Quantifying verbal behavior in search of grammatical creativity

Ittamar Erb[1] and **Shahar Spencer**[2] and **Nurit Melnik**[1] and **Gabriel Stanovsky**[2]
[1]The Open University of Israel
[2]The Hebrew University

Relevant UniDive working groups: WG1,WG4

1 Introduction

The relationship between theoretical linguistics and computational linguistics has a long and dynamic history. We propose that at this point in time, NLP tools and methods can be harnessed to advance goals of theoretical linguistics, namely to gain an understanding of the nature of the language system, as one facet of cognition. This approach is implemented in our current project, which focuses on *grammatical creativity*. We are conducting data-driven investigations aimed to determine the extent to which language use is conservative and predictable, or conversely, variable and flexible. This question touches upon fundamental issues in theoretical linguistics, such as the core-periphery distinction, the limits of grammar, and language variation and change.

We explore grammatical creativity at various linguistic levels (syntax, semantics, lexicon). Our working hypothesis is that creativity occurs when speakers go "off script", producing utterances diverging from linguistic conventions. A well-known example is Goldberg's (1995) *sneeze the napkin off the table*, where the typically intransitive verb *sneeze* appears in a ditransitive argument structure.

We aim to develop methodologies that are largely language independent, and therefore opt to use multilingual processing tools, and avoid making idiosyncratic decisions. For a proof of concept, we begin by exploring creative dimensions in English, where models and data are plentiful.

This paper reports on a work-in-progress exploration of creative verbal uses, conducted as part of the project. We perform various quantitative analyses of morpho-syntactic distributions of verbs to find their creative instances. Specifically, we quantify variation in the distribution of (i) individual lemmas across different parts of speech (POS), and (ii) argument-structure (AS) realizations of verbs.

2 Grammatical creativity and linguistic theory

The expressive power of language, enabling people to produce and understand sentences never made or heard before, is a central pillar in modern theoretical linguistics. Still, the nature of the mechanisms underlying it remains a major bone of contention in the field. *Lexicalist approaches* stress the regularity of verbal behavior, and assume that syntactic information is encoded in lexical entries (e.g., Levin and Hovav, 1994). Conversely, *syntactic approaches* emphasize variability and claim that lexical entries are not encoded for their argument structures (e.g., Borer, 2003). Finally, constructional approaches reject a strict dichotomy between syntax and lexicon, allowing both abstract and rich syntactic representation of linguistic signs (e.g., Goldberg, 1995).

While all approaches provide theoretical machinery to model variation in the syntactic behavior of verbs, they differ in the prominence they attribute to the flexibility of verbs in language use. A lexicalist approach predicts the syntactic configurations of verbs to be individually represented, and their range of variation, relatively rigid; unconventional uses are deemed to be marginal. Conversely, a syntactic approach allows an unconstrained range of syntactic realizations of lexical items, predicting syntactic flexibility. A constructional approach allows a range of variance in the syntactic realization of different lemmas, as some may be encoded with syntactic information, and others might not. These predictions will be put to test in our project.

3 Quantifying verbal behavior

The current paper focuses on two linguistic dimensions at the center of the debate: a *morpho-syntactic* dimension which targets the distribution of lemmas across different POSs, and an *argument structure* dimension which concerns the range of argument structures in which verbs appear. For each dimension, we rate verbal lemmas for their *rigidity* and *flexibility*: a lemma restricted to a certain POS/AS is considered rigid, while a lemma

showing a high degree of variance is considered flexible. This analysis is then utilized for detecting creative instances. Assuming that unexpectedness and surprise are distinctive features of creativity (Runco and Jaeger, 2012; Simonton, 2012) we propose the following hypothesis: Linguistic expressions with unconventional uses of rigid lemmas (in either dimension) are likely to be deemed creative.

3.1 Methodology

Our main computational tool is spaCy[1] - an NLP suite with features such as lemmatization, dependency parsing (using a variant of Universal Dependencies (Nivre et al., 2016)) and POS tagging. We use SpaCy to analyze the Blog Authorship Corpus (Schler et al., 2006), an open source dataset consisting of 681,288 posts and over 140 million words. We chose this domain as it is likely to have high variation in the vocabulary and the topics discussed and allow many creative occurrences, representing natural language use "in the wild".

