(De)constructing paradigmaticity in syntax: An information-theoretic approach

Ryan Ka Yau Lai
Department of Linguistics
University of California, Santa Barbara
kayaulai@ucsb.edu

1 Purpose and motivation

The notions of *paradigmatic* and *syntagmatic relations* are central to linguistics. Traditionally, two linguistic forms are *paradigmatically* related if they fall in the same grammatical slot and can substitute for each other, and *syntagmatically* related if they occur next to each other. For example, in mainstream American English, modals *may* and *can* have a paradigmatic relationship since they share a syntactic position, as seen in the fact that they occur in similar environments yet do not co-occur: one says *I may go* or *I can go*, but not *I may can go*. By contrast, the modal *may* and perfect auxiliary *have* have a syntagmatic relationship as they co-occur in distinct syntactic positions, e.g. *They may have eaten*.

Paradigmatically related forms covering a similar semantic domain may form a closed set, or *paradigm*. Paradigms are well-studied in morphology. For example, English adjective inflections *-er* (comparative), *-est* (superlative), and -∅ (no suffix, positive) have all the hallmarks of a clear paradigm: They appear in the same morphological 'slot' (after an adjectival root) and cannot co-occur, there are no other forms that go into this slot, and they are semantically in opposition with each other. Yet paradigms are often less clear-cut in syntax: As Lehmann (2015) observes, whether a form belongs to a larger, less paradigmatic set of forms (e.g. English secondary prepositions like *within*) or a smaller, more tightly integrated set (e.g. English primary prepositions like *from*) is a matter of degree, suggesting that membership in a paradigm is often gradient rather than clear-cut.

One source of complexity that has rarely been addressed in the previous literature comes from cases where some forms belonging to similar functional domains may co-occur only occasionally, and thus appear marginally paradigmatic. For example, the negative and affirmative potential modality markers 唔 *m4* and 得 *dak1* in Cantonese resultatives are often described as though they form a paradigm (e.g. Matthews 2006), yet occasional examples of co-occurrence between the two forms do exist:

字	寫	唔	得	靚
zi6	se2	m4	dak1	leng3
character	write	NEG	DAK	Pretty

'(If one) cannot write with pretty handwriting …'[1]

By contrast, the focus marker も *mo* and case marker を *o* in Japanese are usually classified as two different particle types (focus vs. case marker), yet co-occurrence between them is rare and mostly found in writing.

Based on such cases, this study proposes a new approach for examining this issue which re-casts the conventionally categorical contrast between paradigmatic and syntagmatic relationships as a gradient information-theoretic notion, pointwise mutual information:[2]

$$PMI = \log_2 \left(\frac{P(\text{forms 1 \& 2 co} - \text{occur})}{P(\text{form 1})P(\text{form 2})} \right)$$

The more strongly negative the PMI between two forms appearing in the same constructional environment, the more paradigmatic the relationship, and if $PMI \geq 0$, it is not paradigmatic at all. Paradigmaticity between forms in this approach is thus the *degree of mutual exclusivity* between them. This allows us to describe the aforementioned cases of marginal paradigmaticity.

[1] Taken from an online forum.

[2] This approach is intended for cases of gradient mutual exclusivity only. To test for *categorical* cases, one may test for $P(\text{forms 1 \& 2 co} - \text{occur}) = 0$; cf. Stefanowitsch (2008).

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 1-5.
Amherst, Massachusetts, June 15-17, 2023

This study then extends this gradient measure to evaluate the degree to which paradigmatically related forms constitute a paradigm. This is done by taking the distribution of PMIs between any pair of members in the paradigm, then taking a summary statistic like the mean, median or maximum. This complements notions of gradient paradigmaticity from previous work (e.g. Lehmann 2015, Diewald & Smirnova 2010), which focus on factors like *semantic* dependence, by extending the notion to *formal* co-occurrence probability.

2 Proposed methodology

Before the analysis, one must first identify a set of k forms from similar semantic domains that occur in the same constructional context, e.g. Japanese postnominal particles. Each instance of the construction examined containing the forms under investigation is then extracted from a corpus and modelled as a k-dimensional Bernoulli random vector, which is 0 when a form is absent and 1 when present (ignoring ordering information).

To estimate PMI, the forms' (co-)occurrence probabilities can be estimated in multiple ways. With the maximum likelihood estimator (MLE), the probability of (co-)occurrence is just the empirical proportion of times in which form(s) (co-)occur in the corpus. Yet since zero empirical co-occurrence probabilities are common in this type of situation, MLEs can be undefined, making it hard to compare across undefined values (which can be due to small n or truly high paradigmaticity), and usual Wald tests and confidence intervals based on asymptotic normality fail.

