Natural Language Engineering (2024), **30**, pp. 385–416
doi:10.1017/S1351324923000220

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Polish natural language inference and factivity: An expert-based dataset and benchmarks

Daniel Ziembicki[1], Karolina Seweryn[2,3] and Anna Wróblewska[3]

[1]Department of Formal Linguistics, University of Warsaw, Warsaw, Poland, [2]NASK - National Research Institute, Warsaw, Poland, and [3]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland
Corresponding author: Daniel Ziembicki; Email: daniel.ziembicki@uw.edu.pl

(Received 4 January 2022; revised 27 February 2023; accepted 2 April 2023; first published online 1 June 2023)

Abstract

Despite recent breakthroughs in Machine Learning for Natural Language Processing, the Natural Language Inference (NLI) problems still constitute a challenge. To this purpose, we contribute a new dataset that focuses exclusively on the factivity phenomenon; however, our task remains the same as other NLI tasks, that is prediction of entailment, contradiction, or neutral (ECN). In this paper, we describe the LingFeatured NLI corpus and present the results of analyses designed to characterize the factivity/non-factivity opposition in natural language. The dataset contains entirely natural language utterances in Polish and gathers 2432 verb-complement pairs and 309 unique verbs. The dataset is based on the National Corpus of Polish (NKJP) and is a representative subcorpus in regard to syntactic construction [V][że][cc]. We also present an extended version of the set (3035 sentences) consisting more sentences with internal negations. We prepared deep learning benchmarks for both sets. We found that transformer BERT-based models working on sentences obtained relatively good results ($\approx$ 89% F1 score on base dataset). Even though better results were achieved using linguistic features ($\approx$ 91% F1 score on base dataset), this model requires more human labor (humans in the loop) because features were prepared manually by expert linguists. BERT-based models consuming only the input sentences show that they capture most of the complexity of NLI/factivity. Complex cases in the phenomenon—for example, cases with entitlement (E) and non-factive verbs—still remain an open issue for further research.

Keywords: Semantics; Pragmatics; Natural language inference; Presupposition; Factivity

1. Introduction

Semantics is still one of the biggest problems of Natural Language Processing.[a] It should not come as a surprise; semantic problems are also the most challenging in the field of linguistics itself (see Speaks 2021). The topic of presupposition and such relations as entailment, contradiction, and neutrality are at the core of semantic-pragmatic research (Huang 2011). For this reason, we dealt with the issue of factivity, which is one of the types of presupposition.

The subject of this study includes three phenomena occurring in the Polish language. The first of them is **factivity** (Kiparsky and Kiparsky 1971; Karttunen 1971b). The next phenomenon is the three relations: **entailment**, **contradiction**, and **neutrality** (ECN), often studied in Natural Language Inference (NLI) tasks. The third and last phenomenon is utterances with the following syntactic pattern: "**[verb][że][complement clause]**." The segment *że* corresponds to the English segments *that* and *to*. The above syntactic structure is the starting point of our research. In order to look at the phenomenon of presupposition, we collected the dataset based on the National Corpus

[a]See relatively new and important work related to this topic: (Richardson *et al.* 2020; Zhang *et al.* 2020).

of Polish (NKJP). The NKJP corpus is the largest Polish corpus which is genre diverse, morphosyntactically annotated and representative of contemporary Polish (Przepiórkowski *et al.* 2012).[b] Our dataset is a representative subcorpus of the NKJP. The analysis of our dataset allowed us to make a number of findings concerning the factivity/non-factivity opposition. We then trained models based on prepared linguistic features and text embedding BERT variants. We investigated whether the modern machine learning models handle the factivity/non-factivity opposition.

Thus, in this paper, our contributions are as follows:

- gathering a new dataset *LingFeatured NLI*, based on fully natural utterances from (NKJP). The dataset consists of 2432 "verb-complement" pairs, that is sentences (denoted as *Theses* in the following text and examples) with their clausal/complements (denoted as *Hypotheses* in the following text and examples). It was enriched with various linguistic features to perform inferences of the utterance relation types, that is entailment, contradiction, neutral **(ECN)** (see Section 3). To the best of our knowledge, it is the first such dataset in the Polish language. Additionally, all the utterances constituting the dataset were translated into English.
- creation of an extended version of a dataset by adding negation to original sentences. This dataset contains 3035 observations and is about 25% larger than the original dataset. The purpose of assembling new sentences was, in particular, to increase the number of observations belonging to class C (from 107 to 162).
- analyzing the above dataset and presenting conclusions on the phenomenon of presupposition and the **ECN** relations.
- building ML benchmark models (linguistic feature-based and embedding-based BERT) that predict the utterance relation type **ECN** (see Section 5).[c]

In the following, Section 2 describes theoretical background of the factivity phenomenon. Then, Section 3 introduces our new dataset *LingFeatured NLI* with a commentary on its language background, annotation process, and features. Section 4 shows the observations from the dataset. The next Section 5 shows our machine learning modeling approach and experiments. Further, in Section 6, we analyze the results and formulate findings. Finally, we summarize our work in Section 7.

2. Linguistic background

2.1. Linguistic problems

In the linguistic and philosophical literature, the topic of factivity is one of the most disputed. The work of Kiparsky and Kiparsky (1971) began the never-ending debate about presupposition and factivity in linguistics. This topic was of interest to linguists especially in the 1970s (see, e.g., Karttunen 1971b; Givón 1973; Elliott 1974; Hooper 1975; Delacruz 1976; Stalnaker, Munitz, and Unger 1977). Since the entry of the term *factivity* into the linguistic vocabulary in the 1970s, there have been many, often mutually exclusive, theoretical studies of this phenomenon.

Karttunen's (2016) article with the significant title *Presupposition: What went wrong?* emphasized this fact. He mentioned that factivity is the research area that raised the most controversy among linguists. It should be clearly stated that no single dominant concept explaining the phenomenon of factivity has emerged so far. Nowadays, new research on this topic appears constantly (see, e.g., Giannakidou 2006; Egré, 2008; Beaver 2010; Tsohatzidis 2012; Anand and Hacquard

[b]The NKJP does not contain too many speech data, but it is currently the biggest and the most representative corpus for contemporary Polish.

[c]The dataset, its translation and the source code, and our ML models are attached as supplementary material and will be publicly available after acceptance of the article.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

2014; Kastner 2015; Djärv 2019). The clearest line of conflict concerns the question in which area to situate the topic of factivity: semantics or pragmatics. There is also a dispute about discursive concepts (see, e.g., Abrusán 2016; Tonhauser 2016; Simons *et al.* 2017). An interesting example from our research point of view is the work of (Djärv and Bacovcin 2020). These authors argue against the claim that factivity depends on the prosodic features of the utterance, pointing out that it is a lexical rather than a discursive phenomenon.

In addition to the disputes in the field of linguistics, there is also work (see, e.g., Hazlett 2010; Hazlett 2012) of a more philosophical orientation which strikes at beliefs that are mostly considered accurate in linguistics, for example that *know that* is a factive verb. These works, however, have been met with a distinctly polemical response (see, e.g., Turri 2011; Tsohatzidis 2012).

In summary, **the first problem** to note is the clear differences in theoretical discussions of the phenomenon of factivity-based presupposition. These take place not only in the older literature, but also in the contemporary one.

Theoretical differences are related to **another issue**, namely the widespread disagreement about which verbs are factive and which are not. A good example is a verb *regret that*, which, depending on the author, is factive or not or presents a different type of factivity from the paradigmatic in the class of factive expressions verb *know that*.[d]

The explosion of work on presupposition in the 1970s and the multiplicity of theoretical concepts resulted in the uncontrolled growth of terminological proposals and changes in the meaning of terms already used. The term *presupposition* has been ambiguous since the 1970s, and this state of matters persists today. Terms such as *factivity*, *presupposition*, *modality,* or *implicature* are indicated as typical examples of ambiguous expressions. Problems of terminology are **the third problem** to be highlighted in the work on factivity. It is important to note the disturbing phenomenon of transferring terminological issues to the NLI. Reporting studies analogous to ours will bring attention to this difficulty.

A final point to note is the lack of linguistic papers that provide a fairly exhaustive list of factive, non-factive, veridical, etc. expressions.[e] There is also a lack of comparative work between languages in general. This kind of research is only available for individual, selective expressions (see, e.g., Özyıldız 2017; Hanink and Bochnak 2017; Jarrah 2019; Dahlman and van de Weijer 2019).

2.2. Key terms

We, therefore, chose to put more emphasis on conceptual issues. The concepts most important to this study will now be presented, primarily factive presupposition.

2.2.1. Entailment, contradiction, neutral

Let's start by introducing an understanding of the basic semantic relations:

- **Entailment**: **H** must be true if **T** is true;
- **Contradiction**: **H** must be false, if **T** is true;
- **Neutral**: **H** may be false or true if **T** is true (Chierchia and McConnell-Ginet 2000).

where **T (Thesis)** is an utterance and **H (Hypothesis)** is an item of information. In our dataset consisting of "verb-complement" pairs, **H** is a clausal/complement of a sentence **T**.

