Manovaad: A Novel Approach to Event Oriented Corpus Creation Capturing Subjectivity and Focus

by

Lalitha Kameswari VA, Radhika Mamidi

in

Language Resources and Evaluation Conference

Report No: IIIT/TR/2020/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2020

Manovaad: A Novel Approach to Event Oriented Corpus Creation Capturing Subjectivity and Focus

Lalitha Kameswari*, Radhika Mamidi*
*Language Technologies Research Centre
IIIT Hyderabad,Telangana, India
v.a.lalitha@research.iiit.ac.in, radhika.mamidi@iiit.ac.in

Abstract

In today's era of globalisation, the increased outreach for every event across the world has been leading to conflicting opinions, arguments and disagreements, often reflected in print media and online social platforms. It is necessary to distinguish factual observations from personal judgements in news, as subjectivity in reporting can influence the audience's perception of reality. Several studies conducted on the different styles of reporting in journalism are essential in understanding phenomena such as media bias and multiple interpretations of the same event. This domain finds applications in fields such as Media Studies, Discourse Analysis, Information Extraction, Sentiment Analysis, and Opinion Mining. We present an event corpus "Manovaad-v1.0" consisting of 1035 news articles corresponding to 65 events from 3 levels of newspapers viz., Local, National, and International levels. Using this novel format, we correlate the trends in the degree of subjectivity with the geographical closeness of reporting using a Bi-RNN model. We also analyse the role of background and focus in event reporting and capture the focus shift patterns within a global discourse structure for an event. We do this across different levels of reporting and compare the results with the existing work on discourse processing.

Keywords: Event Corpus, News, Subjectivity, Bidirectional Recurrent Neural Network, Focus Shift, Discourse Analysis

1. Introduction

News discourse has always been considered as an objective and formal linguistic form of discourse. There have been multiple definitions and opinions on what consists news. Contrary to the commonly accepted notion of news being just an objective factual account of the important happenings around us, many linguists argue that there is much more to it. (Fowler, 1991) explains that "News is not a natural phenomenon emerging straight from 'reality', but a product. It is produced by an industry and shaped by the relations between the media and other industries." Van Dijk (2013) characterises news, not as a "picture of reality", but as a "frame through which the social world is routinely constructed." Information provided by the news reports has an important role in leading the public in society, and newspapers serve as the primary source of textual news content.

Newspapers have been instrumental in shaping people's social, cultural and political views. (Mencher and Shilton, 1997) talk about the various layers of reporting in the newspapers:

- Layer I - Basic facts
- Layer II - Details and description
- Layer III - Interpretation

When we look at different reports for the same event through different sources, we can observe that no two are completely identical. The same event can have reports belonging to different layers as described above. The significance associated with an event in terms of features such as focus, perspective, inclination and details is dependent largely on the closeness of the event with the person or the organisation reporting it.

One way of looking at the closeness is the geographical distance between the location of the reporter and the location where the event has taken place. On the basis of this, we have classified the news articles for each event into local, national and international reports. We hypothesise that subjectivity in reporting gradually increases as one moves from international to local levels of reporting. This was based on the idea that in local reporting, the reporters usually have first hand knowledge of the event and add details and opinions based on their personal interpretation, which lead to subjectivity in the article. The first part of our work deals with experimenting on our corpus to evaluate the validity of this hypothesis.

The second part of our work deals with capturing the chronological shift in the focus of event reporting in each of the three levels of articles. Van Dijk (2013) presents a hierarchical theory of news discourse to model how paragraphs operate as units of discourse structure within news articles to capture the importance of events within a story. We aim to extend this further, and observe how each article, or a group of articles focus on sub-events, and how this focus varies with time, leading to a series of focus shifts which together capture the whole event and its overall focus. Similar to our subjectivity hypothesis as above, we claim that the extent of focus shift would be the most in local news articles.

2. Related Work

2.1. Subjectivity Analysis

The analysis of discourse using news, blogs and social media to study the sentiments and opinions has been an interesting topic which has opened many avenues for research. Godbole et al. (2007) built a large-scale sentiment analysis system for news and blog entities, and showed how public sentiment varies with time.

