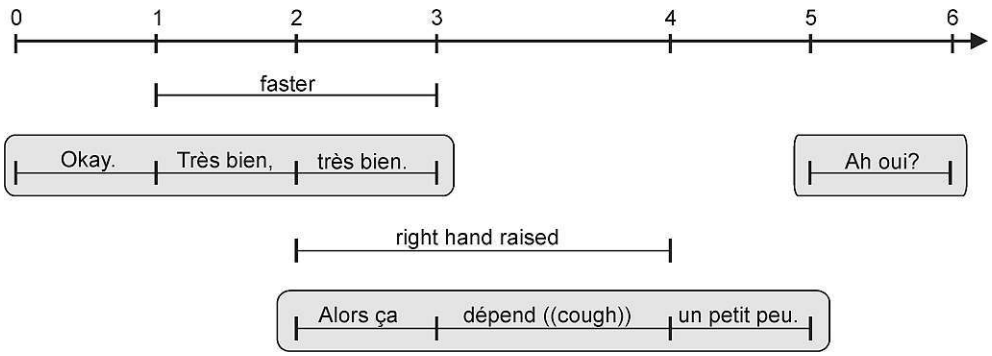
- The document order of individual annotations does not match the order in which the corresponding phenomena occur in the transcribed discourse.
- Likewise, elements having a close semantic relationship, like the orthographic and phonetic transcriptions in the last two tiers, are not necessarily close to one another in the document.
- The dependency between annotations in tiers of type T and tiers of type A is not explicitly represented in the document structure.
- Since the division of annotations is motivated by the temporal structure of the discourse, the boundaries of individual annotation elements may cut through linguistic entities. This is the case, for example, for the utterance "Alors ça dépend ((cough)) un petit peu.", which is distributed across three `<event>` elements in order enable the representation of different simultaneity relations in the discourse.

14 One resulting disadvantage is that certain XML techniques (like XPath queries) can become inefficient for such documents because the techniques are optimised for processing tree structures, whereas the principal structure of the document is not represented in the document tree. Another disadvantage is that the (manual) insertion of additional markup, such as with the help of a standard XML editor, becomes difficult because the elements of the document do not behave as in a "normal" (i.e. written) text. As a basis for a transformation to a TEI-conformant form, this kind of document organisation is thus not ideal. A first question on the way to a TEI-based standardisation therefore is whether an equivalent XML representation of the data model can be found which does not suffer from the same drawbacks.

2.3. A TEI Representation of EXMARaLDA's Data Model

15 My suggestion is to derive such an equivalent representation on the basis of the concept of a *segment chain*. With respect to the EXMARaLDA data model, a segment chain can be defined as any maximally long, temporally connected sequence of annotations in a tier of type T. The above example contains three such segment chains, marked with grey boxes in figure 4.

Figure 4: Combing annotations into segment chains



16 These segment chains–which loosely correspond to an entity often called a *turn* or a *speaker contribution* —have three important structural properties:

- They are implicitly contained in the data model and can be automatically derived from it.
- They re-combine the character data of linguistic entities (words and utterances) from tiers of type T, which were separated in the data model due to temporal considerations (temporal overlap of annotations) into a superordinate entity.
- Since annotations in tiers of type A will, by definition, not cross the boundaries of such segment chains, each such annotation can be assigned to exactly one segment chain.

17 Subsuming all annotations in tiers of type A under "their" segment chain and ordering segment chains by their start points, a document can thus be constructed whose document order is globally analogous to the actual sequence of events in the transcribed discourse, whose elements locally behave like normal written text, and in which dependent annotations are grouped together with the annotations they depend on.

18 Chapters 3 (Elements Available in All TEI Documents), 4 (Default Text Structure), 8 (Transcriptions of Speech), 16 (Linking, Segmentation, and Alignment) and 17 (Simple Analytic Mechanisms) of the P5 Guidelines furnish all the elements necessary to represent such a document in TEI. More specifically, the following elements can be used:

- `<person>` inside a `<particDesc>` to define speakers
- `<when>` inside a `<timeline>` to define the timeline
- `<div>` to group segment chains and corresponding annotations
- `<u>` to represent the actual segment chains[6] with a `@who` attribute assigning this element (and its `<spanGrp>` siblings) to a speaker
- `<anchor>` inside `<u>` with `@synch` attributes pointing to `<when>` elements to represent the internal temporal structure of a segment chain
- `<spanGrp>` to group annotations of the same type (i.e. coming from the same tier)
- `<span>` inside `<spanGrp>` with `@from` and `@to` attributes to represent dependent annotations and their position in the timeline
- `<incident>` to represent the remaining annotations coming from tiers of type D

