Introducing Linguistic Analysis

Jeannine Beeken

Workshop DH19
The Case Of Interview Data
9 July 2019, Utrecht

Welcome to
Introducing Linguistic Analysis - Free Tools

SketchEngine at
https://auth.sketchengine.eu/#login

User Manual at
https://www.sketchengine.eu/user-guide/user-manual/

Overview

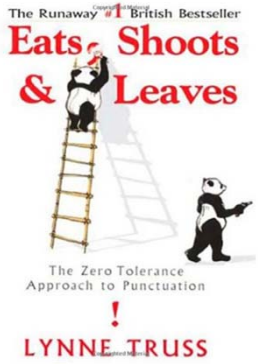
- Human communication: verbal vs. non-verbal
- Verbal communication: oral vs. written ((no) spelling; S2T – T2S)
- Key challenges for linguistic tools analysing human communcation
  - Separate/Disambiguate: tokenizers, sentence splitters, parsers, NER, WSD
  - Cluster/Group: stemmers, lemmatizers, synonyms, MWE (collocations, idioms)
  - Control/Reduce: stopwords, keyword extraction, summarizers
- Tasks

Verbal vs. non-verbal

- Verbal: spoken (sounds, speech) vs. written (text, silent)
- Non-verbal: signs, body language and symbols (silent)
- Examples:
  - Triple/quadruple negation ‘no’ or ‘stop’; quotes
  - Excerpt from a silent film https://www.youtube.com/watch?v=4QrR_1NW_w4 (1.10min)
  - Onomatopoeia (rooster)

Verbal communication: oral vs. written

- Speech vs. silence
- (no) spelling: punctuation marks, capital letters, apostrophe, brackets
  - Your first/family name with/without capital
  - Paris Jackson, Orlando Bloom, mother(-)of(-)P/pearl
  - Homophones: seas, sees, seize; friar, fryer; nun, none; grease, Greece
  - Silent letters (speech): knead, kneed (need); knows (noes, nose)
  - Homographs: minute, lead, object, tear
  - Homophone + homograph: arms, ball, spring, duck, watch, ring



- Speech-to-text (STT) – Text-to-Speech (TTS)

  Fragment from Victor Borge on Phonetic Punctuation (4 mins):
  https://www.youtube.com/watch?v=eixevXANKAo

Key challenges for linguistic tools

- **Separate/Disambiguate**

  tokenizers, sentence splitters, POS taggers, parsers, NER, WSD

  - we/are, its, it/'s, were not, were/n't, brother-in-law
  - We prepared well./ No, no!/She said: "I repeat: come back, please."
  - walk(s), murder, answers, fish, whisper, clean, free
  - (the old car) (has been sold) (in Utrecht)
  - Paris Jackson, WHO, 9 July 2019, £50, 99%
  - homonyms and polysemes (40% of English vocabulary)
    - 'social' as used in 'socialism', 'socialize', 'social media'
    - 'foot' as in 'body part' 'the end of a bed', unit of measurement'
    - The black doctor recevied black money to re-create the black death using black widows. All went black for me then.
    - Warren arrived early this evening (OED: 6x11x6x8=3168)

Key challenges for linguistic tools

- **Cluster/Group**

  spell checkers, lemmatizers, synonyms, MWE - collocations, idioms, (pro)nominal coreferencing

  - run, runs, running, runner, runners; child, children, childish
  - am, are, is, be, being, was, were, been; bear, bore, born; catch, caught
  - pretty, beautiful, attractive; pound, sterling, quid, roll, dough, moolah, loot, dosh; bucks, jack, greenbacks
  - with respect to, black sheep, lily of the valley, buy time, heavy drinker
  - barking up the wrong tree, not my cup of tea, having a ball
  - Warren arrived early this evening. The presidential candidate was accompanied by her daughter./ I was born in Jamaica. I love my country. I consider it my home, Jamrock. I am a Yardie, coming from Yard.

Key challenges for linguistic tools

- **Control/Reduce**

  stopwords, keyword extraction, summarizers

  - a, an, the, in, of, I, you, going, become, was, can, and, or, which (160+)
  - Our expertise and reputation in data governance and information security, along with our technical data infrastructure helps us maintain our leading position as trusted providers of data management services for research. Sharing best practice, we help to lead international best practice in data management and our resources are utilised around the world. Funded by the Economic and Social Research Council (ESRC), we hold important national data resources including many surveys conducted by the Office for National Statistics, national centres for social research and Census data. (UK Data Archive).
    - Keywords: data, providers, resources, expertise, sharing, governance, surveys, infrastructure, management, centres (http://www.cortical.io/extract-keywords.html)
    - Summary: Our expertise and reputation in data governance and information security, along with our technical data infrastructure helps us maintain our leading position as trusted providers of data management services for research. (http://autosummarizer.com/index.php)

Linguistic Analysis - Free Tool

SketchEngine at https://auth.sketchengine.eu/#login

User Manual at
https://www.sketchengine.eu/user-guide/user-manual/

Your TASKS start now!