For Meeting on 10.03.:

* I have requested access to the Verbmobil corpora and now have it
* I have identified a few papers that I want to read
* I have done the reading and exercise 1 for the proposal
* **Question**: The Verbmobil corpus for german exists in two versions – what is the difference and which version should I use. Version 2 looks more similar to the other?
* I had a quick look at the annotation style books – but they are about 80 pages (a lot of syntax trees!) – so I plan to go over them and compare the differences to the Switchboard corpus
* **Question**: How should I work with the data? – the data files in the corpora directory are zipped. Can I copy them to my personal directoy on DICE or even to my personal machine?

For Meeting on 17.03:

* Switchboard corpus: the trees are in /group/corporapublic/switchboard/nxt/xml/syntax?
* Speech data is in ??
* Switchboard parse trees are saved in xml files
* In the tran code, there is also code for Penntree Bank files

For Meeting on 24.03:

* I have read a few papers on prosody and selected new ones to read
* I now have a project in my Discourse Analysis course, which will be about the topic “how do disfluencies affect prosody” in German vs. English in the Verbmobil corpus
* The speech data is now available on DICE
* Currently I try to get a better overview over the corpora and all of its annotations
* The treebank exists in 3 different formats and it seems that I need information from all of them: penn tree bank files for the trees, xml files for the recording ids, and their annotation file to map from sentence IDs to recording ids
* I need time stamp information but these are not included in the treebank files. They are in the speech corpus. However, I have the following problem:
* Verbmobil exists in 2 versions. Version 1 is very heavily annotated – there I have word and phone segmentations. For Version 2 I can simply not find any word segmentation information. How sure are we that this exists? Would be using a forced aligner be an option?
* I tried to run the parser and extract speech features but I encountered several problems: Pytorch related and when calculating the pauses.

For Meeting on 07.04:

* The minority of sentences have word alignments, but phone alignment – it’s possible to derive word alignment from the phonetic transcriptions
* Stats for German data

About 25k sentences

VM2: 14520 turns that have a tree, speech data, and annotations

VM2 WOR: 23

VM2 MAU: 13686

Total: 14520

Total WOR: 23

Total MAU: 13686

total alignment info: 13709

no alignment info: 811

no annotation: 3795 trees with paths to recordings, but those are not there!

* Stats for English data:

Meeting on 02.06.21:

* I did the Ethics training that SLP students have to do before they start they dis
* I got familiar with the MLP cluster and can now run the parser with and without speech features with the new sample data using GPU
* Timeline:

First 2 weeks in June: get a first working pipeline and initial results: run the parser on switchboard and verbmobil and get results

Last 2 weeks in June: do experiments that I wrote about in my proposal: about different corpus sizes, sentence lengths, languages, forced alignments, …

First 2 weeks in July: start draft writing about these results and start further analysis of the results

Last 2 weeks in July: concentrate on further analysis/ follow up questions based on the results and do write-up in parallel

First 2 weeks in August: concentrate on writing.

* - What needs to happen to get a first working pipeline of your system and initial results? Plan to complete that first, even if some of the components are much simpler than intended for the final version.  
  Replicate Elisabeths results on Switchboard -> get familiar with the switchboard data

->Do the feature extraction

Run the Parser on Verbmobil -> get timestamp data using a Forced Aligner

* - What (if anything) do you need to do after that before you can start running your first main experiments?  
    
  - Which sets of experiments make natural "packages" to be completed together? Which packages are the most important to tester hypotheses (do those first) and which are nice to have but not critical (do after, depending on time).

Questions:

* Do we need to fill in an ethics approval form?
* On the MLP cluster, its not possible to access data that is located on DICE machines or the AFS filesystem. Just to be sure – we all have to copy the switchboard data on the mlp cluster?
* Disk space problem. When extracting acoustic features with kaldi for all the switch board data, the mfcc features alone take 11 GB of space. How to deal with this?

There is no disk quota on the cluster, but this would require copying all the raw switchboard data onto the cluster (probably around 15GB)

* Any weeks on vacation? Any preferred weeks for reading/submitting drafts?