19 Figure 5 shows a TEI-conformant document which uses these elements and is equivalent to the document in figure 3.[7]

Figure 5: TEI representation equivalent to the representation in figure 3 (simplified, see Appendix for the full version)

```
<TEI>
  <teiHeader>
    <profileDesc>
      <particDesc>
        <person xml:id="SPK0">
          <persName>
            <abbr>DS</abbr>
          </persName>
        </person>
        <person xml:id="SPK1">
          <persName>
            <abbr>FB</abbr>
          </persName>
        </person>
      </particDesc>
    </profileDesc>
  </teiHeader>

  <text>
    <timeline unit="s">
      <when xml:id="T1" absolute="00:00:00.0"/>
      <when xml:id="T2" absolute="00:00:00.4"/>
      <when xml:id="T3" absolute="00:00:00.9"/>
      <when xml:id="T4" absolute="00:00:01.4"/>
      <when xml:id="T5" absolute="00:00:02.0"/>
      <when xml:id="T6" absolute="00:00:02.3"/>
      <when xml:id="T7” absolute="00:00:02.6"/>
    </timeline>
  <body>
      <div>
        <u who="#SPK0">
          <anchor synch="#T1"/>Okay. <anchor synch="#T2"/>Très bien,
          <anchor synch="#T3"/>très bien. <anchor synch="#T4"/>
        </u>
        <spanGrp type="sup">
          <span from="#T2" to="#T4">faster</span>
        </spanGrp>
        <spanGrp type="en">
          <span from="#T1" to="#T2">Okay. </span>
          <span from="#T2" to="#T4">Very good, very good.</span>
        </spanGrp>
      </div>
      <div>
        <u who="#SPK1">
          <anchor synch="#T3"/>Alors ça <anchor synch="#T4"/>dépend
((cough))
          <anchor synch="#T5"/>un petit peu. <anchor synch="#T6"/>
        </u>
        <spanGrp type="en">
          <span from="#T3" to="#T6">That depends, then, a little bit</span>
```

```
        </spanGrp>
        <spanGrp type="pho">
          <span from="#T5" to="#T6">[ɛ̃tipø:]</span>
        </spanGrp>
      </div>
      <incident who="#SPK0" type="nv" start="#T3" end="#T5">
        <desc>right hand raised</desc>
      </incident>
      <div>
        <u who="#SPK0">
          <anchor synch="#T6"/>Ah oui?. <anchor synch="#T7"/>
        </u>
      </div>
  </body>
  </text>
</TEI>
```

3. Micro Structure and Transcription Conventions

20 If, as described above, the macro structure of a transcription is concerned with the way textual elements are organised and put into relation with one another in a transcription document, the micro structure of a transcription can be said to specify the form and semantics of the textual elements themselves. Whereas macro structure is defined by tool developers and represented in file format specifications, micro structure is defined by transcribing linguists and represented in transcription conventions. There are numerous, if not countless, such conventions, most of which are specific to a single corpus or project and have never been published for a larger audience. Among those that *have* been published in some form or other are the following:

- HIAT (Halbinterpretative Arbeitstranskriptionen: Ehlich and Rehbein 1976; Rehbein et al. 2004), a system widely used in the functional pragmatics research community
- GAT (Gesprächsanalytisches Transkriptionssystem: Selting et al. 2009), a system widely used in the German conversation analysis research community for the transcription of German, and cGAT (Schütte and Schmidt 2010) an adaptation of GAT used for transcription in the FOLK corpus
- CHAT (Codes for Human Analysis of Transcripts: MacWhinney 2000), a system widely used in the child language research community, and CA-CHAT, an adaptation of CHAT to CA (Conversation analysis, Sacks et al. 1978) for use in conversation analysis
- DT1 (Discourse Transcription: DuBois et al. 1993), a system used for transcription of the Santa Barbara Corpus of Spoken American English
- Convention ICOR (ICOR 2007), a system used for the French CLAPI database
- GTS (Göteborg Transcription Standard: Nivre et al. 1999), a system used for the Spoken Swedish Corpus at Göteborg University

21 As will be detailed in the following subsections, these conventions have a lot in common. Although some of them claim to be "unified systems" (GAT) or even "standards" (GTS), they exist more or less independently of one another. In contrast to the situation with tool formats, there have been few attempts to establish "interoperability" between transcription conventions; real standardisation efforts have, to my knowledge, not been

undertaken at all. The present paper is not a place to carry out a full comparative analysis of the systems that would be needed for such a standardisation effort. Instead, I will restrict myself to discussing some commonalities and differences by using examples and working under the assumption that the same method can be transferred to other aspects of the systems. Schmidt (2005a) carries out a more comprehensive and detailed analysis of two of the systems mentioned here (HIAT and GAT).

3.1. Commonalities and Differences

22 Perhaps the most fundamental commonality among the conventions is that they depart from standard written orthography in order to motivate and explain their rules for representing spoken language in the written medium. An important consequence of this is that the entity "word" is present in all the conventions with more or less the same meaning, namely that of a word as defined by standard orthography. Two other basic entities shared by all the conventions are unfilled pauses and audible non-speech events like breathing, laughing or coughing. Furthermore, all of the conventions specify ways to represent uncertainty in transcription (sometimes with the possibility to provide alternatives to an uncertain part) and to represent incomprehensible passages. I will call these five elements the *basic building blocks* of transcription conventions.

23 Another class of entities to be found in most systems consists of prosodic characterisations of words or parts thereof. This class can comprise phenomena like (emphatic) stress or lengthening of syllables. Finally, most systems define entities which summarise words and other basic building blocks into larger units analogous (but explicitly not identical) to the sentence in written language.

24 Taking these commonalities as a starting point, I will illustrate some important differences between the conventions using the set of examples in figure 6 in which a fictitious stretch of speech is transcribed according to five different transcription systems.[8]

Figure 6: Transcriptions of the same stretch of speech according to five different conventions

HIAT	((coughs)) You must/ you (should) let • it be. ((laughs)) Pleease!
GAT	((coughs)) you must- you (should/could) let (-) it be; ((laughs)) plea:se-
CHAT	&=coughs you must... you should let # it be. &=laughs please!
DT1	(COUGH) you must-- you <X should X> let .. it be. @@ please?
cGAT	((coughs)) you must you (should/could) let (-) it be ((laughs)) please

25 Obviously, some variation is due only to symbolic differences among the conventions. Thus, HIAT, GAT and cGAT describe non-verbal incidents ("coughs") in double parentheses, whereas CHAT marks such descriptions with the prefixed symbols `&=` and

DT1 chooses capital letters between single parentheses and, additionally, has special predefined symbols for certain such incidents (laughing is represented by the symbols `@@`). Similarly, each system has its own symbol(s) for representing a short, unmeasured pause: the bullet • in HIAT, the symbols `(-)` in GAT and cGAT, the hash sign # in CHAT, and two full stops (periods) in DT1.

