Feedback

Speech Processing, first assignment, November 2015

Marking process

- 3 markers were trained by me for this specific assignment, and given examples from a previous year (marked by me)

- Parts I and II

  - UG - marker has taken this course for credit in the past

  - PG - marker has taken this course for credit in the past

- Part III - all UG+PG marked by the same person in (not one of the above)

- marking took ~50 minutes per assignment

Moderation process

- All moderation done by the lecturer

- Moderation (done separately for UG and PG):

  - inspecting mark distributions for class as a whole

  - quickly inspecting **every** individual assignment

  - making a few minor changes to individual marks

  - global scaling of marks (separately for parts I&II, part III)

- *Your final overall mark will not equal the sum of the marks in the marking scheme: it will be about 10% higher, due to moderation*

Getting the most out of the feedback

- Read **every** comment, quickmark and grade

  - remember that markers are working "against the clock" and so their comments may be **terse** (but do not interpret this as being rude)

- Ask the **lecturer** for clarification if there is anything you do not understand

  - e.g., via the Forum (Assignment 1 > Feedback)

  - note: you are not able to directly talk to the markers

What you should receive

post-moderation final grade, also with any late penalty applied

quickmarks (hover to see full text)

highlighted text with comments

bubble comments

text comments

UG: comments per section

structured marking

pre-moderation total score

Annotations and marks

- Quickmarks

  - UGs: linked to marking scheme

- Highlighting

  - green=good

  - yellow=query

Annotations and marks

- Text comments

  - either visible text

  - or bubbles to click

- Marking scheme

  - UGs: extra comments and links to quickmarks for Parts I & II

Part I+ II
Fair attempt at describing the p
to underlying theory are presen

3	I: Text processing (5)	ⓘ
A bit more detail would have been good, and figures!		3

5	I: Pronunciation & prosody (5)	ⓘ
Fair attempt, but lacking detail and figures.		3

Most common failing

- Literature review

  - failing to be **critical** (average mark was 5 / 10)

Other common failings

- **Word count**

  - not required to use *all* 3000 + 1000 words

  - but 1500 + 500 words is probably too short to demonstrate your full understanding of the material unless you're a brilliant writer

- Describing Festival (as a piece of **software**), rather than talking about the **problem** of Text-to-speech

Detailed feedback

- **Content**

  - feedback slides from last year were structured according to content - you should already have read those

  - use the Forums to get further feedback on specific points in your assignment, and to continue to improve your understanding of the material

- **Writing**

  - This year's slides focus on scientific writing, to complement the above

Structure

Table of Contents

Use a consistent numbering scheme

Formatting

The generated waveform can be listened through the following link: https://www.dropbox.com/s/0fkqgp44ftxb9es/assignment1.wav?dl=0

Provide friendly URLs, and try not to line wrap them

Formatting

Background

For the purposes of this work, the Festival v...

festival config.scm. This voice uses a unit-selection t...

one example of each diphone. This means that in the ...

diphone, which are then selected and concatenated based on ...

appear, and how well they would concatenate together.

The voice used is Scottish English, so the ut... tested were all in English, and the

intonation and accent are based on Scottish ...

Section 1: Lab report

Part I: Stepping through the synthesis process in Festival

Methodology

This report was written based on tests on various tens of utterances synthesised through the

Festival system. Various sentences had to be tested in order to make the different parts of the

synthesis process work. When possible, the examples reported below will refer to the same utterance,

Hard to decode the meaning of bold / underlining / font size / indentation. Better just to number headings (1, 1.1, etc)

Incorrect bibliography style

Webpages

English Oxford online dictionary: http://www.oxforddictionaries.com

SpeechZone forum: http://speech.zone/forums/

give date of access for online material (because it's subject to change)

speech.zone. (2015). *Forum Topic: 'Duration and Intonation'*. Retrieved October 26, 2015, from http://speech.zone/forums/topic/duration-and-intonation/

Like this

Incorrect bibliography style

[7] Black, A., Lenzo, K. and Pagel, V., 1998, Issues in Building General Letter to Sound Rules

[8] Dick R. van Bergem, 1994, REFLECTIONS ON ASPECTS OF VOWEL RE-DUCTION

[9] Christopher D. Manning, 2011, Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?

[10] Steven J. DeRose, 1988, Grammatical Category Disambiguation by Statistical Optimization
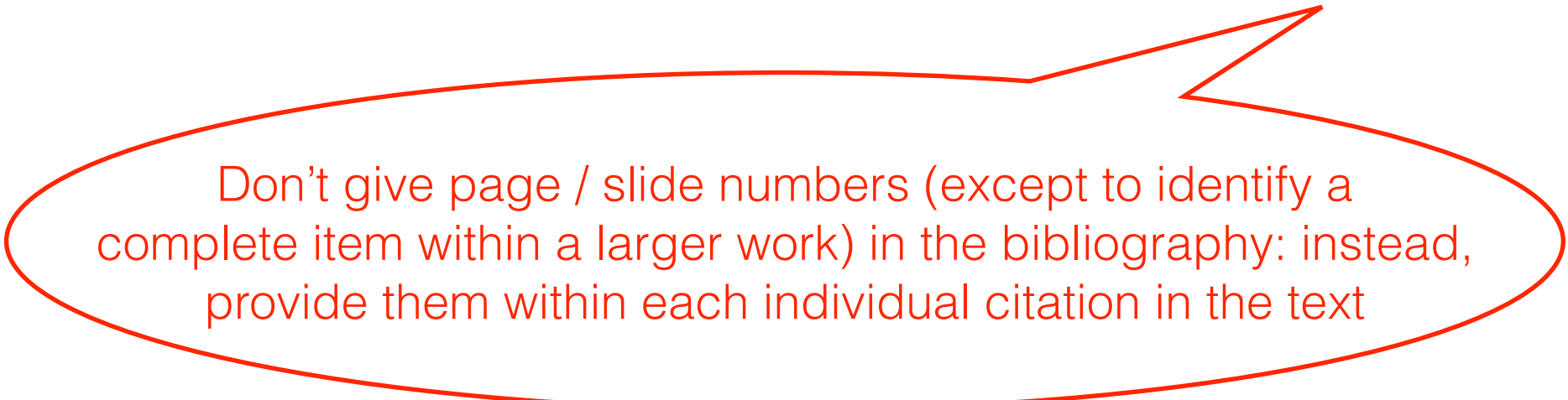
Missing details in every entry.
Are these books or papers? How could we find them?

15

Incorrect bibliography style

Jurafsky, D. & Martin, J. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, N.J.: Dorling Kindersley (India) Pvt., licensees of Pearson Education in South Asia. p. 54-57, 141-144, 257f, 268-272

King, S. (2015). *Speech Synthesis Lecture Pack.* University of Edinburgh. Slides 7, 23, 28, 53-55

King, S. (2015). *Feedback slides from 2014-15 (examples of good and bad work).* Slide 30

Don't give page / slide numbers (except to identify a complete item within a larger work) in the bibliography: instead, provide them within each individual citation in the text

Incorrect bibliography style

King, S. (2015a). Spee... slide 32.

King, S. (2015b). ...e, slide 36.

King, S. (2015c). Speech processing: Speech synthesis lecture, slide 64.

King, S. (2015d). Speech processing: Speech synthesis lecture, slide 9.

Use of 'a', 'b',… would be correct way to distinguish different items from the same author(s) in the same year.

Just one entry per item. Provide the slide number at the point where you cite the item in the text.

Incorrect bibliography style

[1] Alan Black et al, The Festival Speech Synthesis System, V1.4, 1999

[2] Simon King, 2015, SP Lecture Pack 2, Slide 12

[3] Simon King, 2015, SP Lecture Pack 2, Slide 16

[4] Jurafsky and Martin, "Speech and Language Processing", 2009, pg 250, Ch 8.0, Fig 8.2

[5] Simon King, 2015, SP Lecture Pack 2, Slide 17

[6] Jurafsky and Martin, "Speech and Language Processing", 2009, pg 250, Ch 8.0

[7] Alan Black et al, The Festival Speech Synthesis System, V1.4, 1999, Ch 14.1

[8] Alan Black et al, The Festival Speech Synthesis System, V1.4, 1999, Ch 14.2

[9] Alan Black et al, The Festival Speech Synthesis System, V1.4, 1999, Ch 15.1

[10] Jurafsky and Martin, "Speech and Language Processing", 2009, pg. 251, Ch 8.1

[11] Jurafsky and Martin, "Speech and Language Processing", 2009, pg. 251, Ch 8.1.1

[12] Jurafsky and Martin, "Speech and Language Processing", 2009, pg. 252, 8.1.2

[13] Alan Black et al, The Festival Speech Synthesis System, V1.4, 1999, Ch 13.1

[14] Alan Black et al, The Festival Speech Synthesis System, V1.4, 1999, Ch 15.2

[15] Jurafsky and Martin, "Speech and Language Processing", 2009, pg. 140-141, Ch. 5.5

[16] Jurafsky and Martin, "Speech and Language Processing", 2009, pg. 256, Ch. 8.1.3

Another example

Incorrect bibliography style

This last section of the assignment is dedicated to the summary, critical review and comparison/contrast of two different papers, one on Prosody and the other on Intonation:

- Ann Syrdal, Gregor Möhler, Kurt Dusterhoff, Alistair Conkie and Alan W Black (1998). "Three Methods of Intonation Modeling", in Proc. 3rd ESCA Workshop on Speech Synthesis, pp. 305-310

- Cameron S. Fordyce and Mari Ostendorf (1998). "Prosody Prediction for Speech Synthesis using Transformational Rule-based Learning" in Proc. Int. Conf. on Spoken Language Processing (ICSLP) 98.

