LANGUAGE CHANGE IN HINDI FROM A PANINIAN PERSPECTIVE

Aniket Kali*[1] and Gerald Penn[1]

[1]Corresponding Author: aniket.kali@mail.utoronto.ca
[1]Department of Computer Science, University of Toronto, Toronto, Canada

We examine the development of Hindi from Sanskrit using a spectral analysis algorithm. These methods are exceptionally good at drawing out inconsistencies and discrepancies in a dataset, and are therefore well suited to questions around language change. Using a Hindi corpus developed from the Hindi/Urdu Treebank Project (Bhat et al., 2017) annotated in a Paninian (Sanskrit based) scheme, we contrast Panini's *karaka* role scheme against Hindi grammatical markers, such as postpositions. More broadly, we contribute a novel approach to uncovering and analyzing specific changes in language through time.

1. Introduction

The most reliable means of documenting language change through time is by directly considering examples drawn from documents distributed over the history of that language. There may, however, be another way. Many languages come with linguistic traditions of their own. If these traditions come to be interpreted prescriptively, the community that participates in the tradition may attempt to accommodate an inherited (semi-)formal theory to their contemporary language, either to demonstrate that the language is still licensed by the tradition, or simply because that is the only alternative that comes to mind for analyzing contemporary language. These accommodations may not be acknowledged or documented, but it may be possible to infer them from example explanations or derivations.

In this paper, we seek to investigate the possibility of such inference to document changes to Hindi syntax relative to the Paninian Sanskrit tradition. Using spectral analysis techniques, we analyze a Hindi corpus that has been annotated according to Paninian conventions by The Hindi/Urdu Treebank Project (Bhat et al., 2017). By treating Hindi as a kind of "noisy" Sanskrit, and imposing a grammar devised for Sanskrit onto it, we show that a simple technique for noise reduction can in fact reveal certain specific changes. In particular, as we describe in detail below, it has been able to reveal changes between the locative and genitive roles and their grammatical markers, as well as a failure to classify the reflexive pronoun *apna*, given the diversity of reflexives in Sanskrit. It also highlights the development of ergativity and Differential Object Marking, both of which developed well after classical Sanskrit.

While numerous mathematical approaches to language change have focused on modelling the system of syntactic change itself (Niyogi & Berwick, 1997; Kodner & Cerezo Falco, 2018), or have sought to understand its cognitive or semantic implications (Hamilton, Leskovec, & Jurafsky, 2016; Habibi, Kemp, & Xu, 2020), to our knowledge, the vast majority have never been used as a means of discovery. Those that have, some of them with very similar methods to the one used here (Belkin & Goldsmith, 2002; Belkin & Niyogi, 2003; Thaine & Penn, 2019) do not rely on a parallel linguistic tradition as a means of identifying changes. We hope this work will contribute to advancing just such an approach.

2. Grammar and Corpus

The *Ashtadhyayi*, written c. 350 BCE by Panini, is perhaps the first serious grammar of a language - in its case, Sanskrit. It is admired by linguists of all persuasions for its wide range of descriptive devices, and its degree of formal rigour (Kiparsky, 2009).

One of the work's defining traits is its use of *karaka* roles, abstract case markings with some connection to deep/semantic arguments, at least as they appear in the corpus we used, although the correspondence between theta-roles and *karaka* roles in Panini has been disputed elsewhere (Houben, 1999). Kiparsky (2009) articulates three principles of *karaka* theory: (1) every *karaka* role must be "expressed" by a morphological element, (2) no *karaka* role can be expressed by more than one morphological element, and (3) every morphological element must express something. Hindi is a morphologically poorer language than Sanskrit. As Bharati, Bhatia, Chaitanya, and Sangal (1998) have outlined, this requires an analysis using inflections, postpositions, and auxiliary words rather than one with morphology alone. Bhat et al. (2017) did such an analysis by hand in their corpus.

