Using a Concordance for Discourse Research

Objective

The primary objectives of this tutorial are:

- to show what a concordance program is, and what it can do for your discourse research—especially for exploring connections and correlations between patterns in spoken discourse and patterns in grammar and lexicon

- to show how to to take advantage of the systematic representation in discourse transcriptions of such discourse features such as turn-taking, simultaneous speech, vocalizations, pauses, and prosody

- to introduce users to the specific research potential of the type of discourse transcription information found in the Santa Barbara Corpus of Spoken American English, and in other similar transcriptions

Background

A concordance program lets you search through your discourse data for all the instances of a given word, and then see the surrounding context for each instance that is found. The result is called a key-word-in-context concordance. You type in the word you are interested in (the key word), and in a few seconds you have a list of neatly lined up examples of that word as found in your data. You can then sort this list of words in a variety of ways, in order to get a clearer picture of how it relates to its context. No matter how you sort the words, their context remains available to you.

You can also search for patterns of words, using various "wild-card" symbols to represent the combinations of words you are interested in. With a little cleverness in designing your search (and assuming your discourse data are in a systematic discourse transcription format), you can also look for such discourse features as turn completions, discourse particles or vocal noises word that typically co-occur with turn beginnings, interruptions and simultaneous speech, and so on.

In addition, you can get certain statistics on, for example, what other words most frequently collocate with your key word, and how many instances of each collocation occur in your data.

You can print out your search results, or save them in a file to be printed out with your favorite word processor. Or you can just select specific examples you would like to save (e.g. to include in a paper you are writing), and copy them directly into a separate window that you open for your word processor.

MonoConc is such a key-word-in-context program, designed for Windows. It is easy to use, fast, and intuitively organized, and it is available for use in the UCSB Linguistics Lab. Other comparable Windows concordance software to consider is WordSmith (from Oxford University Computing) and ShoeBox (from the Summer Institute of Linguistics). (ShoeBox also allows interlinear grammatical analysis and glossing; facilitates automatic glossing; and can sort data into "dictionary" format.) (For more information about MonoConc and corpus linguistics in general, contact Michael Barlow at athel@aol.net, or http://www.ruf.rice.edu/~barlow/corpus.html, or http://www.nol.net/~athel/athel.html.)

General Instructions

From the Windows **Start** menu, start MonoConc by selecting **Programs**, then **MonoConc**. Or if you have an icon for MonoConc on your desktop, you can double-click on that. (To save the discourse examples you find, you may also want to open a separate window for WordPad, or for your favorite word processor.) Maximize the MonoConc window so that it fills your screen, to allow you to see more examples at once, and more discourse context.

We will be working with transcriptions from the Santa Barbara Corpus of Spoken American English (SBCSAE or SBCorpus) as our discourse data. To access this data, do the following. Within the MonoConc window, load a corpus by telling MonoConc what directory you want, and which files, as follows: From the menu, select **File**, then **Load Corpus**. A dialog box will appear. Make sure the "List files of type" window says "All Files (*.*)" Specify the appropriate drive for directory you want, by clicking on the drop-down arrow for "Drives", and highlighting the appropriate drive. Specify the appropriate folder, by double-clicking on the folder(s) listed in the folder tree window. To use the conversational (and other) transcriptions of the Corpus of Spoken American English available on the UCSB Linguistics Department network, , this would probably be something like **\Data\English\SBCorpus\TextOnly.** Select all the files in this directory, by highlighting them (drag over the listed files with the mouse). Select **OK**.

One or more text files will now be displayed on your screen, each in a separate window. You can look over the transcriptions if you like, but you don't have to. To clear your screen so you can concentrate on concordancing, select **Corpus Text/Hide All**. Even though the file (or files) will no longer be visible, they will still be used in whatever concordance searches you do during this session.

To exert greater control over your searches, you may want to modify the various options that MonoConc offers under the headings of **Concordance**/**Search Options** and **Frequency/Frequency Options**. (For more information, see the tutorial section on Setup Options.) But for getting started, the MonoConc settings in the Linguistics Lab should be sufficient. (I have set up the various options so as to facilitate working with spoken discourse transcriptions.)