Mourad and Darwish (2013) focused on Subjectivity and Sentiment Analysis (SSA) for Arabic language. They

adopted a random graph walk approach to extend the Arabic SSA lexicon using Arabic-English phrase tables, and used different features for both subjectivity and sentiment classification including stemming, part-of-speech tagging, as well as other discourse specific features.

Liu and others (2010) discuss the two stage sentiment and subjectivity classification approach at different granularities (document and sentence levels) using different machine learning approaches along with different ways to construct the required data resources (corpora and lexicon).

Wahl-Jorgensen (2013) conducted a study on subjectivity and story-telling traits in Pulitzer Prize-winning articles. Their results indicated that despite the continued prominence of the ideal of objectivity in scholarly and journalistic debates, award-winning journalistic stories are in fact pervaded by subjective language in the form of what linguists refer to as "appraisals", as well as the narrative construction of emotive appeals.

Other studies have specifically been conducted on phenomena such as sensationalism (Arbaoui et al., 2016), emotionalism (Richards and Rees, 2011), and expressing subjective opinions in the news (Steele and Barnhurst, 1996). These traits of subjectivity in journalism can be regarded as deviating from the objectivity norm of journalism (Schudson, 2001).

However, to our knowledge, there hasn't been any study which aims to systematically capture the dependency between subjectivity in reporting of the same event and the closeness of the person/organisation who is reporting it.

2.2. Focus in Discourse Analysis

The notion of focus has immense importance in the study of information structure and language representation. Brown et al. (1983) described focus as the element of a sentence which contributes new and non derivable information. Since then, there have been several semantic and syntactic studies on extracting focus, but most of them are limited to sentence or paragraph level. We could not find any linguistic work which aims at analysing trends in focus, such as focus shifts, within a some global context.

Grosz and Sidner (1986) proposed an attentional state model for analysing discourse structure. The model has two constituent substructures- a global one and a local one. The local-level component is modeled with centering and the global level component of the attentional state is modeled with a stack, called the focus-space stack.

Our approach towards analysing focus shifts within a global context for an event is the first of its kind, and we can relate it to the attentional state model by considering each article to a particular event as an addition to the stack, which consists of the background of the event provided by all the articles before it in a chronological order.

3. Corpus

A large corpus is the primary requirement to effectively analyse the trends and arrive at notable conclusions, as well as to apply suitable machine learning techniques. Since there is no existing event based corpus focusing on various levels of reporting news, we created a new corpus "Manovaad-v1.0"[1], which consists of various events and the news articles corresponding to those events reported in the local, national, and international newspapers. Our aim was to create a balanced and representative corpus consisting of events and the corresponding articles belonging to different categories such as Sports, Trade, Politics, Entertainment, Natural calamities, Science, etc.

3.1. Corpus Collection

"Manovaad-v1.0" consists of 65 events which have occurred in the period of 2016 to 2018. Newspaper articles in English were collected for each event by scraping from various standard newspaper archives with relevant keyword searches and date filters. At the end of this step, we had 1,100 news articles.

3.2. Corpus Classification and Filtering

Articles for each event were classified into following three categories:

1. **International** - If the article was featured in the International/World News column in some newspaper circulated in various countries. These articles usually report events which have a global significance.
   e.g: The Facebook - Cambridge Analytica Data Scandal of 2018.[2]

2. **National** - If the article was reported only in a particular country, usually pertaining to national level issues within a country such as politics, sports, business, etc. These articles indicate that an event is of importance, usually for the whole country, i.e beyond just the local site of occurrence.
   e.g: The Section 377 verdict of the Indian Supreme Court which led to striking down of the ban on consensual gay sex in India.[3]

3. **Local** - If the article was reported in the Local (State/District/City) editions of some newspapers, or exclusively in the local newspapers of a state.
   e.g: Agitation in West Bengal's Darjeeling district in India for a separate Gorkhaland state.[4]

It has to be noted that throughout this paper, whenever we refer to some newspaper as Local, National, or International - these are clearly the geographical points of reference, and do not denote the reputation or the stature of the publishing house. For instance, an article from a popular newspaper such as the New York Times would be considered as a local article for some event happening within the New York City. Another article from the same paper might be considered International if it talks about some event in India.

