Transcription Manual for
The Student-Transcribed Corpus of Spoken American English

version: 18 February 2023

Contents

Section A: Formatting

Section B: Tokenization

Section C: Fillers, disfluencies and performance issues

Section D: Specific transcription conventions

Section E: Capitalization

Section A: Formatting

1. Encoding

1.1. All corpus text files are UTF-8 encoded.
1.2. The characters in the spoken text of the text files are restricted to the following set:

- lower case standard Latin letters [a-z],
- upper case standard Latin letters [A-Z],
- the punctuation signs . (full stop), , (comma), ? (question mark), - (hyphen),
- the inverted undirected single commas ' and inverted undirected double commas ".

That means that all other symbols, such as numbers, (1,2,3), umlauts (ä, ö), tildes (ñ), accents (é, â), every-day and mathematical abbreviations symbols (+, #, %, $) etc., are not found in the corpus at all. Foreign words that include such special characters are adapted so that they only include the characters outlined.

Examples:
not:
El Niño, the mélange
but:
El Nino, the melange

1.3. The following characters are found only in time stamps:

- numbers, (1,2,3), squared brackets [] and colons :.

1.4. The transcripts are saved with the .txt file extension.
1.5. The audio files corresponding to every transcript text file is saved with the .mp3 file extension.

2. Time stamps

2.1. The corpus is completely time-stamped. A time stamp indicates at what point in time the following transcribed text appears in the corresponding audio file. It consists of a set of three numbers indicating hour, minute and seconds of the audio file respectively. The three numbers are separated by colons and surrounded by squared brackets.

Example:
[00:09:53]
A time stamp indicating the following text can be found at nine minutes, fifty-three seconds into the audio file.

2.2. Time stamps are positioned at the beginning of new lines. In other words, they are the first element of the transcribed sentences.
2.3. A tab space is inserted between the time stamp and the subsequent text material.
2.4. The seconds counter of the time stamp indicates only full seconds. Thus, seconds are not rounded up to the next full second. For example, a time such as 4 minutes, 22 seconds and 890 milliseconds is indicated as [00:04:22], not as [00:04:23].

Section B: Tokenization

1. Purpose

1.1. A sentence token is a complete, coherent syntactic unit.
1.2. The terms ‘sentence token’ or ‘token’ for short will be used interchangeably.
1.3. The term ‘tokenization’ is used here to refer to the systematic division of the transcribed speech into sentence tokens.
1.4. All speech in all transcripts are split into tokens, i.e. they are fully tokenized.
1.5. The purpose of this section is to explain in detail how tokenization is implemented.

2. Token format

2.1. Every token starts on a new line.
2.2. Every token is introduced by a time stamp and followed by a tab space (see A.2.3).
2.3. Every token has a final punctuation sign, either a full stop, `.`, or a question mark, `?`. In other words, the final character of every line must be one of those two symbols (see F.1).

Examples:

```
[00:00:00]   Welcome to this week's "Top Stock Picks".
[00:00:02]   I'm Tracey Ryniec.
[00:00:03]   And I'm joined at the chairs this week by
             Sheraz Mian.
[00:00:06]   And we have a couple of interesting stocks.
[00:00:09]   One is an old Dow component.
[00:00:11]   And the other one is a weight loss company.
```

Every new token starts on a new line. Every token is introduced by a time stamp (shown in light grey). There is a tab space between the time stamp and the beginning of the token. The final character of every line is a final punctuation sign, here a full stop (shown in dark grey).

The following cases illustrate incorrect transcription that are not used in the text files.

not:

```
[00:15:03]Today, it's the other way round.
[00:15:03] Today, it's the other way round.
```

but:

```
[00:15:03]   Today, it's the other way round.
```

In the first two, wrong formats, the text begins immediately after the time stamp or following a space after the time stamp. The correct format has a tab between time stamp and the text of the tokens.

not:

```
[00:01:04]   Stop it!
```

but:

```
[00:15:03]   Stop it.
```

The first, wrong format uses an exclamation mark as the final punctuation. The final punctuation signs can only be full stops or question marks.

3. General definition of ‘token’

3.1. The notion of sentence token is defined in terms of three necessary and sufficient characteristics. A token is a chunk of text that

(i) includes at least **an overt subject**,
(ii) as well as **a finite verb**, either an auxiliary or a lexical main verb,
(iii) and is contained within a **complete**, **independent main clause**, which means
a) it can potentially stand on its own and “feels complete” (i.e. it includes all of the non-optional, required, important, argumental, selected elements) and
b) it does not fulfill a syntactic function in relation to another element.

Examples:

```
[00:17:14]      All politicians lie.
```

A minimal sentence token. It consists only of a subject (in green) and a finite verb (in yellow). These elements occur in a main clause, which means that it can stand on its own and does not have a syntactic function in relation to another element.

```
[00:06:26]      Is the only place on top of this hill?
[00:06:31]      It is.
```

Another minimal sentence token. It has a subject (in green) and a finite verb (in yellow). These elements occur in a main clause that “feels complete” in this context (the predicate “the only place” is understood from the preceding sentence) and that is independent.

The following sentences exemplify common cases that violate the definition of a token and that are therefore *not* transcribed as their own tokens.

Examples:

```
[00:19:11]      That law was advanced under Ronald Reagan
                and   was signed by the first George Bush.
```

The string highlighted in orange **does not have a subject** (missing subject in green) even though it does have a finite verb (in yellow) and occurs in an otherwise complete, independent main clause. It is therefore not transcribed as a sentence token but instead is grouped together with the preceding material.

```
[00:01:14]      Some of us need a lot, others   less.
```

The string highlighted in orange has a subject (in green) but **lacks a finite verb** (missing verb in yellow). Otherwise, it occurs in an independent main clause. It is therefore not transcribed as a sentence token but grouped together with the preceding material.

```
[00:00:46]      So, she was a California high ... but anyway,
                high school teachers should have almost no
                limits in regards to what they speak about
                with their students.
```

The string highlighted in orange has a subject (in green) and a finite verb (in yellow). However, it occurs inside a chunk that **does not “feel complete”**: the non-optional predicate of *was* is not uttered in its entirety. The string is therefore not transcribed as a sentence token but instead is grouped together with the following material.

```
[00:23:22]      He said, 'I can guarantee the ceremony will
                be dry.'.
```

The string highlighted in orange has a subject (in green) and a finite verb (in yellow) occurring inside a complete main clause. However, the main clause takes on a syntactic function with respect to another sentence element: It functions as the complement (object, content) of the very *said*. Therefore, the string is not transcribed as a sentence token but instead is grouped together as direct speech with the preceding material

3.2. A token usually includes not just an overt subject and a finite verb, but several additional elements (dependents, constituents, phrases). (These elements have a direct or indirect syntactic relation to the finite verb.) The following is a fairly comprehensive, but by no means exhaustive, list of potential elements with a syntactic function that occur inside a larger token.

- (1) complements (simple, non-propositional, non-clausal), such as objects, indirect objects, nominal or adjectival predicates, complement prepositional phrases etc.

  Examples:
  [00:02:45] They make Moen faucets.
  [00:00:13] He likes it.
  [00:00:27] I gave her some nice sleepwear.
  [00:00:18] My name is Richard Alley.
  [00:01:21] That's pretty cheap.
  [00:00:15] I'm from Connecticut.

- (2) modifiers (simple, non-propositional, non-clausal) of all types and in a wide sense, such as adverbial elements, negation, non-complement prepositional phrases, extent and measure phrases, focus markers, secondary predicates, question adjuncts (*wh*-phrases), left dislocations, pragmatic / discourse markers, connectives, fillers, interjections, etc.

  Examples:
  [00:02:30] They beat as well.
  [00:01:33] It still continues.
  [00:03:08] They ship worldwide.
  [00:03:11] They're not super cheap.
  [00:00:30] The students wouldn't have studied then.
  [00:09:31] In particular, it's a Wall Street Journal article.
  [00:09:46] They've been trading for twenty one years.
  [00:09:06] We've gone from fifty dollars to eight hundred and thirty dollars.
  [00:19:54] Real estate is going up twenty percent a year, every year.
  [00:26:59] It's only fair.
  [00:08:40] I did see you naked.
  [00:06:08] How can they taper?
  [00:54:20] The clients who sold their foreign stocks, they never bought them back.
  [00:27:47] So, it's kind of like democracy.
  [00:13:36] Well, sure, everybody would want that.
  [00:04:41] So, you know, I mean, I can deal with it.
  [00:14:28] Yeah, it might be his legacy.

  For fillers (e.g. *you know*, *I mean*, *right*), see C.3
  For interjections (e.g. *yeah*, *ouch*, *eh*), see D.4

- (3) conjunctions introducing a main clause, *and*, *but*, *or* (see B.4 below).

  Example:
  [00:00:43] And I'll take your questions.

- (4) non-finite main verbs, the non-finite marker *to*, and adverbial particles, co-occurring with auxiliaries (modal, perfect, progressive, passive), other grammaticalized verbal expressions, phrasal verbs, serial verb constructions, etc.

  Examples:
  [00:02:44] They were booming.
  [00:06:09] Paul Volcker was appointed.
  [00:18:41] The dollar has responded.
  [00:19:26] Somebody was being interviewed.
  [00:15:57] People need to go bankrupt.
  [00:14:56] We have to pay off debt.
  [00:03:03] You gotta get new ones.
  [00:02:58] It sold off.
  [00:08:21] Interest rates go up.
  [00:09:48] Let's go check the other store.

- (5) complement (argument) clauses, both finite and non-finite, embedded with no subordinating words, or with grammatical complementizers such as *that, if, for* etc.

  Examples:
  [00:06:52] You'd think that they would do that.
  [00:02:59] I'm wondering if the room is gonna be big enough.
  [00:17:50] I didn't think I could add that much value.
  [00:05:45] They don't want to admit it.
  [00:29:28] You can see them saying all this stuff.
  [00:12:20] He had a job working as a waiter.
  [00:14:17] It takes time for conclusions to be agreed upon.

- (6) modifying (adjunct) clauses, finite and non-finite, that are introduced by subordinating words such as *since, if, as, when, because, although, so, so that, that, (in order) to, for, after, who, which, where* (for relative clauses), etc.

  Examples:
  [00:08:29] And over a very long period of time, weeks and weeks, you might come back down, if you were still alive.
  [00:07:10] The only year in which U.S. corporations bought back more stock was in 2007.
  [00:14:16] And not that I'm necessarily a big fan of Kennedy, but Kennedy said it.
  [00:01:19] They are created in order to mislead the public about how bad inflation really is.
  [00:28:13] The debt lasts about two and half years, where literally one third of that seven and a half trillion is due in under a year.

- (7) appositives, parentheticals, clarifications, closer definitions, corrections, epithets, titles, etc. These are elements that elaborate on a previous element in some way.

  Examples:

  [00:05:01] I mean, they beat the crap out of this company, the only way they know how to do it, with lawyers and with fines.

  (The highlighted element describes how the "beating" event happened exactly)

  [00:14:35] They're gonna steal it from the rich, the rich greedy people who don't really deserve it anyway.

  (The highlighted element gives a more detailed explanation of *the rich*)

  [00:11:48] And you get money, five K. or so per turbine per year.

  (The highlighted element defines more closely the actual quantity of *money*)

  [00:13:23] Figure out what you want, what works for you.

  (The highlighted element may correct, define or give an alternative to *what you want*)

3.3. Sentence elements can be coordinated (with *and*, *but*, *or* or nothing). Coordinated elements form part of a larger token and not independent tokens (except for independent main clauses with an overt subject and a finite verb, in accordance with the general definition of 'token').

Examples:

[00:10:53] Make sure you like, █ comment and subscribe.

[00:36:48] It's submergence and loss of fresh water.

[00:12:21] And that step is already and constantly taken with women and girls.

[00:02:05] And to top our cupcakes, we'll be using these yellow, █ red, █ green and orange candy gems.

[00:21:27] So, we're kind of training or retraining.

[00:14:15] If I took that wig off, and if I took the eyelashes off, then what it's gonna do is force the writer to write like a woman.

Illustration of coordinated finite verbs, predicates, modifiers, non-finite verbs, modifying clauses etc. The conjunctions (including coordination with nothing) are highlighted in red.

3.4. Disfluencies, i.e. incomplete chunks of speech, are grouped together with preceding or following material into a complete token (see section C).

Examples:

[00:24:56] I ... it's heavy wat ... it's wa ... well, it's heavy water.

[00:02:58] Everything about gaming ... there's just something I love about every single game.

[00:01:55] So the Earth is four billion ... and three billion years ago, there was a major change in the ocean.

[00:04:40] Trump isn't gonna be last in the oratorical regard, hopefully the other things.

The highlighted elements are disfluencies (incomplete or fragmentary chunks).

4. Clarification: Tokenization of coordinated main clauses

4.1. The correct tokenization of coordinated main clauses follows from the general definition of ‘token’ outlined above. The following section clarifies potential problems.

4.2. Coordinated main clauses with overt subjects. Main clause tokens can be introduced by the conjunctions *and*, *but*, *or*. Following the general principle of tokenization, sequences of main clauses introduced by such conjunctions are split into separate tokens.

Example:

not:

```
[00:27:28]    But now we need more troops down there and
the government might end up having to resort to a draft
ER in the next ER administration, but in any event, the
war is gonna get bigger, it's gonna cost more money, and
it's gonna require more troops.
```

but:

```
[00:27:28]    But now we need more troops down there.
[00:27:30]    And the government might end up having to
              resort to a draft ER in the next ER
              administration.
[00:27:36]    But in any event, the war is gonna get
              bigger.
[00:27:39]    It's gonna cost more money.
[00:27:41]    And it's gonna require more troops.
```

The example above includes a series of main clauses, several of which are introduced by conjunctions (in red). Nevertheless they are all complete main clauses with a subject (in green) and a finite verb (in yellow). They are therefore separate tokens.

4.3. Coordinated main clauses without overt subjects. Coordinated main clauses sometimes lack an overt subject. In this case, only the verbal part of a clause is conjoined, and not the whole sentence. The clauses may be introduced by overt conjunctions (*a and b, a but c, a or d)*. However, they can also be coordinated without an overt conjunction (*a*, *b and c*, *a,b,c*). Following the general principle of tokenization, clauses with missing overt subjects under coordination are not treated as separate tokens but are grouped together with the main clause that includes the overt subject.

Examples:

not:

```
[00:05:59]    Anyway yesterday C.N.B.C. actually came out.
[00:06:01]    And did a big piece.
```

but:

```
[00:05:59]    Anyway yesterday C.N.B.C. actually came out
              and did a big piece.
```

This example includes two main clauses linked by a conjunction (*and*, in red). However, only the first clause includes an overt subject (*CNBC*, in green). The subject of the second clause is understood to be identical with it, but is not actually expressed. In other words, the first verb (*came*, in yellow), but not the second finite verb (*did*, in yellow) occurs with an overt subject (in green). The two sentences are therefore analyzed as one single token.

not:

```
[00:04:47]   And they beat the crap out of this guy.
[00:04:49]   Just smashed him.
[00:04:50]   Hit him on the head with the phone.
[00:04:52]   Beat the crap out of him.
[00:04:53]   Put him in the hospital ER because he was
             saying negative things about this stock.
```

but:

```
[00:04:47]   And they beat the crap out of this guy, just
             smashed him, hit him on the head with the
             phone, beat the crap out of him, put him in
             the hospital ER because he was saying negative
             things about this stock.
```

This sentence includes a sequence of finite verbs (in yellow), *beat / smashed / hit / beat / put*, which all share the subject of the first sentence, *they* (in green). The sentences are therefore grouped together into one token.

5. Clarification: Tokenization of potential main clauses

5.1. There are several instances of difficult tokenization that involve potential independent main clauses. This section clarifies this issue.

5.2. Main clause with a subject and a finite verb may be appositives elaborating on a previous element. In this case, it is subjective whether this main clause should be regarded as rather independent (and so should form its own token), or whether this main clause is tightly connected to the element it clarifies (and so should not form its own token). The transcriber has to decide such cases on an individual basis.

- The **default** is to transcribe such main clauses as independent tokens.
- **Exceptional circumstances** that may lead to the transcription of such main clauses as part of a larger token are: the explained element is short or would often be used anaphorically (*this*, *that*); there is a certain parallelism with previous tokens that suggest apposition; phrases such as “namely” or “that is to say” feel very natural between explained element and main clause.

Example:

The following passage can be transcribed in two ways.

In the area of climate change, it's this, if we put heat-trapping gas in the atmosphere, it will trap heat.

Option 1:

```
[00:00:19]   In the area of climate change, it's this.
[00:00:21]   If we put heat-trapping gas in the
             atmosphere, it will trap heat.
```

Option 2:

```
[00:00:19]   In the area of climate change, it's this,
             if we put heat-trapping gas in the
             atmosphere, it will trap heat.
```

This sentence involves a main clause (*If we put… ill trap heat*) with a subject (green) and a finite verb (yellow) that could potentially stand alone. This main clause should therefore form its own token (Option 1). However, the main clause can also be interpreted as a closer explanation of *this* (purple). It can therefore be transcribed together with the previous token (Option 2). (In this case, the transcriber decided on the second option).

5.3. Interrogatives such as *you know what* can either be a fixed idiom to be transcribed within a larger token or an independent main clause questions forming its own token. The transcriber has some subjective room to decide such cases on a case-by-case basis.

- If there is no marked intonation and no break between the idiomatic questions and the subsequent material at all, they are transcribed within the larger token. Usually the question is just a rhetorical device to convey something like, "wait, let me clarify, let me simplify, I have changed my mind."
- If there is a clear question intonation and / or a break between the question and the subsequent material, they are transcribed as independent tokens. Usually the question is a genuine question, inviting a definite answer to a problem or open point.

Option 1:

```
[00:05:59]     Okay, you know what, I like recording all
               my let's plays.
```

Option 2:

```
[00:25:09]     And, you know what?
[00:25:10]     As somebody who isn't running for office,
               because I don't know how in the world I
               would win with my reputation, I ... I maybe
               ... maybe I'm a hypocrite.
```

Illustration of two different kinds of tokenization of the string *you know what*.

5.4. There are cases of coordinated main clauses where both conjuncts depend on an initial epistemic adverb, most importantly *maybe*. In this case, the second main clause feels so dependent on the epistemic adverb that both clauses can be transcribed as one token.

Option 1:

```
[00:01:29]     And maybe I was just bad at it.
[00:01:30]     And there's better techniques to do it.
```

Option 2:

```
[00:01:29]     And maybe I was just bad at it and there's
               better techniques to do it.
```

Illustration of two different kinds of tokenization of the patterns 'maybe clause and clause.' The second option was chosen for this sentence.

6. Clarification: Tokenization of potential subordinate clauses

6.1. The correct treatment of subordinate clauses follows from the general definition of 'token' given above. Since subordinate clauses fulfil a syntactic function in relation to another element (complement clauses, modifying clauses etc.), they never form tokens. The following section clarifies issues arising in connection with the tokenization of subordinate clauses.

6.2. Certain subordinate clauses may feel relatively independent. In fact, some speakers may use some traditionally subordinating words to introduce independent main clause. This concerns in particular the words ***because, although* and *though***. However, these words **are always transcribed as part of a larger token** by default, not as initial words of independent tokens.

Examples:

not:

[00:01:36] They can continue ER to sell bonds to the public at very, very low rates of interest.

[00:01:39] **Because** they would demand a much higher inflation premium.

but:

[00:01:36] They can continue ER to sell bonds to the public at very, very low rates of interest **because** they would demand a much higher inflation premium.