Note: In MonoConc, *spaces matter*. MonoConc treats each space as a "word" boundary: a word is thus any string of symbols bounded by spaces. In the examples of concordance searches given below, you are often asked to type several words or symbols; always type the spaces exactly as they appear on the handout. (When you see a space in an example, it is always exactly one space: there are no cases, in this tutorial, where you will need to type in two spaces in a row.)

You are now ready to start concordancing!

Words and affixes in discourse

The questions which are introduced below are intended to be suggestive of the kinds of research questions you can use a concordance program to ask. With these models to start you off, there are unlimited ways you can use your imagination to extend the questions that you ask in your own research.

1. When is the discourse marker 'anyway' is used in discourse?

Select **Concordance**, then **Search**. Where it says "enter pattern to search for", type the word **anyway**. What do you see? Looking over the data, what generalizations can you make?

Pick a line that looks interesting to you, and click on this line with your mouse. Notice that a larger amount of context for this line immediately appears in the upper half of the screen. To enlarge this discourse context window so you can see more of the prior (or following) discourse context on your screen, use your mouse to click on the horizontal line separating the two halves of the screen, and drag this line down . Use the scroll bars to look over the discourse context for your example (up to several pages of context).

Try clicking on several other lines. Notice that for each example you click on, the relevant context appears in the upper screen, and the keyword is highlighted in color. Also, each time you select a line, the name of the computer file which that example comes from is displayed on the status line (i.e. in the lower left corner of the MonoConc window). The total number of words included in your corpus for this search is also indicated on status line, towards the right.

If you find a specific portion of one of these discourse examples that is of particular interest to you, you can cut it out and save it by first highlighting the relevant portion with your mouse, and then copying it into your word processor. (Notice that the tab symbol, which MonoConc displays as a black rectangle, is properly displayed as a tab once you paste the text into your word processor.)

When you are ready to go on to the next search, you do *not* have to close the window containing the previous search results. Just leave it on your screen, so you will be able to compare it with later search results if you like. You can have as many search result windows on your screen as you want, as they do not seem to affect MonoConc's performance.

2. When is the word 'well' is used in discourse?

Select **Concordance/Search**, and type as the search string: **well**. (Just the word **well**, don't type the punctuation.) What do you see? Use the scroll bars to scroll through the data. What patterns can you identify?

Now examine the data in a more systematic way. Sort the search results you just got for 'well' according to the first word that appears to the left of the word 'well', by selecting **Sort/1st Left**. Scroll through the data. What do you notice now? How does the sorting make the discourse patterning clearer?

Now **Sort** by **1st** (word to the) **Right**. What do you notice? How informative is this sort order compared to '1st Left', for this word?

Now let's look at a more general overview of the discourse patterns involving this word. Select

Collocation, then **Collocate Frequency**. How does this summarize more systematically what you already saw informally in your sorts of the data? What can you see here that you might not have noticed with just a sort?

3. What would a more detailed view of the most frequent collocations involving the word 'well' tell me about how it is used in discourse?

Looking over the Collocate Frequency from the previous search, you notice that some words to the immediate left and right of the word 'well' occur with much higher frequency than others. You may also get a sense that the word 'well' has several different functions. To some extent these can be teased apart by looking at in more detail at the specific collocations that 'well' participates in. To do this, decide on three or four words from among the top ten collocations listed under '1st Left', and do a follow-up search for each one.   For example, select **Concordance/Search**, then type in the phrase **yeah well**. Next try **okay well**. Then do similar searches for several of the other top ten collocates, including words to the right (e.g. **well I**). How does this let you zero in on the function(s) of 'well'?

4. What if I want to find out what words collocate with my original collocations?

You can continue the cycle by asking for collocate frequency data on your collocates themselves. For example, first do a **Concordance/Search** for **well I**. Then select **Frequency/Collocate Frequency** (or type **CTRL+f**). What patterns of collocation do you see here? What generalization can you make about the verbs involved? You can also **Sort** the search results by **1st Right** and scroll through the examples to get a more detailed picture. (Always sort the discourse data by making your search results window active. Don't try to sort the statistics window -- it doesn't make much sense to sort these statistics, and MonoConc won't let you do it.)