Don't give full bibliographic details within the text. Put them in the bibliography, and **cite** them in the text.

Incorrect bibliography style

[3] Jurafsky, D. and Martin,J.H.,(2009). *Speech and Language Processing Second Edition.*
New Jersey: Pearson Education

The edition is not part of the title.

Use author's surname and initials, not full given name(s).

[1] Chapter 4.5; Taylor, Paul Alexander. Text-To-Speech Synthesis. Cambridge, UK: Cambridge University Press, 2009. Print.

Incorrect bibliography style

Bibliography
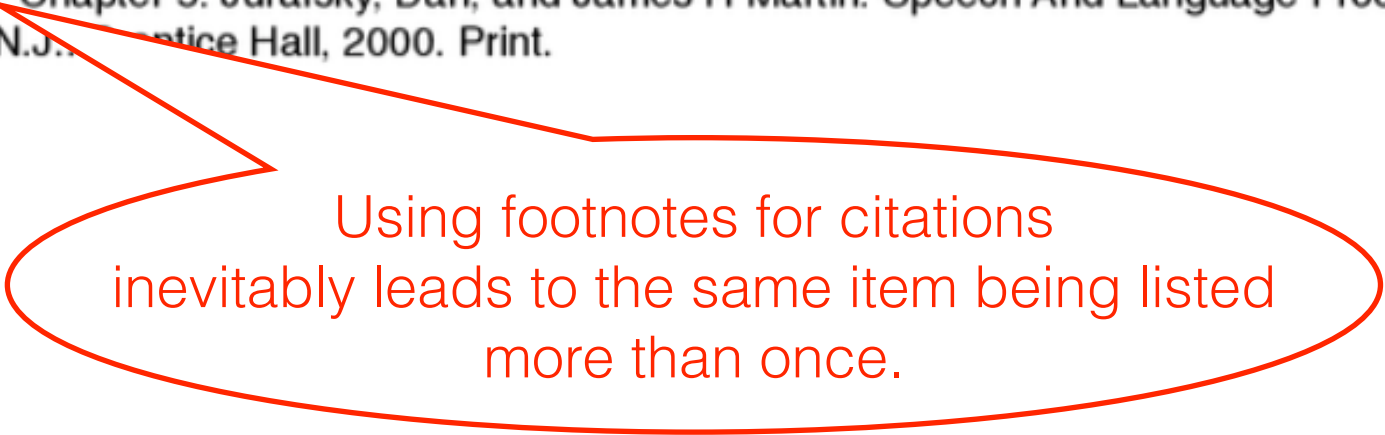
- King, (2015). Speech Synthesis Lecture Pack. University of Edinburgh.
- Jurafsky and Martin, (2009). *Speech and Language Processing Second Edition*. New Jersey: Pearson Education.
- Black, Taylor, and Caley, (1999). *The Festival Speech Synthesis System, System documentation, Edition 1.4, for Festival Version 1.4.0.*
- Black, *The Festival Speech Synthesis System*, The Centre of Speech Technology Research, The University of Edinburgh. http://www.cstr.ed.ac.uk/projects/festival
- Tull and Rutledge (1993). *Linear Predictive Synthesis of Vowels for Pitch Enhancement of Female Geriatric Esophageal Speech*. Northwestern ...ersity.

No need to itemise as a list

Suboptimal bibliography style

[1] Chapter 5. Jurafsky, Dan, and James H Martin. Speech And Language Processing. Upper Saddle River, N.J.: Prentice Hall, 2000. Print.

[2] Chapter 5. Jurafsky, Dan, and James H Martin. Speech And Language Processing. Upper Saddle River, N.J.: Prentice Hall, 2000. Print.

Using footnotes for citations
inevitably leads to the same item being listed
more than once.

Poor choice of items to cite

[8] Wikipedia contributors, "Homograph," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Homograph&oldid=687077559. Accessed October, 2015.

Don't cite Wikipedia unless there really is no other source available.

Good bibliography

for version 1.4.0. http://www.cstr.ed.ac.uk/projects/festival/manual/

Clark, R.A.J., Richmond, K., & King, S. (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, Vol.49, 317–330.
http://www.cstr.ed.ac.uk/downloads/publications/2007/clarkrichmondking_specom2007.pdf

Donovan, R.E. (1996). *Trainable Speech Synthesis*. Ph.D. Thesis. Cambridge: Cambridge University Engineering Department. http://mi.eng.cam.ac.uk/reports/svr-ftp/auto-pdf/donovan_thesis.pdf

Donovan, R.E., & Eide, E.M. (1998). The IBM Trainable Speech Synthesis System. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP) 1998*, Vol.5, 1703–1706. Sydney, Australia.

Hunt, A.J., & Black, A.W. (1996). Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database. In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1996*, Vol.1, 373–376. Atlanta, GA, USA.

[Jurafsky...] ...eech and Language Processing: A... ...tional Linguistics, and Speech Recognitio... ...NJ, USA: Prentice Hall.

King, [S.] (20...). Speech Processing (LASC11065): lecture slides on speech synth... University of Edinburgh.

Klatt, D.K. (1979). Synthesis by rule of segmental durations in English sen... In: Lindblom, B., & Öhman, S. (eds.), *Frontiers in Speech Communication Research*, 287–29... ...ndon: Academic Press.

Taylor, P. (2008). *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.
http://svr-www.eng.cam.ac.uk/~pat40/ttsbook_draft_2.pdf

Full details for every entry

Always use the published version where it exists, not an online draft

24

Incorrect citation style

tags, and predicting where phrase breaks should occur"[King, Tokenisation and Normalisation recording]. The text processing step is essential to dis-

Don't give title within the text. Put it in the bibliography.
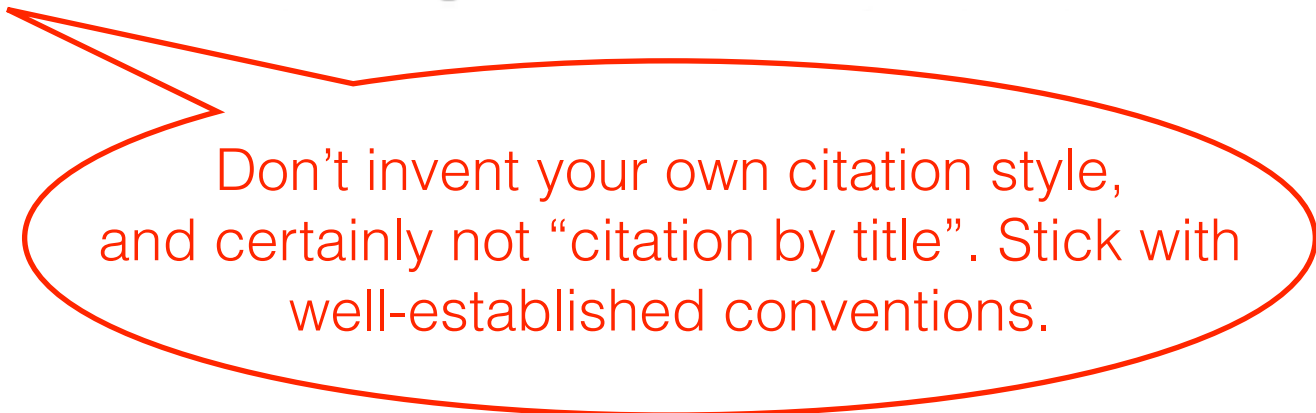
Don't give month within the text, but do give year in full (e.g., 2015).

been "trained automatically on a very large set of labeled data" [King: Oct 15, slide 32]. It can be effectively used as a decision tree for determining

Incorrect citation style

Investigating the Limitations of Concatenative Synthesis (ILCS)

ILCS carried out three experiments with a concatenative text-to-speech (TTS) system.

