RE: automatically collecting psycholinguistic features of Swedish words from corpora?

Paridon, Jeroen van

Tue 2019-04-30 11:41 AM

To: Montero-Melis, Guillermo <Guillermo.MonteroMelis@mpi.nl>;

The reason I used OpenSubtitles instead of Wikipedia here is two-fold, by the way:

1. OpenSubtitles frequencies seem to be better than Wikipedia for predicting behavioral data (cf. Brysbaert's SUBTLEX work)
2. I only have the lemma/POS data for this corpus

Jeroen

From: Montero-Melis, Guillermo <Guillermo.MonteroMelis@mpi.nl>
Sent: dinsdag 30 april 2019 11:33 AM
To: Paridon, Jeroen van <Jeroen.vanParidon@mpi.nl>
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Jeroen,
This is great, thanks! I made a copy of the files and took a quick look and most of it seems pretty clear. Some quick questions:

- What's the source for this data? Swedish Wikipedia? Subtitles? Is there any description of the database somewhere online?
- What's the difference between the "normal" and the filtered versions of the files?

I'll get back to you if I have more questions later.

Best,
Guillermo

From: Paridon, Jeroen van
Sent: Tuesday, April 30, 2019 10:24:42 AM
To: Montero-Melis, Guillermo
Subject: RE: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Guillermo,

The files you need are in scratch/jerpar/for_guillermo
The files with prefix filtered_ contain only the frequencies relevant for your verbs, but I've included the full datasets for Swedish as well, in case you want them. I haven't computed the grammatical ambiguity for each word, but your student should be able to do that herself by dividing the number of verb role occurrences by the total number of occurrences for a word, right?
Just let me know (or come by) if you have any questions.

Kind regards,

Jeroen

From: Montero-Melis, Guillermo <Guillermo.MonteroMelis@mpi.nl>
Sent: maandag 29 april 2019 5:41 PM
To: Paridon, Jeroen van <Jeroen.vanParidon@mpi.nl>
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

Excellent, thanks!
Guillermo

From: Paridon, Jeroen van
Sent: Monday, April 29, 2019 4:52:29 PM
To: Montero-Melis, Guillermo
Subject: RE: automatically collecting psycholinguistic features of Swedish words from corpora?

Let's say tomorrow? (It'll take a little bit of time to write the code, and then some minutes for the script to trawl through 2 GB of text.)

Jeroen

From: Montero-Melis, Guillermo <Guillermo.MonteroMelis@mpi.nl>
Sent: maandag 29 april 2019 3:24 PM
To: Paridon, Jeroen van <Jeroen.vanParidon@mpi.nl>
Cc: Margareta Majchrowska <margareta.majchrowska@isd.su.se>
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

Ah, thanks, Jeroen, that's excellent news!
Do you have an idea of how long if will take for a first result (esp. for lemma frequency): today? some days? a week? more?
Knowing this will help me plan ahead.
cheers
g

From: Paridon, Jeroen van
Sent: Monday, April 29, 2019 3:19:02 PM
To: Montero-Melis, Guillermo
Subject: RE: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Guillermo,

I checked, and it would appear that I can extract lemmatized+POS frequencies.
I'll give it a go and then we can see if the results look sensible. We can always resort to extracting all the verb forms later, if necessary.

Kind regards,

Jeroen

From: Montero-Melis, Guillermo <Guillermo.MonteroMelis@mpi.nl>
Sent: maandag 29 april 2019 10:51 AM
To: Paridon, Jeroen van <Jeroen.vanParidon@mpi.nl>
Cc: Margareta Majchrowska <margareta.majchrowska@isd.su.se>
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Jeroen,
I'm attaching the list of Swedish verbs. Could you please see which of measures 3-7 (see history below) are easy for you to obtain?

Two thoughts:
1) We are only interested in verbs, so: If the corpus is lemmatized *and* you can figure out grammatical category (part of speech, i.e. verb/noun/...), then the relevant frequency measure would be for the verb lemmas.
2) If the corpus isn't lemmatized, the easiest option might be for us (my student and I) to write the different forms of the verbs (past tense, etc). In that case, we are back to word frequency and that should be easy.

Cheers,
Guillermo

From: Paridon, Jeroen van
Sent: Thursday, April 11, 2019 11:08 PM
To: Montero-Melis, Guillermo
Cc: Margareta Majchrowska
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

If you're going to be out of the office for a while, maybe we should meet tomorrow before the lab meeting. I can make sure to come in at 9:30, if that works for you.
I'm happy to help with points 3-7, of course. With regards to collecting the 8-13 norms in Swedish: any data is of course better than nothing.

