Examining Linguistic Biases in Telegram with a game theoretic analysis *

Sviatlana Höhn, Nicholas Asher, Sjouke Mauw

▶ **To cite this version:**

Sviatlana Höhn, Nicholas Asher, Sjouke Mauw. Examining Linguistic Biases in Telegram with a game theoretic analysis *. 3rd Multidisciplinary International Symposium on Disinformation in Open Online Media (MISDOOM 2021), Oxford Internet Institute, Sep 2021, Oxford (virtual), United Kingdom. pp.1-15. ⟨hal-03328712⟩

HAL Id: hal-03328712

https://hal.science/hal-03328712v1

Submitted on 30 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Examining Linguistic Biases in Telegram with a game theoretic analysis*

Sviatlana Höhn[1] [0000−0003−0646−3738]
, Nicholas Asher[2] [0000−0002−7689−8246], and Sjouke Mauw[1] [0000−0002−2818−4433]

[1] DCS/SnT, University of Luxembourg, Esch-sur-Alzette, Luxembourg
{sviatlana.hoehn|sjouke.mauw}@uni.lu
[2] IRIT, France Nicholas.Asher@irit.fr

Abstract. Selective formulations and selective reporting of facts in political news are deliberately used to create particular identities of different political sides. This becomes evident in media dialogue reporting about political conflicts. In contrast to most NLP-based studies of linguistic bias, we engage critically with its nature, aiming at a later de-biasing or at least raising awareness about linguistic bias in political news. We found inspiration in conversation analysis (CA), membership categorisation analysis (MCA) and a game-theoretic approach to discourse called epistemic message exchange (ME) games. We identified three types of bias: selective reports about facts, selective formulations when reporting about the same facts, and different histories built up by the differences in the first two. We extend the epistemic ME games model with findings from a qualitative study.

Keywords: Linguistic bias · Epistemic message exchange games · Political news · Membership categorization analysis.

1 Introduction

Different political parties use different formulations to describe the same events in order to create different opinions and attract voters. Neutral, unbiased descriptions are difficult to find. The most recent critical survey on bias in NLP by Blodgett et al. [6] emphasises that the majority of scholarly articles on NLP-based bias analysis or detection fail to engage critically with the nature of bias. One particular element that needs analysis is the role of the discourse or conversational structure in the framing bias.

Our **research objective** is to apply a formal model of linguistic bias [2] that takes into account the dynamics and structure of discourse to descriptions of events after the 2020 Belorussian elections in state and opposition media, which offer an interesting use case for models of bias. We show that the formal model explains causal factors of linguistic bias in our data. We also show how the model captures aspects of approaches to bias in conversational analysis, in particular the important function of labeling. Linguistic labels help to give content to *types*, a key element, in epistemic ME games.

*We thank the ANR PRCI grant SLANT, the Luxembourgish National Research Fund, INTER-SLANT 13320890 and the 3IA Institute ANITI funded by the ANR-19-PI3A-0004 grant for research support.

Labelling is also a strategic device. In political news, biases are used purposefully and consciously; labelling is not systematic, but invented and opportunistic; we also show they evolve over time by considering the extended discourse structure of the interactions between different media sources. We show in Sec. 4.1 how labelling choices manipulate meaning in order to construct particular identities for various actors and make actions towards particular social categories accountable. Following [9, p. 6], "a warranted analysis of the contextual meaning of the categorisation is based only on evidence in the text analysed." We have tried to interpret the use of labels cautiously.

2 The Basic Model

Most NLP-based bias detection models work with the definition of linguistic bias as *"a systematic asymmetry in word choice" reflecting "social-category cognitions*" [5]. The key idea is that, using social category labels for individuals communicates category and stereotype-congruent information, which can be benevolent or harmful [8].

For us, linguistic bias manifests itself not only at the lexical level but at the discursive level as well. In addition, linguistic bias is the product not only of choices of the author of a text but also of its interpreter, and the choices the author makes are geared to how the interpreters will understand them. This reflects a theme of conversational analysis, on which speakers construct their utterances for a specific recipient in a specific context (recipient design). To frame the issue of bias, let us suppose that our author $\mathcal{A}$ wants to convey information about some event or object e, which we formalize as the set of formulas $F(x)$, satisfiable by e.[3]

As shown in [2], an author's bias reveals itself in part in what set of facts $C_F(x) \subset F(x)$ about e she chooses to convey, what lexical choices she makes to describe $C_F(x)$ (lexical semantics), how those lexical contents combine together (compositional semantics) and how they weave their descriptions of elements of $C_F(x)$ into a consistent and coherent story, narrative or what [4] call a *history*. To build a history, the author must link the chosen basic facts together with semantic or what are known as *discourse relations* that convey causal, temporal, or thematic information. A narrative should make clear how each object or event in $C_F(x)$ chosen by the author fits into a coherent whole. The author's bias thus manifests itself at the level of lexical semantics, compositional semantics and discourse semantics.

To give an illustration, compare Ex. 1 posted by a Belorussian state news channel ONT and Ex. 2 posted by an opposition news channel TUT.BY. The screenshots from the videos that were part of the messages look very similar, they were recorded at the same time and at the same place, the so-called Square of Changes[4]. The corresponding text messages, however, emphasize different aspects of the events reported. While Ex. 1 complains about noise caused by "an aggressive minority" on a weekend, Ex. 2 reports about people who chanted the opposition slogans and refers to the video, leaving the interpretation to the reader; it does not mention "virtually completely barricaded yard

[3] We say *satisfiable* because $\mathcal{A}$ may of course choose to convey falsehoods about e; our only constraint is that $\mathcal{A}$ only conveys content that is logically and semantically consistent (i.e. does not violate selectional or other restrictions).