Don't invent your own citation style, and certainly not "citation by title". Stick with well-established conventions.

Incorrect citation style

The system, described in the paper written by Donovan and Eide, use approach combined with dynamic programming to synthesis speech. First

Better style to talk about the method, not the paper itself. "Donovan and Eide" is not a proper citation: needs the year in parentheses.

Don't use authors' names as a section heading. Focus on the method that they present.

Andrew J. Hunt and Alan W. Black

The system employed in this paper is the CHATR synthesis system, and is an example of the concatenative approach. The paper goes into general background of

Incorrect citation style

Cite both authors when there are two. For more than two, either list all or give first author and use "et al".
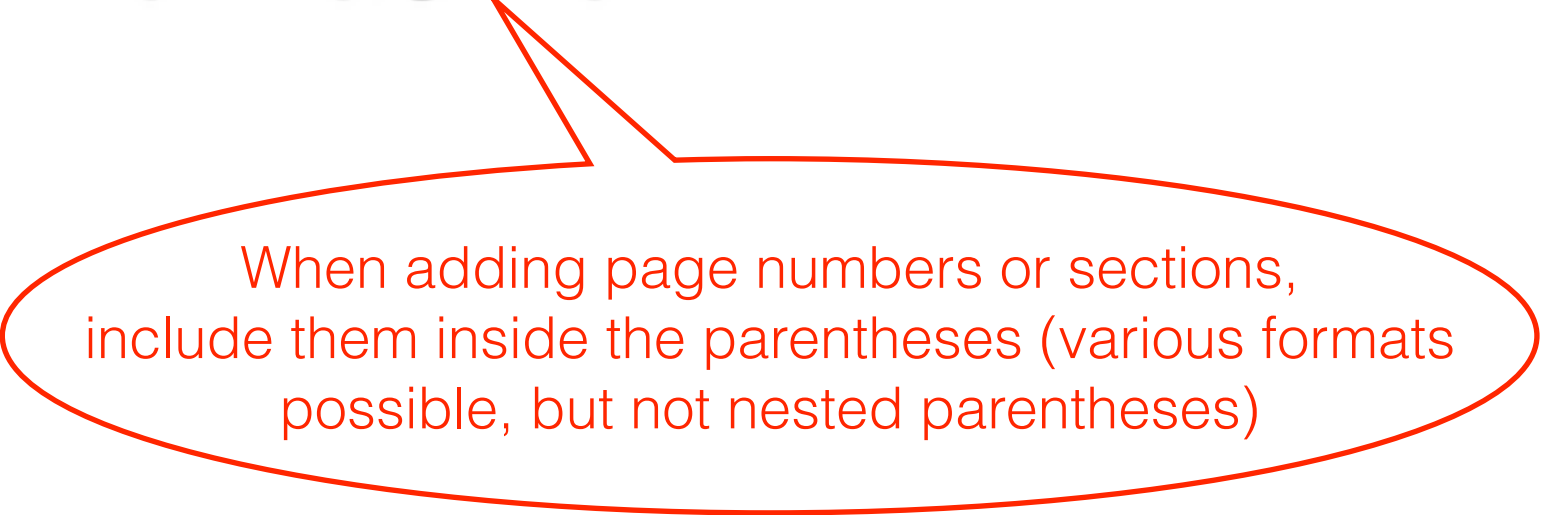
boundaries (Syrdal 1998). The final m

2.1 As a result, Fordyce and Ostendorf generate much more incisive conclusions than Syrdal.

Don't assume the first author did all the work: always mention all authors, or use "et al". Always cite correctly, including the year.

Incorrect citation style

(Taylor (2009), p. 47).

When adding page numbers or sections,
include them inside the parentheses (various formats
possible, but not nested parentheses)

Quoting and attribution

inference that the abstract notion of emotion is reflected by F0, duration and amplitude. It should be pointed out that adding expressivity to a synthetic voice has two aspects, namely, the naturalness and arousable effect. The expressiveness model chosen the system's realisation should be in order to convey the intended expression. Edgington uses the copy synthesis[6] method. This method is appropriate for testing to which extent a given synthesis method can reproduce the same auditory

---

[6] A straightforward approach to control the acoustic correlates of a given emotion. The parameters that can be controlled in the synthesizer (e.g., F0 and duration) are measured in an expressive recording, and used directly as the input to a synthesizer (Schröder 2009).

Fails to use quote marks.

University of Edinburgh
School of Philosophy, Psychology and Language Sciences
Speech Processing
Page 16

ExamNo: B073793

impression as a human speech sample; however, the generalisability is obviously low (Schröder 2009). Combining voice quality and prosody of different expressive styles, Montero et al. and Audibert et al. found that there is not yet a clear picture which emotions predominantly rely on voice quality and which can be recognised based on prosody. It may even be that different speaker strategies exist to express the same emotion.

30

Quoting and attribution

The CART will use contextual data, e.g. a phoneme is word-initial, in order to arrive at a correct pronunciation. At this stage the system will have arrived at a pronunciation for each word as it would be spoken in isolation, the "citation form". (King (2015), slides 27-29).

Original source

fails to use quotation marks, but does cite the source

Post-lexical rules

- The lexicon and letter-to-sound rules arrive at a pronunciation for each word *as it would be spoken in isolation*, known as the "**citation form**"

Quoting and attribution

uses decision trees; it does so in the 'phrasify' process to correctly tag pauses. Decision trees are built by finding the question whose resulting partition is purest, splitting training data at each question.[2] By purity, we mean 'entropy', where 'low entropy' means highly predictable. This is a
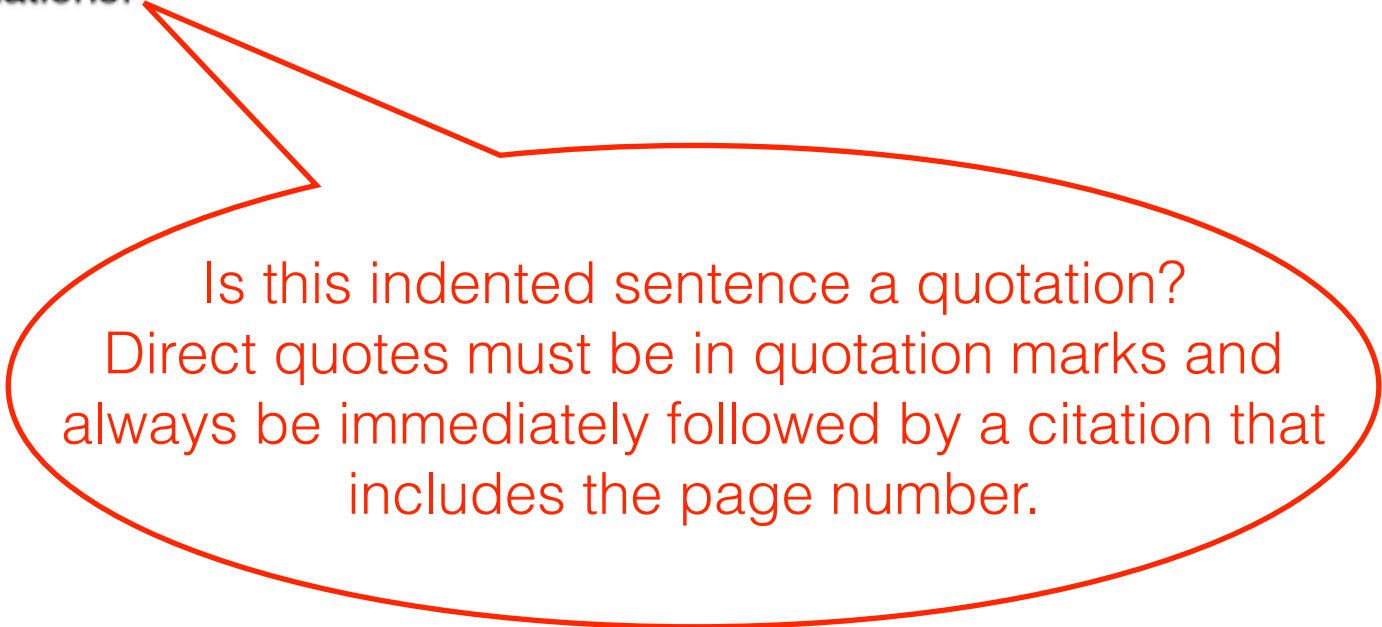
Another unquoted quotation. Adding a citation (here, by footnote) is not enough.