For Meeting on 09.06.21:

Notes on training on SWBD:

* It takes just over 10 hours to train the model without speech features
* F1=90.92 on dev set, F1= 90.78 on test set (slightly better than reported in paper)

Notes on training on VMENG:

* Currently, I use an arbitrary train,dev,test split which is not random – how is it in swbd?
* It takes 2,5 hours to train without speech features
* Fscore: 91.70 on test set; 93 on dev set (slightly better than on SWBD) [these are the results using the wrong trees!]
* Using the right trees: F1 = 94.75 on dev set, F1 = 92.05 on test set
* This is slightly better that the results in the paper – why? Maybe the proportion of one-word sentences is higher? (like “okay”, “great”, “bye” – from a first sight)
* At the moment, I only deal with sentence-based parsing and ignore turn based parsing.
* Training with speech features on VM sample data (around 100 utterances): Fscore of 38.36 on dev-set after 50 epochs) – comparable with SWBD sample data (F1: 28)

Notes on training on VMGER:

* Training with speech features on VM sample data: F1 = 42 (works and is comparable)
* F1 = 94.34 on dev set, F1=94.16 on test set (topological field model annotations are very regular, this could help with parsing!)

Notes on alignment:

* (Running the Aligner requires the corpus to be of a specific shape. I don’t have space for this, so currently I run it locally on my machine.)
* Multiple trees/transcriptions are linked to the same wav file. This could be a problem for forced alignment.
* Currently, I use a quick and dirty solution for alignment:
  + I use a pretrained acoustic model and an existing pronunciation dictionary. Training an acoustic model on the data and using pronunciation dictionaries based on the VM data could improve things
  + Transcriptions are based on the trees. But the trees don’t include every detail in the speech (e.g. filled pauses). I think I should use the transcriptions that are given in the original VM corpus distribution
* Aligning 150 files takes about 1 minute, aligning 25k sentences takes a couple of hours
* Around 200 turns cannot be aligned, I can’t find a solution for this

Notes on extracting kaldi features:

* Scripts are written for extracting features from sph files. Converting wav to sph takes a lot of space (I would guess about 40 GB).

For Meeting on 16.06:

* I have 2 major updates: 1) I now have a working solution for running the parser with speech features on all of the ENG and GER VM data and 2) I extended my code to also work for turn-based parsing
* VMENG:

|  |  |  |
| --- | --- | --- |
|  | SU-based | Turn-based |
| Text only (test) | 91.78 | 90.07 |
| Text+prosody (test) | ~~92.30~~ | ~~92.15~~ |
| Text only (dev) | 94.45 | 91.75 |
| Text+prosody (dev) | ~~94.71~~  94.64 | ~~94.28~~  94.16 |

* VMGER:

|  |  |  |
| --- | --- | --- |
|  | SU-based | Turn-based |
| Text only (test) | 94.14 | 93.16 |
| Text+prosody (test) | ~~94.39~~ | ~~94.55~~ |
| Text only (dev) | 94.15 | 92.85 |
| Text+prosody (dev) | ~~94.29~~  94.27 | ~~94.08~~  94.07 |

* Important aspects are the same as in SWBD setting
* Open questions:
  + Discrepancy between dev (around 94) and test scores (around 92) in VMENG – on the one hand, makes sense because of tuning – but in German, no discrepancy 🡪 make new, updated corpus statistics (done)
    - Discrepancy in VMENG makes sense: Avg. sentence length is larger (12) in test set than in dev set (8) (and train set (8))
    - No discrepancy in VMGER also makes sense: avg. sentence length is about 8 everywhere
  + Performance drop is larger (around -4) in SWBD compared to VM (around 1-2) in turn-based parsing – why?
    - 🡪 corpus statistics about turns (#SUs should be comparable, but what about turn length and complexity, disfluencies (!))?
    - One possible reason could be the small domain
  + Why is the performance on German data so high?
    - (Topological field annotations) – actually, it should be about 94 also in ENG.
  + Less training data (22k sentences vs. 90k; 12k turns vs. 49k) but no drop in F1?
    - Sanity check: less data in SWBD causes a drop: F1=87
    - Domain is smaller (2 possible scheduling scenarios vs. telephone speech on 70 topics) 🡪 smaller vocabulary and repeating phrases
      * What if I sample only SUs of 2 scenarios? 🡪 Not possible to match the size of VM
      * Sample from 10 most common topics?
    - ~~Less speakers (?)~~
    - Less complex tree annotations ?
* Next steps:
  + 1) Train on SWBD[reduced and comparable in size] + original alignments
  + 2) Train on SWBD[reduced and comparable in size] + MFA alignments
  + 3) Compare performance across different sentence lengths and impact of individual prosodic features as done in the paper
* Questions:
  + Arbitrary test-dev split – maybe random sample?
  + 10 models with different random seeds?
  + Generating turn-based input and the RAM issue