Of course you could continue this cycle still further by selecting **Concordance/Search** and specifying three words as the string to search for. But as you add more words the number of "hits" (examples of your search string that are found in your data) will decrease rapidly, and you may not find enough examples to get a generalization. This of course will depend to a great extent on the size of your corpus.

5. What patterns do you find in the way the noun 'reason' is used?

Search for **reason**. What do you see? But notice that this leaves out plurals. To remedy this we can make use of MonoConc's "wild-card" symbols.   For example, the symbol ***** represents any string of symbols. (There are also wild-card symbols that allow you to search for exactly one character; for one or zero characters; and for several words, e.g. between 2 and 5 words.)

So **Search** for **reason***. What do you get? Notice that the use of a wild card means that your search picks up several different search terms. To see more clearly how the various search terms pattern, **Sort** by **Search Term**. Scroll through the results. What do you see?

To weed out unwanted items, such as (in this case) verb forms, highlight the lines in question with the mouse, and then use **Control-d** to delete them from the listing.

Now examine the remaining data in a more systematic way. **Sort** by **1st Left**. What do you notice? **Sort** by **1st Right**. What do you notice?

Do **Frequency/Collocate Frequency**. How does this summarize what you already saw in your sorts of the data?

Notice that the statistics provided here cover just the words that remained in your window at the time you initiated the Collocate Frequency calculation. The irrelevant items that you just deleted using **Control-d** are not included, which is of course a good thing. For the most accurate picture of your data, you should do any necessary weeding out of irrelevant examples *before* you calculate your statistics.

6. When is the verb 'come', in all its forms, used in discourse?

You could get many of the verb forms by **Search**ing for **com***. Try this. Then **Sort** by **Search Term.** What do you see?

You could get rid of the many unwanted items by highlighting them with the mouse, then deleting using **Control-d**. But notice that this is a bit of a hassle, and you still wouldn't have the past tense form **'came'.** Instead, you can specify individually each form of the verb **'come'** that you *do* want, using the **Append Search** option, as follows:

Select **Concordance/Search,** and search for **come**. Then, while you are still in this window, select **Concordance/Advanced Search** (or **CTRL+a**), and check the box for **Append Search**. Now search for **comes**. Notice that this *adds* the instances of the word 'comes' to the end of your previous listing: it puts both sets of examples into the same window, rather than making a new window. For present purposes, this is useful.

Now do another Append Search, by selecting **Concordance/Advanced Search** (or **CTRL+a**), and again checking the box for **Append Search**. (You have to check this box each time you do an Append Search.) Now search for **came**. Again, the hits for 'came' will accumulate in the same window as before. Finally, do an **Append Search** for **coming**.

Now that you have all of the verb forms of 'come' in one listing (in one window), examine the data in a more systematic way. **Sort** by **1st Left**. What do you notice? **Sort** by **1st Right**. What do you notice?

Do **Frequency/Collocate Frequency**. How does this summarize what you already saw in your sorts of the data? How does having all the verb forms in one window provide a clearer picture than if the data were separated into four separate windows?

7. When is the discourse particle 'um' used?

Search for **um**. What do you see? **Sort** by **1st Right**. **Sort** by **1st Left**. What do you see?

But this leaves out lots and lots of cases, because of the different ways **um** is uttered -- for example, this word is often lengthened. To fix this we can use the wild card that allows for one or zero characters, that is, **%**, plus the wild card for an unlimited number of characters (including zero), i.e. *****. (You may need to check to make sure that these characters are indeed the current wild cards used by MonoConc, and that they are *not* listed among the "characters to treat as delimiters". Both functions can be specified by selecting **Concordance/Search Options**. and making sure the right characters appear in the appropriate boxes.)

Search for **u%%m***. Then **Sort** by **Search Term**. What's the difference? How does this change your

understanding of the role that the word **um** serves in discourse?

