Computationally Deriving Language-Internal Factors with Bipartite Networks

Daniel Duncan*

1 Introduction

Many sociolinguistic variables are constrained by the lexical semantics of an element in the linguistic environment surrounding the variable. For example, Spanish subject pronoun expression is often described as conditioned by the semantic class of the verb (Orozco and Hurtado 2021), while variation in the Russian genitive correlates with the semantics of the possessor nominal (Naccarato et al. 2021). The English alternative embedded passive (AEP), also known as the 'needs washed' construction, is likewise said to be constrained by semantic factors. The AEP contrasts with the canonical embedded passive (EP) through the lack of *to be* between the matrix verb and participle it selects. It has been primarily attested with three matrix verbs (Edelstein 2014, Murray et al. 1996, Murray and Simon 1999, 2002): *need*, *want*, and *like* (1). This set of matrix verbs is much more restricted than that of the EP. Murray and Simon (1999) suggest that this restricted usage is subject to some sort of semantic constraint. However, in contrast to the restricted set of attested matrix verbs, recent work has attested the AEP with a wider range of matrix verbs (2).

(1) a. The car **needs** (to be) washed.
    b. The dog **wants** (to be) fed.
    c. The baby **likes** (to be) cuddled.
(2) a. I don't think he **deserves** fired. (Duncan 2021:485)
    b. Gretchen **loves** petted and is a great lap cat. (Duncan 2021:486)
    c. She **hates** petted and only comes near us when she's starving. (Duncan 2021:486)
    d. Surely Lineker **requires** fired for that. (Strelluf 2022:66)
    e. Your radiator could **use** flushed. (Tenny 1998:596)

In light of examples such as these, it is clear that the description of the AEP is incomplete. In fact, there is a relatively basic set of research questions to be explored: what matrix verbs are permitted in the AEP, and what factors constrain this?

However, the simplicity of these questions is deceiving. Following Murray and Simon (1999), we would hypothesize there is a semantic constraint on matrix verb productivity. There are three difficulties here. Firstly, where in the construction is this constraint? In this paper, I will assume the constraint is in the lexical semantics of the matrix verb, but this does not mean every possible semantic constraint has been explored. Secondly, a general difficulty with coding semantic factors is that factor levels can be rather fuzzy and problematic (Orozco and Hurtado 2021). Thirdly, because we do not know what the possible matrix verbs in the construction are, we do not know what the factor levels of a semantic constraint should even be. For these reasons, to examine what constrains matrix verb productivity of the AEP demands more consideration than meets the eye.

This paper proposes that the latter two difficulties of category fuzziness and unknown factor levels may be addressed through computational modeling. Specifically, I suggest that given properly structured data, a model can cluster items together approximately by meaning. These clusters, in turn, are effectively levels of a semantic factor that can subsequently be applied to variationist data. In effect, a model may be used to reproducibly group items within a fuzzy category without presupposed knowledge of what the groups should be. I illustrate one such approach to computationally deriving language-internal factors for study of the AEP. I link EP matrix verbs (and thus potential AEP matrix verbs) to the participles they select in a weighted bipartite network. Quantitative network metrics yield a measure of verb productivity, while a community detection algorithm groups matrix verbs based on the participles they select for. I show that this latter factor clusters verbs by

*Thanks to Loretta Bushall, Mary Robinson, and the reviewers and audience of NWAV50 and the Linguistics Association of Great Britain and Northern Ireland for helpful feedback on this work. This project was supported by BA/Leverhulme Small Research Grant SRG2021\210047.

semantic likeness. In applying these derived factors to tens of thousands of acceptability ratings, I demonstrate that computationally derived language-internal factors can make intuitive sense, significantly correlate with linguistic data, and contribute to our understanding of a linguistic phenomenon. Given this, I argue that this approach has useful applications for the study of linguistic variation beyond the AEP and beyond semantic constraints.

2 Bipartite Network Models

This section provides background on the use of bipartite network models in variationist sociolinguistics thus far, describes the application of them to a linguistic context, and demonstrates the construction of a network linking matrix verbs to embedded participles.

2.1 Background

Social network analysis has a longstanding role in variationist sociolinguistics, as how an individual connects to the rest of their contacts, as well as the density of ties within the network as a whole, influences patterns of language variation and change in predictable ways (Milroy 1987). Network analysis, as exemplified by Milroy (1987), has typically been performed by linking individuals to other individuals they are in contact with. In such networks, individuals are only linked to one another when they directly interact with one another on a regular basis.

