

Chapter 3. Segmenting the Data

In this chapter, you will segment the data streams you have selected for analysis into units appropriate to the phenomenon of interest. After learning about the units characterizing various kinds of verbal data, you will acquire and prepare your data, segment it, and then move and label it in preparation for coding.

■ Introduction to Segmenting

In verbal data analysis, a stream of language is first segmented and then, in a separate and independent step, each segment is selected and coded. The unit used for segmentation is explicitly defined and often linguistic in nature. As we will discuss later in this chapter, this unit may be the sentence, the clause, the t-unit, the topic, or any other unit appropriate to but distinct from the code that is eventually assigned to it. This approach to segmentation has two distinctive features that set it apart from other analytic techniques.

■ The Role of Rules in Segmenting

The first distinctive characteristic of segmentation in verbal data analysis: The decision about where to segment data is less a matter of judgment and more a matter of rules or structure. Where does the sentence end? Where does the topic change? Where is the end of the clause? While segmentation “errors” can occur—i.e., two people may segment the same string of language in two different ways—such lack of agreement should arise not out of differences in judgment, but out of fatigue or inattention, either of which be easily corrected.

By contrast, in content analysis, which can deal with a wider variety of data, analysts may use a unit of analysis that is not rule-based in nature. In her review of attempts to unitize in content analysis, for example, Neuendorf (2016) found researchers segmenting by topic change, by joke, by instance of violent behavior, and even by instance of cigarette smoking. Neuendorf cautioned that, “whenever researchers or coders are required to identify message units separately from the coding of those units, a unique layer of reliability assessment is in order.” (2016, Kindle Locations 1707-1709). She is suggesting, that is, that an assessment of agreement between researchers should be carried out for segmentation—although she reports that few studies actually do so.

By keeping segmentation rule-based, verbal data analysis eliminates the need for an assessment of agreement at the segmentation stage. In many cases, what starts out as a segmenting choice fraught with the need for analytic judgment can be revised to function in a more rule-based manner. In the study by Morris (2009) cited by Neuendorf (2016), for example, the researcher started out by defining the joke as a set of comments organized around a cohesive target, a definition that required a great deal of analytic judgment. But Morris later switched to a more rule-based definition, defining jokes as comments that elicited audience laughter, one that required very little judgment on the analyst’s part.

If we can identify instances of a phenomena using a set of unproblematic rules, then we can use it for segmentation as discussed in this chapter. If we find, however, that we cannot avoid analytic judgment, it may be more appropriate to treat segmentation as a coding issue as discussed in the next chapter. But even if we decide to go this route, we will still segment our data before coding for the reasons we discussed further below.

■ The Separation of Segmenting and Coding

The second distinctive characteristic of segmenting in verbal data analysis is the way it supports focused decision making by separating segmentation and coding into two independent steps. That is, the analyst first segments the entire data set using some well-defined unit of analysis, and only then goes back to code that data, segment by segment.

By contrast, in most qualitative approaches to coding, segmenting and coding are often combined into a single step: The analyst both selects and codes a single piece of data before going on to select/code the next piece of data. When the segmenting and coding decisions are linked in this way, the analyst must keep in mind both the question of where a topic begins and ends and the question of what the topic is. This is a heavy cognitive burden and can lead to a lot of variability between coders.

Thus the most obvious benefit of segmenting a stream of language independently of coding relates to the desirability of producing coding judgments that are systematic and replicable. By segmenting independently in advance of coding, analysts only face one decision at a time: Where does this segment begin and end? Is this segment an X? An agreement among coders on the second question is more simply defined as coming to the same decision about a given segment. Disagreements are consequently minimized.

A second benefit of segmenting data independently of coding relates to being able to measure the relative distribution of any given code. If a single similarly sized unit is used for segmentation, then we have a general sense of what it means to say that 20% of participants' discourse concerned fashion. If, on the other hand, segmenting is linked to the coding decision, then one excerpt coded as fashion may be a few sentences in length; another may be two pages in length. To say, then, that 20% of such variable excerpts concerned fashion does not convey as clearly how much of a participant's focus was actually on fashion.

A final benefit of segmenting data independently of coding is rhetorical. As rhetoricians, we believe that language does work as well as conveys meaning. For this reason, the linguistic unit we choose for analysis is significant. If, for example, we are interested in what domains of knowledge a participant is drawing on in his or her discourse, the appropriate linguistic segment may well be the noun phrase because it is with this unit that the work of naming is accomplished. If, on the other hand, we are interested in the nature of the moves a participant is making, the appropriate unit of segmentation may well be the main clause with all its subordinate clauses because it is at this level of language that the work of rhetorical moves takes place.

■ Avoiding Potential Pitfalls in Segmenting

Our goal in segmenting data in verbal data analysis is to choose the unit, usually linguistic, where our phenomenon “lives.” Jokes, for example, usually take longer than a single sentence to tell, so if they are the phenomena in which we are interested, we would not choose the sentence or any of the even smaller units of language that we discuss later in this chapter. Instead, we would probably want to look at some of the longer units of language.

Pragmatically speaking, we want to choose a unit of segmentation that will allow us to divide the stream of language in a way that each segment will fit into one and only one category of our coding scheme. Making this choice can be tricky, however, since we often segment before we have our coding scheme worked out. In this case, then, we will need to rely on our intuitions about the phenomenon of interest.

To choose a unit of segmentation using our intuitions, we begin by identifying a salient difference across the contrast in our data. If, for instance, we are interested in kinds of negotiations that take place in email exchanges, we begin by looking through our data for instances of particularly successful negotiations, or failed negotiations, or protracted negotiations, and so on. The goal here is not yet to define what a negotiation is, but to observe the unit of language over which a negotiation occurs. When does the negotiation start? When is it over? Once we have two or three instances of negotiation, we can then review the units of analysis (described later in this chapter) to determine the one that seems to be where our phenomenon lives.

When we choose a unit of segmentation that matches our phenomenon, we will have a situation like that represented in Figure 3.1. Suppose we are interested, metaphorically speaking, in segmenting the “stream” of animals we see walking by on our daily walk. In Figure 3.1, we see this stream of pets divided into units by animal. When a coder looks at the second full unit in the stream and tries to code it—to determine, “Is it a cat?” or “Is it a dog?,” the task is relatively easy, because the entire phenomena (the cat) is included in the unit.

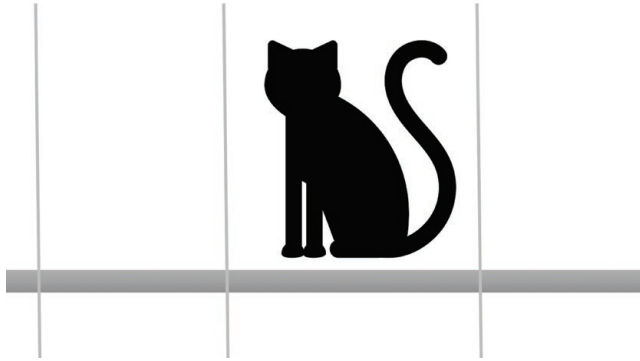


Figure 3.1: Matching the unit of segmentation to the phenomenon of interest.

If, however, the unit we choose is too small for the phenomenon of interest, we will encounter coding situations like that represented in Figure 3.2. When the coder encounters the first unit, she may or may not code it as *cat* (“Could it be a dog?”) because the information is ambiguous. She will, however, probably code the next unit as a *cat* because the information there is more complete. In general, using a too-small unit of segmentation leads to more disagreement in coding decisions—some analysts will say *cat* while others say *dog*. It also fails to provide a good sense of how frequently our phenomenon occurs—do we have just one cat or two in Figure 3.2?

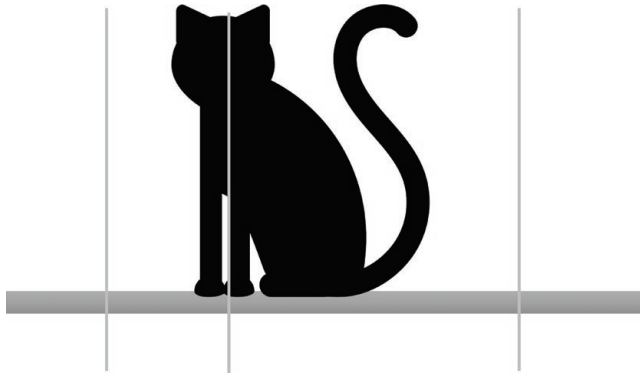


Figure 3.2: The unit of segmentation too small for the phenomenon of interest.

