A computational model of S-selection

Aaron Steven White
Johns Hopkins University

Kyle Rawlins
Johns Hopkins University

We develop a probabilistic model of S(emantic)-selection that encodes both the notion of systematic mappings from semantic type signature to syntactic distribution—i.e., projection rules—and the notion of selectional noise—e.g., C(ategorial)-selection, L(exical)-selection, and/or other independent syntactic processes. We train this model on data from a large-scale acceptability judgment study assessing 1000 English clause-taking verbs in 50 distinct syntactic frames. We find that this model infers coherent semantic type signatures.

To explore the type signatures the model discovers, we employ two case studies of verbs that take both question and declarative complements: (i) cognitive factive verbs and (ii) communicatives verbs. In case study (i), we find the model prefers a direct mapping from type signatures to syntactic frames: one type maps to only questions and the other maps to only declaratives.

This finding raises a worry that what the model discovers is syntactic in nature. To address this worry, we show in case study (ii) that the model finds type signatures that project onto both questions and declaratives; but these type signatures are only found for communicative (and experiencer object) verbs. This demonstrates that, in principle, the model could have found a comparable type signature for cognitive verbs, but it didn't. We suggest these findings militate against type-casting accounts of factives' distributional multiplicity, favoring a polysemy or ambiguity account.

Introduction. Many clause-taking verbs are *distributionally multiplicitous*—e.g., factive predicates can take both question and nonquestion complements (Hintikka, 1975; Karttunen, 1977).

(1) John didn't know {that, whether} Mary was home.

Distributional multiplicity is compatible with—but does not imply—multiplicity in S(emantic)-selection—i.e. *semantic multiplicity* (See Grimshaw 1979; Pesetsky 1982 among many others). Mismatches between distributional multiplicity and semantic multiplicity can arise due to systematic mappings from a single semantic type to multiple syntactic types, as with concealed questions (2), or due to apparent 'noise'—i.e. selectional behavior that is not obviously related to semantic type—as with null complement anaphora (3).

(2) John knows {the time, what time it is}.

(3) John {knows, *recognizes}.

With sufficient data about the surface selectional behavior of a language's verbs, recent advances in computational modeling should in principle allow inference from a verb's syntactic distribution to its semantic type signature(s). We investigate this possibility here.

Contribution. We develop a probabilistic model of S-selection that encodes both the notion of systematic mappings and selectional noise (cf. White, 2015). We train this model on data from a large-scale acceptability judgment study assessing 1000 English clause-taking verbs in 50 distinct syntactic frames. We find that this model infers coherent semantic type signatures.

To explore the type signatures the model discovers, we employ two case studies of verbs that take both question and declarative complements: (i) cognitive factive verbs and (ii) communicative verbs. In case study (i), we find the model prefers a direct mapping from type signatures to syntactic frames: one type maps to only questions and the other maps to only declaratives.

This finding raises a worry that what the model discovers is syntactic in nature. To address this worry, we show in case study (ii) that the model finds type signatures that project onto both questions and declaratives; but these type signatures are only found for communicative (and experiencer object) verbs. This demonstrates that, in principle, the model could have found a comparable type signature for cognitive verbs, but it didn't. We suggest these findings militate against type-casting accounts of factives' distributional multiplicity, favoring a polysemy or ambiguity account.

Background. The distribution in (1) could arise (i) because *know* is polysemous, combining with both Q(uestion)-type and P(roposition)-type denotations (cf. George, 2011, Ch. 4); (ii) because it combines with only one type and, e.g., the other type is covertly casted (Groenendijk and Stokhof, 1984; Heim, 1994; Ginzburg, 1995; Lahiri, 2002; Spector and Egre, 2015; Uegaki, 2015); or (iii) because *know* is ambiguous between a variant that selects one type and one that selects the other (Karttunen, 1977). Distributional multiplicity is not constrained to finite selection. Many predicates can take both finite and nonfinite complements (Pesetsky, 1991; Wurmbrand, 1998, 2014; Landau, 2000; Moulton, 2009; Stephenson, 2010; Grano, 2012; White, 2014).

(4) a. John {saw, heard} that Mary left.
b. John {saw, heard} Mary leave.