[4] https://en.wikipedia.org/wiki/Square_of_Changes

and entrances to it". Ex. 3 posted by the same opposition news channel reports about law enforcement bodies who use stun grenades. The state channel ONT does not say anything about using stun grenades on that day.

Screenshot Example 1

Screenshot Example 2

Example 1. ONT 15.11.2020 12:48 https://t.me/ontnews/21504

On Sunday, ordinary people want to rest, but the protesters don't think about them, the people who live in these houses! Noise, yelling, wild chants and car horns, a virtually completely barricaded yard and entrances to it... A typical example of how an aggressive minority can poison the life of an entire city.

Example 2. TUT.BY 15.11.2020 12:04 https://t.me/tutby_official/19429

Minsk. This is what the Square of Changes looked like at 1:45 p.m.
People chanted "We believe, we can, we win!"

Example 3. TUT.BY 15.11.2020 12:21 https://t.me/tutby_official/19429

Law enforcement bodies arrived at the "Square of Changes" in Minsk. They use stun grenades - eyewitnesses report four explosions.

An additional parameter in bias comes from the interpreter. The author may choose ambiguous expressions or leave certain discourse connections unspecified. It is then the interpreter's role to resolve the ambiguities and create a coherent history about e. It is in the interaction of author and interpreter choices that the game theoretic side of bias becomes clear. For $\mathcal{A}$ will make her choices in the light of how she thinks her interpreter I will construe those choices, in particular how I chooses to resolve the ambiguities and to fill in the underspecified elements. And in turn I will make her choices based on her beliefs about $\mathcal{A}$. Biases are concretized and conveyed via the interaction of $\mathcal{A}$'s and I's conversational strategies.

Our model of bias comes from the game-theoretic framework of Epistemic Message Exchange (ME) games [4], but it has links to membership categorisation analysis (MCA) [17, 19] and conversation analysis (CA) [18] . In particular we use the MCA of labeling [16, 17]. Labels incorporate a range of prototypical associations that authors and their interlocutors can exploit to draw inferences. Categories may have constitutive features that at least partially define the denotation of the label, but also occasioned features, features that members of the category on occasion possess that one can exploit for strategic purposes. Following [6], we explicitly include the effect of linguistic bias

in our formal model, i.e. what bias is harmful in what way and to whom. MCA has been successfully applied to understand identity construction in language [10, 14]; but it is difficult to formalize let alone operationalize. Epistemic ME games, on the other hand, build on a sophisticated formal analysis of discourse and conversational structure, which permits us to capture important insights of MCA for bias [4].

3 The Formal Model of Bias

[4] formalizes the intuitive picture presented in Sec. 2 in the following way. A Message Exchange (ME) game involves two players 0 and 1, each with a set of discourse moves, V_0 and V_1. Formally,

Definition 1. *A* Message Exchange game (ME game)*, $\mathcal{G}$, is a tuple $((V_0 \cup V_1)^\infty, \mathcal{J})$ where $\mathcal{J}$ is a Jury.*

In the definition, the Jury determines which player (or players) has achieved her goal in the conversation; in other words, it fixes the winning conditions in an objective fashion for the players. The Jury is typically an agent distinct from the players 0 and 1 of a ME game, but we can also sometimes identify the Jury with one of the players.

Definition 2 (Jury). *The* Jury *of an ME game is a tuple $\mathcal{J} = (Win_0, Win_1)$ where $Win_i \subset (V_0 \cup V_1)^\infty$ for each i.*

An ME game proceeds in turns where, by convention, player 0 starts the game by playing move x_1, player 1 follows with x_2, player 0 then plays x_3 and so on. These moves are understood to be formulas of a language V representing the semantic content of natural language conversational turns; as such they will include not only formulas representing individual items in $C_F(e)$ but also the semantic relations holding between them, as in Semantic Discourse Representation Theory (SDRT) [3, 4]. This results in the sequence $x_1x_2x_3\ldots$. Given our language V, this sequence is a concatenation of formulas from $V_0 \cup V_1$, where concatenation is viewed as conjunction. Consider the conversation between two conversationalists, our 0 and 1 in Ex. 4.

Example 4. 30.09.2020 14:15 Belarus Seychas https://t.me/belarusseichas/12032
Basketball player Elena Levchenko was sentenced to 15 days of administrative arrest. Shouts of "Shame" were heard in the hall.

(1) a. ρ_1 = (*Basketball player Elena Levchenko was sentenced to 15 days of administrative arrest. Shouts of "Shame" were heard in the hall* 0)
b. ρ_2 = (*Do you know why they shouted "Shame"?*, 1)

Assume player 0 plays the sequence ρ_1. This sequence yields a formula of V_0—a pair consisting of the V formula together with the index 0 for player $[(\langle \pi_1 : \phi_1 \rangle \wedge \langle \pi_2 : \phi_2 \rangle \wedge \mathcal{R}(\pi_1, \pi_2)), 0)]$ where π_1 and π_2 mark *elementary discourse units* or EDUs given by the two sentences in (1-a), and $\mathcal{R}$ is a relation on such discourses. Player 1 then plays the sequence ρ_2 which translates into a formula of V_1, itself a pair consisting of a formula in V for the EDU introduced by the question paired with 1. This results in the sequence $\rho_1\rho_2$. This motivates the following definition of a play of an ME game.

