Surprisal and Interference Effects of Case Markers in Hindi Word Order

Sidharth Ranjan
IIT Delhi
sidharth.ranjan@cse.iitd.ac.in

Rajakrishnan Rajkumar
IISER Bhopal
rajak@iiserb.ac.in

Sumeet Agarwal
IIT Delhi
sumeet@iitd.ac.in

Abstract

Based on the Production-Distribution-Comprehension (PDC) account of language processing, we formulate two distinct hypotheses about case marking, word order choices and processing in Hindi. Our first hypothesis is that Hindi tends to optimize for processing efficiency at both lexical and syntactic levels. We quantify the role of case markers in this process. For the task of predicting the reference sentence occurring in a corpus (amidst meaning-equivalent grammatical variants) using a machine learning model, surprisal estimates from an artificial version of the language (*i.e.*, Hindi without any case markers) result in lower prediction accuracy compared to natural Hindi. Our second hypothesis is that Hindi tends to minimize interference due to case markers while ordering preverbal constituents. We show that Hindi tends to avoid placing next to each other constituents whose heads are marked by identical case inflections. Our findings adhere to PDC assumptions and we discuss their implications for language production, learning and universals.

1 Introduction

Language universals encode distributional regularities across languages of the world. This study is motivated by the well known correlation between case marking and increased word order flexibility (Sapir, 1921; Blake, 2001), often expressed as an *implicational universal*[1]. The origin of such universals has been the topic of a long-standing debate in linguistics and cognitive science (Fedzechkina et al., 2012). As the cited work expounds, one view is that language universals emerged due to constraints specific to the system of language and not related to the other cognitive faculties (Chomsky, 1965; Fodor, 2001). Another view is that languages evolved over time as a consequence of cognitive mechanisms and pressures linked with language use. Thus, cognitive biases related to processing (Hawkins, 2004), learnability (Christiansen and Chater, 2008) and communicative efficiency (Jaeger and Tily, 2011) have been proposed as underlying the systematic similarities and divergences between natural languages.

The Production-Distribution-Comprehension (PDC) account of language processing proposed by MacDonald (2013) is an integrated theory of language production and comprehension that seeks to connect language production with typology and comprehension. It is broadly in the spirit of the second view regarding linguistic universals described above and posits production difficulty as the *sole* factor influencing linguistic form. Hence it is an unconventional approach, contrasting radically with alternative accounts of language use in which language forms are shaped by constraints on language acquisition processes or considerations of facilitating language comprehension for the listeners. Based on PDC assumptions, we formulated two hypotheses linking processing efficiency, case marking, and word order choices at the level of individual speakers (as opposed to the population level) in Hindi, a language having predominantly SOV word order. Hindi has a rich system of case markers along with a relatively flexible word order (Agnihotri, 2007; Kachru, 2006) and thus adheres to the implicational universal stated at the outset.

The PDC principle *Easy First* stipulates that more accessible words are ordered before less accessible words. Accessibility of a word is influenced by its ease of retrievability from memory.

[1] "Given X in a particular language, we always find Y" where X and Y are characteristics of the language (Greenberg, 1963).

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 30–42
Minneapolis, USA, June 7, 2019. ©2019 Association for Computational Linguistics

Inspired by the stated PDC principle, our first hypothesis is that Hindi tends to optimize for processing efficiency at both lexical and syntactic levels. We investigate the role of case markers in this process by comparing the processing efficiency of natural Hindi and an artificial version of Hindi without case markers. Based on the PDC principle of *Reduce Interference*, our second hypothesis is that Hindi orders constituents such that phonological interference caused by case marker repetition is minimized. Interference is the idea that entities with similar properties (like form, meaning, animacy, concreteness and so forth) cause processing difficulties when they occur in proximity. A long line of research attests the role of interference in both production (Bock, 1987; Jaeger et al., 2012) and comprehension (Van Dyke and McElree, 2006; Van Dyke, 2007).

In order to test the stated hypotheses, we deploy a machine learning model to predict the reference sentence occurring in the Hindi-Urdu TreeBank (HUTB) corpus (Bhatt et al., 2009) of written text[2] (Example 1a below), amidst a set of artificially created grammatical variants expressing the same proposition (Examples 1b-1c). Case markers are shown in bold for illustration purposes.

(1) a. isse pehle jila upabhokta adalat **ne**
this before district consumer court ERG
28 April 1998 **ko** apne faisle **mẽ**
28 April 1998 ON own decision LOC
company **ko** nirdesh diyaa thaa ki...
company DAT direction gave COMPL

Earlier, the District Consumer Court had directed the company in its decision on April 28, 1998 that ...

b. isse pehle jila upabhokta adalat **ne** apne faisle **mẽ** 28 April 1998 **ko** company **ko** nirdesh diyaa thaa ki...

c. jila upabhokta adalat **ne** isse pehle apne faisle **mẽ** 28 April 1998 **ko** company **ko** nirdesh diyaa thaa ki...

The variants above have two adjacent *ko*-marked constituents, potentially causing interference during production. So the PDC account would not prefer these sentences on account of production difficulty and instead prefer the reference sentence above. The possibility that speakers chose the reference sentence above so that it would facilitate comprehension for listeners (compared to variant sentences which might be harder to interpret) is not considered by the PDC account.

We quantified processing efficiency using surprisal, originally proposed as a measure of language comprehension difficulty by Surprisal Theory (Hale, 2001; Levy, 2008). Consequently, we introduced surprisal estimated from n-gram and dependency parsing models into a logistic regression model for the task of predicting the reference sentence. Our choice of surprisal is inspired by Levy and Gibson (2013), who point out that the desiderata for PDC to become a theory of powerful empirical import is that it should make *quantitative* and *localized* predictions about incremental processing difficulty at each word. They highlight the fact that such a theory already exists, *viz.* the Surprisal Theory of language comprehension mentioned above. A perusal of the literature on information density in language production suggests that surprisal is a reasonable choice to model production difficulty as well.