26 The conventions also vary in what phenomena are represented in the transcription. Thus, the lengthening of the vowel in the word "please" is indicated in HIAT through a reduplication of the vowel symbol and through the insertion of a colon in GAT (this being another case of symbolic variation), but it is not represented at all in the other three systems. Similarly, transcriber uncertainty with respect to a given word can be marked in HIAT, GAT, cGAT and DT1 (through single parentheses in the first three and through a pair of `<X` and `X>` in the latter), but only GAT and cGAT also provide the possibility to specify one or more alternative transcriptions for an uncertain word (added inside the parentheses after a slash).

27 While symbolic and other variation discussed so far remain on the level of basic building blocks, a last type of variation is more complex and concerns the way basic transcription units are organised into larger structures. This type of variation is visible in the punctuation symbols used in figure 6, specifically:

- HIAT divides the stretch of speech into two entities called utterances. Utterances are pragmatic units of speech, identified and classified according to function-based criteria, most importantly their mood. The first utterance is terminated by a full stop (period), indicating that it is in declarative mood, while the second is terminated by an exclamation point, marking its mood as exclamative. A third punctuation symbol—the forward slash behind the word "must"—indicates a self-repair but does not act as an utterance terminator. Note that in contrast to all other systems, HIAT uses capitalisation of words at the beginning of utterances.
- GAT divides the same stretch of speech into three entities called intonation phrases. Intonation phrases are prosodic units of speech, identified and classified according to form-based criteria, most importantly their intonation contour. The first and third intonation phrases are terminated by a hyphen, indicating a level final pitch movement. The second intonation phrase is terminated by a semicolon, which stands for a falling final pitch movement.
- CHAT proceeds similarly to HIAT, but has three utterances instead of two. The first is terminated by an ellipsis symbol (three dots), marking it as an interrupted utterance. The other two are marked by a full stop (period) and an exclamation point, making them declarative and emphatic, respectively.
- The corresponding entities in DT1 are called intonation units. The first is terminated by two hyphens (an interrupted intonation unit), the second one by a full stop (period) (a terminative intonation unit), and the third one by a question mark (an "appeal").
- cGAT, finally, does not group basic building blocks into larger entities at all.

28 If the information codified in transcription conventions is to be standardised, these different kinds of variation between the systems must be taken into account. Ideally, a standard should make sure that pure symbolic variation is harmonised by mapping different surface forms onto standard single form, and that all other variation is expressed in a manner that conserves the original diversity while still making it possible to process transcriptions from different sources on a common basis.

29 I think that the TEI Guidelines furnish all the necessary elements for such a standardisation; at least the following elements from chapters 3 (Elements Available in All TEI Documents), 4 (Default Text Structure), 8 (Transcriptions of Speech) and 17 (Simple Analytic Mechanisms) will be necessary to adequately represent transcriptions according to any of the above conventions:

- `<w>` and `<c>` to mark up individual words and punctuation characters (unless the semantics of a punctuation character is already represented through another mechanism in the markup), possibly with an attribute `@type` to characterise a word as a repaired form, as an assimilated form, etc. or to note that a character represents a lengthened phoneme
- `<pause>` with a `@dur` attribute and `<incident>` with a `<desc>` child to represent pauses and non-speech events
- `<unclear>` elements, possibly with a superordinate `<choice>` element to represent uncertain transcriptions and alternatives
- `<seg>` elements with a `@function` attribute to provide the general name for such units in the respective conventions (such as utterance vs. intonation unit) and a `@type` attribute to capture the specific characterisation of that unit (such as declarative vs. interrupted)

30 Using these elements, the `<u>` elements in the example from figure 5 (which follows the HIAT convention) could be marked up as shown in figure 7.

Figure 7: TEI marked up version (according to HIAT) of the transcription from figure 5 (simplified)

```
<TEI>
  <!-- [...] -->
  <body>
    <div>
      <u who="#SPK0">
        <anchor synch="#T1"/>
        <seg function="utterance" type="declarative">
          <w>Okay</w>
        </seg>
        <anchor synch="#T2"/>
        <seg function="utterance" type="declarative">
          <w>Très</w>
          <w>bien</w>
          <c>,</c>
          <anchor synch="#T3"/>
          <w>très</w>
          <w>bien</w>
        </seg>
        <anchor synch="#T4"/>
      </u>
      <!-- [...] -->
    </div>
    <div>
      <u who="#SPK1">
        <anchor synch="#T3"/>
        <seg function="utterance" type="declarative">
          <w>Alors</w>
          <w>ça</w>
          <anchor synch="#T4"/>
          <w>dépend</w>
          <incident>
            <desc>cough</desc>
          </incident>
          <anchor synch="#T5"/>
          <w>un</w>
          <w>petit</w>
          <w>peu</w>
        </seg>
        <anchor synch="#T6"/>
      </u>
      <!-- [...] -->
    </div>
  </body>
  <!-- [...] -->
</TEI>
```

31 If the same stretch of speech is transcribed according to different conventions, the resulting TEI markup will be the same with respects to elements like <w>, <incident>,

and <pause> where there is only symbolic variation, but it can differ with respect to elements like <seg> where there is a "real" difference between the systems. Figure 7 shows possible markup for three of the examples from figure 6.