Definition 3 (Play). *A* play ρ *of an ME game is a sequence in* $(V_0 \cup V_1)$.

ρ can be underspecified like that for ρ_1 above where the semantic connection between the sentencing and the shouts is left open. This motivates the following:

Definition 4 (History). *A* history h *of an ME game is a play that is a semantically fully specified unit.*

Given a play ρ, $\mathcal{H}(\rho)$ denotes the set of all histories generated by specifying or removing ambiguities in ρ. $\mathcal{H}(\rho)$ can contain multiple, distinct, even incompatible histories. For example there are at least two possible histories for the play ρ_1 in (1-a): (i) one in which shouts of shame are a Result of the *sentencing* by the institution—and hence the shouts of shame are directed towards the sentencing institution; (ii) one in which the shouts of shame are an Acknowledgment and Comment on the player's behavior that led to the sentencing and thus directed towards her.

(2) Histories for ρ_1 in (1)

a. $\mathsf{h}_1(\rho_1) = [(\langle \pi_a : \phi_a \rangle \wedge \langle \pi_b : \phi_b \rangle \wedge \mathsf{res}(\pi_a, \pi_b)), 0]$

b. $\mathsf{h}_2(\rho_1) = [(\langle \pi_a : \phi_a \rangle \wedge \langle \pi_b : \phi_b \rangle \wedge \mathsf{ack}(\pi_a, \pi_b)), 0]$

Let $|\rho|$ denote the number of turns in a play ρ and $|\mathcal{H}|$ denote the same for $\mathcal{H}$. We let $\mathcal{P}$ (resp. $\mathcal{H}$) denote the set of all plays (resp. histories).

Definition 5 (Winning plays/histories). *A play* ρ *(or history* h*) is said to be* winning *for player i if* $\rho \in Win_i$ *(or* $\mathsf{h} \in Win_i$*).*

Players' strategies are an important element for developing and conveying biases. A strategy of player i tells us how i reacts to player $1-i$'s moves.

Definition 6 (Pure strategy). *A pure strategy* σ_i *for player i in an ME game is a function from the set of* $(1-i)$*-plays to moves in* V_i^+*, the finite positive sequences in* V_i^**. That is,* $\sigma_i \colon \mathcal{P}_{(1-i)} \to V_i^+$*. Let* S_i *denote the set of strategies for player i and let* $S = S_0 \times S_1$.

Let $\rho = x_0 x_1 \ldots$ be a play in an ME game and let $\rho_j = x_0 x_1 \ldots x_j$ for $j > 0$ be the set of prefixes of ρ. We say that ρ conforms to a strategy σ_i of player i if for every $(1-i)$-play ρ_j, $x_{j+1} = \sigma_i(\rho_j)$. Given a finite play ρ, we let S_i^ρ denote the set of all strategies σ_i of player i such that ρ conforms to σ_i and let S^ρ denote the set of all strategy pairs (σ_0, σ_1) such that ρ conforms to (σ_0, σ_1).

To see some examples of strategies, let's return to (1). Suppose 0 has played ρ_1; one strategy of 1 is to play a clarification question $\rho_{2'}$ like *did you mean that the shouts of "Shame" were addressed to the court?* to understand better which history $\mathsf{h}_1(\rho_1)$ of (2-a) or $\mathsf{h}_2(\rho_1)$ of (2-b) was intended. Another strategy is to assume that the intended history was (2-a) and to ask for an explanation of why there were shouts of "Shame". It is this latter strategy that conforms to the actual play in ρ_1, ρ_2 of (1).

We now turn to the epistemic component of ME games. Players' beliefs, or the subjective probabilities they assign to plays, moves, and strategies affect how they reason in an ME game, i.e. what they say or how they react to some conversational turn. And for this, a player's beliefs must include beliefs about other players's strategies and beliefs about them. This nested structure of higher order beliefs (beliefs about beliefs) can be

expressed in different ways, but a natural way to do this is to exploit the type of a player [11]. The type of a player i is a property of the player that encodes his behaviour, the way he strategizes, his personal biases, etc. The $i-types$ for a player i are the possible properties, possible behaviors relevant to the ME game, that i could instantiate. Rubrics like "protester" and "police" describe types that we will use below. We will assume probability distributions, written $\Delta(A)$, for sets of types or strategies A. We will assume types for the players of our game as well as of the Jury.

Crucial to our view of bias, the beliefs of the players affect what content they get from a message and how those messages affect their beliefs. Following [4], we separate out the effect of types both on beliefs about other players and on interpretations of a conversation that result in particular histories.

Definition 7 (Belief function). *For every play* $\rho \in \mathcal{P}$ *the* (first order) belief $\hat{\beta}_i^\rho$ *of player* i *at* ρ *is a pair of functions* $\hat{\beta}_i^\rho = (\beta_i^\rho, \xi_i^\rho)$ *where* β_i^ρ *is the* belief function *and* ξ_i^ρ *is the* interpretation function *defined as:*

$$\beta_i^\rho \colon T_i \times \mathcal{H}(\rho) \to \Delta(T_{(1-i)}) \times \Delta(S^\rho_{(1-i)}) \times \Delta(T_\mathcal{J})$$
$$\xi_i^\rho \colon T_i \times T_{(1-i)} \times T_\mathcal{J} \to \Delta(\mathcal{H}(\rho))$$

The (first order) belief $\hat{\beta}_\mathcal{J}^\rho$ *of the Jury is described by a similar pair of functions.*