Information density and surprisal are mathematically equivalent and both quantify the contextual predictability of a linguistic unit. But surprisal is based on different theoretical assumptions about resource allocation in comprehension. Recent research has demonstrated that reduction phenomena at both lexical (Frank and Jaeger, 2008, verb contraction) and syntactic (Jaeger, 2010, *that*-complementizer choice) levels exhibit the drive to minimize variation in information density across the linguistic signal. Moreover, instances of the same word which have greater predictability tend to be spoken faster and with less emphasis on acoustic details (Bell et al., 2009; Pluymaekers et al., 2005). The work cited above uses lexical frequencies or n-gram models over words to estimate contextual predictability. More recently, Demberg et al. (2012) showed that syntactic surprisal estimated from a top-down incremental parser is positively correlated with the duration of words in spontaneous speech, even in the presence of controls including word frequencies and trigram lexical surprisal estimates. Crucial to our study, words which are predictable in context have been interpreted to be more accessible in recent research (Arnold, 2011).

The results of our experiments show that reference sentences tend to minimize both trigram and dependency parser surprisal in comparison to their variants. Further, we show that the prediction accuracies of surprisal estimates derived from an artificially created version of Hindi without

[2] We concede that the use of written data (due to the lack of a publicly available Hindi speech corpus) is a major limitation of our study.

any case markers are significantly worse than the corresponding surprisal estimates based on natural Hindi. This experiment demonstrates the crucial contribution of case markers towards the predictive ability of surprisal and confirms our first hypothesis. Subsequently, we demonstrate that Hindi tends to avoid placing together constituents whose heads are marked by the same case marker. Moreover, incorporating predictors based on adjacent case marker sequences in a statistical model significantly improves model prediction accuracy over an extremely competitive baseline provided by n-gram and dependency parser surprisal. Phonological interference is a plausible explanation for this phenomenon and lends credence to our second hypothesis. The Hindi sentence comprehension literature provides only limited support for interference involving case marker sequences (Vasishth, 2003). Hence, it is plausible that this effect is a factor confined to the production system and not related to considerations of language comprehension. Further research using spoken corpora and spontaneous production experiments need to be performed in order to validate the psychological reality of our findings. Given that symbols used in the Hindi orthography have a direct correspondence with the sounds of the language (Vaid and Gupta, 2002), we expect speech to behave similarly.

Our main contribution is that we broaden the typological base of the PDC account of language processing, leveraging its connection with the well established surprisal theory of language comprehension. Levy and Gibson (2013) state that surprisal would enable PDC to be implemented computationally, thus facilitating hypothesis testing on a wide range of linguistic phenomena cross-linguistically. To this end, we set up a computational framework consisting of standard tools and techniques from the field of Natural Language Generation (NLG). Methodologically, the task of referent sentence prediction is a relatively novel way of studying word order and is inspired from the surface realization component of NLG. Recently, using a similar setup, Rajkumar et al. (2016) showed the impact of dependency length on English word order choices.

In this paper, Section 2 provides necessary background and Section 3 provides details of our data sets and models. Section 4 presents our experiments and their results. Finally, Section 5 summarizes the conclusions of our study and discusses the implications of our results for language production and learning.

Marker	**Case** (Gloss)	**Grammatical Function**
ϕ	nominative (NOM)	subject/object
ne	ergative (ERG)	subject
ko	accusative (ACC)	object
	dative (DAT)	subject/indirect object
se	instrumental (INS)	subject/oblique/adjunct
$ka/ki/ke$	genitive (GEN)	subject (infinitives) specifier
$m\tilde{e}/par/tak$	locative (LOC)	oblique/adjunct

Table 1: Hindi case markers (Butt and King, 1996).

2 Background

This section offers a brief background on Hindi word order and case marking, surprisal and core assumptions of the PDC account.

2.1 Hindi Word Order and Case Marking

A long line of work (Butt and King, 1996; Kidwai, 2000) has shown that scrambling in Hindi is influenced by factors like discourse considerations (topic, focus, background, and completive information), semantics (definiteness and animacy), and prosody (Patil et al., 2008). Hindi follows the head-marking strategy where case markers are postpositions which attach to noun phrases and encode a range of grammatical functions like subject and object (see Table 1 and case markers in bold in Examples 1a and 2a).

2.2 Surprisal Theory

The Surprisal Theory of language comprehension posits that fine-grained probabilistic knowledge (attained from prior linguistic experience) helps comprehenders form expectations about interpretations of the previously encountered structure as well as upcoming material (Hale, 2001; Levy, 2008). The theory defines surprisal as a measure of comprehension difficulty. In this work, we used the following definitions of surprisal:

1. **n-gram surprisal**: Mathematically, n-gram surprisal of the $(i{+}1)^{th}$ word, w_{i+1}, based on a traditional n-gram model is given by $S_{i+1} = -\log P(w_{i+1}|w_{i-n+2}, ..., w_{i-1}, w_i)$, as defined by Hale (2001). We estimated n-gram surprisal via trigram models (n=3) over words trained on 1 million sentences from the EMILLE corpus (Baker et al., 2002) using the SRILM toolkit (Stolcke, 2002) with Good-Turing discounting.

2. **Dependency parser surprisal** was computed using the probabilistic incremental dependency parser developed by Agrawal et al. (2017), based on the parallel-

processing variant of the *arc-eager* parsing strategy (Nivre, 2008) proposed by Boston et al. (2011). This parser maintains a set of the k most probable parses at each word as it proceeds through the sentence. The probability of a parser state is taken to be the product of the probabilities of all transitions made to reach that state. This parser can thus be used to define a measure of *dependency parser surprisal*: for the i^{th} word in a sentence, we first define the *prefix probability* α_i as the sum of probabilities of the k maintained parser states at word i:

$$\alpha_i = \sum_{\text{top } k \text{ derivations } d \text{ leading to word } i} Prob(d) \qquad (1)$$

The dependency parser surprisal at word i+1 is then computed as:

$$S^{syn}_{i+1} = -\log(\alpha_{i+1}/\alpha_i) \qquad (2)$$

The dependency parser surprisal of the $(i{+}1)^{th}$ word is computed as the negative log-ratio of the sum of probabilities of maintained parser states at word $i+1$ to the same sum at word i. We estimated it using a corpus of 12,000 HUTB projective trees.