After classifying articles as described above, we manually checked for the relevance of the articles to the event, and

[1] Derived from Sanskrit. *Manovaad* means Subjectivity. *Mano*: Related to the mind (subjective opinions, interpretations, etc.), *Vaad*: Expressing

[2] https://en.wikipedia.org/wiki/Facebook-Cambridge_Analytica_data_scandal

[3] http://bit.ly/2XcCX7J

[4] https://en.wikipedia.org/wiki/Gorkhaland#2017_agitation

removed the irrelevant articles which were scraped. This step was carried out to ensure that the limited efficiency of keyword search and date filters in various archives did not affect the quality of our corpus. We also removed the events which have articles in only one of the three categories. This is because, we require at least two levels of reporting for every event for further analysis. After filtering, we have 250 local articles, 510 national articles, and 275 international articles. On the whole, "Manovaad-v1.0" consists of 7,04,762 words and 1,035 articles corresponding to 65 events. To our knowledge, this is the first and the largest event oriented corpus till date.

3.3. Challenges in Corpus creation

1. Most of the local news articles were reported in the local language, and there were only few local news articles in English.

2. Images, videos and advertisements embedded in the webpages of the archives made it difficult to obtain the textual content of the article.

3. Many newspapers either had no archives, or had digital archives where the newspapers were stored as images, and hence it was difficult for us to retrieve the data for many events.

4. Methodology

4.1. Subjectivity Evaluation

The first objective of this paper is to analyse the trends in the degree of subjectivity of reporting at different levels. For this, we used a Subjectivity Classifier[5] implemented using a Bidirectional Recurrent Neural Network proposed by (Schuster and Paliwal, 1997) as shown in Figure 1.
The original network structure was taken from Augenstein et al. (2016) who proposed a Bidirectional Conditional LSTM encoding model for stance detection.

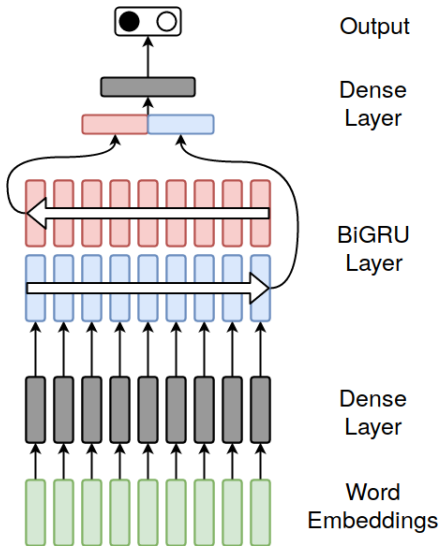

Figure 1: Network Structure of the Subjectivity Classifier

[5] https://github.com/fractalego/subjectivity_classifier

Since this classifier was trained on movie reviews which have ample scope for subjective opinions, we preferred this over the other few models trained on corpora such as the Stock market news. However, we wanted to check the validity of this classifier for identifying subjectivity in event based news.
For this purpose, we collected 100 articles from various domains such as Politics, Sports, Business, Entertainment etc. and each article was given to three annotators to rate the subjectivity of the article on a scale of 1-10 (1 for completely objective and 10 for completely subjective). The same set of articles were now given to the classifier which gave them a subjectivity score between 0 to 1, based on the fraction of subjective sentences in the article. After averaging the annotator scores and normalising them to a scale of 0 to 1, we compared the mean absolute difference between the classifier value and human value. Our results were highly satisfactory with 0.045 as the magnitude of mean absolute difference. The inter annotator agreement (IAA) was found to be 0.73. This confirmed our view that this classifier was suitable for our corpus.
In the next step, we similarly used the classifier to assign a subjectivity-score to each article in our corpus. The mean subjectivity scores were assigned to local, national, and international categories which were calculated as the average of all the subjectivity scores of the corresponding articles. If no article was present in a particular category for an event, it was given a value of "NULL". We also noted down the maximum subjectivity-score and the minimum subjectivity-score for each category.