The example above contains a subordinate clause introduced by the subordinating word *because* (*because they would demand...*). It must therefore be included within the previous main clause.

not:

[00:00:56] ER, it is a short-term top.

[00:00:58] **Although** I don't think ER it's gonna be a short-term top ER for months.

but:

[00:00:56] ER, it is a short-term top **although** I don't think ER it's gonna be a short-term top ER for months.

This example contains a subordinate clause introduced by the subordinating word *although* in red (*although I don't think...*). The subordinate clause is treated as a dependent element contained in the higher clause.

6.3. Subordinate clauses can form exceptional tokens in a small number of special cases, most importantly as answer fragments (see below B.10.10).

6.4. Degree complement clauses depending on degree words such as *so*, *such* etc. (*so cold that it froze*) are transcribed within the token of the degree words. This is in accordance with the general principle of tokenization.

Example:

[00:03:27] Something hit Mars so hard that a little rock was launched into space.

However, the complementizer of the complement clause can be omitted (*so cold it froze*). As a consequence, ambiguity may arise between an interpretation of a clause as a degree complement clause without a complementizer or an independent main clause token.

Example:

Option 1:

[00:04:56] So, you know, they're so tiny it's kind of unbelievable.

Option 2:

[00:04:56] So, you know, they're so tiny.

[00:04:58] It's kind of unbelievable.

Transcribers have to decide such cases on an individual basis.

7. Clarification: Tokenization of coordinated subordinate clauses

7.1. The correct tokenization of coordinated subordinate clauses follows from the general definition of ‘token’ outlined above. In general, the tokenization is unproblematic. There are, however, some ambiguous cases. This section clarifies issues surrounding this issue.
7.2 Subordinate clauses can be coordinated with conjunctions such as *and*, *but*, *or*. In this case, the subordinate clause conjuncts are all tokenized together as one unit within their superordinate clause as one token.
In particular, grouping several conjoined subordinate clauses together into the same token is unambiguous in the following cases.

(a) The subordinating word is repeated.

Example:

[00:01:39] It's one of the most engaging and powerful demonstrations of **how** the earth's climate is changing and **how** carbon dioxide is increasing.

This example includes a finite subordinate clause introduced by *how* (*how the earth's climate is changing*, in blue). It is conjoined with another subordinate clause using a conjunction (*and*, in red). The second clause also has an overt subordinating word, a second *how* (*how carbon dioxide is increasing*, also in blue). Hence, the grouping of the two finite subordinate clauses into the same token is unambiguous.

(b) The conjunct clauses are headed by non-finite verbs.

Example:

[00:01:38] All I can do is **tell** you some of the things that I've witnessed and then **hope** that it may help.

This example includes one main clause with a subject (*all I can do*, in green) and a finite verb (*is*, in yellow). In addition there is a non-finite predicate headed by an infinitive (*tell*, in blue). The non-finite predicate is conjoined with another clause using a conjunction (*and*, in red). The second clause cannot be an independent token because it is also headed by a non-finite, not a finite, verb (*hope*, also in blue). Hence, the grouping of the two non-finite conjuncts into one token is unambiguous. (Analyzing the second conjunct as an independent token would be impossible because it does not have a subject as well.)

7.3. However, it can also be a subjective choice whether a string should be transcribed as several conjoined subordinate clauses in one token (*because a and b*), or as a subordinate clause and a main clause in several tokens (*because a. And b.*). Such ambiguity arises if the subordinating word of a finite subordinate word is not repeated after a conjunction. In this case, the material is tokenized as it seems to make best sense subjectively. The transcriber has to decide such cases on an individual basis.

Example:

Option 1:

[00:07:23] The only thing they can do is recognize ER **that** there's a huge problem and they need to get out of the way.

Option 2:

[00:07:23] The only thing they can do is recognize ER
that there's a huge problem.
[00:07:26] And they need to get out of the way.

This example includes a subordinating word, *that*, shown in bold face, which introduces a subordinate clause (*that there's a huge problem*, in blue). It is followed by a conjunction (*and*, in red) and another clause (*they need to get out of the way*, in blue).The second clause can plausibly be conjoined with the first clause ('recognize two things: that there is a huge problem and (that) they need to get out of the way.' However, it can also be interpreted as an independent token ('recognize only one thing: that there is a huge problem,' followed by a new thought, namely '(I think) they need to get out of the way.' The choice between the two transcriptions is subjective to some degree. (In this case, the transcriber decided on the first option.)

7.4. The correct resolution of the issue described in 7.3. is subjective. However, some guidelines can be used to decide which tokenization is more appropriate:

- **Context** may make it quite clear that the two clauses depend on the same head in their superordinate clause and are therefore coordinated subordinate clauses belonging to the same token.
- There may be a certain degree of **parallelism** and repetition between the two clauses making it clear that they are indeed coordinated subordinate clauses belonging to the same token.

The following sentences exemplify potentially ambiguous cases that have been found to be better tokenized as a single unit based on context and parallelism.

Examples:

[00:06:12] And they showed **that** there was periods of
time when oil was rising and there was
periods when oil was falling.

This example includes one main clause with a subject (*they*, in green) and a finite verb (*showed*, in yellow). In addition, there is a subordinating word, *that*, which introduces a subordinate complement clause shown in blue (*that there was periods of time when oil was rising*, in blue). The clause is conjoined with another clause using a conjunction (*and*, in red). The second clause does not occur with a subordinating word (*there was periods when oil was falling*, also in blue). However, the subordinating word is understood to be *that*, interpreted from the preceding subordinate clause because the two clauses are quite similar. Hence, context makes it clear that the second clause in blue is on the same level as the first clause in blue. Both the first and the second clause in blue are the complement of the finite verb (*showed*, in yellow), meaning 'showed [that there was periods when oil was rising] and (sowed that) [there was periods when oil was falling]'. Therefore, the second clause does not form an independent token.

[00:04:56] But one amazing thing it does is it
preserves a very detailed record of the
climate history **because** the snow falls on
the ice, that snow compresses, and then it
traps air in little, tiny bubbles.

This example includes one main clause with a subject (*one amazing thing it does*, in green) and a finite verb (*is*, in yellow). In addition, there is a subordinating word, *because*, which introduces a subordinate adverbial clause shown in blue (*because the snow falls on the ice*, in blue). The clause co-.occurs with two additional clauses, the second of which has a conjunction (*and then*, in red). The two clause do not occur with a subordinating word, i.e. no repetition of *because*. However, the subordinating word is understood to be repeated from the preceding subordinate clause because the three subordinate clauses have a logical connection, first snow falls, then it compresses, then air bubbles are trapped. Hence, context makes it clear that the second and third clause, also in in blue, is on the same level as the first one. The three clauses are subordinate to the subordinator *because*, meaning ‘[because the snow falls ...] (and because) [that snow compresses] and then (because) [it traps air in little, tiny bubbles]’. Therefore, the second and third clauses do not form an independent token.

8. Tokenization of sandwiched and gapped clauses

8.1. This section covers the tokenization of three types of structures:

- Sandwiched clauses (*She came, that’s clear, to London*) (8.2.)
- Right-node raising (*She came, and he came, to London*) (8.3.)
- Gapping (*She came to London, and he to Berlin*) (8.4)

8.2. Sandwiched (parenthetical) clauses. Sandwiched (or parenthetical) clauses are complete main clauses inserted into another clause. Abstractly speaking, a parenthetical clause `X` is contained within a clause `Y`, `[Y ... [ X ] ... ]`. Put differently still, a parenthetical clause occurs in between two sentence elements of a higher clause (e.g. between the subject and the finite verb, between a subordinate clause and an object, etc.). Sandwiched main clauses are not transcribed as an independent token, but as part of the token in which it is contained. Sandwiched clauses are separated by commas.

Examples:

[00:00:52] **I really**, there's a few encounters with the real world, **came** to understand how science is just a tiny, tiny part of this change.

This example includes a complete main clause shown in purple with a subject (*there*) and a finite verb (*‘s*). It is embedded within a larger token shown in blue. Specifically the sandwiched clause is placed between the subject and an adverb (*I really*) and the finite verb (*came*) of the higher main clause, shown in bold face.

[00:01:31] In fact, the Chinese economy **in 2010**, that's next year, **will** be back up to close to eleven percent.

This example includes a complete main clause shown in purple with a subject (*that*) and a finite verb (*‘s*). It is embedded within a larger token shown in blue. Specifically the sandwiched clause is placed between a temporal adjunct (*in 2010*) and the finite verb (*will*) of the higher main clause, shown in bold face.

[00:09:53] **If Americans can't borrow money**, and that's what's gonna be the end of this, right, **the people who have been lending us the money** are going broke and going out of business.

This example includes a complete main clause shown in purple with a subject (*that*) and a finite verb (*‘s*). It is embedded within a larger token shown in blue. The sandwiched clause is placed between an initial adverbial subordinate clause

(*If Americans can't borrow money*) and the subject (*the people who have been lending us the money*) of the higher main clause, shown in bold face.

[00:32:04] The reason they don't wanna make it public is **because it will embarrass them**, that's why, **and because they don't want somebody like me on "The Daily Show"**.

This example includes a complete main clause shown in purple with a subject (*that*) and a finite verb (*'s*). It is embedded within a larger token shown in blue. Specifically the sandwiched clause is placed between a subordinate clause (*because it will embarrass them*) and another subordinate clause conjoined to the first with *and* (*and because they don't want somebody like me on "The Daily Show"*) of the higher main clause, shown in bold face.

8.3. Right node raising. Right node raising describes a construction that involves the start of a clause, interrupted by another, incomplete clause, followed by the completion of the clause, where some elements are interpreted as part of both the first and interrupting clauses. In other words, a right node raised element Z completes both a starting clause Y and a subsequent interrupting clause X, [Y ... _ [X _]] Z. Put differently still, a displaced dependent belongs to both a higher clause and an incomplete, sandwiched clause. The second incomplete, clause is often introduced by a conjunction like *and*. Right node raising structures are transcribed as one token. The incomplete clauses are separated with commas.

Examples:

[00:02:06] It's my opinion, and it was my opinion back then, that we would not stop making cars if General Motors went bankrupt.

The complement clause *that we would not stop making cars if General Motors went bankrupt* shown in pink, depends on the main clause *It's my opinion* shown in blue. However, another clause, *and it was my opinion back then*, shown in purple, intervenes between the core of the main clause and the complement clause. It is sandwiched between the two. The complement *that*-clause also depends on the intervening clause, *it was my opinion ... that we would not stop*. The *that*-clause is the complement of both *It's my opinion* as well as *it was my opinion back then*. Right node raising structures like this are transcribed as one single token.

[00:05:27] We borrowed, and we spent, too much money.

The phrase *too much money*, shown in pink, is the object of both *borrowed* in a higher clause, in blue, and of *spent* in an incomplete, sandwiched clause, shown in purple. Right node raising structures like this are transcribed as one single token.

[00:00:40] All of us here have some experience with the ocean, whether it's sitting on a beach or, if you're lucky, maybe snorkeling over, or diving over, a coral reef.

The phrase *a coral reef*, shown in pink, complements both the preposition *over* in a higher non-finite clause, in blue, and the preposition *over* in an incomplete, sandwiched non-finite clause, shown in purple. Right node raising structures like this are transcribed as one single token.

8.4. Gapping. Gapping is here used as a wide cover term for structures in which at least the finite verb and potentially other material are missing in a clause, but are understood, or can be reconstructed, from a preceding clause. In abstract terms, a complete clause X forms the basis for the reconstruction of the meaning of an incomplete clause Y, [X (complete)], [Y (incomplete)]. The incomplete clauses are tokenized together with the complete clause on which their meaning depends. The incomplete clauses are separated with a comma.

Examples:
Bill drank beer, Sally wine.
The above sentence is understood as *Bill drank beer, Sally ~~drank~~ wine*. The finite verb *drink* is missing in the second incomplete clause and reconstructed on the preceding clause. The incomplete clause therefore forms one token with the previous, complete clause.

[00:00:01] The NASDAQ is up seven points, the Dow
 down fifty three points.
The above sentence is understood as *The Dow ~~is~~ down fifty three points*. The finite verb *is* is missing in the second incomplete clause and reconstructed on the preceding clause. The incomplete clause therefore forms one token with the previous, complete clause.

[00:22:57] So, not the borrower, but the lender,
that's who's gonna be at risk, and also
the American tax payer .
The above sentence is understood as *And also the American tax payer ~~is~~ ~~who's gonna be at risk~~*. The finite verb, *is*, as well as the entire rest of the clause after the subject, *who's gonna be at risk*, are missing ("stripping"). The incomplete clause therefore forms one token with the previous, complete clause.

[00:00:15] And then will it be the airlines, and
then some real estate companies maybe?
The above sentence is understood as *And then ~~will it be~~ some real estate companies maybe?* The finite verb, *will*, as well as other elements, the subject *it* and the non-finite verb *be*, are reconstructed on the previous material. The incomplete clause therefore forms one token with the previous, complete one.

[00:01:15] Initially, when they proposed it, it
applied to future ER college students.
[00:01:21] Then it applied to current students, and
now, the past.
The above sentence is understood to mean *and now ~~it applies to~~ the past ~~students~~*. The finite verb, *applies*, as well as other elements, the subject *it* and the preposition *to*, are reconstructed on the previous material. The incomplete clause therefore forms one token with the previous, complete clause.

[00:14:17] It takes time for conclusions to be agreed
upon, and details sorted out.

The above sentence is understood to mean *and ~~it takes time for~~ details ~~to be~~ sorted out.* The finite verb, *takes*, as well as other elements, the subject *it* and the object *time* as well as the complementizer *for* and the chunk *to be*, are reconstructed on the previous material. The incomplete clause therefore forms one token with the previous, complete one.

Clarification: A clause's missing material must include the finite verb for it to count as gapping and be grouped together with the preceding complete clause as one token. Otherwise, the normal tokenization rules apply. In particular, if an adjunct clause (*if, when, in order to*, etc.) is felt to belong to a preceding as well as a following clause, it is transcribed only with the first, clause and a new token starts for the second clause.

<u>Examples:</u>

not:

[00:12:32] If the confidence number comes out higher than Wall Street had thought, there's usually a big rise in the stock market, and the dollar strengthens.

but:

[00:12:32] If the confidence number comes out higher than Wall Street had thought, there's usually a big rise in the stock market.

[00:12:40] And the dollar strengthens.

not:

[00:05:23] When you're counting on a log scale, you don't count one, two, three, you count one, ten, a hundred.

but:

[00:05:23] When you're counting on a log scale, you don't count one, two, three.

[00:05:27] You count one, ten, a hundred.

not:

[00:09:20] And in order to create that impression, he ignored corporations, and he ignored the government itself.

but:

[00:09:20] And in order to create that impression, he ignored corporations.

[00:09:25] And he ignored the government itself.

In the examples above, there is a clause, shown in grey, that may be felt to be incomplete because it may be said to miss an adjunct clause, such as the *if*-clause, the *when*-clause or the *in order to*-clause in the first, second and third example respectively. However the supposedly incomplete clause has a subject (in green) and a finite verb (in yellow). They are therefore treated as their own, independent token. A possibly missing adjunct clause is not enough to argue for a gapping interpretation. Instead, a finite verb must at least be missing.

9. Tokenization of direct speech

9.1. Direct speeches are chunks of complete clauses or fragmentary phrases from quotes, citations, readings or other imitations of someone else’s speech.
9.2. Direct speech is placed in single inverted, undirected commas, ' ... '.
9.3. The first letter of direct speech is capitalized.
9.4. If the direct speech consists of a fragment, for instance, a nominal phrase, an interjection etc., there is no final punctuation sign within the quotation marks ' ... '. If the direct speech involves a complete sentence, either a declarative sentence or a question, the direct speech includes a full stop or a question mark, respectively, '.' or '?', indicating the end of the direct speech sentence.
9.5. As always, there will be a token final punctuation sign for tokens that embed direct speech. If a direct speech sentence comes at the end of a token, as is frequently the case, that means that there can be two final punctuations next to each other ' ... .'. or ' ... ?'. – one for the end of the direct speech sentence, the other for the end of the whole token.
9.6. Direct speech is often introduced by a clause with **a verb of saying**. In this case, the preceding clause and all the direct speech after the verb of saying form one token. A comma is used to introduce direct speech. This is true for fragments as well as declarative sentences or questions.

Examples:

```
[00:01:51]     And she answered, 'A cat'.
```

This example includes direct speech (*A cat*). The direct speech is introduced by a verb, *answered*. The direct speech is a fragments, namely just a nominal phrase. It forms one token with the preceding clause.

```
[00:03:54]     It says, 'The outlook on the U.S.
               triple A. credit rating was raised to
               stable.'.
```

This example includes direct speech (*The outlook on the U.S. triple A credit rating was raised to stable*). The direct speech is introduced by a verb of saying, *says*. The direct speech is a complete declarative clause. It forms one token with the preceding clause. The token thus involves the punctuation sequence .'. at the end, one full stop for the direct speech sentence, one for the whole token.

```
[00:20:10]     And I said, 'Well, how much do you
               make?'.
```

This example includes direct speech (*Well, how much do you make?*). The embedding main clause has a verb of saying, *said*. The direct speech consists of a question. It forms one token with the preceding clause. The token thus involves the punctuation sequence ?'. at the end, one question mark for the direct speech question, the other for the whole token.

```
[00:09:08]     The scholarship which says, 'Using this
               knowledge makes us better off.' has in
               it integrated an assessment model.
```

This example includes direct speech (*'Using this knowledge makes us better off.'*). The embedding main clause has a verb of saying, *says*. The direct speech consists of a declarative clause. It forms one token with the preceding clause. The direct speech sentence has its own punctuation mark, a full stop. The embedding clause continues after it and then finishes with its own full stop.

9.7. Direct speech after a **verb of saying** may consist of more than one sentence, i.e. a sequence of sentences. In this case, all sentences are included within the direct speech. Every sentence occurs with its own final punctuation sign.

Example:

[00:18:18] And you tell him, 'Hey, I'm gonna have a baby. I'll see you in a few months. Save my job.'.

This example includes a quotation consisting of three main clauses (firstly, *I'm gonna have a baby*, secondly, *I'll see you in a few months*, thirdly, *Save my job*). The direct speech comes after a verb of saying, *tell*. The direct speech chunk is treated as a single unit and forms one token together with the preceding clause.

9.8. Direct speech can sometimes be introduced, or depend on, words that are **not verbs of saying**. These include subordinators such as *that*, *how*, quotative *like*, nouns such as *question*, *answer*, and others. Direct speech after such items can sometimes feel somewhat incoherent, but is treated analogously to direct speech after verbs of saying. The direct speech can consist of non-clausal phrases or complete sentences.

Examples:

[00:03:09] And I was like, 'Probably'.

This example includes direct speech (*Probably*). It is introduced by quotative *like*. The direct speech is a fragment, i.e. not a complete sentence with a subject and a finite verb but just an adverb and so does not appear with its own final punctuation. It forms one token with the preceding clause.

[00:13:28] You know, he talked about how, 'I want an America that's not dependent on Saudi oil.'.

This example includes direct speech (*I want an America that's not dependent on Saudi oil*). The direct speech is introduced by an unusual expression, *talked about how*. The direct speech is a complete declarative clause and so occurs with its own final full stop. It forms one token with the preceding clause.

9.9. Direct speech depending on words that are **not verbs of saying** can also consist of a sequence of fragments and sentences. As before, the entire sequence is transcribed as one direct speech and every complete sentence occurs with its own final punctuation sign.