The Hindi-language fragment of the Hindi/Urdu Treebank Project (Bhat et al., 2017) contains 11,600 sentences and 242,600 tokens. This corpus is annotated using an extension of the Universal Dependencies scheme: roles like 'vmod' have been subdivided into 'k1' (agent), 'k2' (patient), and so on. These 'k*' roles were inspired by traditional, Paninian *karaka* roles. We evaluate the relationship between the actual distribution of Hindi morphemes and clitics and *karaka* roles using spectral analysis.

3. Singular Value Decomposition

The corpus annotation scheme of Bhat et al. (2017) uses postpositions to indicate role assignments to syntactic phrases. We define a matrix $\mathbf{A}$ that aligns postpositions with roles, in order to analyze discrepancies in their correspondence. Each row $\mathbf{x}_i$ represents a role, with n total roles. Then row $\mathbf{x}_i = [x_1, ..., x_j, ..., x_m]$ where x_j is the number of times the j-th postposition is the final element in a phrase marked with the i-th role, with m total postpositions.

We extracted a list of roles and postpositions from the corpus to construct this matrix. We only counted postpositions at the end of phrases, as these determine the role, and counted phrases without postpositions as inflected with a NULL postposition. In the case of inflection by gender and number, e.g., the possessive *ka*, which appears as *ki* (singular or plural feminine) or *ke* (plural masculine), we count it as the stem *ka*. This provides consistency, as gender and number are determined by the possessee and have no bearing on the role itself. Nonetheless, we analyze occurrences of the form *ke*+PSP as is (where PSP is the POS tag for postposition), since this is a postpositional construct in Hindi that is unrelated to the possessive. In the case of pronouns, Hindi also permits postpositions to attach as morphology, which we extract and analyze in the manner described above.

In our presentation, we retain all roles found in the corpus, but include only the top 26 postpositions, to reduce clutter. These make up about 99.9% by occurrence. In practice, including the rest has a negligible effect on our analysis. This leaves us with a total of n=65 roles and m=26 postpositions. A complete list of roles and postpositions, and their frequencies, are found in the supplementary materials.[1]

Singular-Value Decomposition (SVD) takes a co-occurrence matrix as input and returns component vectors that capture the most salient dimensions of the dataset, as measured by variance, together with weights, called singular values, that describe the importance of each dimension. Here, we present only the dimensions corresponding to the six largest singular values and project the vectorized roles onto them pairwise in a series of two-dimensional visualizations. We used `scikit-learn`'s implementation of SVD.

The j-th component in a vector represents the relevance of the j-th postposition. Once normalized, entries with an absolute value close to 1 are the most relevant, and those close to 0, irrelevant. Negative entries indicate an anti-correlation, which can still be very relevant.

Examining the projection of roles onto components outlines the relationship between particular roles and postpositions. Components with a single dominant postposition should attract roles highly related to the functional use of that postposition. Likewise, components containing multiple postpositions may either attract multiple roles, or a single multi-purposed role, indicating an analytical gap.

4. Results

As stated above, the most relevant parts of a component are those entries with values furthest from 0. We provide the first three postpositions ranked this way, in Table 1. These entries already tend rapidly toward 0, which is where the vast majority of entries lie anyhow. As we will explain, there are particular reasons that these postpositions appear on the components that they do, and that certain roles associate with them.

[1] https://hdl.handle.net/1807/117650

Table 1. The first six components from our spectral analysis. The three most extremal postpositions of each component are given, as determined by the absolute value of their singular value.

Component	Postposition 1	Postposition 2	Postposition 3
c_0	*ka* (0.97)	NULL (0.21)	*ne* (0.06)
c_1	NULL (0.9)	*ne* (0.32)	*ka* (-0.22)
c_2	*mein* (0.92)	*par* (0.37)	*tak* (0.01)
c_3	*ne* (0.86)	*ko* (-0.41)	NULL (-0.24)
c_4	*ko* (-0.88)	*ne* (-0.36)	NULL (0.25)
c_5	*ke liye* (0.99)	*se* (-0.11)	*ke karan* (0.0)

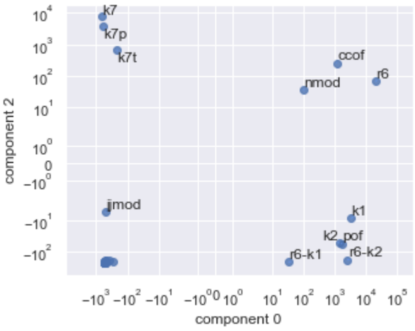

Figure 1. Projection onto components 0 and 2.