The final possibility of mismatching the unit of segmentation to the phenomenon of interest arises if the unit is too big. This situation is represented in Figure 3.3. In this case, the coder sees two different instances of the phenomenon of interest, both a cat and a dog, in the same unit. How should it be coded? In systems of qualitative analysis that allow more than one code to be applied to a unit, we might be tempted to code it both ways, but this would be problematic for verbal data analysis in ways we will discuss further in Chapter 4.

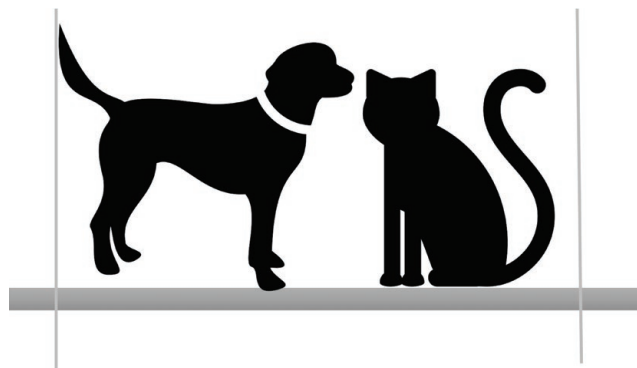


Figure 3.3: *The unit of segmentation too big for the phenomenon of interest.*

Another option, if we really want to make sure we see all of the cats in our analysis, would be to include a rule such as, “If you see a cat in the unit, code it as *cat* no matter what else you may see there.” The result of using a rule like this is to underreport dogs and other pets in favor of identifying all of the cats. While this might be an acceptable outcome for some purposes, we usually try to avoid this situation by getting the best match of unit to phenomenon of interest that we can at the segmentation stage.

Of course, in verbal data analysis, we are not coding for non-verbal phenomena like cats and dogs. But these instances are intended to illustrate conceptually some of the potential problems that can later arise from inappropriate segmenting, so that you can try to avoid them at this early stage of analysis.

■ Basic Units of Language

In this section, we describe some of the basic units of language into which any stream of verbal data can be segmented. Then, in subsequent sections we discuss units characteristic of specific types of verbal data. Finally, we close the chapter by describing the process of segmentation and pointing to some of the issues you may encounter and ways to handle them.

■ T-Units

Syntactically, a stream of language is structured as a set of t-units, the smallest group of words which can make a move in language. By move, we mean the work that a piece of language does to advance communication. In traditional argument, rhetorical moves advance the reader or listener along the path the speaker or writer is constructing (Kaufer & Geisler, 1991). More generally, a rhetorical move can be understood as the work done by language to fulfill any communicative purpose (Biber & Kanoksilapatham, 2007, p. 23).

A t-unit consists of a principle clause and any subordinate clauses or non-clausal structures attached to or embedded in it. The following are all t-units:

I ran to the store because we needed flour for the cake for Martha's birthday.

Jen is the mail carrier who replaced the one we liked.

Walking is my favorite form of exercise, the one with the least impact.

T-units are one of the most basic units of language. If the phenomenon in which you are interested is associated with the moves that a speaker or writer makes, the t-unit may be the most appropriate unit for your segmentation. You might, for example, look at each t-unit in the transcript of a meeting of an engineering design meeting for the kind of move it makes: descriptions, proposals, questions, evaluations, and so forth. You could also look only at t-units that make proposals. The length of t-units has also been used as a measure of syntactic maturity (Kellog, 1978).

To segment a stream of language into t-units, begin by finding the first inflected verb that has a subject. In the following sentence, the first inflected verbs with subject is I ran:

I ran to the store because we needed flour for the cake for Martha's birthday.

Next, look for the next inflected verb with a subject. If it stands on its own, then segment the stream at the most logical place between them, as shown in the following:

I ran to the store.

We needed flour for the cake for Martha's birthday.

If the second inflected verb you does not stand alone, keep it together with the first one:

I ran to the store because we needed flour for the cake for Martha's birthday.

In this example the two clauses with inflected verbs are linked with the subordinate conjunction *because*. As a result, the second clause cannot stand alone and would not be divided from its main clause:

I ran to the store because we needed flour for the cake for Martha's birthday.

If you are working with language that is spoken or with informal written language, you may find that some subordinate clauses are made to stand alone and should be treated as its own t-unit:

David: I ran to the store this morning.

Josh: Why'd you do that?

David: Because we needed flour for the cake for Martha's birthday.

Here David's answer is a subordinate clause but because it is spoken by a separate speaker, it would be segmented from the other speaker's question.

Keep in mind that many inflected verbs do not stand independently because they do not have their own subjects and should instead be kept with their main clauses:

Needing flour, I ran to the store.

I ran to the store and fell.

In the first of these t-units, the -ing form of need takes its subject from its main clause (*I*) and thus cannot stand alone from it. In the second, the inflected verb *fell* shares a subject with *ran* and thus stays in the same t-unit.

If you have trouble distinguishing between main and subordinate clauses, you may want to take a refresher course online. Here are a few you might consider:

- Clauses: the Essential Building-Blocks, Capital Community College Foundation, <http://grammar.ccc.commnet.edu/grammar/clauses.htm>
- Identifying Independent and Dependent Clauses, Purdue Online Writing Lab, https://owl.purdue.edu/owl/general_writing/punctuation/independent_and_dependent_clauses/index.HTML

■ Clauses

Clauses are the smallest units of language that make a claim—that predicate something—about an entity in the world. A clause is a group of words containing a subject—the entity—and a predicate—the claim being made about the subject. When clauses stand alone, they are said to be independent. When they make sense only in conjunction with an independent clause, they are said to be dependent. As we have already seen, an independent clause with all of its dependent clauses makes up the t-unit. The following are all independent clauses:

the committee requested the prior report from the president
once upon a time two children were lost in the woods

The underlined language in the following are all dependent clauses:

He refused when the committee requested the prior report from the president.

She told us that two children were lost in the woods.

If your phenomenon of interest is related to the claims that a speaker or writer makes about the world, the clause may be the right unit of analysis for you.

To segment a stream of language into clauses, begin as with t-units by

finding the first inflected verb that has a subject. Then look for the next inflected verb with a subject. Segment the stream at the most logical place between them:

He refused

when the committee requested the prior report from the president

The resulting segmented stream will consist of a mix of independent clauses and dependent clauses. Any stream of language that has been segmented using t-units can easily be further subdivided into clauses.

■ Noun Phrases

Noun phrases are the units of language which pick out objects in the world, both concrete objects and those which are abstract. In clauses, noun phrases can serve as subjects, but they may take other roles as well, as the following examples suggest:

That cat is obnoxious.

The day I was born was cold.

June is a hot month in Kentucky.

If your analysis is concerned with what is being spoken or written about, you may want to use the noun phrase as your unit of analysis. Noun phrases can help you identify the domains of knowledge from which speakers or writers draw, the worlds of discourse through which they move.

Choosing to analyze noun phrases is inherently selective—you make the decision not to look at the predicates that make up the clauses that, in their turn, make up the stream of verbal data. If you are going to look at noun phrases, you will probably want to choose some slightly larger unit (such as the clause) by which to segment the data, and then look at each noun phrase within that segment.

The easiest way to segment your discourse by noun phrase is to break the discourse up by clause, and then underline each noun phrase you find within each clause.

Critical care patients have often suffered a “disturbance” to the normal operation of their physiological system;
this disturbance could have been generated by surgery or some sort of trauma.

When you give coders data with this kind of selective segmentation, make sure that they understand that they are to code only the language that is underlined.

■ Verbals

Verbals are the unit of language which convey action, emotion, existence. Inflected verbals—those with tense—fill the predicate slot in clauses, both independent and dependent, as in the following:

When you back up your hard drive regularly, you prevent data loss.

Other verbals come as reduced verb phrases:

The purpose for backing up your hard drive should be obvious.

In this example, “backing up” is actually serving as part of a noun phrase, but it is clearly a reduced form of the verbal “back up” used in the previous example. Some verbals like “back up” are idiomatic combinations of verb forms with prepositions, back + up, where the meaning is quite different from the sum of the parts. In these cases, the verbal is the entire idiom.