In contrast to such networks, a bipartite network connects members of two distinct categories. One common use of such networks is to link individuals to institutions (Dodsworth and Benton 2019, see Figure 1). In such networks, individuals may be considered connected to one another when they interact with the same institution, regardless of whether they are in direct contact with one another. Bipartite networks may be weighted or unweighted. An unweighted network considers links between category members to be a binary: either there is a link, or there is not. On the other hand, a weighted network includes each separate link between category members, even if links are repeated.

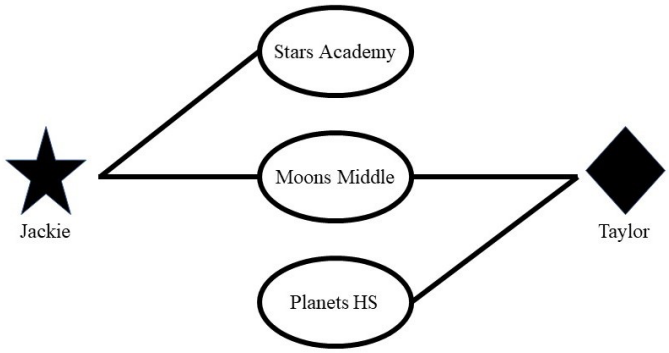

Figure 1: Toy bipartite network model linking speakers to educational institutions.

Use of bipartite networks in variationist sociolinguistics thus far has been primarily to link individuals to educational institutions in unweighted networks (Dodsworth and Benton 2019, Labov et al. 2016). Mapping such networks in a manner like in Figure 1 can illustrate patterns not readily apparent from other approaches to education effects on local variation. For example, Labov et al. (2016) use a bipartite network to find that young Philadelphians' short-*a* systems are strongly linked to secondary school attendance. Those who attended parochial Catholic schools have the traditional Philadelphia short-*a* split, while those who attended public schools and prestigious private institutions have the nasal short-*a* system typical of many American Englishes.

Networks may be analyzed using quantitative measures as well. In their network analysis of education and linguistic production in Raleigh, NC, Dodsworth and Benton (2019) use measures of structural cohesion and structural equivalence to examine retreat from the Southern Vowel Shift (SVS). In addition, they use a community detection algorithm to group speakers with who attended similar sets of schools. This effectively is a method of obtaining a language-external factor, which they show is correlated with SVS retreat.

2.2 Use in Deriving Language Internal Factors

Although bipartite networks in sociolinguistics have thus far been used to link individual speakers to a group or institution, the categories need not be limited to this scenario. Because bipartite networks simply link members of one category to members of another, the two levels of the network model could comprise ordered ngrams, parts of a construction, or another kind of linguistic object with two identifiably distinct components. In this sense, the first component is one category, with a set of lexical items, morphemes, etc., as members, while the second component is the other category that is being linked to.

To construct a bipartite model for a linguistic object necessitates a corpus of text in which the object appears with regularity. The model itself would be constructed by compiling an exhaustive list of category membership for each component within the sample and entering co-occurrences of category members into a matrix. Such an approach has applications to natural language processing (Nastase et al. 2015), as well as linguistic research more generally. For example, Shirtz uses bipartite network models in his (2019) exploration of nominal predication in ten Indo-Iranian languages.

Here I suggest that a similar approach is useful for studying the AEP. The construction has two mandatory components in the matrix verb and participle it selects. We can thus create a bipartite network linking members of these two categories (Figure 2). As seen, the network I suggest here connects matrix verbs which select the same participle.

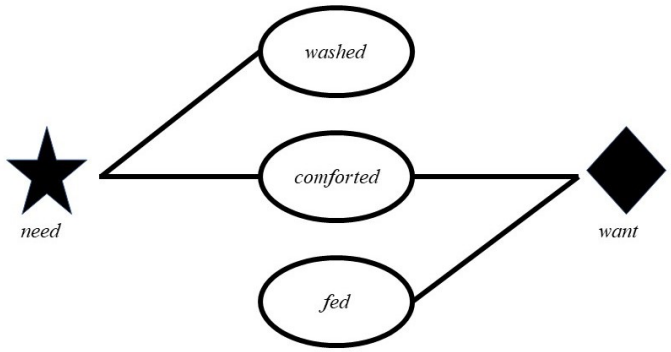

Figure 2: Toy bipartite network model linking matrix verbs to embedded participles.