We consider roles mapped onto the components c_0 and c_2 first, shown in Figure 1. Each dot is a role, where the most relevant roles are those furthest away from the main cluster. We do not annotate the main cluster for this reason in any figure, although we do provide a list of cluster roles in the supplementary materials. It is worth noting that frequency plays a part in where roles land: a very frequent role can end up seeming more relevant than it is, with a weak correlation exaggerated by its prevalence. On the other hand, SVD remains exceptionally good at highlighting lower frequency roles that are actually relevant.

Considering c_2 first, the roles k7t (location in time), k7p (location in space) and k7 (location elsewhere) as the most prominent. Given that c_2 is mainly composed of *mein* (in) and *par* (on), this makes total sense: the locative postpositions relate to the locative roles. Considering the other five components, it is striking that neither *mein* nor *par* are relevant in any. Nor, as we will see, are the locative roles as a group so clearly related to any other component. While this is a more straightforward case, it also reinforces the empirical validity of this analysis.

We consider c_0 next, where the role r6 (genitive) and its variant r6-k2 (complex predicate with object) dominate, accompanied by k1 (agent), likely due to sheer frequency. Notice that c_0 is composed primarily of *ka* (possessive marker), along with NULL, albeit more weakly. The relationship between *ka* and r6 is clear, with the effect of relating the possessive marker to the genitive role.

r6 appears with the NULL marker in the corpus almost exclusively with the reflexive pronoun *apna*, which is marked in the corpus as r6, as are other possessive pronouns. The inchoate category of "reflexive pronoun" only entered linguistics in the late 19th century, partly due to comparative evidence from Sanskrit (Orqueda, 2019). Other possessive determiners do exist in Sanskrit, although historically it has always been more common to use the genitive case of the corresponding pronoun, which thus exhibits no agreement with its head noun. The reflexives stand out in Sanskrit by being so heterogeneous in any account, including the Paninian one, as they include an indeclinable (*swayam*), a noun (*atman*), a declinable determiner (*svah*) that, unlike every other possessive declinable determiner in Sanskrit, is not depronominal, and in certain cases, the verbal middle voice (known to the traditional grammarians as *atmanepada*, arguably translatable as the reflexive voice). Against this backdrop, just about any choice of label for Hindi's *apna* would be a stretch, and indeed the r6 label belies the fact that *apna* is declinable and, if it is de(pro)nominal at all, not transparently so.

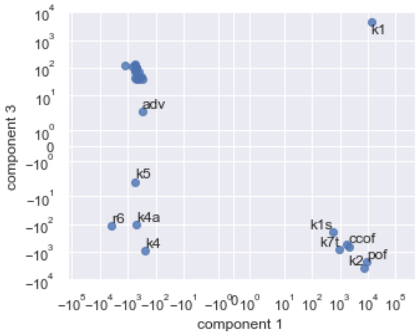

Figure 2. Projection onto components 1 and 3

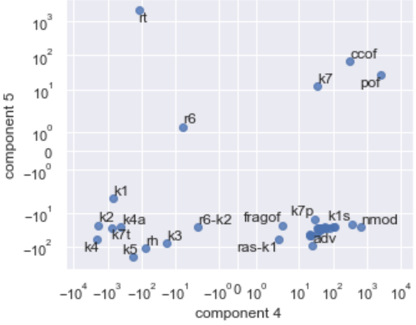

Figure 3. Projection onto components 4 and 5

Before we discuss the next group of components, we find it necessary to provide some background on ergativity in modern Hindi. In Hindi, the subject of an intransitive clause and the object of a transitive clause are treated the same way, whereas the subject of a transitive clause is marked with *ne*. But Hindi is split-ergative: we only see the ergative pattern in the preterite and perfect tenses (Verbeke & De Cuypere, 2009). The origin of this construction is an open question with differing hypotheses, but crucially, it lacks any precedent in Sanskrit's case system itself (Anderson, 1977; Butt & King, 2004; Verbeke & De Cuypere, 2009), presenting a problem for annotating Hindi using a Paninian scheme.