Example:

[00:12:45] So the question arises again, 'Have we passed a point of no return? Is ice melt sure to increase so that A.MOC shutdown is a foregone conclusion?'.

This example includes a quotation consisting of two questions (firstly, *Have we passed...* secondly, *Is ice melt sure...*). The direct speech is an appositive on the noun, *question*. Since the direct speech consists of two questions, they both occur with their own question marks. The direct speech chunk is treated as a single unit and forms one token together with the preceding clause.

9.10. Direct speech can also occur **without any introductory words** at all. Direct speech can then only be recognized by contextual clues, such as a change in the voice of a speaker. Such independent direct speech units can consist of non-clausal phrases or fragments. In this case, they are grouped together into one token with the preceding material.

Example:

[00:18:33] Even the Pledge of Allegiance has ER been in the ER the news recently, ER 'And to the republic for which it stands'.

This example involves a direct speech segment (*And to the republic...*). It is a fragment because it does not have a subject and a finite verb to form a token on its own. Therefore, it does not occur with its own final punctuation sign. The direct speech is grouped together with the preceding material into one token.

9.11. Direct speech that is **independent of any introductory words** may be a complete sentence. In this case, the direct speech forms its own token. There will then be no final punctuation sign outside of the quotation marks. Indications are a mocking or imitating voice, an obvious shift in speaker perspective, or reading and writing speech.

Example:

[00:31:47] It's not like there's not enough space on the internet.

[00:31:50] 'Oh, the internet is too crowded.'

[00:31:52] It will cost them nothing to make my video public.

This example involves a direct speech segment that consists of a main clause (*Oh the internet is too crowded.*). There is no introductory word at all. Instead, the pragmatic context makes it clear that this is a direct speech segment. The relevant unit is the speaker's mocking imitation of people who think that there is not enough space on the internet. The direct speech segment forms its own, independent token. There is no full stop outside of the quotation marks.

9.12 Direct speech **without any introductory words** may also consist of a sequence of sentences. In this case, every sentence forms its own token in accordance with the general definition of 'token.' Every independent direct speech token of that kind will occur with their own quotation marks and without a final punctuation sign outside of the quotation marks. Again, reasons for such speech may be mockery, perspective shift, reading and writing, etc.

Example:

[00:03:13] Eventually investors are gonna get tired.

[00:03:13] 'Okay, we've been growing revenues for ten years, twenty years, thirty years.'

[00:03:13] 'We need a profit.'

[00:03:13] 'Where is the profit?'

This example includes independent direct speech without any introduction. It includes three main clauses (firstly, *Okay, we've been growing* .., secondly, *We need a profit*, thirdly a question, *Where is the profit?*). The voice of the speaker changes in a mocking way. Each sentence forms its own token. They each have quotation marks but no final punctuation sign outside of the quotation marks.

10. Exceptional tokens

10.1. Definition. Exceptional tokens are tokens with their own time stamp starting on a new line that nevertheless do not conform to the general principle of tokenization. The following points list the most common types of exceptional tokens.

10.2. Imperatives. Overt subjects are usually missing in imperative sentences. Nevertheless, imperative sentences form independent tokens. Imperatives are independent tokens even in cases where the imperative and the subsequent clause may be felt to have a tight connection.

Example:

not:

[00:28:31] Now imagine this, what's gonna happen to the cost of servicing that debt if interest rates go to six percent or seven percent?

but:

[00:28:31] Now imagine this.

[00:28:32] What’s gonna happen to the cost of servicing that debt if interest rates go to six percent or seven percent?

The first part of this example is an imperative (with the verb *imagine* in yellow). Even though there is no overt subject, it forms its own token.

not:

[00:02:17] Believe it or not, Western science did not reach that conclusion until the middle of the nineteenth century.

but:

[00:02:17] Believe it or not.

[00:02:18] Western science did not reach that conclusion until the middle of the nineteenth century.

The first part of this example is an imperative (with the verb *believe* in yellow). Even though there is no overt subject, it forms its own token. A tight connection may be felt to hold between the expression *believe it or not* and the subsequent clause. Nevertheless, the imperative is tokenized independently.

[00:00:00] Let's see.

The verb *let* is an imperative in this sentence and so heads an independent token.

However, it can be difficult to decide if an imperative should form its own exceptional token (as in the examples above), or if they are lexical fillers, like *I mean*, *you know* (see below, C.3.1). This concerns in particular the words *Look* and *See*, but also others. There is some subjective choice for the transcribers to decide if such imperatives are independent exceptional tokens, or if they should be transcribed as fillers together with the following material. Independent imperatives should be long (2 or more words) and be pronounced fully and often with a salient intonation Fillers should be short (1-2 words) and may have a rapid, reduced pronunciation or unremarkable intonation.

Example:

[00:00:54] ER but look, I mean, this is a company.

here, the imperative *look* is transcribed as a filler together with the following material

10.3. Expressions of greetings, farewell and thanking. Material that expresses a greeting, a farewell or thanks can be combined into one token even if a subject or finite verb is not present. The token includes as much material as can reasonably be grouped together with the greeting, farewell or thanks, such as repetitions (*hello, hello*), temporal specifications (*see you later*), degree specifications (*thank you very much*), etc.

Examples:

```
[00:00:00]   Hi.
[00:00:00]   Hi everybody.
[00:00:01]   ER welcome ER welcome to my show.
[00:00:57]   See you then.
[00:17:26]   So thank you.
[00:04:54]   Anyway, thanks everybody for listening.
```

10.4. Sentences with missing light, functional material. Sentences may lack light, functional material and can still be transcribed as a token. Light, functional material refers to:

(1) the first person subject *I*,
(2) the non-thematic, dummy subject *it*,
(3) the non-thematic, dummy subject *there*,
(4) the auxiliary verb *be*,
(5) the auxiliary verb *have* and
(6) other personal pronouns if their reference is quite generic, such as *we*, *you*, *they*.

If a sentence does not include any of these items but could be made complete by the addition of nothing more than these elements, it can be transcribed as an independent token.

The following examples illustrate exceptional tokens with missing light material.

Examples:

missing *I*:

```
[00:14:22]   ER don't wanna get myself into trouble.
```

This sentence does not have an overt subject. The first person singular pronoun *I* seems to have been dropped. The material can be reconstructed as a main clause, ***I*** *don’t wanna get myself into trouble*. It is therefore analyzed as an independent sentence token.

missing dummy subject *it*:

```
[00:07:13]   Doesn't matter whether the government actually
             takes the money or takes the purchasing power.
```

The main clause does not have an overt subject. The dummy subject *it* seems to have been dropped. The material can be reconstructed as a main clause, ***It*** *doesn't matter whether the government actually takes the money or takes the purchasing power*. It is therefore analyzed as an independent sentence token.

missing auxiliary verb *be*:

```
[00:09:24]   The Dow now down fifty three points.
```

This sentence does not have a finite verb, but it can easily be reconstructed with the copula *be* as *The Dow* **is** *now down 53 points*. It is therefore analyzed as an independent sentence token.

missing dummy subject *there* and *be*:

```
[00:09:11]   Lots of red and green ER for my stock
             symbols.
```

This sentence token does not have a main verb or subject in its main clause. However, the intended meaning can be recovered by analogy with the common *There is* existential construction. The sentence means, ***There is*** *lots of red and green for my stock symbols.* It is therefore treated as one independent token.

missing dummy subject *it* and *be*:

```
[00:26:14]   Interesting that before the communists took
             over, they had no weather problems, right?
```

The main clause can be reconstructed as ***It is*** *interesting that...* and it is therefore treated as one token.

missing dummy subject *there* and *have*:

```
[00:18:05]   Never been a paper currency that didn't
             become worthless.
```

This sentence token does not have a finite verb or a subject in its main clause. However, the intended meaning can be recovered by analogy with the *There is* existential construction. The sentence means, ***There has*** *never been a paper currency that didn't become worthless.* It is therefore treated as one independent token.

missing dummy subject *they* and *be*:

```
[00:01:44]   It was fascinating.
[00:01:45]   Nice, interesting people.
```

The main clause can be reconstructed as ***They were/are*** *nice, interesting people* and the phrase is therefore treated as one token.

Other examples

```
[00:04:19]   One more analogy to the stock market.
[00:11:52]   Pays a lot better than the steers down at
             the bottom.
[00:02:22]   Well, ER hard to say.
```

However, wherever the transcription of such sentences as exceptional, independent tokens can be avoided, an alternative transcription is preferred. They are a resort option, not a default. In particular, clauses with missing light functional material are transcribed where possible as gapping structures (see above B.8.4), grouped together with a preceding complete clause that allows the reconstruction of the missing material.

Examples

not:

```
[00:00:01]   The NASDAQ is up seven points.
[00:00:03]   The Dow down fifty three points.
```

but:

```
[00:00:01]   The NASDAQ is up seven points, the Dow down
             fifty three points.
```

Here, the second clause is incomplete as it misses the auxiliary *be* (*The Dow ~~is~~ down fifty three points).* However, the preceding clause has a parallel structure and involves the verb *be* overtly (*The NASDAY* **is** *up seven points*). The clause with the missing light functional material is therefore grouped together with this preceding sentence and not transcribed as its own independent token. This is the preferred, default transcription wherever possible.

not:
[00:21:13]   ER and you can see it kind of ... some
weird stuff over here too.
[00:21:17]   Same kind of thing.
but:
[00:21:13]   ER and you can see it kind of ... some
weird stuff over here too, same kind of thing.
Here, the phrase *same kind of thing* is interpreted as appositive on *some weird stuff* ('some weird stuff namely / to explain / to repeat: same kind of thing') rather than as an independent token with missing subject and copula ('This is the same kind of thing').

10.5. The conditional imperative and consequence construction. The conditional imperative and consequence construction refers to patterns of the form imperative clause, followed by another clause, where the imperative is interpreted as a condition and the subsequent clause as a consequence. Such structures can be transcribed as one token not as two. The main clause is frequently introduced by the conjunction *and* but not necessarily. The imperative may occur with an overt subject, such as *you*. The imperative and the following sentence are separated by a comma.

Examples:
[00:07:39]   Just blame it on the weather, **and** you've got a
pass.
[00:10:46]   Stand with Rand, **and** you got a better chance than
standing again ER with ER ... with Mitt Romney.
[00:06:58]   You stick to the Equator **and** you'll be fine.
[00:01:48]   I mean, you look at his numbers with women, he's
having a eleven point slide so far.
These examples show an imperative, shown in orange, followed by *and*, shown in bold face, and a second clause, shown in blue. The imperative is interpreted conditionally ('if you just blame it on the weather...', 'If you stand with Rand...'). The two sentences are therefore grouped together into one token.

10.6. The correlative *Not only* construction. Structures that include the expression *not only* plus a clause, followed by a second clause, are treated as one token, not two. The second clause may be introduced by words such as *but* or *but also* but is also often just a main clause. The clause introduced by *not only X* is separated from the main clause with a comma. The second main clause may be triggered by the unit *not only* occurring inside a subordinate clause.

Examples:
[00:11:52]   But not only should we hate them because they're
rich, but we should also hate them because of how they got
rich.
This example is a correlative *not only* constructions continued by *but also* in the main clause (in red). It is transcribed as one token.

[00:00:22]   You know, not only are these books number one
and two on Amazon, and they've been that way for a month, but

```
number three is not even close as far as where the numbers
are.
```

This example is a correlative *not only* constructions continued by *but* in the main clause (in red). It is transcribed as one token.

```
[00:20:17]    Yet not only do they not understand that,
they're actually praying for that to happen.
```

This example is a correlative *not only* construction (in red) continued by a main clause without a particular introduction. It is transcribed as one token.

```
[00:01:29]    So young people need to not only understand the
problem, ER they need to understand realistic solutions.
```

10.7. The correlative *Either ... or ...* construction. Structures that include the expression *either* plus a clause, followed by *or* and a second clause, are treated as one token, not two. The two clauses are not separated by a comma. Commas may only appear before the *or* clause if they are independently required.

Examples:

```
[00:14:07]    They're either not gonna get paid or they're gonna
              get paid in money that has no value.
```

This example is a correlative *either* ... *or* construction (in red). The second *or* clause is not separated from the first *either* clause by a comma.

```
[00:09:28]    Well, either healthcare costs went up, which is
              not good, or we're just sicker.
```

This example is a correlative *either* ... *or* construction (in red). A comma is required for the clause-adjoined *which* relative clause (*which is not good*), so that the second *or* clause is separated from the first *either* clause by a comma as an accidental consequence.

10.8. The correlative *The ... the ...* construction. Structures that include the expression *the* plus a comparative, followed by a second *the* and a comparative, are treated as one token, not two. The second instances of *the* is separated by a comma.

Examples:

```
[00:01:58]    Well, the higher the rating, the more money it's
              gonna be able to get for the bonds.

[00:15:26]    The longer this thing goes on, the worse it's
              ultimately gonna get.

[00:07:00]    The hotter the colors, the more the change.
```

10.9. Question fragments. Questions can consist of incomplete sentence, involving *wh*-words without a subsequent complete clause involving a subject and a main verb. Examples are *how about X?*, *Why?*, *Where?* etc. Instances of such incomplete questions are transcribed as independent tokens. They are closed off with a question mark punctuation (`?`).

Examples:

```
[00:02:30]    Why?
[00:04:02]    Why not?
```

```
[00:12:22]   How exactly?
[00:04:43]   How come?
[00:14:50]   How about food?
[00:12:30]   Alright, how about that one?
```

These sentences do not have a finite verb. Nevertheless they are analyzed as independent tokens because they instantiates the common, lexicalized *How about X* construction.

10.10. Answer fragments. Replies to questions often consist of material that does not include a main clause with a subject and finite verb. Such fragments are transcribed as independent tokens. The token includes as much material as can reasonably be grouped together.

Examples:

```
[00:11:12]   Would you do that?
[00:11:13]   No.
```

The answer in the second token is a single word (no) forming its own token. It is the end of the recording.

```
[00:10:38]   What is this right here?
[00:10:41]   Residential.
```

A single word answer fragment that cannot be grouped with any following material.

```
[00:35:51]   Why do people know that communism is bad?
[00:35:53]   Because it's failed everywhere.
```

The second token provides an answer to the question in the first one. It is a finite subordinate clause introduced by *because*. There is no main clause with a subject and finite verb. Thus, the subordinate clause exceptionally forms a token on its own.

```
[00:15:55]   Is it gonna change anything for the working
             stiffs?
[00:15:58]   No, because everybody else is gonna keep
             doing what this guy is doing.
```

In this examples, the second token provides an answer to the question in the first one. It is introduced by the interjection *No* followed by a subordinate clause introduced by *because*. Since there is no main clause with a subject and finite verb here, the two elements form one exceptional token. That is to say, *No* is grouped together with the following material since it makes good sense together.

```
[00:00:33]   So, are prices going to go up?
[00:00:35]   Yeah, ultimately.
```

Here, the second token provides an answer to the question in the first one. It includes the interjection *Yeah* and the adverb *ultimately*. Since there is no main clause with a subject and finite verb here, the two elements form one exceptional token.

Note that answer fragments are not grouped together with sentences that can form tokens on their own. Instead, such fragments form their own exceptional tokens.

Example:

not:

```
[00:02:43]   Would you say, 'This is what's telling us
             what the market's doing in the long term.'?
```

[00:02:47] No, you'll look at the longer data set.

but:

[00:02:43] Would you say, 'This is what's telling us what the market's doing in the long term.'?
[00:02:47] No.
[00.02:47] You'll look at the longer data set.

In this example, the interjection *no* is not grouped together with the subsequent material because the following material is a main clause that can form a token on its own.

10.11. Fragmentary exclamatives. Fragmentary exclamatives are constructions with initial *wh*-elements such as *what* or *how*, but no subject or finite verb. They occur with a final full stop.

Examples:

[00:08:24] Jeez, man, what a back-up.
[01:05:19] How intense.

10.12. The *is is* construction. The *is is* construction consists of an initial nominal phrase, such as *the problem*, *the issue*, etc., followed by a form of *be* (most commonly *is*), followed by a second form of *be* (again, most commonly *is*), and a focused clause. The two instances of *is* are separated by a comma. Sometimes it can be difficult to distinguish between a disfluent repetition of two *is* (indicated with …) and a real *is is* construction (indicated with comma).

Example:

[00:00:42] But the reality is, is we need half an inch pretty consistently here.

10.13. The open *so* construction. The word *so* can be used in token-final position as a rhetorical device to invite a listener to draw a supposedly obvious conclusion. This pattern is transcribed by separating the word *so* with a comma from the preceding material, a three period ellipsis marker consisting of three, separate full stops, . . . and finally a token-final full stop.
Short words, like the interjection *yeah*, may follow the open *so* construction. In this case, the final full stop is replaced by a comma but the three period ellipsis sign is maintained.

Examples:

[00:01:51] And then, ER this ... they beat estimates four straight quarters, so ... .
[00:00:20] There's about eight hundred and seventy of those around the country, so ... .
[00:02:10] I don't really read, so ... , yeah.

10.14. Overt corrections across token boundaries. Overt corrections (see C.5) can sometimes refer to material found in one or several tokens above the current one. In such cases, the overt correction can form its own token. The material in the token is kept as short as possible and is not joined together with other material to form a token.

Example:

[00:21:56] So, when those winds relax, that water relaxes back across the Pacific, spreads warm water westward.
[00:22:02] Rising air follows the warm water.
[00:22:04] I'm sorry.

```
[00:22:05]    Eastward, spreads warm water eastward.
```

An example of an overt correction (in green) referring to material found in an earlier, non-adjacent token (in yellow). In this case, the overt correction forms its own token.

10.15. Continuations across token boundaries. A fragment may continue a sentence from several (more than one) tokens earlier. In this case, the fragment may form its own token. (Alternatively, such utterances may be described with parenthetical clauses, see B.8.1).

Examples:

```
[00:25:53]    And progressively you have your P.H.D. student,
              your master student, some undergraduate intern.
[00:25:58]    That looks dangerous.
[00:26:01]    I would not do that.
[00:26:07]    And of course your post-doc determined to find
              his or her own way ER asserting independence.

[00:00:27]    It's really just ER some merging problems that
              I could probably fix with ER maybe some ... ER
              some fixing of ... of laying these highways and
              things and also using the T. Plus Plus mod.
[00:00:37]    Now we have that back again.
[00:00:38]    I haven't really done much with that since I
              got it back again, so ... .
[00:00:42]    ER things like this, where we could probably
              have, for example, the highway come in and then
              have the exit after the entry

[00:03:57]    Alright, so then next to that picture, we have
              Drew, who's in our group.
[00:04:01]    Hey, Drew.
[00:04:02]    ER we're friends.
[00:04:04]    And ER my roommate, Amber, and some of my other
              friends.
```

The last token (beginning in green) continues an earlier token (in yellow). Even though the last token does not have a subject and a finite verb in a main clause, it can still form its own token in this context

10.16. The *clause, be X* amalgamation. A complete clause (*I'm having a nice time*) , often with a demonstrative element, can be continued with a form of *be* and a predicate (, *is what I'm saying*.). Effectively, the complete clause also functions as the subject of *be*. The amalgamation of clause, *be* and predicate are transcribed as a single token. There is a comma before *be*.

Examples:

```
[00:01:15]    We were gusting up to fifty nine, was the
              highest that I saw there.
[00:01:19]    That's the problem with revisionist history, is
              that sometimes it goes too far.
[00:11:28]    Well, that's the thing, is because to me, most
              movies that you see now are Hollywood.
```

10.17. <u>*Which* plus a complete clause</u>. *Which* can introduce clause-adjoined relative clauses (*I saw him, which is nice*). However, *which* can also introduce complete clauses, where the relativized element in the clause-adjoined relative clause is present (*I saw him, which this is nice*). Such structures are transcribed with a three period ellipsis sign after *which*.

