Title: acquisition, correlation, visualization and use of lexico-syntactic and semantic features, from Swedish transcribed interactions

Purpose

The purpose of this project is threefold:

1. To conduct a *literature review* in the area of feature extraction from conversation data and spoken language transcriptions (particularly by people with various forms of neural, mental or cognitive impairments such as aphasia, dementia, autism and schizophrenia). A lot of research papers have been published during the last years in this area and the goal in this thesis proposal is to focus on a set of relevant features used in the various described experiments discussed in such literature. Present the most relevant research around this topic and, perform an analysis of the methodologies and acquisition techniques used in these studies. The goal is not to find *all* existing studies conducted around the world, since this would be an impossible goal to achieve. Instead, to focus on the ones in high impact journals and conferences (relevant papers can be available from e.g. PubMed/NCBI).

2. Based on step-1, above, *implement* a (large) set of some of the described lexico-syntactic and semantic features in the papers from step-1 (see "Problem Description" for several examples). The extraction of these features will be acquired from Swedish data, transcribed spoken dialogues (for an example see below). Correlations between features (this can accomplished using ANOVA or Pearson's coefficient) and statistical analysis should be performed for comparison reasons which will be reported, visualized and discussed. The features can be further used as a training input for training and testing various types of supervised machine learning classifiers.

3. Therefore, as an *application scenario*, consider using some existing implementation of such a classifier(s) (e.g. Support Vector Machine: SVM) by using the extracted features from step (2) above. The application would be used to differentiate between transcribed spoken dialogues (interviews) with people that either developed or not dementia years after the interviews were taken (a *facit* will be available). Therefore the classifier can be a simple binary one with two class labels e.g. *yes* and *no* or other labels we can agree on.

In sum you should:

- implement a set of features from the data sets
- model the features (each dialogue sentence will be written as an *n* element vector of features values)
- work and discuss which the most significant features are
- select a "good" set of features and discuss their predictive power
- run various evaluation rounds with different feature sets
- evaluate your results (since you know in advance whether the corpus examples come from a person that developed or not a neurogenerative condition years after the interview was taken) e.g. by "saving" some random sample of sentences for the evaluation of the classifier. As a baseline you may use the most frequent dialogue type, e.g. *no* if most examples come the set of interviewed persons that didn't develop dementia.

Background and problem description

Supervised machine learning, e.g. for predictive analytics, requires input data in the form of feature-value pairs. Such pairs can be acquired from various sources depending of the application in mind and the context in question. In this work you will implement and acquire

a set of such features that will be used in supervised machine learning experiments.

For the project it's advisable to preprocess the input data (an example is provided in Appendix A). You may also use the Språkbanken pipeline (<http://spraakbanken.gu.se/eng/research/infrastructure/korp/corpus- pipeline>) to part-of-speech annotate and syntactically parse the data. For the features to be used you can get inspiration from various published research papers, such as for instance Roark et al., 2011; Le, 2010; Le at al., 2011; Rentoumi et al., 2014; but also based on other literature you will identify from step-1 described in the "Purpose" section above; inspiration about features you can also get from implemented packages such as "*koRpus: An R Package for Text Analysis*". A selection of features can for instance be:

- *Lexical richness* (i.e. the proportion of type lemma per tokens – measures the lemmatised word type normalised by all words)
- the proportion of *simple sentences*. (a sentence that contains only one finite verb)
- the proportion of *complex sentences* (a sentence that contains more than one finite verb)
- *Scoring measures*
    - *Yngve scoring*
    - *Frazier scoring*
- *Dependency distance*
- *Density measures*
    - *Idea density (*or *proposition density*; i.e. the average number of ideas expressed per words used)
    - *Content density*
- the *proportion of subordinate and coordinate phrases to the total number of phrases*
- the *mean dependency distance* (the sum of its individual distances divided by the number of its dependencies)
- *Filler ratio* – filler sounds such as 'hm' and 'ehm' are used by people in spoken language when they think (if these are present)
- *Semantic depth* (of tokens using the lexical-semantic network Saldo; i.e. the distance between a lexical entry and the PRIME. The example below (Heimann Mühlenbock, 2013:114) shows two paths or depths for the Swedish noun *läge*:

Lexeme	Depth	Path in Saldo		
läge	3	← plats	← var	← PRIM
läge	3	← situation	← vara	← PRIM

- …

Such features can be used to create and train a/different supervised model(s) that can be applied to your text data in order to (possibly) differentiate between the transcribed spoken dialogues. You should also perform some statistical analysis of the features and rank their predictive power as previously discussed (step-3 in the "Purpose" section).

Recommended skills
It is free to use any programing language and programing environment(s), such as WEKA, R.

Supervisors
Dimitrios Kokkinakis and possibly other researchers from Språkbanken or Chalmers.

Resources
SALDO v2.3:
<http://spraakbanken.gu.se/swe/resurs/saldo>
MaltParser: <http://www.maltparser.org/mco/mco.html>
Språkbanken's annotation tool: <https://spraakbanken.gu.se/sparv/>

References

- Bucks RS. et al. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, vol. 14.
- Anthony Habash. 2012. *Language analysis of speakers with dementia of the alzheimer's type*. PhD thesis. University of North Carolina Wilmington, USA.
- Xuan Le. 2010. *Longitudinal Detection of Dementia through Lexical and Syntactic Changes in Writing*. Masters Thesis, University of Toronto, Canada.
- Xuan Le, Ian Lancashire, Graeme Hirst and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Oxford JournalsArts & Humanities Digital Scholarship in the Humanities.* Vol 26:4, pp. 435-461.
- Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. *Spoken Language Derived Measures for Detecting Mild Cognitive Impairment.* IEEE Trans Audio Speech Lang Processing. 19(7): 2081–2090.
- Rentoumi V., Raoufian L., Ahmed S., de Jager CA. and Garrard P. 2014. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *J Alzheimers Dis.* 42 Suppl 3:S3-17. doi: 10.3233/JAD-140555.

Appendix A

```
Filnamn: AXXX09
Längd: 1.09.16 tim

...
P: nej jag skiter i den (skratt)
I: ja precis, alltså
P: ja det är ju rätt intressant faktiskt (skrattar)
I: precis, det har du aldrig varit med om en sån här (skrattar)
P: nej
...
P: mm, ja
I: för den är ju väldigt liksom
P: ah
I: bygger mycket på den i vården.
P: det är det
I: så det, det är jag och är det nånting annat du undrar över?
P: nej
...
I: ja just det (skrattar till)
P: åh, gud
I: mm
P: nej så att då så blev det ju som det blev för vi skyddade ju oss bara första gången och så
I: (skrattar till)
P: knäppt, knäppt, knäppt
I: ja
P: nej för jag menar (harklar till) det var tufft alltså och jobba (hostar) dag, kväll, dag, kväll
I: mm, tiderna var jobbiga?
P: jajaja
...
```