If you are interested in the schema being invoked by your speakers or writers, you will want to select the verbal as your unit of analysis. The verbal, “back up your hard drive regularly,” for example, invokes a schema related to computer use in the same way that “went on a date” invokes a courtship schema. Using verbals, you can track the way schemas shift through your data set.

Because of their relationship to schemata, verbals are often indicative of genre choices, or shifts within genres from one part to another. In news reports, for example, hot news is often presented using present perfect tense:

The government has announced support for the compromise bill.

Details are then presented in simple past tense:

The compromise was worked out in committee yesterday.

In a similar fashion, in narratives, main events tend to be in simple past tense:

She went to see him one day and she said, “Has anybody been to see you?”

while really significant events may be presented in the historic present:

And he says, “No, but a right nice young lady came to see me.”

Background events tend to be in progressive form:

This friend of mine brought these photographs out, of the family through the years, and, passing them around, and he’s looking at them, and he said, “Oh! That the young lady that came to see me when I was in bed.”

If you are interested in genre, you may want to consider the verbal as a unit of segmentation. A complete list of verb tenses with examples can be found at <https://www.grammarly.com/blog/verb-tenses/> if you need to refresh your memory.

The easiest way to segment your discourse by verbal is to break the discourse up by clause, and then underline any verbal you find within each clause.

Critical care patients

have often suffered a “disturbance” to the normal operation of their physiological system; this disturbance

could have been generated by surgery or some sort of trauma.

The critical care physician

is to maintain certain patient state variables within an acceptable operating range.

Often the physician

will infuse several drugs into the patient

to control these states close to the desired values.

Notice that here in this scientific language, you find quite a few reduced verbals; that is, verbals that are reduced from full clauses. “Acceptable operation range,” for example, is a reduction of the clause, “a range that is acceptable to operate within” and “desired values” is a reduction of the clause, “values

that are desired.” In identifying the verbals, underline any phrase that can be expanded to a full verb. And remember, when you give coders data with this kind of selective segmentation, make sure that they understand that they are to code only the language that is underlined.

■ Topical Chains

Topical chains in both spoken and written interactions are what allow participants to understand their discourse as being about something. To some extent, the topic of a discourse can be established by how it points to or indexes objects in the world; listeners and readers will understand language which points to the same object to be somewhat cohesive. But the true workhorses of cohesion are the topical chains that writers and speakers establish.

Topical chains are constructed out of continuous units—either the t-units (frequent in formal discourse) or clause (not uncommon in informal discourse); each continuous unit may either start a new topic or continue the same topic as the one before it. When a writer constructs a long topical chain, or when two interlocutors work together to extend one another’s thoughts through a long topic chain, they develop the complexity of the topic.

Topical chains are often held together by the following kinds of referentials:

- personal pronouns: it, they, he, she, we, them, us, his, her, my, me
- demonstrative pronouns: this and that, these and those
- definite articles: the
- other expressions: such, one
- ellipses and repetition

In oral discourse, the boundaries of topical chains are often marked (*OK, well*).

If you are interested in the conceptual complexity of discourse, the extent to which a topic is developed, the depth of interaction on a topic, you may want to use the topical chain as your unit of analysis. You may also want to use the topical chain as a unit of analysis if you wish to do a selective analysis of discussions that concern a specific topic. Next to the t-unit, the topical chain is one of the most useful units for segmenting language.

To segment a stream of language into topical chains, it's easiest to begin by segmenting the stream of language into t-units for formal language or clauses for informal language. Then begin to study the way the referential language works to introduce topics or refer to topics already introduced, looking for breaks in the topical chain.

In the conversation shown in Figure 3.4, for example, Speaker A introduces the topic of keywords in clause 2 with the indefinite phrase “keywords.” In clause 3, he uses “several” to refer back to it. In clause 4, “one” also refers back to it; in clause 5, “they” again refers back to it; in clause 7, “keywords” is actually repeated for the first time since clause 2, but this time with the definite article (“the keywords”) to let us know that it is continuing the same topic.

All of these referentials indicate the topical connection among clauses 1 through 7, and it is not until clause 8 that we fail to find a reference back to keywords and instead see the introduction of a potential new topic, “the papers”; clause 10 takes up the papers topic with “them.” Clauses 10 and 11 briefly introduce a potentially new topic of abstracts but with just a few exceptions, the topic of papers is referred to over and over again through to the end of the excerpt, even across four speaker changes.

Speaker A:

- 1 The way I work with sources is
- 2 I go on the Web, to the Library Search and put in keywords [KEYWORDS] and
- 3 I use several [KEYWORDS]
- 4 so there's one [KEYWORD] for Chemistry
- 5 and there's one [KEYWORD] for Engineering
- 6 because they [KEYWORDS] don't all cover the same journals.
- 7 And you put in the keywords [KEYWORDS]
- 8 and you find the papers [PAPERS]
- 9 And you go get them. [PAPERS] [...]
- 10 reading abstracts [ABSTRACTS]
- 11 so—I always read the abstract [ABSTRACT] first

12 and I see if they [PAPERS] are useful.
13 And then when I get the paper,—[PAPER]
14 read the abstract, [ABSTRACT] and
15 I read the conclusions
16 before I decide if I'm gonna read the rest of the paper.
[PAPER]
17 And then I have—
18 let me open up my file cabinet
19 see that I have folders of by topic,
20 so I work with composites
21 so there will be some [PAPERS] on nanoparticles polymer
[that] faculty or (which is thermocetics?) or, you know, catego-
rize them [PAPERS]

Speaker B:

22 So you actually print them [PAPERS] out
23 after you have a look at the abstract [ABSTRACT] on the web
24 and then print it [PAPER] out
25 if you think it's [PAPER] interesting
26 and you file it [PAPER]?

Speaker A:

27 Right—
28 read it [PAPER] and
29 file it [PAPER]

Speaker B:

30 Read it [PAPER]
31 and file it [PAPER]

Speaker A:

32 And I often hand them [PAPERS] to my students

Figure 3.4: Following the topical chains in a conversation.

By using referentials to track and label the topics, as in Figure 3.4, we can clearly see where breaks in the topical chain occur. When you first begin to segment by topical chain, you may find it useful to track and label topics as we have done. With some practice, however, you will be able to sense the breaks without this kind of extensive annotation.

■ Units in Conversation

In addition to the basic units that characterize all verbal data described above, specific kinds of verbal data have additional regularities that can be exploited as units of segmentation. In this section, we look at regularities in conversation that suggest a variety of units of segmentation.

■ Conversational Turns

Conversations are made up of *turns*. For the most part, only one speaker talks at a time, although there is often some overlap at the edges. Much can be learned from looking at the turns in a conversation, particularly by speaker. How turns are allocated among possible speakers tells a great deal about relative power in a conversation: Who speaks most often? Whose turns are longest? Whose turns initiate new topics? If you are interested in phenomena of power, you may well want to look at the turn as a unit of segmentation.

To segment a stream of language into turns, simply segment at the borders between speakers.

■ Conversational Sequences

Conversation does not take place through the random accumulation of speaker's turns. Instead it is organized by its participants into *sequences*. A conversational sequence can be thought of as a joint project undertaken by two or more speakers, using language. It is made up of the following components.

1. Initiation: the first speaker proposes a “joint project”
2. Response: the speaker responds to the proposal
3. Follow-up: the speaker acknowledges the response

There exist a variety of kinds of joint projects, with routinized initiations that call for expected responses. A question, for example,

Speaker 1: What time is it?

generally receives a reply that contains the information requested:

Speaker 2: Six-thirty

and that is followed up by some acknowledgment:

Speaker 1: Thanks.

Other routinized exchanges include greetings:

Kate: Hi.

Ron: Hi.

and invitations:

June: Can you come to my party Saturday night?

Nance: Sure.

June: Great!

There are many more.

While initiations in conversational sequences call for a certain preferred response, interlocutors need not give the preferred response. In fact, interlocutors have four options when faced with a conversational proposal in the form of an initiation:

1. Compliance: Interlocutor takes up the proposed project
2. Alteration: Interlocutor proposes an alternative project
3. Declination: Interlocutor declines the proposed project
4. Withdrawal: Interlocutor withdraws from considering the proposed project

These four options are roughly ordered in terms of the first speaker's preferences. That is, first speakers hope their interlocutors will comply, but if not, perhaps propose an alternative:

Don: How 'bout dinner on Saturday?