Intuitively, by fixing a type for the players and the Jury, the respective interpretation function says how they interpret the current play; that is, what are the probabilities that they assign to each possible history arising from the current play. The belief function returns the beliefs about the types and the strategies of the other players and/or the Jury given a history and a particular player type; together the interpretation and belief functions show a *codependence between beliefs and interpretation.*[5]

We now have the pieces to define our tool for analyzing linguistic bias:

Definition 8. *An* Epistemic Message Exchange game (Epistemic ME game)*,* $\mathcal{G}$*, is an ME game, with set of types for the players and the Jury and belief functions for* 0*,* 1 *and the Jury, as defined in Definition 7.*

In some cases, the beliefs or the interpretations of the players or the Jury may be independent of one or more components or those components may be fixed.[6] In that case we can simplify our notation. For example, player i's beliefs concerning the type of player $(1-i)$ and her strategies might be independent of what player i believes about the type of the Jury. In that case the belief of i is the function $\beta_i^\rho : T_i \times \mathcal{H}(\rho) \to \Delta(T_{1-i}) \times \Delta(S^\rho_{(1-i)})$. We will simplify the interpretation function similarly.

Let's return to Ex. (1) to see how types and interpretations might play out in a very simple scenario. Suppose we have two types for 0, roughly one, t_0^e according to which 0 intended to link π_b to π_a via the discourse relation of Result and another type t_0^r according to which 0 intended to link π_b to π_a via Acknowledgement. Suppose 1 only

[5] Using the definitions of first order beliefs, S, the set of strategies, and types, [4] define higher order beliefs, beliefs that players or the Jury have about the beliefs of other players (and the Jury) and fill out the epistemic picture of our players.

[6] For a definition of independence see [4].

has one type. In that case, the play ρ_1 together with $\beta_1^\rho \colon \mathcal{H}(\rho) \to \Delta(T_0)$ determines a probability distribution over the types for 0. In turn these types via $\xi_1^\rho \colon T_0 \to \Delta(\mathcal{H}(\rho))$ determine a probability distribution over the two histories (2-a) and (2-b) for player 1. [4] shows how such distributions evolve as a conversation proceeds.

4 Analysing Linguistic Bias with ME Games

For our empirical study, we have focused on descriptions of events in Belorussian Telegram channels—including official Belorussian news channels ONT NEWS, BelTA and Pool Pervogo, and opposition channels BelSAT, Belarus Seychas, and TUT.BY with posts in Russian and Belorussian languages. The analysis is mostly based on the dataset of 140.388 Telegram posts (76.918 opposition and 63.470 state); 109.721 posts contain text (58.976 opposition and 50.745 state). In what follows, the protests will be the event e, and the tweets will select sets of formulas $C_F^t(x)$ to construct histories, depending on the type t of the author.

Media publications through these channels build up a dialogue between conflicting parties with distinct strategies. The recipients are distributed over time and space (not restricted to e.g. a single TV discussion), and there are multiple groups within the recipients. As Fig. 4 shows, the dialogue evolves over time. The quantity of protester contributions (dotted line) follows an inverse power law with intermittent peaks reflecting the increase of activities during the regular weekend marches, and also some extraordinary events such as the inauguration of Lukashenko on September 23, strikes on October 26, the death of Roman Bondarenko on December 11 (peak from 15/11). The most tweets from the opposition come right after the elections and with the first protests, and then gradually die down. Government posts (in black) stay relatively constant with certain peaks and gradually come to dominate in number the opposition posts. We analyze this

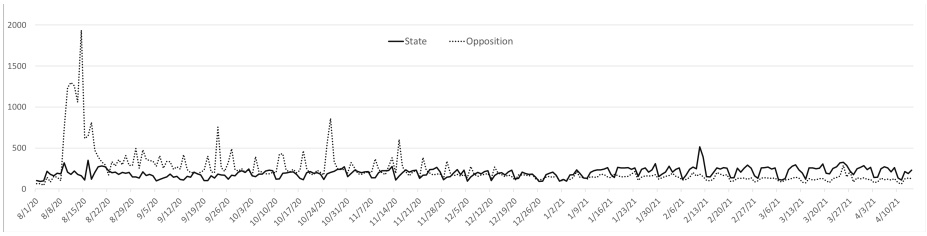

Fig. 1. Dynamics of the conversation

extended conversation as an ME game $\mathcal{G}$ between the protesters (0) and the government (1). Player i constructs a narrative during his turn either about some contemporaneous event or as a reply and "counter-narrative" to player $1-i$'s narrative on some previous turn. We'll consider two types t_0 for PROTESTER and t_1 for the type POLICE, which includes the government controlled media. $\mathcal{G}$ is zero sum; the winning condition for player i is to convince his readership that her history, which portrays the type t_{1-i} of player $1-i$ in negative terms, is the correct one. The Jury of this ME game are the interpreters or readers of the posts, either be of type t_0 or t_1, of the contribution. This Jury

will assign either favorable or unfavorable ratings to the author's play at a given turn. Player i wins in a game with Jury of type t_j just in case i has more favorable ratings than Player $1-i$. Figure 1 suggests from the number of posts that arguably player 1 wins $\mathcal{G}$.

We will examine on the strategies that the players use in this game. We focus first on how authors try to invest the type of their opponents with content. We then turn to the dialogue-like structure of this exchange.

