Chapter 10

# The challenges of segmenting spoken language

*Nikolaus P. Himmelmann*

## Introduction

The core of a language documentation as conceived of in this book consists of a corpus of audio or video recordings of more or less naturally occurring communicative events with annotations and commentary. As already discussed in Chapter 9, the most basic form of annotation is a transcription of the linguistic utterances contained in the recording. Transcriptions of spoken language involve a number of decisions regarding the representation of relevant features of the speech event (e.g. the question of whether to use a narrow phonetic transcription or a practical orthography to represent phonological segments). One major decision pertains to the units into which the continuous flow of spoken language is to be segmented.

There are four major segmentation levels for spoken language, two of which are dealt with at length in Chapter 9 and will not be further discussed here. These are (phonetic or phonological) segments and speaker turns, i.e. utterances produced by different speakers (see Sections 2.1–2.3 and 2.6 in Chapter 9, respectively). The present chapter is concerned with the following two segmentation issues:

1. a middle-sized transcription unit, delimited by empty spaces, which represents a basic unit in terms of meaning, grammatical function, or sound structure, typically a morphosyntactic or phonological word.
2. higher-level transcription units, indicated by various kinds of punctuation marks and by the spatial arrangement of larger units on a page (lines, indentation for a new paragraph, etc.), representing a stretch of discourse that coheres in terms of intonation and/or pragmatic import and/or syntactic structure. Typical units of this type include intonation units, clauses, sentences, and paragraphs.

The first level is addressed in the literature on morphology and orthography. Major issues relevant for documentary linguistics are summarized in

Section 1. Our main concern, however, will be with level 2 units because there is very little agreement and much confusion as to how to proceed on this level of segmentation. Section 2 will be devoted to this issue.

Before we take a closer look at level 1 and 2 transcription units, two general remarks are in order. First, transcription practice on all segmentation levels is very strongly influenced by the writing systems for European languages, which evolved over more than two millennia. In reflecting on transcription practices, it will thus be instructive to take a look at writing practices at earlier stages of the development of the modern European systems as well as at the major writing traditions outside Europe (see, for example, Daniels and Bright 1996 or Coulmas 2003). In a classic paper, Ochs (1979) reviews some biases inherent in the European writing tradition which may adversely affect the analysis when uncritically adopted in transcribing spoken interactions.

Second, if you happen to be able to work with native speakers who are literate in a dominant language and may thus be able to work independently on transcriptions, it will be very instructive to document such independent transcriptions as primary data. In the initial phase, the transcripts may often be difficult to interpret because they appear to be full of inconsistencies and lack the indication of higher level units (transcripts can go on for pages without a single punctuation mark or indentation to show the beginning of a new unit). Over time and usually influenced strongly by the practices of the researcher(s) or the dominant writing culture, a more consistent and "orderly" set of transcription practices may emerge which in turn may feed directly into an emerging literacy in the speech community. Documenting this process will be of great interest for many reasons, including the fact that such transcripts may provide independent evidence for native speaker intuitions about segmentation units such as words or sentences.

## 1. Segmenting 'words'

It is a matter of controversy whether and to what extent the 'word' is a basic structural unit in all languages. There are also differing reports as to whether native speakers have intuitive knowledge regarding word boundaries. In many literate societies, native speakers have relatively clear ideas about wordhood, but their perception of word boundaries is largely based on the orthographic conventions familiar to them (a word is 'what one writes between spaces'). In many non-literate societies, speakers are also able to segment utterances into form-meaning pairings of word-like sizes (as when

asked to 'dictate' an utterance to a researcher not yet familiar with the language). The consistency with which such segmentation is performed, however, varies greatly between individual speakers and speech communities, depending in part at least on the overall structure of the language. Thus, segment size in 'dictation' (i.e. speaking slowly and very articulately for the benefit of an outsider) may vary between a syllable or a (metrical) foot and a phrase. In a similar way, historically evolved conventional orthographies often show considerable variation and inconsistency in indicating word boundaries (compare, for example, English *blackfish* with *black snake* (with initial stress) or *cannot* with *may not*).

However, it would be wrong to conclude from the inconsistencies observed in many orthographies as well as in native speaker behavior that variation here is totally arbitrary and that 'word' is not a useful unit, having no cognitive validity whatsoever for speakers in non-literate communities. Instead, it is important to note that variation and inconsistency in delimiting word boundaries pertains to a well-known set of phenomena, most importantly compounds such as *blackfish* and *black snake*, clitics (e.g. /nt/ in English *shouldn't*), particle constructions such as English *put off*, and lexicalized phrases (e.g. *forget-me-not, whatsoever*, *kick the bucket*). Disregarding these problem areas, it probably holds true that speakers of all languages have clear intuitions about "smallest, completely satisfying bits of isolated 'meaning' into which the sentence resolves itself," as Sapir (1921: 34) put it. Thus, there never seems to be any doubt about the fact that clear affixes such as *-ing* in English *sing-ing* are part of a single word form *singing*. And conversely, there is no doubt about the fact that a unit such as *book on the table* is phrasal, consisting at least of two words (*book* and *table*), while the wordhood of *on* and *the* may be less clear.

Consequently, native speaker input will provide the major source for segmenting continuous discourse into word-sized chunks. In the problem areas, however, it will in general not be possible to rely exclusively on this input. Rather, it will be necessary to devise a set of criteria to be adhered to when segmenting units involving clitics, compounds, and the like. Before we turn to these, it will be worth emphasizing a point already made at the end of the preceding section. A documentation should include clear evidence as to how native speakers handle word boundaries, both in the clear and the unclear cases. This may be done by including recordings of acts of 'dictation' (for example, recording a transcription session where the native speaker listens to a previously made recording and dictates it in workable chunks to the transcriber) or by including specimens of unedited transcrip-

tions in those instances where speakers are able to provide these themselves (usually based on the literary skills acquired for a dominant language).