Jen: Sorry. Can't. But what about tonight?

If no alternative is possible, the speaker hopes at least to get a declination that includes a reason for declining:

Don: How 'bout dinner on Saturday?

Jen: Sorry. Can't. I am going out of town to see my mom this weekend.

From the speaker's point of view, the worst response is a withdrawal:

Don: How bout dinner on Saturday?

Jen: You've got to be joking.

Lots of information can thus be gained by looking at the conversational sequence as a unit of segmentation. If interlocutors routinely give dispreferred responses rather than preferred responses, for example, it is a sign of a lack of cooperation.

The nature of conversational sequences can also shift significantly with context. With question-answer sequences for example, the speaker is not supposed to know the information being requested. In school, however, teachers routinely ask for information they already know and then use their follow-up turn to evaluate the student's answer:

Teacher: Who knows the capital of New York?

Jean: New York City

Teacher: Good guess, Jean, but not quite right. Johnny?

Johnny: Albany?

Teacher: Good!

This IRE (Initiation-Response-Evaluation) sequence is so closely tied to the context of school that even adults long out of school will feel like they are in school if subjected to this sequence structure. Other contexts appear to have their own specific sequences as well. Thus, if you are interested in looking for variations in context, such as from teacher-directed discourse to peer-to-peer collaboration—the conversational sequence may be your best unit of segmentation.

To segment a conversation into sequences, begin by locating an initiation (questions, invitations, etc.). If several initiations are repeated within the same

speaker's turn as rephrasings of each other, then treat them as a single initiation. Mark the segmenting boundary prior to the initiation either immediately before it, or, if the previous t-units were used by the same speaker to introduce the initiation, then right before these introductory t-units.

To locate the segmenting boundary following the initiation, examine the nature of the response:

1. No Response: If the initiation does not receive a response from a second speaker, divide after the initiation when silence is noted in the transcript.
2. Response Only: If a response is given by the second speaker and not commented on by any other speakers, then divide after the response. Responses can take more than one t-unit.
3. Response + Comment: If a response is given by a second speaker and then commented on either by the first speaker or by other speakers, then divide after these comments. In general, all comments by speakers on previous speakers' turns should be included in the sequence. A comment is related material, but has to be related to what comes immediately before it.

■ Interview Responses

Often implicitly, interviewers select the *response* as their unit of segmentation. That is, they look only at what is said in response to interview questions rather than at the interview as a whole. Such a move often helps to focus on the situation or person of interest, but it should never be done without considering the extent to which these responses have been shaped by conversational imperatives set up by the interviewer's questions.

Interviews are often structured according to an interview schedule—that is, with certain fixed questions that are asked of all those interviewed or are asked repeatedly of the same person over time. In such situations, it is possible to use the question as a unit of segmentation: to look at all responses to the same question. Since questions often direct respondents to particular topics, this unit of segmentation will help to focus on phenomenon related to topic.

No guarantee exists, however, that the same topic will not have come up

elsewhere in an interview. Thus, if you are concerned to be comprehensive, you may want to begin with the answers to certain questions and then move outward to look for the same topics elsewhere in the interview transcript.

■ Units in Text

Written texts have a variety of characteristics, some associated with conventions of publication, others with conventions of typography, and still others associated with the rhetorical interactions with readers at a distance. All of the following can serve useful purposes as units of analysis for textual data.

■ The Text

Perhaps the most obvious unit for analyzing textual data is the text itself. Unlike the stream of conversational data which must often be bounded in some arbitrary fashion for the purposes of analysis, written texts often have well-established boundaries. In a classroom, for example, students generally write and bind (with staples or paperclips) individual texts separately: the boundaries of individual student “papers” are seldom hard to determine. In published formats, conventions exist for separating individual texts: the chapters of an edited volume, the articles in a magazine or journal, the stories in a newspaper.

From the writer’s point of view, many phenomenon occur at the level of the text: the quality of the text, the genre of the text, the implied audience for the text. From the reader’s point of view, texts also have a variety of characteristics that can be examined: their persuasiveness, their familiarity, their importance, and so on. If you are concerned with any of these or similar phenomena, from either the writers’ or the readers’ point of view, you may find the text itself a good unit with which to segment textual data.

■ Genre Elements

Most texts belong to families of texts we call genres. While genres are not rigid, texts in certain genres do tend to share common features and common

structures. Genres represent a typified response to a typified rhetorical situation. They thus exhibit many typified features: typified moves, typified relationships to audience, typified reading patterns, typified publication venues.

You can use specific information you have about a genre to select or analyze specific genre-related elements—the abstracts of research articles, the response of readers to scientific articles, and so on. You may even use marked sections as the unit of segmentation when you want to look at kinds of rhetorical moves that tend to happen in certain places. You might, for example, look at the opening section of research articles to examine citation patterns since these openings often contain reviews of the literature. Or, looking for the same phenomenon, you might examine all sections in which citations are made.

Opening sections are also good places for looking at phenomena related to the voice of a piece or the relationship defined between author, reader, and context. Other times, you will want to skip opening sections and choose text from middle sections. Letters, for example, tend to have routinized openings that precede getting to the real issues with which the letter deals.

To segment a written text using a genre element, use the distinguishing features of the element to locate its boundaries in your discourse. Salutations in letters, for example, appear on paper in just one or two places. Comments on online news articles always come after the news story itself. If you want to segment by section, use the headings and our spacing within the discourse as your boundaries.

■ Typographical Units

Texts are structured by their layout with a variety of characteristics, any of which can be used as a unit of segmentation. As units, they can serve useful purposes as ways of selecting data when the phenomenon of interest is assumed to be regularly distributed through the text and you simply need some way of selecting part of the data.

You might choose, for instance, every third sentence, every fifth paragraph, every other page, or the first ten lines of each section. Keep in mind

that typographical units are relatively meaningless rhetorically. While paragraphs, for example, may be used for topical development by some writers, many writers simply break a paragraph based on relative length. If you want to use a rhetorically meaningful unit for segmentation, one in which the language does work of some kind, avoid using typographical units. But typographical units can be quite handy as a way of making a stratified or random selection of textual data.

To segment your data using a typographical unit, insert line breaks after each unit. Paragraphs, of course, may already have line breaks. To segment by sentence, you can use a search and replace function to replace periods (.) with a period followed by a carriage return.

■ Other Selective Units

■ Indexicals

Indexicals provide language with ways to anchor interactions to the specific context in which they occur. The essential indexicals are *I*, *here*, and *now*. With *I*, the speaker or writer points to him- or herself. With *here*, he or she anchors the discourse in place, and with *now*, he or she anchors it in time. Many other expressions depend upon our ability to identify the essential indexicals. For example, we cannot comply with the following sentence:

Bring the book tomorrow.

without identifying the implicit *I* (to surmise what book might be relevant), the *now* (to figure out what is tomorrow), and the *here* (to know where to bring it). The essential indexicals scope out the beginnings of a common ground that interlocutors share, a common ground that can be increasingly enriched by further interactions.

Speakers and writers use the demonstrative pronouns, *this* and *that*, *these* and *those*, to point to objects locating them physically or metaphorically with respect to the here of the discourse:

Not this one; that one.

These indexicals can also give you a handle on the extent to which interlocutors are coordinating with each other.

If your analysis is concerned with understanding the development and nature of the common ground that speakers or writers create with their interlocutors, you may want to use one or more of the indexicals as your unit of segmentation:

1. Pronouns: *I, he, she, it*
2. Demonstratives: *this* and *that*, *these* and *those*
3. Adverbs: *here, now, today, yesterday, tomorrow*
4. Adjectives: *my, his, her*

The easiest way to segment your discourse by indexical is to break the discourse up by clause, and then underline any indexical you find within each clause.

■ Personal Pronouns

Personal pronouns—*I, me, you, he, she, him, her, they, and them*—point to the world of interlocutors in which a speaker or writer takes as common ground. As we have already seen, pronouns are indexical. Focusing on the personal pronouns as a specific kinds of indexical can give you clues about the scope of the human world in which writers or speakers see themselves as acting. Looking specifically at first person pronouns (*I, me*) can help you to examine the agency of the speaker or writer. Personal pronouns can be looked at for themselves (how many time did the speaker use *I*?), for what they refer to (Where did the speaker talk about her family), or they can be used to select other phenomenon for analysis (what kind of verbals does the speaker attribute to herself).