We develop a model for each dimension. For the *morpho-syntactic* dimension, we compute a co-occurrence matrix of all verbal lemmas in the corpus and open-class POSs (Verb, Noun, Proper Noun, Adjective). The *argument-structure* dimension is restricted to verbs occurring in complete clauses. We compute for each verb the number of times that it occurs with a particular subset of the set of non-subject core dependents {DOBJ, PREP, DATIVE, XCOMP, CCOMP}.

We explore the flexibility of each lemma by calculating the entropy of its distribution across the categorical variables associated with each dimensions (POSs for the morpho-syntactic dimension, dependency sets for the AS dimension), where lower entropy indicates rigidity. We hypothesize that creativity will be found when rigid lemmas are used non-canonically.

To test our hypothesis, we identified for each dimension the 20 most rigid lemmas, and for each lemma we retrieved a sentence in which it was used in its least frequent configuration. Each instance was rated on a 3 point creativity scale by our team.[2] Furthermore, we compared the two sets against a random sample of 20 sentences for evaluation.

[1] https://spacy.io/

[2] At this point, the evaluation was performed intuitively, but we plan to elicit judgments from naive native speakers.

3.2 Preliminary results

For the morpho-syntactic dimension, we extracted verbal instances of lemmas canonically associated with other POSs, e.g., denominal verbs (Clark and Clark, 1979). We filtered the data for lemmas occurring as verbs less than 50% of the time, with a minimum threshold of 5 tokens. For the 20 lemmas with the lowest entropy, we randomly retrieved a corpus sentence in which each lemma occurred as a verb.

Of the 20 sentences, 7 were found to be potentially creative (1a). In six cases the use of the lemma as a verb was not frequent, albeit not unconventional (1b). The rest of the cases were typos (e.g., *planing* instead of *planning*) and wrong analyses of spaCy (e.g., miss-tagging *lots* as a verb).

(1) a. it's hip [...] to **lower case** your type.
b. at least TRY to **reason** with the man

For the argument-structure dimension we focused on verbs which alternate between at least two sets of non-sentential complements ({DOBJ, PREP, DATIVE}). For each of the 20 lemmas with the lowest entropy score, we randomly retrieved a single corpus sentence in which the verb appears in the least frequent argument structure.

Only three sentences were found to be potentially creative (2a). However, 11 sentences were retrieved based on erroneous analyses. Thus, for example, only the DOBJ complement of a ditransitive occurrence of *rid* was identified as a complement (2b). Other indexing issues involved non-standard language (2c).

(2) a. tHey aCtually **aRe tO me**
b. I do put time aside [...] to help **rid** the society of those who plan to destroy it.
c. mom says '**gimme** a drag of that'

3.3 Discussion

The comparison between the two sets and the random sample was instrumental. The sentences of the latter set were all judged to be standard, and only two were parsed incorrectly. We conclude that the higher rate of potential creativity and indexing errors in the two sets is not coincidental, as we are in principle looking for *tail phenomena*. This suggests that our heuristics capture less conventional expressions.

Nevertheless our goal is to reach higher rates of precision in the two tasks and extend the methodology to other dimensions. Our main challenges

are to tease apart creativity from rarity, as well as from human and indexing errors.

References

Hagit Borer. 2003. Exo-skeletal vs. endo-skeletal explanations: Syntactic projections and the lexicon. *The nature of explanation in linguistic theory*, 31:67.

Eve V Clark and Herbert H Clark. 1979. When nouns surface as verbs. *Language*, pages 767–811.

Adele Goldberg. 1995. *Constructions. A Construction Grammar approach to argument structure*. University of Chicago Press, Chicago.

Beth Levin and Malka Rappaport Hovav. 1994. *Unaccusativity: At the syntax-lexical semantics interface*, volume 26. MIT press.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Mark A Runco and Garrett J Jaeger. 2012. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Dean Keith Simonton. 2012. Taking the us patent office criteria seriously: A quantitative three-criterion creativity definition and its implications. *Creativity Research Journal*, 24(2-3):97–106.