As for the problem areas, it will be useful to distinguish two separate, though clearly interrelated issues: problems of analysis and questions of orthographic representation. Problems of analysis are widely discussed in the morphological literature, both in textbooks and specialist work (see, for example, Matthews 1991: 206–222; Basbøll 2000; Haspelmath 2002: 148–162; and the contributions in Dixon and Aikhenvald 2002). Here it will suffice briefly to introduce the basic issue and some useful terminology.

In most languages, there are different criteria for defining words and these criteria can be in conflict with each other. Major conflicts often arise between phonological and morphosyntactic criteria for defining words, giving rise to two different 'types' of words, i.e. the *phonological word* and the *morphosyntactic* (or *grammatical*) *word (form)*. Thus, for example, English *shouldn't* is clearly a single phonological word as seen by the fact that it carries only one stress and /nt/ does not fulfill the phonotactic requirements of a minimal word form in English (among other things, an English word has to have at least one vowel). But *shouldn't* clearly also comprises two morphosyntactic words as seen by the fact that it consists of two constituents which are separable from each other (as in *Why should you not apply?*).

In those instances where the phonological and morphosyntactic criteria define units of different sizes – a common but by no means universal occurrence – all possible interrelationships of the units thus defined are attested: A phonological word may comprise two or more morphosyntactic words (as in the case of English *should=n't*). Conversely, a morphosyntactic word may comprise two or more phonological words. Apart from the long morphosyntactic words found in polysynthetic languages, this is also common in some types of reduplication which involve the complete lexical base (or a significant part of it) as in Malay *rumah-rumah* 'houses'. One reason for considering this form as two phonological words is that /hr/ is a consonant cluster otherwise not attested in Malay phonological words. Finally, Dixon and Aikhenvald (2002: 29f.) report two instances where some phonological words consists of one morphosyntactic word plus part of a second morphosyntactic word, that is, the formation of phonological words here "ignores" morphosyntactic word boundaries.

While best known, conflicts in determining wordhood do not only arise from the application of criteria at two different levels, phonological and morphosyntactic. They may also arise by the application of different criteria

at the same level. That is, two phonological features or rules may not target the same unit, giving rise to two types of phonologically defined words (and similarly for morphosyntactic words). Woodbury (2002: 91–97) provides an example from Cup'ik.[1]

Turning now briefly to the issue of orthographic representation, it is a widely accepted and used practice to write items which clearly are single words as separate items delimited by spaces on either side and not to use any further means of orthographically indicating wordhood. As for problematic items such as compounds, clitics, and lexicalized phrases, the western writing tradition offers essentially three options for representing these orthographically. One may write problematic items as single units as in *shouldn't, blackfish*, or *whatsoever*, thus emphasizing their wordhood but obscuring their constituency. Or one may write them separately as in *black snake* and *kick the bucket*, thus making their constituents and original phrasal structure more easily recognizable but also rendering them orthographically indistinguishable from productively formed (compositional) phrases. Finally, one may write them with a hyphen as in *forget-me-not* in an attempt to convey both word-like coherence and phrasal transparency.[2]

No widely accepted principles or practices exist as to how to represent the typical problem cases. Both conventional writing systems and practical orthographies developed by descriptive linguists differ widely in this regard. Thus, while in English noun-noun compounds such as *clothes peg* are often written apart, in German they are regularly written as a unit (*Wäscheklammer*). Similarly, in the Northern Philippine language Iloko enclitics are regularly written together with the preceding word as in *Surátemon!* (suráten=mo=en 'write =2SG=now') 'Write it!' (Rubino 2005: 334), while in Tagalog, a Central Philippine language, clitics are generally written as separate items, hence *Isulat mo na!* 'Write it!'.

Sometimes there are good reasons for either option. In the Philippine case, for example, Iloko clitics tend to fuse with their hosts to a much larger extent than Tagalog clitics, which mostly appear in the same shape regardless of the host. Hence, writing the Iloko clitics together with their hosts provides for an orthographic representation of (phonological) words which is close to their actual articulation. But very often there are conflicting motivations for both options which are difficult, if not impossible, to resolve in a totally consistent and systematic fashion. A good example for this state of affairs is provided by the lively debate concerning the principles of orthographic wordhood in German which has accompanied the development of the modern German writing system from its beginnings and continues to be

a matter of considerable controversy. Thus, this issue is, once again, one of the most contested aspects of the last orthography reform in German writing countries (see Jacobs 2005 for a recent attempt to resolve the problems in a principled manner).

In dealing with cases of problematic orthographic wordhood, it will be useful to keep the following considerations in mind:

- Issues of orthographic representation usually have to be resolved by taking into account non-linguistic factors such as learnability or already established neighboring orthographies, as discussed in detail in Chapter 11. Of course, the (practical) orthography used in transcriptions does not have to be identical to the practical orthography used in, or developed for, the community. But in most instances it will not be feasible to use two practical orthographies in parallel. Hence, the non-linguistic factors will also play a role for the orthography used in transcription.
- While in writing no major difference exists, in reading it appears to be easier to process shorter simplex units which have to be combined into a larger unit (as when one has to determine that *clothes peg* is a compound and not a phrase) than to break down longer complex units into their constituent parts (as in the case of Iloko *surátemon*). Note that this 'principle' is contravened by the principle that whatever clearly forms a single, phonological *and* grammatical word should be written together. Hence, there are no orthographies which write clear affixes consistently as separate items.[3]
- It is a widespread, though by no means universal, practice to base orthographic wordhood on the criteria for the grammatical word wherever phonological and grammatical wordhood are in conflict. For example, clitics are widely represented as orthographically independent items. However, there may be indications for the opposite option, e.g. when clitics show fusional tendencies (as in the Iloko example above) or when particles are separable from the verb with which they form a grammatical unit (cp. *to put off the meeting* vs. *to put it off*).

## 2. Intonation units, 'paragraphs' and more

The segmentation of continuous spoken discourse at levels higher than the orthographic word is rarely, if ever explicitly, addressed in descriptive linguistics. That is, it usually remains a mystery as to how exactly the author(s) arrived at the format of a transcript published in a text collection or in the

appendix of a grammar. Most transcripts are presented with sentence and paragraph structure, with standard punctuation (commas, full stops, indenting) indicating major units. But with few exceptions (for example, Heath 1980: 2–5 [see also Heath 1984: 589–619] or Himmelmann and Wolff 1999: 83, 98f.), the authors usually remain silent as to how the various boundaries implied by these marks have actually been determined.