Figure 8: TEI markup for examples from figure 6 (simplified)

CHAT

```
<u>
  <seg function="utterance" type="interrupted">
    <incident>
      <desc>coughs</desc>
    </incident>
    <w>you</w>
    <w>must</w>
  </seg>
  <seg function="utterance" type="declarative">
    <w>you</w>
    <w>should</w>
    <w>let</w>
    <pause dur="short"/>
    <w>it</w>
    <w>be</w>
  </seg>
  <seg function="utterance" type="emphatic">
    <incident>
      <desc>laughs</desc>
    </incident>
    <w>please</w>
  </seg>
</u>
```

DT1

```
<u>
  <seg function="intonation_unit" type="interrupted">
    <incident>
      <desc>cough</desc>
    </incident>
    <w>you</w>
    <w>must</w>
  </seg>
  <seg function="intonation_unit" type="terminative">
    <w>you</w>
    <unclear>
      <w>should</w>
    </unclear>
    <w>let</w>
    <pause dur="short"/>
    <w>it</w>
    <w>be</w>
  </seg>
  <seg function="intonation_unit" type="appeal">
    <incident>
      <desc>laughs</desc>
    </incident>
    <w>please</w>
  </seg>
</u>
```

cGAT

```
<u>
  <incident>
    <desc>coughs</desc>
  </incident>
  <w>you</w>
  <w>must</w>
  <w>you</w>
  <choice>
    <unclear>
      <w>should</w>
    </unclear>
    <unclear>
      <w>could</w>
    </unclear>
  </choice>
  <w>let</w>
  <pause dur="short"/>
  <w>it</w>
  <w>be</w>
  <incident>
      <desc>laughs</desc>
   </incident>
   <w>please</w>
</u>
```

4. Application of this Standard Format

32 Having defined a proposal for a TEI-based standard, I will now turn to the question of how to use it in practice. Most importantly, this means thinking of ways in which transcribers can efficiently produce standard, conformant transcriptions. Ideally, they will continue to be able to use the tools they are familiar with and to focus on the transcription task itself rather than on issues related to XML and TEI encoding.

33 These requirements are relatively easy to meet as far as the macro structure of transcriptions is concerned: the format illustrated in figure 5 is isomorphic to EXMARaLDA's tool format. This format, in turn, is compatible to a large extent with all the other tool formats mentioned in Section 2 because of the import and export routines built into EXMARaLDA and several other tools. By virtue of transitivity, making all tools compatible with the format in figure 4 is therefore simply a matter of defining a one-to-one mapping between one tool format and the TEI format. In order to ensure maximal portability, this mapping should be accomplished with an XML-only approach using XSL stylesheet transformations. XSL stylesheets which transform an EXMARaLDA transcription into an equivalent TEI representation and vice versa have been made available on the EXMARaLDA website at http://www.exmaralda.org/tei.html. The stylesheets have also been integrated into the EXMARaLDA editor, where the transformations can be carried out using the tool's import and export functions. For

formats from other tools, either a direct mapping could be defined in an analogous manner, or EXMARaLDA could be used as an intermediary representation.

34 The requirements are harder to meet for the micro structure of transcriptions. Most commonly used tools (FOLKER being an exception) do not provide a way of directly representing micro structure in their file formats. While the markup expressing the micro structure could be added manually in a generic XML editor after a tool's format has been converted to the TEI representation of figure 5, this procedure would be rather inefficient since it requires a second tedious manual processing step after the actual transcription has been completed. A more efficient way is to automatically derive the micro structure markup from the regularities formulated inside the transcription conventions. This is possible if we interpret some of the symbols defined by a convention as an implicit (and non-standardised) markup and formulate an algorithm—a parser—to transform this implicit markup into explicit, TEI-conformant XML markup. Figure 9 exemplifies this process for the HIAT example from figures 5 and 7.

Figure 9: **Parsing for micro-structure**

1. Unparsed <u>

```
<u>
  <anchor synch="#T3"/>Alors ça <anchor synch="#T4"/>dépend ((cough))
  <anchor synch="#T5"/>un petit peu. <anchor synch="#T6"/>
</u>
```

2. Character data of unparsed <u>

```
Alors ça dépend ((cough)) un petit peu.
```

3. Parsing: Transforming implicit to explicit markup

Alors␣ça␣dépend␣((cough))␣un␣petit␣peu.␣ [9]

```
<seg function="utterance" type="declarative">
  <w>Alors</w><w>ça</w><w>dépend</w>
  <incident><desc>cough</desc></incident>
  <w>un</w><w>petit</w><w>peu</w>
</seg>
```