4.2. Focus Identification

We adopt the definition of linguistic focus given by Halliday (1967) as the most prominent information conveyed through the discourse element. Yang et al. (2017) conducted cognitive experiments on dynamic construction of discourse representations. They asserted that discourse comprehension involves creation of a mental representation in which only the focused parts of the discourse are more attended to than the other information.
However, unlike the usual method of studying focus independently without considering some global variables, we attempted to organise events and the corresponding articles at each level in a chronological order. For focus identification of each article, we used the Python NLTK (Bird et al., 2009) and Gensim (Řehůřek and Sojka, 2010) to generate the summary, considering the headline and the article. Then, we manually arrived at the focus for each article from its summary and the additional context provided from the previous articles related to the event wherever necessary as shown in the example in Figure 2.

5. Observations and Results

5.1. Subjectivity Analysis

After getting the subjectivity scores in all the categories, we observed that there was a significant growth in the values of the subjectivity scores as we moved from international to local articles. For events in which there were no "NULL" values for any category, we considered all the three subjec-

Article:
With the buses going off-road, the TSRTC continued their strike for the second day. The buses were confined to the depots while hundreds of passengers stranded at the bus stations. Meanwhile, around 160 employees out of 50,000 withdrew the strike and joined the duties on Saturday evening after the government warned to sack the employees participating in the strike. The buses were run amidst the tight security on Saturday. However, the passengers faced the striking effect although the government arranged alternate measures and stranded the bus stations for hours. On the other hand, the private vehicles made the most of the strike by charging passengers extra. While the ticket from Hyderabad to Karimnagar was Rs 190, the passengers paid Rs 600-800.

Generated summary after lemmatization and removing stop words:
bus were confined to depot. hundred of passenger stranded at bus station. around 160 employee out of 50,000 withdrew strike and joined duty on Saturday evening after government warned to sack employee participating in strike. passenger faced striking effect although government arranged alternate measure and stranded bus station for hour.

Elements of focus generated manually
Buses unavailable - Passengers stranded - Govt warns employees and arranges alternate measures

Figure 2: Extracting focus from an article

tivity values. For the other events, we considered the two categories with positive subjectivity scores.

No.	Event	Inter-national	National	Local
1	Mithali Raj creates a record of 6000 ODI runs in cricket	0.24	0.35	0.52
2	Actress Priyanka Chopra marries singer Nick Jonas	0.05	0.24	0.36
3	Two Indians win Ramon Magsasay award in 2018	0.19	0.42	NULL
4	Ochki cyclone hits Tamil Nadu and Kerala in India	0.18	0.21	0.33

Table 1: Some events and their subjectivity values in the three categories

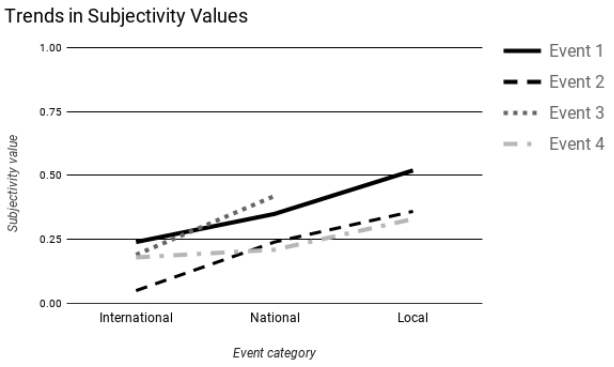

Figure 3: Graph showing trends in subjectivity values of events in Table 1

For instance, Table 1 shows the subjectivity values for 4 events from our corpus and Figure 3 graphically represents the increase in subjectivity scores. We also observed that the difference between maximum and minimum subjectivity scores increased as we moved from international to local articles.

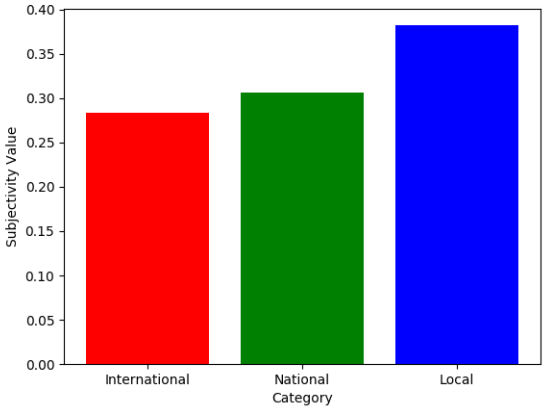

Figure 4: Average Values of Subjectivity for each level as observed in our corpus

Figure 4 shows the average subjectivity values for the three levels. We see that there is an increase from International (0.283) to National (0.306) to Local (0.382). These results are in agreement with our earlier hypothesis mentioned in Section 1.