Two quantitative analyses of such a bipartite model of the AEP are likely to yield useful language-internal factors. The first is degree centrality. By counting how many participles are connected to a given matrix verb, this measure represents how productive the matrix verb is in the construction. In an unweighted model, this PRODUCTIVITY measure simply captures how many participles a matrix verb selects for, while in a weighted model it captures matrix verb frequency in addition to the information in an unweighted model. Calculating degree centrality for a bipartite model of the AEP is thus a principled way to obtain a language-internal factor similar to lexical/lemma frequency, which is focused solely on how matrix verbs are used within the construction.

The second useful quantitative analysis is application of a community detection algorithm to determine which sets of matrix verbs select similar sets of participles. This is effectively a method for selecting sets of matrix verbs that share a degree of SEMANTIC LIKENESS. Consider, for example, that a dog may *like*, *love*, or *hate* (*to be*) *petted*. While these matrix verbs are not exact synonyms, they all share an element of sentiment toward an action. However, one would be unlikely to say that a dog or another subject *scrambles* or *forgets* (*to be*) *petted*, both for discourse-pragmatic reasons and because the matrix verbs do not share this sentiment element. If a researcher intended to sort verbs into semantic categories by hand, the former group of matrix verbs could thus be coded separately from the latter. In a bipartite network constructed from a sample of language use, the distinction between groups of matrix verbs such as these would emerge as common matrix verb/participle pairs would be more likely to appear than unlikely pairs. By dividing matrix verbs into semantic likeness groups algorithmically, this analysis thus obtains an otherwise fuzzy language-internal factor in a principled, rigorous manner.

Both of these factors—verb productivity via degree centrality and semantic likeness via community detection—are best derived from a weighted bipartite network model. With respect to verb productivity, this is because a weighted model explicitly includes verb frequency information, while an unweighted model includes this information implicitly and incompletely. Consider two verbs which each select the same set of participles. If one verb is far more frequent than the other, in any given sample this verb is likely to be observed co-occurring with more of the set of participles than the infrequent verb. In this situation, an unweighted model in which the more frequent matrix verb co-occurs with more participles is simply showing that the more frequent matrix verb is more frequent. However, how much more frequent this matrix verb is compared to the other has been lost. In contrast, a weighted bipartite model keeps both the different range of selected participles and the magnitude of frequency difference, thus providing a more accurate picture of verb productivity. Keeping frequency information is likewise important for deriving semantic likeness; if a given matrix verb co-occurs quite frequently with a subset of the participles it selects, it likely is most semantically like matrix verbs that also select for that subset.

While quantitative analysis of a bipartite network model of the AEP or another linguistic feature can thus yield language-internal factors that may be applied to further analysis, it is important to emphasize that such factors are sensitive to the input data. Because such bipartite networks are constructed from corpus data, this means that the choice of sample is a crucial consideration. It must be a large enough sample to contain the linguistic object at a high enough frequency to yield an informative network, as well as a sample that is reasonably representative of the variety whose data will be analyzed using the derived factors. At the same time, the community detection algorithm in particular is sensitive to the membership of the two model components, as missing data will result in group membership being delineated differently than if the data were included. This means that when the network model includes an open class, care must be taken search for the linguistic feature as systematically as possible. Furthermore, any decisions regarding inclusion of category members in the model should be documented as thoroughly as possible.

2.3 Network Construction

In this section, I discuss the construction of a bipartite network and the derivation of verb productivity and semantic likeness factors through quantitative analysis of the network. In order to be most relevant to data concerning usage and acceptability of the AEP, this model would ideally be of AEP usage in a large sample of speech from a population that uses the construction. Given that the AEP is underattested in the literature, this is of course not possible; because embedded passives are rather uncommon, a very large corpus is necessary to find sufficient numbers of the feature to construct the network model. However, corpora of a sufficient size are of national varieties of English, which means that a noncanonical linguistic feature will not appear in sufficient numbers. For this reason, I instead construct a bipartite network model linking matrix verbs to participles in the canonical EP. This is likely to be a good approximation given that researchers often treat the EP and AEP as variants identical in meaning (Duncan 2019, Murray et al. 1996, Strelluf 2020, 2022). However, I acknowledge that this contributes additional uncertainty and error to the derived language-internal factors.