[d]See on the one hand the works of (Karttunen 1971b; Öztürk 2017; Dietz 2018), on the other hand (Egré 2008; Dahlman 2016; Grigore *et al.* 2016).

[e]We have in mind such lists that could be found in the strictly linguistic literature. In this one, unfortunately, there is still rather a dispute about which particular verbs are factive and which are not. Indeed, there are lists of different types of verbs, but these are based on annotations, often by non-specialist annotators (see, e.g., Nairn, Condoravdi, and Karttunen 2006).

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

2.2.2. Information

The information we are interested in is that transmitted by means of spoken sentences, which are utterances. The mechanisms involved in this transmission may be either of a *purely semantic* (lexical consequences) or *pragmatic nature* (conversational/scalar implicatures) (Grice 1975; Sauerland 2004). Three examples of information are shown below.

<T, H> Example 1. *(Entailment)*
***T**: Only then did they realize that they had been cut off from their only way out.*
***H**: They had been cut off from their only way out.*

<T, H> Example 2. *(Contradiction)*
***T**: He was wrong to say that it was just a normal room.*
***H**: It was just a normal room.*

<T, H> Example 3. *(Neutral)*
***T**: Statists believe that people can be demanding towards the state.*
***H**: People can be demanding towards the state.*

In Example 1, the entailment is lexical in nature because it is founded on the factive verb *realize that*.

The contradiction in Example 2 is also lexical in nature. In Example 3, on the other hand, we have the neutrality: there is nothing in **T**'s utterance that guarantees either a lexical entailment or a contradiction, and there are no pragmatic mechanisms that would fund either of these two relations.

2.2.3. Negation

We take into account in our dataset the occurrence of negation, specifically internal negation. We distinguish it from the so-called external negation, which is not relevant to the phenomenon of factivity. The examples of these two types of negation can be found below.

<T, H> Example 4. *(Entailment)*
***T**: He did not know that he had opened the door.*
***H**: He had opened the door.*

<T, H> Example 5. *(Neutral)*
***T**: It is not the case that he knew he had opened the door.*
***H**: He had opened the door.*

In Example 4, internal negation was used, and in Example 5, external negation. The utterance in Example 4 implies **H**, whereas in Example 5 does not imply **H**. The source of this difference is the different types of negation used. The external negation applied to the utterance 5 makes it possible for all the meaning components of the verb to be within its range, so **H** does not follow from Example 5.

2.2.4. Presupposition

We understand the term *presupposition* as follows:

DEF. If the utterance **T** entails **H** and **H** is insensitive to internal negation in **T**, whether currently occurring or potential, then **H** is the presupposition of the utterance **T**.

We named all such information as presuppositions, regardless of their detailed nature. Thus, **presuppositions can have both semantic and non-semantic grounds**. In the literature, one

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

can find lists of expressions and constructions that are classically called *presupposition triggers* (Levinson 1983). Below is an example illustrating a presupposition based on a factive verb.

<T, H> Example 6.
***T**: [PL] Ona wie/nie wie, że należy podać hasło.*
***T**: [ENG] She knows/does not know that a password must be provided.*
≫ ***H** [PL] Należy podać hasło.*
≫ ***H** [ENG] A password must be provided.*

Information **H** in Example 6 (with and without internal negation) is the presupposition of the utterance because this information is insensitive to internal negation in **T**.[f] Presupposition **H** is guaranteed by the factivity property of the verb *wiedzieć, że/know that*.

The relation between a semantic entailment and a semantic presupposition is shown in Example 7.[g]

<T, H> Example 7. ***T**: The driver managed to open the door before the car sank.*
⊨ /≫ ***H**a The driver managed to open the door before the car sank.*
⊨ ≫ ***H**b The driver tried to open the door before the car sank.*
⊨ /≫ ***H**c The driver opened the door before the car sank.*
⊨ ≫ ***H**d The car sank.*

The utterance in Example 7 semantically entails **H**a-**H**d because the definition of this type of entailment is fulfilled: if **T** is true, then **H** must be true. This is because in the case of the implicative verb *manage that* (Karttunen 1971a), some of its meaningful components are not within the scope of internal negation, for example, the information that someone tried to do something. Apart from that, the information **H**b and **H**d is also presuppositions of the utterance. They meet the defining conditions of a presupposition: they are insensitive to—in this case, potential—internal negation. In other words, we treat presupposition as a certain subtype of entailment.

Analyzing the difference between semantic presupposition and non-semantic presupposition, consider the following utterances in Examples 8 and 9.

<T, H> Example 8.
***T**: She was not told that he was already married.*
≫ ***H**: He is already married.*

<T, H> Example 9.
***T**: She didn't know that he was already married.*
≫ ***H**: He is already married.*

Both utterances in Examples 8 and 9 entail information **H**. Moreover, in both cases, **H** is not within the scope of internal negation, so the information **H** is their presupposition. However, **the foundations of these presuppositions are radically different**. In Example 8, **H** is guaranteed by the appropriate prosodic structure of the utterance, whereas in Example 9, the presupposition **H** has a semantic grounding—it is guaranteed because of the factive verb *know that*.[h,i] In other words, the entailment in Example 8 is not lexical, and in Example 8, it is.

[f]The following question may be asked: how can one justify that the truthfulness of the complement in this case is insensitive to internal negation in **T**? It is beyond the scope of the paper to provide such a justification, but it is worth noting that it can be difficult to provide an adequate reasoning. In our work, we were guided by the findings of the paper (Ziembicki 2022).

[g]"⊨"—lexical entailment; "≫"—presupposition; "/≫"—no presupposition

[h]For an entailment relation to take effect in Example 9, the utterance **T** must be uttered with the appropriate intonation contour.

[i]Note that a segment other than *that* in Examples 8 could result in a lack of entailment, for example *whether*: the sentence *She was not told whether he was already married* does not semantically guarantee the entailment of the sentence complement.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

2.2.5. Factivity

It is worth noting the occurrence of the following four terms in the NLI literature: *factivity*, *event factuality*, *veridicality*, and *speaker commitment*. These terms, unfortunately, are sometimes understood differently depending on the work in question. In the presented study, we use only the term *factivity*, which is understood as an element of the meaning of particular lexical units. Such phenomena as speaker "degrees of certainty" are entirely outside the scope of the research presented here. We assume that presupposition founded on factivity takes place independently of communicative intentions; it may or may not occur: there are no states in between. For comparison, let's look at how the term "event factuality" is understood by the authors of the FactBank dataset:

"(. . .) we define event factuality as the level of information expressing the commitment of relevant sources towards the factual nature of events mentioned in discourse. Events are couched in terms of a veridicality axis that ranges from truly factual to counterfactual, passing through a spectrum of degrees of certainty (Saurí and Pustejovsky 2009: p. 231)."

In another paper, the same pair of authors provide the following definition:

"Event factuality (or factivity) is understood here as the level of information expressing the factual nature of eventualities mentioned in text. That is, expressing whether they correspond to a fact in the world (. . .) (Saurí and Pustejovsky 2012: p. 263)."

It seems that the above two explanations of the term event factuality are significantly different. They are also composed of other terms that require a separate explanation, for example *discourse*, *veridicality*, *spectrum of degrees of certainty*, and *level of information*.

Note that the second quoted fragment also defines factivity; the authors apparently put an equality mark between "event factuality" and "factivity" Reading the specifications of the FactBank corpus and the instructions for the annotators leads, in turn, to the conclusion that Saurí and Pustejovsky understand factivity as (a) a subset of factuality, (b) in a "classical way," as a property of certain lexical entities (Saurí and Pustejovsky 2009).

In the presented approach, factivity is understood precisely as a semantic feature of specific lexical units. In other words, it is an element of their meaning. According to the terminology used in the literature, we will say that factive verbs are *presupposition triggers*. We understand the term *factive verb* as follows:

DEF. A verb **V** is an element of a set of factive units if and only if the negation-insensitive part of the meaning of **V** includes the information **p**.

The information **p** in the utterances with the structure we analyzed "[verb][że][complement clause]" is a complement clause. Using the category of verb signature, it can be said that factive verbs belong to the category (+/+).[j] Examples 10 and 11 illustrate presuppositions based on the meaning of the factivity verb *be aware that*.

<T, H> Example 10.
T*: She was not aware that she had made a fatal mistake.*
⊨ ***H****: She made a fatal mistake.*

<T, H> Example 11.
T*: She was aware she had made a fatal mistake.*
⊨ ***H****: She made a fatal mistake.*

Information **H** follows semantically from both Example 10 and its modification as Example 11. This information is beyond the scope of internal negation and is, therefore, its presupposition.

[j]Factive verbs are usually designated by this signature, for example (Lotan, Stern, and Dagan 2013). We take our understanding of the concept of verb signature as (Lotan *et al.* 2013): *Each signature has a left sign and a right sign. The left sign determines the clause truth value of the predicate's complements, when the predicate is in positive contexts (e.g., not negated), while the right sign applies in negative contexts.*

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

The foundation of this presupposition is the presupposition trigger in the form of the verb *be aware that*. Neither the common knowledge of the speakers nor the prosodic properties of the utterances are irrelevant to the fact that **H** in the above examples are presuppositions.