Jeroen

From: Montero-Melis, Guillermo
Sent: Thursday, April 11, 2019 10:57:27 PM
To: Paridon, Jeroen van
Cc: Margareta Majchrowska
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Jeroen,
Thanks, that's very useful!

So here are my thoughts:
- If 3-6, and possibly 7, are easy for you, it'd be great if you could help us with that.
- By 7 my guess is that they indeed mean that kind of noun/verb/adjective ambiguity; they don't specify it, but it's the only thing that makes sense given the context (they checked these things somewhere, god knows where, so it has to be information that is more or less easily available for English)
- For 8-13, we'll go on with our norming idea then, since I am doubtful I will find any of these norms for Swedish. We can be pragmatic about it and collect norms from 20 people or so. For this, I think (hope) it will suffice.

Tmw we have our lab meeting 10-12. I'll leave early in the afternoon and next week I won't come in. So it'd have to be the week after. Are you going to be around for the last 3rd of April / beginning of May?

Thanks again!

Best,

Guillermo

From: Paridon, Jeroen van
Sent: Thursday, April 11, 2019 8:27 PM
To: Montero-Melis, Guillermo
Cc: Margareta Majchrowska
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Guillermo,

3, 5, and 6 would be trivial for me to retrieve if you send me a list of words. 4 is likely trivial as well, although I'd need to check if the lemmatized corpus I have for Swedish is large enough to get reliable numbers. For 7, do you mean grammatical ambiguity in the sense that a word like "lift" can be both verb and noun, or do you mean phrase-length (or longer) units that can be interpreted multiple ways? (The former I can probably find out for you, the latter would be next to impossible.)

8 through 13 are not trivial, unfortunately. There is a method for matching semantic dimensions in different languages, but it's computationally intensive and very much a work in progress (one that I won't get to work on very much until after I finish my thesis). If you can find any of these norms in Swedish (even if it's for entirely different words) that would make it easier because then I can just extend the norms within the language, instead of having to try to predict across languages.

Either way, we can discuss this in person if that's helpful. Tomorrow I'll be in my office between 10:00 and 11:45 for certain; outside those hours I might be in meetings.

Kind regards,

Jeroen

From: Montero-Melis, Guillermo
Sent: Thursday, April 11, 2019 5:36:12 PM
To: Paridon, Jeroen van
Cc: Margareta Majchrowska
Subject: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Jeroen,
(cc-ing Margareta, a Master's student in Stockholm who is working with me on this project)

Let me ask you a question following up on something you mentioned the other day during our meeting. I want to know how difficult / time demanding it would be for you to do your magic using your vector space tools to solve an issue of stimuli validation. Let me explain:

For a replication study, we were planning to collect some norms for our Swedish materials (the original study was conducted in English; I am attaching the study we are replicating for reference, Shebani and Pulvermüller, 2013, *Cortex*; **S&P** for short)

We would like to match two lists of words along the following variables (cf. p.225, Table 1 in S&P):

1. Number of letters
2. Number of phonemes
3. Word frequency
4. Lemma frequency
5. Bigram frequency
6. Trigram frequency
7. Grammatical ambiguity

8. Valence
9. Arousal
10. Imageability
11. Visual relatedness
12. Body relatedness
13. Action relatedness

Unfortunately, S&P don't cite any sources about where they took this information from. In any case: Some of these are trivial (1) or foreseeably easy to obtain even for Swedish (2-4). Others are slightly more tricky, at least for me (5-7). For 8-10 it's easy to find English norms (Brysbaert's stuff). For 11-13 we've struggled a bit more to figure out who collected them, and thus we are not entirely sure what the numbers mean (probably ratings on Likert scales, but on which exactly? Anyway, that's a separate issue).

My approach right now was to collect data for 8-13 ourselves: set up an online survey form and have 20 native speakers or so rate our critical words along these dimensions (probably interspersing them with other random words to not bias the ratings). But I just talked to Markus and he told me you might be able to get some of these measures for Swedish using a computational approach? Is it the case? And if so, how much work would it take? What information would you need to do that? E.g., for the Visual Relatedness ratings, suppose we find a study that has collected such norms (à la Brysbaert); would that be enough to get a measure of this for Swedish as well? I guess it'd help me to know how you would go about...

Perhaps it's easier to discuss this in person, so let me know if you have a minute and I can drop by your office.

Cheers,
Guillermo