5.2. Trends in Focus Shift

Following are the trends at each level we could observe in the events which had coverage at the local level. We illustrate the findings through an example of the devastating floods in India in the state of Kerala in 2018[6].

- In the local articles, we were able to identify and clearly demarcate the chronologically sorted sequence of articles into sub-units which had related but varying focus, as shown in Table 2.
- In the national articles, this trend was observed, but to a much smaller extent i.e, the shift in focus between the last and first article was very less when compared to the local articles. Also, the time period for which the event was covered in various articles was very less when compared to the local level, as shown in Table 3.
- There was no focus shift in the international articles, since they were usually limited to stating the facts about an event, and had no further coverage after the initial reporting, as shown in Table 4. .

These observations are in agreement with the famous Structure Building Framework of discourse comprehension in which Gernsbacher (2013) proposed two main stages in creating a structure corresponding to each discourse. Similarly, we observed that the chronologically sorted articles at the local level were divided into two categories.
The articles belonging to the first category correspond to the initial stage, which aims at creating a basic foundation for cognitive structure corresponding to that event. This

[6] https://en.wikipedia.org/wiki/2018_Kerala_floods

Time Period	Main Focus
11th - 13th Aug	Event initiation and description of the damage being caused
14th -16th Aug	Rescue operations
17th - 19th Aug	Support from various state governments
20th - 27th Aug	Damage: deaths, casualties and financial loss
27th Aug - 1st Sep	Role of various political and other bodies
4th - 11th Sep	Damage caused to various industries and their specific support measures
12th Sep - 27th Oct	Post flood analysis - damage, causes, displacement of people, etc.
28th Oct - 7th Nov	Employment and Livelihood options for victims
8th - 24th Nov	Complaints and political accusations among various local and national parties
25th Nov - 4th Dec	Finances - loss, expenditure, compensation, etc.
5th Dec - 10th Jan	Rehabilitation measures

Table 2: Focus shifts in the local articles - August 2018 to January 2019

Time Period	Main Focus
9th - 12th Aug	Initiation of event, damage caused
12th - 20th Aug	Response of national level political parties
20th - 26th Aug	Solidarity and support from the country and other parts of the world

Table 3: Focus shifts in the national articles

was mainly done by the key words, facts and primary ideas in the initial reports.

The articles belonging to the second category correspond to the secondary stage of the framework, also known as the "mapping stage", where the readers further develop the cognitive structure representing the event by mapping new information onto the already existing foundation (given information).

Most of the articles belonging to this category had a considerable increase in the frequency of conceptual/event anaphora including pronouns and the definite article. Similarly, many headlines were observed to be elliptical which omitted the mentioning of keywords or phrases which were understood by the reader from the existing cognitive structure of that event. For instance, the headline of an article related to the unrest caused in the state of West Bengal in India for a separate state of Gorkhaland read, "Continues on the 70th day", which left it to the reader to deduce that it was related to the Gorkhaland agitation, from the existing cognitive structure which contained the information that the agitation was already going on since the past 69 days.

Time Period	Main Focus
12th-13th August	Floods caused severe damage in Kerala, India

Table 4: Focus shifts in the International articles

On the whole, our results agreed with the initial hypothesis that the closest level of reporting i.e the local articles would have the maximum shift in focus. So, we can say that the reader is ultimately building a mental superstructure of a single event by linking all the related articles that show the shift in focus in a global discourse context.

6. Conclusions and Future Work

In this paper, we present a novel method of event oriented corpus creation along with our corpus "Manovaad-v1.0". This is a layered corpus in terms of the classification of articles into three categories. When compared to the other existing news corpora in English such as the North American News Text Corpus[7], The AQUAINT Corpus of English News Text[8] which are flat and have no layered representation, our method provides more insight into the content of the corpus. This additional information has proved to be useful in giving us novel parameters such as focus shifts and subjectivity analysis to understand the flow of an event in a multifaceted manner.