For meeting on 23.06:

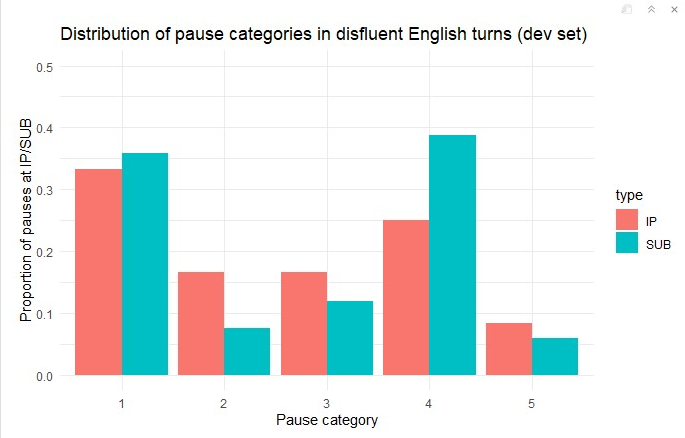
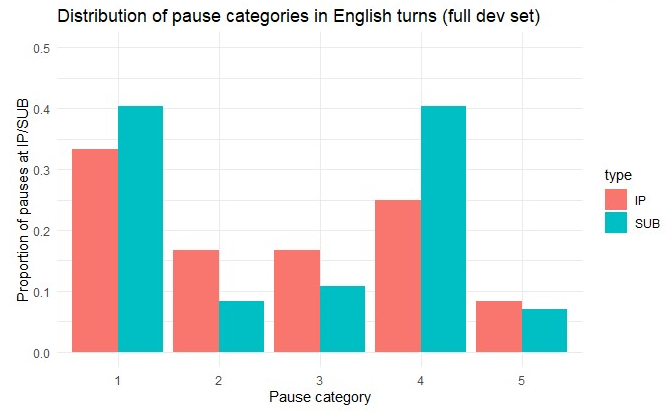
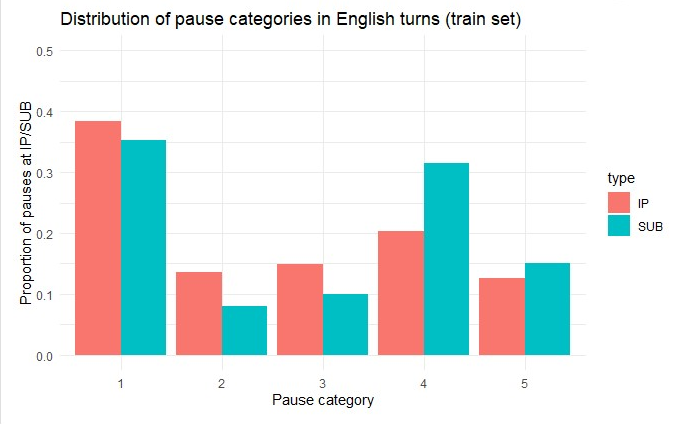
* Statistics about the turns (filled pauses don’t count as disfl.):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Split |  | SWBD | VMENG | VMGER |
| Train | Mean turn length | 13.34 | 15.21 | 13.75 |
| Mean SUs | 1.80 | 1.89 | 1.72 |
| Mean # disfl./turn | 0.43 | 0.37 | 0.24 |
| Percentage disfl. turns | 0.24 | 0.23 | 0.17 |
| Dev | Mean turn length | 16.29 | 19.23 | 14.09 |
| Mean SUs | 1.86 | 2.21 | 1.78 |
| Mean # disfl./turn | 0.70 | 0.23 | 0.23 |
| Percentage disfl. turns | 0.30 | 0.18 | 0.16 |