To use personal pronouns to segment your discourse, underline each pronoun and then break the discourse right before each one. Or you may choose to segment your discourse first by some larger comprehensive unit such as the t-unit or clause, and then underline each personal pronoun within each of these larger units.

■ Modals

Modals provide language users with a way to indicate the attitude or stance of the writer or speaker toward the message he or she is conveying. The stance can range from bald assertion:

Sally will leave tomorrow.

to assertions with less definite status

Sally might leave tomorrow.

Sally could leave tomorrow.

Sally will probably leave tomorrow.

Sally will certainly leave tomorrow.

In general, modality can communicate probability (she might go tomorrow) advisability (she ought not go tomorrow), or conditionality (she would have gone yesterday). Modality is often conveyed through the modal auxiliary verbs: *might*, *may*, and *must*, *can* and *could*, *will* and *would*, *shall* and *should*, *ought*. Modality can be conveyed in many other ways however as the following lists suggest:

1. Modal auxiliaries: *might*, *may* and *must*, *can* and *could*, *will* and *would*, *shall* and *should*, *ought*
2. Conditionals: *if*, *unless*
3. Idioms: *have to*, *need to*, *ought to*, *have got to*, *had better*, *need to*
4. Adverbials: *probably*, *certainly*, *most assuredly*
5. Verbs: *appear*, *assume*, *doubt*, *guess*, *looks like*, *suggest*, *think*, *insist*, *command*, *request*, *ask*

All modals convey information about the level of obligation or certainty that speakers or writers associate with the content of what they are saying. If you are interested in tracking the degree of certainty with which interlocutors assess their claims, you may want to use modals as a unit of segmentation.

The easiest way to segment your discourse by modality is to break the discourse up by clause, and then underline any modal you find within each clause.

■ Metadiscourse

Metadiscourse is the part of discourse that talks about the discourse: the meta-discourse. If you can imagine that a text has a primary channel in which information is conveyed, the metadiscourse forms a background channel through which the writer talks to the readers to tell them how to understand and interpret the text.

There are two primary kinds of metadiscourse. Textual metadiscourse directs the reader in understanding the text. Textual connectives such as *first*, *next*, and *however* help readers recognize how the text is organized. Illocution markers like *in summary*, *we suggest*, and *our point is* point to the kind of work the writer is trying to do. Narrators such as *according to*, *many people believe that*, and *so-and-so argues that* let readers know to whom to attribute a claim. Textual metadiscourse is directly related to the rhetorical awareness exhibited in the text, and can be used as a unit of segmentation when you are concerned with rhetorical sophistication.

A second kind of metadiscourse is interpersonal, and serves to develop a relationship between writer and reader. Validity markers such as hedges (*might*, *perhaps*), emphatics (*clearly*, *obviously*), and narrators (*according to*) give the reader guidance on how much face value to give to the claim with which they are associated. Other attitude markers like *surprisingly* and *unfortunately* communicate the writer's attitude toward the situation and invite the reader to share the same stance. Commentaries such as *as we'll see in the following section* and *readers are invited to peruse the appendix* are more extended directions to the reader.

Interpersonal metadiscourse is directly related the degree to which a text shows evidence of audience awareness. Interpersonal metadiscourse can vary by genre, by rhetorical sophistication, and by the degree of comfort an writer has with the audience addressed. If you are interested in phenomenon related to audience, you may well wish to look at interpersonal metadiscourse as a unit of segmentation.

The easiest way to segment your discourse by metadiscourse is to break the discourse up by t-unit, and then underline any metadiscourse you find within each t-unit.

Exercise 3.1: Test Your Understanding

Within each group below, match the unit in the first column with the kind of phenomenon it can be used to study in the second column. (You can download this exercise at <https://wac.colostate.edu/books/practice/codingstreams/>).

Group 1

- | | | | |
|---|---------------------------------------|---|--|
| 1 | the text in written interactions | a | the certainty or obligation of claims |
| 2 | t-units in language | b | the domains of knowledge a writer or speaker refers to |
| 3 | modals in language | c | rhetorical awareness in text |
| 4 | metadiscourse in written interactions | d | the moves a writer or speaker makes |
| 5 | noun phrases in language | e | quality of writing |

Group 2

- | | | | |
|----|-----------------------------------|---|--|
| 6 | responses in interviews | f | frequency or distribution of textual phenomenon |
| 7 | sentences in written interactions | g | context of interaction |
| 8 | topical chains in language | h | the perceptions, feelings, beliefs of an interviewee |
| 9 | sequences in conversation | i | the complexity and depth of ideas |
| 10 | verbals in language | j | the schemata a writer or speaker uses |

Group 3

- | | | | |
|----|--|---|--|
| 11 | turns in conversation | k | the claims a writer or speaker makes |
| 12 | genre elements in written interactions | l | relative power among speakers |
| 13 | personal pronouns in language | m | the common ground and/or coordination between participants |
| 14 | clauses in language | n | typified action in writing |
| 15 | indexicals in language | o | references to the human world |

For Discussion: How can you see the phenomenon at work in each unit?

■ Memo 3.1: Unit of Segmentation

Find two to three sections of your data that show the kind of phenomenon that is of interest to you. Carefully note what you are seeing.

Next, look for several sections that show the absence or the opposite of this phenomenon. Also carefully describe what you are seeing.

Finally, consider the kind of unit over which this phenomenon occurs in your selection, reviewing the options for segmenting outlined earlier in this chapter.

Document your choice of unit of analysis, its relationship to the phenomenon that interests you, and your reasons for rejecting other units of analysis that might also have been appropriate.

■ Segmenting the Data

The goal of segmenting data to produce a file where each unit is separated from the next by a single paragraph break. Such data will be easy to move into an analysis program for further manipulation.

■ Segmenting Comprehensive Units

Segmenting by comprehensive units is perhaps the easiest task. Comprehensive units are those units which include the entire stream of language. That is, every word in the discourse is included in a segment to be further analyzed. All streams of language, for example, can be divided into t-units, clauses, or topical chains and every word in the stream will be part of some t-unit, clause, or topical chain. The following discourse has been segmented by t-unit (The paragraph symbol, ¶, is shown in order to clearly indicate the paragraph breaks).

Critical care patients have often suffered a “disturbance” to the normal operation of their physiological system;¶

this disturbance could have been generated by surgery or some sort of trauma.¶

The critical care physician is to maintain certain patient state variables within an acceptable operating range.¶

Often the physician will infuse several drugs into the patient to control these states close to the desired values.¶

For example, in the case of critical care patients with congestive heart failure, measured variables that are of primary importance include mean arterial pressure (MAP) and cardiac output (CO).¶

Secondary variables which are monitored, but not regulated as tightly as the primary variables, include heart rate and pulmonary capillary wedge pressure.¶

The physician uses her/his own senses for other variables that are not easily measured, such as depth of anesthesia, and often infers them from a number of measurements and patient responses to surgical procedures.

See Procedure 3.1 for more information on segmenting using comprehensive units.

■ Segmenting Conversational data

Segmenting conversational data poses some special challenges. This data often takes the following form with the name of the speaker, a colon, and then the actual conversational snippet:

P: okay ... ah ... in terms of your overall plan ... then ... where do you move from here ... after you finish extracting information ...

J: which is going to be a chore ... considering ...

P: it's going to take a while right ...

If you are moving your data into Excel, you will want the speaker identification in one column and the actual conversation, divided into your chosen units, in an adjacent column. In the Excel example given in Figure 3.5, for example, the



Procedure 3.1: Segmenting Using Comprehensive Units

<https://goo.gl/1jf8Up>


1. Working with a single stream of language in a word processing program, place your cursor at the break between one segment and the next.
2. Hit enter.

Speakers names (P and J) are in the first column. In the second column, each clause appears, one line at a time.

See Procedure 3.2 for more information on segmenting conversational data.

P:	okay ... ah ... in terms of your overall plan ... then ... where do you move from here ...
	after you finish extracting information ...
J:	which is going to be a chore ... considering ...
P:	it's going to take a while right ...
	okay ... ah ... in terms of your overall plan ... then ... where do you move from here ... after you finish extracting information ...

Figure 3.5: Conversational data, segmented by clause and moved into Excel.