2.3 Production-Distribution-Comprehension (PDC) Account

The **Production** component of the PDC account posits three factors of production ease. 1. *Easy First*: Relatively more accessible (ease of memory retrieval and conceptual salience) or available elements are produced earlier in the structure. 2. *Plan Reuse*: Speakers tend to repeat previously used or mentioned structures due to syntactic priming. 3. *Reduce Interference*: Speakers tend to choose words which do not interfere with other words in the utterance plan. These factors compete with each other during the production process to mould language forms.

The **Distribution** component states that the distribution of structures in natural languages reflects a bias towards having a greater number of structures which are easier to produce. Thus PDC attributes the greater frequency of subject relative clauses compared to object relatives across languages to production ease. Finally, the **Comprehension** part of the PDC approach proposes that language comprehension reflects the statistics of the input (*i.e.*, production patterns) perceived by language users. Thus, according to PDC, the greater difficulty involved in comprehending object relative clauses compared to subject relatives (Gibson, 2000) is because of the lower exposure to object relatives by virtue of their lower frequency in the linguistic input to comprehenders. Levy and Gibson (2013) puts forth the idea that surprisal (estimated from corpora) is naturally very compatible with the PDC assumption described above. Maryellen MacDonald and colleagues validate PDC predictions using a series of experiments related to relative clause production and comprehension in many languages (Gennari and MacDonald, 2008, 2009; Gennari et al., 2012).

3 Data and Models

Our data set consists of 8736 reference sentences corresponding to labeled, projective dependency trees in the Hindi-Urdu TreeBank (HUTB) corpus of written Hindi (Bhatt et al., 2009). We generated variants for each reference sentence by randomly permuting the preverbal constituents of the root node of its dependency tree. We selected trees whose roots were verbs. For example, in the tree depicted in Figure 1 (corresponding to Example 1a), we reordered the preverbal constituents immediately dominated by the verb *diyaa* and obtained the variants shown in Examples 1b and 1c. In order to eliminate ungrammatical variants, we excluded variants containing dependency relation sequences of the root word not present in the corpus of HUTB gold standard trees. Dependency relation sequences like *k7t-k1*, *k1-k7t*, *k7t-k7* and *k7-k4* in Figure 1 simulate grammar rules used in grammar-based surface realization systems. We obtained 175801 variants after filtering.

In order to mitigate the imbalance between the number of reference and variant sentences, we transformed the data set using a technique described in Joachims (2002). As per this technique, a binary classification problem can be converted into a pairwise ranking problem by training a classifier on the difference between the feature vectors of a reference sentence and its syntactic choice variants:

$$\mathbf{w} \cdot \phi(Reference) > \mathbf{w} \cdot \phi(Variant) \qquad (3)$$

$$\mathbf{w} \cdot (\phi(Reference) - \phi(Variant)) > 0 \qquad (4)$$

In Equation 3 above, the $Reference$ data point is predicted to outrank the $Variant$ data point when the dot product of the feature vector of the reference with $\mathbf{w}$ (learned feature weights) is greater than the corresponding product of the variant. The same can be written (Equation 4) as the dot product of $\mathbf{w}$ with the feature vector difference being positive. We created ordered pairs

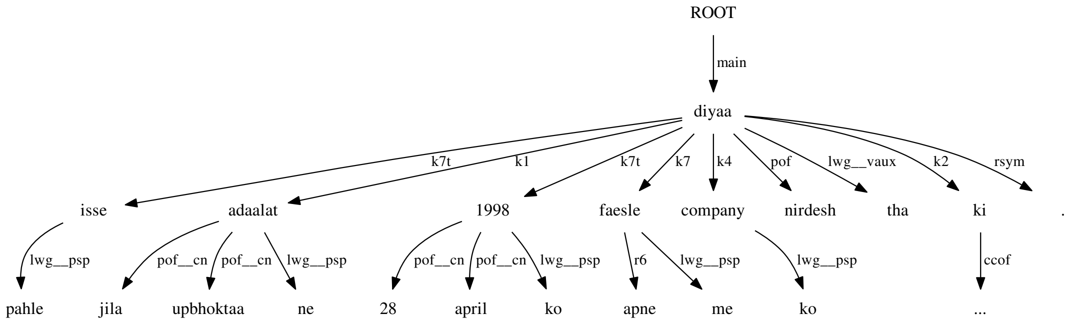

Figure 1: Example HUTB dependency tree (Table 8 in the Appendix provides a glossary of dependency relations)

consisting of the feature vectors of reference and variant sentences. Every reference sentence in the data set was paired with each of its variants (Examples 1a-1b and Examples 1c-1a constitute two such pairs). Then the feature values of the first member of the pair were subtracted from the corresponding values of the second member. Pairs alternate between *reference-variant* (coded as "1") and *variant-reference* (coded as "0"), resulting in a data set consisting of an equal number of classification labels of each kind (see Appendix for a detailed illustration).

4 Experiments

In this section, we describe three experiments to test our hypotheses on the transformed version of our data set consisting of 175801 data points using a logistic regression model. The goal is to predict "1" and "0" labels (as described in the previous section) using a set of cognitively motivated features. We calculated lexical and dependency parser surprisal feature values over entire sentences by summing the log probabilities of the surprisal values of individual words. We carried out 27-fold cross-validation; for each run, a model trained on 26 folds (consisting of 1 fold for hyperparameter tuning) was used to generate predictions about the remaining fold (100 training iterations using lbfgs solver in python *scikit-learn* toolkit-v0.16.1).