An alternative is additive smoothing, from which one can obtain Dirichlet-based posterior intervals. Rather than smoothing word counts directly, I treat the Bernoulli random vectors as categorical random variables with 2^k categories (one for each combination of forms), so smoothing can be applied to each of these co-occurrence categories. With a sizeable form inventory, most co-occurrences will be extremely infrequent, so smoothing would be excessive using standard hyperparameter choices like $\alpha = 1$. I thus cap the maximum number of forms appearing at the maximum *attested* number M, leaving $\sum_{i=1}^{M}\binom{k}{i}$ categories, and use $\alpha = \left[\sum_{i=1}^{M}\binom{k}{i}\right]^{-1}$ to concentrate density near zero. The following section describes two applications to Japanese.

3 Case studies

Japanese final particles. I examined the most common final particles (FPs), also known as utterance particles, in Japanese using the Nagoya University Conversation Corpus (Fujimura et al. 2012). A spoken corpus was chosen since such particles are pervasive in conversation.

All but three of the FPs studied have previously been organised into paradigms (Hasunuma 2015):

Type A	Type B	Type C	Other
ka, *sa*, *wa*	*yo*, *i*	*ne*, *na*	*mon, tte, kke*

In Hasunuma's original presentation, only *ka* and *sa* are linked to (i.e. possibly precede) the Type B particles, which in turn are linked to Type C particles. She additionally includes in Type A the masculine-indexing particles *zo* and *ze*, which are excluded from this study (along with other gender-indexing particles like *kasira*) because of low frequency in the corpus.

The corpus was processed with spaCy (Honnibal & Montani 2017) to identify non-sentence-initial FP sequences before punctuation marks, excluding uses of final particles in isolation, but including uses of final particles after nominals within a sentence. Smoothed pairwise PMI estimates are shown in Figure 1.

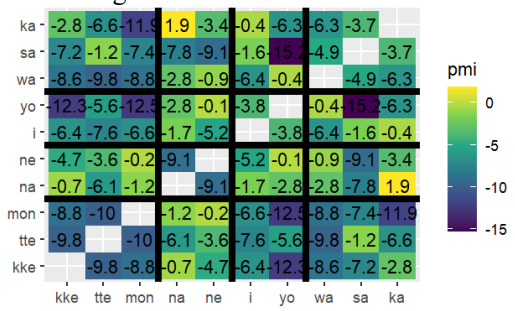

Figure 1 Heatmap of PMIs between FPs studied.

As expected, within the three paradigms (i.e. the squares along the diagonal of the heatmap), PMI values are moderately to strongly negative. *Sa* also has low PMI values with most Type B and Type C particles, consistent with Hasunuma's characterisation. Yet considering that *sa*'s PMI values with non-members of its paradigm (e.g. *kke*, *ne*) are often *smaller* than with members *ka* and *wa*, its paradigmatic strength with *ka* and *wa* may be not as strong as traditional paradigms indicate. Indeed, occasional examples of co-occurrence between *sa* and *ka* are frequently found, for example, in the combination *tte iu ka sa*, which seem not to be predicted by Hasunuma's account:

リスク	取りたくない	って	いう
risuku	tori-taku-nai	tte	iu
risk	take-want-NEG	QUOT	say
か	さー		
ka	saa		

'Is it that I don't want to take risks, as you say?' [data128, line 266]

Japanese postnominal particles. The following forms were studied (cf. Vance 1993):

Case particles	*ga*, *no*, *o*, *de*, *ni*, *kara*, *he*, *yori*, *made*, *to*
Information-structural (IS) particles	*bakari*, *dake*, *sika*, *mo*, *datte*, *sae*, *sura*, *wa*, *koso*, *kurai*, *gurai*, *hodo*, *nado*

A corpus of Wikipedia articles (National Institute of Information and Communications Technology 2014) was chosen for this study, since rare co-occurrences between some particles (e.g. *o mo*, (*minna*) *mo ga*) are more common in writing than in conversational texts. The corpus was again processed with spaCy and postnominal particles were selected. Results are shown in Figure 2. Several interesting patterns were found.