If one happens to have access to the original recording underlying the published transcript, one will almost immediately notice that in fact quite a lot of editing and interpretation is involved in arriving at the "clean" published form. False starts, repetitions, and hesitations ('uhm' and the like) are usually edited out. Decisions as to what to include in a single clause and sentence are usually based on semantics and, if available, morphosyntactic evidence. But more often than not, such decisions are also influenced by what a sentence in written English looks like (or whatever written language the editor is most familiar with). Given this mixture of variables, many of which are difficult to handle in a consistent manner, it is almost unavoidable that decisions regarding sentence and paragraph structure become almost arbitrary. It is thus highly unlikely that two editors working in this way with the same recording and the same speaker would arrive at a reasonably similar "clean" transcript for publication (to my knowledge, no experiment along these lines has been conducted so far, but it seems reasonably safe to predict this outcome).[4]

The importance of the (edited) transcript resides in the fact that for most analytical procedures (in particular in morphosyntax and semantics but also in phonology) it is the transcript (and not the original recording) which serves as the basis for further analyses. Obviously, whatever mistakes or inconsistencies have been included in the transcript will be carried on to these other levels of analysis, perhaps not always causing major harm but clearly introducing unknown variables into these further analyses. This problem may become somewhat less important in the near future inasmuch as it will become standard practice to link transcripts line by line (or some other unit) to the recordings, which allows direct and fast access to the original recording whenever use is made of a given segment in the transcript.

Nevertheless, even with transcripts linked to the recording, one still has to decide on some higher-level unit into which the flow of spoken discourse is to be segmented. As opposed to descriptive linguistics, such segmentation has been a major concern in anthropological linguistics and in (some variants of) discourse analysis, and we will heavily draw on this work in the remainder of this section.[5]

Work in anthropological linguistics such as Tedlock (1983) or Sherzer (1990, 1992) has focused on verbal art where segmentation units above the word such as verse/line, couplet, or stanza tend to be indicated by a host of prosodic, lexical, and grammatical features. The variants of discourse analysis of interest here have mostly been based on everyday speech, mostly narratives and conversation. The basic higher-level segmentation unit identified in most of this work is the *intonation unit* (also known as *tone group*, *breath group*, *intonational phrase*, and the like).[6] The intonation unit roughly corresponds to the line (or verse) in verbal art genres. It is widely held to be the basic unit into which native speakers themselves chunk their utterances, i.e. it is seen as a unit of speech production which in some sense has a psychological reality for the speakers (as opposed to a purely analytic construct "invented" by linguists). In this section, we will first take a closer look at how intonation units can be identified and then briefly discuss the possibilities of identifying even larger units above the intonation unit.

## 2.1. Identifying intonation units

In most languages, evidence for intonation unit boundaries is provided by changes in pitch and rhythm. Evidence from pitch is of three kinds: a) the occurrence of a boundary tone at the end of an intonation unit, i.e. a clearly perceptible change in pitch on the last syllable(s) of a unit; b) a new onset at the beginning of the next unit, i.e. typically a jump in pitch between the offset of one unit and the beginning of the next one; and c) a reset of the baseline which is most clearly visible in the fact that early pitch peaks in the new unit are higher than the final pitch peaks in the preceding one. Major rhythmic evidence is also of three kinds: a) a pause in between two units; b) lengthening of the final segment of a given unit; c) anacrusis, i.e. an accelerated delivery of the unstressed syllables at the beginning of the new unit.[7]

It is rarely the case that all the diagnostics for a boundary listed above can actually be observed at a given boundary in spontaneous speech. In fact, most of the diagnostics are optional, i.e. they do not have to occur at every boundary. Only two diagnostics, i.e. the final boundary tone and the new onset, are, in theory at least, obligatory in many languages.[8] But in spontaneous speech, there are various factors which may make it difficult or impossible to identify relevant phenomena in a given instance (more on these shortly). Nevertheless, at least two or three of the diagnostics will be

present at a given boundary in most instances. That is, between 80–90% of the intonation unit boundaries occurring in spontaneous speech are relatively easily and clearly identifiable, although there may be considerable variation across speakers and genres (boundaries in monological speech are generally easier to identify, in part simply because there is only minimal interference from other speakers).

In practical terms, the two most common and useful diagnostics for boundaries are the final boundary tone and pauses, both of which, however, are not always straightforwardly identifiable. As for pauses, the major problem lies in the fact that not all pauses occur at the boundary of an intonation unit but some types of pauses – widely known as *hesitation pauses* – also occur within intonation units. Some hesitation pauses are easily distinguished from boundary marking pauses by the fact that they involve a rather abrupt stoppage in the stream of speech which often ends in a glottal stop. They often also include some kind of filler (*uhm* and the like) and may be followed by further disfluencies as in *but uhm (0.2) the the sound.* Pauses at intonation unit boundaries, on the other hand, are characterized by complete silence, the audible relaxation of the vocal organs, audible exhalation, and/or an audible breath intake. Apart from hesitation pauses and boundary pauses, a third type of pause needs to be distinguished, namely rhetorical pauses. These may occur as part of a package of features used to put particular emphasis on a given item, as in *That is the most* [pause] *ludicrous idea I have ever heard.* These are much rarer than the other types of pauses and usually are easily distinguishable from them because of other contextual features which signal special emphasis.

As for final boundary tones, these are often only clearly identifiable if the unit ends on one or more unstressed syllables. If the unit ends on a stressed syllable, it may be difficult to distinguish between a pitch change related to stress and a pitch change related to the boundary. A second problem regarding final boundary tones pertains to the fact that more often than not, the inventory of boundary tones in a given language contains a default member which is characterized by the lack of a major pitch excursion, the unit typically ending somewhere in the non-descript middle of a speaker's pitch range. Such instances may be difficult to distinguish from hesitations. And finally, the voice at the end of a unit may become creaky and/or very low in intensity so that actually occurring pitch changes may become hardly perceptible (this, of course, is also the case when actually occurring pitch changes are masked by co-occurring noise such as overlap from another speaker, laughter, etc.).