4.1 Processing Efficiency Experiments

Here, we test the hypothesis that word order choices in language are optimized for processing efficiency by incorporating trigram and dependency parser surprisal as predictors in a logistic regression model. A negative regression coefficient for these predictors would imply that corpus sentences have lower surprisal than variants. For the entire corpus, Figure 2 indicates this trend, where the mean trigram surprisal per sentence of the corpus of reference sentences is lower than the corresponding value of all their variants (Figure 4 in the Appendix depicts the same trend for syntactic surprisal). For the task of predicting HUTB reference sentences, both our surprisal measures have a negative regression coefficient, individually as well as in combination (first three rows of Table 2). This confirms our hypothesis that word order choices optimize for processing efficiency. Given our interpretation of low surprisal as denoting ease of accessibility, our first experiment shows that Hindi

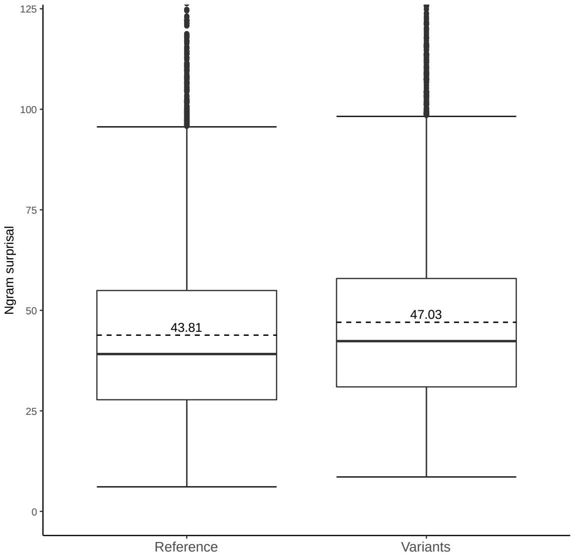

Figure 2: Mean trigram surprisal per sentence of reference and variant sentences (95% confidence intervals indicated)

Predictor(s)	Accuracy%	Weight(s)
Hindi		
Parser surprisal	62.10	-0.43
Trigram surprisal	89.96	-0.81
Trigram + parser surprisal	90.14	-0.98, -0.43
Caseless Hindi		
Caseless parser surprisal	55.03	-0.29
Caseless trigram surprisal	87.73	-0.83
Caseless trigram + parser surprisal	87.81	-0.93, -0.27

Table 2: Classification accuracies of surprisal for natural and caseless Hindi (175801 data points)

speakers tend to produce sentences by ordering preverbal constituents such that more accessible elements are realized first compared to other competing grammatical variants. This is in line with the PDC *Easy First* principle. Further, the classification accuracies indicate that trigram surprisal estimated from the EMILLE corpus is very effective in modelling syntactic choice (89.96% accuracy). For the same task, Ranjan (2015) reported that trigram surprisal estimated from the HUTB itself (smaller quantity of in-domain data) resulted in a lower accuracy of around 85%. In this context, our results show that a bigger n-gram training set can overcome the limitation of being from a different domain. A qualitative exploration revealed that n-gram model surprisal was particularly effective in reference-variant pairs as shown below (case markers shown in bold):

(2) a. Paakistan **ne** brihaspativaar **ko**
Pakistan ERG Thursday at
kathit taur **par** apne yahaan nirmit paramanu
allegedly indigenous nuclear
hathiyaar dhone **me** saksam krooj
weapons capable LOC carrying cruise
missile **ka** pareekshan kiyaa hai.
missile GEN tested

Pakistan has allegedly tested an indigenous cruise missile capable of carrying nuclear weapons on Thursday.

b. brihaspativaar **ko** Pakistan **ne** kathit taur **par** apne ...

The reference sentence (Example 2a with trigram surprisal of 45.60 hartleys) has the ergative-accusative (*ne-ko*) ordering of case-marked nouns compared to the variant (Example 2b with higher trigram surprisal 47.12) having the opposite ordering of nouns. Overall, 6% of a total of 175 HUTB sentences having ergative and accusative case markers exhibit a non-canonical accusative-ergative order (Agrawal et al., 2017). In both sentences above, the case markers in questions are separated by a single word and hence form part of a single trigram. Thus trigram surprisal is able to model the dominant order successfully while dispreferring the opposite order seen in Example 2b. Moreover, dependency parser surprisal has much lower classification accuracy compared to trigram surprisal and has a very negligible impact on performance on top of trigram surprisal. Thus, surprisal estimates from an incremental dependency parser are not effective in modelling constituent order choices. This is slightly unexpected as Agrawal et al. (2017) showed that surprisal estimates derived via the dependency parser deployed in our work accounts for per-word reading times for the Potsdam-Allahabad corpus over and above bigram frequencies. Using a similar setup, Rajkumar et al. (2016) showed that for the task of predicting English syntactic choice alternations, PCFG surprisal performed significantly better than n-gram model surprisal and the impact of dependency length is over and above both the aforementioned surprisal predictors. We are in the process of creating a constituency structure treebank for Hindi and plan to experiment with surprisal derived from a constituent-structure parser very soon. In recently completed work, Ranjan et al. (In Preparation) show that for Hindi, dependency length exhibits a weak effect over and above surprisal for predicting corpus sentences amidst artificial variants. Finally, we examined 1022 reference-variant pairs in our dataset where none of our features was able to predict the reference sentence correctly. We isolated cases involving other factors like given-new orders (30% cases), focus or topic considerations (marked by *hi* or *to* markers constituting 10% of cases) and null subjects (7.5%). Such discourse considerations are not encoded in our surprisal estimates (confined to single sentences) and further research can incorporate information about sentences from the preceding context into surprisal estimates.

Note that when considering the relationship between communicative efficiency and word order choices, there is a potential 'levels' problem (Levy, 2018). At the level of evolutionary timescales and entire populations, one might expect the grammar or distributional properties of the language to be adapted for efficiency. But at the level of an individual speaker's production choices, certain measures of efficiency will in turn depend on the extant distribution of linguistic forms. So there is a potential circularity in trying to assess the validity of such measures. Here we seek to model only the lower of these levels, *i.e.*, individual choices over a human lifetime. Hence, all the non-corpus variants we consider are *grammatical*. We assume that the grammar of the language is held fixed, and within the set of possible word order variants of a sentence licensed by that extant grammar, seek to model why speakers may have a greater propensity to produce some variants over others.