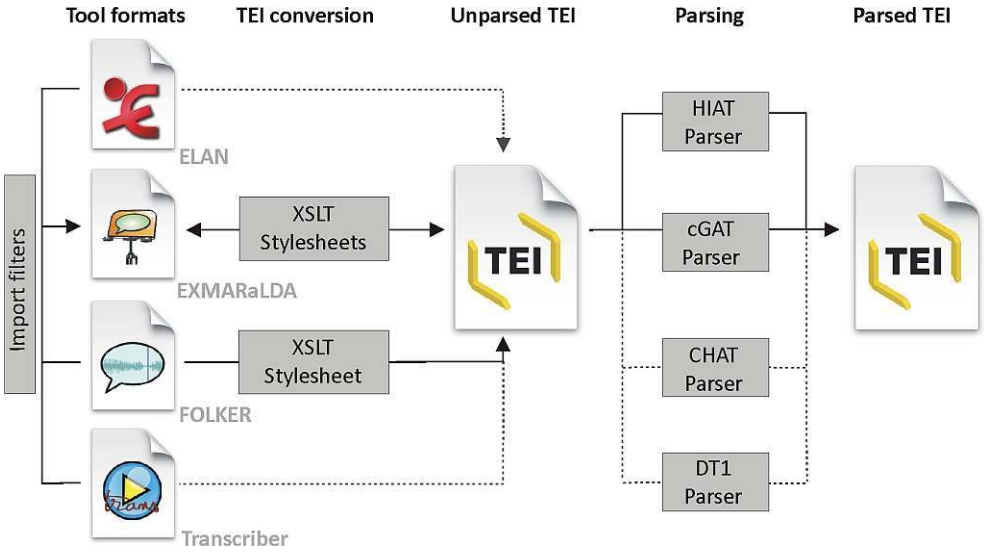
4. Reinserting anchors

```
<u>
  <seg function="utterance" type="declarative">
    <anchor synch="#T3"/>
    <w>Alors</w><w>ça</w>
    <anchor synch="#T4"/>
    <w>dépend</w>
    <incident><desc>cough</desc></incident>
    <anchor synch="#T5"/>
    <w>un</w><w>petit</w><w>peu</w>
    <anchor synch="#T6"/>
  </seg>
</u>
```

35 The implicit markup in this case consists of spaces indicating word boundaries, double parentheses indicating non-phonological descriptions, and the full stop (period) indicating and qualifying an utterance boundary. Of course, in order for the parsing algorithm to work reliably, the symbols interpreted as implicit markup must have been rigidly and unambiguously defined in the respective convention. Luckily, all conventions claim to ensure this unambiguousness in their choice of transcription symbols.[10] The parsing algorithm can then, in principle, be implemented in any technology and does not need to take any prescribed form as long as it produces correct output (a well-formed TEI-compliant XML fragment) for correct input (a string following the rules of a given transcription system).[11] EXMARaLDA has built-in parsing algorithms for HIAT, GAT, cGAT and CHAT which are implemented as finite-state transducers in Java, showing that a very simple parsing technique can be sufficient to deal with several of the transcription conventions mentioned above.

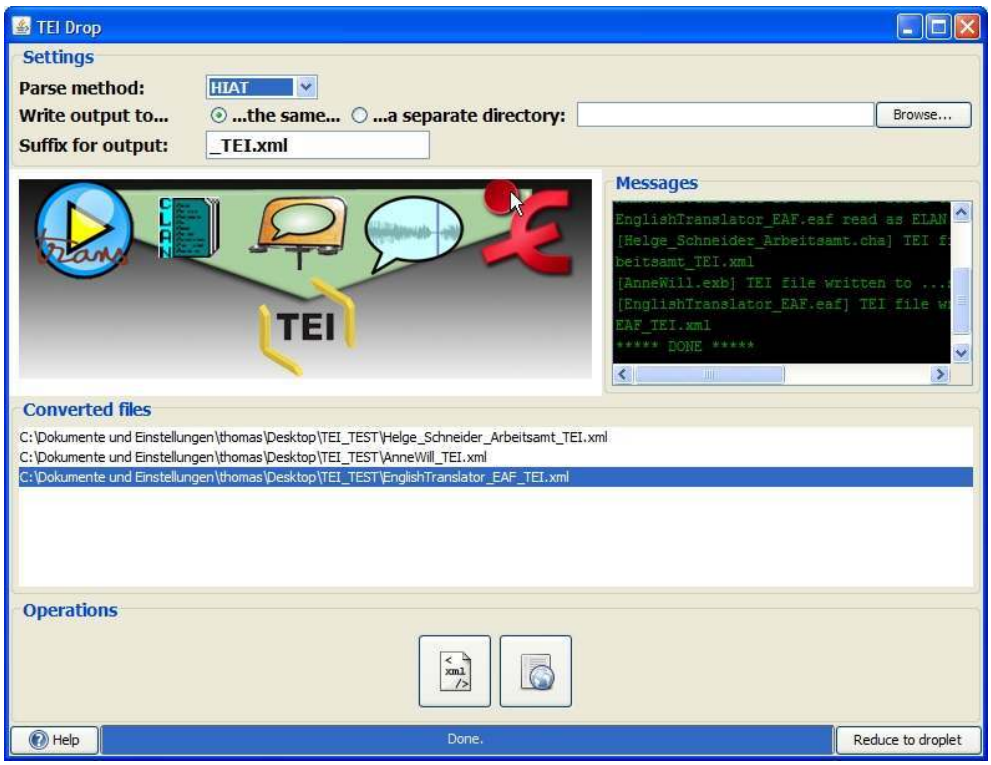
36 Transforming a tool format to a corresponding TEI format in which both macro and micro structure are represented is thus a two-step-process. First, a generic TEI document is produced in which only the macro structure is represented. Second, a parsing algorithm is applied, which adds markup for the micro structure. Figure 10 gives a schematic illustration of the transformation workflow.[12]

Figure 10: Transformation workflow from tool format to parsed TEI document



37 In order to make this transformation workflow available to users in a maximally accessible way, we have written a Java droplet which takes as input any CHAT, ELAN, EXMARaLDA, FOLKER or Transcriber transcription file and transforms it to a TEI file using a set of parameters—the parsing algorithm to be used among them—specified by the user. Figure 11 shows a screenshot of that application, which will be made freely available as a part of the EXMARaLDA tool package.