As for components c_1, c_3, and c_4 (Figures 2 and 3), what is immediately clear is that there exists some relationship between NULL, *ne* (ergative marker), and *ko* (to). These postpositions are almost entirely isolated to within this component group, with the exception of NULL in c_0, which we have already addressed. Looking at Figure 2, k1 and k2 (patient) are consistently prominent throughout. This is rather conspicuous, as *ne* almost exclusively occurs with k1 to mark ergative agents, and *ko*, primarily with k2 but rarely k1. NULL however occurs frequently with both roles, and must therefore logically serve to unify them, although this alone does not explain the linguistic relationship between *ne* and *ko*.

We believe the solution comes through a phenomenon known as Differential Object Marking (DOM). DOM is a process of marking animacy and definiteness in a language. It is present in Hindi, and developed alongside the modern language - well after Sanskrit. In Hindi, its presence is typically expressed with *ko*, and its absence, with NULL (Montaut, 2018). This latter point is crucial, as we can relate it to ergativity: NULL marks both non-ergative agents, and non-animate or non-definite objects. In other words, the development of ergativity and DOM in Hindi - phenomena both not present in Sanskrit - have come to intersect through NULL.

The component c_4 also provides brief additional insight into role mergers connected with *ko*. As Butt and King (2004) note, *ko* is a postposition of many forms, covering the accusative (including DOM as above), dative, and experiencer or dative passive. This is precisely what we observe in Figure 2, with k2, k4 (recipient), and k4a (experiencer), respectively present on the negative side of c_4 where *ko* has a strongly negative singular value. Additionally, k7t appears in this group, because temporal locatives in Hindi can also be marked with *ko*.

Finally, we examine c_5 in Figure 3. The postposition *ke liye* (for) almost always occurs under rt (purpose), whereas *se* (from) occurs primarily with k5 (source). The relationship between these two is rather one-sided: *se* is a rather versatile postposition, and occasionally occurs under rt, but we never observe *ke liye* as expressing k5. *se*, moreover, is actually the third most frequent postposition on the role rt, after *ko*. A *post hoc* analysis of the corpus reveals that under the role rt both postpositions are used to mark adverbials in Hindi, demonstrating a convergence of one aspect of *se* with *ke liye* in this domain.

The reader familiar with Universal Dependencies will notice that we have not commented on nmod (noun participle), pof (part of relation), or ccof (coordination or conjunct), despite their prevalence. This is because in this corpus they either serve as catch-all roles, or annotate more functional aspects of grammar. They have less to do with the *karaka* scheme (Begum et al., 2008; Tandon, Chaudhry, Bhat, & Sharma, 2016), and in turn the aim of this work.

5. Discussion

All three of our findings can be corroborated by existing knowledge among area specialists in Hindi syntactic innovations. This suggests that spectral analysis relative to a prescriptive grammatical structure in which a descendant language is viewed as a noisy version of its ancestor has some promise as a means of discovering as-yet unknown changes.

The success of this work has relied on the alignment of several stars. First, we picked a language pair with a very mature and well-documented linguistic tradition; Sanskrit is the most analyzed language in the world after only modern English. Second, the Hindi/Urdu Treebank Project methodically and, to our great relief, electronically annotated enough Hindi data that SVD could reveal certain anomalies. But most importantly, they did so in accordance with certain aspects of the ancient Paninian tradition that made the application of our proposed method straightforward. There are other approaches to Hindi grammatical analysis. Current analyses of modern Hindi are rooted in Prakritic grammars dating back to 1698 at the earliest (Bhatia, 1987), but these grammars come from a colonial tradition based in the Western classical paradigms of Greek and Latin.