4.2 Case Markers and Processing Efficiency

In order to quantify the exact contribution of Hindi case markers towards the predictive accuracy of syntactic and trigram surprisal, we performed similar experiments using an artificial version of the language (*i.e.*, Hindi without case markers). The sentence comprehension literature demonstrates the vital role of case markers in predicting the final verb in verb-final constructions of languages like German (Levy and Keller, 2013) and Japanese (Grissom II et al., 2016). Moreover, in recent years, deploying artificial languages to test hypotheses about language processing and learning has been in vogue in both connectionist modelling (Lupyan and Christiansen, 2002; Everbroeck, 2003) as well as behavioural experiments (Kurumada and Jaeger, 2015; Fedzechkina et al., 2017). Inspired by the cited works, we created a caseless version of Hindi by removing case markers (those listed in Table 1) from both reference and variant sentences. The caseless equivalents of Examples 2a and 2b discussed in the previous section are given below:

(3) a. Pakistan brihaspativaar kathit taur apne yahaan nirmit paramanu hathiyaar dhone mein saksam krooj missile pareekshan kiyaa hai.

b. brihaspativaar Pakistan kathit taur apne ...

Then we estimated surprisal by stripping off case markers from the EMILLE corpus (trigram surprisal) as well as HUTB trees (dependency parser surprisal) so that our surprisal estimates mirrored the patterns in the caseless version of the language faithfully. Both surprisal measures derived from the caseless version of Hindi perform significantly worse than natural Hindi (last three rows of Table 2). Caseless trigram surprisal does 2% worse, while there is a 7% dip in the performance of caseless dependency parser surprisal (McNemar's two-tailed significance $p < 0.001$ for both measures). Thus the caseless language model is not able to predict the reference sentence shown in Example 3a as it awards higher trigram surprisal (45.21), in comparison to the variant sentence in Example 3b, which has a lower surprisal value (43.74). Figure 3 in the Appendix depicts the lexical surprisal profiles for the examples discussed above (both regular Hindi and caseless equivalents). Dependency parser surprisal also exhibited the same predictions.

Removing any kinds of words (especially function words) will result in a decrease in prediction accuracy. So we compared the prediction accuracy of caseless surprisal with another baseline obtained by removing case markers and all other postpositions (e.g. *ke liye*, *ke dwara*) from both training and test data. Surprisal estimates derived from the case marker and postposition stripped version of Hindi resulted in an extra dip of 0.3% in the accuracy of trigram surprisal and 2.5% for dependency parser surprisal compared to surprisal obtained by stripping just the case markers. Thus even within the set of postpositions, case markers play a significant role in lexical and syntactic predictability and hence processing efficiency. Lack of case markers reduces the overall information content of a sentence for both speaker and hearer. Spontaneous production experiments showed that Japanese speakers tend to omit the optional marker -*o* when the meaning of the sentence is probable in a given context (Kurumada and Jaeger, 2015). However, in cases where the meaning is not plausible, speakers tend to mention the case marker, in spite of entailing greater production effort.

The work of Lupyan and Christiansen (2002) showed that for artificial SOV languages with no case marking, a sequential learning device (Simple Recurrent Network) failed to achieve high accuracy for the task of mapping words to grammatical roles. Their simulations suggest that verb-final languages need a case system for optimal learning as word order is not a reliable cue for grammatical function assignment. Using the miniature artificial language learning paradigm, Fedzechkina et al. (2017) conducted a study where two groups of adult learners were exposed to artificial languages with optional case marking (one fixed order and one flexible order). Learners of the flexible constituent order language produced more case markers than learners of the fixed order language, mirroring typological patterns.

4.2.1 Interference Experiments

In the light of the PDC principle of *Minimize Interference*, we investigate whether interference

Predictor name	Sequence	Distance
ϕ-ne	1	3
ne-ko	1	3
ko-$m\tilde{e}$	1	2
$m\tilde{e}$-ko	1	1
same-seq	0	-
diff-seq	4	-

Table 3: Values of case features extracted from tree in Figure 1.

Case marker sequence	Weight
ϕ-ϕ	**-0.002**
ke-ke	**-0.025**
ko-ko	**-0.291**
$m\tilde{e}$-$m\tilde{e}$	**-0.061**
tak-tak	0.008
par-par	0.231
se-se	0.055
same-seq	-0.009
diff-seq	0.009

Table 4: Learned weights of some case-sequence predictors.

Predictor(s)	Classification accuracy%	Ranking accuracy%
Case distance features	70.79	-
Case sequence features	74.94	-
Random Classifier	-	21.25
Baseline (trigram+parser surprisal)	90.16	55.04
Baseline+Case distance features	90.85***	55.68***
Baseline+Case sequence features	91.13***	56.03***
Baseline+Case distance + sequence features	**91.60*****	**56.16*****

Table 5: Pairwise classification and ranking accuracy (*** denotes McNemar's two-tailed significance $p < 0.001$ over the baseline model).

between NPs whose heads are marked by the same case marker influence preverbal constituent ordering choices in Hindi. Since PDC seeks to link production and comprehension, our experiments are also motivated by prior work on case marker interference in sentence comprehension in SOV languages like Japanese (Lewis and Nakayama, 2001), Korean (Lee et al., 2005) and Hindi (Vasishth, 2003). Our work is directly related to the experiments on identical case marking described in Chapter 3 of Vasishth (2003). In the case of Hindi center-embeddings, this work examined whether NPs having nominal heads marked by identical case markers induce similarity-based interference effects at the subsequent verb as predicted by the Retrieval Interference Theory (Lewis, 1998; Lewis and Nakayama, 2001). The study shows limited support for interference emanating from phonologically similar case markers.