4.1 Identities of Protesters and Police

As we have said, authors of each type build quite different histories. Authors of type t will use $C^t_F(x)$ to make plays ρ^t. One of the principal tasks of a history built by t_i is to build a negative identity for the opponent t_{1-i} and indirectly then to paint t_i in a positive light. The identities of the interacting participants are constructed via their own contributions, the contributions of *all* participants *and* the entire interaction history [20]. The aim in these tweets for authors of type t_0 is to build sympathy and support for the protesters, and one strategy to do this is to depict the type of opposition as evil—though not always true, the adage, the enemy of evil is good, is an effective strategy. We've seen this strategy already at work in our Examples 1 and 2. Both examples use the word *people* to refer to some social categories. For ONT, *ordinary people* are those who are tired from protests, and protesters are an *aggressive minority*, while for TUT.BY *people* are the protesters.

The post in Ex. 5 by a player of type t_0, uses a definite description that brings with it a host of associated negative concepts to develop a strategic attack on t_1 (though the use of *punishers* needs some background history to be properly understood).

Example 5. 09.08.2020 20:56 Belsat https://t.me/belsat/10308

On Masherova Avenue in Minsk, ***people*** *clashed with* ***OMON***[7]*. At least* ***one of the punishers*** *had their head smashed.*

The term *punishers* was used to refer to a Nazi division operating in Belarus during the Second World War; the term thus is associated with a number of other concepts [nazi, soldiers, enemy, aggressors, defenders, partisans, army, ...]. It associates people of t_1 (via, in MCA terms, the membership-categorization device or MCD) with war, not protests. The historical usage and associations of *punishers* in turn define a characteristic activity of t_1 people: *acting with special cruelty against Belorussian people*. By using *punishers* in the context of the description of a protest, the author of t_0 implies that people of t_1 are waging war against people of t_0. Labelling with terms loaded with a historical meaning is thought by the authors to be an effective strategy for painting the opponent in negative terms, and to justify the injuries of the police caused by protesters.

Example 6. 09.08.2020 21:06 https://t.me/belsat/10321

In Mogilev, ***cosmonauts*** *block the streets.*

Attacks by t_i against t_{1-i} do not always use the strategy of depicting t_{1-i} as evil or cruel. In Ex. 6, the word *cosmonauts* refers to OMON officers wearing their full equipment

[7] Otryad Militsyi Osobogo Nasnacheniya, En.: Special police detachment

and helmets, visible in the accompanying photo. The discourse structure of this last example is rather complex as it involves multimodal information. The concept *cosmonauts* evokes a category like space travel and a collection of concepts like [cosmonauts, engineers, scientists, aliens, ...]. The defining property of this class includes *wearing protective clothes and helmets*. This attribute makes them visually similar to OMON police. A certain displacement and ridiculing happens when *cosmonauts* block streets in Mogilev (far away from any space-travel area). This is also an effective strategy to promote t_0, the type of the protesters. By ridiculing the opposition t_1, the author puts t_0 in a position of authority and *gravitas.* Both *punishers* in Ex. 5 and *cosmonauts* in Ex. 6 are marked: *punishers* with anger and fear, and *cosmonauts* with displacement.

The government information sites also use various labeling strategies to characterize their opponents in negative terms. State channels report events in the night from 9th to 10th August (bold added by us) as follows.

Example 7. 10.08.2020 10:57 ONT https://t.me/pressmvd/1890

On the night of 9 to 10 August 2020, ***Focal gatherings of citizens****... were recorded in the country.*

In total, about 3 thousand ***people*** *were detained throughout the country for participating in* ***unauthorized mass events****... As a result of the* ***clashes****, more than 50* ***citizens*** *were injured, as well as 39* ***police officers****, some of whom are currently hospitalized.*

In Minsk at 22.00 in the area of the stele "Minsk - Hero City", ***protesters*** *lit fireworks, threw spikes and nails on the roadway, erected barricades from mobile turnstiles, dismantled paving slabs and threw them and other objects at* ***law enforcement officers****.*

An active resistance to the ***law enforcement bodies*** *was rendered in Pinsk, where a* ***group of aggressively minded citizens****, using pointed stakes, rods, stones and reinforcement bars, tried to organize an attack on* ***police officers****.* ***Some of the citizens*** *taken to the country's medical institutions were in a* ***state of alcoholic intoxication****.*

! It should be noted that military weapons were not used against ***violators****. There are no fatalities.*

The news channel ONT refers to the protests first as *focal gatherings* and *unauthorized mass events*. Crucially no mention is made of why the protesters are gathering. The category *protesters* is used in the third paragraph of the news bulletin, but it assigns to them rather violent properties against *police officers* and *law enforcement bodies*. All in all the protesters are painted in a negative light, which justifies the actions of the police.

Example 8. 10.08.2020 11:54 BelTA https://youtu.be/BjS1uHqbRaY

Video: *We detained the* ***organizers*** *who were* ***hiding*** *and* ***running around the corner****. About three thousand - half of them in Minsk -* ***stoned****, Sergei Nikolaevich*[8]*, there are* ***many drunks****,* ***with drugs****, horror.*

The President of Belarus on the same day describes the events in Ex. 8 and then two days later introduces a new theme (Ex. 9).

