DH2019 UTRECHT UNIVERSITY 8-12 JULY

Utrecht University

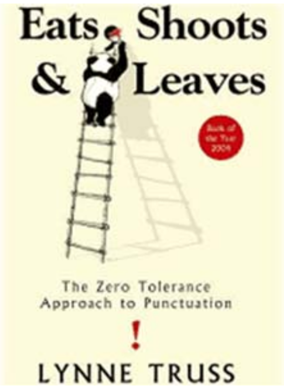
Workshop: The Case of Interview Data, Tuesday 9 July, 09:00-13:00

Session: Introducing linguistic analysis of text: free tools

Background information

Jeannine Beeken, UK Data Archive, University of ESSEX

- Verbal versus non-verbal communication
    - Verbal communication: use of natural language spoken (sound) or written (spelling system)
    - Non-verbal: signs (body language), symbols
    - An example: triple/quadruple negation ‘no’ or ‘stop’
    - An excerpt from a ‘silent’ film (1.10 min) https://www.youtube.com/watch?v=4QrR_1NW_w4

- Oral vs written use of natural language = Spelling versus no spelling
    - Punctuation (marks): comma, question mark, exclamation mark, full-stop; colon and semicolon, quotation marks, brackets.
      Purpose: to clarify meaning of sentences and their components
        - (Example: 

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

    - Morphosyntax or grammar: apostrophe, hyphen, capital letters.
      Purpose in written text: to clarify, disambiguate and cluster

    - Examples: try to pronounce in a *different* way (homophones) – S-to/2-T tools
        - your name with/without capital letters (Cook, Smith, Hall, Hill, Webb, Brown)
        - mother-of-pearl/mother of Pearl with/without hyphens/capital letters
        - Paris Jackson referring to a person or referring to two cities (Jackson Mississippi (river?); geospatial)

        - Other homophones (etymological spelling) –
          1. I, eye, aye (aye, you are right), aye (in voting: Ayes vs. Noes) or ceiling, sealing, seeling; cense, cents, scents, sense; right, rite, wright, write; seas, sees, seize; two, to, too; which, witch; weave, we've; bye, by, buy; cite, sight, site; dear, deer; friar, fryer; none, nun; grease, Greece, higher, hire; leek, leak; bear, bare; it's, its
          2. English words with 'silent letters': knight-night, aisle-I'll-isle, climb-clime, heir-air, hole-whole, hour-our; knead-kneed-need, know-no; knows-noes-nose
    - Examples: try to pronounce *differently* (homographs) T-to/2-S tools
      content, minute, lead, object, record, tear

    - Examples: try to give a *different* meaning to – Disambiguate homophone + homograph)
      arms, ball, bear, bark, current, spring, well, bat, ring, letter, ruler, duck, fall, watch

    - Examples: try to assign a specific PoS (Part of Speech; word class (noun, verb etc.)
      murder, scratch, attack, answer, bomb, brush, care, delay, echo, end, fish, hammer, measure, promise, vote, whisper, clean, free

- Key challenges for linguistic tools
    - Separate or disambiguate
    - Cluster or group
    - Reduce or control

- Common types of linguistic tools for **separation**

    - Tokenizers: separation of words
        - we/are; its; it's becomes it/'s ; weren't becomes were/n't

    - Sentence splitters
        - You prepared well./ No, no!/She said: "I repeat: come back, please."

    - POS taggers
        - noun walk(s), verb walk(s); verb clean, adjective clean

    - Syntactic parsers
        - noun phrases (the old car), verb phrases (has been sold), prepositional phrases (in Utrecht, in Paris)

    - Named entity recognizers (NER)
        - Persons (Paris Jackson), organisations (WHO), geographical entities (Paris, Jackson), numeric expressions (time, date, money, percent)

    - Word sense disambiguation (WSD) used to separate
        - Homonyms

        - Different concepts sharing the same stem: ‘social’ as stem of socialist, socialism, socialize, social (media)

- Common types of linguistic tools for **clustering**
    - Spell checkers and types (reducing x occurrences of the same tokens to 1)
    - Stemmers group words sharing the same stem
        - runs, runs, running, runner, runners
    - Lemmatizers group words being variants of the same lemma (dictionary entry)
        - am, is, was, been, being, were, be
    - Synonym clustering/conceptualization
        - pretty, beautiful, attractive; United Kingdom, UK; jealous, badmind;
        - love, deep affection, fondness, tenderness, warmth, intimacy, attachment, endearment, devotion, adoration, doting, idolization, worship, passion, ardour, desire, lust, yearning, infatuation, adulation, besottedness
        - pound, sterling, quid, roll, dough, moolah, loot, dosh; bucks, jack, greenbacks
    - Multiword expressions (MWE) kept together/having a meaning such as
        - Grammatical collocations
            - on behalf of, with respect to, even though
        - Lexical collocations
            - mother of pearl, distance learning, black sheep of the family, take off, commit suicide, buy time, strong tea, heavy drinker, lily of the valley
        - Idioms
            - barking up the wrong tree, cry over spilt milk, not my cup of tea
        - Concordances (KWIC, n-grams): keywords in context; 2/3/4/5 etc. words before/after
        - Co-occurrence/correlation patterns: the frequency/pattern that words appear in each other’s company: smoking – drinking; school – learning - teaching
    - Semantic wordnets: clusters or sets of linked meanings
    - Syntactic parsers
        - we/played/awfully; they/played/awfully nice; John and Taylor/seemed to be/the best of friends.
    - Pronominal and nominal coreferencing
        - Clinton 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.

- Common types of linguistic tools for **controlled language or reduction**
    - Lists of stopwords, i.e. often the most frequently used words of a language: a, an, the, in, out, of, I, you, us, going, become, was, were, have, can, should, and, or, why,

where, which

- Concept/keyword/topic extraction tools

- Word clouds based on frequency of types/stems/lemmas (also compounds or collocations: on behalf of, black sheep of the family, mother of pearl) (no stopwords)

- Auto-summarizers