I use the Corpus of Contemporary American English (COCA, Davies 2008) as the data source for network construction. At approximately 1 billion words, COCA is large enough to find a wide range of EP tokens for coding. A list of potential EP matrix verbs was compiled primarily from the Penn Treebank 2a guidelines for Raising, Control, and Exceptional Case Marking (ECM) verbs (Taylor 2006). Each of these categories of verb may be followed by an embedded passive. Because the AEP does not contain *to be* and is limited to selecting eventive participles (Edelstein 2014, Tenny 1998), I omitted copular constructions; the modal constructions *ought to*, *have to*, and *have got to*; and the verbs *look*, *seem*, and *appear*, which select a stative participle. To the initial list I added attested AEP matrix verbs *(could) bear*, *(could) use*, *require* (Strelluf 2022, Tenny 1998) and a small set of clear synonyms to verbs in the initial list (*beg*, *beseech*, *demand*, *enjoin*, *entreat*, *exhort*, *implore*, *importune*, *petition*, *request*). These revisions to the initial list yielded a total of 142 potential matrix verbs to search for. For each Raising and Control verb, the following search was run in COCA, with results grouped by lemma (3). Because ECM verbs can only be followed by an embedded passive when passivized, the search context was restricted accordingly (4).

(3) VERB_vv to be _v?n
(4) VERB_v?n to be _v?n

For every search, each matrix verb/participle pair was noted alongside the number of hits.

An R script (R Core Team 2020) was used to transform these results into a weighted bipartite network. Every matrix verb attested in the EP in COCA was included as a column (n=116), with every attested participle included as a row (n=3301). Unattested matrix verbs were excluded from the network.[1] The number of hits for a given matrix verb/participle pair was used as the weighting. Verb productivity was derived by calculating degree centrality; this was effectively the number of hits in COCA for the matrix verb in the EP. The measure ranged from 1 (*concede*, among others) to 43115 (*need*). Beckett's (2016) community detection algorithm was used in the *bipartite* R package (Dormann et al. 2008) to derive semantic likeness. This yielded nine categories of verbs of varying broadness, which can be found in Table 1. Note that group labels are my interpretation of the output. As seen, the semantic likeness categories largely seem to make sense. The core attested matrix verbs *need*, *want*, and *like*, as well as weakly attested matrix verbs *deserve*, *love*, and *require* are in two groups related to volition and necessity.

Group and Approximate Likeness	**Verbs in Group**
A: Volition, Sentiment, Intention	Afford, agree, ask, avoid, bear, beg. care, choose, clamor, decide, decline, demand, deny, deserve, elect, enjoy, favor, hate, hesitate, jump, like, love, manage, mind, move, offer, opt, pledge, prefer, push, risk, struggle, vow, want, wish, imagine, suppose
B: No future event	Refuse, figure, stop
C: Permissivity of future event	Apply, come, mean, negotiate, petition, set out, try
D: Stating, Asserting	Attempt, bother, claim, determine, hope, know, press, profess, promise, report, resolve, seek, strive, swear, assume, consider, declare, deem, estimate, find, perceive, repute, rumor, say, see, show, think, believe, happen, prove, tend
E: Working toward future event	Aim, scramble, remain
F: Necessity	Admit, concede, flock, request, require, serve, sign, make, need
G: Forget	Forget
H: Arranging for future event that will happen	Arrange, begin, discuss, force, intend, learn, proceed, stay, threaten, undertake, vote, wait, allow, hold, judge, continue, fail, start
I: Arranging for future event that may not happen	Plan, propose, rush, stand, cause, expect, project

Table 1: Semantic likeness categories from weighted bipartite network model.

[1] This decision generally meant that extremely infrequent verbs such as *importune* were excluded. However, it also revealed the limitations of using the EP to approximate the AEP. *Could use*, which is attested as a potential matrix verb in Tenny (1998), is unattested in the EP in COCA and was thus excluded. This is part of a general underattesting of other verbs like *prefer* and *enjoy*, for which the more common embedded passive construction would be as in (i). Both the exclusion of unattested verbs and underattesting of others will have influenced the derived factors used here.

(i) The cat enjoys being petted.

To my knowledge it has not yet been noted in the literature that the AEP can vary with the context in (i), in contrast to the well-documented variation with the EP.

3 Application to Acceptability Judgement Survey Data

Thus far, a bipartite network has been used to derive language-internal factors associated with a set of potential AEP matrix verbs. It remains to be seen, however, whether these factors are useful for a variationist analysis. To test this requires data on the AEP which is coded for matrix verb. Because in production data only examples with matrix *need* occur at a high enough frequency for quantitative analysis (see Duncan 2019, Strelluf 2020, 2022), I test acceptability of the AEP with a variety of matrix verbs in a large-scale judgement survey.