<u>Example</u>:

```
[00:01:59]   And normally they're fifteen, which ... actually
             I don't know if that's good or bad.
```

10.18. <u>Other amalgamations</u>. There are many other cases where a phrase functions simultaneously as one function for a preceding and another function for the following speech. There is no intonational break or marked prosodic feature that would indicate that the shared phrase is more closely associated with one over the other speech segment. Wherever this is the case, it is assumed that the phrase in question forms a coherent sentence with the following, not the preceding, speech, and instead the preceding speech is treated as a disfluency, separated off with the three period ellipsis marker.

<u>Example</u>:

```
[00:01:40]   People writing in, 'Hey, Adriene, ER I think
             something up with ... your newsletter's not
             working'.
```

Section C: Fillers, disfluencies and performance issues

1. General remarks

1.1. A coherent transcription of speech involves a number of complicating aspects that are not encountered to the same degree in the treatment of purely written texts. The most important ones are:

- **Fillers** causing interruptions within syntactic units (e.g. *uhm, okay*)
- **Disfluencies**, formulations of fragmentary syntactic units (e.g. *She was … I don't know*)

Additionally, there are a number of other performance issues, e.g. uttering unintended messages (e.g. forgetting a *not*), problematic recordings etc. (e.g. noisy background sounds).
This section details how such difficulties are dealt with.
1.2. A 'filler' is understood to be any word or phrase that is conventionally used to interrupt an utterance. Alternative descriptive terms are 'hedge', 'pausing marking' or 'hesitation.' Fillers are usually needed to give the speaker time to think, pause and plan their speech. They may also be discourse markers with pragmatic meaning rather than purely processing facilitors.
1.3. A "disfluency" is defined as a syntactically incomplete, fragmentary, incoherent chunk of speech. More precisely, a disfluency is a stretch of text that is (a) syntactically incomplete (internally) and so cannot be parsed as its own complete sentence token (for definition of sentence token, see B.3.1) and is (b) syntactically incoherent (externally) and so does not fulfil a syntactic function with respect to the surrounding speech (e.g., it does not function as an object etc., for a fairly comprehensive list of syntactic functions, see B.3.2).

2. The filler ER

2.1. The English filler usually orthographically rendered as *er, erm, uh, um, uhm* etc. is transcribed with a generic marker in the text files, spelled ER (capital E, capital R without punctuation).

Examples:
ER that's where we need jobs.
And ER that ER was very interesting.
And ER it is, you know, guaranteed by that government
The marker ER in the above examples indicates positions where the speakers utters a filler word such as *um, uh, er, erm,* etc.

2.2. ER is transcribed in sequence as often as it can be heard and clearly discerned as separate instances. However, speakers generally only use a single ER in a row and so it is generally found only once. It can sometimes be difficult to distinguish between several instances of ER and a single, long ER.

Example:
the uh, uh, uh chief economist
not:
the ER chief economist
but:
the ER ER ER chief economist
A sequence of clearly discernable instances of *uh* are rendered as a sequence of ERs, not a single ER.

2.3. ER is included irrespective of whether the pausing filler is very long and clearly perceptible or very short and almost inaudible. As long as any kind of filler can be discerned, it is included in the transcription as ER (for some subjectivity, see below C.8.3.)
2.4. ER is not inherently separated by commas. By default, ER does not occur with a comma.

Examples:
not:
ER, the gold stocks finally ER, have come back to life.
but:
ER the gold stocks finally ER have come back to life.

2.5. If ER happens to occur immediately before or after a position that allows or requires separation with a comma according to the transcription guidelines, then it may co-occur with such a comma. The comma does not follow from the use of ER, but from independent comma requirements. (For the use of commas in the transcriptions, see section F.4).

Examples:
So you get a company that gives away their profits, like ER, you know, WhatsApp.
The filler *you know* is always separated with commas. Since in this example, the disfluency marker ER happens to occur just before *you know*, it is followed by a comma.

```
As a result of that, ER the Consumer Price Index was only
up one tenth of one percent.
```

Initial adjuncts like *as a result of that* may optionally be followed by a comma in Standard English. Since the disfluency marker `ER` appears here just after an initial adjunct, is may be preceded by a comma.

2.6. `ER` is preferred in token-initial position, rather than in the last position of a preceding token. In other words, `ER` does not normally occur in token-final position. This is true even if there is a short pause between the first token and `ER` but a relatively long pause between the `ER` and the second token.

Example:

not:

```
[00:01:07]   ER I remember I had tears on my eyes during the
... that election night ER.
[00:01:14]   But it did not solve the problems.
```

but:

```
[00:01:07]   ER I remember I had tears on my eyes during the
... that election night.
[00:01:13]   ER but it did not solve the problems.
```

2.7. If `ER` occurs in token-initial position, the subsequent word starts with a lower-case letter (unless upper case is independently required, e.g. for names, the pronoun *I*, etc. see section E).

Example:

not:

```
[0:00:24]    ER The other thing you mentioned is the places
I worked at.
```

but:

```
[0:00:24]    ER the other thing you mentioned is the places
I worked at.
```

2.8. On rare occasions, a filler such as *uhm* can be used or mentioned (consciously) as a word, rather than be used (rather automatically) as a filler. In such cases, the word is rendered orthographically as closely as possible to what is heard (*uhm*, *err* etc.). In particular, it is spelled `uh` (for the vowel), `err` (with `r` for an *r*-sound), `uhm` (with `m` for an *m*-sound) or `erm` (with both `r` and `m` for an *r*-sound and an *m*-sound).

Example:

```
[00:03:22]   Those are his equivalents of roughly uhm or
             scratching his head.
```

3. Lexical fillers

3.1. Lexical fillers are transcribed verbatim in the text files. The most important ones are: *you know*, *I mean*, *okay*, *alright*, and *right*. They are not transcribed as their own tokens (see B.3.2.(2)).

Examples

`And of course,` **`you know`**`, short term performance really doesn't,` **`you know`**`, mean anything with respect to ER,` **`you know`**`, if you're talking about the fundamentals of the economy.`

A token that includes instances of *you know* used as a filler.

`I mean``, I made a profit.`

A token that includes an instance of *I mean* used as a filler.

3.2. Lexical fillers are always separated off with commas.

Examples:

not:

`In order to you know represent a speculative blow-off`

but:

`In order to,` **`you know`**`, represent a speculative blow-off`

The token includes *you know* as a filler or hedge. It is therefore separated with commas.

not:

`See that's what the Fed is doing.`

but:

`See``, that's what the Fed is doing.`

The token includes *see* as a filler or hedge in initial position. It is therefore separated with commas.

not:

`You're gonna have to earn over one thousand ER dollars a week right ER not eight hundred.`

but:

`You're gonna have to earn over one thousand ER dollars a week,` **`right`**`, ER not eight hundred.`

The token includes *right* as a filler or hedge. It is therefore separated with commas.

3.3. Some speakers use the word *like* as a lexical filler. In this use, it is transcribed with the same conventions as for other fillers.

Examples:

```
[00:29:03]   Like, what is this?
[00:28:39]   It's more of, like, a chillax moment for us.
```

3.4. If a lexical filler or hedge is used in token-final position and is interpreted as a pragmatic question as if to ask a listener for attention, affection or confirmation, typically with a rising intonation, then the token is closed off with a question mark. Frequent fillers occurring in this context are *right*, *alright*, *hey* and *okay*.

Examples:

`It is T.G.I. Friday every day of the week for this guy,` **`right?`**

`And then I'll close with something,` **`okay?`**

3.5. The preferred transcription for fillers that express pragmatic questions such as `right?`, `okay?` is at the end of a token, rather than at the beginning of new one. This is true even if there is a long pause between the first token and the filler but a relatively short pause between the filler and the second token.

Example

not:

```
[00:07:13]   She has a strapless dress on
[00:07:14]   Okay, I didn't even realize that.
```

but:

```
[00:07:13]   She has a strapless dress on, okay?
[00:07:16]   I didn't even realize that.
```

There is a relatively long pause between *okay* and *I didn't*. However, since *okay* seems to be used as a pragmatic question seeking attention in this context, it is grouped together with the preceding not following token.

3.6. Imperatives that are used as discourse markers, such as *listen*, *believe me*, are transcribed as their own tokens, not as lexical fillers within a token. (For imperatives, see B.10.2.)

Example

not:

```
[00:00:32]   And so listen, is a quarter of an inch gonna
             help?
```

but:

```
[00:00:32]   And so listen.
[00:00:33]   Is a quarter of an inch gonna help?
```

The imperative *listen* is used to get the speaker's attention. It is transcribed as its own sentence token, not within a larger sentence token.

4. The disfluency marker ‘three period ellipsis, . . . ’

4.1 The disfluency marker ‘three period ellipsis,’ a sequence of three full stops, is used for the indication of all occurrences of disfluencies (for the definition of disfluency, see above C.1.3). The disfluency marker occurs to the right of the disfluency and is followed by (the rest of) a complete sentence token. That means that a certain amount of text to the left of the disfluency marker can be ignored or crossed off to result in a coherent syntactic sentence token.

Examples

```
[00:20:05]   Suppose the ... the experiment's over now.
[00:04:34]   If I'm s ... is it ... am I blocking it?
[00:01:19]   I feel like I need ... I need more crops.
[00:03:04]   You have to persuade these bureaucrats and
             these ... the Euro ... in Brussels.
[00:10:46]   Let's say this is about three ... this is a
             little more than three hund ... this is about
             three hundred ten.
```

Sections highlighted in red are disfluent chunks of speech. They are followed by the three period ellipsis disfluency marker and then grouped together with (the rest of) a complete sentence token. If the highlighted parts are left out, the result is a normal, non-disfluent sentence token.

4.2. The three period ellipsis marker ... is surrounded by whitespaces to the left and right.

Example:

not:

```
they were running a...a paid broadcast
```

but:

```
they were running a ... a paid broadcast
```

4.3. The three period ellipsis marker ... consists of three distinct characters ... and not of one symbol as provided by some fonts, like … .

Example

not:

```
And ER, you know, I mean, that is … is certainly a leading
indicator ER for the gold price.
```

but:

```
And ER, you know, I mean, that is ... is certainly a
leading indicator ER for the gold price.
```

4.4. If a token begins with a disfluency, the first word after the three period ellipsis marker ... is not capitalized but spelled with a lower-case letter (unless a capital letter is independently required for proper names, etc.).

Example

not:

```
[00:02:42]   But ... But you don't have to.
```

but:

```
[00:02:42]   But ... but you don't have to.
```

4.5. The three period ellipsis marker `...` is not used for any purpose other than the indication of disfluencies. In particular, the ellipsis symbol is *not* used in the following cases.
(a) Three period ellipsis is *not* used to indicate what might be perceived as long pauses. Pauses are not represented in the basic transcriptions at all.

Example:
not:
```
Anyway, also I thought an interesting ER topic to discuss
today would be ... the political debate.
```
but:
```
Anyway, also I thought an interesting ER topic to discuss
today would be the political debate.
```
There is a very long pause between *be* and *the* in the audio file. It is not represented by an ellipsis sign. The pause is not indicated at all.

(b) Left-dislocated phrases (i.e. elements that are resumed in the subsequent clause with a co-referential pro-form or other word) are *not* separated with `...`, but with a comma.

Examples
```
[00:54:20]  The clients who sold their foreign stocks in
two thousand, they never bought them back.
```
This token includes a left-dislocated nominal phrase (*the clients who...*) (in yellow), which is resumed in the subsequent clause with a pronoun (*they*) (in orange). The left-dislocated phrase is therefore separated with a comma (in blue), not with an ellipsis sign.

```
[00:17:27] Now in fact, a lot of the people who are
buying with zero down and using adjustable-rate
mortgages, a lot of these zero-down buyers are actually
buying a house.
```
This token includes a left-dislocated nominal phrase (*a lot of the people who are buying with zero down and using adjustable-rate mortgages*) (in yellow), which is resumed by a full nominal phrase (*a lot of these zero-down buyers*) (in orange). The left-dislocated phrase is therefore separated with a comma (in blue), not with an ellipsis sign.

(c) Three period ellipsis is *not* used for repetitions that are most likely used for emphasis and related purposes (rather than for reasons of hesitation, correction and speech structuring). Instead, such structures are separated with a comma (see section F.4).

Examples:
```
You know, this is a very dangerous, dangerous slope.
And it is truly, truly great to see you here.
And it's economically beneficial, extremely beneficial.
So they've been very, very different.
Well, it’s not really a riot, riot in the normal sense.
```

4.6. The transcriptions do not make use of more fine-grained distinctions between different types of disfluencies. Instead, ***all* disfluencies are indicated with the three period ellipsis marker**. A disfluency as such is seen as a fairly basic notion that is easy to transcribe consistently. Researchers can then use their own ontology of disfluencies on top of this global notion of disfluency. The reason is that more fine-grained distinctions of disfluencies may become more cumbersome to transcribe and may be more subjective to distinguish.

In particular, the transcripts do *not* distinguish between:

a) **corrections**, in which a speaker restarts and entirely "abandons" the previous attempt at expressing an idea, intending for the listener to ignore the message in the disfluency and interpret it as uninformative (repetitions, false starts etc.),

b) **accidents**, in which a speaker fails to formulate a syntactically complete and coherent message for reasons other than the intention to correct (lack of concentration, complicated message, a "nominal" idiosyncratic style, etc.), and still expects the listener to pay attention to the message in the disfluency and interpret it as informative.

Examples

[00:09:43] We draw a ... a ... a curve through this.
[00:59:56] Where is this going to be the back ... best?
[00:30:58] Actually think is ... I think it's valuable.
[00:02:14] It's important to remember that the ocean ... unlike the atmosphere and unlike the biosphere, the ocean reacts very slowly.
[00:16:33] There's an arena player who's like ... anyway, I never read the card text.
[00:13:44] So, Xureila, you gotta ... Xureila, where are you going?
[00:11:56] So, cash ... I mean, we had all kinds of ... I mean, we weren't poor at that time.
[00:01:15] But we have seen through conservative and religious education practices, things like abstinence only prac ... it doesn't work.

Parts highlighted in red illustrate type a), disfluencies which the speaker restarts and probably wants the listener to ignore. Parts in green illustrate type b), disfluencies, which the speaker does not complete for reasons other than correcting and probably wants the listener to pay attention to. This distinction is not made in the transcripts.

Furthermore, the transcripts also do not distinguish between disfluencies that a) involve material that is literally repeated (**repetition**) or that b) involve different material (**false starts**).

Examples:

And ... and ... and that is happening.
on their ... on their ... on their long portfolios
this ... the ... these temporary factors
These are ... it's not managed money.

Disfluencies that are repetitions, shown in purple, or false starts, shown in orange. Both types are indicated with the same three period ellipsis marker in the transcripts.

The transcripts do not distinguish between the **length** of disfluent speech parts, whether they are extremely short, medium sized or very long.

Examples

```
[00:07:45]   I ... I showed this figure.
[00:01:19]   Now ... now this is just gonna be a little
             bit too much.
[00:02:20]   There is no ... there is no way for it to
             connect to this gun debate.
[00:26:03]   And this ... I mean, that's ... I wouldn't do
             that.
[00:06:11]   The reason for that is, you know, when you
             see all the evidence of inflation and ... and
             again, rising oil prices don't cause
             inflation.
[00:09:36]   Then there's that horrible tweet that just
             went out saying, 'If you're a Trump voter, I
             would' ... my brother, who's a Trump voter,
             he would take a bullet for you.
```

Disfluencies of different sizes, ranging from a single sound, one word, three words, a relatively long phrase, a fairly long utterance, to a very long chunk of text that is almost complete. All of them are indicated with the same three period ellipsis marker.

4.7. Clarification: The three period ellipsis marker `...` is used for all disfluencies. This includes syntactically incomplete material that could be described as "**almost complete**", i.e. material that is typically long, includes a subject, finite verb and possibly several other dependents but still lacks some required element that feels somewhat trivial (e.g. just the last part of a phrase, the final element of a compound, a conjunct etc.).
The transcripts do not make use of token-final `...` after incomplete sentences of this kind. Instead, even "almost complete" utterances are treated as disfluencies joined with a three period ellipsis marker to subsequent material to form complete tokens.

Examples:

not:

```
[00:07:17]   I bet Putin's in there right now, like,
             talking to ... .
[00:07:21]   I don ... I don't know who Putin talks to.
```

but:

```
[00:07:17]   I bet Putin's in there right now, like,
             talking to ... I don ... I don't know who
             Putin talks to.
```

not:

```
[00:20:34]   They say, 'Yeah, I'm self-employed. I have
             a little website and ... '.
[00:20:47]   I don't know if these people are making any
             real money.
```

but:

```
[00:20:34]   They say, 'Yeah, I'm self-employed. I have
             a little website and' ... I don't know if
             these people are making any real money.
```

not:

```
[00:09:06]   The media is trying to say that this is a
             problem for Wall Street and that it's not
             gonna affect Main Street or that we gotta
             make sure it doesn't affect Main ... .
[00:09:14]   This is nonsense.
```

but:

```
[00:09:06]   The media is trying to say that this is a
             problem for Wall Street and that it's not
             gonna affect Main Street or that we gotta
             make sure it doesn't affect Main ... this
             is nonsense.
```

Examples of almost complete disfluencies. Like other cases, they are not transcribed with a token final three period ellipsis (shown in red), but as a long disfluency (highlighted), joined to subsequent material to form a complete token.

<u>Other examples:</u>

```
[00:11:52]   He says that, 'ER could she not be used
             to' ... but Fox News literally will have
             young conservatives on.
[00:04:57]   And, you know, as I look around, talk to
             my friends, talk to my colleagues, walk
             around the city, just working through this
             sort of well-off Western consumer culture,
             ER it seems really obvious to me that we're
             at risk of being too complacent, and that
             the risk of being too scared is ... it just
             seems silly to me to think about becoming
             too scared.
```

Where direct speech is incomplete, the default position for the three period ellipsis marker is outside of the single inverted commas (see F.5).

<u>Example:</u>

```
[00:09:49]   She said, 'I don't know if' ... I can't
             remember what she said.
```

For one exception that uses the three period ellipsis marker in token final position, see the “open *so* construction” described in B.10.13.

4.8. **Mispronounced, incomplete or unintelligible** words are treated by default as disfluencies with the offending words rendered as closely to their phonetic reality as possible using the English alphabet. The resulting disfluency is joined to the following material with `...` as per usual.

As a consequence, there are no special labels or indications of unintelligible words in the transcripts. That means that there are no markers like [inaudible], (unintelligible), {incomprehensible} or similar items. Instead, such passages are normally treated as disfluencies.

Mispronounced, incomplete or unintelligible words must be disfluencies if they are single word mistakes that are corrected immediately afterwards.

Examples:

```
[00:19:19]   I didn't foc ... forecast this short-term
             rise.
[00:02:42]   How much is a bottle of water worth in the
             dessert when you're starv ... dying of
             thirst?
[00:00:50]   And then she really grmp ... freaked out.
```

Examples of mispronounced (*foc…*), incomplete (*starv…*for *starving*) or incomprehensible (sounds like *grmp*) words. They are disfluencies and connected to subsequent material with `...`.

Incomplete or unintelligible words can also come at the end of a longer chunk of text that are then corrected immediately afterwards. Again, they are treated as disfluencies.