In summary, presuppositions can be either lexical or pragmatic in nature. What they have in common is that they are insensitive to internal negation in **T**. We treat presuppositions founded on factive verbs as lexical presuppositions. If information **H** is a lexical presupposition of an utterance *T*, then *T* entails **H**. These relations are independent of the speaker's communicative intention; it means that the speaker may presuppose something unconsciously or against his own communicative intention.

3. Our dataset

The historical background of this paper is gathered from (Cooper *et al.* 1996; Dagan, Glickman, and Magnini, 2005, 2013). These works established a certain pattern of construction of linguistic material, consisting in pairs of sentences: *thesis* and *hypothesis* (<**T, H**>).

3.1. Input material sources and extraction

The material basis of our dataset is the National Corpus of Polish Language (NKJP) (Przepiórkowski *et al.* 2012). We used a subset of NKJP in the form of the PCC Polish Coreference Corpus (Ogrodniczuk *et al.* 2014), which contains randomly selected fragments from NKJP and constitutes its representative subcorpus. We did not add any prepared utterances—our dataset consists only of original utterances found in the PCC. Moreover, the selected utterances have not been modified in any way—we did not correct typos, syntactic errors, or other difficulties and issues. Thus, the language content remained entirely natural, not improved artificially.

We automatically annotated with Discann all occurrences of the phrase *że* (*that* | *to*) as in Example 12.[k]

<T, H> Example 12.
T: *[PL] Przez lornetkę obserwuję, że zalane zostały żerowiska bobrów.*
T: *[ENG] I can see through binoculars that the beaver feeding grounds have been flooded.*

From more than 3500 utterances, we left only those that satisfied the following pattern: "[verb] [że] [complement clause]." It required a manual review of the entire dataset. Thus, we obtained **2320** utterances that constitute the empirical basis of our dataset.

3.1.1. <T, H> pairs

In this work, the original source of **T** utterances is NKJP, and **H** is complement clauses of **T**.

<T, H> Example 13.
T: *[PL] Myśleli, że zwierzęta gryza się ze soba. [NKJP]*
T: *[ENG] They thought the animals were biting each other.*
H: *[PL] Zwierzęta gryza się ze soba.*
H: *[ENG] The animals were biting each other.*

From 2320 utterances, we created 2432 <**T, H**> pairs (309 of unique main verbs). We have not paired the sentences randomly. In any particular <**T, H**> pair, **H** is always the complement clause of the **T** utterance from the pair. In some utterances, the verb introduced more than one complement—in each such situation, we created a separate <**T, H**>. For each sentence,

[k]Discann is a web application conceived as a tool for tagging connections between text and metatext. *http://zil.ipipan.waw.pl/Discann*

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

we extracted a complement clause manually. Our manual work included, for example, removing fragments that were not in the range of the verb—see Example 14.

<T, H> Example 14.
T: *He said I am—I guess—beautiful.*
H: *I am—I guess—beautiful.*

Annotators <**T, H**> are the core of our dataset. We assigned specific properties (i.e., linguistic features) to each pair. In the following, we presented these linguistic features with their brief characteristics.

3.2. Linguistic features

3.2.1. Entailment, contradiction, and neutral (ECN)

Each of <**T, H**> pairs was assigned one of three relations: entailment, contradiction, and neutral. The occurrence of features is not balanced, that is *entailment* class states 33.88% of the dataset, *contradiction*—4.40%, *neutral*—61.72%. The lack of balance is a characteristic of all linguistic features in the dataset. Due to the fact that our dataset is a representative subcorpus of the NKJP, we assume that the same distribution of relations is found in the NKJP itself.

3.2.2. Verb

In each utterance, the experienced linguist manually identified the verb that introduced the **H** sentence.

Despite appearances, this was not a trivial task. Often identifying a given lexical unit required deep thought and verification of specific delimitation hypotheses. We assume that an adequate delimitation of language entities is a precondition and also a necessary condition for further linguistic research. The difficulty in determining the form of a linguistic unit is primarily due to the difficulty of establishing its argument positions. However, there are no universal linguistic tests to distinguish between an argument and an adjuncts. In the case of verbs, it is not uncommon for two (or more) different lexical units to be hidden under a graphically equal form. For example, we distinguish between *czuć, że/feel that*, which is epistemic and non-factive, and *czuć, że/feel that*, which is purely perceptual and factive (Danielewiczowa 2002) (see Examples 15, 16, and 17). In addition to these two verbs, we also identified the verb *CZUĆ, że* which is epistemic, factive, and must take on a non-contradictory sentence stress.

<T, H> Example 15. *(czuć*$_1$*, epistemic, non-factive)*
T: *[PL] Czuł, że już nigdy jej nie zobaczy.*
T: *[ENG] He felt that he would never see it again.*

<T, H> Example 16. *(czuć*$_2$*, perceptual, factive)*
T: *[PL] Czuł, że po jego ręce chodzi pajak.*
T: *[ENG] He felt a spider walking on his hand.*

<T, H> Example 17. *(czuć*$_3$*, epistemic, factive)*
T: *[PL] CZUŁ, że patrza na niego, choć stał odwrócony plecami.*
T: *[ENG] He FELT they were looking at him, although he stood with his back turned.*

Another difficulty can be caused, for example, by verbs with the shape *wspominać/mention*. We distinguished between wspominać, że$_1$ (non-factive, speech, non-epistemic) and wspominać, że$_2$ (factive, speech, epistemic):

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

<T, H> Example 18. *(wspominać, że$_1$)*
***T**: [PL] W każdym razie ja kontaktowałem się tylko z nim i nigdy nie wspominał, że ktoś inny bierze udział w transakcji.*
***T**: [ENG] In any case, I only contacted him and he never mentioned that anyone else was participating in the transaction.*

<T, H> Example 19. *(wspominać, że$_2$)*
***T**: [PL] Przyjechał na Ślask z Tomaszowa Mazowieckiego w 1967 roku. Wspomina, że wtedy praca była wszędzie.*
***T**: [ENG] He came to Silesia from Tomaszów Mazowiecki in 1967. He recalls that at that time work was everywhere.*

We identified a total of 309 verbs. A list of these verbs can be found in a publicly accessible repository.[1]

3.2.3. Verb type

We assigned one of two values: *factive*/*non-factive* to all verbs. From the linguistic side, it was the most difficult part. This task was done by a linguist writing his Ph.D. thesis on the factivity phenomenon (Ziembicki 2022). The list was checked with the thesis supervisor, and in most cases, these people agreed with each other, but not in all cases. Finally, 81 verbs were marked as factive and 230 as non-factive.

3.2.4. Internal negation

For each utterance, we marked whether it contains an internal negation. About 95% of utterances did not contain explicit negation words, and almost 5% of sentences did.

3.2.5. Verb semantic class

We have distinguished four semantic classes of verbs: epistemic (*myśleć, że*/*think that*), speech (*dodać, że*/*add that*), emotive (*żałować, że*/*regret that*), and perceptual (*dostrzegać, że*/*perceive that*). Most verbs were hybrid objects, for example epistemic speech. The class name was given due to the given dominant semantic component. If the verb did not fit into any of the above classes, the value *other* was given.

3.2.6. Grammatical tense

In each utterance, we marked the grammatical tense of the main verb and the complement **H**. Note that the tenses in Polish much differ from tenses in English. Polish has only three tenses: past, present, and future. It is worth adding that Polish verbs have an aspect: imperfective and perfective. Perfective verbs reflect completion of an action. In contrast, imperfective verbs focus on the fact that an action is being performed.

3.2.7. Utterance type

We labeled the type of utterance as indicative, imperative, counterfactual, performative, interrogative, or conditional. Below are examples of this type of speech.

(indicative) She knew it was raining.
(imperative) Know that it is raining.
(counterfactual) If I hadn't looked in the window, I wouldn't have known it was raining.

[1]We treated the aspect pairs as two verbs. The Polish language, in a nutshell, has only two aspects: perfect and imperfect.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

(performative) I give you the name "falling rain."
(interrogative) Does it rain?
(conditional) If it rains, the walkways are wet.
All **T** utterances have been translated into English, see Appendix B.

3.2.8. Extended LingFeatured NLI dataset

For additional experiments, we added new sentences with or without internal negation. The procedure for creating new sentences was as follows: (1) in the original sentences with negation it was manually removed, (2) in sentences without negation it was manually added, (3) necessary corrections were made if the procedure of adding/removing negation caused semantic/syntax defects. For many utterances, steps (1)–(2) resulted in such a distortion of the original utterance that it was unsuitable for the dataset. The expanded version of the dataset contains 603 new utterances.

3.3. Annotation

3.3.1. Expert annotation

Among linguistic features assigned to pairs <**T, H**> the most difficult and essential to identify were factivity/non-factivity and logic relations ECN. Whether a verb is factive was determined by two linguists who are professionally involved in natural language semantics. They achieved more than 90% agreement, with most doubts arising when analyzing verbs of speaking, for example *wytłumaczyć, że*/*explain that*. The final decisions on identifying which verb is factive were made by a specialist in the field.