 **Procedure 3.2: Segmenting Conversational Data**

<https://goo.gl/1jf8Up>

- Working with your stream of conversation in a word processing program, turn off **Autocorrect** in your **Preferences** under the **Word** menu item.
- Replace the colons after speaker names with colon + tab.
- Place your cursor at the break between one segment and the next.
- Hit enter to insert a carriage return and then add a tab before the second unit

The results should look like the following with -> represeting tabs:

P:->

okay ... ah ... in terms of your overall plan ... then ... where do you move from here ...

->

after you finish extracting information ...

J:->

which is going to be a chore ... considering ...

P:->

it's going to take a while right ...

■ Segmenting for Selective Units

Unlike comprehensive units, such as t-units and clauses, selective units include just part of the stream of language. As we suggested earlier, in the first approach to segmenting for selective units, you simply underline each selective unit and then segment the discourse right before each one. Using our cats and dogs metaphor from earlier in this chapter, this method of segmentation is equivalent to dividing the stream right before each animal's nose, so that later you can code each wet nose as a dog and each dry nose as a cat.³

In a more relevant example of this approach to selective segmentation, this one taken from Rick Steves' online guide on Internet Security for Travelers (<https://www.ricksteves.com/travel-tips/phones-tech/internet-security-for-travelers>), we have underlined modals for further analysis and placed each one on its own line in Excel, ready to be coded in the adjacent column:

1	<u>If</u> you're taking your devices on the road, be aware that gadget theft is an issue in Europe. Not only	
2	<u>should</u> you take precautions to protect your devices from thieves, but you	
3	<u>should</u> also configure them for maximum security so that	
4	<u>if</u> they are stolen, your personal data	
5	<u>will</u> stay private.	

This way of segmenting for selective units not only clearly communicates to coders which text is supposed to be considered in coding (the underlined words), but also supplies them with the full context to support that coding.

As you can see from this example, however, the text shown on each line is rather arbitrary and meaningless. And we will often encounter discourse that has long passages without a selective unit, as in this passage further along in Steves' article:

³ I now realize that this rule of thumb does not actually work for coding cats and dogs, but I was taught it when I was very young and it does illustrate the point nicely.

21	Many laptops have a file-sharing option. Though this setting is	
22	<u>likely</u> turned off by default, it's a good idea to check that this option is not activated on your computer so that people sharing a Wi-Fi network with you	
23	<u>can't</u> access your files	
24	(<u>if</u> you're not sure how, do a search for your operating system's name and "turn off file sharing"). Newer versions of Windows have a "Public network" setting (choose this when you first join the network) that automatically config-ures your computer so that it's less susceptible to invasion. Once on the road, use only legitimate Wi-Fi hotspots. Ask the hotel or café for the specific name of their network, and make sure you log on to that exact one. Hackers sometimes create bogus hotspots with a similar or vague name (such as "Hotel Europa Free Wi-Fi") that shows up alongside a bunch of authentic networks. It's better	
25	<u>if</u> a network uses a password (especially a hard-to-guess one) rather than being open to the world.	
26	<u>If</u> you're not actively using a hotspot, turn off Wi-Fi so that your device is not visible to others.	

Not only is this first way of segmenting selective units unrelated to meaning, but it presents problems for understanding the frequency of your phenomenon. In this example, the length of the segments is arbitrary, ranging from four words in segment 23 to 106 words in segment 24. Thus to give a sense of the relative frequency of modals, we could not rely on the number of segments as a basis for our analysis; that is, there is no communicative value in saying that there was on average one modal per segment, since definitionally we have insured that there will be one modal per segment. To give a better sense of frequency, then, we will have to use some other base metric, saying, for example, that there were five modals in 185 words, or an average of 2.7 (5/185) modals per 100 words.

A second problem with this first kind of segmentation arises if you should wish to code your data in a second way, a not uncommon strategy as we'll see

in the next chapter. Going back to our cats and dogs example, suppose we decide we want not only to code for cats and dogs that are in our stream, but also for the kinds of flowers we see. Since the natural segmentation unit for flowers is the plant, we could go back and resegment our stream in an entirely different way than by the nose in order to code for flowers, but this would be a tremendous amount of work.

As this metaphorical example suggests, a second and simpler approach to segmenting for selective units is often called for. This second approach involves picking a comprehensive unit to start with and then using underlining to identify the selective units. In our cats and dogs and plants example, this might mean segmenting our stream by property lot, and then coding each lot first for pets and second for plants in bloom. We might have to deal with the problem of a few lots that had both cats and dogs (perhaps by adding a category for *both*), but this approach to segmenting would allow us to look for relationships between pet ownership and the state of the lot's landscape.

Coming back to the Rick Steves example, this second approach to selective segmentation approach would involve segmenting the discourse first by clause, and then underlining each modal within the clause:

1	<u>If</u> you're taking your devices on the road,	
2	be aware	
3	that gadget theft is an issue in Europe.	
4	Not only <u>should</u> you take precautions to protect your devices from thieves,	
5	but you <u>should</u> also configure them for maximum security	
6	so that <u>if</u> they are stolen,	
7	your personal data <u>will</u> stay private.	

This kind of segmentation allows you to describe the frequency of modality with a statement such as, "71% (5 of 7) of clauses contained modals," which does give a more informative sense of their frequency. In the later section with fewer modals, segmenting first by clause and then underlining the modals would produce the following:

1	Many laptops have a file-sharing option.	
2	Though this setting is <u>likely</u> turned off by default,	
3	it's a good idea to check	
4	that this option is not activated on your computer	
5	so that people sharing a Wi-Fi network with you	
6	<u>can't</u> access your files	
7	(<u>if</u> you're not sure how,	
8	do a search for your operating system's name	
9	and "turn off file sharing").	
10	Newer versions of Windows have a "Public network" set- ting	
11	(choose this	
12	when you first join the network)	
13	that automatically configures your computer	
14	so that it's less susceptible to invasion.	
15	Once on the road, use only legitimate Wi-Fi hotspots.	
16	Ask the hotel or café for the specific name of their network,	
17	and make sure	
18	you log on to that exact one.	
19	Hackers sometimes create bogus hotspots with a similar or vague name (such as "Hotel Europa Free Wi-Fi")	
20	that shows up alongside a bunch of authentic networks.	
21	It's better	
22	<u>if</u> a network uses a password (especially a hard-to-guess one) rather than being open to the world.	
23	<u>If</u> you're not actively using a hotspot,	
24	turn off Wi-Fi	
25	so that your device is not visible to others.	

Now we can say that the same five modals occur over 25 clauses or at a rate of one every five clauses.

When you use this kind of combination of comprehensive and selective segmentation, you may find that more than one example of the selective unit occurs within the comprehensive unit. The following passage, for example, has been segmented by clause and then for noun phrases. Each of the clauses contains more than one noun phrase:

1	<u>Critical care patients</u> have often suffered a “ <u>disturbance</u> ” to the <u>normal operation of their physiological system</u> ;	
2	<u>this disturbance</u> could have been generated by <u>surgery</u> or <u>some sort of trauma</u> .	

Our interest here is not, of course, whether the clauses have noun phrases, but what kind of noun phrases; perhaps we want to code each noun phrase for the use of everyday language or medical jargon. This would require us to pull out each noun phrase on a separate line for coding. Ideally, our data would look like that shown in Figure 3.6 once in Excel:

A	B	C	D	F
clause #	noun phrase #	Clause/Noun phrase	Code 1	Code 2
1		<u>Critical care patients</u> have often suffered a “ <u>disturbance</u> ” to the <u>normal operation of their physiological system</u> ;		
	1a	Critical care patients		
	1b	a “disturbance” to the normal operation of their physiological system;		
2		<u>this disturbance</u> could have been generated by <u>surgery</u> or <u>some sort of trauma</u>		
	2a	this disturbance		
	2b	surgery		
	2c	some sort of trauma		

Figure 3.6: Data first segmented comprehensively by clause and then selectively by noun phrase.

This data is set up so that 1) the noun phrases are numbered in column B (1a, 1b, 2a, 2b, 2c) and can be coded using column D; and 2) the clauses are numbered in column A (1, 2) and can be coded in column F. The greyed-out cells in each column help to tell the coder what data not to code. We will describe the procedure for formatting this kind of data before moving it into Excel in the section on Moving the Segmented Data.

■ Creating a Segmenting Style

A file full of text segmented using paragraph breaks can be difficult to read:

Critical care patients have often suffered a “disturbance” to the normal operation of their physiological system; this disturbance could have been generated by surgery or some sort of trauma.