3.1 Methods

Three test sentences approximating the AEP surface form were constructed for each of the 116 verbs attested as EP matrix verbs in COCA. To minimize effects of other language-internal factors influencing acceptability ratings, stimuli for most verbs were in present tense, positive declarative contexts (5a). Where necessary, manner adverbs or *by*-phrases were included to force an eventive reading of the test sentence. Exceptions to the present declarative context included ECM verbs, which were themselves passivized to enable embedding a passive (5b), and the pair *bear* and *stand*, which were preceded by the modal *could* (5c-d). Thirty-seven fillers were constructed from the canonical EP, as well as transitive constructions with clefted objects (6). These fillers were designed to be acceptable to all speakers, regardless of if they accept the AEP.

(5) a. This paperwork requires completed.
    b. The dangerous package was determined produced by a disgruntled employee.
    c. The toddler couldn't bear parted from her teddy bear.
    d. His room could stand cleaned up a bit.
(6) a. The fence needs to be painted.
    b. Here is the form that I need filled out.

In total, there were 385 stimuli that tested acceptability on a five-point Likert scale. These were divided into four surveys, (130 test items/survey for three, 133 test items in one). Each survey included the constructed AEP sentences with *need*, *want*, and *like* as matrix verbs, one quarter of the remaining constructed AEP stimuli, and the 37 filler sentences expected to be rated as acceptable.

Survey distribution was in line with a larger aim (not discussed here) to observe geographic patterning in acceptability of the AEP in both the United States and United Kingdom. The surveys were distributed via Prolific Academic, a distribution platform for academic research surveys which enables the fair payment of participants (£2.50 for an estimated 20 minute task, or £7.50/hour). In keeping with best practice outlined by Wood et al. (2019), each of the four surveys had 500 participants from the US (2000 total); given the relative population and area difference of the UK, each survey had 100 participants (400 total) from this country. Prolific enables researchers to prescreen participants for demographic characteristics; all participants were required to be native English speakers, fluent English speakers, to have been born in one of the countries, and to currently live in the same country. Because the AEP is particularly attested in Scotland and Northern Ireland (Strelluf 2022), which have relatively low populations, the UK sample oversampled these regions such that 35% of participants came from them. Spelling and lexical items in the sentences were adjusted as necessary for each country. Survey items were presented in a randomized order. Altogether, 225,000 grammaticality judgements were obtained for analysis (187,500 US/37,500 UK)

3.2 Results

In reporting survey results, my primary goal is to evaluate the utility of the language-internal factors derived computationally from the bipartite network. The main way to do so is to test for whether the verb productivity and semantic likeness factors significantly influence acceptability ratings. Of course, finding a correlation between a factor and acceptability ratings does not mean that the effect of the factor is real or meaningful in and of itself. When using significance testing to evaluate these factors, it is thus imperative—as it is in the course of any data analysis—to confirm that any correlations are reasonable. Bearing in mind that the AEP appears to be most commonly used with *need*

as a matrix verb, we expect that if the derived factors are viable, they will point toward test sentences with matrix *need* being particularly highly rated by participants. Thus, if verb productivity is significantly correlated with acceptability ratings, we expect a positive correlation because *need* has the highest verb productivity value in the set. Likewise, if semantic likeness is significantly correlated with acceptability ratings, we expect to see the group including *need* to be favored relative to other groups.

Given that the acceptability judgement survey was distributed across the entirety of the US and UK, it is to be expected that many or most participants do not actually have the AEP in their grammars. However, the questions of how acceptability of the AEP is constrained and which matrix verbs are viable in the construction are only relevant to speakers who have the AEP. Participants were therefore filtered by their ratings for test sentences with matrix *need*. Because this is the most common context in which the AEP occurs, a high rating for such sentences was taken to be the most viable proxy for having the AEP available. Participants whose average rating for test sentences with matrix *need* was greater than 3.67 (in effect, those who rated at least two of the three test sentences as acceptable on the five-point scale) were kept for further analysis (210 UK participants yielding 19,686 ratings, 988 US participants yielding 92,631 ratings).

To evaluate the utility of the derived factors, I take an intentionally simple approach to the data. I use linear mixed effects regression with the *lme4* package in R (Bates et al. 2015) to test whether language-internal factors influence acceptability ratings. Verb productivity (log-scaled), semantic likeness group, and SYNTACTIC TYPE of the matrix verb (Raising, Control, or ECM) were included as fixed effects, with participant and test items as random intercepts and the order in which items were presented included as a by-participant random slope. Verbs like *need*—Raising verbs in semantic likeness Group F—were used as a baseline condition. Any effects of language-external factors or interactions between language-internal and/or language-external factors were set aside for future research. Below, I focus my reporting of model results solely on the verb productivity and semantic likeness constraints.