Semantic relations ECN were established in two stages. The dataset itself consists of single-sentence contexts, but the annotation involved an analysis of the utterances in context—annotators checked the relevant utterances in the NKJP. The first stage was annotated by two linguists, including one who academically deals with the issue of verb semantics. They achieved 70% agreement in the annotation in terms of inter-annotator agreement given by Cohen's Kappa. Significant discrepancies can be observed for relations of contradiction, as opposed to entailment. Then, those debatable pairs were discussed with a third linguist, a professor specializing in natural language semantics. The result of these consultations was the final annotation—the gold standard.

3.3.2. Experiment with non-professional annotators

We checked how the gold standard created in this way would differ from the annotations of non-professional linguists—a group of annotators who are not involved professionally in formal linguistics but have a high linguistic competence. The criteria for selecting annotators were the level and type of education and a pre-test. Thus, the four selected annotators were as follows: (A1) cognitive science student (3rd year), (A2) and (A3) annotators with master's degree in Polish Studies, (A4) Ph.D. in linguistics (see Table 1).

Each annotator was given the same set of <**T,H**> pairs from the set (20% of the total set). The task of each annotator was to note the relation between **T** and **H**. There were four labels to choose from *entailment*, *contradiction*, *neutral,* and "*?*."[m] The annotation instructions included simplified definitions of key labels—as we presented in Section 2.2.

Annotators were asked to choose "*?*," if (1) they could not indicate what the relation was, or (2) they thought the sentence was meaningless, or (3) they encountered another problem that made it impossible for them to choose any of the other three labels. Especially important, from our point of view, is the situation (1). The idea was to reserve it for <T, H> pairs whose semantic relation is

[m]However, the utterances annotated as "?" in the GOLD label, were not taken for training and testing ML benchmarks.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Table 1. Inter-annotator agreement given by Cohen's Kappa ($\alpha = 0.05$). Ex—an expert who made the gold standard, A1–A4—non-expert linguists

	Ex	A1	A2	A3	A4
Ex	1.00	0.65	0.60	0.38	0.61
A1	0.65	1.00	0.59	0.29	0.51
A2	0.60	0.59	1.00	0.47	0.70
A3	0.38	0.29	0.47	1.00	0.52
A4	0.61	0.51	0.70	0.52	1.00

dependent on prosodic features—like accent, which determines focus and topic (see Partee 2014). Let's look at an Example 20:

<T, H> Example 20.
T*: Let's not say [that] these projects are supposed to end in a constitutional change.*
H*: These projects are supposed to end in a constitutional change.*

There are two possible situations in Example 20: (a) the sender wants to hide the information from **H** (label: *entailment*) and (b) the sender does not want to say **H** because, for example he wants to make sure that this is true first (label: *neutral*).

Let us specify that the way in which "?" was used was, however, different for the creation of the Gold standard, which was not based on the work of annotators but of a linguist specializing in the topic of factivity: the "?" sign applied only to condition (1) and it was used only six times. Two of these are given below:

<T, H> Example 21.
[PL] No wie pani. . . jemy sobie czekoladę, a pani mówi, że będzie nowym naczelnym. . . ja w śmiech. . .
[ENG] Well, you know. . . we're eating chocolate, and you say you're going to be the new editor. . . I'm laughing. . .

<T, H> Example 22.
[PL] ale też jest fajnie pokazane właśnie to że ona nie jest dojrzała jakby no nie zupełnie
[ENG] but it's nicely shown, too, that in fact she is not really mature, not entirely so to speak

In the first example, neither the utterance itself nor the context (several sentences before and after) make it possible to determine which ECN relation is occurring. In the second example, an analogous situation takes place, and, in addition, it is not clear exactly which main verb introduces the sentence complement: it is not known whether it is factive.

Inter-annotator agreement with the dataset gold standard was calculated using R implementation of Cohen's Kappa. It was in the range of 61%–65%, excluding the worst annotator whose Kappa was below 52% with all other annotators.[n] Table 1 summarizes the inter-annotator agreement among four non-expert linguists and one of the experts preparing the dataset. The conclusions of the annotation performed and described above are provided in Section 6.1.

3.4. Related datasets

We will now review some of the most recent and similar works. Table 2 collates several analogous datasets and highlights the tasks solved in the publications.

[n] A few annotation examples are given in our supplement A.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Table 2. Datasets on NLI and their linguistic features

Dataset	Main characteristics	Task and models in ML
Our dataset *LingFeatured NLI*	Language: Polish. Base version: 2467 original sentences; number of unique main-clause verbs: 309 (30 most common verbs cover 55.41% of the dataset). Extended version: 3035 observations. Additional features: verb, grammatical tense of verb, verb type (factive/non-factive), verb semantic class (epistemic, speech, perceptual, etc.), occurrence of internal negation, grammatical tense of complement clause, utterance type (indicative, performative, interrogative, etc.) gold label—logical relation: ECN	NLI (logical relation: entailment, contradiction, neutral), BERT-based models
CommitmentBank (de Marneffe *et al.* 2019)	Language: English. 1200 examples of naturally occurring discourse segments, 48 different clause-embedding predicates. Data source: extracted from three corpora of different genres: the Wall Street Journal (WSJ, news articles), the fiction component of the British National Corpus (BNC, fiction) and Switchboard (SWBD, dialogue). Annotation method: Annotations from at least eight self-reported native English speakers, using a questionnaire created on Amazon's Mechanical Turk Platform Additional features: lemma, genre, modal type, embedding, morfological features, information whether main-clause verb is factive, gold label—relation: ECN	On average 85% F1 for BERT-based model—results reported in (Jiang and de Marneffe 2019)
RCB dataset from RussianGLUE benchmark (Shavrina *et al.* 2020)	Language: Russian. 2715 examples. Data source: Hand-corrected sentences from the Taiga corpus. Only news and fiction parts of Taiga and only frequently used words. Annotation method: Three human annotators.	38%–45% F1 for BERT-based models and TF-IDF
Dataset published in (Ross and Pavlick 2019a)	Language: English. 1500 sentence pairs, covering 137 unique verbs (212 sentences with factive verbs) (+1500 with artificially added negations). Data source: MNLI corpus. Annotation method: Human judgments on Amazon Mechanical Turk. Raters label on a 5-point likert scale. Additional features: lemma of verb, the type of complement, verb signature (positive and negative environment), gold label—relation: ECN (5-point likert scale)	62% for non-negated sentences, 29% for negated sentencesBERT NLI, for factive verbs
NOPE (Parrish *et al.* 2021)	Language: American English. Main corpus: 2386 examples; adversarial: 346 examples. Data source: extracted sentences with presupposition triggers from the Corpus of Contemporary American English (COCA, Davies 2010). 12.9% of examples: wrote an altered version of the trigger sentence by making small edits in order to make the subsequent annotations possible. Annotation method: collected 5 probability judgments from participants via Amazon Mechanical Turk (MTurk). Participants provided ratings from 0.0 to 100.0 by adjusting a non-linear slider that allowed greater precision at the slider's edges.	The results are divided between negated and non-negated trigger sentence and ECN classes. For E and non-negated sentence—RoBERTa and DeBERTa achieve for negated and E—above 68%, and for other N combined with C between 32% up to 38%.
MEANTIME (Minard et al., 2016a)	Language: English, Spanish, Dutch. 2096 event mentions and linked them to a total of 1717 instances. Data source: The English section of the corpus: English Wikinews articles. Annotation method: The English documents have been annotated by six trained annotators. Spanish, and (partially) Dutch: automatically project the annotations available in the English texts onto the translated texts.	No benchmark
UW (Lee *et al.* 2015)	Language: English. 4243 sentences Data source: gathered data on CrowdFlower. Annotation method: non-experts annotators, Likert 7-point scale (from 3 to −3)—a scale of −3 (certainly did not happen) to 3 (certainly did)	Event factuality prediction task—regression model SVR with custom features

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Table 2. Continued.

Dataset	Main characteristics	Task and models in ML
UDS-IH2 (Rudinger *et al.* 2018)	Language: English. 20,580 total predicates. Data source: extension of the UDS-IH1 dataset (White *et al.* 2016). Annotation method: 32 unique annotators through Amazon's Mechanical Turk	Event factuality prediction task—bidirectional-LSTM-based models
FactBank (Saurí and Pustejovsky 2009)	Language: English. 208 documents and contains a total of 9488 manually annotated events. Data source: built on top of the TimeBank corpus (88%) (Pustejovsky *et al.* 2006) and A-TimeML Corpus (12%). Annotation method: annotated by a pair of annotators, both of whom were undergraduates competent in linguistics, and adjudicated by the authors	Factuality prediction by annotators
IMPPRES (Jeretic *et al.* 2020)	Language: English. 25,500 sentence pairs. Data source: semi-automatically generated Annotation method: semi-automatically, without annotators, made by the authors of the set	Presupposition controls, presupposition triggers a bag of words model, InferSent (Conneau *et al.* 2017), and BERT BERT is the best model in almost every setting and group

3.4.1. Data source

The linguistic material in our dataset was extracted from NKJP. It thus belongs to the class of datasets whose source is a linguistic corpus, such as the Corpus of Contemporary American English or the British National Corpus. A different approach to the data source is presented by for example (Jeretic *et al.* 2020), in which sentences are semi-automatically generated. In contrast to datasets such as (Parrish *et al.* 2021), we did not make any changes to the source material. The only changes we have made relate to the extended version of the dataset (see Section 3.2.8).