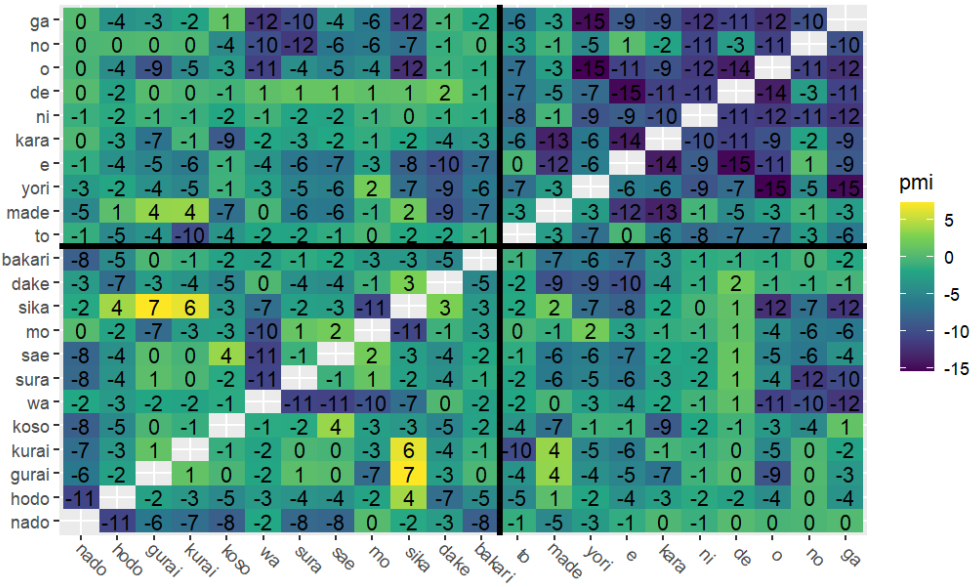

Figure 2 Heatmap of PMIs between postnominal particles studied using additively smoothed estimates. PMIs are clearly much lower among case markers than among information-structural markers.

Firstly, PMI values are very low among pairs of case markers outside *no*, *made* and *to*, with 95th percentiles of Dirichlet posteriors having mean -9.46 and maximum -4.05, indicating high paradigmaticity. The three exceptions can be readily explained: *made* can also be information-structural with a meaning like English *even*; *to* can also be conjunctive or quotative; and although *no* is always a genitive case marker postnominally, it is special among case markers since it indicates a relationship between a nominal and another nominal. IS particles as a whole are certainly not a paradigm: 95th quantiles of PMI posteriors go up to 7.93 (for *gurai* / *sika*). The information-theoretic particle *wa* is a clear exception in having very low PMIs with many other markers, likely because it is a topic marker whereas markers like *mo* and *sura* have focus-like functions. Note that most of the combinations of IS particles outside of *sika*, *mo* and *wa* are in fact unattested, despite many of the pairs having fairly high negative PMI after additive smoothing. This suggests that the unattestedness of those pairings may be due to the low frequency of both particles, rather than high paradigmaticity.

Secondly, looking across the two categories of particles, the IS particles *sika, mo, wa* have far lower PMIs with core case markers (i.e. nominative *ga*, accusative *o*) than non-core ones (e.g. ablative *kara*, locative *ni*). For *ga* / *o*, 95th percentiles of the posteriors never go above -4; for *kara* / *ni*, 5th percentiles of posteriors never go below -2. This is consistent with the observation that core case markers can have information-structural meaning (Nakagawa 2020), suggesting that they are more paradigmatically related to IS markers than non-core case markers, which primarily serve to mark the role of the nominal in the predicate, and hence

would be more likely to need to co-occur with particles specialised for IS work.

Interestingly, these patterns are not as clear when examining raw estimated joint probabilities instead of PMI. For example, the maximum likelihood estimate of the log-probability of *o* + *mo* is around -8, comparable to most non-core case particles + *mo*. Yet *o* + *mo* has a much lower PMI than non-core case + *mo*. This is because *o* is much more frequent than non-core case markers, and looking at joint probability alone ignores the much higher baseline frequency of *o*, which contributes to a relatively high frequency of *o* + *mo*. By contrast, PMI successfully considers this baseline frequency. This highlights the value of an information-theoretic approach, and aligns with work arguing for association measures that compare actual co-occurrence probability to expected ones over raw frequencies (Gries 2005), particularly for *non*-co-occurrence (Stefanowitsch 2008).

4 Conclusion

I showed that syntactic and paradigmatic relationships, hitherto seen as binary categories, can be fruitfully studied as a continuum, where occasionally co-occurring forms in similar semantic domains can still be non-categorically paradigmatic. The method was applied to two cases of Japanese particles, revealing some patterns not statable in categorical approaches. The study adds to existing computational research (Salle & Villavicencio 2019) suggesting negative PMI encodes syntactic information.

In this study, PMI is used to describe linguistic *behaviour*, rather than model some aspect of cognition. However, as Stefanowitsch (2008) has argued, if speakers find that the less frequent two forms occur compared to the baseline expectation of their co-occurrence frequency, the more noticeable it will be to language users, leading to *negative entrenchment*, which serves as direct evidence of constraints on linguistic form. Extending Stefanowitsch's argument, then, below-chance co-occurrence can lead language users to mentally represent forms as more strongly belonging to a single paradigmatic slot. Future work can test the current PMI approach against cognitive correlates of paradigmaticity, such as whether structural priming effects are stronger when a form in the prime and a corresponding form in the target have more negative PMI.