* VM contains less disfluent turns in the dev set – explains smaller performance drop in turn based parsing
* NOTE: these stats are outdated, see source files instead
* Looking at individual prosodic features:

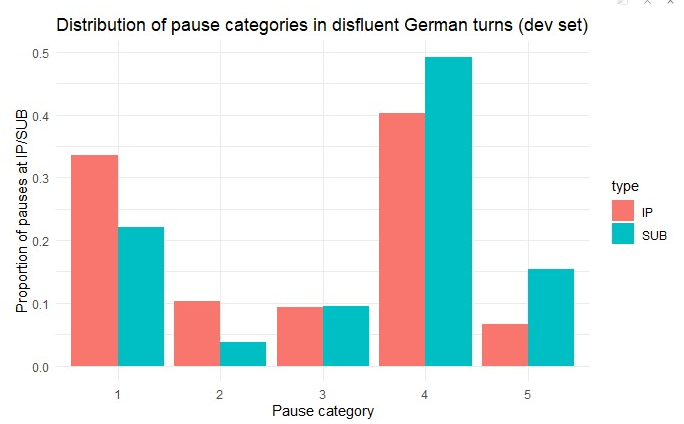
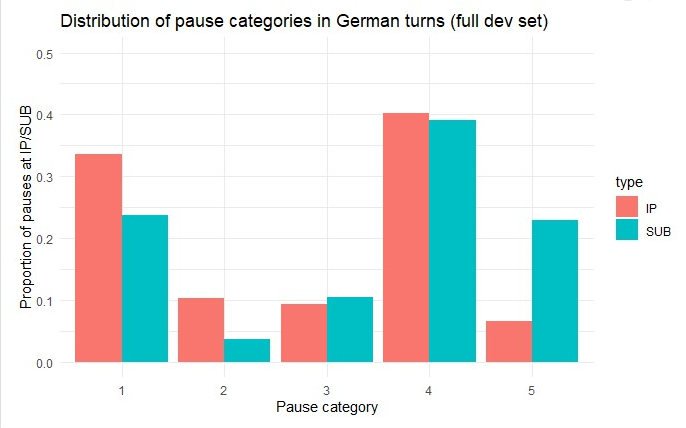
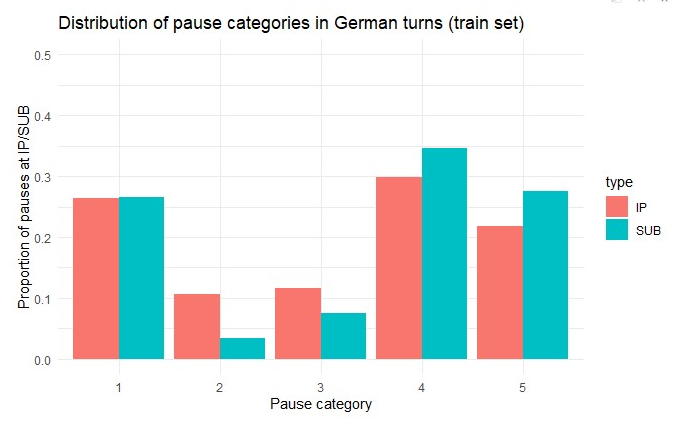
VMENG:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SU-based | Turn-based |  | Fluent | Disfluent |
| Text only | 94.45 | 91.75 |  | ~~92.32~~  92.29 | ~~90.21~~  89.65 |
| Text + prosody (all) | ~~94.71~~  94.64 | ~~94.35~~  94.16 | ~~∆+2.6~~  ∆+2.41 | ~~95.05~~  ~~94.83 (∆+2.51)~~  94.89 (∆+2.6) | ~~92.44~~  ~~92.30 (∆+2.09)~~  91.26(∆+1.61) |
| Pitch only | 94.59 | 93.92 | ∆+2.17 | ~~94.71 (∆+2.39)~~  94.74(∆+2.45) | ~~91.76 (∆+1.55)~~  90.72(∆+1.07) |
| Fbank only | 94.72 | 94.02 | ∆+2.27 | ~~94.87 (∆+2.55)~~  94.86(∆+2.57) | ~~91.67 (∆+1.46)~~  90.71(∆+1.06) |
| Duration only | 94.78 | 91.85 | ∆+0.10 | ~~92.40 (∆+0.08)~~  92.30 (∆+0.01) | ~~90.32 (∆+0.11)~~  90.06 (∆+0.41) |
| Pause only | ~~94.62~~  94.75 | ~~94.12~~  92.07 | ~~∆+2.37~~  ∆+0.32 | ~~94.76~~  ~~92.71 (∆+0.39)~~  92.61(∆+0.32) | ~~92.36~~  ~~90.31 (∆+0.10)~~  89.94(∆+0.29) |
| Pitch,fbank,pause | 94.53 | 94.20 | ∆+2.45 |  |  |
| Pitch, fbank | 94.77 | 94.13 |  |  |  |
| Fbank, dur, pause | 94.72 | 94.20 |  |  |  |
| Pitch,fbank,dur | 94.52 | 93.97 |  |  |  |