Examples:

```
[00:06:06]   And it manifested itself to low self-estreem ...
             in low self-esteem.
[00:07:29]   I'll talk about ice shee ... ice caps.
[00:27:08]   I'm gonna turn anya ... switch hemispheres on
             you now and look at Antarctica.
```

Examples of mispronounced (*estreem*), incomplete (*shee…* for *sheet*) or unidentifiable (*anya* perhaps for *I'm gonna*) words. They turn the preceding material into disfluencies and are connected to subsequent material with `...`.

Incomplete or unintelligible words can also come at the end of a long chunk of text, introducing a disfluency, which is, however, not corrected. Instead, the subsequent material is not verbally related to the disfluency.

```
[00:40:33]    I couldn't talk to my father weli ... when he
              behaves like a monkey.
[00:10:49]    So yeah, I hope you all are happ ... I can't
              speak.
[00:23:58]    Like, you know, for cau ... and, again, I
              understand that there's negative conflation.
```

Mispronounced (*weli*…for *really*) incomplete (*happ* for *happy*) or unidentifiable (*cau* perhaps for *conflation*) words. They result in disfluencies that are not picked up or corrected. Instead, the disfluencies are joined to subsequent, new material with . . . .

4.9. Clarification: The three period ellipsis marker . . . is used for all disfluencies. This includes syntactically incomplete material that could be described as "**almost complete**", i.e. material that is typically long, includes a subject, finite verb and possibly several other dependents but still terminates in an incomplete or unintelligible word that feels somewhat trivial (e.g. just the last syllable of a word, a barely unintelligible word after a very long chunk of text etc.).
The transcripts do not make use of token-final . . . after incomplete sentences of this kind. Instead, even "almost complete" utterances are treated as disfluencies joined with a three period ellipsis marker to subsequent material to form complete tokens.

Examples:

not:

```
[00:22:55]    If you watch the whole four hours, your
              impression of me would be no different than if
              you just watched the first seventy se ... .
[00:23:21]    So you know I'm being genuine.
```

but:

```
[00:22:55]    If you watch the whole four hours, your
              impression of me would be no different than if
              you just watched the first seventy se ... so you
              know I'm being genuine.
```

This example has a long utterance terminating in the incomplete word *se*… (most likely for the *seconds*). As a consequence, the entire preceding material is treated as a disfluency. It is joined to the following material to form a complete token with . . . .

not:

```
[00:13:25]    My songs are on Spotifa ... fy.
[00:13:28]    ER you can download them for free.
```

but:

```
[00:13:25]    My songs are on Spotifa ... fy ... ER you can
              download them for free.
```

This example involves an almost complete sentence which, however, terminates in an incomplete word (*fy*, obviously for *Spotify*). As a consequence, the preceding material is regarded as a disfluency and joined to the subsequent material with . . . .

For exceptions in which corrections of mispronounced words are possible, see below C.7.
For confusion between disfluencies and non-linguistic noises, see below C.8.3.

5. Overt corrections

5.1. Overt corrections indicate that the speaker himself or herself assesses their previous speech to be incorrect and to be ignored or amended. Typical overt corrections are short, usually single word expressions, such as *or*, *rather*, *sorry*, *no*, but can also be longer expression, such as *or rather, I meant to say, excuse me, oh no, I meant*.
5.2. If the overt correction follows and corrects an actual disfluency, i.e. material that is incomplete, fragmentary or incoherent and so cannot be integrated into the rest of the token without the three period ellipsis marker . . . (the most common case), then the overt correction **is itself transcribed as a disfluency** followed by another three period ellipsis marker . . . . The general pattern is: a disfluency + . . . + overt correction + . . . + subsequent speech. As a consequence, the elimination of all disfluencies leads, as before, to a coherent sentence. Note that there is usually no marked pause between these kinds of overt corrections and the subsequent speech. As before, the three period ellipsis marker can thus not be interpreted as a pause.

Examples

[00:05:29] And that's a big change ... sorry ... a small change in P.H..

[00:03:35] This last summer ER Tyler, Texas and Lubbock, Texas had the lowes ... or ... record low high temperature in July.

[00:00:34] I went to work for R.C.A. Astro Electronics, which became General Electric, and then Lo ... no ... Martin Marietta, and then Lockheed Martin.

[00:10:02] But there ... there really have been tragedies ER in the collection of ... of ER ... or ... deployment of our Arctic and Antarctic field campaigns.

[00:11:21] They haven't aged whe ... or ... ng ... gained wisdom.

[00:04:34] But this is data on the percentage of boys ... or ... the ... sorry ... the number of boys out of a hundred thousand in the population.

Examples of sentence tokens involving overt corrections. They form disfluencies followed by . . . . All disfluencies above are highlighted. The overt corrections are shown in red. If the disfluencies, including the overt corrections, are eliminated, a coherent sentence is left. The last example is particularly complex as an overt correction (*not full bottle*) includes a parenthetical close (*hello, wake up*), which is therefore included within a disfluency.

5.3. If the overt correction follows normal, non-disfluent speech, i.e. material that is complete and so can be integrated into the rest of the token without the three period ellipsis marker (a rarer case), then the overt correction is integrated into the token as a dependent and separated with a comma. (Such cases are not always easy to distinguish from elaborations, and so this type of overt correction and elaborations are transcribed in the same way.) The general pattern is: material + , + overt correction. There can be substantial subjectivity in deciding whether material before a correction counts as disfluent or not.

Examples

```
[00:06:21]   So the scale was made by a brewer, or chemist
             rather.
[00:03:53]   And the very first thing, or second thing, he
             said was this.
[00:03:48]   We are capable, as authors of that degradation,
             to undo it or to stop it.
[00:17:58]   And by doing that, you eliminate any merging or
             merging problems, I should say.
[00:19:03]   This is gonna be maybe a little bit an office
             corridor or office area back out this way.
```

5.3. If the overt correction can form its own token (and the preceding material can also form a token), then it is transcribed as its own token and is not grouped together with other material.

Examples

not:

```
[00:13:12]   This is hotter ... or no, I meant ... colder.
```

not:

```
[00:13:12]   This is hotter, or no, I meant colder.
```

but:

```
[00:13:12]   This is hotter.
[00:13:13]   Or no, I meant colder.
```

This example involves an overt correction Or no I meant. However, since it can be analyzed as a sentence token and the preceding material is also non-disfluent, This is hotter, the overt correction is transcribed as its own sentence token.

```
[00:31:21]   ER I would be remiss if I did not mention ER
             glaciers and ice shee ... ice caps.
[00:31:27]   Excuse me.

[00:03:27]   I'm using a white wine.
[00:03:29]   And, I mean, actually ... I'm sorry.
[00:03:32]   You can use white wine.
[00:03:32]   I actually had some cherry wine that I'm
             throwing in this.
```

5.4 Additional material that is introduced because of an overt correction, such as affirmative interjections (*yes*, *yeah*) etc., are transcribed as part of the non-disfluent sentence token.

Example

```
[00:08:41]   So five meters ... sorry ... yeah, five meters
             in a century, ER oceans have gone up in our
             past.
[00:04:03]   Feel that full bottle ... body ... not full
             bottle ... hello, wake up, full body stretch.
```

5.5. Overt corrections across token boundaries are transcribed as their own exceptional tokens. See B.10.14 for a closer explanation.

Example

```
[00:02:21]    My sister's five years old.
[00:02:23]    She just started school.
[00:02:25]    Oh no, so, six years old.
```

6. Co-occurrence of fillers and disfluency markers

6.1. The filler ER and the disfluency marker three period ellipsis ... may co-occur. ER is regarded as a filler rather than as a sign of a disfluency. It is not always easy to establish the relative order of ER and ... (either ... ER or ER ...). By default, the order is ER ..., but transcribers have a subjective choice to decide whether ER belongs to the disfluent material and signals a stop to the previous incomplete or incorrect utterance (default), or whether ER belongs to the subsequent material and signals hesitation or time needed to plan the next utterance (less common).

Examples:

```
[00:00:19]    That's ER ... that didn't sell it for her.
[00:02:11]    I mean, right now that currency is pegged to
              ER ... to the U.S. dollar.
[00:13:50]    I saw how ER ... how young people affected
              Obama and Bernie Sanders' campaigns.
[00:04:37]    We are in a huge deficit ER ... in ... in a
              war.
```

Examples of disfluencies co-occurring with the filler ER. The highlighted material shows the disfluencies indicated with the three period ellipsis ... . By default, ER occurs inside the disfluency. However, ER can also occur inside non-disfluent material.

6.2. The ellipsis marker ... and lexical hedges (*you know*, *I mean*, etc.) may co-occur. The position of the ellipsis sign before or after the lexical hedge depends on a subjective feeling of where the hedge best adjoins.

not:

```
we've certainly, you know, we've been in a bull market
```

but:

```
we've certainly ... you know, we've been in a bull
market.
```

not:

```
anybody who truly wants to conserve, right, it's not the
number of dollars that you wanna conserve
```

but:

```
anybody who truly wants to conserve, right ... it's not
the number of dollars that you wanna conserve
```

7. Correction of mispronounced words

7.1. Mispronounced words are not normally corrected. Instead, they form disfluencies, where their phonetic form is approached as closely as possible with letters of the English alphabet (see above C.4.8, C.4.9). This is the default treatment of mispronounced words.
7.2. However, in certain cases, it is preferrable to correct mispronounced words rather than to convert them into disfluencies. Correction of mispronounced words is permissible if:

- The mispronunciation is relatively trivial, involving typically a single sound, at worst a few sounds, occurring quite commonly (missing *d/t*, slurring of a vowel, anticipation of a sound, confusion between *r/l* etc.).
- The intended sense and form of the mispronounced word is reasonably clear.
- The mispronounced word is not contained in a disfluency for reasons other than its mispronunciation, e.g. because it is repeated or corrected immediately afterwards.

Correction of mispronounced words is further encouraged if the offending word occurs in the middle of an otherwise unremarkable token so that its treatment as a disfluency would create a number of disfluencies (before and after the mispronounced words), thus impeding the readability of the transcripts. Nevertheless, there can be substantial subjectivity on behalf of the transcriber as to if and when a mispronounced word can and should be repaired.

Examples:
The word *pray* is pronounced as *play*
Option 1: (disfluency)

```
[00:03:01]   I mean, they wanna play ... to God ... they
             wanna go to church.
```

Option 2: (correction)

```
[00:03:01]   I mean, they wanna pray to God.
[00:03:03]   They wanna go to church.
```

This utterance includes a mispronounced word. It is caused by a single sound (*l* for *r*). Its intended sense is clear from the context, 'pray.' There is no other reason to transcribe it as a disfluency. In addition, its treatment as a disfluency would require the postulation of two disfluencies (before *play* and *to God*), which would make it more difficult to read the transcript. Consequently, this is an example of a mispronunciation that can and should be corrected. Hence, the transcript chooses option 2.

The word *underlain* is pronounced as *under la lain*
Option 1: (disfluency)

```
[00:08:19]   And there were numerous salt lakes under la
... lain ... by taliks, which is unfrozen permafrost, okay
... so, let me just have a look here, okay?
```

Option 2: (correction)

```
[00:08:19]   And there were numerous salt lakes underlain
by taliks, which is unfrozen permafrost, okay?
[00:08:30]   So, let me just have a look here, okay?
```

This utterance includes a mispronounced word. It is caused repetition of two sounds (*la*). Its intended sense is clear from the context, 'underlain.' There is no other reason to transcribe it as a disfluency. In addition, its treatment as a disfluency would require the postulation of two disfluencies (after *lain* and after *permafrost*, *okay*), which would make it more difficult to read the transcript. Consequently, this is an example of a mispronunciation that can and should be corrected. The transcript chooses option 2.

Other examples of corrections:

The speaker pronounces a phrase as *high accracy*. The transcript corrects:

high accuracy

A phrase sounds like *tocks of hurricanes*, probably with the /k/-sound anticipated from *hurricanes* replacing the /p/-sound of *tops*. The transcript corrects:

the tops of hurricanes

A word is pronounced as *radi-ive* in the sound file. This could be a new or technical word, but the transcriber decided to repair this word to *radiative*, the word closest in sense found in standard dictionaries. The transcript corrects:

radiative

A sound file does not include the final *-s* of *news*. The transcript corrects:

the news on T.V.

and then gear up in Jackson ER ville

7.3. There is no indication that a correction has taken place. (For example, there is no system such as [play/pray] to indicate the mispronounced and corrected version of the word as in the Switchboard corpus.) The corrections are implemented without any further notice.

7.4. Non-standard kinds of grammatical expressions (e.g. non-standard agreements) are not regarded as mispronunciations and are therefore not repaired. In general, it is too hard to decide if relevant forms reflect an unintended mistake or dialectal / idiolectal features of the speaker.

Examples:

Paula don't think so.

This example includes a non-standard agreement between *Paula* and *don't*, where standard English would use *doesn't*. Since this is a kind of grammatical expression, it is not corrected, but spelled *don't* in the transcript.

How has this arised?

This example includes a non-standard past form, *arised*, where standard English would use *arisen*. Since this is a kind of grammatical expression, it is not corrected, but spelled *arised* in the transcript.

[00:11:25]   And it's just a whole nother experience

This example include the non-standard form *nother*.

7.5. Slips of the tongues are not corrected or indicated in the transcripts. Slips of the tongue are here understood as any utterance that includes words or phrases conveying something that clearly were not intended, accidental untruths, incidents of a person misspeaking. Often, they involve mistaking one of a pair of terms expressing opposites.

Examples:

Hillary Clinton, ER who will very likely ER be the Republican nominee.

What was meant was 'the Democratic nominee.' The mistake is not indicated or corrected.

if businesses pay their employees less, these businesses have less money to invest.

What was meant was 'if businesses pay their employees more.' The mistake is not indicated or corrected.

8. Non-linguistic and semi-linguistic noises

8.1. There is no indication of **non-linguistic noises**, which are purely accidental in the speech situation, such as applause, sneezing, background sounds, ambient music, computer mouse clicks, squeaking of a chair, phone calling, clinking of glasses, etc.

8.2. Likewise, there is generally no indication of **semi-linguistic noises**, which are unlikely but not impossible to convey some communicative intent, such as laughter, coughing, clearing one's throat, other throat sounds, sighing, clicks of the tongue, audible breath, being out of breath, sniffing, etc. It would often be hard to decide if such noises are accidental and non-linguistic or intentional and linguistic. It would also be extremely difficult to transcribe these sounds consistently.

Examples:

not:

```
[00:01:53]  There been some areas of cooling. {cough}
```

but:

```
[00:01:53]  There been some areas of cooling.
```

The speaker coughs after this utterance. This is not indicated in the transcript.

not:

```
[00:00:37]  And I think it's possible, hu-hah, to
            solve the big problems.
```

but:

```
[00:00:37]  And I think it's possible to solve the big
            problems.
```

This sentence includes an audible chuckle, which could perhaps be transcribed as *a-ha*, *ha-ha*, *hu-ha*, etc. It might be used for various reasons, perhaps to lighten the mood, as a response to something funny happening in the situational context, etc. However, as a non-linguistic or semi-linguistic sound, it is not transcribed at all.

not:

```
[00:01:53]  The reason, ehee, it's urgent is because
            of the inertia of the system.
```

but:

```
[00:01:53]  The reason it's urgent is because of the
            inertia of the system.
```

This utterance involves an audible sound which could perhaps be orthographically rendered as *he*, *hu*, *ehe*, *ehu*. It is perhaps used to express something like "You see," a certain form of exasperation, a jocular form of frustration, a not entirely serious way of saying something serious, etc. However, since it is not clear if this sound is accidental or intentional, since the sound is difficult to transcribe, and since the meaning of the sound is not conventionalized enough to be understood consistently, it is not transcribed at all.

8.3. It can sometimes be difficult to decide if a sound is non-linguistic or semi-linguistic and so not transcribed, or if it is an instance of *uhm* transcribed as the filler `ER` (see C.2.3.) or a mispronounced word transcribed as a disfluency (see C.4.7., C. 4.8.) or an interjection (see D.2). The default is to transcribe sounds as `ER` or as a disfluency or interjection wherever reasonably possible rather than to interpret them as non-linguistic or semi-linguistic. The transcribers have some subjective choice in this respect.

Section D: Specific Transcription Conventions

1. General Remarks

1.1. Unless otherwise regulated, all words are spelled in accordance with standard English orthography.
1.2. The transcripts use American English spelling conventions.
1.3. Some of the most important aspects of American English spellings concern the following orthographic variables:

verbs in *–ize,* not *–ise*:	*apologize, recognize*, not *apologise, recognise*
nouns in *–or* not *our*:	*color*, not *colour*
nouns in *–er* not *–re*:	*center, meter* not *centre, metre*
single *-l-* not double *-ll-*:	*spiraling, modeling* not *spiralling, modelling*
individual lexical items:	*gray,* not *grey, maneuver*, not *manoeuvre, check* not *cheque*

2. Numerals

2.1. All numbers are spelled out in full in the transcripts. All numbers are transcribed as pronounced. Every number item is spelled separately. Hyphens or other punctuations signs are not used. The following sections clarify and illustrate this general guideline.
2.2. Cardinal numbers from 1-19 are spelled out in full as single words. Likewise, multiples of ten are spelled out in full as single words, such as *twenty*, *thirty*, *forty,* etc.

Examples:

not:
```
2 or 3 days ago
```
but:
```
two or three days ago
```

not:
```
down 10 or 15 percent on the gold portion
```
but:
```
down ten or fifteen percent on the gold portion
```

not:
```
what is available from the Sun, 2 40,
```
but:
```
what is available from the Sun, two forty,
```

2.3. The number 0 is spelled `O.` (an upper case letter *o*, followed by a full stop, not the numeral symbol *0*) when pronounced *oh*. However, it is transcribed `zero` or `null` when pronounced *zero* or *null*.

Examples:
```
[00:05:38]   That's really politics one O. one.
[00:07:07]   I have zero patience.
```

2.4. Complex cardinal numbers from 21 to 99, which consist of several words, such as 35 (*thirty five*), 23 (*twenty three*) etc., are spelled out in full as separate words. They have a space between their components, not a hyphen or any other connector.

Examples:

not:
```
49 of the 50 states
```
but:
```
forty nine of the fifty states
```

not:
```
if you die and you're f ... 55
```
but:
```
if you die and you're f ... fifty five
```

2.5. Large numbers that indicate orders of decimal magnitude are spelled out in full as separate words, such as *hundred, thousand, million, billion, trillion*.

Examples:

not:
```
600,000,000,000 dollar bailout
```
but:
```
six hundred billion dollar bailout
```

not:
```
go to 10,000, 100,000, 1,000,000
```
but:
```
go to ten thousand, a hundred thousand, a million
```

2.6. Combinations of numbers to form complex numbers are spelled out in full. Every component number is spelled separately.

Example:
```
six thousand three hundred forty nine jobs
```

2.7. Years numbers are spelled out in full like any other number. Year numbers are written exactly as heard. For example, a text file may include the pronunciation of a year as. It is then transcribed correspondingly. Every component number is written separately surrounded by whitespace. Years without the millennia or centuries pronounced are also transcribed as heard.

Examples:
```
nineteen ninety nine
back to what it was in nineteen oh two
two thousand one
twenty twenty
ER two thousand fourteen, fifteen, sixteen, the warmest years on record
in two thousand ER one hundred.
```

2.8. Numbers (e.g. year decades) with the plural endings *–s* are likewise spelled out in full.

Examples:

```
the nineteen seventies were bad
the late nineties
```

2.9. Fractions are spelled out. No punctuation signs are used to separate independent words in fractions. In particular, fractions do not require slashes, periods or hyphens.

