DOMAIN - Source

- reddit ← typos, Native S. vs NON-N, upvotes
- transcribed press conferences (EU)
- copyright-free poems/lyrics
- Wikipedia ← Scientific Vs ~~Vulgar~~ colloquial, topic moderated
- old movie scripts
- magazines
- children's books → might get nonsense words, Fairytales
- personal emails
- journal entries from 2018 (Shakespeare, King James Bible)
- Court cases
- podcasts (w/ copyright)

CLEAN

- Wiki → formulae + greek letters, Currency symbols, regex ⇒ read out or take out
  - foreign names/h-pronounce w/ feeling
- children's books + fairytales → Names, Countries, nonsense } feed to dict. + pop
- Shakespeare → w/ CMU dict, poetry > plays

- Court cases → [role:] regex/chatgpt
  - redacted X
  - read out utters ⇒ normalize
  - Uhms... etc. ↳ check if in dict + book of law
  - +stutter
  - swear words
  - hateful speech ← vet the cases
  - custody cases → children?

RICHNESS

- # unique diphones (♦), not already found
- normalize by # diphones
- weight by diphone freq
- consider sentence position/word position
- prosody - sentence type
- Seperate lists of diphones for
  - pop diphones as encounter

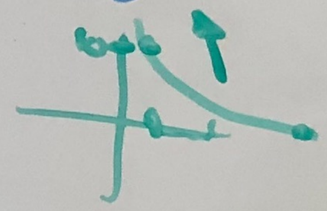
☺

Step 1: Source - Out of copyright books, travel blogs,
Speech zone, open access books, Wikipedia
AI-generated text, Common Crawl, twitter, stackoverflow

Parliament Preceeding

Step 2: Clean → filter away too long sentences → text-normalisation
(readability) & short
programming → formatted text & non-Eng
hastags & emojis
numbers
dates
acronym...
(synthesiser's frontend)
↓ further (selection)
diversify prosody by
sentence types & phrase positions

Step 3:



- diphones (weighted according to English freq. according to Zipf) i.e. how many we have.
- length / # of words
- question / types of prosodic att.
- neighboring diphones
- stress / unstressed
- pitch / f0 low / high

Full coverage dynamic cost

the kat sat
[t → k ae t s ae t]
↓ ↓ ↓
200 = k-ae-stressed
0.1 + 'ə + z

STEP 1.

SOURCE: • TED TALKS + non-comm License + readability + quality, transcription
- big
- narratives

• NOW corpus + free + recent (2010)

• Amazon Q&A data + just cite
- informal
- names ...

STEP 2.

• F.E.

- Out-of-vocab words, names
- Remove too long/short
- clean away markups, timestamps
- Get rid of question marks, etc. (exclaimation)

STEP 3.

• weighted sum of interested features
- phoneme (type) coverage
- stress/unstress
- position in syllable
- position of punctuation
- phonetic context

$\sum w_i a_i$

- compare w/ existing data higher - if it adds variety

Source Text:

Tweets
- Short
- Text Norm is awful
- Prosody ↑

Academic Papers
- Domain Specific
- Prosody ↓
- Hard to Access

"General Science"
Textbooks
- Variety of Domain

EAS System
- Short

Myths

~~Emails~~

Esp-to-speed

Clean Text

Year & Dates
Figure Numbering
Named Entities.
Equations

Individual Blocks
~~Pronunciation~~
Punctuation.
Figure Caption.
Measurement Units

Abbreviation

Score [0 - 1] (Use python G2Pen) The cat sat.
Num. of phones/diphones

[v1, v2] [variety, length] Tally phonemes

SK θ ə k æ t s æ t s K = 10
= 20

$\alpha \theta_1 + \beta \theta_2$

DL: [~~~ ~~~ ...]
100

school website

Government website: gov.uk

public dataset: [ ] speech. LibriTTS

~~[illegible]~~ readme

Step2 clean

HTML

1) Script to find text

2) text normalization: abbrevation? capitalized?