* This is the same as in SWBD:
  + Pitch and Intensity yield high performance (help distinguishing disfl. From SU boundaries)
  + Duration cannot outperform text only model (probably function word issue) (?)
  + Pattern in duration only is the same: no difference on fluent sentences, slight improvement on disfluent ones.
  + Pause improves only slightly, however, a bit more when it comes to disfluent sentences (compared to SWBD)
* (This is different)
  + Pause improves performance! – This makes sense because:
    - Pauses are a cue to sentence boundaries, but also to interruption points
    - Pauses can signal both, can be confusing for the model
    - However, VMENG doesn’t contain many disfluencies, so long pause 🡪 sentence boundary is a good rule
    - 51.32% of SU boundaries have a long pause (> 0.05) <-> 50% of disfluencies have a long pause (pause at interruption point)
    - 
    - IPs are not as clearly “marked” by certain pause categories as in German, however 2 and 3 seem to signal IPs – 1 and 4 more SUBs
    - 1 and 4 are the most dominant categories for both IPs and SUBs (which is the same for German)
    - 
    - When fluent sentences are present, the whole distribution doesn’t change much (as opposed to German)
    - 
    - Apparently, the model learns that 2 and 3 indicate IPs and 4 indicates SUB
    - Percentage of SU boundaries with long pauses: 56.43506870524482 %
    - Percentage of IPs with long pauses: 47.85207321628689 %
    - 🡪difference of 8.59 %
* This is strange:
  + Best performance in SU parsing with duration only -> why?
  + Pitch alone isn’t that helpful in SU parsing
  + Best performance in TB parsing, when duration is not used (however, diff is very small)
* In contrast to SWBD, the prosodic features help more with fluent sentences than with disfluent ones (However, there are only a few disfluent ones anyway)

VMGER:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SU-based | Turn-based |  | Fluent | Disfluent |
| Text only | 94.15 | 92.85 |  | ~~93.73~~  93.62 | ~~91.04~~  90.69 |
| Text + prosody (all) | ~~94.29~~  94.27 | ~~93.81~~  94.07 | ~~∆0.96~~  ∆+1.22 | ~~94.71~~  ~~95.11 (∆+1.38)~~  95.07 (∆+1.45) | ~~91.95~~  ~~91.92 (∆+0.88~~)  91.22(∆+0.53) |
| Pitch only | 94.39 | 93.53 | ∆+0.68 | ~~94.60 (∆+0.87)~~  94.56 (∆+0.94) | ~~91.32 (∆+0.28)~~  90.65 (∆-0.04) |
| Fbank only | 94.25 | 93.66 | ∆+0.81 | ~~94.61 (∆+0.88)~~  94.53(∆+0.91) | ~~91.70 (∆+0.66)~~  91.21(∆+0.52) |
| Duration only | 94.36 | 93.29 | ∆+0.44 | ~~94.24 (∆+0.51)~~  94.10(∆+0.48) | ~~91.32 (∆+0.28)~~  91.01(∆+0.32) |
| Pause only | ~~94.35~~  94.18 | ~~93.61~~  92.81 | ~~∆0.76~~  ∆-0.04 | ~~94.55~~  ~~93.83 (∆+0.10)~~  93.62 (∆+0) | ~~91.68~~  ~~90.69 (∆-0.35)~~  90.51 (∆+0.82) |
| Text + prosody (all, new duration) |  | 93.86 |  |  |  |
| New duration only |  | 93.11 |  |  |  |
| Pitch,fbank,pause | 94.35 | ~~93.68~~  93.76 |  |  |  |
| Pitch, fbank | 94.31 | 93.56 |  |  |  |
| Fbank, dur, pause | 94.36 | 94.07 |  |  |  |
| Pitch,fbank,dur | 94.40 | 93.95 |  |  |  |