(5) a. John {hoped, likes} that Mary left.
b. John {hoped, likes} to leave.

Options (i)-(iii) are not mutually exclusive. A verb may be ambiguous between two or more expressions, each of which could select one or more semantic types. For instance, *know* could have (a) a single variant that selects each of the relevant semantic types, (b) both Q-selecting and P-selecting variants, (c) only one or the other, (d) one that is Q-selecting (but takes P when casted to Q) and one that selects nonfinite denotations, and so on. But crucially, option (ii) is only possible if some semantic type is projected onto two values of a syntactic feature—e.g. [+Q] and [-Q].

Experiment. To gather acceptability judgment data, we used a standard Likert scale task that involved judging acceptability of sentences constructing using a large set of verbs. We selected

1000 verbs using a search seeded by the lists contained in Hacquard and Wellwood 2012; Anand and Hacquard 2013; Rawlins 2013 and 50 syntactic frames, which represented different combinations of syntactic features: complementizer/WH (*that*, *for*, $\emptyset$, *whether*, *D-linked WH*), embedded tense/modality (*past*, *future*, *infinitival*, *present participle*, *bare*), matrix object (*true*, *false*), matrix PP (*to*, *about*, *none*), embedded subject (*true*, *false*), and passivization (*true*, *false*).

Sentences to be judged were constructed by fully crossing the verbs with the syntactic frames; frames were instantiated by inserting *someone* or *something* in place of DPs, *do something* and *have something* in place of infinitival VPs, and COMP+*something happen*+TENSE/MODALITY in place of CPs. Example frames (not the complete set) for two verbs are shown in (6).

(6) a. Someone decided {that, whether} something {happened, happen, would happen}.
b. Someone wanted {for someone, someone, $\emptyset$} to {do, have} something.

Passivized sentences, which all had active counterparts, were included to capture verbs that take expletive subjects without including frames with which many other verbs would be unacceptable. Verbs that only allow expletive subjects were thus predicted to only be good with passivized forms.

727 unique participants were recruited through Amazon Mechanical to rate each of the 50,000 sentences, which were bundled into lists of 50. Five ratings were obtained per sentence.

Computational model. We develop a probabilistic model that uses the above experimental data to infer (i) the semantic type signatures a verb has, (ii) how those type signatures are projected onto syntactic frames, and (iii) what feature of each syntactic frame the mappings in (ii) are sensitive to. Each of these objects are treated as unobserved representations that our model must discover.

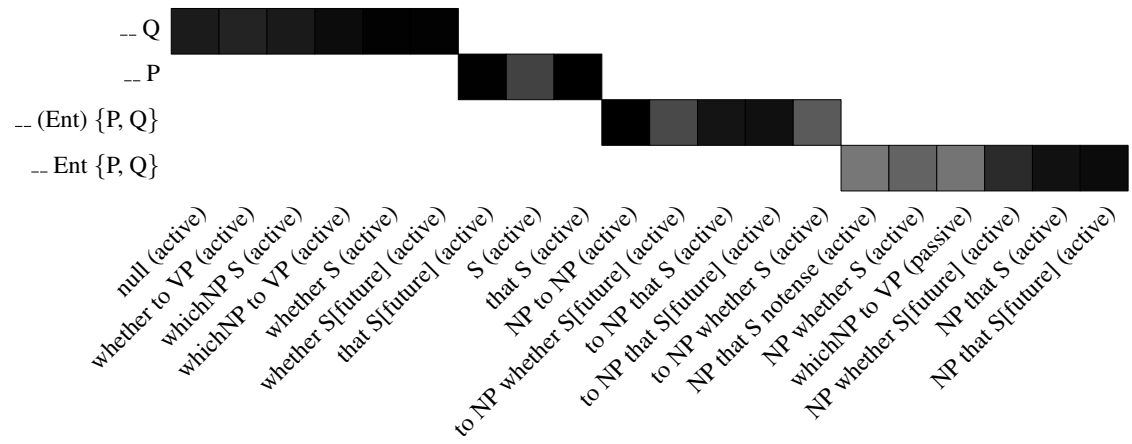
We represent objects (i)-(iii) as matrices of probabilities. Each cell s_{vt} of the semantic type signature matrix $\mathbf{S}$ corresponds to the probability that verb v has type signature t; each cell p_{tr} of the projection matrix $\mathbf{P}$ corresponds to the probability that type signature t projects onto syntactic frame r; and each cell m_{kt} of the mapping matrix $\mathbf{M}$ corresponds to the probability that type signature t is sensitive to syntactic feature k. Verbs' syntactic distributions $\hat{\mathbf{D}}$ and syntactic frames' features $\hat{\mathbf{F}}$ are also encoded as matrices of probabilities, with $\hat{d}_{vr}$ the probability that verb v can occur in syntactic frame r and $\hat{f}_{kr}$ the probability that syntactic frame r has syntactic feature k. These are related to $\mathbf{S}$, $\mathbf{P}$, $\mathbf{F}$ by the fuzzy product logic disjunctive normal forms in (7) and (8).