3.4.2. Number of observations

LingFeatured NLI has a similar number of observations (sentences/utterances) to (de Marneffe, Simons, and Tonhauser 2019; Ross and Pavlick 2019a; Parrish *et al.* 2021) and (Minard et al., 2016b). Compared to these datasets, the LingFeatured NLI stands out for the number of lexical units (main verbs). For example, the CommitmentBank contains 48 different clause-embedding predicates, whereas our dataset contains 309 main verbs. Because the LingFeatured NLI is a random subcorpus of the NKJP, the main verbs have different numbers of examples of use (see Table 3).

As for factive verbs, we noted 82 entities of this type in our dataset. In comparison, CommitmentBank notes only 11 factive verbs (see also Table 3).

3.4.3. Annotation methodology

An important difference between our dataset and some others is the lack of use of the Likert scale. This scale is used, for example, in (Ross and Pavlick 2019a).[o] The central term of this work is *veridicality*, which is understood as follows: "*A context is veridical when the propositions it contains are taken to be true, even if not explicitly asserted.*" (Ross and Pavlick 2019a, p. 2230). As can be seen, the quoted definition also includes situations in which the entailement is guaranteed by

[o] A Likert scale is a psychometric scale commonly involved in research that employs questionnaires. It is the most widely used approach to scaling responses in survey research, such that the term (or more fully the Likert-type scale) is often used interchangeably with rating scale, although there are other types of rating scales.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Table 3. Factive verbs in CommitmentBank and their Polish equivalents in LingFeatures NLI

Factive verb in CB [ENG]	Polish equivalents in LingFeatures NLI [PL]
Bother	przeszkadzać (0%)
Find	okazać się (2.43%)
Forget	zapomnieć (0.25%)
Know	wiedzieć (5.88%)
Learn	nauczyć się (0%), dowiedzieć się (0.78%)
Notice	zauważyć (.90%)
Realize	zdać sobie sprawę (0.20%), uświadomić sobie (0.37%)
Recognize	rozpoznać (%), zauważyć (0.90%), zdać sobie sprawę (0.20%)
Remember	pamiętać (1.44%)
See	widzieć (2.14%)
Understand	rozumieć (0.58%)

factive verbs. The authors pose the following research question: "whether neural models of natural language learn to make inferences about veridicality consistent with those made by humans?" This is a very different question from the one posed in this paper. Ross and Pavlick used Likert scales to conduct annotations employing unqualified annotators (Ross and Pavlick 2019a). They then checked the extent to which the models' predictions coincide with the human annotations obtained. Unlike these authors, we do not use any scales in annotation, and the object of the models we train is to predict real semantic relations, not those that occur as judged by humans. What we mean is that the task of the model is to predict the relations that objectively occur in reality, not those that occur in the judgment of the annotators. Indeed, at the operational level, these two different tasks merge, because the human being is ultimately behind the identification of the reality relations. For this reason, it does not use a Likert scale and the gold standard is created by specialists in lexical semantics.

3.4.4. Number of presupposition types and verb types

Many datasets deliberately include several linguistic phenomena, for example several types of presupposition, such as (Parrish *et al.* 2021) or distinguishes many different signatures of main verbs, for example (Ross and Pavlick 2019b).[P] A similar approach is in (Rudinger, White, and Van Durme 2018), that is factive verbs are one of several types of expressions that are of interest. Such a decision is dictated by the deliberate action of the corpus authors. Our set is somewhat different in this respect. Our linguistic interests are focused on one phenomenon of presupposition—factivity understood as a lexical phenomenon. In order to take a closer look at it in Polish, we decided to select a random sample of a specific syntactic construction, namely the previously described [V][że/that|to][compound sentence]. The result is linguistic material in which both the lexical and non-lexical presupposition of interest occurs. The latter occurs when the verb is neither factive nor veridical (e.g., mówić/tell), and the truth of the subordinate sentence takes place anyway because it is guaranteed by mechanisms of pragmatic nature. This allowed us to observe, in particular, the

[P]They distinguish between eight pairs of signatures, for example, (+/+) (*realize that*), (+/− (*manage to*), and (−/+) (*forget to*).

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

proportions of occurrence of different types of feature sets, primarily <factive verb, E>, <non-factive verb, E>, <factive verb, N>, and <non-factive verb, N>. In other words, LingFeatured NLI at the annotation level contains different sets of linguistic features with varying frequency of occurrence and this constitutes its annotative core.

We are aware that the number of verb distinctions is sometimes significantly higher in similar papers and our dataset could also include them. The decision to use only a binary distinction (factive vs. non-factive) is dictated by several interrelated considerations. First, there are no lists of Polish verbs with signatures assigned by linguists. Secondly, the preparation of such a list of even several dozen verbs is a highly specialized task. It may be added that there are still disputes in the literature about the status of some high-frequency verbs, for example *regret that*. Third, we are interested in the real features of lexical units, and not in the "textual" ones, that is those developed by non-specialist annotators, using the committee method. The development of signatures by unqualified annotators would be pointless with regard to the research questions posed. The type of linguistics used in this work is formal linguistics, which investigates the real features of a language, unlike survey linguistics, which collects the intuitions of speakers of a language.[q]

4. Implications from our dataset

The dataset created allows a number of observations to be made. We will now discuss the most important of these from the point of view of the issue of factivity/non-factivity opposition and ECN relations.

4.1. Frequency of ECN relations

As the LingFeatured NLI Dataset is a representative subcorpus of the NKJP, this leads directly to the conclusion that in the syntactic construction under study ([V][że/that|to][cc]), the NKJP contains significant differences in the frequency of occurrence of ECN relations (see Table 4). Particularly striking is the low occurrence of contradictions (4.4%). A preliminary conclusion could therefore be that in the syntactic construction under study classes Entailment and Neutral are most common in Polish. In various research, national corpora are used to calculate the frequency of words or collocations in language (see, e.g., Grant 2005). NKJP is the largest representative dataset for Polish which contains texts from various sources to reflect the use of the language. To the best of our knowledge, currently this is the most reliable way to verify the frequency of analyzed semantic relations.

4.2. Factivity/non-factivity and ECN relations

Firstly, the distribution of features in the dataset indicates that, in the vast majority of cases, factive verbs go with an entailment relation (24.4% of our dataset), and non-factive verbs with a neutral relation (61.1%)—see Table 5. Other utterances, that is the pairs <**T, H**>, in which, for example, despite a factive verb, there is neutral, or despite a non-factive verb, there is entailment, constitute a narrow subset of the dataset (14.5% – 352 utterances in the dataset). Table 6 contains examples of such pairs. In the experiments carried out these kinds of <**T, H**> pairs will pose the biggest problem for humans and models—the best model accuracy of 62.87% (see Table 9 and Sections 3.3 and 6).

[q]See (Hsueh, Melville, and Sindhwani 2009; Ipeirotis, Provost, and Wang 2010) on the problems of low-quality annotation with Amazon Turk and how to solve them.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Table 4. Distributions of features in *LingFeatured NLI* and *LingFeatured NLI* Extended dataset in Polish version

Features	Distribution base dataset	Distribition extended dataset
Target/GOLD – logic relations	*Entailment* 33.88%; *contradiction* 4.40%; *neutral* 61.72%	*Entailment* 41.12%; *contradiction* 5.34%; *neutral* 53.54%
Verb type (factivity)	Factive 24.96%; non-factive 75.04%	Factive 32.92%; non-factive 67.08%
Grammatical verb tense	Past 36.18%; present 52.22%; future 3.08%; none 8.51%	Past 37.79%; present 50.31%; future 3.13%; none 8.76%
Utterance type	Indicative 90.17%; performative 2.43%; rule 2.22%; interrogative 1.97%; imperative 1.93%; counterfactual 0.66%; conditional 0.62%	Indicative 90.84%; performative 2.60%; rule 1.85%; interrogative 1.98%; imperative 1.68%; counterfactual 0.53%; conditional 0.53%
Verb semantic class	Epistemic 51.85%; speech 38.03%; perceptual 1.81%; emotive 1.40%; other 6.91%;	Epistemic 51.12%; speech 34.37%; perceptual 2.54%; emotive 2.17%; other 9.75%;
Occurrence of internal negation	Occurs in 4.93%; does not occur in 95.07%	Occurs in 5.60%; does not occur in 94.40%
Tense of complement clause	Past 23.23%; present 49.01%; future 14.80%; other 12.95%	Past 24.58%; present 50.05%; future 13.31%; other 12.06%
#observations	2432	3035

Table 5. Contingency table consisting of the frequency distribution of two variables

	Base LingFeatured NLI		Extended LingFeatured NLI	
	Factive	Non-factive	Factive	Non-factive
C—contradiction	0	107	0	162
E—entailment	593	231	984	262
N—neutral	14	1487	15	1612
Total	607	1825	999	2036

Let's recap—in 85.5% of the pairs of the whole dataset, entailment co-occurs with a factive verb or the neutrality co-occurs with a non-factive verb.