* This is the same as in VMENG:
  + Pitch and Intensity only improve upon the text-only model
* Duration seems to be somewhat helpful (less than other features, but more than in VMENG)
  + Why?
* Similar as in the SWBD setting, pause is useless – BUT different: helps only with disfluent sentences
  + Is the reason the same?:
    - Argument in the paper: long pauses can signal SU boundaries OR interruption points 🡪 long pauses occur in 30% of SU boundaries AND interruption points 🡪 model confuses them
    - Here:
      * 72.61% (!) of SU boundaries have a long pause (> 0.05) <-> 56.07% of disfluencies have a long pause (in the full dev set)
    - Ja… - constructions at the end and at the beginning of sentences explain the pauses
    - This plot sheds light on what is happening:
    - 
    - Longer pauses (4 and 5) signal SUB vs. shorter pauses (1,2) signal IP
    - 
    - Why is pause not helpful on the full dev set?: 40% of SUB have a pause (4) and 40% of IP have a pause (4) when fluent sentences are present 🡪 model confuses those and pause isn’t so helpful anymore (?)
    - 
    - 2 and 3 could be an indicator for IP; 4 and 5 for SUB
    - Percentage of SU boundaries with long pauses: 69.81375820863387 %
    - Percentage of IPs with long pauses: 63.07765602667937 %
    - 🡪 difference of 6.81 %
    - What the model learns seems to be similar to VMENG, however in this particular dev set, disfluent sentences are dominantly marked by pause 4 (in over 50% of SUBs)
* This is different:
  + Pitch doesn’t help with parsing disfluent sentences – why?
  + Duration is somewhat helpful

Results on SWBD:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SU-based | Turn-based |  | Fluent | Disfluent |
| Text only |  | 86.09 |  | 89.89 | 84.25 |
| Text + prosody (all) |  | 90.90 | ∆+4.81 | 93.63 (∆+3.74) | 89.58(∆+5.33) |
| Pitch only |  | 90.71 | ∆+4.62 | 93.46 (∆+3.57) | 89.37 (∆+5.12) |
| Fbank only |  | 90.29 | ∆+4.2 | 93.30 (∆+3.41) | 88.83 (∆+4.58) |
| Duration only |  | 86.24 | ∆+0.15 | 89.94 (∆+0.05) | 84.44 (∆+0.19) |
| Pause only |  | 86.21 | ∆+0.12 | 90.09 (∆+0.2) | 84.32 (∆+0.07) |

Regarding pitch:

* the pitch contour before an interruption point is generally “flat or slowly falling” (Shriberg, 2001, 161), while SU boundaries are characterized by a boundary tone, generally corresponding to a fall or rise.
* the mean warped NCCF value for pre-interruption point words is significantly higher than the value for SU-final words
* the log-pitch with POV-weighted mean subtraction (~F0) is significantly lower at interruption points than at SU boundaries