The following example from a spontaneous English narrative[9] illustrates some of the features of intonation units mentioned above (see also Figure 1). In the first unit, you can hear a brief hesitation pause where the speaker audibly does not release the vocal organs right after the initial *and*, which is a very typical place for hesitation pauses to occur. With regard to pitch, the unit ends somewhere in mid range without a clear rise or fall, which is indicated here with a semicolon (;). The second unit ends on a clear rise which, however, occurs on a stressed syllable and hence combines characteristics of an accentual tone and a rising boundary tone (rising boundary tones are marked by a slash /). In the last unit, on the other hand, the final (rising) accent tone is on *strong*, which is followed by a clear fall to the lower bottom of the speaker's pitch range (170–180 Hz in this story), a very clear example of a falling final boundary tone (final falls are indicated by a backslash). The numbers in parentheses indicate pause length in seconds.[10] In contrast to the hesitation pause at the beginning of the first unit, these pauses are completely silent. Note, finally, that the speaker starts each unit in the lower mid of her pitch range (around 230 Hz), which in each instance involves a jump up or down from the pitch target reached at the end of the preceding unit (new onset).

(1) PEAR-L-36FF

36. and (0.4) you see his hand sometimes at close up ; (1.1)
37. uh snatching the pears from the tree / (0.8)
38. and you hear the sound really: strongly \ (0.8)

The following example is a bit more complicated and illustrates two of the most common difficulties that may occur in determining intonation unit boundaries. These are false starts/self-repairs, as in units 49–51 of the following example, and latching, i.e. two units occur in immediate succession, without an audible break intervening, which is indicated by an equal sign in parentheses (=) instead of a pause duration at the end of lines 49–51:

(2) PEAR-L-48FF

48. he climbs down the ladder / (0.5)
49. and he puts a couple of the pears– (=)
50. well: (=)
51. as he's standing there ; (=)
52. couple of the pears fall \ (0.4)

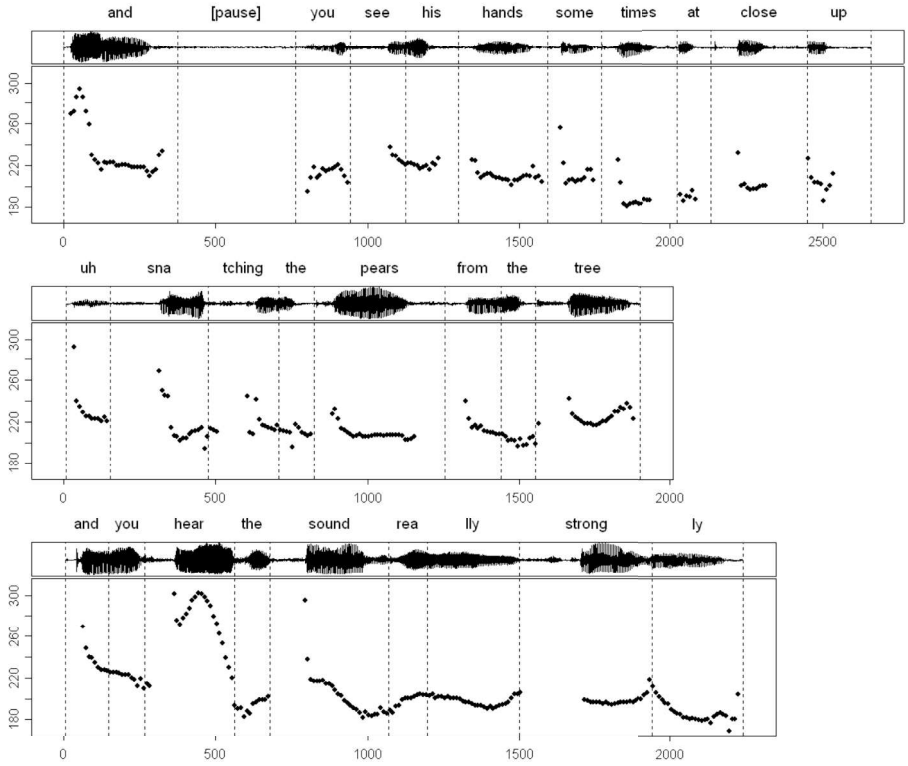


*Figure 1.* Waveform and fundamental frequency[11] for example (1)

Going briefly through this example line by line (see Figure 2), the intonation unit in line 48 is very easily identifiable since it ends on a clearly identifiable boundary tone (strong rise on the final unstressed syllable of *ladder* followed by a boundary pause with audible breath intake). Unit 49 illustrates the phenomenon of self-repair where the speaker interrupts herself as she starts pronouncing the final fricative of *pears*, breaks off before finishing this segment (signaled by a dash –), and immediately restarts in mid range with a slightly lengthened *well* (lengthening is indicated by the colon :), which here functions as a lexical repair marker. Then she immediately starts the repair unit (51) which ends on a clear fall across the final two syllables (*ing there*). This fall, however, does not reach the bottom of her range (it ends around 195 Hz) and is therefore marked here by a semicolon. The final unit again starts without an audible pause preceding it. This unit ends on a fall on the final (stressed) syllable to the bottom of her pitch range.