Quoting and attribution

therefore, it doesn't have the same issues as decision trees when it comes to training data.[2]

Transformations are ordered; later transformations are dependent on the outcome of earlier transformations.

Is this indented sentence a quotation? Direct quotes must be in quotation marks and always be immediately followed by a citation that includes the page number.

Quoting and attribution

9. **Post Lexical rules:**
   - Command: **(PostLex myutt)**
   - Utility of the Process: PostLex is a module which is run after accent assignment but before duration and F0 generation. This is because knowledge of accent position is necessary for vowel reduction and other post lexical phenomena and changing the segmental items will affect durations. The post-lexical rules indicate the context in which the words are being spoken making the speech sound more natural.
   - Method: The PostLex first applies a set of built input utterance, which are set up on a per voice post-lexical procedures are again take by CARTs rule is applied is as follows:

   The 's in English may be pronounced in a number of different ways depending on the preceding context.

   (a) If the preceding consonant is a fricative or affricative and not a palatal labio-dental or dental a schwa is required (e.g. bench's) otherwise no schwa is required (e.g. John's).

   (b) Also if the previous phoneme is unvoiced the "s" is rendered as an "s" while in all other cases it is rendered as a "z".For our English voices we have a lexical entry for "'s" as a schwa followed by a "z".

   Once the context is determined, the post-lexical rules are applied

*Unattributed text from several sources; **this is plagiarism***

Quoting and attribution

The rule based approach uses a target interpolation scheme with accent and boundary markers that are ToBI labels. Targets are placed with reference to syllable structure, within a pitch range specified by top and base lines which are derived from the speaker's speech using a CART tree. In the test, contour was predicted based on the segmentation and the ToBI labels for the test utterances.

The Tilt approach uses Tilt accents estimated from the accents marked in the speech database and parameterized using parabolic approximation. CART trees are trained to related segmental and prosodic features to each of the parameters. The parameters being: pitch accents, boundary tones, connections and appropriate set of parameters is predicted using th was predicted from the derived Tilt accents and bou passed on to the synthesizer module for processing.

Summarising the paper by extractive summary, without attribution of the source text; **this is plagiarism**

Quoting and attribution

are suggested: weight space search and regression training. [3] For weight space search, we determine the distance of our chosen set of units from the natural waveform using the "objective distance function". This is repeated for many utterances and weight sets and the most consistently performing weight set is chosen. Regression training uses the objective distance function to calculate the acoustic difference between one phoneme and all the other instances of that same phoneme in the database. It then retains the twenty best units and uses linear regression to predict the weights based on this.

This summary is too close to the source text; **this is poor style.**

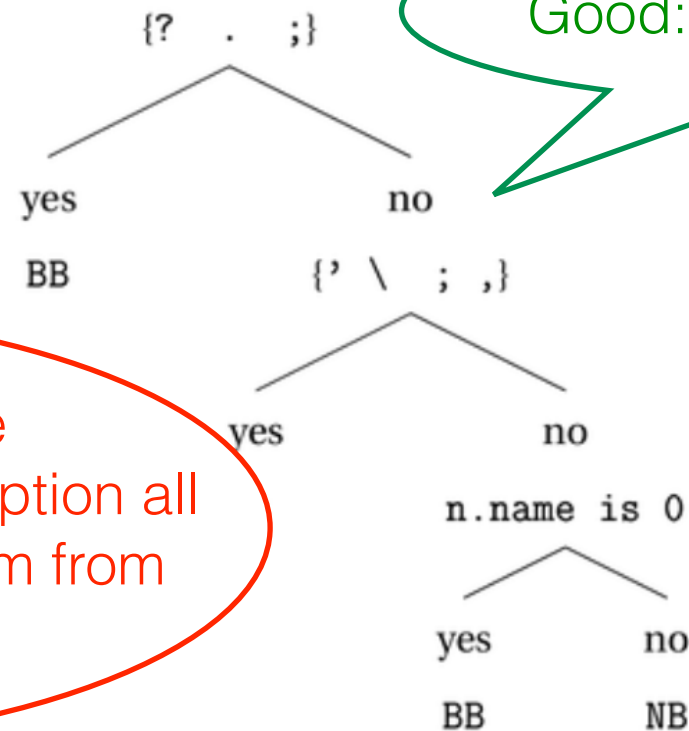# Figures that could be better

```
(set! simple_phrase_cart_tree

'

((R:Token.parent.punc in ("?" "." ":"))
  ((BB))
  ((R:Token.parent.punc in ("'" "\"" "," ";"))
   ((B))
   ((n.name is 0)
    ((BB))
    ((NB))))))
```

Figure 2. Simple Phrase CART Tree

Why not actually draw the tree!?

Figures that could be better

They are assigned according to a word's following character based on the CART:

Good: draws the tree

"Inline" figures should be avoided: always number and caption all figures, so you can refer to them from elsewhere in the text

This relatively simple CART only considers punctuation and, in the last node, checks for an
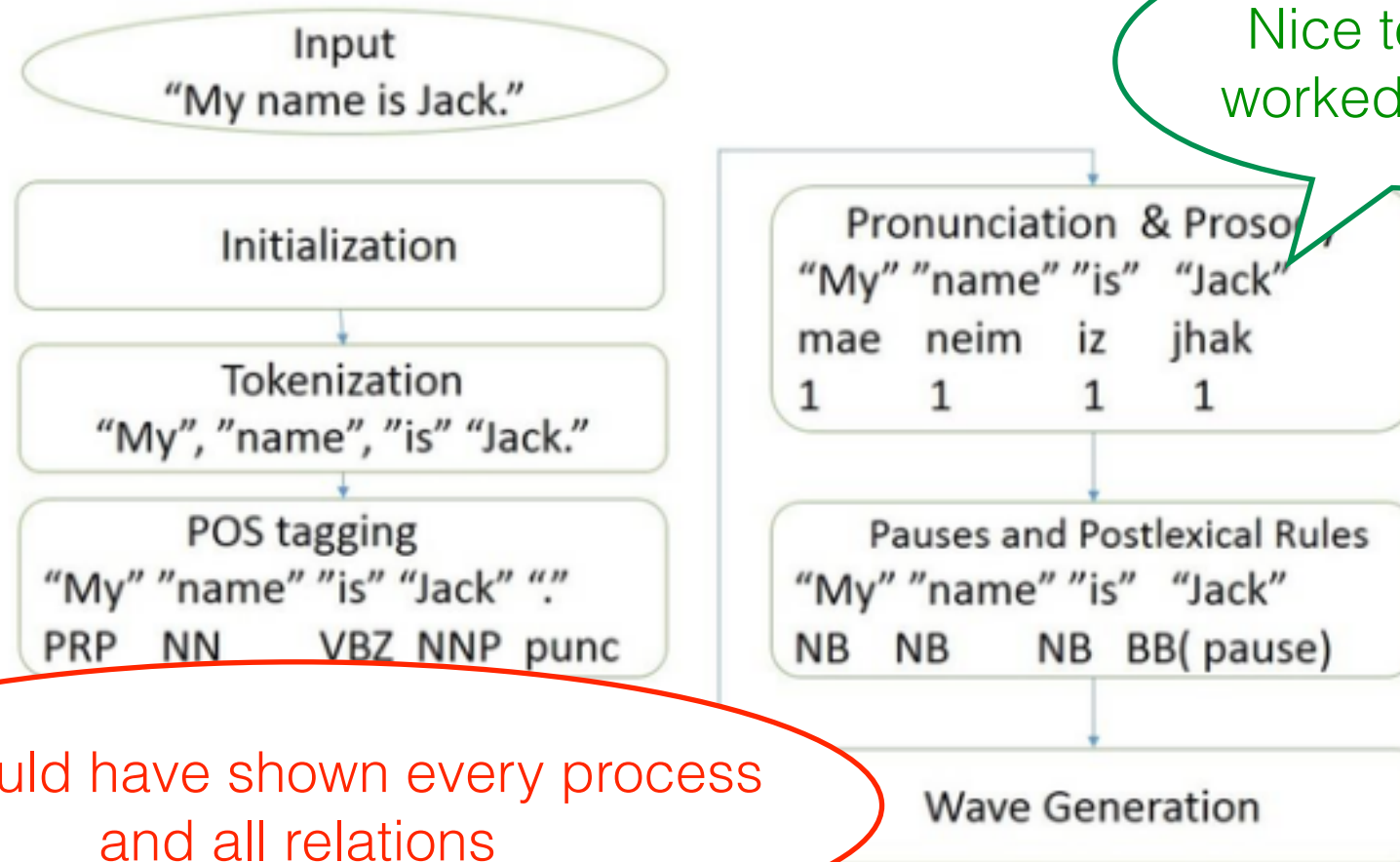
Figures that could be better

Input
"My name is Jack."