Figure 11: Screenshot of TEI Drop



5. Summary, Conclusion and Outlook

38 In this paper, I have formulated a proposal for standardising spoken language transcription with the help of the TEI Guidelines. The proposal consists of two principal components. First, a TEI-conformant format is defined that is structurally equivalent to the formats written by several widely used transcription tools and which represents the macro structure of the transcription in a form that is well-suited for standard XML processing. Second, implicit markup contained in the character data of such documents is transformed to explicit TEI conformant markup using a parsing algorithm that embodies the formal regularities of a transcription convention. The resulting document then represents both macro and micro structure of the transcription in a TEI-compliant way. A droplet application enables users to carry out the transformation from tool format to TEI format and the parsing of the TEI format according to a specific transcription convention in a user-friendly way.

39 The route to standardisation formulated here can be viewed as a synthesis of work in three areas related to spoken language transcription: tool development, TEI encoding, and transcription conventions. All three can be said to have as one of their goals unification or harmonisation of similar practices, but each of them foregrounds a different aspect in that goal.

40 Tool developers usually aim at defining data models and formats which are both general and flexible enough to be used for different data types and different research interests while at the same time specific enough to allow for efficient processing of the data. As the present paper has shown, the solutions they have developed to meet these requirements are sufficiently interoperable to become the first ingredient of the standardisation effort.

41 The goal of the TEI is to provide a common tag set for the representation of texts in digital form where spoken language transcriptions are simply viewed as "texts of a special kind". Again, the present paper has shown that the existing solutions—as formulated in the P5 version of the Guidelines—are comprehensive and detailed enough to adequately represent commonalities and differences between transcription formats and conventions. They can thus become the second ingredient of the standardisation effort.

42 The situation is a little less clear for the third ingredient, the transcription conventions. Here, the present paper has shown—as a proof of concept at least—that existing conventions are sufficiently systematic to become the basis for a parsing algorithm. However, the formalisations required to derive such an algorithm are usually not explicitly defined inside the conventions but have to be inferred from a potentially error-prone interpretation of an informal text. Likewise, the distinction drawn here between symbolic and other variation among transcription conventions, though arguably very important for standardisation, is not a topic that the conventions themselves deal with at greater length. It seems, therefore, that in this area, the idea of formal standardisation has not yet gained as much ground as in the area of tools and the TEI. If the approach suggested here is to become the basis of a full-grown standard, most work will probably remain in standardising transcription conventions.

43 If we assume that such a full-grown standard can be agreed upon eventually, the task of the example researcher from the introduction will become considerably easier. He will be

dealing with only a single format, which rests on a well-defined and well-documented basis, namely the TEI Guidelines. Inside that format, pure symbolic variation between different transcription conventions will be levelled out, and "genuine" theory-motivated variation will be retained in a manner which facilitates a common processing of data from different sources. Moreover, new data in the same form will easily be produced because transcribers will continue to use established technology and established conventions for their task. Last but not least, the fact that the proposed standard for *spoken* language transcription draws from the same set of TEI elements as many other actual or proposed standards in the field of *written* language, such as the Corpus Encoding Standard (CES, see http://www.xces.org/), also opens a potential for common processing of spoken and written data.

BIBLIOGRAPHY

Barras, C., E. Geoffrois, Z. Wu, and M. Liberman. 2001. "Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production." *Speech Communication* 33: 5-22.

Bird, S., and Mark Liberman. 2001. "A formal framework for linguistic annotation." *Speech Communication* 33: 23-60.

Boersma, P., and D. Weenink. 2010. "Praat: doing phonetics by computer."Version 5.1.05. http://www.praat.org/.

Crowdy, S. 1995. "The BNC spoken corpus." In *Spoken English on computer: transcription, mark-up and application*, ed. G. Leech, G. Myers and J. Thomas, 224–235. Harlow: Longman.

DuBois, J., S. Schuetze-Coburn, S. Cumming, and D. Paolino. 2003. "Outline of Discourse Transcription." In *Talking Data: Transcription and Coding in Discourse Research*, ed. J. Edwards and M. Lampert, 45–89. Hillsdale, NJ: Erlbaum.

Ehlich, K. and J. Rehbein. 1976. "Halbinterpretative Arbeitstranskriptionen (HIAT)." *Linguistische Berichte* 45: 21–41.

Ehlich, K. 2003. "HIAT: A Transcription System for Discourse Data." In *Talking Data: Transcription and Coding in Discourse Research*, ed. J. Edwards and M. Lampert, 123–148. Hillsdale, NJ: Erlbaum.

Goedertier, W., S. Goddijn, and J.-P. Martens. 2000. "Orthographic Transcription of the Spoken Dutch Corpus." In *Proceedings of the Second International Conference on Language Resources and Evaluation* (LREC 2000), 909-914. http://lands.let.kun.nl/old/cgn.old/2000_01.pdf.

Groupe ICOR. 2007. *Convention ICOR.* Lyon: Université de Lyon. http://icar.univ-lyon2.fr/documents/ICAR_Conventions_ICOR_2007.doc.

Kipp, M. 2001. "Anvil: A Generic Annotation Tool for Multimodal Dialogue." In *Proceedings of the 7th European Conference on Speech Communication and Technology*, 1367-1370. Eurospeech. http://www.dfki.de/~kipp/public_archive/kipp2001-eurospeech.pdf.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk.* Mahwah, NJ: Erlbaum. http://childes.psy.cmu.edu/manuals/chat.pdf.

Nivre, J., J. Allwood, L. Grönqvist, E. Ahlsén, M. Gunnarsson, J. Hagman, S. Larsson, and S. Sofkova. 1999. *Göteborg Transcription Standard.* Version 6.3. Department of Linguistics, Göteborg University. http://www.ling.gu.se/projekt/tal/doc/transcription_standard.html.