There are likely to be a few unwanted items here, which you could weed out by hand. Or better, try searching for **u%%m%%**, then **Sort** by **Search Term**.

Turn-sensitive phenomena

8. What words tend to occur in turn-initial position?

Recall that the symbol for a speaker attribution label is a colon (in the conventions of Discourse Transcription 1 [DT1]), or a semi-colon (in DT2). So we can expect that this symbol should occur whenever a new speaker starts talking, because of the speaker label. We can take advantage of this fact in our searches. We'll use the wild card ***** to stand for any string of symbols -- in this case, it will give us any speaker's name, because these always end with a colon (in the Discourse Transcription system).

Do a **Concordance/Search** for ***:** (that is, for asterisk followed by colon, with no spaces). (This is for data transcribed according to DT1. For transcriptions in DT2, do the search with an asterisk followed immediately by semi-colon: ***;** .)

Select **Frequency/Collocate Frequency** (or type **CTRL+f**). Look over the column of the '1st (word to) Right'. This corresponds to the first word immediately following a speaker label, that is, the first word of a "turn". (Notice that this is actually a pretty crude approximation to real turns, since it does not separate out backchannels, as one could argue should be done.) What words are most frequent in turn-initial position?

To see examples of these turn-initial words in context, return to the search results (discourse data) window for ***:**. Then **Sort** by **1st Right**, and scroll to an interesting (frequent) word, as indicated in your Collocate Frequency listing.

9. What words typically occur in turn-final position?

View the **Collocate Frequency** for speaker labels as above. Look over the column headed '1st Left'. Why does this give you turn-final words? What words are the most frequent in this position? To see examples in context, first go back to the search results (discourse data) window above (i.e. for ***:**). **Sort** by **1st Left**, then scroll to an interesting word

10.What words occur as single-word turns (e.g. backchannel, etc)?

Search for ***: * *:** (Remember to type the spaces just as they appear here, that is, one space between the first and second "word", and one space between the second and third "word". Because spaces count as word boundaries, MonoConc will interpret this search string as a sequence of three "words" in exactly this order: a speaker label, followed by exactly one word, followed by another speaker label. In other words, a turn of exactly one word in length.) Scroll through the words -- what patterns do you notice?

11.What words occur in short turns (two-word to four-word turns)?

Select **Concordance/Search Options**. Set the **"matches between"** variable to stand for 2 to 4 words. (You may need to check to make sure that the **"matches between"** symbol is **&**, and that it does not appear on the list of "Characters to treat as delimiters". You can do this at the same time, i.e. by selecting **Concordance/Search Options**.)   Select **OK**.

Search for ***: & *:**   What do you see?

When done, you may wish to select **Concordance/Search Options** again to reset the **"matches between"** numbers back to what they were before.

Other wild-card uses

12.When do compound verbs and participles occur?

You could search for everything containing a hyphen (using ***-***), but this would produce too many unwanted items, including lots of truncation symbols (double hyphens). So instead.... (First check **Concordance/Search Options** to make sure that **$** is the wildcard symbol for "exactly one" character, and ***** is the symbol for "zero or more" characters. Also check that neither symbol is listed under "characters to treat as delimiters".)

Search for **$*-$*ing**

Search for **$*-$*ed**

Search for **$*-$*s**

What do you find this way?

13.How do genitives and definites interact?

Use **Concordance/Search Options** to set the **"matches between"** numbers to 1 to 7 words. Then, do **Concordance/Search** for **the & of the**. What does this kind of search give you?

14.How do speakers use special voice qualities and prosodic modifications?

Use **Concordance/Search Options** to set the **"matches between"** numbers to 1 to 10 words. **Search** for **<* & *>**   What do you see?

To look more narrowly at just special qualities (i.e. as used by speakers imitating another's speech, making fun of someone's dialect, and so on), **Search** for **<VOX & VOX>**   What are these speakers doing?