Initialization

Tokenization
"My", "name", "is" "Jack."

POS tagging
"My" "name" "is" "Jack" "."
PRP   NN   VBZ NNP punc

Pronunciation & Prosody
"My" "name" "is" "Jack"
mae   neim   iz   jhak
1   1   1   1

Pauses and Postlexical Rules
"My" "name" "is" "Jack"
NB   NB   NB   BB(pause)

Wave Generation

Nice to have a worked example

But could have shown every process and all relations

Figure 2. How My name is Jack. is processed in Festival. Modules and output are shown in the boxes (not all are illustrated in the chart).

Figures that could be better

id _12 ; name m ;
id _13 ; name ae ;
id _15 ; name n ;
id _16 ; name ei ;
id _17 ; name m ;
id _19 ; name i ;
id _20 ; name z ;
id _22 ; name jh ;
id _23 ; name a ;
id _24 ; name k ;

Verbatim output! Could write this phone sequence on one line.

Figures that could be better
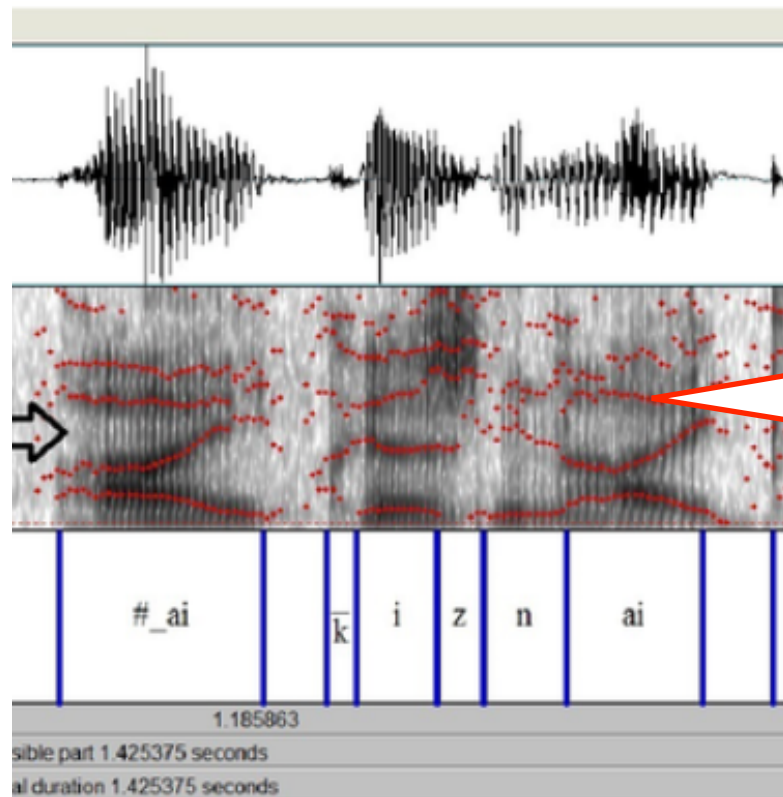
Praat: turn off the formant and F0 tracks, unless you are referring to them

#_ai   k̄ i z n ai

1.185863

sible part 1.425375 seconds

al duration 1.425375 seconds
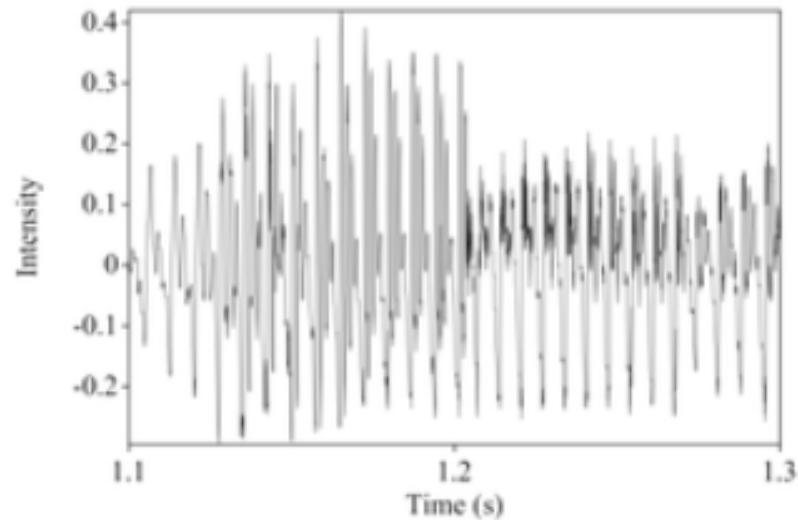
Figures that could be better

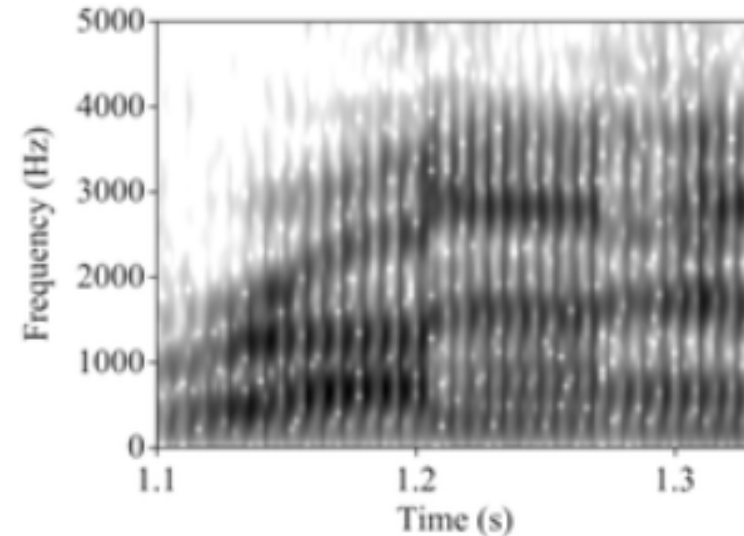Figure 3. Join between diphones seen as discontinuity in the waveform envelope.

Figure 4. Spectra do not align well at join. Burst of energy across many frequencies visible as dark vertical band at centre of this section of spectrogram.

Mark the precise point of interest, or at least mention it ("at 1.2s") in the caption

Good examples of an audible join

Figures that could be better
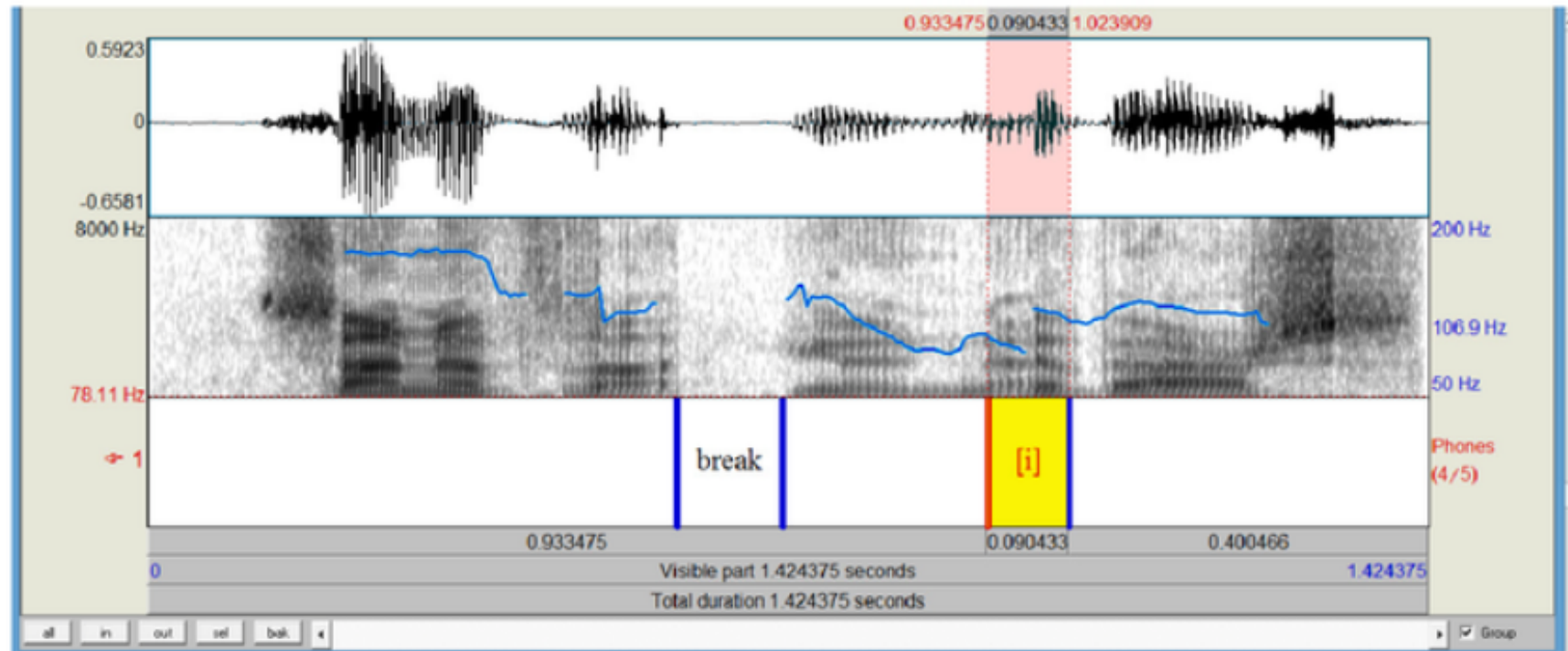
Note to graph 2: waveform, spectrogram and notes added to the sentence:

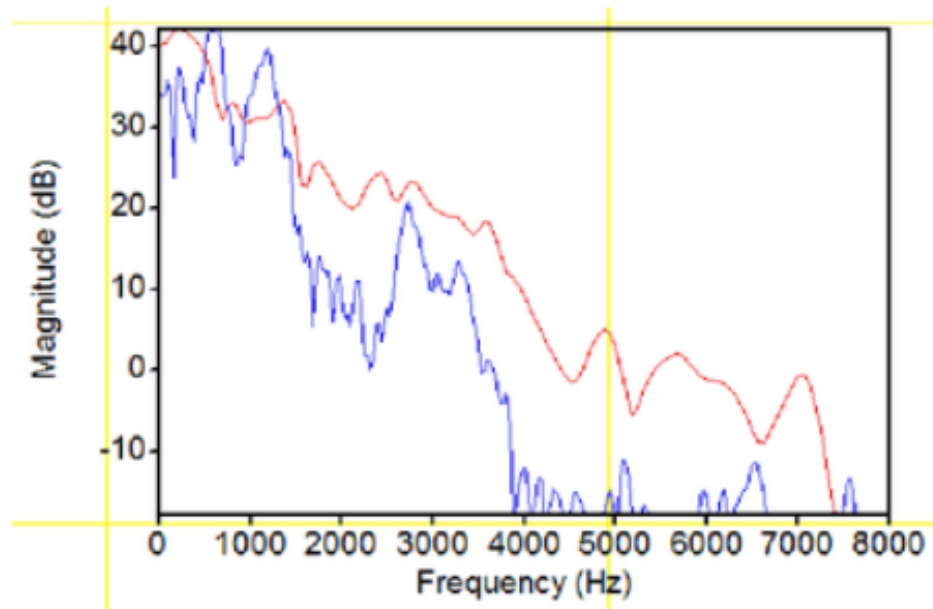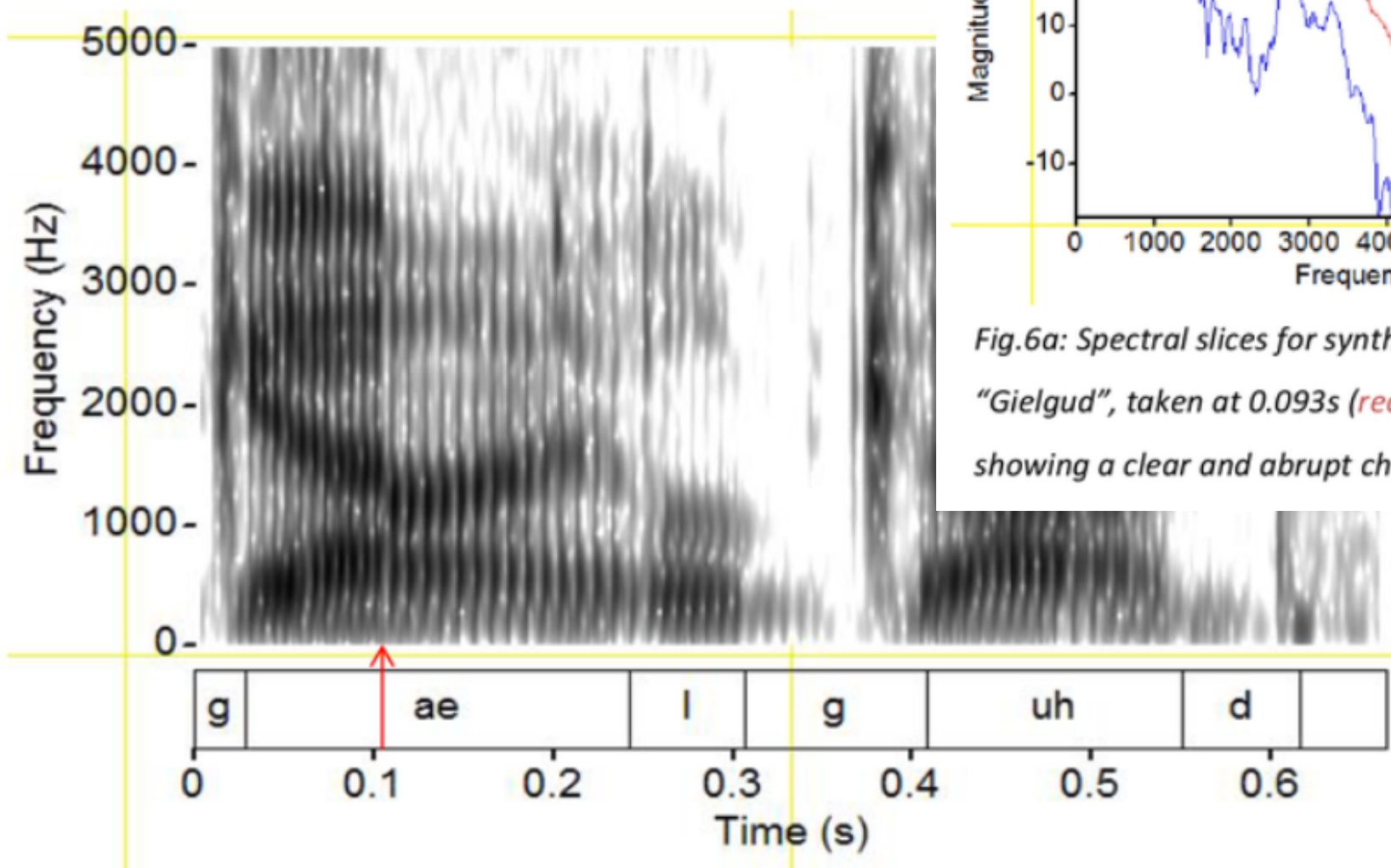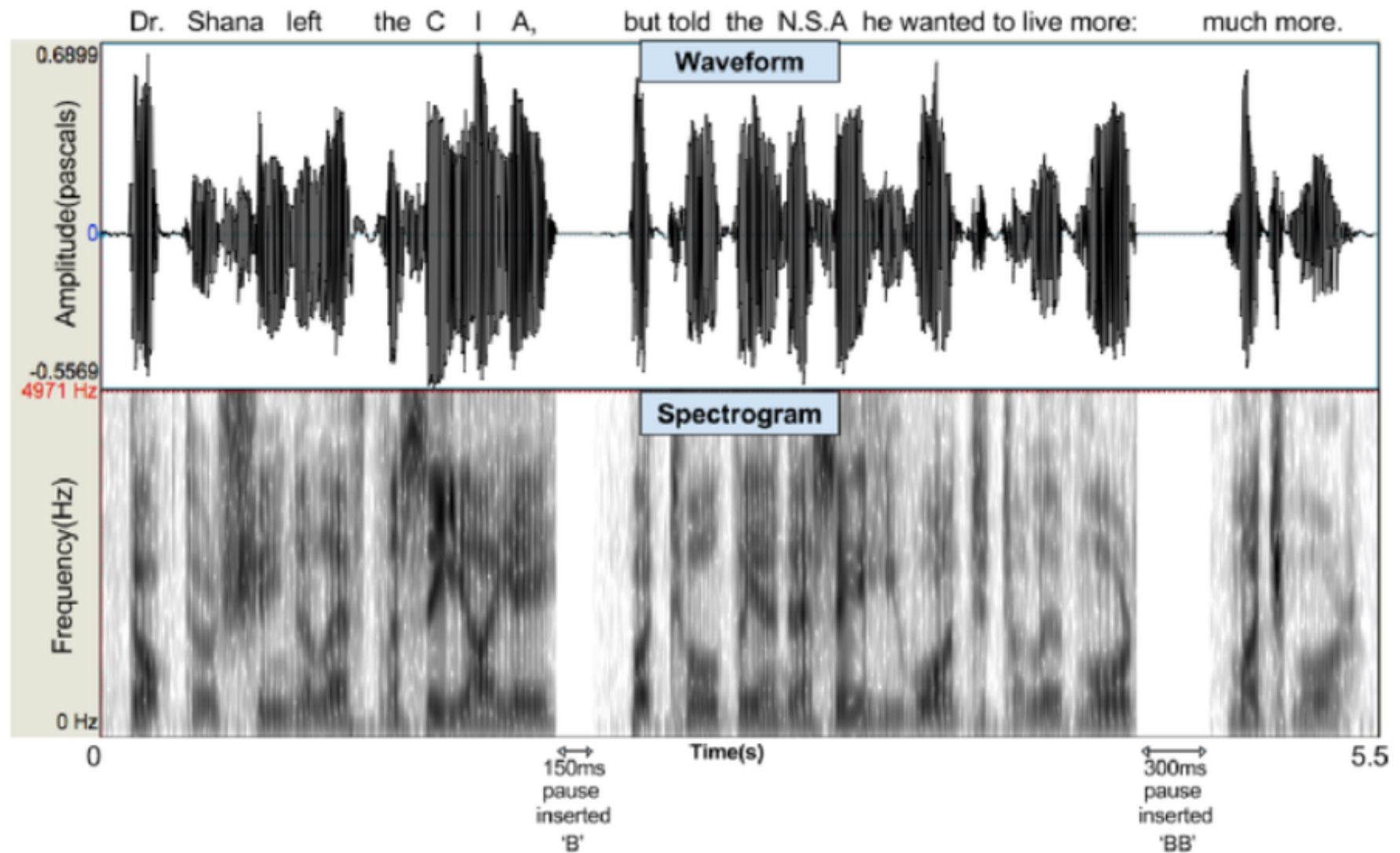"Son of the universe"