It is an open question as to whether the method proposed here can be extended to pairs of cognate languages other than direct ancestor-descendant pairs. There are a variety of more sophisticated methods for spectral analysis than SVD, the algorithms for which were worked out in the 1960s. Some of them would perhaps reveal anomalies between more diverse pairs, or more subtle anomalies with ancestors and descendants. Having located an anomaly, its interpretation furthermore remains non-trivial. We relied on the assistance of Professor Miriam Butt, an area specialist, who was immediately able to interpret what the anomalies were. The present authors would not have been able to do this themselves, even with considerable effort.

It should also be noted that we did not uncover every change from Sanskrit to Hindi, nor even every noteworthy change by conservative estimates, such as the development of the dative passive in Hindi. All of these shortcomings remain areas for further investigation.

Acknowledgements

Our many thanks to Professors Miriam Butt and Dipti Sharma for their input and feedback on the development of this work, and the reviewers for their comments.

References

Anderson, S. (1977). On mechanisms by which languages become ergative. *Mechanisms of Syntactic Change*, 317-363.

Begum, R., Husain, S., Dhwaj, A., Sharma, D. M., Bai, L., & Sangal, R. (2008). Dependency annotation scheme for Indian languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing* (p. 722-726).

Belkin, M., & Goldsmith, J. (2002). Using eigenvectors of the bigram graph to infer morpheme identity. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (pp. 41–47).

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*(6), 1373-1396.

Bharati, A., Bhatia, M., Chaitanya, V., & Sangal, R. (1998). Paninian grammar framework applied to English. In *South Asian Language Review* (Vol. 8). Creative Books.

Bhat, R., Bhatt, R., Farudi, A., Klassen, P., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., Vaidya, A., Vishnu, S. R., & Xia, F. (2017). The Hindi/Urdu treebank project. In *Handbook of Linguistic Annotation* (pp. 659–697). Springer.

Bhatia, T. (1987). *A history of the Hindi grammatical tradition: Hindi-Hindustani grammar, grammarians, history and problems.* Brill.

Butt, M., & King, T. (2004). The status of case. In V. Dayal & A. Mahajan (Eds.), *Clause Structure in South Asian Languages* (p. 153-198). Springer.

Habibi, A., Kemp, C., & Xu, Y. (2020). Chaining and the growth of linguistic categories. *Cognition*, *202*(104323).

Hamilton, W., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proc. 54th ACL* (pp. 1489–1501).

Houben, J. (1999). 'Meaning statements' in Panini's grammar: on the purpose and context of the Ashtadyayi. In *Studien zur Indologie und Iranistik* (Vol. 22, p. 23-54).

Kiparsky, P. (2009). On the architecture of Panini's grammar. In *Sanskrit Computational Linguistics* (p. 33–94). Springer.

Kodner, J., & Cerezo Falco, C. (2018). A framework for representing language acquisition in a population setting. In *Proc. 56th ACL* (pp. 1149–1159).

Montaut, A. (2018). The rise of differential object marking in Hindi and related languages. In *Diachrony of differential argument marking* (p. 281–313). Language Science Press.

Niyogi, P., & Berwick, R. (1997). A dynamical systems model for language change. *Complex Systems*, *11*(3).

Orqueda, V. (2019). *Reflexivity in Vedic.* Brill.

Tandon, J., Chaudhry, H., Bhat, R., & Sharma, D. M. (2016). Conversion from Paninian karakas to universal dependencies for Hindi dependency treebank. In *Proc. 10th Linguistic Annotation Workshop, 54th ACL* (pp. 141–150).

Thaine, P., & Penn, G. (2019). Vocalic and consonantal grapheme classification through spectral decomposition. In Y. Haralambous (Ed.), *Graphemics in the 21st century* (pp. 367–386). Fluxus.

Verbeke, S., & De Cuypere, L. (2009). The rise of ergativity in Hindi: Assessing the role of grammaticalization. *Folia Linguistica Historica*, *43*, 367-389.