Example 9. 12.08.2020 14:24 Pool 1, https://t.me/pul_1/1250 *Lukashenka: "The basis of all these* ***so-called protesters*** *are* ***people with a criminal past*** *and are* ***unemployed*** *today. There is no job, which means they can "walk the streets and avenues".*

[8] Lebedev, Executive Secretary of the CIS

Examples 8 and 9 characterize protesters as unemployed alcoholics, addicts and criminals. What is used as an occasional feature in Ex. 7 (*being in a state of alcoholic intoxication*) becomes at least a tight feature in Ex. 8 (*stoned, many drunks, with drugs*). Ex. 9 uses the description of activity *walk the streets and avenues* as a justification of the presence of people on the streets. This activity in turn, is a consequence of the people's unemployment. The marker *so-called* modifies *protesters* giving it an ironic or sarcastic reading *distancing* its meaning on this occasion from its normal one [12]. Table 1 summarizes the way t_0 and t_1 are characterized.

types	**by t_0 (protesters)**	**by t_1 (police)**
t_0 **(protesters)**	people, protesters	people, citizens, protesters, violators, so-called protesters, people with criminal past, unemployed, sheep, drunks
attributes		stoned, drunk, with drugs, people with criminal past, unemployed, aggressively minded, in a state of alcoholic intoxication
actions	gather in the center of Minsk, clash with OMON, smash heads of OMON, build barricades, throw bottles at OMON, break through the cordons	lit fires, threw spikes and nails on the roadway, built barricades, dismantled paving slabs and threw them and other objects at police, attack police officers, use pointed stakes, rods, stones and reinforcement bars, hiding, running around the corner, are being controlled, do not understand what they are doing
t_1 **(police)**	riot police, OMON, punishers, cosmonauts	law enforcement bodies, law enforcement officers, police officers
actions	clash with protesters, use flash bangs, block streets	did not use military weapons, detain organisers

Table 1. Labeling of *protesters* and *police* by opposition and state channels.

To sum up, different news channels use different labelling strategies, picking out different defining features for our types, to get complex messages across.

4.2 Interaction between PROTESTER and POLICE

We have examined strategies by the players in our game that are used on individual turns to convince their readership. Here we detail strategies for player i's replies to previous turns by $1-i$. Player 1 POLICE plays the move from Ex. 10:

Example 10. 10.08.2020 13:24 ONT t.me/ontnews/13864

*We identified calls from abroad. The calls came from Poland, UK and Czech Republic, they controlled our - excuse me - **sheep**: they do not understand what they are doing, and they are being controlled.*

In response, player 0 PROTESTER plays the move as illustrated in Ex. 11, in which a photo of a person holding a piece of white cardboard with text written on it in red letters conveys the message that protesters reject the attributes, such as *unemployed* and

sheep, assigned to them by an author of type t_1 in the previous message. The visual is an effective strategy; rather than the author verbally rejecting the negative labels provided by player 1 of type t_1, it is a winsome, smiling protester who is conveying the message rejecting the government's labelling strategy. In addition, she is using the symbolic colors of the opposition (red and white) to do it.

Example 11. 13.08.20 13:37 Belarus Sejchas https://t.me/belarusseichas/5827

Today in Minsk.

Text on the picture:
We are not sheep, we have jobs.

Another strategy by player 0 to attack the histories proposed by 1 is to point out inconsistencies and contradictions. Ex. 12 illustrates this. Part of 1's strategy is to attack the identity of player 0 via attributes not related to political content such as employment, alcohol consumption, bad parenting, and affiliation to particular profession. But this conflicts with other labelling strategies.

Example 12. Belsat 12.01.2021 12:26 https://t.me/belsat/38767

Video: "*The basis of this* ***protest*** *is made up of* ***these IT people*** *who are* ***snickering****, excuse me,* ***who were nearly kissed the ass****...*"

Text: *Wait, but some* ***alcoholics****,* ***drug addicts*** *and* ***parasites*** *come out to protest. Apparently,* ***state television*** *is finally* ***confused*** *in its own versions.*

4.3 Dynamics and bias hardening

Asher et al. [2] use the model explained in Sec. 3 to predict that interpreters' biases become more entrenched through the co-dependence of belief and interpretation: prior beliefs or the distribution over types will guide I to a particular interpretation. In turn, that interpretation can reinforce those initial beliefs over time. We see empirical evidence of bias hardening in the corpus.

One strategy is to appropriate a label from an opponent and reassign it in negative connotations. For example, opposition channels use the label 'unbelievable' as adjective to emphasize the bravery of the Belorussian protesters: *https://t.me/belarusseichas/7388* from 14/08/2020 shows a video of a peaceful demonstration with the text "*Unbelievable people*". State channels, however, then re-use this to label protesters 'the unbelievables' and to link it with actions described as meaningless or aggressive (Ex. 13).

Example 13. ONT 19.10.2020 18:37 https://t.me/ontnews/19205

Protests of the 'fighters' have long ceased to be peaceful. The participants intentionally take to the streets and provoke ordinary citizens, throwing themselves with aggression at those who do not agree with their views. The footage shows the unbelievables starting fights and doing everything they can to heat up the situation in society.

Over time, *the unbelievables* becomes synonymous with *pointless aggressive protests* and is used alone without further explanations. Fig. 2 shows the successful appropriation of the term *unbelievable* by state media. These examples show the non-cooperative

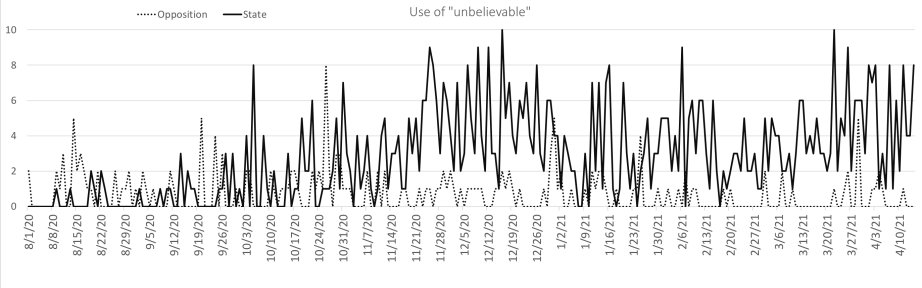

Fig. 2. State takes up the use of 'unbelievable' to label protesters

nature of the dialogue between state and opposition players. Participants do not co-construct meaning; rather, they present different versions of meaning to the readership or Jury, typed as POLICE and PROTESTER. We showed in Sec. 4.1 how selective formulations create different identities of protesters and police. The sequential organisation of those selective formulations results in different interpretations of entire conversations.