In order to investigate interference caused by case markers in syntactic choice, we designed features based on case markers and incorporated them into our logistic regression model. For each dependency tree, we introduced two types of features associated with preverbal constituents of the root verb. 1. *Case-sequence features*: Counts of case marker sequences associated with the heads of a pair of adjacent constituents. We also introduced generic case-sequence features *same-seq* and *diff-seq* to model the overall trend. For each tree, these features denote the total number of identical and different case markers sequences associated with pairs of adjacent constituents. 2. *Case-distance features*: Number of intervening words between heads of the constituents of root verbs. Here, the feature name is obtained by combining the case markers associated with the constituent heads in question. Constituents which do not have case marked heads are marked as ϕ in order to model the fact that languages often use adverbial elements or other non-case marked arguments to separate case marked constituents. Table 3 illustrates our case features based on the dependency tree in Figure 1 corresponding to Example 1a.

In isolation, the case-sequence and case-distance features exhibit accuracies around 70% (second column of Table 5). The case sequence and distance features together induce a significant accuracy increase of 1.5% (McNemar's two-tailed significance $p < 0.001$) over a baseline model consisting of lexical and dependency parser surprisal as features. Though this might be a small increase when considered in isolation, we would like to note that our baseline model is extremely competitive (90.16% accuracy). Even dependency parser surprisal did not confer considerable performance gains over and above trigram surprisal as discussed earlier. So in this context, the contribution of case features is noteworthy.

Subsequently, we examined the learned weights of our case sequence features (Table 4) in our best model containing surprisal and all the case marker features. A negative weight is associated with four of the seven identical case marker sequences as well as the *same-seq* feature encoding the overall pattern across all case markers. These negative weights lend support to our hypothesis that Hindi shows a dispreference for placing together constituents whose heads are marked using the same case inflection. Interference due to repetition of phonologically identical case markers may be a plausible explanation for this phenomenon. However, three other case marker sequences have a positive weight and hence indicate a tendency towards adjacency. These three case markers are much lower in frequency in the HUTB compared to the other four and might not represent the dominant tendency. However, future inquiries need to explore the role of case-based facilitation (Logačev and Vasishth, 2012). Since our features are not sensitive to clause boundaries, conclusive evidence for phonological interference will emerge only after controlling for clause boundaries.

The best model (baseline + case marker features) picked the reference sentence (Example 1a) while the baseline model erroneously selected the artificially generated variants (Examples 1b and 1c). The reference sentence has two *ko*-marked constituents separated by intervening constituents. In contrast, the variant sentences have two adjacent *ko*-marked constituents, potentially causing interference. These examples also highlight the ambiguous nature of the *ko*-marker in denoting several functions in Hindi. As noted by Ahmed (2006), *ko* marks both accusative and dative case on objects (*company* in the cited examples) as well as dative subjects. In addition, it also occurs on spatial and temporal adjuncts (as in *28 April 1998*). In these examples, since *ko* marks both dative case and temporality, interference might be purely phonological in nature and not related to the actual grammatical function being marked. Further, we calculated the ranking accuracy of our main models, *i.e.*, the percentage of times a model ranked the reference sentence compared to all its variants. Table 5 (column 3) indicates that introducing case marker features into the baseline model induced significant ranking accuracy gains (McNemar's two-tailed significance $p < 0.001$). So our best model ranked Example 1a as the best sentence among all the other variants. Our classification and ranking results suggest that the PDC *Reduce Interference* principle of production ease is a valid constraint in constituent ordering.

In Hindi sentence comprehension, Vasishth (2003) explored the idea of positional similarity (Lewis and Nakayama, 2001), whereby the position of otherwise syntactically indiscriminable NPs in the structure contribute to interference at the subsequent verb. So he compared reading times at the innermost verb in the sequences of constituents with heads marked by ne-se-ko-ko and ne-ko-se-ko inflections. However, there was no significant difference in reading times between these sequences, thus offering no support for positional similarity during comprehension. This is the experimental condition which is most closely linked to our work. Interpreted in conjunction with our findings, case marker interference in Hindi appears to be a constraint on production rather than comprehension.

5 Discussion

Our main findings are broadly in line with two of the production ease principles of the PDC account. Our first experiment shows that the Hindi language orders words to optimize production ease (quantified using surprisal) at both lexical and syntactic levels, consistent with the PDC *Easy First* principle. Our second experiment shows that case markers make a significant contribution towards the predictive accuracy of both syntactic and trigram surprisal in choosing the reference sentence amongst grammatical variants denoting the same meaning. The role of surprisal and case markers in conferring accessibility needs to be investigated more thoroughly in future work. Finally, our third experiment shows that Hindi tends to disprefer constituent sequences with heads case marked by identical case markers, as predicted by the PDC principle of *Reduce Interference*. However, the lack of case marker interference in Hindi comprehension necessitates further inquiries into the PDC account, which conceives the lexico-syntactic statistics of production data (result of biases in utterance planning) as guiding comprehension processes. Thus, overall, we would like to conclude that certain aspects of PDC are validated by our experimental results. Further computational inquiries will be facilitated by formulating an algorithmic sketch of a process model outlining the causes of mismatches between production and comprehension. Finally, the PDC account conceives word order variation in languages of the world as emerging from an interplay of the three PDC production principles. Crucially, PDC conceives learning biases to be production biases, *i.e.*, speakers learn forms which are easier to produce (MacDonald, 2013). Future inquiries can explore whether learning outcomes are indeed consistent with typological patterns described by language universals.

6 Acknowledgements

We are grateful to the anonymous reviewers of this workshop, NAACL-2019 and Sigmorphon-2018 for their feedback. The second author acknowledges support from IISER Bhopal's Faculty Initiation Grant (IISERB/R&D/2018-19/77).

References

Rama Kant Agnihotri. 2007. *Hindi: An Essential Grammar*. Essential Grammars. Routledge.

Arpit Agrawal, Sumeet Agarwal, and Samar Husain. 2017. Role of expectation and working memory constraints in hindi comprehension: An eyetracking corpus analysis. *Journal of Eye Movement Research*, 10(2).

Tafseer Ahmed. 2006. Spatial, temporal and structural uses of urdu ko. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the 11th International LFG Conference*. CSLI Publications, Stanford.