Example:

not:

```
literally less than 1/3 of that
```

but:

```
literally less than one third
```

2.10. Ordinal numbers are always spelled out in full. Every element of the ordinal number is spelled separately. Hyphens are not used.

Examples:

```
the twenty first century
It is Friday, August twenty first 2009
```

2.11. Decimal points or other symbols are not used within numerals to indicate. Instead, the transcripts include the words that a speaker actually chooses to express decimal places.

Examples:

not:

```
7.5 trillion is due in under a year
```

but:

```
seven and a half trillion is due in under a year
```

not:

```
revenues were 20.8 million
```

but:

```
revenues were twenty point eight million
```

2.12. Numerals with the suffix *-fold* are spelled as a single word, not with spaces or hyphens.

Example

not:

```
eight fold
eight-fold
```

but:

```
eightfold
```

3. Interjections

3.1. A fixed set of transcribable interjections are assigned a standardized, specific orthographic form. Items are included in this set only if their pronunciation is relatively unique and if their meaning is conventionalized enough to allow a fairly objective identification. If an interjection from the list appears in speech, it must be transcribed. If some other sound appears that is not in the list, it must not be transcribed.

Nevertheless, it can sometimes be difficult to decide if the relevant sounds really instantiate the relevant interjection, or if they are some other kind of noise, e.g. a different interjection or a non-linguistic sound. There is some subjectivity in the transcription of interjections.

3.2. The primary guide for the transcription of interjections is form, not meaning. For example, if an interjection is audible as *hey*, it will be transcribed as `hey`, even though intended interjections could have been *okay* or maybe even *alright* given a context. In other words, ambiguous interjections are transcribed as heard by default. There is however some subjectivity on part of the transcriber as to how best to transcribe ambiguous interjections.

3.3. The following lists the interjections that are used in the transcriptions in alphabetical order:

	Transcription	**Pronunciation**	**Meaning**
1	`a-hah` (see `uh-huh`)	This interjection consists of two open *a*-like vowels, the second pre-aspirated and typically carrying greater stress. /ˌaˈhɑ/, /ˌəˈhɑ/, /ˌaˈhɑ̃/, etc. The sound is not always easy to distinguish from `uh-huh`.	It is used to convey sudden insight, a sudden realization, confirmation, surprise, a revelation or understanding, the attainment of a goal.
	`[00:00:16]   A-hah, so you admit it.`		
2	`ahh`	A long, monophthongal open vowel sound, /a:/. It often occurs with high falling or a level high intonation.	This is a multi-purpose interjection. It is most commonly used to express (i) satisfaction, positive sentiment, positive attitude towards something, joy, pleasure (usually with a falling intonation) or (ii) surprise, shock, negative sentiment, astonishment (usually with a high flat or rising intonation).

	`[00:07:01]   Ahh, the metro, I like the metro.`
	`[00:02:43]   Ahh, what an adult male thing to do.`		

3	`'ah`	A monophthongal open, short vowel sound with a marked, initial glottal stop, /ʔa/, often a falling intonation.	It expresses accidents, exhaustion, being annoyed, or an involuntary reaction to something negative.
	`[00:06:19]   'Ah, tripping people over here.`		

4	aww (see naww)	A long mid or low back vowel, /ɑː/, /ɒː/, /ɔː/.	The interjection suggests that the speaker finds something sentimental, cute, adorable.
	[00:04:24] Aww, you're so adorable.		
5	bam (see ba-bam)	A bilabial voiced plosive followed by an open vowel or mid vowel in front of the interjection bam, /ba bæm/, /bɐ bæm/, /bəbæm/, etc.	The expression expresses joy at success, having managed or finished something.
	[00:00:12] Yes, bam, that's how you do it.		
6	ba-bam (see bam)	A bilabial voiced plosive followed by an open vowel or mid vowel in front of the interjection bam, /ba bæm/, /bɐ bæm/, /bəbæm/, etc.	The expression expresses joy at success, having managed or finished something. It's is probably a strengthening of bam.
	[00:03:29] But we did it, ba-bam, ba-bam.		
7	bla	This item is pronounced /blɑː/ or similarly. It is often repeatedly.	The word indicates trivial, banal or annoying speech. It is repeated as often as it is audible.
	[00:59:37] I was imagining multiple legendary minions in my hand, multiple small copies, bla bla bla bla bla.		
8	eek	Pronounced /iːk/	This interjection expresses unpleasantness, disgust, or that the speaker does not want to do something.
	[00:04:18] Eek, I hate it.		
9	eh (see hey)	The diphthong /eɪ/ or sometimes a monophthongized version /eː/.	This is a multi-purpose interjection. For instance, it can be used to seek approval or confirmation by the audience, or to complain, express exhaustion. In token final position it can co-occur with a question mark.
	[00:01:44] You just saw that, eh?		

10	'eh (see meh)	A short front vowel, /ɛ/, /e/, etc. Often with an initial glottal stop, /ʔɛ/, /ʔe/, etc.	It estimates something as “so so”, “not great” or “low to average”. It expresses belittlement, that something is not a big deal, that something is not particularly impressive or important. It is regarded as a reduced version of *meh*.
	< [00:02:21] How was it? > [00:02:21] 'Eh. [00:26:15] 'Eh, no big change.		
11	eww (see urgh) (see heuh) (see meuh)	A long *u*-sound, usually preceded by a palatal glide, /juː/, /uː/.	It often expresses disgust or related emotions, repulsion, difficulty, deviation.
	[00:03:54] I don't wanna see that, eww.		
12	gee (see jeez)	Pronounced /ʤiː/ or similarly. There is no consonant after the vowel.	The word is used to introduce a reaction to something, often surprise, admiration, irony, frustration, anger.
	[00:11:14] Gee, I don't know actually.		
13	heuh (see urgh) (see eww) (see meuh)	This interjection is pronounced with initial aspiration, usually clear, but may also be glottalized. The following vowel is central, and long and often nasalized. It often carries rise-fall or some other salient intonation, /hɜː/, /hœ̈ː/, /hɐː/, /hʌː/, /ʔɜː/, /ʔœ̈ː/, /ʔɐː/, /ʔʌː/, etc.	It expresses exhaustion, difficulty, regret, complaint, despair, and other negative emotions, but is not as strong as disgust or rejection (unlike *eww* and *urgh*). It can also be used to indicate breathing out to relax, conclude or begin a new topic. It is largely synonymous with meuh.
	[00:01:18] Heuh, why did I do that? [00:00:59] Heuh, feels good to take a breath.		
14	hey (see eh)	This interjection is pronounced /heɪ/.	It is used to draw attention to something. It can also be used as a greeting.
	[00:16:56] Hey, what are you doing? [00:02:48] So, hey, you guys.		

15	hff (see uff) (see pff)	A clearly audible, typically long *f*-sound, possibly pre-aspirated /f/, /fː/, /hf, /hfː/.	The interjection typically expresses difficulty, trouble, the fear that something might be difficult or go wrong. It can also be used to show that something is unpleasant or to show compassion for someone who is going through something difficult. It is perhaps a strengthened form of a sigh (not transcribed) or a shortened form of uff.
	[00:00:36] Hff, let's see if we can run it up.		
16	hoo (see woo) (see woo hoo)	An aspirated back vowel, /huː/.	The interjection is usually used to indicate mirth, relief or joy.
	[00:03:24] Look at that, hoo.		
17	hhoah	An ingressive gasping sound, i.e. /huː↓/, /haː↓/, /hə↓/. Etc.	The interjection is used to express horror, shock, or surprise.
	[00:03:24] Hhoah, dangerous.		
18	hah (see huh)	An audible initial aspiration, *h*, followed by a stressed, short, low mid vowel, /ha/. It can be difficult to distinguish from *huh*, but is generally shorter and the vowel is less back.	It is used to express triumph, victory, positive emotion, catching someone in the act.
	[00:02:21] Hah, gotcha.		
19	huh (see hah)	An audible initial aspiration, *h*, followed by a back vowel /hɑ/, /ha/. The vowel can be laryngealized. The *h*-sound can be quite long.	It is used to express realization that something has happened, reaction to news, mild surprise. At the end of an utterance it can be used as a tag to seek confirmation or attention.
	[00:06:05] So, I'm like, 'Hey, huh, it's a win-win situation.'. [00:08:29] Oh man, twenty million dollars, huh?		

20	jeez (see gee)	Pronounced /ʤiːz/ or similarly. There must be a consonant after the vowel.	The word is used to introduce a reaction to something, often surprise, admiration, irony, frustration, anger.
	[00:11:14] Jeez Louise, calm down already.		
21	meh (see 'eh) (Note: not transcribed as neh.)	A nasal with a short front vowel, most commonly a bilabial nasal, but sometimes also another kind of nasal /mɛ/, /me/, /nɛ/, /ne/, etc.	It estimates something as “so so”, “not great” or “low to average”. It expresses belittlement, that something is not a big deal, that something is not particularly impressive or important.
	< [00:00:00] It's beautiful. > [00:00:01] Meh.		
22	meuh (see urgh) (see eww) (see heuh)	This interjection is pronounced with an initial nasal. The following vowel is central, and long and often nasalized, /mɜː/, /mœ̈ː/, /mɞː/, /mʌː/, etc.	It expresses exhaustion, difficulty, regret, complaint, despair, and other negative emotions, but is not as strong as disgust or rejection (unlike *eww* and *urgh*). It is largely synonymous with heuh.
	[00:00:01] Meuh, I really don't wanna do this.		
23	mmh (see mmh-mmh) (Note: not transcribed as hm, hmm, mh, etc.)	A long, bilabial nasal sound /mː/. There may be pre-aspiration. /hmː/. Both versions are transcribed the same. The interjection can also be pronounced with a glottal stop instead of *h*, /ʔm/ /hʔm/, etc. with a falling ton, lots of air escaping from the air, and a very short nasal.	This interjection has at least two distinct functions. On the one hand it can indicate thoughtfulness, reflection, pensiveness. It then often occurs with a falling intonation. On the other hand it can be used to express positive attitudes or affirmation. In particular, it can mean positive feelings towards food or drink, meaning ‘delicious’. It then often occurs with a rising or fall-rise intonation.
	< [00:00:00] Do you read a lot? > [00:00:01] Mmh. [00:03:43] Mmh, this seems to be right. [00:00:13] You can just toss the caffeine back one at a time, mmh, delicious.		
24	mmh-mmh (see mmh)	Two bilabial nasal sounds in a row, often the second longer than the first, /m mː/. Sometimes the first sound may be glottalized, /ʔm m/.	This interjection conveys affirmation, agreement, or paying attention.

25	mpwah	A kissing sound	The sound of a kiss, used to show affection, love, etc.
	[00:15:12] But you guys are the best, mpwah.		
	< [00:00:38] He's quick to remind us that he went to the best schools. > [00:00:40] Mmh-mmh, and he's learned nothing in them.		
26	nah (see no)	An alveolar nasal followed by a long or short low back or front vowel, /nɑː/, /nɑ/, /nɛ/ etc.	This interjection expresses the opposite, saying "no", saying that something did not happen, disapproval or incredulity.
	[00:06:39] And then, nah.		
27	naww (see aww)	An *n*-sound followed by a long mid or low back vowel, /nɑː/, /nɒː/, /nɔː/.	It conveys sympathy, strong feeling of compassion or attraction, finding something cute, adorable, praiseworthy, or sympathetic.
	[00:02:19] Naww, bless you for that.		
28	no (see nah)	An n-sound followed by an o-like diphthong or monophthong, /noʊ/, nəʊ/, /nɔː/, etc.	The standard English interjection to deny something.
	[00:07:12] No, I don't think so.		
29	och (see urgh)	A short *o*- or *a*-vowel and a voiceless velar or palatal fricative, /ɔx/, /ɒx/, /ɔʝ/, /ɒʝ/, etc.	It is most commonly used to express annoyance, grief, stress.
	[00:01:32] Och, she adopted two kids after she went into remission. [00:00:40] Och, I, myself included, have fallen into this trap.		
30	oh (see ooh)	This interjection has an *o*-sound, /oʊ/, /o/, /ɔ/. The expression can be quite short, but also long /oʊː/, /oː/, /ɔː/	This is a multi-purpose item. It can be used, among other things, for surprise, insights, realization, questioning.
	[00:01:08] Oh, my X Box just turned on 'cause I said, 'Bus time'.		

31	oof (see uff) (see heuh)	A relatively long *u*-sound (rather than a typically short *u*-sound as for uff), possibly pre-aspirated, with a final weak *v*-sound, (rather than a strong *f*-sound as for uff) possibly reduced to /h/ or /ç/, etc., /uːv/, /uːç/, /uːh/, /huːv/, /huːç/, /huːh/,	It expresses nervousness, mild anxiety, anticipation of something hard or stressful though not necessarily hurtful or repelling.
	[00:00:18] Oof, I'm, like, shaking 'cause I'm, like, excited but nervous.		
32	ooh (see oh)	A single, monophthongal high back vowel, /uː/, /u/, /ʊ/. It must be an *uh*-like sound, not an *oh*-like sound. It can sometimes be quite loud and prominent.	The interjection is used to express positive surprise, a good outcome, coming up with a good idea, wonder, amazement or unexpectedness, pleasure, joy.
	[00:04:17] 'Augmented ore shaft', ooh, futuristic ore shafts, nice. [00:06:59] Ooh, maybe we can add a Hollywood sign?		
33	oops	Pronounced as /uːps/ or similarly.	The interjection *oops* is used to convey a small error, a slip-up, something unexpected.
	[00:36:52] Oops, that's not what I meant.		
34	ouch (see ow)	Pronounced /aʊʧ/.	The interjection is used to indicate pain, empathy for pain, or metaphorical pain.
	[00:33:32] Ouch, that hurt.		
35	ow (see ouch)	Pronounced /aʊ/.	The interjection is used to indicate pain, empathy for pain, or metaphorical pain.
	[00:57:56] Ahh, Despi, she's my little lap cat. [00:58:03] ER yeah, ow, Despi, ahh, ow. [00:58:06] She's scratching me.		
36	pff (see hff)	This interjection involves a clear, voiceless *f*-sound, perhaps sometimes	This interjection expresses derision, belittlement, worthlessness.

		audible as an *s*-sound, preceded by a plosive, typically *p*, but perhaps also a glottal stop or *b* (unlike *hff*, which does not have such a marked sound in front of *f*), /pf/, /fs/, /ʔs/, /ʔs/, etc.	If the initial plosive is quite weak, it can be synonymous with *hff*.
	[00:15:14] Pff, who cares about that?		
37	phew (see whew)	An *f*-sound followed by a glide and a high back vowel, /fjuː/.	The interjection normally expresses relief, a happy feeling that something luckily went well even though it might not have.
	[00:15:14] Phew, that was close.		
38	shh	A *sh*-sound, often marked in comparison to the surrounding speech and long, /ʃ/, /ʃː/.	An order to be quiet, or other discourse functions surrounding silence.
	[00:12:33] Shh, be quiet.		
39	uff (see hff)	This is a short high back vowel followed by a clearly audible *f*-sound, /ʊf/.	The interjection typically expresses difficulty, trouble, the fear that something might be difficult or hard to express or explain. It can also be used to show that something is unpleasant or to show compassion for someone who is going through something difficult.
	[00:00:00] Uff, it's so hot right now.		
40	uh-huh (see uh-uh) (see a-hah)	It consists of two central vowels, typically short and low, the second not involving a glottal stop, but rather a glottal fricative, often nasalized salient tone, /ˌəˈhə/, /ˌʌˈhə/, /ˌə̃ˈhə̃/, etc. The sound is not always easy to distinguish from a-hah.	It is used to confirm, to say ‘yes’, to express confirmation, understanding, agreement or realization.
	<[00:03:19] You got a new single, "Fall Crazy".> [00:03:22] Uh-huh.		

41	uh-uh (see uh-huh)	It consists of two successive, somewhat indistinct low-back vowels, frequently nasalized, either both or just the second involving an initial glottal stop.	This expression is used to say *no*, give a negative answer or convey disagreement, negation or refusal in an emphatic, emotional or colloquial way.
	[00:03:22] Everybody, when they got the December number, was saying, 'Uh-uh, this is a fluke.'. [00:05:53] Is there any investigation of the rating agencies? [00:05:56] None, none, uh-uh, nothing.		
42	urgh (see eww) (Note: not transcribed as eugh.) (see heuh) (see meuh)	The interjection is pronounced with a central or back vowel, often diphthongized, ending in a uvular, laryngeal or glottal (maybe velar) sound, /ɜːʀ/, /ʊɜʀ/, /ɜːʕ/, /ʊɜʕ/, /ɜːx/, /ʊɜx/, /ɜːh/, /ʊɜh/.	It often expresses disgust or related emotions, repulsion, deviation. It can also convey difficulty, uncertainty, trouble in explaining something or finding an answer.
	[00:06:21] Urgh, you're gross.		
43	whew (see phew)	An *h*-like sound (possible a palatal fricative, /ç/), followed by a glide and a long high back vowel, /hjuː/, /hçjuː/, /çjuː/.	It is used to show sympathy without offering advice. It shows compassion, understanding for a difficult situation.
	[00:02:41] Whew, that's really intense.		
44	whoops	A *w*-sound followed by a back vowel a *p*-sound and an *s*-sound, /wʊps/	It expresses a minor accident, something that went wrong, a small incident.
	[00:37:24] Whoops, that should be increased by ninety three percent.		
45	whoa	A *w*-sound followed by a lowering diphthong /woə/, /woɐ/, or a backing diphthong /wəʊ/, /woʊ/ or	It is used to indicate a range of emotions, including, glee, mirth, relief, joy, being impressed, but also warning, excessiveness, asking to

		a monopthongized mid back vowel /woː/, /wɔː/. The *w*-sound may sometimes sound like a *b*-sound.	slow down. It has to do with the idea that something is happening or comes in an excessively great manner.
	[00:00:11] Whoa, a big thank you to Barbie. [00:03:49] People said, 'Whoa, it's done.'.		
46	woo	A *w*-sound followed by high back vowel, /wuː/, or similar.	The interjection is usually used to indicate mirth, relief or joy.
	[00:03:43] Woo, excellent. [00:04:49] Woo, we're back, okay?		
47	woo hoo (see woo) (see hoo)	A combination of the interjections woo and hoo, the latter possibly repeated several times.	The interjection is usually used to indicate mirth, relief or joy. It is a strengthened form of woo.
	[00:03:19] Woo hoo hoo, was a close one.		
48	wow	A *w*-sound followed by a raising diphthong, /waʊ/.	It expresses awe-struck surprise, or unexpectedness, astonishment, being impressed.
	[00:02:59] Wow, that's a small tree.		
49	yay	a palatal glide followed be fronting diphthong, or a palatal coda, /jeɪ/, /jɛj/.	The interjection expresses joy, mirth, positive excitement.
	[00:02:40] Yay, I'm so excited.		
50	yeah (see yup, yes)	A palatal glide followed by a centralizing diphthong, or front or low back monophtong, /jɑ/, /ja/ /jeɐ/, /jeə/, /jeː/, etc.	A colloquial way of saying “yes”, giving agreement, finding one’s opinion confirmed, giving approval.
	[00:01:42] Yup, yeah, that happened. [00:12:17] Yeah, this is all covered really well.		
51	yes (see yup, yeah)	Pronounced /jɛs/ or similarly.	The standard interjection for giving agreement.
	< [00:02:31] You're coming? > [00:02:32] Yes, sure.		