Setup options

This section offers some suggestions on how to modify the various options that MonoConc offers. The most significant options are those found under the headings of **Concordance** (**Search Options**) and **Collocation** (**Frequency Options**).   These options can make a big difference in the efficiency of your searches, and even in their accuracy. So it is well worthwhile to give careful thought to choosing the most effective settings for your research. In many cases this will involve changing the settings slightly as you ask different research questions.

Choosing the most effective settings for your research is easier once you have gained some experience using concordances. You may wish to work with concordances for a while first, and then, when you know better what you are looking for, pay attention to the details of setting the options.

General Setup Options

Select **Concordance** to set up the **Search Options** the way you want them. Some suggestions, oriented toward discourse research:

Set the **"Max search hits"** to the maximum, i.e. 32000

Set the **"Context characters"** to 120 or so, meaning 120 characters to the left of your search term, and 120 characters to the right. (Choose whatever number fills up your screen.)

Set the **"matches between"** character to **&**.

Set the **"matches between"** numbers to **1 to 5**.

Set the **"wildcard character"** for **"0 or more"** to *****.

Set the **"wildcard character"** for **"0 or 1"** to **%**.

Set the **"wildcard character"** for **"exactly 1"** to **$**.

Delimiters & Skipping Characters

The following are suggested initial *default values* for delimiters and skipping characters for the published transcriptions in the Santa Barbara Corpus of Spoken American English (e.g. for transcriptions using **Du Bois et al. 1993** conventions **[**=DT1]). Type the characters into the appropriate spaces in the dialog box. (Do not type any spaces between the characters, nor at the beginning or end.)

- ◆ Characters to treat as delimiters: _-+0123456789.
- ◆ Skipping characters: []~^`\/#=%!?,@
- ◆ Neither skipping nor delimiters: $&{}<>():
- ◆ MonoConc characters (defaults): &?%*

15. What difference does the choice of delimiters make to the way a search turns out?

To understand what a difference your list of **"characters to treat as delimiters"** makes, first select **Concordance**, then **Search Options**. Erase all the characters found in the box labeled **"characters to treat as delimiters"**, and then then type the following characters (with *no spaces between them*) into the box in their place: **; ! # ^ + / \ | ` ~ "**

Then **Search** for the word **mhm**. How many instances of the word are there in the data, according to this search?

Now add (type in) the following additional characters in the **"characters to treat as delimiters"** box (again without any intervening spaces): **0123456789** Repeat the **Search** for the word **mhm**. Now how many instances of the word are there in the data, supposedly? Why did the change occur?

Now add in the following characters and repeat the search: **[ ]**   What change is there in the number of **mhm**'s found in the data, and why?

Finally, add in the intonation-unit final symbols: **, . ?** Repeat the **Search** for the word **mhm**. How many additional instances of **mhm** turn up this time? Why?

To round out the picture, this time keep the same set of delimiters, but broaden the search formula. That is, **Search** for the formula **m%h%m***. Now **Sort** the results by **Search Term**. How many **mhm**'s are there that fit this pattern? What are the new instances that appear?

Now let's try a trick to look for some more **mhm**'s that may not have been found yet by our previous searches. Search for **hm** (*not* **mhm**). Then **Sort** by **Search Term**, and look through the data. Do you see some instances of **mhm** here? Were they found in previous searches? Why or why not? Why do they show up in this search for **hm**?

Now remove the numerals and square brackets **[ ] 0123456789** from the list of delimiters. **Search** for ***m%%h%%m***   Then **Sort** by **Search Term**. Remove any inappropriate items by highlighting them and deleting with **Control-d**. Are there new instances of **mhm** that were not turned up by previous searches? What kinds?

16. How does your choice of delimiters affect your research? Why might you need to do a search more than once, with different sets of delimiters?