Is this a caption?

Good figures

Fig.6a: Spectral slices for synthesized utterance "Gielgud", taken at 0.093s (red) and 0.122s (blue), showing a clear and abrupt change in spectral shape.

Fig.6b: Spectrogram for "Gielgud", extracted from synthesized utterance henry5, showing a clear and abrupt change in sound quality at 1.05s (indicated by arrow).
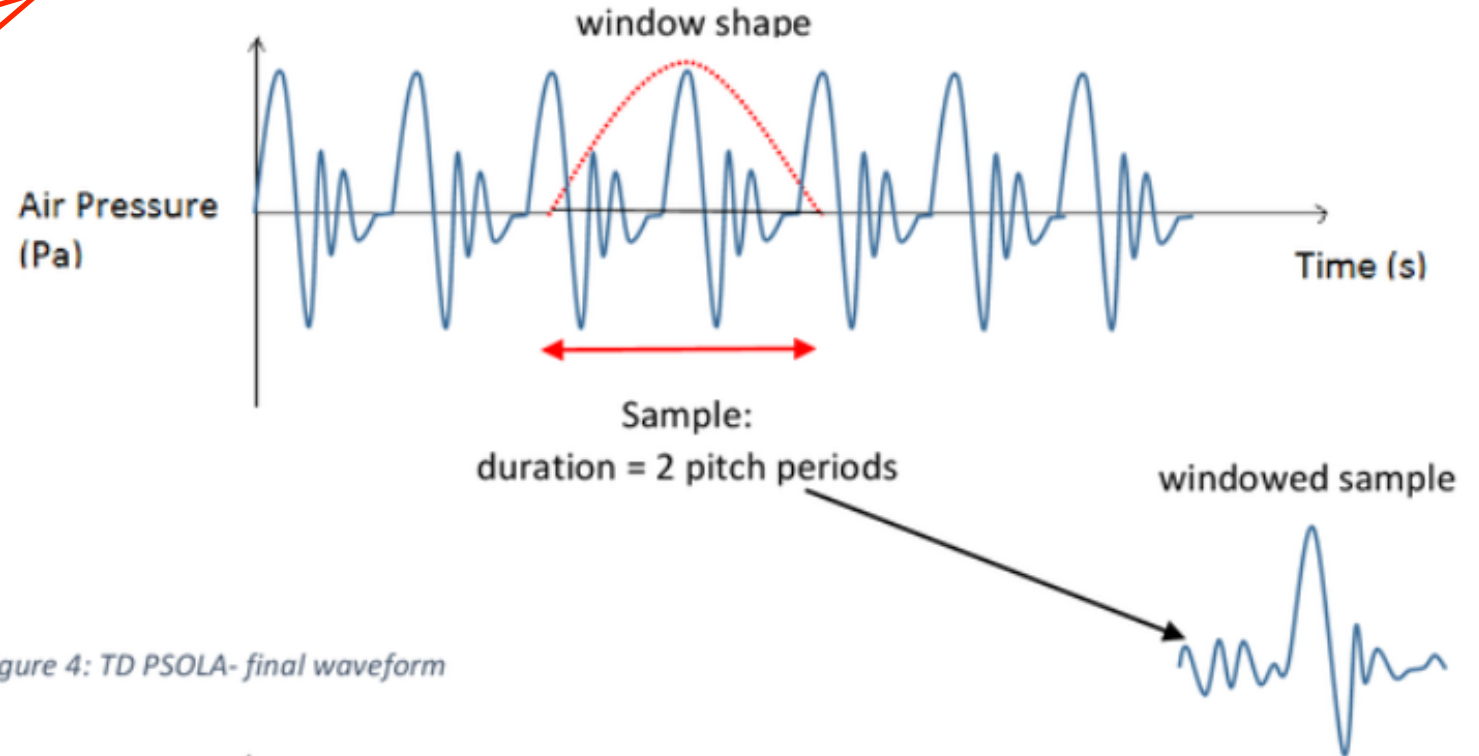
Good figures

Fig. 1.2: Test2 sentence with 150ms 'B' pause after "," and 300ms 'BB' pause after ":"

Good figures

Figure 3: TD PSOLA- selecting a sample to adjust and add

window shape

Air Pressure (Pa)

Time (s)

Sample:
duration = 2 pitch periods

windowed sample

captions could be a bit better though

Figure 4: TD PSOLA- final waveform

Air Pressure (Pa)

Time (s)

46
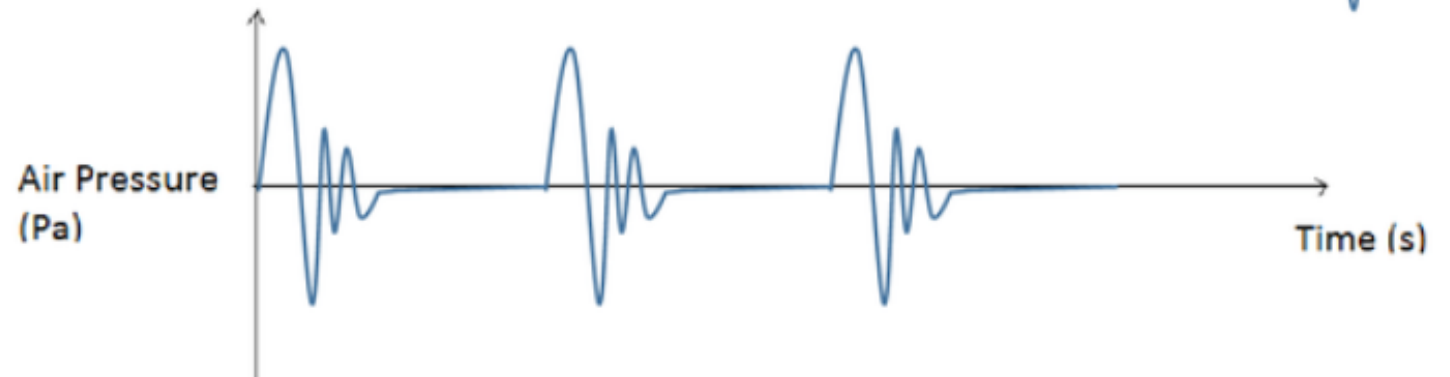
Examples that could be improved

1. It's no use (pos nn) to ask to use (pos vb) the telephone
2. Do you live (pos vbp) near a zoo with live (pos jj) animals?
3. I prefer bass (pos nn) fishing to playing the bass (pos nn) guitar.
4. The wind (pos nn) is blowing hard.
5. I have to wind (pos vb) my clock.
6. The singer made a low bow (pos nn) to the audience.
7. Maria placed a red bow (pos nn) on the birthday gift
8. All the students are present (pos jj) today.
9. The boss will present (pos vb) the award
10. Please close (pos vb) the door.
11. The boy sat close (pos rb) to his uncle
12. The rope was wound (pos nn) around
13. The soldier received a wound (pos nn) in the
14. I don't know if I will live (pos vb) or die.
15. Last night I saw the band play live (pos vbp) in concert.