As seen in Figure 3, there is a clear positive effect of verb productivity on acceptability ratings. The effect is quite similar among both US and UK participants (UK $\beta = 0.096$, US $\beta = 0.082$, $p < 0.0001$ for both). Note the distribution of ratings; most verbs received high acceptability ratings from some participants. The verb productivity effect is therefore not simply a matter of *need, want*, and *like* being highly productive, highly rated outliers. Rather, the verb productivity effect reflects a continuous distribution of ratings.

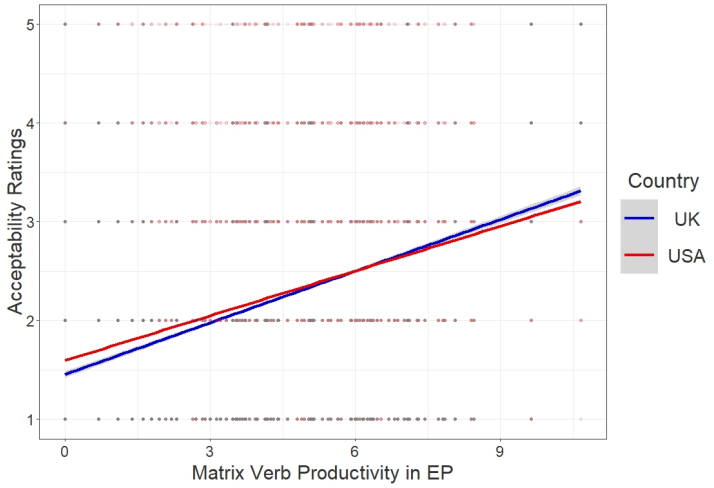

Figure 3: Effect of verb productivity (network degree centrality) on AEP acceptability.

In addition, there were clear effects of semantic likeness group using the factor as originally derived (Figure 4). As seen, Group F, which includes *need*, is rated highest, followed by Group D and Group A (which includes *want* and *like*). Again, the US and UK participants rated the sentences quite similarly across the semantic likeness groups.

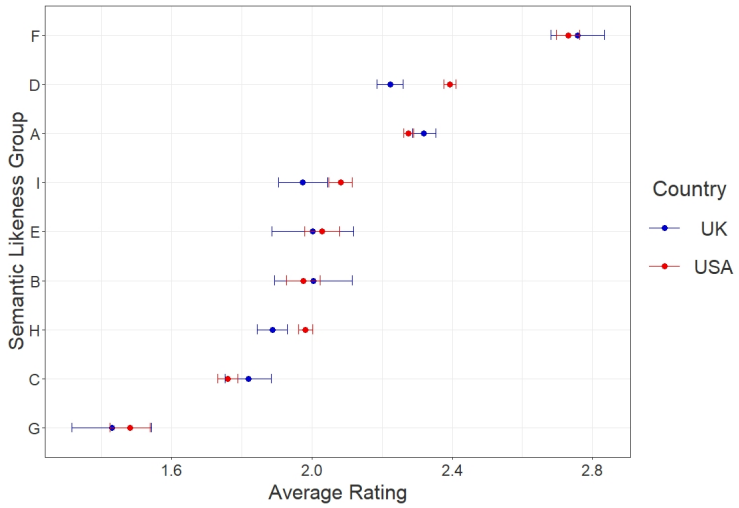

Figure 4: Effect of semantic likeness group on AEP acceptability.

The factor in this original form significantly influenced acceptability ratings. However, the overall effects of individual groups were similar after accounting for the other model effects. This factor was therefore recoded to combine semantic likeness groups with similar effect sizes. The nine factor levels were reduced to four: Groups A/F (now including *need*, *want*, and *like*), Groups B/C (permissivity or refusal of a future event), Groups D/E (assertion or working toward a future event), and Groups G/H/I (*forget* and arrangements for a future event). The model with this recoded factor (using Raising verbs in Groups A/F as a baseline condition) found that Groups A/F received significantly higher acceptability ratings than Groups B/C (UK: $\beta = -0.279$, $p = 0.0116$; US: $\beta = -0.317$, $p = 0.0014$), Groups D/E (UK: $\beta = -0.295$, $p = 0002.$; US: $\beta = -0.196$, $p = 0.0055$), and Groups G/H/I (UK: $\beta = -0.450$, $p < 0.001$; US: $\beta = -0.392$, $p < 0.0001$).