(7) $\hat{d}_{vr} = \bigvee_t s_{vt} \wedge p_{tr} = 1 - \prod_t 1 - s_{vt}p_{tr}$ (8) $\hat{f}_{kr} = \bigvee_t m_{kt} \wedge p_{tr} = 1 - \prod_t 1 - m_{kt}p_{tr}$

Finally, $\text{logit}(\hat{\mathbf{D}})$ is related to the experimental data $\mathbf{D}$ by a cumulative link logit (ordinal) model with parameters for each participant, and $\hat{\mathbf{F}}$ is related to hand-coded (binary) syntactic features $\mathbf{F}$ by a Bernoulli distribution. The ordinal model simultaneously controls for (a) differences in participants' use of the Likert scale and, more importantly, (b) selectional 'noise'—i.e. selectional behavior that is not obviously related to semantic type—which might arise from C-selection (Grimshaw, 1979), L-selection (Pesetsky, 1991), or an independent syntactic module (*ibid*). The Bernoulli model ensures that there is a preference for type signature-to-syntactic frame projection rules that map onto natural syntactic classes, though as we show, this preference is not so strong as to disallow projection onto competing syntactic features (e.g., C[+Q] and C[-Q]).

This model is a natural relaxation of the standard set-theoretic treatment of selection in the following sense: using elementary results from Boolean algebra (Stone, 1936), the set-theoretic treatment of semantic type can be isomorphically encoded as a Boolean ring and, for each verb, our model searches over distributions on that Boolean ring. Similarly, our model's representation of projection rules can be seen as a distribution on boolean ring homomorphisms.

One thing this specification does not yet provide is the number of type signatures the model should assume. To determine this, we use a standard method for finding the optimal number of



types in **S**, **P**, and **M**. This method involves fitting the model with different number of type signatures and assessing how well the model fits the data. Allowing for more type signatures will always allow the model to fit the data better, so models with more type signatures should be penalized in proportion to the number of type signatures, scaled by the number of verbs, number of frames, and number of syntactic features. This measure (the Akaike Information Criterion) is commonly employed in similar situations and is statistically well-founded (Gelman et al., 2013).

Results. The optimal number of type signatures determined by the above procedure is 12, which gives a lower bound on the number of type signatures needed to explain the distributional facts (and hopefully a good approximation of the true number of unique type signatures these verbs have).

This means that each verb is associated with 12 different probabilities, one for each type signature. These probabilities are interpretable as the model's certainty that a particular verb has a particular type signature. Similarly, each syntactic frame is associated with 12 different probabilities. These probabilities are interpretable as the model's certainty that a particular type signature projects onto (some syntactic features) in that frame. The above graphs show four of these 12 type signatures, to which we have assigned an interpretation on the basis of the verbs that they characterize, the model's representation of the projection rules, as well as the feature mappings. (Note that frames with less than 70% probability for the four listed types are not shown.)

Looking first to the top two rows, we see a type signature that projects onto only question frames (__ *Q*) and another that projects onto only declarative frames (__ *P*). The first of these is furthermore associated with inquisitive verbs—*wonder*, *question*, *ask*, *investigate*, *research*, *study*—and (semi)factive verbs—*know*, *figure out*, *find out*, *discover*, *realize*, *understand*, *forget*, *remember*. The second of these is associated with the same (semi)factive verbs as well as nonfactive, noninquisitive verbs—*think*, *believe*, *say*, *expect*, *claim*, *hope*. This suggests that, if this model is truly finding type signatures, (semi)factives have two distinct type signatures as opposed to one unified type signature that projects onto multiple syntactic features.