Rehbein, J., T. Schmidt, B. Meyer, F. Watzke, and A. Herkenrath. 2004. "Handbuch für das computergestützte Transkribieren nach HIAT." *Working Papers in Multilingualism* 56: 1–78. http://www.exmaralda.org/files/azm_56.pdf.

Sacks, H., E. Schegloff, and G. Jefferson. 1978. "A Simplest Systematics for the Organization of Turn Taking for Conversation." In *Studies in the Organization of Conversational Interaction*, ed. J. Schenkein, 7–56. New York: Academic Press.

Selting, M., P. Auer, D. Barth-Weingarten, J. Bergmann, P. Bergmann, K. Birkner, E. Couper-Kuhlen, A. Deppermann, P. Gilles, S. Günthner, M. Hartung, F. Kern, C. Mertzlufft, C. Meyer, M. Morek, F. Oberzaucher, J. Peters, U. Quasthoff, W. Schütte, A. Stukenbrock, and S. Uhmann. 2009. "Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)". *Gesprächsforschung* 10: 353–402. http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf.

Schmidt, T. 2005a. *Computergestützte Transkription: Modellierung und Visualisierung gesprochener Sprache mit* texttechnologischen *Mitteln.* Frankfurt: Peter Lang.

Schmidt, T. 2005b. "Time-based Data Models and the Text Encoding Initiative's Guidelines for Transcription of Speech." *Working Papers in Multilingualism* 62: 1–32. http://www.exmaralda.org/files/SFB_AzM62.pdf.

Schmidt, T., S. Duncan, O. Ehmer, J. Hoyt, M. Kipp, M. Magnusson, T. Rose, and H. Sloetjes. 2009. "An Exchange Format for Multimodal Annotations" in *Multimodal Corpora*, ed. M. Kipp, J.-C. Martin, P. Paggio, and D. Heylen, 207–221. Berlin: Springer.

Schmidt, T. and W. Schütte. 2010. "FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction." In *Proceedings of the Seventh International conference on Language Resources and Evaluation* (LREC 2010), 2091–2096. http://www.exmaralda.org/files/LREC_Folker.pdf.

Schmidt, T. and K. Wörner. 2009. "EXMARaLDA: Creating, analysing and sharing spoken language corpora for pragmatic research." *Pragmatics* 19: 565–582.

Wittenburg, P., H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. "Elan: A Professional Framework for Multimodality Research." In *Proceedings of the Fifth International conference on Language Resources and Evaluation* (LREC 2006). http://www.lat-mpi.eu/papers/papers-2006/elan-paper-final.pdf.

APPENDIXES

Appendix: Full example of (unparsed) TEI transcription

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:xs="http://www.w3.org/2001/
XMLSchema">
  <!-- TEI Header is only used in a rudimentary fashion here -->
  <!-- Should be supplemented with additional information -->
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title/>
      </titleStmt>
      <publicationStmt>
        <p/>
      </publicationStmt>
      <sourceDesc>
        <recordingStmt>
          <!-- the recording to which the transcription refers -->
          <!-- it was necessary to introduce an attribute @url here -->
          <!-- so that the actual digital file could be referenced -->
          <recording type="audio" url="./PaulMcCartney.wav"/>
        </recordingStmt>
      </sourceDesc>
    </fileDesc>
    <profileDesc>
      <particDesc>
        <person xml:id="SPK0" sex="1">
          <persName>
            <abbr>DS</abbr>
          </persName>
        </person>
        <person xml:id="SPK1" sex="0">
          <persName>
            <abbr>FB</abbr>
          </persName>
        </person>
      </particDesc>
    </profileDesc>
    <revisionDesc>
      <change when="2011-01-19T13:41:42.515+01:00">
        Created by XSL transformation from an EXMARaLDA basic transcription
      </change>
    </revisionDesc>
  </teiHeader>

  <text>
    <!-- timeline with timepoints used as anchors inside the transcription
-->
    <!-- the absolute times are offsets into the recording specified above
-->
    <timeline unit="s" origin="#T1">
      <when xml:id="T1" absolute="00:00:00"/>
```

```
      <when xml:id="T2" absolute="00:00:00.4"/>
      <when xml:id="T3" absolute="00:00:00.9"/>
      <when xml:id="T4" absolute="00:00:01.4"/>
      <when xml:id="T6" absolute="00:00:02"/>
      <when xml:id="T5" absolute="00:00:02.3"/>
      <when xml:id="T7" absolute="00:00:02.6"/>
      <when xml:id="T0" absolute="00:02:56.96"/>
    </timeline>
    <body>
      <!-- the first segment chain -->
      <div>
        <!-- the transcribed text from the primary tier -->
        <u who="#SPK0">
          <anchor synch="#T1"/>Okay. <anchor synch="#T2"/>Très bien,
          <anchor synch="#T3"/>très bien. <anchor synch="#T4"/>
        </u>
        <!-- additional annotations from a sup (=suprasegmentals) tier -->
        <spanGrp type="sup">
          <span from="#T2" to="#T4">faster</span>
        </spanGrp>
        <!-- additional annotations from an en (=English translation) tier
-->
        <spanGrp type="en">
          <span from="#T1" to="#T2">Okay. </span>
          <span from="#T2" to="#T4">Very good, very good.</span>
        </spanGrp>
      </div>

      <!-- the second segment chain -->
      <div>
        <u who="#SPK1">
          <anchor synch="#T3"/>Alors ça <anchor synch="#T4"/>dépend
((cough))
          <anchor synch="#T6"/>un petit peu. <anchor synch="#T5"/>
        </u>
        <spanGrp type="en">
          <span from="#T3" to="#T5">That depends, then, a little bit</span>
        </spanGrp>
        <spanGrp type="pho">
          <span from="#T6" to="#T5">[ɛ̃tipø:]</span>
        </spanGrp>
      </div>