Our unique way of looking at events and analysing subjectivity from various proximity levels makes our corpus stand apart from the other Subjectivity Corpora such as the MPQA 3.0 (Deng and Wiebe, 2015), which mainly recognizes sentiment and polarity towards entities and events. Our method of presenting the corpus can also greatly improve the search efficiency in large event databases and archives by enabling category based or event based search filters in addition to the existing filters such as date, author or newspaper.

6.1. Future Work

Using "Manovaad" and its future versions with even more data, many machine learning algorithms can be applied and the results can be very useful in various fields of discourse analysis.

While working on subjectivity and focus shift, we observed phenomena such as ellipsis, presupposition and implicature in the headlines and articles. We plan to extend the use of our corpus to study the above mentioned pragmatic notions and understand their role in the news articles.

[7] https://catalog.ldc.upenn.edu/LDC95T21

[8] https://catalog.ldc.upenn.edu/LDC2002T31

We are currently working on efficient presupposition identification techniques in event analysis which can contribute to better results in arriving at focus shifts. This way, we aim to extend the work on focus shifts and structure building framework to contribute to the domains of journalism and computational psycholinguistic studies on cognition and language understanding.

7. Acknowledgements

We would like to thank Meghana Bommadi, Varshit Battu and our team of annotators for their valuable contribution.

8. Bibliographical References

Arbaoui, B., De Swert, K., and van der Brug, W. (2016). Sensationalism in news coverage: A comparative study in 14 television systems. *Communication Research*, page 0093650216663364.

Augenstein, I., Rocktäschel, T., Vlachos, A., and Bontcheva, K. (2016). Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Brown, G., Brown, G. D., Brown, G. R., Gillian, B., and Yule, G. (1983). *Discourse analysis*. Cambridge university press.

Deng, L. and Wiebe, J. (2015). MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado, May–June. Association for Computational Linguistics.

Fowler, R. (1991). *Discourse and Ideology in the Press*. Routledge.

Gernsbacher, M. A. (2013). *Language comprehension as structure building*. Psychology Press.

Godbole, N., Srinivasaiah, M., and Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. *Icwsm*, 7(21):219–222.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Halliday, M. A. (1967). Notes on transitivity and theme in english: Part 2. *Journal of linguistics*, 3(2):199–244.

Liu, B. et al. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Mencher, M. and Shilton, W. P. (1997). *News reporting and writing*. Brown & Benchmark Publishers.

Mourad, A. and Darwish, K. (2013). Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 55–64.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Richards, B. and Rees, G. (2011). The management of emotion in british journalism. *Media, Culture & Society*, 33(6):851–867.

Schudson, M. (2001). The objectivity norm in american journalism. *Journalism*, 2(2):149–170.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Steele, C. A. and Barnhurst, K. G. (1996). The journalism of opinion: Network news coverage of us presidential campaigns, 1968–1988. *Critical Studies in Media Communication*, 13(3):187–209.

Van Dijk, T. A. (2013). *News as discourse*. Routledge.

Wahl-Jorgensen, K. (2013). Subjectivity and story-telling in journalism. *Journalism Studies*, 14(3):305–320.

Yang, X., Zhang, X., Wang, C., Chang, R., and Li, W. (2017). The interplay between topic shift and focus in the dynamic construction of discourse representations. *Frontiers in Psychology*, 8:2184.

9. Appendix

Following is the list newspapers from which we collected the data for our corpus creation.

1. Times of India
2. The Hindu
3. The Indian Express
4. India Today
5. Al Jazeera
6. Pakistan Daily
7. Hindustan Times
8. ESPN Sports Network
9. BBC
10. New York Times
11. Ahmedabad Mirror
12. Deccan Chronicle
13. The Guardian
14. The Economic Times
15. Daily Mail
16. Business Standard
17. The Telegraph
18. France 24
19. The Independent
20. Telangana Today

21. Daily News and Analysis (DNA)
22. CNN - News18
23. Jakarta Post
24. Reuters
25. Mylapore Times
26. Mathrubhumi
27. Times of Israel
28. Mangalorean
29. Atlantic
30. The Hans India
31. The Tribune
32. Daily Express UK
33. Houston Chronicle
34. The Financial Express
35. Huffington Post
36. The Myanmar Times
37. The Burma Times
38. Evening Standard
39. The Irish Times
40. The Siasat Daily