Example 14. Belarus Seychas 17.10.2020 15:03 https://t.me/belarusseichas/13064 *This is nothing new. It's just that the yabatskas stipulate the amount they get for participation in pro-Lukashenko events*

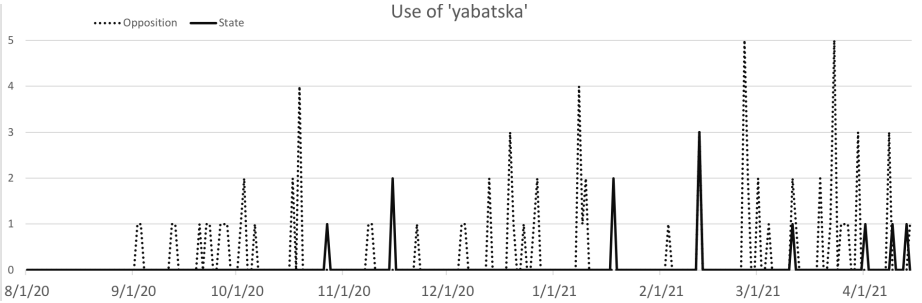

Fig. 3. Opposition takes up the use of 'yabatska' to label state supporters

Similarly, the opposition successfully appropriates the term *yabatska* which is a composite of *ya-mi-batskka, En.: me-us-father*. Originally used to express solidarity with the president (e.g., *https://t.me/belta_telegramm/15842*), the term taken up by the opposition media to describe state supporters as uncultivated, uninformed and unable to think critically (Ex. 14). Fig. 4.3 shows how the opposition appropriated the label *yabatska*.

4.4 The "neutral" view point

As shown in Sec. 3, a play may develop different histories depending on the type of the interpreter (see discussion of Ex. 4). Thus, the histories h^{f_0} and h^{f_1} constructed by interpreters of the two types may also differ, even though they both arise from the interpretation a single play ρ, as in Ex. 15:

Example 15. 09.08.2020 20:11 Belsat https://t.me/belsat/10272

*About **20-30 locals**, including **children**, gathered in the park on Hrybaedava Street (Minsk) near Stella. Two **paddy wagons** came to the park, **police officers** said that **people** had 2 minutes to **go away**, after which they started to **push people out of the park**. A Belsat correspondent witnessed how one **person** was* detained *but later **released**. After trying to **evict people from the park**, the **paddy wagons left** and the **people returned** to the park. There are 3 ambulances on duty, near to them **people in plainclothes** are standing.*

Ex. 15 labels persons as *locals* and *children*. The action attributed to them *gathered in the park* together with these labels evoke a scene of leisure activity in which parents and children are going to the park (some of those people were with children, they must be parents to those children)—which all sounds innocuous.

The narrative then changes as *paddy wagons* and *police officers* enter the scene and the actions now contrast with the scene of leisure activity evoked above. The police arguably confront the people by saying *that people had 2 minutes to go away* and then by starting to *push out people of the park*. These events need to be related to the gathering in a coherent history, but the author does not explicitly say why the paddy wagons arrived or the police acted in this way. The police actions towards *locals, children and people* arguably cause us to interpret a *gathering in the park* as an undesired event.

At this point an interpreter has two interpretive strategies with two distinct semantic relations relating the sentence contents: (1) The gathering was illegal and hence the actions of the police are a natural and legitimate Result; OR (2) The gathering is legal and in Contrast the police are acting in a wrong way.

These readings depend on the readers' prior beliefs and political preferences. Government supporters would read it as "police prevented escalation", opposition supporters would read it as "government uses power for oppression". This example shows how interpreters or readers contribute to a biased reading by inferring semantic relations between discourse units to form a coherent narrative.

The last sentence then introduces *people in plainclothes* which are arguably not the same as *locals* or just *people*, although *locals* usually wear plainclothes. The attribute *plainclothes* refers to the appearance of people who are supposed to wear something different but wear plainclothes, probably in order to hide their identity. Again depending on the type of the interpreter, this second paragraph has two messages: (1) a reassurance: police officers are still there, protecting law and order; OR (2) a warning: if you go to this park, you might by observed by the *people in plainclothes* or even *detained*.

Even messages that use only unmarked references to police and protesters, such as Example 16, will be colored by the reader's bias in a positive or negative way.

Example 16. 09.08.2020 21:04 Belsat https://t.me/belsat/10317

*In the center of Minsk **OMON** uses flash bangs against **protesters**.*

5 Discussion and Conclusions

Our corpus confirms [13] observations concerning *gatekeeping* or *selection* bias (the choice of a channel to report an issue or not), *coverage* bias (how much space in the media is dedicated to an event) and *framing* bias (the way a fact is presented). Our model, however, sharpens the notions of framing and coverage biases by linking them to strategies at the lexical and discursive level that can be opportunistic and evolve over time. We also see the confirmation of different levels of granularity in our corpus: category-level, message-level and media source level [1, 7].