As mentioned before, paradigmaticity in previous work has mostly been explored in qualitative, semantic ways (Lehmann 2015, Diewald & Smirnova 2010). In particular, semantic opposition is the defining feature of paradigmaticity: the different members represent different values of a semantic dimension, and the presence of one member necessarily negates the rest (e.g. Sabar 2018, Diewald 2020). The PMI-based metric of this study adds a quantitative, formal dimension to paradigmaticity. Rather than constituting a fully independent dimension of paradigmaticity, however, I believe this study's conception of paradigmaticity as the degree of mutual exclusivity complements the semantic conception. In the case where two forms belong to a tight paradigm where elements encode fully incompatible values of a certain semantic dimension, then one would not expect them to co-occur at all, since co-occurrence would create a contradiction. If the semantic dimensions encoded by two forms are highly correlated such that the values encoded by the two forms are *rarely* compatible, then one may expect the forms to co-occur only rarely too. For instance, consider our second case study: PMI is higher between non-core case particles and IS particles, which have little semantic overlap, than between core case particles and IS particles, since core case particles shade into the functional domain of information structure. Future work can further examine the relationship between the semantic conception and the formal one proposed in this paper.

Other planned future work includes extending to paradigms with more members, and channelling PMI-based measures to *detect* paradigms from a parsed corpus, rather than simply testing existing paradigms. I also hope to examine how paradigmaticity varies across time and space, for example how multiple modals came to be ungrammatical in mainstream English, yet were retained in various non-hegemonic varieties (Montgomery & Nagle 1993). This will involve extending the current multivariate Bernoulli model to allow for covariates and random effects, and building more structure in the model (e.g. through statistical copulas).

References

Diewald, Gabriele & Elena Smirnova. 2010. Paradigmaticity and obligatoriness of grammatical categories. *Acta Linguistica Hafniensia* 42(1). 1–10. https://doi.org/10.1080/03740463.2010.486911.

Diewald, Gabriele. 2020. Paradigms lost – paradigms regained: Paradigms as hyper-constructions. In Lotte Sommerer & Elena Smirnova (eds.), *Constructional Approaches to Language*, vol. 27, 278–315. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/cal.27.08die.

Fujimura, Itsuko, Shoju Chiba & Mieko Ohso. 2012. Lexical and grammatical features of spoken and written Japanese in contrast: Exploring a lexical profiling approach to comparing spoken and written corpora. In *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*, 393–398.

Gries, Stefan Th. 2006. Some proposals towards a more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 191–202. https://doi.org/10.1515/zaa-2006-0209.

Hasunuma, Akiko. 2015. Shūjoshi ‘sa' no honshitsu-teki kinō: ninshiki-teki modariti to no kyōki kankei ni chakumoku shite. *Nihongo Nihon bungaku* (25). 1–27.

Honnibal, Matthew & Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Lehmann, Christian. 2015. *Thoughts on grammaticalization* (Classics in Linguistics 1). 3rd edition. Berlin: Language Science Press.

Matthews, Stephen. 2006. On serial verb constructions in Cantonese. In R. M. W. Dixon & A Y Aikhenvald (eds.), *Serial verb constructions: a cross-linguistic typology*, 69–87. Oxford: Oxford University Press.

Montgomery, Michael B. & Stephen J. Nagle. 1993. Double modals in Scotland and the Southern United States: Trans-Atlantic inheritance or independent development? *Folia Linguistica Historica* (vol. 14,1–2). 91–107. https://doi.org/10.1515/flih.1993.14.1-2.91

Nakagawa, Natsuko. 2020. *Information structure in spoken Japanese: Particles, word order, and intonation*. Berlin: Language Science Press.

National Institute of Information and Communications Technology. (2011). Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles ver.2.01.

Sabar, Nadav. 2018. *Lexical meaning as a testable hypothesis: the case of English look, see, seem and appear* (Studies in Functional and Structural Linguistics 75). Amsterdam; Philadelphia: John Benjamins Publishing Company.

Salle, Alexandre & Aline Villavicencio. 2019. Why So Down? The Role of Negative (and Positive) Pointwise Mutual Information in Distributional Semantics. *arXiv:1908.06941 [cs]*. http://arxiv.org/abs/1908.06941.

Stefanowitsch, Anatol. 2008. Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics* 19(3). https://doi.org/10.1515/COGL.2008.020.

Vance, Timothy J. 1993. Are Japanese Particles Clitics? *The Journal of the Association of Teachers of Japanese* 27(1). 3. https://doi.org/10.2307/489122.