For Meeting on 30.06:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Findings regarding** | **SWBD** | **VMENG** | **VMGER** | **Notes** |
| Pauses |  | 56 % of SUBs and 48 % of IPs have long pauses (train set, see above for further analysis) | 70 % of SUBs and 63 % of IPs have long pauses (train set, see above for further analysis) | Both are characterized by long pauses. However, the numbers are higher in VM and SUBs > IPs.  Pause is slightly more helpful for parsing in VM (see analysis above);  Difference in German dev set is larger. Hypothesis: model predicts more sentence breaks than in English, therefore, pause is not as useful overall |
| 33 % of SUBs and 37 %  of IPs have long pauses | 51 % of SUBs and 50 % of IPs have long pauses (dev set, see above for further analysis) | 73 % of SUBs and 56 % of IPs have long pauses (dev set, see above for further analysis) |
| Duration  (1. Normalized by mean) |  | words preceding SU boundaries are lengthened on average (normalized duration: 1.20), and those preceding interruption points  even more so (normalized duration: 1.5)  Note: SUBs are only sent-medial in this calculation; train set | words preceding SU boundaries are lengthened on average (normalized duration: 1.14), and those preceding interruption points  even more so (normalized duration: 1.53)  Note: SUBs are only sent-final in this calculation; train set | Pattern is very similar across datasets and languages. Duration is similarly helpful (only a bit). |
| words preceding SU boundaries are lengthened on average (normalized duration: 1.18), and those preceding interruption points  even more so (normalized duration: 1.41)  diff = 0.23 | words preceding SU boundaries are lengthened on average (normalized duration: 1.32), and those preceding interruption points  even more so (normalized duration: 1.66)  Note: SUBs are only sent-final in this calculation; dev set  Diff = 0.34 | words preceding SU boundaries are lengthened on average (normalized duration: 1.19), and those preceding interruption points  even more so (normalized duration: 1.47)  Note: SUBs are only sent-final in this calculation; dev set  Diff = 0.28 |
| Duration (2. normalized by max) |  | SU-final words have  mean value of 0.84, while words directly before  the interruption point have a mean of 0.53  (train set) | SU-final words have  mean value of 0.83, while words directly before  the interruption point have a mean of 0.52  (train set) | Pattern is very similar across datasets and languages. Duration is similarly helpful (only a bit). |
| SU-final words have  mean value of 0.86, while words directly before  the interruption point have a mean of 0.50 | SU-final words have  mean value of 0.86, while words directly before  the interruption point have a mean of 0.56  (dev set) | SU-final words have  mean value of 0.85, while words directly before  the interruption point have a mean of 0.49  (dev set) |
|  | 20.6 percent of train set SUs end in a function word, while the word before an IP is a function word 54.5 percent of the time | 17.3 percent of train set SUs end in a function word, while the word before an IP is a function word 70.2 percent of the time |  |
| 21.9 percent of development set SUs end in a function word, while the word before an IP is a function word 51.6 percent of the time | 22.5 percent of development set SUs end in a function word, while the word before an IP is a function word 53.6 percent of the time | 18.4 percent of dev set SUs end in a function word, while the word before an IP is a function word 79.4 percent of the time |
| Pitch (NCCF) |  | the mean warped NCCF value for pre-interruption point words is significantly **higher** (-0.47868) than the value for SU-final words (-0.432547), and also higher than the overall average value (-0.42311) across the  train set. | the mean warped NCCF value for pre-interruption point words (-0.55 in train) is significantly **higher** than the value for SU-final words (-0.47 in train), and also higher than the overall average value (-0.46 in dev) across the  train set. |  |
| the mean warped NCCF value for pre-interruption point words (-0.45 in dev, -0.44 in train) is significantly higher than the value for SU-final words (-0.42 in dev, -0.47 in train), though somewhat  lower than the overall average value (-0.39 in dev) across the  development set. | the mean warped NCCF value for pre-interruption point words is significantly **higher** (-0.53283092) than the value for SU-final words (-0.43039488), and also  higher than the overall average value (-0.47054214) across the  development set. | the mean warped NCCF value for pre-interruption point words (-0.54 in dev) is significantly **higher** than the value for SU-final words (-0.49 in dev),  and also higher than the overall average value (-0.48 in dev) across the  development set. |