All of these are POS error examples. Better to provide fewer examples, examine each in more depth and/or provide more errors in in other categories instead.

Good example, but go further…

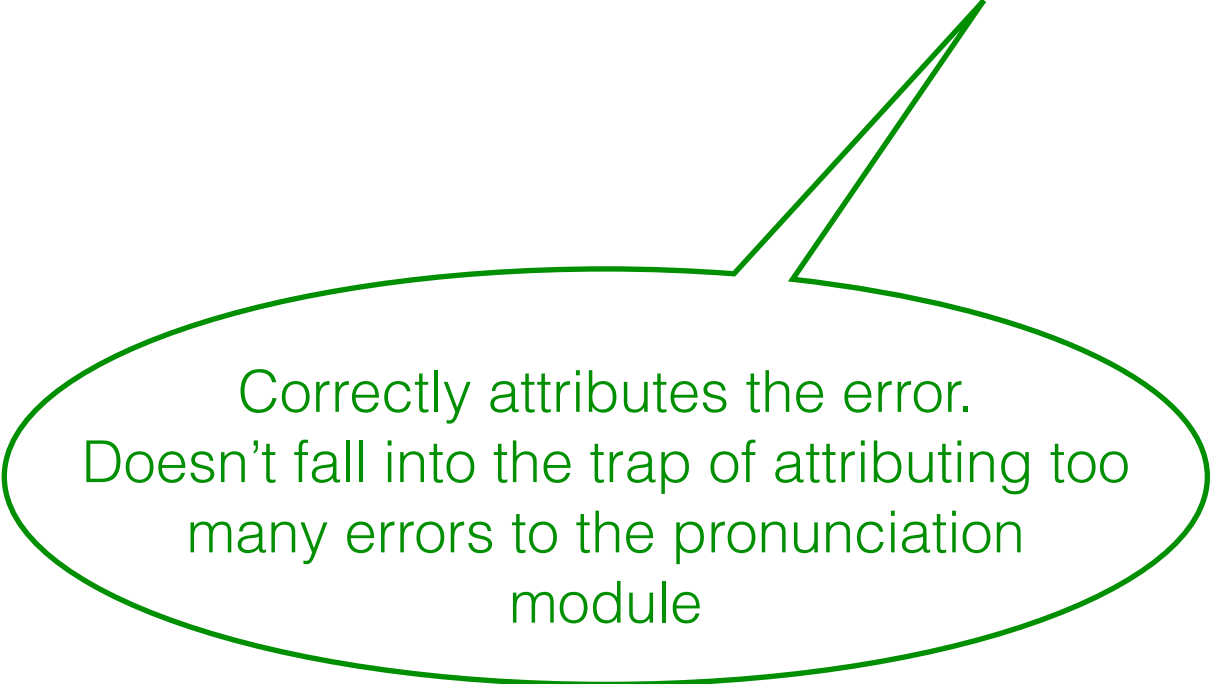**Example 1. 'Dogs stretch their legs.' (http://bit.ly/Dogs_legs)**

In the utterance above in the 'PostLex' module *Festival* did not apply allophonic assimilation, i.e. devoicing of a fricative (v) was not realised, yielding [doogz strech] instead of more natural [doogs strech]. Analogous examples that returned this kind of mistake are 'I have to go.' which was returned as [hev t@] instead of [hef t@] and 'It has to change.' realised as [haz t@] instead of [has t@]. It should be pointed out, however, the vowel reduction has been in all of the above examples reducing in reduction of a [uu] sound to a schwa [@].

Good description of what goes wrong.

But go further: how might you solve this problem? Which part of Festival would you modify, and how?

Well-chosen examples

Festival had trouble disambiguating numbers. Utterances like '£39.99' were parsed as 'thirty-nine-pounds-dot-ninety-nine'. These type of errors cannot be considered pronunciation errors since they happened at the tokenization level.

Correctly attributes the error.
Doesn't fall into the trap of attributing too many errors to the pronunciation module

Well-chosen examples

(1) *youtuber*

(2) *cybersecurity*

Fitspiration returns 1,520,000 Google results

Wording that could be improved

To generate speech, Festival uses Linear Predictive Synthesis, a method which consists in getting diaphones, process and concatenate them to match the phonemes in the Utterance. Linear Predictive Synthesis uses the residual waves of each diaphone to build, through a filter, the correct pitch accent for syllables in the sentence, adjusting their duration and modifying their F0 in the process.

Too many different things packed into two long sentences. Better to unpack it into a few sections / paragraphs and use simpler sentences.
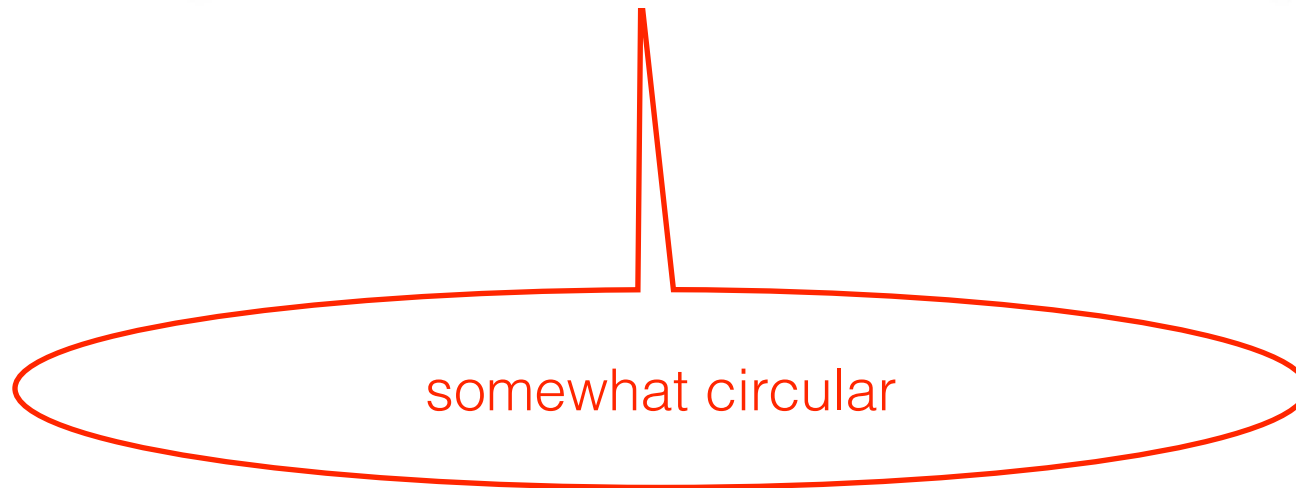
Wording that could be improved

A phrase-break should either be after a word or not.

The results were successful

Vague or imprecise wording

Tokenization

This is the first step of processing and it is to covert the string into tokens. Token is a list of trees and it can be regarded as a list of tokens which are read from string. Basically every

somewhat circular

Describing a process, without demonstrating understanding

- Token_POS

This command does not create a new relation, but updates the one created at the previous stage by adding identification information to it. For example, the command decides whether a numerical token is a year, a date, etc.

- Token

This command converts each token into a word (or some words) and creates a new relation *Word* to store them. So, at this stage all the abbreviations and digits are changed into corresponding words.

- POS

This command uploads the relation *Word* by adding speech for each word.

- Phrasify

This command creates a new relation *Phrase*, which stores the information about the prosodic phrases contained in the utterance. Each phrase in the relation is the root for the tree whose leaves are the words of the phrase.

Just a list of what happens, without say **how** it is done

How to do better next time

- **Compare** your first assignment with these feedback slides

  - the markers cannot annotate every individual error or potential improvement: so now you could <u>add your own feedback</u> (or swap with a classmate)

- **Think** about how to go <u>above and beyond</u> the instructions for the assignment

  - interesting experiments of your own invention (always driven by a clear hypothesis or research question)

  - novel analyses of the data / models / results, etc

- **Draft** your second assignment well before the deadline, then <u>mark it yourself</u>

  - what *mark* would you give it?

  - what *comments* would you write on it?