· Acronyms/initialisms

Spelled-out

· Decide line length/segment/Numbe

Step 3

· Use Arctic as baseline

· Count diphone occurrances

· Use that as weights

(less frequent → higher score)

· Give "?"/"!" more points...

· Normalize by total # of diphones

twitter corpus (?)

reddit

Wikipedia

Spotify (podcasts)

interviews

+ data
conversational
modern

- NSW (too many)
- toxic

TTS standard language
- not conversational
- copyright
- topic domain

REMOVE "#"s "@"s Links
- tweets already split data into readable chunks
- rank by length
- expand? "jk" → "just kidding" but different. ES: "jkhjkhj...." "lol" ???
- "AI", "NLP"?
- multilingual tweets?
- typos

---

⑤ - Perplexity - range
(BUT we want to keep rare sounds...)

② - Sentence length
+ character length
+ syllables
→ ③ Prioritise rare diphones

① - how much 'cleaning' has been done

④ - Prioritise questions for prosody range

Source of txt:

	COCA	BNC
eadability	21st but too long 1/2	outdated (20th) 1/2
ize	1 billion 1/2	100 million 1/2
pyright	✓ =	✓ =
VSWs	✓	✓

Clean the text:

- ~~Shorten sentences?~~
- Filter out hard to pronounce words & misspellings.
- Include sentences with uncommon and NSWs.
- Filter according length (10)
  ↳ different domains.
- sentences with: numbers, homographs, etc.
- Diverse punctuation contexts

Richness measure ↑

- frequency of words → proportion to real life
  MLE
- Depth of syntactic tree (variation) (n° nodes + level
- Diphones, capturing contextual effects
- No of breaks
- N° of diphones → $\frac{c(\text{dif} \cdot \text{diph})}{c(\text{diph})}$

different
Machine learning methods
(Random Forest → importance of features)

$f(x) = 6($ $)$

Sources of Text

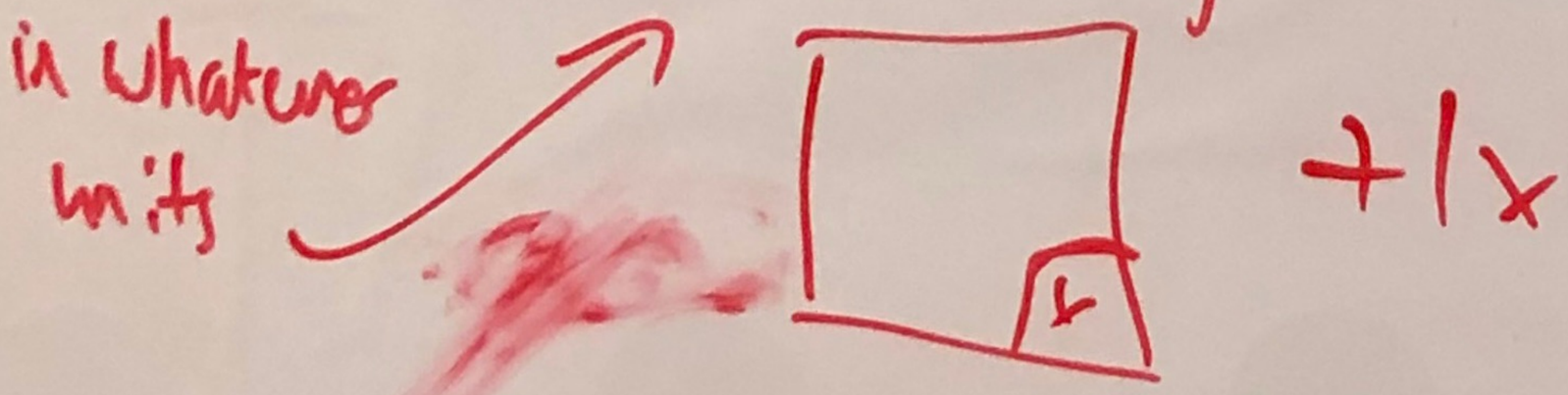
1. Parliament Proceedings. [Europarl] (lots of NSWs, natural ✓,
2. Wikipedia (lots of NSWs, lots of processing,
3. Dictionary (unnatural)
4. Twitter (very NS, ↓natural, ↑preprocessing, ↓
5. Encyclopedia (domain ✓, natural, ↓ NSWs, ↑ validity
6. News (natural speech, domain ✓, NSWs
7. Call centre scripts(
8. Committee meeting minutes