Second, if the verb was factive, then the entailment relation occurred in 97.70%. And if the verb was non-factive, the neutrality occurred in 81.50%. It means that pairs of features <factivity, entailment>, and <non-factivity, neutrality> very often co-occur with each other, especially the first pair. This means that such phenomena as cancelation and suspension of presuppositions are marginal in our dataset.[r] Below are examples of such utterances:

<T, H> Example 23.
[PL] Skad wiesz, że kupił Gosia?
[ENG] How do you know he bought Gosia?

<T, H> Example 24.
[PL] Mniejsza o incognito, może nie zorientowaliby się, że to ja.
[ENG] Never mind the incognito, maybe they wouldn't know it was me.

[r]See, for example, the paper (Abrusán 2016), which discusses the phenomena behind these terms.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Table 6. Hard utterances in our dataset

T: [ENG] *How do you know he bought, Gosia?* [PL] *Skad wiesz, że kupił, Gosia?*
H: [ENG] *He bought.* [PL] *Kupił.*
GOLD—Neutral, verb—Factive
T: [ENG] *I read that Gabrysia was crying when she discovered that her daughter Tygrysek had lied, and this is such a moral compass for me.* [PL] *Czytam, że Gabrysia płakała,kiedy odkryła, że jej córka Tygrysek skłamała, i jest to dla mnie taki moralny azymut.*
H: [ENG] *Her daughter Tygrysek had lied* [PL] *Jej córka Tygrysek skłamała*
GOLD—Neutral, verb—Factive
T: [ENG] *Ernest and Agnieszka didn't plan that they would have a big female family.* [PL] *Ernest i Agnieszka nie planowali, że będa mieli wielka, babska rodzinę.*
H: [ENG] *Ernest and Agnieszka have a big female family.* [PL] *Ernest i Agnieszka maja wielka, babska rodzinę.*
GOLD—Entailment, verb—Non-factive

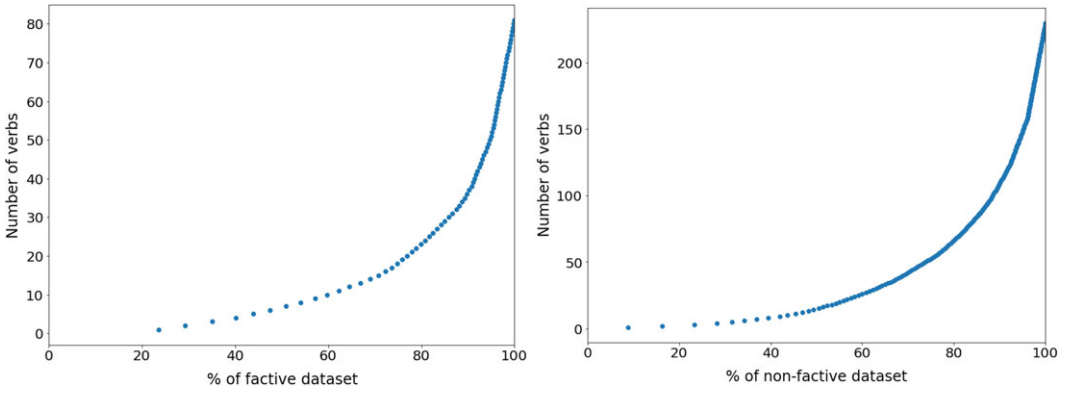

Figure 1. Relationship between the number of the most frequent verbs and the coverage of dataset. Left: The analysis of factive subsample. Right: The analysis of non-factive subsample.

<T, H> Example 25.
[PL] W polu Kryteria: kolumny Imie wpisz w cudzysłowie imię osoby, o której wiesz, że i tak nigdy nie odda pożyczonych filmów (rysunek 4.20).
[ENG] In the field Criteria: column Name, enter in inverted commas the name of a person about whom you know that they will never return borrowed films (Figure 4.20).

4.3. Frequency of verbs

In the dataset, significant differences can be observed regarding the frequency of occurrence of particular main verbs and the coverage by these verbs of the entire dataset (see Figure 1 and Table 4). According to our dataset, only 10 factive verbs with the highest frequency account for the occurrence of 60% of all occurrences of such expressions, and 10 non-factive verbs with the highest frequency account for nearly 45% of all occurrences of non-factive verbs (see Figure 1).

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

5. Machine learning modelling: experiments and results

We used the dataset described above (Base and Extended versions) for experiments with ML models. The models we built aim to simulate human cognitive abilities. The models trained on our dataset were expected to reflect high competence—comparable to that of human experts—in recognizing the relations of entailment, contradiction, and neutral (ECN) between an utterance **T** and its complement **H**. We trained seven models:

(1) Random Forest with an input of the prepared linguistic features (LF)—Random Forest (LF),[s]
(2) fine-tuned HerBERT-based models for only main verbs in sentences as inputs—HerBERT (verb),
(3) model (2) with input extended with linguistic features listed in Table 12—HerBERT (verb+LF),
(4) fine-tuned HerBERT-based model for the whole input utterance **T**—HerBERT (sentence),
(5) model (3) with input extended with linguistic features listed in Table 12—HerBERT (sentence+LF),
(6) fine-tuned PolBERT-based model for only main verbs in sentences as inputs—PolBERT (verb),
(7) fine-tuned PolBERT-based model for the whole input utterance **T**—PolBERT (sentence).

We employed HerBERT (Mroczkowski *et al.* 2021) and PolBERT (Kłeczek 2020) models instead of BERT (Devlin *et al.* 2019), because they are trained explicitly for Polish. Regarding verb models, the input to the model is text consisting only of a main verb. Each model was trained using 10-fold cross-validation with stratification in order to avoid selection bias and estimate standard deviation. Due to the fact that class C is sparse, selecting cross-validation allows us to obtain more reliable results. Each fold has the same proportion of observations with a given labels due to stratification.

In addition, we have prepared a rule-based baseline model from the relations shown in Table 5 that works as follows.

- if verb is factive, then assign entailment (E);
- if verb is non-factive, than assign neutral (N).

Table 7 shows the models' results achieved on the first-seen data (data unknown for a model). The values in the table represent mean and standard deviation of metrics, respectively. F1 score in binary setting is harmonic mean between model precision and recall. In multiclass situation, it is calculated per class and overall metric is average score. Here F1 score was calculated as weighted F1, due to large imbalance between classes. Weighted F1 score is calculated by taking the mean of all per-class F1 scores while considering each class's support.[t]

The parameters of models and their training process are gathered in Table 8. The precise results of Random Forest for different feature sets and the feature importance plots are given in Table 12 and in Figure 2. Table 9 summarizes the results of our models for the most characteristic sub-classes in our dataset: entailment and factive verbs, neutral and non-factive verbs, and the other cases.

Table 10 presents models' results on extended dataset. Training procedure is the same as previously. We also show the analysis of hard cases in Table 11.

The overall results indeed show very high performance of models in Entailment and Neutral classes: Entailment—87% to 92% in F1 score, and Neutral—91% to 93.5% in F1 score (see Table 7).

[s]Random forest implementation from sklearn python package (Breiman 2001).
[t]*https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html*

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Table 7. Classification results of entailment (E), contradiction (C), and neutral (N). Linguistic features comprise: verb, grammatical tense of verb, occurrence of internal negation, grammatical tense of complement clause, utterance type, verb semantic class, verb type (factive/non-factive). F1 score depicts weighted F1 score. LF denotes linguistic features. The results for baseline are calculated over the entire input dataset

		F1 score (%)				
Model	Input	All	C	E	N	Accuracy (%)
Baseline	Factivity/non-factivity	83.27	0.00	82.88	89.42	85.53
Random Forest	LF	**90.56 ± 2.11**	39.89 ± 23.30	91.96 ± 2.50	93.35 ± 1.32	91.32 ± 1.84
HerBERT	Verb	89.45 ± 1.00	39.21 ± 17.57	90.50 ± 1.93	92.42 ± 0.75	90.21 ± 1.06
HerBERT	Verb + LF	90.52 ± 2.37	33.00 ± 31.97	**92.45 ± 1.88**	**93.53 ± 1.36**	**91.53 ± 1.89**
HerBERT	Sentence	88.51 ± 1.06	48.33 ± 11.93	88.34 ± 1.61	91.46 ± 0.08	88.57 ± 1.11
HerBERT	Sentence + LF	89.92 ± 1.23	26.25 ± 17.00	92.27 ± 1.84	93.15 ± 0.71	90.95 ± 1.15
PolBERT	Verb	88.38±0.94	35.26±20.10	89.02±2.08	91.80±0.70	89.27±1.05
PolBERT	Sentence	87.75±1.36	**51.05±11.49**	87.09±2.65	90.71±0.78	87.75±1.12

Table 8. Model and training parameters

	Model	Parameters
1	Random Forest	sklearn implementation with 100 trees (n_features = 100, max_depth = 20, random_state = 123, class_weight={'C': 2, 'E': 1, 'N': 1} and default other parameters)
2	HerBERT (verb)	32 – batch size, 10 – epochs, 1e-5 – learning rate (Pytorch implementation of Adam)
3	HerBERT (verb+LF)	Predictions of model (2) combined with linguistic features preprocessed with one hot encoding, sklearn implementation of Multi-layer Perceptron
4	HerBERT (sentence)	32 – batch size, 13 – epochs, 1e-5 – learning rate (Pytorch implementation of Adam)
5	HerBERT (sentence+LF)	Predictions of model (4) combined with linguistic features preprocessed with one hot encoding, sklearn implementation of Multi-layer Perceptron
6	PolBERT (verb)	32 – batch size, 4 – epochs, 1e-4 – learning rate (Pytorch implementation of Adam)
7	PolBERT (sentence)	32 – batch size, 4 -– epochs, 1e-4 – learning rate (Pytorch implementation of Adam)

Similar relation can be observed on extended dataset: F1 score of Entailments varies from 87% to 92% and from 89% to 92% for Neutral class. More precisely, the models achieved very high results (93% up to 100% in accuracy) for the sentence pairs containing lexical relations—subsets entailment and factive, neutral, and non-factive in Table 9. However, it gained very low metrics (47% up to 62.9%) on those with a pragmatic foundation, which are drastically more difficult—subset "Other" in Table 9. The results for the extended version of the dataset are similar and can be seen in detail in Table 11. Moreover, in both cases, the models' performance is much better than the rule-based baseline.