4 Discussion and Conclusions

As noted above, my primary goal in reporting on survey results here is in evaluating the utility of computationally derived language-internal factors. However, it is worth briefly mentioning what these results indicate about the AEP. Firstly, the continuous distribution of acceptability ratings indicates that the AEP is likely much more productive than previously assumed with respect to matrix verbs licit in the construction. The semantics of the matrix verbs do indeed constrain acceptability (cf. Murray and Simon 1999); a broad class of matrix verbs related to necessity, volition, and sentiment are favored as matrix verbs compared to other possible verbs. Verbs that are more productive matrix verbs in the EP are more acceptable as AEP matrix verbs. These factors alone are sufficient to explain *need* and *want* being strongly attested: both are in the broad semantic class that favors acceptability and are the two most productive EP matrix verbs. The factors are less able to explain why *like* has been attested as a third canonical matrix verb (Edelstein 2014, Murray and Simon 2002), as it is roughly as productive in the EP as other verbs in the favoring semantic class such as *deserve*. Given the relative rarity of the AEP in production, this may simply be a matter of a linguist happening to be present for usage with matrix *like*, but not usage with another matrix verb.

Below, I assess the approach of using a bipartite network model to derive language-internal factors and consider future applications of this technique. In doing so, it is important to remember that the factors were derived independently from the judgement survey. As such, they could have been applied to production data or any other type of linguistic data quantified in a similar manner. Thus, while the evaluation below concerns usage of derived factors in experiments, it carries implications for methodology in variationist approaches to other data as well.

4.1 Evaluation of Method

Computationally derived factors are viable if, like human-defined factors, they are able to contribute

to interpreting our data in a reasonable manner. By this standard, both factors derived from a bipartite network model, whether obtained through basic network measures or application of a community detection algorithm, are a success. Verb productivity, obtained from the degree centrality measure, correlates with acceptability of the AEP such that matrix *need* and *want* are predicted to be highly rated compared to other verbs. This matches the widespread attestation of these matrix verbs in the construction. Semantic likeness, obtained through community detection, shows that the classes of matrix verbs which include *need*, *want*, and *like* are more acceptable than others. This again reflects these matrix verbs' attestation in the literature. Although the factor levels for semantic likeness include a broad range of verbs, this range includes most of the other matrix verbs occasionally attested in the AEP, suggesting that this broadness may be a feature rather than a bug.

The computationally derived factors thus reflect attested patterns of AEP usage while offering additional insight into the productivity of the construction. They are successful on a more intuitive level as well. The semantic likeness groups largely make sense: the verbs that are grouped together do seem to share some elements, and groups can be discussed rather straightforwardly in prose labels. It is evident that verbs were grouped together by semantic commonalities, which means that we were able to obtain fuzzy categories for a factor with unknown factor levels through a rigorous computational approach. Verb productivity is likewise intuitive. In a weighted network degree centrality is effectively frequency within a corpus. While a simple measure, the use of a bipartite network offered a principled reason for quantifying productivity in this way.

While this paper has demonstrated that useful language-internal factors may be derived computationally, there are two points of caution. The first is that like any language-internal factor, computationally derived factors are only viable to the extent that the factor represents a reasonable hypothesis about how variation is structured. The second consideration is while community detection is a viable method for rigorously grouping items in an otherwise fuzzy or subjective category with unknown factor levels, this approach is highly dependent on the data in the network. The choices of what to include or exclude, whether to use a weighted or unweighted model, and what data source is used to construct the network all influence the overall shape of the network, and therefore influence the properties of any factors derived from the network. In this sense, while bipartite networks may be used to rigorously categorize a fuzzy factor, fuzziness and human input are not entirely eliminated under this approach.

4.2 Outlook for Further Applications

This paper suggests that bipartite networks may be used to computationally derive language-internal factors. This is most useful when the factor in question is fuzzy, or when the factor levels are unknown. Using the AEP as a test example, I show that the factors derived from such networks can make intuitive sense, significantly correlate with linguistic data, and contribute to our understanding of a linguistic phenomenon. Although I use derived factors in the analysis of acceptability judgement data, these factors could quite straightforwardly be applied to production data.

The use of bipartite networks to derive language-internal factors need not be limited to analysis of the AEP. The success of the approach lends itself to extension to other sociolinguistic variables in which two categories can be linked to one another. Some variables are quite similar to the AEP in this regard. For example, genitives involve two nominals that are linked. Variable production of the genitive is present and conditioned by semantic factors (e.g., Naccarato et al. 2021). A straightforward extension of the approach introduced here would be to create a bipartite network linking one set of nominals to another within a corpus and then use community detection to obtain semantic likeness groups for one of the sets of nominals.