      <!-- an incident from a nv (=nonverbal) tier describing nonverbal
behaviour -->
      <incident who="#SPK0" type="nv" start="#T3" end="#T6">
        <desc>right hand raised</desc>
      </incident>

      <!-- the third segment chain -->
      <div>
        <u who="#SPK0">
```

```
          <anchor synch="#T5"/>Ah oui? <anchor synch="#T7"/>
        </u>
      </div>

    </body>
  </text>
</TEI>
```

NOTES

1. And the examples in Table 1 are still homogeneous insofar as they are all orthographically (rather than phonetically) transcribed corpora of spontaneous (rather than read or prompted), multi-party (rather than monological) speech. This type of corpus is typically used in conversation analysis and related fields. If we add to the picture spoken language corpora used in speech technology or phonetics and phonology, even more variation in transcription techniques will have to be taken into account. It is doubtful, however, whether a standardisation across such a diverse spectrum of practices is feasible at all. This paper therefore concentrates on the type of spoken language corpora exemplified in Table 1.

2. In a way, CHAT is an exception to this because it is the name both of the data format used by the CLAN tool and of a transcription convention. However, the CHAT format and the CHAT convention can be clearly separated conceptually. Thus, it is possible to use the CHAT format with a different transcription convention and to use the CHAT convention with a different format.

3. It is by no means uncommon to use such tools for transcription. However, the resulting data are more or less unstructured texts, and this lack of explicit structure makes them ill-suited for a standardisation effort.

4. Further tools belonging to the same family are: the TASX annotator, tools from the AG toolkit and WinPitch.

5. The data models can therefore all be understood as special types of annotation graphs as defined by Bird & Liberman (2001).

6. Note that the definition given in the TEI Guidelines for the `<u>` element – "a stretch of speech usually preceded and followed by silence or by a change of speaker" – is compatible with the way it is used here to represent a segment chain. The name "utterance", however, may not be too lucky a choice for this element since some transcription conventions use the same name to denote a much more specific entity of speech (see next section).

7. There are of course many possible alternative representations which also conform to the TEI Guidelines. However, as Schmidt (2005b) and others have repeatedly argued, processing of the data is much facilitated by selecting one option out of the many and disallowing all others. For example, the document in Figure 4 might just as well connect a `<u>` to the timeline by giving it a `@start` and an `@end` attribute. The representation chosen here is not in any way superior or inferior to that alternative, but it is still important to minimise variation by explicitly declaring one alternative as the preferred one.

8. The examples use a selection of the conventions' rules only. Proficient users of the respective conventions may disagree on some details of what is transcribed here and how it is transcribed, and the example is certainly not a realistic one. Remember, though, that the aim here is to *exemplify* some differences between the systems, not to fully and precisely describe them.

9. Implicit markup is printed in bold face here. The symbol ␣ represents a space.

10. E.g. MacWhinney (2000) for CHAT: "Codes, words, and symbols must be used in a consistent manner across transcripts. Ideally, each code should always have a unique meaning independent of the presence of other codes or the particular transcript in which it is located."
11. Since the algorithm relies on the regularities defined in the transcription conventions, any incorrect input (a string violating the convention) should lead to an error in parsing, indicating the non-validity of the input string with respect to the conventions. In the tool described below, such parsing errors will be signalled to the user, and an unparsed TEI version will be produced as output.
12. Solid lines stand for existing conversion routes; dashed lines indicate additional possible conversion routes.

ABSTRACTS

This paper formulates a proposal for standardising spoken language transcription, as practised in conversation analysis, sociolinguistics, dialectology and related fields, with the help of the TEI guidelines. Two areas relevant to standardisation are identified and discussed: first, the macro structure of transcriptions, as embodied in the data models and file formats of transcription tools such as ELAN, Praat or EXMARaLDA; second, the micro structure of transcriptions as embodied in transcription conventions such as CA, HIAT or GAT. A two-step process is described in which first the macro structure is represented in a generic TEI format based on elements defined in the P5 version of the Guidelines. In the second step, character data in this representation is parsed according to the regularities of a transcription convention resulting in a more fine-grained TEI markup which is also based on P5. It is argued that this two step process can, on the one hand, map idiosyncratic differences in tool formats and transcription conventions onto a unified representation. On the other hand, differences motivated by different theoretical decisions can be retained in a manner which still allows a common processing of data from different sources. In order to make the standard usable in practice, a conversion tool—TEI Drop—is presented which uses XSL transformations to carry out the conversion between different tool formats (CHAT, ELAN, EXMARaLDA, FOLKER and Transcriber) and the TEI representation of transcription macro structure (and vice versa) and which also provides methods for parsing the micro structure of transcriptions according to two different transcription conventions (HIAT and cGAT). Using this tool, transcribers can continue to work with software they are familiar with while still producing TEI-conformant transcription files. The paper concludes with a discussion of the work needed in order to establish the proposed standard. It is argued that both tool formats and the TEI guidelines are in a sufficiently mature state to serve as a basis for standardisation. Most work consequently remains in analysing and standardising differences between different transcription conventions.

INDEX

Keywords: digital infrastructures, spoken language, standardization, transcription

AUTHOR

THOMAS SCHMIDT

thomas.schmidt@uni-hamburg.de
Research Centre on Multilingualism, University of Hamburg
Thomas Schmidt holds a PhD in German linguistics and text technology from the University of Dortmund. He currently is a Principal Investigator at the Research Centre on Multilingualism at the University of Hamburg. His research interests include corpus technology, spoken language corpora, and computational lexicography.