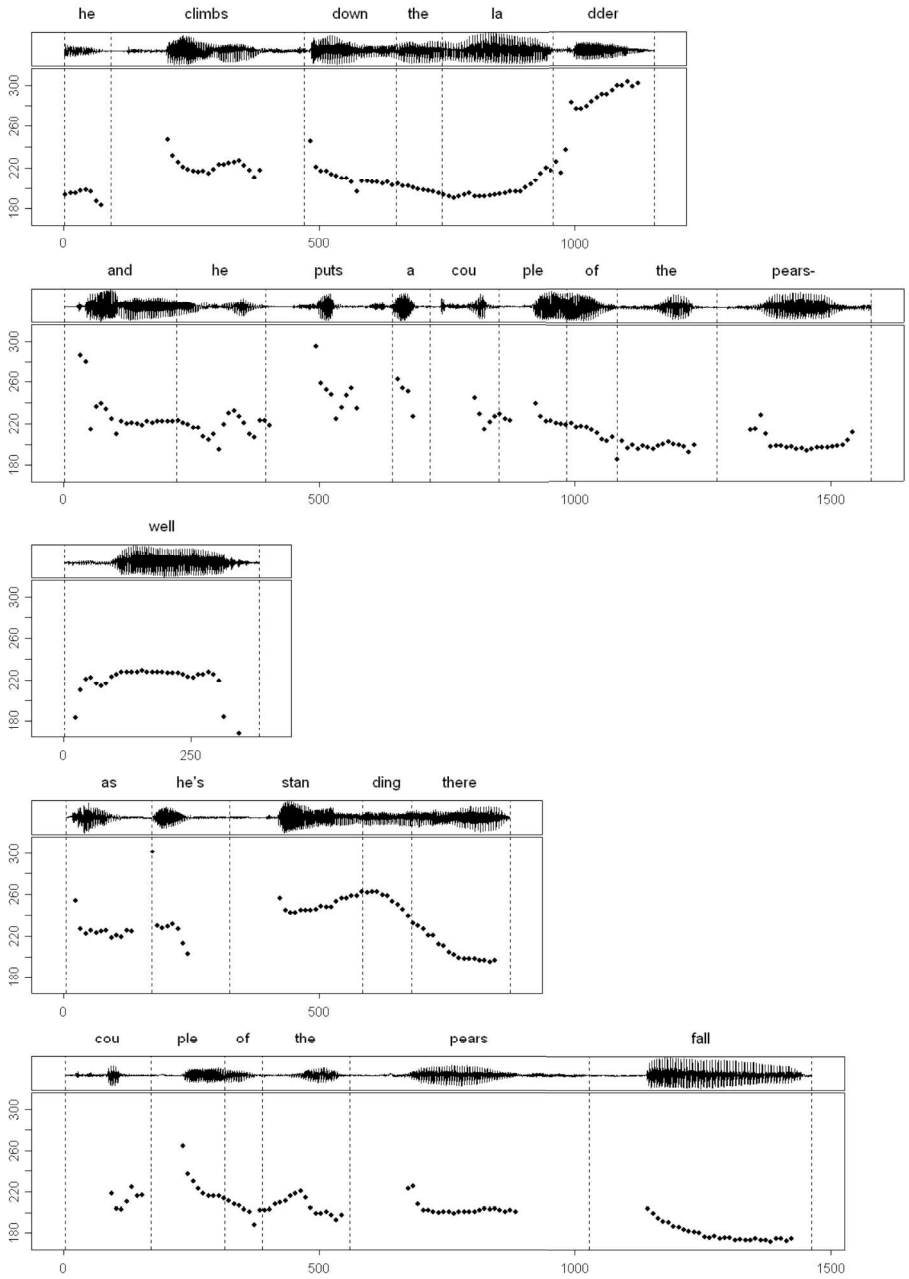


*Figure 2.* Waveform and fundamental frequency for example (2)

Latching as in units 49–51 often causes some problems in that the other indicators for intonation unit boundaries become then all important. Thus, e.g., at the end of unit 51 there is a clear fall across two unstressed syllables, which is interpreted here as a boundary tone. But, importantly, unit 52 does not start with a clearly new onset of pitch. Instead, the pitch continues without any audible interruption. Hence, the only reason for assuming a boundary between 51 and 52 is the fall at the end of 51.

Self-repairs are often easily recognizable by the abrupt break-off of the word under way. They are more difficult to identify when the break-off occurs after the word or construction currently under way has been finished. In such instances they may be difficult to distinguish from intonation units that do not end on a clearly identifiable boundary tone.

Lexical repair markers such as *well* in unit 50 and other kinds of so-called discourse markers such as *and then, you know, I think, let me see* pose a minor practical problem in that it is often not clear whether they should be considered intonation units of their own (as in unit 50 above) or whether they are part of the preceding or following unit (that is, in the example above units 50 and 51 could also be combined into a single unit: *well: as he is standing there*). The prosodic evidence for either option is often not very clear. In the case of tags as in *and he sort of slips, you know* the prosody can actually be somewhat complicated in that there may be clear indications for the end of an intonation unit before the tag but no evidence for a new onset on the tag. However, for the practical purposes of a base transcript in a language documentation nothing much depends on how these elements are represented. As usual, the main concern here should be with consistency, i.e. to put them all in units of their own or to include them in the unit they appear to belong to (in a few instances it may not be a straightforward exercise to determine whether this is the preceding or following unit).

In this regard, it may also be noted that coordinating and subordinating conjunctions in many languages allow three options of prosodic packaging. They may either occur together with the second conjunct (see unit 49 in (2) above) or the subordinate clause they introduce, as in:

(3) he didn't notice / (0.3)
*because* he was busy picking pears \

Or they may occur at the end of the first conjunct or the matrix clause, as in:

(4) he didn't notice *because* / (0.3)
he was busy picking pears \

The third alternative is to have them form an intonation unit of their own:

(5) he didn't notice / (0.3)
*because* ; (0.7)
he was busy picking pears \

In this last case, there will often be no clear-cut boundary tone at the end of the intermediate intonation unit. Arguably, instances such as (5) can often also be analyzed as instances of (3), i.e. as a single intonation unit with a hesitation pause following the initial word or phrase: *because (0.7) he was busy picking pears*.