1. scrape, normalise, repeat → script periodic increases
2. ✓
3. tomar preprocessing, HTML / SPEECH
4. Tools available? time cost ~~X~~ 6. ASR? reliability?
7. is it real? ~~X~~

'What is our sentence'
W D+ I E a w @ sent@ns ?
S =
1 0

Max diphones not in DB.
Context
prosodic position
sentence position.
Word position
Stress.
prioritise rarer diphones

Sentence length



Step 1: Sources of Text

- News Articles (Domain: Sports)
  - BBC
  - AP
  - NPR
- TED Talks transcripts

↳ and transcripts (radio news)

Step 2: Clean the Text

- Split punctuation
- remove sentences that are too long, duplicate words
- tokenize
- normalize (eg. Dr., 1-0 score)

Step 3: Richness Measure

this is a laptop

dhis iz @ laeptap

- new diphones normalised by $\frac{\text{no. of diphones}}{\text{no. of diphones}}$
  - reverse Zipf distribution
- diphones in new context
  - position in word
  - position in utterance
  - within question: preceding 3 diphones or whole sentence?

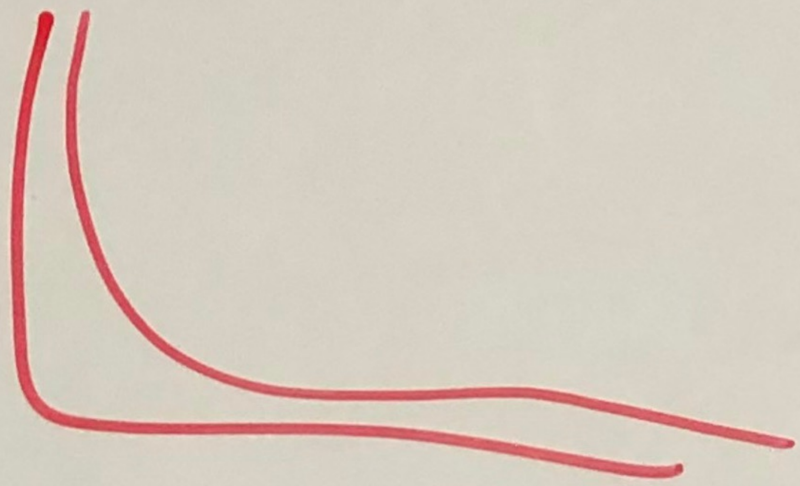
STEP 1

SOURCE: articles, wikipedia, GPT, podcast transcripts, film transcripts?, → theatre

STEP 2

CLEAN: pre-process text (remove NSWs?, segment by puncuation, discard every sentence w/ symbols, remove sentences w/ words not in dictionary. Keep sentences of length 5-15 words,

~~the~~ cat slept on ~~the~~

[the ]

domain specific vocab /embeddings

①

radio shows / podcasts
news transcriptions (copyright?)
british national corpus
CNN, daily mail
[redacted] Scribe (british national corpus)
weatherforecast corpus

---

②

beautiful soup
for not-friendly
formatted text

FLESCH
[redacted] - grab
the readability

use Language model
to calculate perplexity

nltk - tokenize, → NSWs (ver[redacted]
[redacted] split into lower case
sentences

transcribed things were
designed to be spoken,
should be easy to read

chat gpt

---

③

weighted average of context

questions (prosody)
stress
phrase initial/final

run text through G2P
to find all diphones for
maximal coverage

BPE
to find
common
morphemes

prosody
↳ through
punctuation
! ?

prosody + diphone coverage