Jennifer E. Arnold. 2011. Ordering choices in production: For the speaker or for the listener? In Emily M. Bender and Jennifer E. Arnold, editors, *Language From a Cognitive Perspective: Grammar, Usage, and Processing*, pages 199–222. CSLI Publishers.

Paul Baker, Andrew Hardie, Tony McEnery, Hamish Cunningham, and Robert Gaizauskas. 2002. Emille: a 67-million word corpus of indic languages: data collection, mark-up and harmonization. In *Proceedings of LREC 2002*, pages 819–827. Lancaster University.

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 186–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barry J. Blake. 2001. *Case*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press.

Kathryn Bock. 1987. An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language*, 26(2):119 – 137.

Marisa Ferrara Boston, John T. Hale, Shravan Vasishth, and Reinhold Kliegl. 2011. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349.

Miriam Butt and Tracy Holloway King. 1996. Structural topic and focus without movement. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the First LFG Conference*. CSLI Publications, Stanford.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, volume 11. The MIT press.

Morten H. Christiansen and Nick Chater. 2008. Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489509.

Vera Demberg, Asad B. Sayeed, Philip J. Gorinski, and Nikolaos Engonopoulos. 2012. Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 356–367, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ezra Van Everbroeck. 2003. Language type frequency and learnability from a connectionist perspective. *Linguistic Typology*, 7(1):1–50.

Maryia Fedzechkina, T. Florian Jaeger, and Elissa L. Newport. 2012. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902.

Maryia Fedzechkina, Elissa L. Newport, and T. Florian Jaeger. 2017. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive Science*, 41(2):416–446.

Janet Dean Fodor. 2001. Setting syntactic parameters. *The Handbook of Contemporary Syntactic Theory*, pages 730–767.

A. Frank and T.F. Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Cogsci. Washington, DC: CogSci*.

Silvia P. Gennari and Maryellen C. MacDonald. 2008. Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, 58(2):161 – 187.

Silvia P. Gennari and Maryellen C. MacDonald. 2009. Linking production and comprehension processes: The case of relative clauses. *Cognition*, 111(1):1 – 23.

Silvia P. Gennari, Jelena Mirkovi, and Maryellen C. MacDonald. 2012. Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology*, 65(2):141 – 176.

Edward Gibson. 2000. Dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, Language, brain: Papers from the First Mind Articulation Project Symposium*. MIT Press, Cambridge, MA.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, Mass.

Alvin Grissom II, Naho Orita, and Jordan Boyd-Graber. 2016. Incremental prediction of sentence-final verbs: Humans versus machines. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 95–104. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press.

Florian Jaeger, Katrina Furth, and Caitlin Hilliard. 2012. Incremental phonological encoding during unscripted sentence production. *Frontiers in Psychology*, 3:481.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage information density. *Cognitive Psychology*, 61(1):23–62.

T. Florian Jaeger and Harold Tily. 2011. Language processing complexity and communicative efficiency. *WIRE: Cognitive Science*, 2(3):323–335.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

Y. Kachru. 2006. *Hindi*. London Oriental and African language library. John Benjamins Publishing Company.

Ayesha Kidwai. 2000. *XP-Adjunction in Universal Grammar: Scrambling and Binding in Hindi-Urdu: Scrambling and Binding in Hindi-Urdu*. Oxford studies in comparative syntax. Oxford University Press.

Chigusa Kurumada and T. Florian Jaeger. 2015. Communicative efficiency in language production: Optional case-marking in japanese. *Journal of Memory and Language*, 83(Supplement C):152 – 178.

Sun-Hee Lee, Mineharu Nakayama, and Richard L. Lewis. 2005. Difficulty of processing Japanese and Korean center-embedding constructions. In M. Minami, H. Kobayashi, M. Nakayama, and H.Sirai, editors, *Studies in Language Science*, volume Volume 4, pages 99–118. Kurosio Publishers, Tokyo, Tokyo.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126 – 1177.

Roger Levy and Edward Gibson. 2013. Surprisal, the pdc, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, 4(229).

Roger P Levy. 2018. Communicative efficiency, uniform information density, and the rational speech act theory.

Roger P. Levy and Frank Keller. 2013. Expectation and locality effects in german verb-final structures. *Journal of Memory and Language*, 68(2):199 – 222.

Richard L. Lewis. 1998. Interference in working memory: Retroactive and proactive interference in parsing. CUNY sentence processing conference.

Richard L. Lewis and Mineharu Nakayama. 2001. Syntactic and positional similarity effects in the processing of Japanese embeddings. In *Sentence Processing in East Asian Languages*, pages 85–113, Stanford, CA. CSLI.

Pavel Logačev and Shravan Vasishth. 2012. Case matching and conflicting bindings interference. In Monique Lamers and Peter de Swart, editors, *Case, Word Order and Prominence: Interacting Cues in Language Production and Comprehension*, pages 187–216. Springer Netherlands, Dordrecht.

Gary Lupyan and Morten H. Christiansen. 2002. Case, word order, and language learnability: Insights from connectionist modeling. In *In Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 596–601. Erlbaum.

Maryellen C. MacDonald. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology*, 4(226):1–16. Published with commentaries in Frontiers.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Comput. Linguist.*, 34(4):513–553.

Umesh Patil, Gerrit Kentner, Anja Gollrad, Frank Kügler, Caroline Féry, and Shravan Vasishth. 2008. Focus, word order and intonation in Hindi. *Journal of South Asian Linguistics*, 1(1):55–72.

Mark Pluymaekers, Mirjam Ernestus, and R Harald Baayen. 2005. Lexical frequency and acoustic reduction in spoken dutch. *The Journal of the Acoustical Society of America*, 118(4):2561–2569.

Rajakrishnan Rajkumar, Marten van Schijndel, Michael White, and William Schuler. 2016. Investigating locality effects and surprisal in written English syntactic choice phenomena. *Cognition*, 155:204–232.

Sidharth Ranjan. 2015. Investigation of locality effects in hindi language production. Master's thesis, Indian Institute of Technology (IIT) Delhi. Unpublished thesis.

Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. In Preparation. Locality and surprisal effects in hindi preverbal constituent ordering.