As a general rule of thumb, it may be of help to remember that intonation units are in some sense planning units for the speaker and rarely include more than 5–7 content words (2–3 words in highly polysynthetic languages). In fact, it has been suggested by Chafe (1994; see also Pawley and Syder 2000) that each intonation unit contains only a single bit of new information (which is also known as the *one-new-idea-at-a-time hypothesis*). Thus, with regard to spontaneous speech, overly long intonation units making reference to several new participants or activities not mentioned before should be regarded with some suspicion. This rule of thumb, however, does *not* hold true for more ritualized forms of speech which often contain large formulaic chunks that have been memorized. Similarly, units containing quoted direct speech are often significantly longer than the average intonation unit in a given speech event.

The planning load to be managed by the speaker also manifests itself in the following phenomenon widely observed in spontaneous monologic speech (in particular narratives of various types but also procedural texts). At the beginning of a narrative or similar genre, there tend to be lots of hesitations and false starts as well as a mixture of longish and very short intonation units, while later on, the delivery will become more fluent and rhythmically spaced. This is probably due to the fact that at the beginning of an extended monologue speakers have to deal with a higher planning load, since apart from putting together individual intonation units, they also have to develop and implement an overall plan for the delivery of their story. In terms of transcription and segmentation, this means that identifying intonation units at the beginning of a monologue is often more difficult and cumbersome than later on, and it may be a good idea to start the segmentation of a narrative a minute or two into the telling and turn to the beginning only after the rest of the recording has been dealt with.

A somewhat different problem pertains to the fact that when transcribing spontaneous speech in a language one understands very well, there is a strong tendency for semantic and syntactic factors to interfere with one's perception of prosodic boundaries. That is, indications for prosodic boundaries within clauses or noun phrases tend to be missed and, conversely, there is a tendency to hear prosodic boundary signals at, e.g., clause boundaries when in fact there are none. A well-known example for these tendencies is the fact that clause-internal pauses are often not perceived and at the same time, pauses are "heard" at clause boundaries when according to the instrumental evidence there aren't any. It is, therefore, important to control for these interferences by instrumentally crosschecking a sample of the boundaries marked auditorily (checking all boundaries acoustically will normally not be feasible because it would be too time consuming). Otherwise, one ends up with boundaries based on a mixture of prosodic, semantic, and syntactic criteria which tend to lead to irresolvable inconsistencies.

Note in this regard that the diagnostics listed above in part pertain to offset phenomena and in part to onset phenomena. In almost all instances, these two align in the sense that where there is an offset, there is also an onset. However, this need not be the case. Speakers may choose to start a new unit, providing all the evidence for new units (most importantly, a new onset), without having properly finished the preceding one (which then remains a fragment). Furthermore, and this is even less common, they may also construct a new unit as a continuation of the preceding one although the preceding unit was in fact "properly closed". This latter case is illustrated in the following example from the same Pear Story:

(6) PEAR-L-88FF

88. because he looks Hispanic \ (0.7)
89. probably a Mexican: ; (1.3)
90. worker being exploited by some landlord / (1.5) ((laughs))

The unit of interest here begins with line 89. While there is no clear final boundary tone at the end of this line, the final *n* of *Mexican* is lengthened (about 200 ms) and followed by a long pause with audible breath intake, both being clear indications of an intonation unit boundary. However, the first word of line 90, *worker*, is produced as if it were a direct continuation of the preceding unit. There are no indications whatsoever for a new onset. On the contrary, the pitch of the first syllable continues very precisely the pitch of the final *n* of *Mexican*, which is quite remarkable given the long

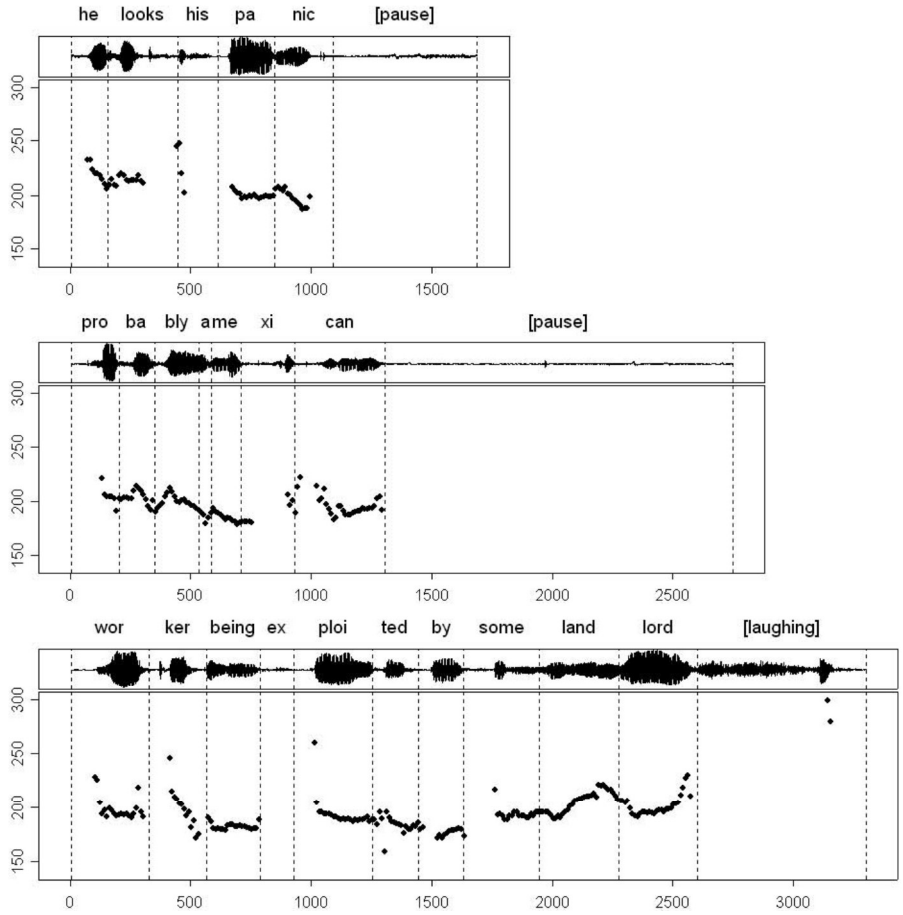


*Figure 3.* Waveform and fundamental frequency for example (6)

pause in between.[12] Note that in the transcript in (6), no attempt has been made to capture this very special relation between the two units, which arguably could also be considered a single intonation unit. It would appear to be of such rare occurrence that it is not feasible to introduce special conventions for this case.