This conclusion is bolstered by the existence of the bottom two type signatures (__ *Ent* {*P, Q*} and __ *(Ent)* {*P, Q*}), which project onto frames with an internal NP or PP argument. Both of these frames project onto both question and declarative frames—the first being associated with communicatives like *tell* and the second being associated with communicatives like *say*. This suggests that our model does not just find shallow syntactic generalizations but that it can recognize deeper explanatory variables (e.g., type signatures) that bind together distinct syntactic features.

Conclusion. We have provided an unusually large-scale experiment on the selectional behavior of clause-embedding predicates, together with a novel computational model that successfully infers interpretable semantic type signatures from the distribution of verbs in the experimental data.

References

Anand, Pranav, and Valentine Hacquard. 2013. Epistemics and attitudes. *Semantics and Pragmatics* 6:1–59.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian data analysis*. CRC press.

George, Benjamin Ross. 2011. Question embedding and the semantics of answers. Doctoral Dissertation, University of California Los Angeles.

Ginzburg, Jonathan. 1995. Resolving questions, II. *Linguistics and Philosophy* 18:567–609.

Grano, Thomas Angelo. 2012. Control and restructuring at the syntax-semantics interface. Doctoral Dissertation, University of Chicago.

Grimshaw, Jane. 1979. Complement selection and the lexicon. *Linguistic Inquiry* 10:279–326.

Groenendijk, Jeroen, and Martin Stokhof. 1984. On the semantics of questions and the pragmatics of answers. *Varieties of formal semantics* 3:143–170.

Hacquard, Valentine, and Alexis Wellwood. 2012. Embedding epistemic modals in English: A corpus-based study. *Semantics and Pragmatics* 5:1–29.

Heim, Irene. 1994. Interrogative semantics and Karttunens semantics for know. In *Proceedings of IATL*, volume 1, 128–144.

Hintikka, Jaakko. 1975. Different Constructions in Terms of the Basic Epistemological Verbs: A Survey of Some Problems and Proposals. In *The Intentions of Intentionality and Other New Models for Modalities*, 1–25. Dordrecht: D. Reidel.

Karttunen, Lauri. 1977. Syntax and semantics of questions. *Linguistics and philosophy* 1:3–44.

Lahiri, Utpal. 2002. *Questions and answers in embedded contexts*. Oxford University Press.

Landau, Idan. 2000. *Elements of control*, volume 51. Springer.

Moulton, Keir. 2009. Natural selection and the syntax of clausal complementation. Doctoral Dissertation, University of Massachusetts, Amherst.

Pesetsky, David. 1982. Paths and categories. Doctoral Dissertation, MIT.

Pesetsky, David. 1991. Zero syntax: vol. 2: Infinitives.

Rawlins, Kyle. 2013. About 'about'. In *Semantics and Linguistic Theory*, volume 23, 336–357.

Spector, Benjamin, and Paul Egre. 2015. A uniform semantics for embedded interrogatives: An answer, not necessarily the answer. *Synthese* 192:1729–1784.

Stephenson, Tamina. 2010. Control in centred worlds. *Journal of semantics* 27:409–436.

Stone, Marshall H. 1936. The theory of representation for Boolean algebras. *Transactions of the American Mathematical Society* 40:37–111.

Uegaki, Wataru. 2015. Interpreting questions under attitudes. Doctoral Dissertation, MIT.

White, Aaron Steven. 2014. Factive-implicatives and modalized complements. In *Proceedings of the 44th annual meeting of the North East Linguistic Society*, ed. Jyoti Iyer and Leland Kusmer, 267–278. University of Connecticut.

White, Aaron Steven. 2015. Information and incrementality in syntactic bootstrapping. Doctoral Dissertation, University of Maryland.

Wurmbrand, Susanne. 1998. Infinitives. Doctoral Dissertation, Massachusetts Institute of Technology.

Wurmbrand, Susi. 2014. Tense and aspect in English infinitives. *Linguistic Inquiry* 45:403–447.