For interest's sake, you might want to consider how well **mhm** correlates with turn beginning (before and after it). Compare the number of **mhm**'s found in the previous search with the number for the following two searches:

*: *m%%h%%m*

m%%h%%m *:

*: *m%%h%%m* *:

Skipping Characters

In recent versions of MonoConc Pro, the option of specifying "skipping characters" has been introduced. Select **Concordance/Search Options**, and type the relevant characters into the box marked **Skipping characters**. If you specify a character as a skipping character, this means that a MonoConc search will ignore it, acting for all intents and purposes as though the symbol were not present in the word. This is very useful if there is a laugh symbol in the middle of a word, or the symbol for lag/lengthening, or an overlap bracket. This way you can have your cake and eat it too: you can have detailed transcriptions that acknowledge the reality that in discourse things sometimes happen in the middle of a word, and you can also achieve easy word recognition in your words searches. For this reason, it is useful to experiment with which characters to specify as skipping characters, following the lines suggested by the previous discussion of delimiters. (Note that under **Advanced Search**, you need to check the box marked **Use skipping and equal characters**.)

The following characters are recommended to be specified as "skipping characters" for data in Discourse Transcription format, whether DT1 (Du Bois et al. 1992) or DT2 (Du Bois 2004). They are recommended as initial default values, but there is room to experiment with additions to and subtractions from the list.

- Skipping characters: []~^`\/#=%!?,@
- Neither skipping nor delimiters: $&{}<>():
- MonoConc characters (defaults): &?%*

Stop List

Set up a "**Stop List**" of "words" to be ignored in your searches. (Depending on your research goals, this could include pseudo-words like **((BEEP))**, **<<FOOTSTEPS,** speaker attribution labels, and possibly even laughter, breathing, and so on.) To do this, first select **Collocation** , then **Frequency Options**. Check the box marked "**Content words only**", then select **Edit**. Look at the words in the stop list. **Add** or **Remove** words individually if you need to, one to a line.

If necessary, **Load** a stop list from a file. For example, **Load** the file named **stoplist.txt**, which you may find located in the appropriate corpus directory (such as **\Data\English\SBCorpus\StopList**).

If you need to do a lot of work on the stoplist, you may find it useful to use an ASCII text-editor like the Windows programs Notepad or WordPad to edit a list (as an ordinary plain text file in ASCII format). You can then **Load** this file into MonoConc.

Note that for some purposes you might want to include vocal noises such as **(H), (TSK)**, and so on, as words, for purposes of looking at collocate frequency information. If so, you should remove them from the stoplist you are using, or check the box **"Count all words"** under **Frequency/Frequency Options**.

17. What difference does a Stop List make to how the results of a search turn out?

To understand what a difference your Stop List makes, first select **Collocation**, then **Frequency Options**, then check the box marked "**Content words only**". Then **Search** for the word **um**. Get the **Collocate Frequency** (under **Collocation**). Take note of the most frequent words that co-occur with **um**. What patterns do you see?

Now select **Collocation** again, then **Frequency Options**, then check the box marked **"Count all words"**. Then **Search** again for the word **um**. (It is necessary to make a new search in order to get MonoConc to use the new Frequency Options settings you have just selected.) Again, get the **Collocate Frequency** (under **Collocation**). Now take note of the most frequent "words" that co-occur with **um**. What is the difference here? Which of the two pictures of the data you have just created is more enlightening for the word **um**? Are there other words for which the opposite choice might be better?

Regular Expressions

"Regular expressions" represent a powerful way to analyze your discourse data. They are somewhat technical in nature, and may require some getting used to. They are included here as an option for more advanced computer users.

The following "regular expressions" may be useful for searches in the Santa Barbara Corpus, Parts 1 and 2, 1st edition (i.e. transcribed following conventions in Du Bois et al. 1993). The format of the formulas is that used in the concordance program MonoConc, but similar formulas may be useful for carrying out this kind of search in other programs as well.

Regular expressions are part of the "**Advanced Search**" function in MonoConc, which you can access by pressing **CTRL+a**. For more information, consult the help function in MonoConc: **\Info\Help**, and search for "**Regular expression**".

Please note that it will make a big difference in these searches whether or not you have checked the box "**Use skipping and equal characters**". Try it both ways to see what happens. Also, you may wish to experiment with the choice of characters to be specified as skipping characters, and with the choice of delimiters.