The ability to identify intonation unit boundaries auditorily needs some practice, and it is a highly recommended exercise for anyone planning to undertake a language documentation to transcribe a number of recordings of spontaneous speech in his or her own language (both monologues and

conversations) in order to get a "feeling" for working with spoken language and also for the amount of work and time involved in transcribing it. The level of detail to which features of spoken language are included in a transcript varies significantly across various transcriptions conventions (see Edwards and Lampert 1993 for a survey). The conventions used in the transcription examples given above are loosely based on the ones proposed by DuBois et al. (1993), which are fairly simple and widely used in spoken discourse research.

Since the transcripts included in a language documentation are only intended to provide a starting point for further analysis in different frameworks, it is recommended to be rather sparse with regard to the inclusion of such features as voice quality, speech tempo, laughter, and so on. Pauses will in general not be measured instrumentally but simply indicated by some convention such as a (.) = short pause and (..) = longer pause. The number of boundary tones distinguished should also be restricted to an easily manageable number. In the conventions used above, the only differences indicated are: clear rise (/), clear final fall to the bottom of the speaker's range (\), and everything else (;), which includes falls to lower mid range as well as level ending units. More detailed annotation schemes will inevitably increase the number of problematic decisions to be made and, in the case of boundary tones, a more detailed schema will normally only make sense when the phonological structure of the intonation unit has been analyzed in detail.

While the conventions used in spoken discourse research may thus be a bit too detailed and cumbersome for the amount of transcription involved in a language documentation and should be further simplified along the lines just indicated, it is highly recommended to include all kinds of hesitations and false starts in a base transcript since these may prove to be crucial for various interpretative and analytical tasks. Omitting hesitations and false starts from transcripts can in fact lead to major errors of analysis. In Tolai, for example,[13] one may get the impression from heavily edited transcripts that the form of the article is *a* for subjects and *ra* for objects, thus involving a case-like distinction in grammatical relation marking. However, listening closely to spontaneous speech and preparing adequate transcripts makes it clear that this alternation has nothing to do with grammatical relation marking but pertains to pausing: *a* is the form of the article after a pause (and at sentence boundaries) while *ra* is used when no pause precedes. This becomes obvious when transcripts include all pauses, making it clear that *a* is also used before objects provided a pause precedes.

Furthermore, repair strategies may yield important evidence for morphosyntactic structure in that they generally target morphosyntactic units rather than some arbitrary number of syllables or segments. Thus, e.g., some types of self-repair recycle the complete word, phrase, or clause that the speaker abandoned before completing it and thus provide evidence for the viability of these structural units, as seen in the following example (again from the Pear Story):

(7) I assume <this take pla-> this is taking place in California ; (0.3)

Here the speaker begins a complement clause (*this take pla-*), breaks off half-way into the word *place* and then restarts at the beginning of the complement clause. See Marandin and de Fornel (1996), Fox et al. (1996), and Apothéloz and Zay (1999) for further discussion and exemplification.

## 2.2. Evidence for paragraphs/episodes

Spoken discourse does not simply consist of a sequence of intonation units. Instead, when listening to a coherent stretch of spoken discourse, it is quite clear that some intonation units "belong closer together" than others, forming units larger than a single intonation units. The nature of these units and the boundaries separating them is not yet well understood, and there is a large variety of terms in use for referring to them, including *paragraph, (spoken* or *prosodic) sentence, episode, utterance, intonation unit complex*, etc. (these terms have various readings and, depending on the framework, may refer to units of different sizes).

To date it remains unclear as to whether speakers of unwritten languages have strong and clear intuitions about these units. I am not aware of any reports concerning such intuitions in the literature, and the issue does not seem to have been investigated systematically. Reports by experienced fieldworkers provide conflicting evidence. According to some reports, there are native speakers who are very consistent in marking something which can be called a 'sentence boundary'. Other fieldworkers have quite the opposite experience of speakers producing transcripts and written texts which go on for pages without a single indication of sentence or paragraph structure (I myself belong to the latter group).

Note that the issue here is not 'clausehood'. Speakers often have reasonably clear and consistent intuitions about the fact that a (finite) verb forms

some sort of unit with its arguments and at least some of the more peripheral adjuncts.[14] The issue here pertains to intuitions about which clauses together form larger sentence-like units, including both what from a grammarian's point of view are main and subordinate clauses. To give just one example for possibly conflicting evidence in this regard, in languages which allow for extended chains of subordinated or nominalized clause constructions such as the converb constructions found in Turkic or Papuan languages, some speakers will accept or even propose major boundaries at points within the chain which grammatically speaking are sentence-medial forms.

It may thus be the case that with regard to higher-level segmentation, in at least some languages the native speaker's position is not very different from that of a non-native researcher. It is in fact likely that both draw on the same kind of evidence when attempting to determine the boundaries of higher-level units. In the rare instances where it is explicitly discussed, the evidence for such boundaries usually involves a mixture of semantic, pragmatic, and prosodic factors. Semantic-pragmatic criteria include, for example, changes relating to time and space of the setting (*the next morning*, *arriving at the river*) and a change of topic or subject. The most important prosodic phenomena occurring at such higher-level boundaries are: a) a boundary tone signaling finality (usually a strong fall to the lower bottom of the speaker's range); b) long pauses, i.e. pauses that are distinctly longer than the pauses occurring at the end of a paragraph-internal boundary (this appears to hold statistically when comparing pause lengths across a sufficiently large corpus, but is of little help in making decisions in individual instances); c) reset in declination, i.e. the baseline reaches its absolute minimum at the end of a paragraph and the new paragraph starts with a higher baseline as seen in the level of onsets and low and high tonal targets; d) a particular pitch pattern at the beginning of the unit, often associated with some special lexical expression introducing a new paragraph (something like *after this happened ...*).