| Pitch (log pitch) |  | No significant difference. | the log-pitch with POV-weighted mean subtraction (~F0) is significantly **higher** at interruption points (-0.02558534) than at SU boundaries -0.04654488)  (train set) |  |
| the log-pitch with POV-weighted mean subtraction (~F0) (-0.0003 in dev, -0.04252539 in train) is significantly lower (**higher**) at interruption points than at SU boundaries (-0.03943205 in dev, -0.04413732 in train)   * this is confusing, lower vs. higher | the log-pitch with POV-weighted mean subtraction (~F0) is significantly **higher** at interruption points (-0.03160787) than at SU boundaries (-0.05917872)  (dev set) | p = 0.003  (not significant in dev) |
| Pitch (derivative of log pitch) |  | Significant in train set  At IP (-0.00695) lower than  At SUB (-0.0041) | Significant in train set  At IP (-0.00396139) lower than  At SUB (0.00222442) |  |
| No differences. | No significant difference. | No significant difference (dev set) |
| Intensity (overall energy) |  | SU-final words (0.84, train) **higher** than in all other words (0.81, train) (significant); words before IP (0.83 significant as well) | SU-final words (0.8141928, train) **higher** than in all other words (0.79165782, train) (significant); words before IP (0.80801121, significant as well) |  |
| No differences ?  significant difference between Pre-SUB words and Pre-IP words;  SU-final words (0.73, dev) **higher** than in all other words (0.71, dev) – significant;  Pre-IP words (0.75, dev) **higher** than in all other words (0.71, dev) – significant;  🡪why not mentioned in the paper? | No significant difference between Pre-SUB words and Pre-IP words;  SU-final words (0.84, dev) **higher** than in all other words (0.83, dev) – significant;  Pre-IP words (0.84, dev) **higher** than in all other words (0.83, dev) – significant; | No significant difference between Pre-SUB words and Pre-IP words;  SU-final words (0.84, dev) **higher** than in all other words (0.83, train) – significant;  Pre-IP words (0.84, dev) **higher** than in all other words (0.83, dev) – significant; |
| Intensity (energy in lower half of freq.) |  | SU-final words have a significantly **lower** mean value for lower-frequency intensity (0.982) than all other words (0.989)  (p = 0, not sure if significant), while words before the interruption  point (1.07) | SU-final words have a significantly **higher** mean value for lower-frequency intensity (1.09075024) than all other words (1.04936419)  (p < 0:001), while words before the interruption  point (1.04704509, train, significant) |  |
| SU-final words have a significantly higher mean value for lower-frequency intensity (1.41852566 in dev, 1.45 in train) than all other words (1.34366463 in dev, 1.37864307 in train)  (p < 0:001), while words before the interruption  point do not  (train or dev set?, I assume dev) | SU-final words have a significantly **lower** mean value for lower-frequency intensity (0.78526733) than all other words (0.93392735)  (p < 0:001), while words before the interruption  point do not (not significant) | No significant difference between Pre-SUB words and Pre-IP words;  SU-final words (0.87, dev) **higher** than in all other words (0.82, train) – significant;  Pre-IP words (0.88 dev) **higher** than in all other words (0.82, dev) – significant; |
| Intensity (energy in higher half of freq.) |  | SU-final words (2.13, train) **lower** than in all other words (2.23, train) (significant); words before IP (2.05, significant as well) | SU-final words (0.93897636, train) **lower** than in all other words (1.14250908, train) (significant); words before IP (1.01524441, significant as well) |  |
|  | No differences ?  significant difference between Pre-SUB words (2.44) and Pre-IP words (2.22);  SU-final words (2.44, dev) significantly different than in all other words (2.47, dev) – significant;  Pre-IP words (2.22, dev) **lower** than in all other words (2.47, dev) – significant; | significant difference between Pre-SUB words (2.18) and Pre-IP words (2.02);  SU-final words (2.18, dev) not significantly different than in all other words (2.18, dev) – significant;  Pre-IP words (2.02, dev) **lower** than in all other words (2.18, dev) – significant; | significant difference between Pre-SUB words (0.53) and Pre-IP words (0.79);  SU-final words (0.53, dev) **lower** than in all other words (0.71, dev) – significant;  Pre-IP words (0.79, dev) **lower** than in all other words (0.71, dev) – significant; |  |

Question:

- Train/dev set in analysis?

- re-use material from proposal in thesis?