Additionally, the overall ML modeling results show that HerBERT sentence embedding-based models are at a much higher level than non-expert linguists. Nevertheless, they did not achieve the results of professional linguists (by whom our gold standard was annotated). Feature-based models achieve slightly better results (mean accuracy across folds of 91.32%), although not for the contradiction relation (mean accuracy of 39.89%). However, the weak result for this relation is due to a small representation in the dataset (only 4.4% cases, see Tables 4 and 5). Moreover, the

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

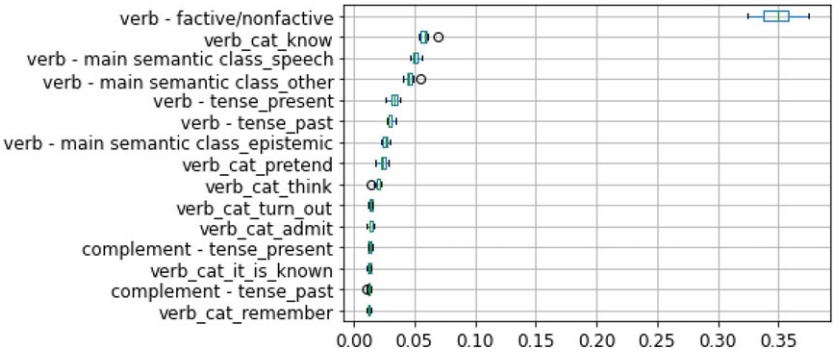

Figure 2. Impurity-based feature importance of feature-based Random Forest trained using base LingFeatured NLI. The chart shows the English equivalents of Polish verbs: *know*/*wiedzieć*; *pretend*/*udawać*; *think*/*myśleć*; *turn out*/*okazać się*; *admit*/*przyznać*; *it is known*/*wiadomo*; *remember*/*pamiętać*.

variance of the ML results is between 0.08% up to 2.5% across different folds in cross-validation for the overall results and the easier classes (E,N). However, the variance for C class (contradiction) is very high—from 11.93 even up to 31.97%. Once again, this occurs because the phenomenon appears very rarely, in the dataset we have only a few cases (i.e., 107 cases).

We prepared the gold standard dataset based on annotations made by experts in formal semantics (see Section 3). Additionally, we compared the model results with non-specialist annotator as an experiment. Note that the performance of models trained on such a training set achieves significantly higher results than those obtained by non-experts annotators in the test annotation performed.

Further, models with verb embedding vs the entire sentence representations are better. However, they require manual extraction of the main verb in the utterance because sometimes it is not apparent. The examples of such a difficult extraction are as follows.

<T, H> Example 26. *(Neutral)*
T: *[PL] Czuł, że inni zbliżali się do niego, ale nie był tego pewien.*
T: *[ENG] He felt others getting closer to him, but he wasn't sure.*

<T, H> Example 27. *(Entailment)*
T: *[PL] Czuł, że maszynistka nie spuszcza zeń oczu.*
T: *[ENG] He felt that the typist wasn't taking her eyes off him.. . .*

Consideration of the difference between the main verbs in the above examples requires attention to suprasegmental differences. In Example 26, the non-factive verb *czuć, że* is used. In contrast, in Example 27, a different verb is used, namely the factive *CZUĆ, że*, which necessarily takes on the main sentence stress (see, e.g., Danielewiczowa 2002). Note that from the two verbs above, we should also distinguish the factive perceptual verb *czuć, że*.

<T, H> Example 28. *(Neutral)*
T: *[PL] Lawirował na granicy prawdy, lecz przez cały czas czułem, że kłamie.*
T: *[ENG] He was teetering on the edge of the truth, but at all times I felt he is lying.*
H: *[PL] Kłamie.*
H: *[ENG] He was lying.*

<T, H> Example 29. *(Entailment)*
T: *[PL] – Dziękuję – odpowiedział, czujac, że policzek zaczyna go boleć.*
T: *[ENG] – "Thank you," he replied, feeling his cheek starts to hurt.*
H: *[PL] Policzek zaczyna go boleć.*
H: *[ENG] His cheek starts to hurt.*

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Table 9. Results in the most characteristic subsets in our test base dataset: entailment and factive, neutral, and non-factive, and the other cases. Values present model accuracy [%]. LF denotes linguistic features

Model	Random Forest	HerBERT (sentence)	HerBERT (sentence + LF)	HerBERT (verb)	HerBERT (verb + LF)	PolBERT (verb)	PolBERT (sentence)
Entailment and factive verbs	99.81 ± 0.56	93.00 ± 2.76	**100.00 ± 0.0**	97.58 ± 2.54	**100.00 ± 0.0**	94.86 ± 3.95	92.62 ± 3.31
Neutral and non-factive	**98.19 ± 1.12**	93.01 ± 1.83	96.77 ± 1.23	95.75 ± 1.95	97.04 ± 1.28	95.81 ± 2.40	91.12 ± 2.86
Others	47.76 ± 10.18	62.87 ± 6.04	50.81 ± 7.32	53.81 ± 9.71	52.15 ± 9.71	50.44 ± 8.81	**63.65 ± 8.86**

Table 10. Classification results of entailment (E), contradiction (C), and neutral (N) using the **extended *LingFeatured NLI*** dataset. Linguistic features comprise: verb, grammatical tense of verb, occurrence of internal negation, grammatical tense of complement clause, utterance type, verb semantic class, verb type (factive/non-factive). F1 score depicts weighted F1 score. LF denotes linguistic features. The results for baseline are calculated over the entire input dataset

		F1 score (%)				
Model	Input	All	C	E	N	Accuracy (%)
Baseline	Factivity/non-factivity	83.17	0	87.66	88.02	85.54
Random Forest	LF	87.29 ± 1.38	53.94 ± 5.38	87.99 ± 2.11	90.08 ± 1.01	87.58 ± 1.35
HerBERT	Verb	85.66 ± 2.02	43.18 ± 9.95	87.22 ± 2.92	88.69 ± 1.63	86.23 ± 1.89
HerBERT	Verb + LF	86.24 ± 1.41	42.09 ± 12.90	87.95 ± 1.80	89.35 ± 1.33	87.05 ± 1.35
HerBERT	Sentence	**90.59 ± 1.35**	**67.44 ± 5.93**	**91.88 ± 1.33**	**91.90 ± 1.31**	**90.61 ± 1.31**
HerBERT	Sentence + LF	89.55 ± 1.83	44.55 ± 13.75	89.40 ± 2.70	90.21 ± 1.55	88.07 ± 1.97
PolBERT	Verb	85.48 ± 1.97	40.85 ± 12.18	86.83 ± 2.57	88.90 ± 1.64	86.22 ± 1.93
PolBERT	Sentence	89.21 ± 2.61	64.56 ± 10.84	90.52 ± 2.09	90.67 ± 2.71	89.16 ± 2.84

Table 11. Results in the most characteristic subsets in our test **extended *LingFeatured NLI*** dataset: entailment and factive, neutral, and non-factive, and the other cases. Values present model accuracy [%]. LF denotes linguistic features

Model	Random Forest	HerBERT (sentence)	HerBERT (sentence + L F)	HerBERT (verb)	HerBERT (verb + L F)	PolBERT (verb)	PolBERT (sentence)
Entailment and factive verbs	99.99 ± 0.30	98.23 ± 1.96	**100.0 ± 0**	96.40 ± 1.96	**100.0 ± 0**	96.53 ± 2.21	95.47 ± 2.55
Neutral and non-factive	**96.90 ± 1.25**	92.04 ± 3.40	93.72 ± 3.12	92.31 ± 1.62	94.04 ± 1.36	94.91 ± 4.41	91.37 ± 4.75
Others	25.84 ± 3.47	36.15 ± 13.23	34.06 ± 8.51	**71.42 ± 7.65**	37.96 ± 8.23	32.24 ± 13.42	66.24 ± 8.58

In Example 28, the epistemic verb is used, while in Example 29, the main predicate is the perceptual verb. The former is non-factive and the latter is factive.