As usual when applying a fairly heterogeneous set of diagnostics, there are many instances where these diagnostics provide conflicting evidence, some (a final fall and a long pause, for example) indicating a major boundary, others (no topic change, continued declination) indicating continuity. To date, there is no agreement as to how to resolve such conflicts.

In working on transcripts, there are three points to keep in mind. First, for many analytical procedures higher level boundaries are irrelevant (obviously, they are not irrelevant when looking at conjunctions, discourse markers, and the like). Hence, in many instances it may be preferable not to

indicate any such boundaries rather than marking them in a haphazard and unsystematic way. Second, if one decides to indicate such boundaries, consistency is of paramount importance which is usually helped by explicitly listing the diagnostics and their relative rank. Finally, it is important to keep in mind that units in spoken language are often quite different from those in written language. For example, taking final falls as a major diagnostic, it is not uncommon that units thus delimited in German or English narrative are of extremely varied size. That is, a very long paragraph consisting of 37 intonation units may be followed by another one which consists just of one intonation unit, the next one comprising ten intonation units, and so on.

From these remarks and observations, it follows that for reasons of time economy it will in general not be feasible to attempt a systematic segmentation into higher level units of all recordings when working on transcriptions within a language documentation project. Obviously, whenever there are clear indications for such higher-level structure, these should be explicitly noted and commented upon. Furthermore, it will be useful to document the various segmentation stages applied to those texts which have been chosen for publication and are edited both by native speakers and researchers in the process.

## 3. Conclusion

This chapter has surveyed two major segmentation issues in transcribing spoken discourse. With regard to segmenting words, the primary source of information will be native speaker intuition which, however, has to be supplemented by an explicit convention for transcribing problematic items such as clitics, compounds, and lexicalized phrases. This convention will be based on phonological and morphosyntactic criteria for wordhood, but will also have to take into account non-linguistic factors in deciding on the representation of problematic items. The segmentation into intonation units, on the other hand, will be based primarily on auditory impression, listening for the boundary signals produced by the speaker. The auditory impression should be repeatedly checked acoustically (instrumentally) in order to contravene biases introduced by the semantics and pragmatics of the utterances transcribed or, in the case of a non-native speaker doing the transcription, by one's native prosodic system, which may be tuned to a somewhat different set of boundary signals. Depending on the amount of recordings to be processed within a documentation project, segmentation at levels higher than

the intonation unit will often not be feasible for reasons of time economy. However, inasmuch as native speakers themselves indicate such higher-level segments, these should of course be preserved as part of the annotations stored with the recording of a given event.

**Acknowledgements**

I am grateful to my co-editors and Eva Schultze-Berndt for useful discussion and comments on an earlier version of this chapter. Special thanks to Jan Strunk for preparing the figures and to Louisa Schaefer for help with the Pear Story data.

**Notes**

1. See also the work on word domains done in the AUTOTYP framework (http://www.uni-leipzig.de/~autotyp).
2. In principle there is an almost limitless number of further possibilities for indicating different types of words (word-like coherence) by using additional symbols in place of a hyphen, thus having complex words with '&' (*fair&play*), ones with '=' (*should=nt*), ones with '_' (*tittle_tattle*), and so on. But there are severe limits on how many of such extra symbols can be used consistently by writers and parsed by readers without constantly checking the conventions. It is probably not by chance that there are few, if any, practical orthographies which have gone beyond the three ways of dealing with wordhood orthographically just mentioned (written together, written with a hyphen, written separately).
3. A possible exception is the Japanese writing system, where lexical elements are represented in Chinese characters (Kanji) while morphological elements which arguably can be considered suffixes are consistently written as orthographically separate items (in Hiragana, one of the two syllabaries). This distinction is often reflected even in Roman transcriptions (using spaces or hyphens).
4. Often native speakers are also involved in the process of editing transcripts of spontaneous speech for publication. They usually tend to prefer very clean forms which are similar in structure and appearance to the forms of written language they are familiar with. See Mosel (2004b) for discussion.
5. See Serzisko (1992) for a thorough review and discussion of the discourse analysis literature on segmenting spoken language.

6. The major alternative is the *turn constructional unit* used in Conversation Analysis which, however, is not easily identifiable on the basis of a simple, all-purpose operational procedure. See Ford et al. (1996) for some discussion.
7. See Chafe (1994), Schuetze-Coburn (1994), Ladd (1996), Cruttenden (1997), or Wennerstrom (2001) for a more detailed discussion of the intonation unit and its boundaries.
8. The major exception here are prototypical lexical tone languages, i.e. languages where (almost) every syllable inherently carries a lexical tone. In such languages, there may be either no boundary tone (as has been claimed, for example, for Yoruba) or the boundary tones interact with the lexical tone of the unit-final syllable, resulting in a modification of this lexical tone (e.g. Chinese or Thai).
9. This and the following segments are from a Pear Story (Chafe 1980) by a female speaker of American English recorded by the author. Thanks to Wallace Chafe for the permission to use the pear film. Wave files containing the segments are available at this book's website.
10. In documentary work, it will in general be neither feasible nor necessary to measure the length of pauses instrumentally. See further below.
11. Fundamental frequency (also known as "F zero") is the acoustic measure for the rate of vibration of the vocal cords when producing voiced sounds. It corresponds quite closely to pitch, which is an auditory/perceptual category. But fundamental frequency and pitch perception may diverge and hence need to be distinguished (see Laver 1994: 450ff., for discussion and exemplification).
12. As noted with regard to the transition from unit 51 to 52 in example (2), continuation of pitch level also occurs in latching. But as soon as there is even just a very short boundary pause, there is typically also a clearly new onset of pitch.
13. Thanks to Ulrike Mosel for providing this example (cp. Mosel 1984: 17).
14. Obviously, the consistency and strength of such intuitions depends in part on the typological profile of a language. In so-called non-configurational and, in particular, in polysynthetic languages, intuitions about which words together form a clause may be less clear and rather similar to the vague ideas about 'sentencehood' reported for some languages with relatively tight and hierarchically organized clause structure.