In addition, our game theoretic model captures the important role of the recipient/interpreter of messages, even when those are neutrally formulated. We have shown how different histories arise from different although valid interpretations of the same message. In contrast to previous bias-detection work based on static models [15, 8], our model is able to deal with dynamics in linguistic bias. We explained how biases can harden into established opinion and exploit strategies for appropriating terms from an opponent for one's own discourse purposes. In line with [6], we have shown these discourse purposes can encode normative reasoning and specific rationales for physical harm and restrictions of freedoms of the opponents. Our qualitative study shows that complex historical background and values of the target recipients, which we can express as types in our model, play an important role in bias construction and detection. None of the existing static models is able to capture these factors. Embedding-based approaches may be helpful to find collocations and discover asymmetries based on word choice even for an unknown set of labels [8], however, they cannot discover omitted facts or details, nor are they able to express the dynamics of the conversation. The training corpora used for any machine-learning approach only provide a snapshot view on the histories, and the models need to be continuously retrained on new data in order to learn new labels. The opportunistic nature of labelling, as explained in Sec. 4.3, questions lexicon and embedding-based approaches. In addition, seemingly neutral labels (such as *people* vs. *ordinary people* vs. *people in plainclothes*) are usually not considered as potentially biased, especially in lexicon-based approaches, such as [1].

Finally, our study issues a challenge to automated de-biasing of political news. A completely neutral viewpoint does not exist. If it were to exist, it should be non-selective in terms of issues to report (what is important enough to be reported?), equally covering (all parties must have access to all channels equally), and non-selective in terms of formulations (lexical choice) and details (how complete is the picture?). Examples in Sec. 4.1 illustrate this finding.

This study has empirical limitations: we analysed only one political event in only one country. Although we understand the import of the cultural context, more comparison is needed with other events of a similar controversial degree, e.g. Navalny protests in Russia, the US Black Lives Matter movement, anti-Corona restrictions movement (mis)used by right radicals and protests in Hong-Kong, just to name a few. Analyzing such data is also technically difficult. Messengers like Telegram typically become the main source of communication in many political conflicts. They galvanize public opinion and can move masses of people in real time. However, messages with photos and videos pose challenges to computational analysis beyond those from newspaper articles.

References

1. Aleksandrova, D., Lareau, F., Ménard, P.A.: Multilingual sentence-level bias detection in Wikipedia. In: Proc. RANLP 2019. pp. 42–51 (2019)
2. Asher, N., Hunter, J., Paul, S.: Bias in semantic and discourse interpretation. Linguistics and Philosophy (2021), in press
3. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge University Press (2003)
4. Asher, N., Paul, S.: Strategic conversations under imperfect information: epistemic message exchange games. Journal of Logic, Language and Information **27**(4), 343–385 (2018)
5. Beukeboom, C.J., Burgers, C.: Linguistic bias. In: Oxford Research Encyclopedia, Communication. Oxford University Press (2020)
6. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in NLP. In: Proc. 58th ACL Meeting. pp. 5454–5476. ACL (2020)
7. Chen, W.F., Al-Khatib, K., Wachsmuth, H., Stein, B.: Analyzing political bias and unfairness in news articles at different levels of granularity. arXiv preprint arXiv:2010.10652 (2020)
8. Ferrer Aran, X., van Nuenen, T., Such, J., Criado Pacheco, N.: Discovering and Categorising Language Biases in Reddit (2020)
9. Freiberg, J., Freebody, P.: Applying membership categorisation analysis to discourse: When the 'tripwire critique' is not enough. Critical discourse analysis: An interdisciplinary perspective pp. 49–64 (2009)
10. Gibson, W., Roca-Cuberes, C.: Constructing blame for school exclusion in an online comments forum: Membership categorisation analysis and endogenous category work. Discourse, Context & Media **32**, 100331 (2019)
11. Harsanyi, J.C.: Games with incomplete information played by "bayesian" players, i–iii part i. the basic model. Management science **14**(3), 159–182 (1967)
12. Härtl, H.: Name-informing and distancing sogenannt 'so-called': Name mentioning and the lexicon-pragmatics interface. Zeitschrift für Sprachwissenschaft **37**(2), 139–169 (2018)
13. Lazaridou, K., Krestel, R.: Identifying political bias in news articles. Bulletin of the IEEE TCDL **12** (2016)
14. McLay, K.F.: Geeks, gamers, and girls: revealing diverse digital identities with membership categorisation analysis. Discourse: Studies in the Cultural Politics of Education **40**(6), 946–961 (2019)
15. Reddy, R.R., Duggenpudi, S.R., Mamidi, R.: Detecting political bias in news articles using headline attention. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 77–84 (2019)
16. Sacks, H.: On the analyzability of stories by children. In: Directions in Sociolinguistics. The Ethnography of Communications, pp. 325–345. Holt, Rinehart and Winston, NY (1972)
17. Sacks, H.: Lectures on Conversation, vol. 1 and 2. Blackwell (1995)
18. Schegloff, E.A.: Sequence Organization in Interaction: Volume 1: A Primer in Conversation Analysis. Cambridge University Press, 1 edn. (2007)
19. Schegloff, E.A.: A tutorial on membership categorization. Journal of pragmatics **39**(3), 462–482 (2007)
20. Spranz-Fogasy, T.: Interaktionsprofile: Die Herausbildung individueller Handlungstypik in Gesprächen. Radolfzell: Verlag für Gesprächsforschung (2002)