

Project proposal for Computational Semantics class

Zi Huang, Hamit Kavas, Albert Lleó, Clàudia Martínez

November 12, 2019

1 Goal

To predict the use of *to*-infinitives and gerunds as complements of a pre-defined set of verbs, with which the choice of infinitive or gerundive complement only affects the meaning **minimally**.

	<i>to</i> -inf	<i>-ing</i>
like (in all forms)	11643	741
like (not in base form) ¹	2897	541
begin	18386	2617
continue	9742	911
hate	322	234
love	914	373
prefer	1947	96
propose	1157	191
start	6019	6666
attempt	6326	22
intend	6297	272
not bother	143	11

Table 1: Raw frequency in BNC (SketchEngine)

The table shows that BNC will provide enough data for this task. *To*-inf is generally preferred to *-ing* (the case of *start* then is very interesting). It is said that *to*-inf is the unmarked form.

One thing we can show with **cosine similarity** is that among the verbs that take both infinitive and gerund as complements, we can distinguish between two classes: one that distinguishes two meanings with two forms, and one that does not. For example, we expect that *try to xxx* and *try xxxing* are more different, while *begin to xxx* and *begin xxxing* are close.

2 Patterns to extract

For *to*-inf: for each_verb in verblist: [lemma="each_verb"][word="to"][tag="VV"]

We probably will still try to avoid *would like to*.





For gerunds: for each_verb in verblist: [lemma="each_verb"][tag="VVG"]

If the corpus is parsed, we can extract the complement of these main verbs for which the head is a verb, so that we also get the cases where there is a noun/adv/adj between the main verb and the non=finite one.

3 Features

1. **Animacy** of the subject. (This is not applicable for *like/hate/love/propose/intend/attempt* though, because their subject is naturally animate...)

¹ *Would like* seems to strongly prefer a *to*-inf complement.

- Bresnan used Garretson et al.'s coding practice. Maybe see this paper:
<https://acl-arc.comp.nus.edu.sg//antho/W/W04/W04-0216.pdf>
- Whether **person** is switched in the next utterance. (*To* forms necessarily have Subject Control; *-ing* forms are allowed to have a different subject. I suppose that *-ing* forms will favor a switch of person.)
 - If the corpus is parsed then probably it's not hard to find the subject (see Gemma's suggestion on dealing with personal pronouns - and common nouns can be distinguished according to number)
 - Length** of the non-finite verb. 
 - Whether there is anything in the **constituent headed by the non-finite verb**. The **length** of this constituent: in words or in characters. 
 - Distance to **last mention** of the non-finite verb, in any inflection or derivation. (If longer than 20 or more words, can be seen as "no last mention".)
 - Loop back in the corpus until we find the last occurrence of the verb. Save the number of lines that we loop.
 - ~~How **positive/negative** is the sentence containing target expression. (It is said that *like to* is used when there is a stronger desire. Use a sentiment analysis dataset to grade the words.)~~
 - ~~• If Hamit is interested in this feature. A possible way to do it is to look for a dataset of sentiments and apply their value to calculate for each sentence.~~
 - Whether there is an **adverb** that precedes the main verb. Whether the main verb is **negated**. Whether there is an adj/noun before or after the non-finite verb. 
 - ~~**Semantic class** of the non-finite verb. (Use VerbNet classification.)~~
 - Argument structure** of the non-finite verb. (Could be hard, but if we have the whole constituent, we can distinguish between "no complement", "NP complement" and "PP complement".) 
 - ~~**Aspectual feature** of the main verb. (Or maybe each of them has its own bias...)~~
 - Temporal anchoring**. Whether there are temporal expressions that talk about generic situations or a future time. 