The critical care physician is to maintain certain patient state variables within an acceptable operating range.

Often the physician will infuse several drugs into the patient to control these states close to the desired values.

For example, in the case of critical care patients with congestive heart failure, measured variables that are of primary importance include mean arterial pressure (MAP) and cardiac output (CO).

This problem that can be remedied by applying stylistic formatting to shape the text into a more readable form (see Procedure 3.3). For example, it is often helpful to increase the spacing after each segment in order to distinguish segmenting breaks from simple text wrapping:

Critical care patients have often suffered a “disturbance” to the normal operation of their physiological system; this disturbance could have been generated by surgery or some sort of trauma.

The critical care physician is to maintain certain patient state variables within an acceptable operating range.

Often the physician will infuse several drugs into the patient to control these states close to the desired values.

For example, in the case of critical care patients with congestive heart failure, measured variables that are of primary importance include mean arterial pressure (MAP) and cardiac output (CO).

■ Memo 3.2: Segmenting Procedure

Using the unit of analysis you selected in Memo 3.1, segment four to five pages of your data. Make sure to select typical data—data from the middle of a conversation or text, data from across your built-in contrast, and so on.

Annotate your segmentation, noting where you are uncertain of your segmentation. Consult with an online source or a colleague to help you resolve your uncertainties.

Document your segmenting decisions so that you can maintain consistency as you segment all of your data.



Procedure 3.3: Using a Style to Format your Data

<https://goo.gl/1jf8Up>

To create a new style in Microsoft Word:

1. Select a segment.
2. Format it in the way you want.

For example, you might increase the spacing after a segment to 6 points by placing the cursor in the segment, invoking the **Paragraph** command on the **Format** menu, and increasing the spacing after the paragraph to 6.

3. Click on the **Style Panes** icon to open the **Style Pane**.
4. Click on the **New Style** button.
5. Name your new style.

To apply this style to other segments:

6. Select the other segments to which you want to apply the new style.
7. Then choose the new style from the drop down **Style** menu.

To change a style:

8. Change the style the way you want in one segment.
9. Then in the **Style Pane**, hover over the style name until you see the drop down menu to the right.
10. From that drop down menu, choose **Update to Match Selection**.

Word will automatically apply the new style changes to every segment with that style in your file.

■ Moving the Segmented Data

Once the text has been segmented appropriately in your word processing application, you are ready to move the segmented data into your data analytic application. This procedure differs slightly depending on whether your data is segmented comprehensively or selectively. We start by describing the methods to use with comprehensively segmented data and then review those for data that has been selectively segmented.

■ Moving Comprehensively Segmented Data

If you have segmented your data using a comprehensive unit, you will now want to move it into your analytic application. Procedures for Excel (3.1 and 3.2) and MAXQDA (3.1 and 3.2) provide guidance on this process.



Excel Procedure 3.1: Moving and Numbering Comprehensively Segmented Data into Excel

<https://goo.gl/1jf8Up>

1. Select and copy the data to be moved from Word.
2. Paste it into a worksheet, starting with Column C, leaving Columns A and B free for the labels you will insert later.

Generally speaking, each data stream (interview, transcript, text) should be placed on its own worksheet. Make sure to label the worksheets as you go.

After segments are moved into Excel, you should label them:

3. In Column B, insert numbers starting with 1 and continuing for 3 or 4 segments.
4. Select these cells and drag down to fill the column.
5. In Column A, type a label next to the first segment.
6. Copy the label and select the rest of the cells next to the rest of the data and issue the paste command.

Numbering and labeling segments in this way will insure that each segment has a unique identity in analysis.

If you have conversational data, you will also want to insure that each segment is labeled for speaker.

Exercise 3.2: Try It Out

In word processing, segment the following text by t-unit, move it into Excel or MAXQDA. Make sure to number and label the segments if necessary. Compare your results with others in your class. (You can download this exercise at <https://wac.colostate.edu/books/practice/codingstreams/>).

The language people speak or write becomes research data only when we transpose it from the activity in which it originally functioned to the activity in which we are analyzing it. This displacement depends on such processes as task-construction, interviewing, transcription, selection of materials, etc., in which the researcher's efforts shape the data. Because linguistic and cultural meaning, which is what we are ultimately trying to analyze, is always highly context-dependent, researcher-controlled selection, presentation, and recontextualisation of verbal data is a critical determinant of the information content of the data. Data is only analyzable to the extent that we have made it a part of our meaning-world, and to that extent it is therefore always also data about us. Selection of discourse samples is not governed by random sampling. Discourse events do not represent a homogeneous population of isolates which can be sampled in the statistical sense. Every discourse event is unique. Discourse events are aggregated by the researcher for particular purposes and by stated criteria. There are as many possible principles of aggregation as there are culturally meaningful dimensions of meaning for the kind of discourse being studied.

For Discussion: What issues did you have to resolve to do this segmenting?



MAXQDA Procedure 3.1: Importing Comprehensively Segmented Data

<https://goo.gl/1jf8Up>

Moving comprehensively coded data into MAXQDA is very easy.

1. In a new project in MAXQDA, from the **Documents** menu, select the segmented files you want to import.
2. Click **Open**.

Generally speaking, each data stream (interview, transcript, text) should be placed on its own document. Make sure to label the documents with identifying information as you go.


Each file will be imported and automatically numbered by segment. MAXQDA will also automatically keep track of the source of each segment. Thus, you do not need to do any additional numbering or labelling.

Moving Selectively Segmented Data

If you have segmented first with a comprehensive unit and then with a selective unit, you may want to number your segments *before* moving them. With the method described in Excel Procedure 3.2 and MAXQDA Procedure 3.2, you can automatically and separately number both the comprehensive units and the selective units as shown in Figure 3.7. The process is a bit complicated, but it sure beats doing it by hand.

	A	B	C	D
	clause	noun phrase		
1	#	#	Clause/Nount phrase	Code
2	1		1 Critical care patients have often suffered a "disturbance" to the normal operation of their physiological system;	
3		1 a	Critical care patients	
4		1 b	a "disturbance" to the normal operation of their physiological system;	
5	2		2 this disturbance could have been generated by surgery or some sort of trauma	
6		2 a	this disturbance	
7		2 b	surgery	
8		2 c	some sort of trauma	
9	3		3 this disturbance could have been generated by surgery or some sort of trauma	
10		3 a	this disturbance	
11		3 b	surgery	
12		3 c	some sort of trauma	

Figure 3.7: Selectively segmented data numbered in Excel.

 **Excel Procedure 3.2. Numbering and Moving Selective Segments in Excel**

<https://goo.gl/1jf8Up>

- Working with a stream of language in Microsft Word, underline the selective segments in your comprehensive unit:
Critical care patients have often suffered a “disturbance” to the normal operation of their physiological system
- Create a copy of the comprehensive segment below the comprehensive unit and edit it so that each selective unit is located on a separate line beneath the comprehensive unit:

Continued . . .-



Excel Procedure 3.2: Numbering & Moving Selective Segments (continued)

<https://goo.gl/1jf8Up>

Critical care patients have often suffered a “disturbance” to the normal operation of their physiological system

Critical care patients

a “disturbance” to the normal operation of their physiological system

After you have edited all of your segments:

3. Select all of the segments you want to number, both comprehensive and selective.
4. Select **Outline Numbered** from the **Bullets and Numbering** option under the **Format** menu.
5. Select the third format option (1. 1.1. 1.1.1) and click **Customize**.
6. With Level 1 selected, add a second period to the number format so that it is a number followed by two periods (1..).
7. Select Level 2 and change the **Number Style** to a, b, c
8. Edit the number format so that it is a period followed by a number followed by a letter followed by a period (.1a.) and click **OK**.
9. To move the selective segments to level 2, select and indent them.
10. Check to make sure that the text looks appropriately numbered, with comprehensive segments numbered 1, 2, and so on followed by two periods, and selective segments numbered under their comprehensive segments as 1a 1b and so on with a period before and after the numbering.
11. Save your file as a text file (.txt).