52	yup (see yeah) (Note: not transcribed as yep, yap, etc.)	The word *yup*, with a back or central vowel, and an audible final *p*-sound, /jʌp/, /jɐp/, /jəp/, /jʊp/, /jep/, etc.	A way of saying “yes”, giving agreement, finding one’s opinion confirmed.
	[00:00:32] Alright, yup, that's very busy right there.		

4. Contractions and reductions

4.1. Auxiliaries like *is, have, has, had, would* etc. as well as the negation *not* are often contracted. The relevant contractions are indicated in the text files with an apostrophe. The apostrophe is an inverted, undirected single line, ‘. Otherwise, the contractions follow standard English orthography. It is often difficult to pick up all contractions and to distinguish them coherently from full forms and so there is some subjectivity for these forms.

Examples:

```
[00:16:33]   See, that's what Sean Hannity doesn't
             understand.
```

4.2. The words *going to*, *want to* and *got to* are often contracted to *gonna*, *wanna* and *gotta*. Wherever such a contraction is audible, it is indicated as `gonna`, `wanna` and `gotta` in the text files.

Examples:

not:

```
[00:10:24] They're going to want to buy something with
the profits.
```

But:

```
[00:10:24] They're gonna wanna buy something with the
profits.
```

The audio clearly includes the contracted forms *gonna* and *wanna* and they are thus transcribed accordingly.

The form *gonna* can be further contracted, usually after *I*, to *I’ma*. It is then spelled with an apostrophe followed by a single *m* and *a*, `I'ma`.

Examples:

```
[00:05:16]   I'ma be all right.
[00:00:27]   What I'ma do is I'm gonna swipe high.
```

4.3. The word *because* is often shortened to *‘cause*. Wherever the initial segment *be*- is not audible, the word is transcribed as `'cause` in the text files.

Example:

```
[00:16:02]   You're doing it 'cause I said so 'cause I'm
             the boss.
```

4.4. The (predominantly Southern) contracted second person plural pronoun from *you* and *all* is transcribed as `y’all`.

Example:

```
[00:00:03]   How are y'all doing tonight?
```

4.5. The expression *of course* is often reduced to *course*. It is then transcribed `course`.

Example:

```
[00:04:51]   Course that's not necessary.
```

4.6. The form *isn't* can be further reduced to lose the sibilant. If no *s*-sound in *isn't* is audible whatsoever, the word is transcribed *in't*.

Example:

```
[00:04:59]  And in't it a frustrating thing?
```

4.8 The form *got you* can be reduced and fused to *gotcha* and is then transcribed as `gotcha`.

Example:

```
[00:04:59]  Hah, gotcha.
```

4.9. Other contractions or reduction are not indicated in the transcripts.
In particular, the following contractions and reductions are not transcribed:

(a) Shortenings of *them* to *'em* are not indicated in the text files. Instead, *them* is always spelled out in full as `them`.
(b) *Kind of* is not contracted to *kinda* but *kind of* is always spelled out in full `kind of`.
(c) Verbal and other forms ending in *–ing* are not contracted to *–in'*, where they are pronounced with an alveolar instead of a velar nasal, as in *shining* (not *shinin'*), *asking* (not *askin'*) *morning* (not *mornin'*) etc. Instead, such forms are always spelled out in full ending in `ing`.
(d) The conjunction *and* is not shortened to *'n'*, e.g. *two thousand n one*. It is always spelled out in full as `and`.
(e) *Okay* is not reduced to spellings such as *mkay, kay, hey, ey*, but, where clearly identifiable as "okay", is always spelled `okay`.
(f) *You* is not reduced to spellings such as *ya, y'*, etc. (Exception: second person plural *y'all*, see 4.4. above). Instead the pronoun is always spelled out, `you`.
(g) *Out of* is not reduced to *outa*.
(h) *About* is never contracted to *'bout*.
(i) *You know* is never contracted to *y'know*.
(j) *Little* is never contracted to *lil'*.
(k) *Trying to* is never contracted to *tryna*.

4.10. There are a number of euphemistic reductions of *freaking* (or *fucking*) as intensifiers. Those can be spelled in a number of different ways, such as *freaking, frigging, fricking*, etc. according to how it is heard. However, the ending *-ing* is always spelled out in full, never as *in'* (see above 4.9.(c)).

Example:

```
[00:01:55]  It feels like a fricking race car.
[00:01:25]  It is just so fricking sad.
```

5. Abbreviations and letters

5.1. Acronyms and other abbreviations that are <u>pronounced as independent letters</u> from the English alphabet are indicated with capital letters separated by full stops. Some common examples of this kind are: `U.S.` (*United States*, not `US`), `G.D.P.` (*gross domestic product*, not `GDP`), , `C.N.B.C.` (*Consumer News and Business Channel*, not `CNBC`).

<u>Examples</u>

`T.V. shows`	(*television shows*, not `TV`)
`C.P.I.`	(*consumer price index*, not `CPI`)
`I.P.O.`	(*initial public offering*, not `IPO`)

5.2. Initialisms and other abbreviations that are <u>pronounced as a single word</u> and not as separate letters of the English alphabet are spelled as capital letters not separated with full stops. Some examples are: `NATO` (*North Atlantic Treaty Organization*), `NASDAQ` (*National Association of Securities Dealers Automated Quotations*) or `ARM` (*adjustable-rate mortgage*).

<u>Examples</u>

`NOAA` (*National Oceanic and Atmospheric Administration*, pronounced /noʊə/)
`TED` (*Technology, Entertainment, Design* /tɛd/).

5.3. If initialisms that are pronounced as single words have become frequent, fully integrated into the grammatical system of English, and are generally felt to not be initialisms, then these items are spelled like other common nouns. That is to say, they are spelled as one word with lower case letters. For example, the word radar is transcribed `radar`, not `RADAR`, from *Radio Detection And Ranging*, because it is not generally understood as an initialism anymore.

<u>Examples</u>

`sonar` (not `SONAR`, *Sound Navigation And Ranging*)
`laser` (not `LASER`, *Light Amplification by Stimulated Emission of Radiation*)
`wifi` (not `WiFi` or `Wi-Fi`, pun on *hi-fi*, retrospectively analyzed as *Wireless Fidelity*)
`Covid` (not `COVID` or `CoViD` or `CoviD`, from Coronavirus Disease, note capital letter because it is a proper noun)

5.4. Whenever individual letters are pronounced, they are transcribed with upper-case letters and subsequent full stop.

<u>Examples:</u>

not:
`triple A credit rating`
but:
`triple A. credit rating`
not:
`Q one`
but:
`Q. one`
not:
`the F student`

but:
`the F. student`
not:
`the y-axis,the Y-axis,the Y axis`
but:
`the Y. axis`
not:
`three D`
but:
`three D.`

5.5. Clarification: The rule to separate acronyms and letters with full stops may often lead to spellings that are markedly different than conventions in Standard English. For example, the text files would use strings such as `four P.M.` (rather than normal English orthography, which would require lower case letters, *4 p.m.*), etc.

<u>Examples</u>

`P.H.D.`	(*philosophiae doctor*, not *PhD*)
`B.S.`	(*bull shit*, not *bs*)
`H. two O.`	(a chemical element, not H_2O)
`S. and P. five hundred`	(stock market index, usually rendered *S&P500*)

5.6. Some linguistic items are mixtures between acronyms, independent letters, initialism or other strings.
The parts of a word that are pronounced as an acronym or independent letter are transcribed as capital letters with a full stop.
Parts of a word that are an initialism are transcribed in all caps without full stops and joined to the rest of the word without a space.
Parts of a word that are other strings or independent words are transcribed as they ordinarily would be and are separated from the acronym or letter with a space.

<u>Examples</u>

`S.MOC`	(*Southern meridional over-turning circulation*, pronounced /ɛs mɑːk/)
`C.SPAN`	(a TV channel, pronounced /siː spæn/).
`X. Box`	(a video gaming console by Microsoft)
`NASCAR X. Finity`	(a stock car racing series)
`C.O. two`	(a chemical molecule, not CO_2)
`ICE Sat`	(a NASA satellite, where "ICE" is an initialism for *Ice, Cloud, Elevation*, and "Sat" is another string, a clipping of *satellite*)
`I. Pad`	(a produce by the Apple company)

5.7. When the full stop of an acronym or single letter occurs at the end of a token, it will co-occur with and an additional token-final punctuation. In other words, there is one full stop for the acronym or single letter and an additional punctuation sign to mark the end of a token.

<u>Examples:</u>
not:
`It's always fun to come back to where I did my P.H.D.`
but:
`It's always fun to come back to where I did my P.H.D..`

5.8. Abbreviation symbols (e.g., $ % £ ¥ € ₡ ‰) are not used in the corpus at all. Instead all the words they represent are always spelled out in full. For example, the word *percent* is always spelled `percent`, and never `%`. The word *dot* is always spelled out as `dot` and never abbreviated with a dot (`.`) symbol, etc.

Examples:
not:
`$350 billion`
but:
`three hundred and fifty billion dollars`

not:
`the S&L industry`
but:
`the S. and L. industry`

not:
`7.5 trillion`
but:
`seven point five trillion`

not:
`Overstock.com`
but:
`Overstock dot com`

5.9. Other common abbreviation in English involving letters are not used, but always spelled out as full words. Unused abbreviations include *ok*, *vs.*, *Mr.*, *Dr.* etc.

Examples:
not:
`ok`
but:
`okay`
not:
`vs.`
but:
`versus`
not:
`Mr.`
but:
`Mister` (capital letters in titles)
not:
`Dr. Spencer`
but:
`Doctor Spencer` (capital letters in titles)
not:
`Mrs.`
but:
`Missus` (capital letters in titles)

6. Spelling of specific words

6.1. The table below lists some common lexical items and patterns in the transcripts that are often spelled variably by different transcribers. It fixes how these items are treated in the transcripts.
6.2. Table for specific words:

Item	Transcription	Not
alright	`alright`	`all right`
axe, a variant of ask, particularly common in the African-American community	`ax`	`ask, aks`
band-aid	`band-aid`	`bandaid, band aid`
Bluetooth	`Bluetooth`	`bluetooth, blue tooth, Blue tooth Blue Tooth`
Covid (Coronavirus Disease)	`Covid`	`COVID, CoviD, CoViD`
et cetera	`et cetera`	`etc. etcetera`
email	`email`	`E. mail, e-mail`
focused, focusing	`focused, focusing`	`focussed, focussing`
high school	`high school`	`highschool`
high schooler	`highschooler`	`high schooler`
Master (academic degree)	`master, master student, masters`	`Master, Master's, master's`
no one	`no-one`	`no one, noone`
op-ed	`op-ed`	`oped`
Okie doke	`okie-doke`	`Okiedoke, okie doke, okay doke`
real estate	`real estate`	`real-estate, realestate`
skeptic, skepticism	`skeptic, skepticism`	`sceptic, scepticism`
YouTube	`YouTube`	`Youtube, youtube`

Section E: Capitalization

1. General remarks

1.1. By default, all words are written with lower case letters.
1.2. The use of upper case letters is regulated explicitly in the following sections.
1.3. The terms ‘lower / upper case / capital’ are used to describe letter case.

2. Sentence-initial words

2.1 Capital letters are used for the first letter of a new sentence token immediately after the time stamp.

Examples:
not:

```
[00:01:04] now this is i ... this is incredible.
```

but:

```
[00:01:04] Now this is i ... this is incredible.
```

2.2 However, if the first element of the token is the disfluency marker `ER`, the first word is spelled with a lower case letter unless it is also a proper name.

```
[00:00:35] ER that's a big move.
[00:01:08] ER Bill Gross ER wrote an excellent article on
his website.
```

2.3. Capital letters are also used for the first item of material included in single quotation marks (see B.9, below F.5).

```
[00:02:23] The government says, 'Look, our number shows
there is no inflation'.
```

3. Proper names

3.1. Capital letters are used for proper names in accordance with Standard English orthography.
3.2. Normally, proper names denote unique entities or a unique set of entities. However, it is not easy to define precisely which words should count as proper names. As a consequence, there can be some subjectivity with regards to capitalization of proper names.
3.3. Proper names include names of truly unique entities (singular reference existing only once in the world), such as people, countries, states, companies, specific places, celestial bodies, etc.

Examples:

```
Barack Obama
United States, Puerto Rico, the State of New Hampshire
Microsoft
Wall Street
the Sun, the Earth, Mars, Moon
```

3.4. Proper names include common nouns that a speaker uses to pick out a unique entity or unique set of entities even though the common noun itself may not conventionally be used for a unique referent. Capitalization is used to distinguish specific from general readings. The difference may not always be objective.

Examples:

the President	(= the president of the US)
but: the president	(= any president, for instance of a company)
the Presidential election	(=election of the US President)
but: very presidential	(=behaving like a president)
the Depression	(= the Great Depression of the 1930s)
but: a depression	(= a severe recession, or negative psychological state)
Congress	(= the US Senate and House of Representatives)
but: I went to a congress	(= a conference or symposium or meeting etc.)
Democrat	(=a member of, or pertaining to, the US Democratic Party)
Republican	(=a member of, or pertaining to, the US Republican Party)
but: democrat, republican, communist, capitalist	(=a person with, or pertaining to, democratic, republican, communist, capitalist political ideology)
Apple	(=the company that makes, for example, iPhones)
but: apple	(=a piece of fruit)
the U.S. Treasury	(=a unique government institution)
but: U.S. treasuries	(=a debt instrument)
the War on Poverty	(=name of legislation by the Johnson administration)
but: a war on the rich	
the Constitution	(= the constitution of the US)
but: not in a good constitution	(=in bad shape)
the Democratic Convention	(=a specific convention)
but: Democrats at their convention	(=a non-specific convention)

3.5. Capitalization is used for words referring to the Abrahamic god (including exclamations). This rule also affects pronouns, such as *he* or *his*, when they refer to God. However, generic references to another god or gods are spelled with a lower case letter.

Examples:

God
the Lord
oh my God

[00:00:52] It's a privilege to be hated for His name's sake.

[00:02:49] You could say, 'I have no belief in a god.'

3.6. Capitalization is used for more abstract unique entities or set of entities, such as political programs, historical periods, metaphorical places, products, etc.

Examples:

Obamacare	
Cash for Clunkers	(= subsidy program to boost car sales)
Common Era	

```
Industrial Revolution
Wall Street and Main Street
the Promised Land
we're a dealer for Perth Mint Certificate Program
```

3.7. Names of months are always capitalized. This is true even when they’re used in a non-specific way or as plurals.

Examples:

```
February
What a nice March we're having.
one of the best Decembers ever
```

3.8. Geographical terms involving cardinal directions (*North, West, Southern, Eastern* etc.) are usually spelled with a capital letter because they pick out a specific, unique place in the world. However, cardinal directions can also be spelled with a lower case letter if they refer to a non-specific part, occur in the plural, or modify a general noun.

Examples: (upper case)

`from the Midwest all the way down to the South`
(referring to the Midwest and South of the USA)
`winters in the Southwest and extending into the West`
(referring to the Southwest and West of the USA)
`the North Atlantic`
`Northern Australia`
`Eastern Russia`
`the Northern hemisphere`
`the West`
`Western science`
(where “West, “Western” is understood to refer specifically to Europe and America)
`It's in the North End of Boston`
(“North End” is a specific neighborhood in Boston)
`Northeastern urban centers`

Examples: (lower case)

```
warm temperatures further west
temperatures in the easts
easterly trade winds
the southern parts
blow water westward
they make their way west
```

3.9. Similarly, adjectives referring to geographical places are generally spelled with a capital letter (in accordance with standard English orthography). However, when the geographical adjective is felt to be generic, it is spelled with a lower case letter.

Examples:

```
Cuban
the tropical Pacific Ocean
the hot equatorial Sun
```

3.10. Currency names are spelled with a capital letter except *dollar*, *dollars*, which is always spelled with a lower case letter.

Examples:
the New Zealand dollar
the U.S. dollar
Japanese Yen
the Yuan
Bitcoin
the Renminbi
the Deutsch Mark, or the Yen, the Swiss Franc

3.11. Capital letters are used for official titles in connection with a proper name. However, titles that are used as predicates are not capitalized.

Examples:
Mister Roach President Obama
Governor Chris Christie
Fed Chairman Ben Bernanke
but: She may go down in history as the worst Fed chairman.
Senator William Proxmire
but: He would have been a great senator
Attorney General Peter Kilmartin
but: She is running for attorney general.

3.12. Capital letters are never used for physical units. This is true for names of physical units that are obviously based on proper names (different than Standard spelling).

Examples:
a hundred watts
a hundred meters
a hundred kilos
gigatons per year
three degrees celsius
three degrees fahrenheit

3.13. Capital letters can appear within company names in accordance with the company's typical, stylized orthography. Where a company name uses punctuation, this is avoided in the transcripts by using capital letters instead (see F.2.4).

Examples:
YouTube
FedEx
BitPay
WhatsApp
TikTok

3.14. For capitalization in media titles, see below F.6.

Section F: Punctuation

1. General Remarks

1.1. The only punctuation signs used in the corpus are the following (see section A.1):

`.` full stop / dot / period – the three descriptions are used interchangeably
`?` question mark
`,` comma
`'` simple inverted, undirected comma
`"` double quotation marks, inverted, undirected
`-` hyphens, only for compounds

1.2. The punctuation signs colon (`:`) and square brackets (`[…]`) are found exclusively in time stamps (see section A.2). They are not used elsewhere in the transcripts.
1.3. The triangular brackets (<…>) are used to indicate speech by other speech participants that are not the main speech participant of a file. They are not otherwise used in the transcripts.
1.3. Other punctuation symbols are not found in the corpus at all. For example, the corpus does not include semicolons (`;`) exclamation markss (`!`), slashes (`/`, `\`), brackets (`{`, `(`, `}`, `)`), etc.

2. Full Stops

2. 1 The corpus uses full stops (`.`) to indicate the end of an ordinary sentence token by default (see section A.2). Full stops are found at the end of all non-interrogative sentences like declaratives, exclamatives, imperatives, answer fragments etc.

Example:
```
[00:00:41]   ER the worst is yet to come.
```
A declarative sentence token ends in a full stop (in grey).

2.2. Full stops are also used within independent sentences in direct speech transcribed within one token (see section B.9).

Example:
```
[00:02:38]   She said, 'Give me a break. This is nonsense.'.
```

2.3. Full stops are used to indicate letters and acronyms (see section D.5).

Example:
```
[00:11:32]   And he came to J.P.L..
```
This sentence includes one full stop for the letter *L* of the acronym *JPL*, immediately followed by another full stop as the token-final punctuation sign.

2.4. Full stops are not used in any other function in the transcripts. In particular, proper names that have punctuation in their official stylized name are not transcribed with such punctuation. Instead, capitalization is used where appropriate (see E.3.13).

Example:
Not:
```
Musical.ly
```

but:

```
MuscialLy
```

3. Question marks

3.1. Question marks (`?`) indicate the end of interrogative sentences.

Example:

```
[00:01:11]   What's the point of having a ceiling if you raise
             it every time you hit it?
```

An interrogative sentence ends in a question mark (in grey).

3.2 The question mark (`?`) is used for direct matrix clause questions both if they are marked formally by subject-auxiliary inversion, as well as if they are interpreted as questions because of context and a characteristic rising intonation.

Examples:

```
[00:01:52]   And what is it?
```

This sentence is a direct question with subject auxiliary inversion (*is* comes before *it*). Therefore, it occurs with a token-final question mark (in blue).

```
[00:02:11]   ER it's made of cell phone parts?
```

This sentence is interpreted as a question because of the context and because it has a rising intonation at the end. Therefore, it occurs with a token-final question mark (in blue).