The following are suggested initial *default values* for delimiters and skipping characters (for transcriptions using **Du Bois et al. 1993 [DT1]** conventions):

- Characters to treat as delimiters: _-+0123456789.
- Skipping characters: []~^`\/#=%!?,@
- Neither skipping nor delimiters: $&{}<>():
- MonoConc characters (defaults): &?%*

In the table below, the first column specifies the target, that is, the symbol (or an example of it) that you are looking for. The second gives the regular expression, what you type into the search box in MonoConc Pro. The third column gives a gloss, naming the category being searched for.

Target	Expression	Gloss
&	&	IU continued
~	~	pseudo-graph
#	#	real name (LDC Part 1)
!	!	real name (LDC Part 2)
=	=	prosodic lengthening (DT1)
@	@	laugh sign, single
<@	<@	long feature: laughter begins
[	\[	overlap begins
]	\]	overlap ends
@@@	@+	one or more laugh signs
@@@	\b@+\b	one or more laugh pulses, surrounded by word boundary
XXX	\bX+\b	one or more unintelligible syllables, w/ word boundary
@word	@[a-z]+	laughing while speaking a word
,	,	continuing intonation (comma)
?	\?	appeal intonation (question mark)
.	[^\.]\.\s	final intonation (period)
...	\.{3}	pause, medium or long (3 dots)
..	\s\.{2}\s	pause, short (2 dots, with surrounding whitespace)
--	--	truncated IU sign (two hyphens)
--	\s--	truncated IU sign (with whitespace preceding)
--	\S--	truncated IU sign (with NO whitespace preceding)
-	[^-]-\s	truncated word (followed by whitespace)
- --	-\s--	truncated IU in which last word of IU IS truncated
--	[^-]\s--	truncated IU in which last word of IU is NOT truncated
-	[^-]-\s[^--]	truncated word NOT followed by IU truncation
-	[a-z]+-[a-z]+	hyphenated word

Target	Expression	Gloss
((MIC))	\(\([A-Z]+\)\)	comment, one-word
(SNIFF)	[^\(]\([A-Z]+\)[^\)]	vocalism
X-ray	[a-wyz]*x-*[a-wyz]+	word containing one letter X, but not only X (e.g. XX)
(TSK)	\(TSK\)	click
: (TSK)	:\s*\.*\s*\[*[0-9]*\(TSK\)	click at the beginning of turn (ignoring pause, overlap)

The following notations are for the version of the Santa Barbara Corpus which is time-stamped in seconds (i.e. the LDC published version).

Target	Expression	Gloss
13.27	\s[0-9]+\.[0-9]+\s	one timestamp (with surrounding whitespace)
@ 13.27	@+\s[0-9]+\.[0-9]+\s	laugh followed by whitespace, then timestamp
@	@\d*\]*\s[0-9]+\.[0-9]+\s	IUs ending in a laugh, with or without overlap
(H)	\(H\)\d*\]*\s[0-9]+\.[0-9]+\s	IUs ending in (H), with or without overlap (NB: square brackets and numbers are removed from the "delimiters" list]
(Hx)	\(Hx\)\d*\]*\s[0-9]+\.[0-9]+\s	IUs ending in (Hx)
((MIC))	\)\)\d*\]*\s[0-9]+\.[0-9]+\s	IUs ending in comments ((COMMENT))

Expression	Gloss
\s[0-9]+\.[0-9]+\s[0-9]+\.[0-9]+\s	two timestamps in a row (w/ whitespace)
\.\s[0-9]+\.[0-9]+\s[0-9]+\.[0-9]+\s	final intonation (period, then two timestamps)
[?,&_\.\-]\s[0-9]+\.[0-9]+\s[0-9]+\.[0-9]+\s	IUs ending in ? , . _ - &
[^?^,^&^_^\.^\-]\s[0-9]+\.[0-9]+\s[0-9]+\.[0-9]+\s	APPROXIMATELY: IUs with no final boundary tone marker (but note the substantial number of false positives)
\([0-9]*\.+[0-9]*\)	will include all and only pauses in DT2; for example (.) (..) (...) (0.7) (13.29)

[Rev. 12-Dec-2005]