To import your data:

12. From within Excel, put your cursor in cell B2 and invoke the **File > Import** command.
13. Select text file from the file types and click **Import**.
14. Select the file you want to import and click **Get Data**.
15. In Step 1 of the **Text Import Wizard**, select **Delimited** as your file type and click **Next**.
16. In Step 2 of the **Text Import Wizard**, uncheck all delimiters and type a period (.) in the box following **Other:.**
17. In Step 3 of the **Text Import Wizard**, click **Finish**.
18. In the **Import Data dialogue box**, make sure the data will go into an existing worksheet with =\$B\$2 as the destination and click **OK**.

■ Issues in Segmenting

As you segment verbal data, a few issues will arise that may require special handling. In closing this chapter, we call your attention to some of these.

■ Fragments

Particularly if you are working with oral or online conversations, you may need to deal with fragments of language that don't quite add up to the full unit you are using for segmentation. Not only do we encounter the *uhms* and *ohs* with which people fill their speech, but we also hear the fits and starts of unfinished ideas, a clause started and left hanging. You will need to decide how



MAXQDA Procedure 3.2: Importing Selectively Segmented Data

<https://goo.gl/1jf8Up>

1. In Word, underline the selective segments in your comprehensive unit:
Critical care patients have often suffered a “disturbance” to the normal operation of their physiological system
2. Create a copy of the segment and place the selective units beneath the comprehensive unit:
Critical care patients have often suffered a “disturbance” to the normal operation of their physiological system
 Critical care patients
 a “disturbance” to the normal operation of their physiological system
3. Select each of the selective units and change the font color:
Critical care patients have often suffered a “disturbance” to the normal operation of their physiological system
 Critical care patients
 a “disturbance” to the normal operation of their physiological system;

When you import the data into MAXQDA, it will preserve the font colors, and you will be able to tell a coder to code just those segments in black (the comprehensive units).

to handle these fragments and be consistent in your treatment.

You might, for example, treat the monosyllabic back channeling by a second speaker as separate turns as has been done in the following transcript (Li et al., 2010):

Dr.: I'll give a prescription for the codeine.

Pt.: Uhm.

Dr.: You're a pretty damn healthy guy so it shouldn't be a problem.

Pt.: Uhm.

Or you might routinely decide just to include them in the first speaker's turn in square brackets as has been done here:

Dr.: I'll give a prescription for the codeine [Uhm] You're a pretty damn healthy guy so it shouldn't be a problem [Uhm]

If you are trying to capture the back and forth of the conversational dynamics, you should use the first method. But if you intend to code the segments for meaning, you might prefer to use the second method.

With incomplete thoughts, where a speaker has started one way and then started over to continue in a different way, it is probably best to treat them as separate segments:

I was just wondering if ...

I was just wondering when we are planning to leave?

■ Center Embedding

Another issue that you may encounter involves center-embedded clauses. Most of the clauses in English come one right after another and can easily be segmented as in this earlier example from Rick Steves:

- 1 Many laptops have a file-sharing option.
- 2 Though this setting is likely turned off by default,
- 3 it's a good idea to check
- 4 that this option is not activated on your computer

Occasionally, however, one clause is embedded right in the center of another clause:

- 5 ... people who share a Wi-Fi network with you
- 6 can't access your files

In this example (modified slightly from Steves original), the subordinate clause, *who share a Wi-Fi network with you*, is plopped right in the center of another clause, *people can't access your files*. The way we have done the segmentation separates the subject of the clause, *people*, from its verb, *can't access*. You might be tempted to draw the segmentation so that at least the embedded clause is correctly segmented:

- 5 so that people
- 6 who share a Wi-Fi network with you
- 7 can't access your files

But while this segmentation may be more technically correct, it does little to support our intended coding. For the most part, we have found it best to keep the embedded clause with the part of the surrounding clause that it modifies, leaving only the remainder of the clause for a separate segment. However you choose to handle center embedding, be consistent about it.

■ Pronouns

A final issue you may encounter in preparing data for coding involves pronouns. Pronouns pose more difficulties in interpretation than the noun phrases to which they refer. Many references are vague; others refer to persons or things outside of rather than in the text. And, of course, anytime you ask people to take the additional step of deciding what a pronoun refers to before they decide how to code it, there will be increased variation.

To manage this referential complexity, you can take one of two approaches as you segment the data. The first is simply to remove pronouns from coding if your unit of analysis would otherwise indicate that they should be coded. So, for example, if you are planning to code all nominals, you might decide not to code any nominal that was a pronoun. While such a decision might seem to eliminate a lot of data, if the elimination is spread proportion-

ately through your coding categories, the overall patterns will be preserved.

Sometimes, however, you will not be able to eliminate the pronouns because they contain important information about the phenomenon of interest. If, for example, you are looking at references to human agents, you may not want to eliminate pronouns because they disproportionately contain a lot of information about agency in verbal data.

In this case, you may pre-process the verbal data to insert into the data the noun to which the pronoun should be understood to refer. So, for example, if the pronoun “he” is used to refer to Harry, we could insert it as follows:

He [Harry] was taking his dog for a walk.

Resolving pronominal reference in this way in advance of coding will allow the data to be coded with greater consistency. But this technique is inherently tricky: if the referent is unclear or vague, you may need to read too much into the data to resolve it. Thus you may find it best to resolve only those references about which there is no ambiguity.

■ Memo 3.3: Data Set for Analysis

Complete the segmentation of your data set and move it into the data analytic application of your choice.

Document your data set. Create a table that shows how many segments you have for each text/transcript across your build in contrast. Include a verbal description of the table in your notes.

■ Selected Studies Using Segmentation

- Graham, S. S., Kim, S.-Y., DeVasto, D. M., & Keith, W. (2015). Statistical genre analysis: Toward big data methodologies in technical communication. *Technical Communication Quarterly*, 24(1), 70-104. (By paragraph.)
- Imbrenda, J.-P. (2016). The blackbird whistling or just after? Vygotsky’s tool and sign as an analytic for writing. *Written Communication*, 33(1), 68-91. (By sentence.)

- Kuhn, D., Hemberger, L., & Khait, V. (2015). Tracing the development of argumentative writing in a discourse-rich context. *Written Communication*, 33(1), 92-121. (By idea unit.)
- Ngai, C. S. B., & Jin, Y. (2016). The effectiveness of crisis communication strategies on Sina Weibo in relation to Chinese publics' acceptance of these strategies. *Journal of Business and Technical Communication*, 30(4), 451-494. (By genre element, response).
- Shin, W., Pang, A., & Kim, H-Y. (2015). Building relationships through integrated online media: Global organizations' use of brand web sites, Facebook, and Twitter. *Journal of Business and Technical Communication*, 29(2), 184-220. (By genre element—website, Facebook profile, wall post, Twitter profile, tweet).
- Swarts, J. (2015). Help is in the helping: An evaluation of help documentation in a networked age. *Technical Communication Quarterly*, 24(2), 164-187. (By t-unit.)
- Walker, K. C. (2016). Mapping the contours of translation: Visualized un/certainties in the ozone hole controversy. *Technical Communication Quarterly*, 25(2), 104-120. (By sentence.)

■ For Further Reading

- Biber, D., & Kanoksilapatham, B. (2007). Introduction to move analysis. In D. Biber, U. Connor, & T. A. Upton (Eds.), *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins.
- Clark, H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press.
- Heritage, J. (1984). *Garfinkel and Ethnomethodology*. Cambridge, UK: Polity Press.
- Kaufer, D. S., & Geisler, C. (1991). A scheme for representing written argument. *The Journal of Advanced Composition*, 11(1), 107-122.
- Kolln, M., & Funk, R. (1998). *Understanding English grammar*. Boston: Allyn and Bacon.
- Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Los Angeles: Sage.
- Li, H. Z., Cui, Y., & Wang, Z. (2010). Backchannel responses and enjoyment of the conversation: The more does not necessarily mean the better. *International Journal of Psychological Studies*, 2(1), 25-37.
- McCarthy, M. (1996). *Discourse Analysis for Language Teachers*. Cambridge, UK: Cambridge University Press.
- Mehan, H. (1979). *Learning lessons: Social organization in the classroom*. Cambridge, MA: Harvard University Press.

- Morris, J. S. (2009). The Daily Show with Jon Stewart and audience attitude change during the 2004 party conventions. *Political Behavior*, 31(1), 79-102.
- Neuendorf, K. (2016). *The Content Analysis Guidebook* (2nd ed.). London: Sage Publications. Kindle Edition.
- Saldaña, J. (2016). *The coding manual for qualitative researchers*. London: Sage Publications.