Edward Sapir. 1921. *Language: An Introduction to the Study of Speech*. Harcourt, Brace, New York.

Andreas Stolcke. 2002. SRILM — An extensible language modeling toolkit. In *Proc. ICSLP-02*.

J Vaid and Anshum Gupta. 2002. Exploring word recognition in a semi-alphabetic script: The case of Devanagari. *Brain and Language*, 81:679–90.

Julie A. Van Dyke. 2007. Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 33 2:407–30.

Julie A. Van Dyke and Brian McElree. 2006. Retrieval interference in sentence comprehension. *Journal of Memory and Language*, 55(2):157 – 166.

S. Vasishth. 2003. *Working Memory in Sentence Comprehension: Processing Hindi Center Embeddings*. Outstanding Dissertations in Linguistics. Taylor & Francis.

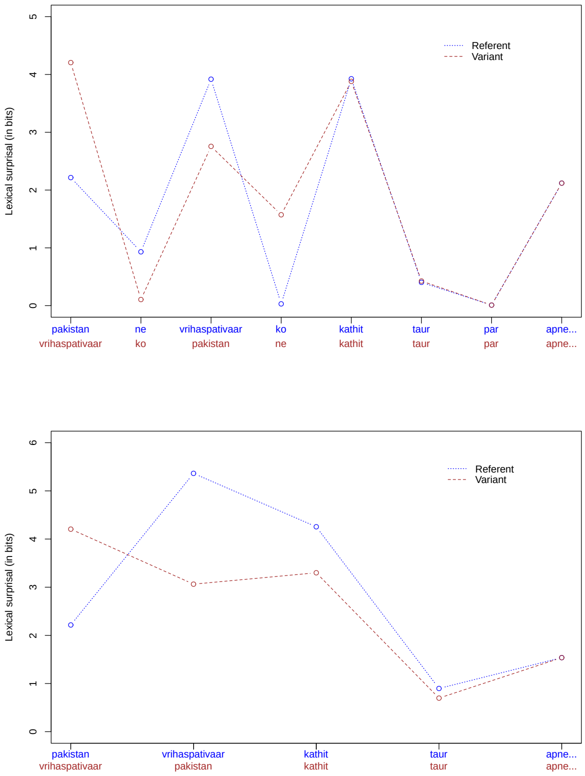

Figure 3: Lexical surprisal profiles of normal and the caseless artificial version of Hindi

Sentence type	Label	Lexical surprisal	Syntactic surprisal
Reference	1	97.45	156.64
Variant$_1$	0	97.69	160.77
Variant$_2$	0	98.25	156.91
Variant$_3$	0	97.97	159.50
Variant$_4$	0	98.16	161.94

Table 6: Original dataset

A Appendix

A.1 Joachims Transformation

Consider the first example in the following Hindi sentences as reference corresponding to *'Jayalalitha has written a letter to the prime minister on this issue'* and remaining as grammatical variants expressing the same idea. Assuming this as a toy dataset, Table 6 denotes their lexical and syntactic surprisal feature values whereas Table 7 represents its Joachims transformation.

Reference [jayalalitha-ne]$_1$ [is mazle par]$_2$ [pradhanmantri-ko]$_3$ [ek patr]$_4$ V ...

Variant$_1$ [is mazle par]$_2$ [jayalalitha-ne]$_1$ [pradhanmantri-ko]$_3$ [ek patr]$_4$ V ...

Variant$_2$ [jayalalitha-ne]$_1$ [pradhanmantri-ko]$_3$ [is mazle par]$_2$ [ek patr]$_4$ V ...

Variant$_3$ [pradhanmantri-ko]$_3$ [is mazle par]$_2$ [jayalalitha-ne]$_1$ [ek patr]$_4$ V ...

Variant$_4$ [is mazle par]$_2$ [pradhanmantri-ko]$_3$ [jayalalitha-ne]$_1$ [ek patr]$_4$ V ...

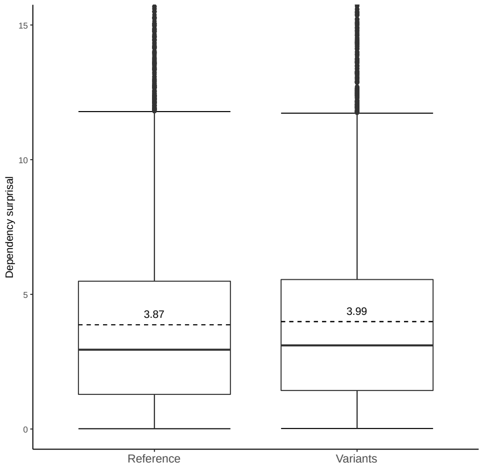

Figure 4: Mean syntactic surprisal per sentence of reference and variant sentences (95% confidence intervals indicated)

New	Label	δ Lexical surprisal	δ Syntactic surprisal
Variant$_1$-Reference	0	0.25	4.13
Reference-Variant$_2$	1	-0.81	-0.28
Variant$_3$-Reference	0	0.53	2.87
Reference-Variant$_4$	1	-0.72	-5.30

Table 7: Transformed dataset

Here are the steps to transform the data set using the Joachims transformation technique.

1. Equal number of ordered pairs of type *(Reference, Variant)* and *(Variant, Reference)* were created.

2. Differences between the feature values of the elements of these ordered pairs were taken (see Table 7).

3. <*Reference-Variant*> pairs were labelled as **1** and <*Variant-Reference*> pairs were labelled as **0**. Here, 1 stands for the correct choice and 0 denotes the incorrect choice.

Label	Dependency relation
Invariant syntactic relations	
k1	subject/agent
k2	object/patient
k4	recipient
k7	location (elsewhere)
k7t	location (in time)
r6	genitive/possessive
Local word group (lwg)	
lwg_psp	postposition
lwg_vaux	auxilliary verb
Symbols	
rsym	symbol relation
Indirect dependency relations	
ccof	co-ordination and sub-ordination
pof	part of units such as conjunct verbs
pof_cn	part of units such as compound noun

Table 8: Glossary of dependency relations