3.3. Questions marks are used to mark questions in direct speech that are transcribed within a larger token (see section B.9, in particular B.9.4, B.9.7, B.9.9, B.9.12).

Example:

```
[00:01:11]   So I ask, 'When is it gonna be over? When will
             finally win?'.
```

Interrogative sentences within direct speech end in question marks (in grey).

4. Commas

4.1. In general, commas are used in observance of rules of Standard English orthography.
4.2. The guidelines distinguish between cases where commas must be used, where commas can be used or not, and where commas must not be used.
4.3. **Commas must be used** in the following cases.
(a) Commas must be used to separate lexical fillers, such as *you know*, *I mean*, etc. (see C.3.2.).
(b) Commas must be used to separate sandwiched, parenthetical clauses (see B.8).
(c) Commas must be used to separate appositive elements, i.e. elements that elaborate, clarify or otherwise depend on some other sentence item (see B.3.2.(7), B.5).

Example:
[00:09:54] Now here I'm introducing a term that might be new to you, tropical cyclones.
[00:00:47] The policies pursued by the Federal Reserve have been debasing the value of our money, the U.S. dollar.

A comma must be used to separate an appositive. Here the phrases "tropical cyclones" and "the U.S. dollar" are appositives on "term" and "our money" respectively.

(d) Commas must be used to separate emphatic repetition such as *very, very good*, *really, really terrible*, etc. (see C.4.5.(c))

Examples:
[00:14:44] When that happens, it's the gold market, the gold market that's really gonna take off.
The phrase *the gold market* is focused by repeating it (*the gold market, not some other market, like the stock market*). It's separated by a comma.

[00:05:16] Seventy five percent of the jobs that the government claims were created were in the service sector, seventy five percent.
The number *seventy five percent* is stressed by repetition.
[00:01:02] It's not terrible, terrible.
[00:01:04] But yeah, it's not good.
The adjective *terrible* is repeated showing with negation that it is not the highest conceivable degree of 'terribleness'.

(e) Commas must be used to separate parenthetical material, like lists, enumerations, asides, elaborations, specifications, clarifications, etc.

Examples:
[00:00:41] It's in the metals, like gold and silver.
Commas are used to separate asides (*metals, like gold and silver*).
[00:00:44] I look at nickel, zinc and uranium prices.
Commas are used to separate elements in lists (*nickel, zinc and uranium*).

(f) Commas must be used to separate clearly non-restrictive relative clauses. This includes clause-final, clause-adjoined relative clauses.

Examples:

[00:02:27] Obama, who was a great president, would never do this.

[00:17:58] No one even has any idea that they're in the race, ER which is a topic for another discussion.

(g) Commas must be used after expressions introducing direct speech (see B.9.6).

(h) Commas must be used in a number of special constructions (see. B.10), such as the 'not only X, but Y' construction (see B.10.6), the correlative 'the more, the better' construction (see B.10.8), the 'is, is' construction (see B.10.12), the open 'so' construction (see B.10.13 etc.).

(i) Commas must be used to separate left-dislocation, i.e. phrases that are later explicitly resumed with a co-referential pronoun or full phrase (see B.3.2.(2), C.4,5.(b)).

Examples:

[00:01:04] And again, most of this inflation indexes, they're not designed to give out honest information about inflation.

[00:17:27] Now in fact, a lot of the people who are buying with zero down and using adjustable-rate mortgages, a lot of these zero-down buyers are actually buying a house.

[00:14:12] These Democrats, they always like to talk about how they admire John Kennedy.

In the examples above, the left-dislocated phrases are shown in blue, the resumption in yellow, and the comma in green.

(j) commas must be used to separate titles subtitles in media names (see below F.6.5).

4.4. **Commas may be used** in the following cases.

(a) Commas are optional after clause initial adjuncts. In this case, the transcribers may make their choice based on intonation, the length of pauses, a change in tone, etc.

Examples:

[00:03:11] So, I'm not sure.

or:

[00:03:11] So I'm not sure.

[00:06:31] As a pro-choice person, I'm not gonna stop them.

or:

[00:06:31] As a pro-choice person I'm not gonna stop them.

(b) Commas can be used after fronted / topicalized elements, especially topic clauses and hanging topics (which can be paraphrased with “as for X”, “as far as X is concerned”).

Examples:

[00:15:24] Whether it's gonna be worse under one or the other, I don't know.

[00:13:58] So whatever they take, they get a lot more than they dole out.

[00:11:44] Now, the only thing that could really stop the U.S dollar from a complete collapse, I mean, just like, you know, eighty or ninety percent or more collapse, the Federal Reserve would have to get out in front of the inflation curve.

[00:02:00] Okay, the child of the sixties thing, I'm not high right now.

[00:02:16] But normal conflicts, kids need to get in fights, and then settle it and go back to playing.

These examples show fronted elements or hanging topics in blue and the optional comma in green.

4.5. **Commas must not be used** in the following cases. The list of cases where commas must not be used is endless. Therefore, the following sections merely highlight the most common cases and potential problems.

(a) Commas must not be used before complement clauses.

Examples:

not:

```
[00:18:59]   They don't want to admit, that oil prices are
             going to rise so sharply
```

but:

```
[00:18:59]   They don't want to admit that oil prices are
             going to rise so sharply
```

English orthographical conventions prohibit commas before complement *that*-clauses.

not:

```
[00:01:33]   And that's probably as likely to be correct,
             as what Yellen thinks about the economy.
```

but:

```
[00:01:33]   And that's probably as likely to be correct
             as what Yellen thinks about the economy.
```

The construction *as likely as something else* forms one constituent with the second *as*-phrase as the complement of the adjective. It does therefore not occur with a comma.

(b) Commas must not be used before adjunct / adverbial clauses (although they may be used after them).

not:

```
[00:13:19] I'm paying almost seven thousand dollars a
month, to rent this house.
```

but:

```
[00:13:19] I'm paying almost seven thousand dollars a
month to rent this house.
```

English orthographical conventions disallows the use of commas before clausal adjuncts.

contrast with:

```
[00:16:56] If there was no welfare, he'd have to work.
```

The initial adjunct *if*-clause can be separated with a comma in accordance with normal rules of English orthography. If the clause is long or if there is a pause after it, the inclusion of a comma is preferred.

(c) Commas must not be used to separate clearly restrictive relative clauses.

Example:

not:

```
[00:00:09] That's all, that I can think about.
```

but:

```
[00:00:09] That's all that I can think about.
```

(d) Commas must not be used between subject and finite verb.

Examples:

not:

My mum, is calling.

but:

My mum is calling.

not:

And all that I can do, is tell you ER what I think.

but:

And all that I can do is tell you ER what I think.

not:

What I love about them, is they're cheap.

but:

What I love about them is they're cheap.

(e) Commas must not be used after token-initial conjunctions like and, but, or, even if there is long marked pause after these conjunctions.

not:

But, he didn't know.

but:

But he didn't know.

5. Single inverted commas

5.1. Single, undirected inverted commas are used in certain contractions, such as *isn't*, *I'm*, *it's*, *'cause*, etc. (see D.4).
5.2. Singe, undirected inverted commas are used to enclose direct speech (see B.9).

<u>Examples:</u>

`[00:03:54]`	`It says, 'The outlook on the U.S. triple A credit rating was raised to stable.'.`

Direct speech (*The outlook on the U.S. triple A credit rating was raised to stable*) is contained within single quotation marks (shown in red).

5.3. Single, undirected inverted commas are also used for mentioning words. In particular, they are used to introduce the appositive content of lexical items such as *word*, *argument* etc. The first item of mentioned phrases is capitalized. They are also used for *By 'X', I mean Y*, where the element X is enclosed in inverted commas.

<u>Examples:</u>

`[00:03:14]`	`And so he covers it with things like, 'Believe me.'.`
`[00:01:27]`	`So that's what 'You know' is.`
`[00:02:49]`	`The word 'Student' never showed up.`
`[00:12:15]`	`I don't even think we should use the word 'Loan' whenever we talk about the U.S. government.`
`[00:12:30]`	`I hate this argument right there, ER, you know, 'Children, they don't have the ability to make good decisions because they're not adults.'.`
`[00:01:38]`	`The answer is 'Yes' and 'No'.`
`[00:00:45]`	`And by 'We', I mean me.`
`[00:01:39]`	`When I say 'Sick' or 'Not feeling well', I just mean, like, I felt very off and weird.`

5.4. Note that constructions that convey quotative or mentioning meaning conventionally do not occur with single inverted commas. For instance, the corpus files transcribe "be called" as *she called him stupid* rather than *she called him 'stupid'*.

<u>Examples:</u>

not:

`They called it 'mobocracy'`

but:

`They called it mobocracy`

not:

`Or no, I meant 'colder'.`

but:

`Or no, I meant colder.`

6. Double quotation marks

6.1. Double quotation marks (`"  "`) are special symbols in the corpus. They always come in pairs, both being inverted and undirected.
6.2. Double quotation marks are reserved exclusively for **media titles**, such as:

- names of books,
- titles of radio shows or podcasts,
- song titles,
- magazine and journal names,
- titles of TV shows,
- named parts of the above categories, e.g., book chapters, poem names, a segment of a TV show, etc.
- video game names,
- official names of YouTube channels etc.

In other words, double quotation marks identify one very specific kind of entity, media titles, and they are not used for any other purpose.
6.3. Token-final punctuation or commas separating the media title from other material in the token appear outside of the double quotation marks.
6.4. Inside the double quotation marks, “important” (lexical) words are capitalized as is typical for titles. The first word within the double quotation marks is always capitalized.

Examples:

`It's like "The Communist Manifesto".`

(Name of a book. Token-final full stop appears outside of the double quotation marks.)

`another live episode of "Wall Street at Nine", ER the midweek market update.`

(Name of a radio show. Comma appears after `"..."` not inside of it. All words except *at* are capitalized.)

`that article in "The Rolling Stone Magazine".`

(Name of a magazine)

`the "Wall Street Journal"`

(Name of a magazine)

`a paper ER that we had in "Science" a few years back`

(Name of an academic journal)

`and then at midnight on "C.N.N. Headline News".`

(Name of a TV program)

`"The Sopranos", you know, on ... on ER on H.B.O.`

(Name of a TV program)

`We have a new segment here called "T.W. Single".`

(Name of a segment of a TV program)

`from "The New York Times", "The Washington Post"`

(Names of two well-known daily newspapers)

`welcome to the "Yoga with Adriene" channel`

(More or less official name of a YouTube channel)

6.5. Punctuation can appear inside the double quotation marks. Commas are used to separate main titles from subtitles. Token-final punctuation is used if the title is an independent main clause or exceptional token.

Examples:

```
I laid it out cold in a book that I wrote called, "Crash
Proof, How to Profit from the Coming Economic Collapse".
I love Band Aid's song, "Do they know it's Christmas?".
```

6.6. Double quotation marks can be embedded within each other. This is the case if the title of a media name includes within it another media name.

Examples:

```
a book called "The Art of "The Incredibles""
```

6.7. If a media title is disfluent, the three period ellipsis marker appears outside of the double quotation marks, and the double quotation marks are repeated if the disfluent title is corrected.

Examples:

```
[00:02:45]   And ER she was learning this song, "Big" ... "Big
             House", by Audio Adrenaline.
```

6.8. Double quotation marks are not used for any purpose other than media titles. In particular, they are not used for political movements, product names or companies names. It can sometimes be difficult to decide if the mention of a name refers to a media enterprise as a company (*The New York Times is laying off two hundred journalists*) or a media name as a title (*"The New York Times" is a newspaper*).

not:

```
Canadian ER gold miner, "Goldcorp"
"Occupy Wall Street"
"The Commissioner of the Internal Revenue Service"
"Suntory Holdings"
"Chrysler"
"WhatsApp"
"Google"
"C.N.N." and "Fox News"
```

but:

```
Canadian ER gold miner, Goldcorp
Occupy Wall Street
the Commissioner of the Internal Revenue Service
Suntory Holdings
Chrysler
WhatsApp
Google
C.N.N. and Fox News
```

6.9. Double inverted commas are used for specific books and other titles, not for generic descriptions or collections of books. In particular, the Christian holy book, *the Bible*, is not put into double inverted commas since it is a collection of books, not a single book.

not:
```
the "Bible"
"The New Testament"
the "Harry Potter" series
```
but:
```
the Bible
the New Testament
```
but: `"The Gospel of Saint John"`
```
the Harry Potter series
```

6.10. Double inverted commas are not used for apps or websites (e.g. WhatsApp, YouTube) or other forms of generic software. They are used for names of games, however.

```
[00:03:02]   And then I have Twitter.
```
but:
```
[00:02:48]   I've "Monopoly", which I play on really long car
rides, "Trivia Crack", which is, like, the latest and
greatest fad.
```
Twitter is regarded as a proper name, an app on a phone, whereas *Monopoly* and *Trivia* crack are interpreted as media titles by virtue of being (video) games.

7. Hyphens

7.1. Hyphens appear in the corpus. Their function is to indicate certain compounds, such as *long-term*, *sell-off*. However, the use of hyphens is not always strictly regulated. There can be substantial subjectivity as to whether a hyphen should be used or not. The following sections outline some general rules about the use of hyphens.

7.2. Hyphens are not used in compounds that are written as single words in standard orthographical conventions. There can be great subjectivity with regards to whether a word is fossilized and conventionalized enough to justify spelling as a single word or if they should occur with a hyphen or with a space. Compounds that are spelled as single words include:

Examples:
```
automakers
broadcast
bullseye
candlelight
cellphone
ecosystem
hashtag
highway
freshwater
greenhouse
iceberg
kilometer
lifetime
meltwater
online
outlet
photosynthesis
```

```
pipeline
rainbow
rainfall
riverbed
skyrocket
southeast
touchscreen
treadmill
upstream
```

7.3. Hyphens are never used in noun-noun compounds. Instead, combinations of noun plus noun are spelled with spaces between the component parts.

Examples:

```
air conditioning
climate scientist
crude oil prices
auto workers in Detroit
Russia was the bread basket of Europe
video gaming industry
sound bite speeches
phone case
sea level rise
the ice melt from Greenland
private sector actors
the Chris Christie workout
```

7.4. Hyphens are not used in phrasal compounds. Instead, complex modifications are spelled with spaces between the component parts. There may, however, be exceptions.

Examples:

```
The British is best mentality
a buy the rumor sell the fact reaction
a do it yourself cell phone kit
not exactly a in a week kind of thing
the most energy using of all animals
```

but:

```
[00:05:16]     And, you know, I guess it’s a little early
               for the I-told-you-sos.
```

7.5. In general, hyphens should not be used twice in a row. Hyphens are usually used only between two words. However, some transcriptions may sometimes include multiple hyphens in a word.

Examples:

```
anti gay marriage
anti money laundering
analog to digital conversion
this last up and down ER modulation
```

the north to south ER injustice
you deserve to get you know what
a dog and pony show
under the table ER deals
this is a bunch of form over substance
there's round the clock coverage for weeks
four inches below average, year to date.
the most up to date shows

7.6. Hyphens should be used for noun compounds with two elements where the final element is an adverbial particle, such as, *up, down, off*, etc.

Examples:

a slow-down
a sell-off
a draw-down of carbon dioxide
a speculative blow-off
a blow-up of that ER blue line
a little run-down
a big turn-around
a set-up

but: leftovers

7.7. Hyphens should be used for compounds of two elements where the two elements are of a different word class, e.g., adverb-noun, adjective-participle, adjective-noun, etc.

Examples:

off-line
it's now open-ended
it's open-access, open-source
present-day conditions
low-wealth people
the randomly-typing monkeys
a very cost-effective way
value-added jobs
a spooky-looking mask
this small-scale ocean mixing
very low-density roads
the deep-water circulation
a very round-about way
on-shore drilling
the Sun-warmed Earth
their same-store sales for January
an all-time record high
his Blu-ray player
so well-prepared
jet-lagged
getting long-term results
adjustable-rate mortgages
self-serving comments

```
by-products
thank-you notes
to out-vote them
oil in Renminbi-terms
a one-way street
a laissez-faire approach
I'm self-employed
income-producing foreign stocks
a full-on communist
They're gonna get blind-sighted.
over-confident
```

7.8. Hyphens are used where they appear in standard orthographic spellings of compounds, brand names, proper names, etc.

Examples:

```
band-aid
```

7.9. Hyphens can be used in conjunction with certain transparent, word-like but functional affixes. It is not always clear if a hyphen should be used with such affix-like words or not. Examples include *anti-* ("against"), *co-* ("with"), *de-* ("adverse"), *inter-* ("in between"), *mega-* ("very high, extreme"), *micro-* ("very small"), *multi-* ("many"), *non-* ("not"), *post-* ("after"), *pre-* ("before"), *re-* ("again"), *sub-* ("under"), *super-* ("very high, extreme"), trans- ("through") *-like* ("similar") *-free* (without), *-ish* ("similar") and others.

Examples:

```
anti-gun laws
anti-N.R.A.
anti-Christianity
     but: antisemitism
the co-pilot
a co-op student
it de-syncs it
pre-Columbian inhabitants of the Caribbean
multi-meter sea level rise
re-investing the interests
re-set
they had re-structured their balance sheets
Bush is more likely ER to get re-elected.
mega-boulders
micro-cloth
post-docs
sub-tropical
multi-year ice
to multi-task
non-hedged
the last inter-glacial period
super-storms
     but: super major problems
this bullseye-like feature
Sub-Saharan Africa
```

trans-dimensional
trans-polar
they're tax-free
Christi-Gate
non-creditworthy borrowers
one thirty-ish

However, hyphens are not used if such strings are fossilized, opaque, and are felt to be an integral part of the word.

reiterate, rely, recognize
(not re-iterate, re-ly, re-cognize)
interrupt, interruptions
(not inter-rupt, inter-ruptions)
transport, transit
(not trans-port, trans-it)
geochemical data
(could probably also be geo-chemical data)
suburban
(could perhaps also be sub-urban)

7.10. Hyphens are not used in combinations of *every*, *some*, and followed by another word like *thing*, *body*, *one*, etc. There is an exception however: The word *no-on*e is always spelled with a hyphen.

Examples:
everywhere
someone
nothing
but:
no-one

7.11. Hyphens are used for compounds with the word *self* as the first element.

Examples:
self-harm
self-esteem
self-centered
self-confident
self-taught
self-employed

7.12. Hyphens are not used with numerals.

Examples:
a seven and a half trillion dollar national debt
a thirty year mortgage
the one hundred year climate response
a fifty year period
if you took a five year snapshot
one meter sea level rise

7.13. Hyphens are not used for numbers indicating the measure or extend of a following unit and predicate, such as *ten kilogram heavy*, *seven years old*, *three foot high* etc., and similar constructions. This is true even where the predicate is plural.

Examples:

a thirty three kilometer thick ice sheet
these single year events
fifteen dollar an hour minimum wage
twenty and thirty year bonds
ninety day T-bills
ten year or twenty year doubling times
six year old children
eleven year olds
a year long subscription

7.14. Hyphens are not used for *best* and *worst* followed by another element.

Examples:

the worst case scenario
the best case outcome

7.15. Hyphens are not used for *quasi* followed by another element.

Examples:

quasi activist
quasi advocate

7.16. Combinations with *mid* are spelled as single words wherever possible.

Examples:

midwinter
the Midwest
midday
my Econ midterm
midcontinent
but:
the mid nineteen forties