Using bipartite networks to derive language-internal factors beyond lexical semantics will be fruitful as well, as semantic factors are not the only categories that may be relevant to variationist analysis but are fuzzy or have an unknown set of factor levels. For example, variables such as English complementizer deletion are influenced by collocations of the subject and matrix verb (Tagliamonte and Smith 2005). However, determining which subject/matrix verb pairs are collocations is rather subjective. Quantitative measures obtained from weighted bipartite network linking subjects to matrix verbs within a corpus may be useful in reliably identifying collocations in a rigorous and replicable manner. At the same time, using bipartite networks to link a construction to a cluster of features (Shirtz 2019) may also prove helpful. The methods demonstrated here thus provide a

useful technique that may improve understanding of a wide range of sociolinguistic variables.

References

Bates, Douglas, Martin Mächler, Ben Bolker and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48.
Beckett, Stephen J. 2016. Improved community detection in weighted bipartite networks. *Royal Society Open Science* 3: 140536.
Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA): 520 million words, 1990-present*. Available online at https://corpus.byu.edu/coca/.
Dodsworth, Robin, and Richard A. Benton. 2019. *Language Variation and Change in Social Networks: A Bipartite Approach*. New York: Routledge.
Dormann, Carsten F., Bernd Gruber, and Jochen Fruend. 2008. Introducing the bipartite Package: Analysing Ecological Networks. *R News* 8: 8–11.
Duncan, Daniel. 2021. A note on the productivity of the alternative embedded passive. *American Speech* 96: 481–490.
Duncan, Daniel. 2019. Grammars compete late: Evidence from embedded passives. In *U. Penn Working Papers in Linguistics 25.1*, ed. A. Creemers and C. Richter, 89–98.
Edelstein, Elspeth. 2014. This syntax needs studied. In *Micro-syntactic variation in North American English*, ed. R. Zanuttini and L. Horn, 242–268. Oxford: Oxford University Press.
Labov, William, Sabriya Fisher, Duna Gylfadottír, Anita Henderson, and Betsy Sneller. 2016. Competing systems in Philadelphia phonology. *Language Variation and Change* 28: 273–305.
Milroy, Lesley. 1987. *Language and Social Networks*. New York: Blackwell.
Murray, Thomas E., and Beth Lee Simon. 2002. At the intersection of regional and social dialects: The case of like + past participle in American English. *American Speech* 77: 32–68.
Murray, Thomas E., and Beth Lee Simon. 1999. Want + past participle in American English. *American Speech* 74: 140–164.
Murray, Thomas E., Timothy C. Frazer, and Beth Lee Simon. 1996. Need + past participle in American English. *American Speech* 71: 255–271.
Naccarato, Chiara, Anastasia Panova, and Natalia Stoynova. 2021. Word-order variation in a contact setting: A corpus-based investigation of Russian spoken in Daghestan. *Language Variation and Change* 33: 387–411.
Nastase, Vivi, Rada Mihalcea, and Dragomir R. Radev. 2015. A survey of graphs in natural language processing. *Natural Language Engineering* 21: 665–698.
Orozco, Rafael, and Luz Marcela Hurtado. 2021. A variationist study of subject pronoun expression in Medellín, Colombia. *Languages* 6: 5.
R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Shirtz, Shahar. 2019. Isomorphic co-expression of nominal predication subdomains: An Indo-Iranian case study. *Journal of South Asian Languages and Linguistics* 6: 59–89.
Strelluf, Christopher. 2022. Regional variation and syntactic derivation of low-frequency NEED-passives on Twitter. *Journal of English Linguistics* 50: 39–71.
Strelluf, Christopher. 2020. Needs+PAST PARTICIPLE in regional Englishes on Twitter. *World Englishes* 39: 119–134.
Tagliamonte, Sali, and Jennifer Smith. 2005. No momentary fancy! The zero 'complementizer' in English dialects. *English Language and Linguistics* 9: 289–309.
Taylor, Ann. 2006. Treebank 2a guidelines. Available online at https://www-users.york.ac.uk/~lang22/TB2a_Guidelines.htm. Accessed 1/26/2023.
Tenny, Carol. 1998. Psych verbs and verbal passives in Pittsburghese. *Linguistics* 36: 591–597.
Wood, Jim, Kaija Gahm, Ian Neidel, Sasha Lioutikova, Luke Lindemann, Lydia Lee, and Josephine Holubkov. 2020. Mapbook of syntactic variation in American English: Survey results, 2015-2019. Ms., Yale University. Available online at http://lingbuzz.net/lingbuzz/005277.

School of English Literature, Language and Linguistics
Percy Building
Newcastle University
NE1 7RU
United Kingdom
daniel.duncan@ncl.ac.uk