Further findings are that the models with inputs comprising text embeddings and linguistic features achieved slightly better results than those with only text inputs. The only exception to this rule is HerBERT with sentence input trained and tested on extended dataset, where the addition of linguistic slightly worsened the results (from 90.59% to 89.55% of F1 score). Besides, we can

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Table 12. Features in classification of entailment, contradiction and neutral. Random forest results with inputs of different sets of features on base LingFeatured NLI

Features	Accuracy (%)	Weighted F1 (%)
Verb – factive/non-factive	85.53 ± 1.78	83.24 ± 1.92
Verb	83.23 ± 2.29	81.29 ± 2.75
Verb, tense of verb, occurrence of negation, tense of complement clause, type of sentence	83.10 ± 1.93	81.67 ± 2.38
Verb, tense of verb, occurrence of negation, tense of complement clause, type of sentence, **semantic class of verb**	86.76 ± 1.44	85.83 ± 1.89
Verb, tense of verb, occurrence of negation, tense of complement clause, type of sentence, **semantic class of verb**, **verb—factive/non-factive**	**91.32 ± 1.84**	**90.56 ± 2.11**

see that—in our base feature Random Forest model (1)—some features make the most significant contribution to our model, that is if the verb is factive or non-factive (see Table 12 and Figure 2). However, the indication of verb tense (see Figure 2) as relatively important for our ML tasks, that is ECN classification, appears to be misleading in the light of linguistic data and requires further analysis. It seems that we are dealing here with spurious correlations rather than with lexical rules of language connecting the verb tense with ECN relations. Deeper linguistic analysis would be advisable here, however, because the relation between the grammatical tense of the verb and the ECN relations may be the result of pragmatic rules and prosodic properties of the utterances. We hypothesize that these are spurious correlations in our dataset because, indeed, present or past tense co-occur more often with a particular class of ECN in our dataset.[u]

6. Analysis of results and discussion

6.1. Expert and non-expert annotation

Inter-annotator agreement of non-expert annotators with the linguists' preparing the dataset gold standard (Kappa of 61%–65%) indicates that the task is very specialized. We did not find patterns of errors made by the annotators. If the goal of human annotation is to identify the real relationships between two fragments, then such annotation requires specialized knowledge and a sufficiently long time to perform such a task.

Note that Jeretic et al. (2020), as part of their verification of annotation in the MultiNLI corpus (Williams, Nangia, and Bowman 2018), randomly selected 200 utterances from this corpus and presented them for evaluation to three expert annotators with several years of experience in formal semantics and pragmatics. The agreement among these experts was low. This provides the authors with an indication that MultiNLI contains a few "paradigmatic" examples of implicatures and presuppositions.[v]

Notice that the low agreement of annotators may also be the result of differences in their beliefs of theoretical nature and their research specialization. In our opinion, the analysis of the issue of human annotation process in such a task as detecting relations of entailment, contradiction and neutral in principle deserves a separate study.

[u]Other names for this issue: *annotation artifacts* (Gururangan *et al.* 2018), *dataset bias* (He, Zha, and Wang 2019), *group shift* (Oren *et al.* 2019). For the problem of spurious correlations in the context of NLI (see, e.g., Dasgupta *et al.* 2018; McCoy, Pavlick, and Linzen 2019; Tu *et al.* 2020).

[v]See Levinson (2001) for typical examples of presuppositions and implicatures.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Table 13. Top 10 verbs broken down into factive/non-factive subgroups. List of all factive and non-factive verbs is available in Appendix D

Factive	Non-factive
wiedzieć/know	mówić/say
pamiętać/remember	myśleć/think
wiadomo [komuś]/it is known	powiedzieć/tell
przyznać/admit, acknowledge	uważać/believe
[epistemic] widzieć/see	okazać się/turn out
cieszyć się/to be glad	mieć nadzieję/hope
przypomnieć [komuś]/remind [someone]	twierdzić/assert
dowiedzieć się/find out, learn	wydaje się [komuś]/it appears [to someone]
zrozumieć/understand	stwierdzić/state
przyznawać/admit, acknowledge	wynikać/imply, follow

6.2. Factivity/non-factivity opposition

It can be concluded that **the opposition factivity/non-factivity for the prediction of ECN relations is relevant in a fundamental way**. A factive verb co-occurs with relation E in 24% in our base dataset (32% in the extended dataset), while a non-factive verb co-occurs with relation N in 61% in our base dataset (53% in the extended dataset). This means that the linguistic material we collected in principle does not require labeling main verbs with other signatures (e.g., +/−; +/o) to achieve high scores for E and N relations (see Table 5 in our paper for precise counts.)

6.3. Verb frequency

It is also worth noting that in the task of predicting ECN relations in the "that|to" structure, we do not need large lists of verbs with their signatures to identify these relations reasonably efficiently. This is due to the fact that a relatively small number of verbs covers a large part of the entire dataset (see Figure 1). Data from other corpora (see, e.g., the attendance statistics for the British National Corpus) suggest that, given a list of 10 English factive verbs and 10 English non-factive verbs with the highest attendance, one might expect high model predictions if the train and test set were constructed from a random, representative sample of BNCs, analogous to our dataset, which is the NKJP sample.[w]

Given the problem of translation of utterances from one language to another, it is therefore sensible to create a multilingual list of verbs with the highest frequency of occurrence. We realize that the frequency of occurrence of certain Polish lexical units may sometimes differ significantly from that of their equivalents in other languages. However, there are reasons to believe that these differences are not significant.[x] A bigger issue than frequency may be a situation where a factive verb in language X is non-factive in language Y and vice versa. Table 13 contains lists the factive and non-factive verbs with the highest frequency in our dataset. We leave it to native speakers of English to judge whether the given English verbs are factive/non-factive.

[w] *http://www.kilgarriff.co.uk/bnc-readme.html*

[x] Compare, for example, the verbs *wiedzieć* and *powiedzieć* with their English counterparts (Davies and Gardner 2010) and many other such verbs realizing the "[V][że/that|to]" structure.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

At this point, it is worth asking the following question: do the results obtained on the Polish material have any bearing on communication in other languages? We think it is quite possible. Firstly, the way of life and, consequently, the communication situations of the speakers of Polish do not differ from the communication situations of the speakers of English, Spanish, German, or French. Secondly, we see no evidence in favor of a negative answer. It is clear, however, that the answer to this question requires research analogous to ours in other languages.

6.4. ML model results

We presented benchmarks with BERT-based models and models utilizing prepared linguistic features. They are even better than the performance of non-specialist annotators. Hence, it follows that annotation of ECN relations by models on a random sample of NKJP would be better than of non-specialist annotators.

However, a few issues remain unresolved in this task, that is utterances with a pragmatic foundation. Other issues to examine are potential spurious correlations (e.g., influence of the verb tense on the model results)—further, deeper analysis of the models and their interpreting. Our results indicate the need for a dataset that focuses on these kinds of cases.

7. Summary

We gathered dataset *LingFeatured NLI* that is representative with regard to particular syntactic pattern "[verb][że (eng: that/to)][complement clause]" and factivity and non-factivity characteristic of the verb (in the main clause). The dataset is derived from NKJP—Polish National Corpus, which itself is representative for Polish contemporary utterances. In this study, we stated a review of the opposition factivity—non-factivity in the context of predicting ECN relations.

The four most important features of our dataset are as follows:

- it does not contain any prepared utterances, only authentic examples from the national language corpus (NKJP). Only the second version of the dataset (extended *LingFeatured NLI*) contains prepared utterances,
- it is not balanced, that is some features are represented more frequently than others; that is *entailment* class states 33.88% of the dataset, *contradiction*—4.40%, *neutral*—61.72%. It is a representative subcorpus of the NKJP,
- each pair <T, H> is assigned a number of linguistic features, for example the main verb, its grammatical tense, the presence of internal negation, the type of utterance, etc. In this way, it allows us to compare different models—embedding-based or feature-based.
- it contains a relatively large number of factive and non-factive verbs.

We then analyzed the data content of our dataset. Finally, we trained the ML models, checking how they dealt with such a structured dataset. We found that transformer BERT-based models working on sentences obtained relatively good results ($\approx$ 89% F1 score on base dataset). Even though better results were achieved using linguistic features ($\approx$ 91% F1 score on base dataset), this model requires more human labor because the features were prepared manually by expert linguists. BERT-based models consuming only the input sentences show that they comprehend most of the complexity of NLI/factivity. Complex cases in the phenomenon—for example cases with entitlement (E) and non-factive verbs—remain an issue for further research because the models performed unsatisfactorily (below 71% F1 score on the base and extended dataset).

Acknowledgements. We thank Przemysław Biecek and Szymon Maksymiuk for their work on another NLI dataset and valuable remarks on conducting experiments and interpretability approaches. We also want to thank Karol Saputa, who implemented preliminary source code for the machine learning models we reused and redesigned in our experiments.

Also, we are grateful for many Students from the Faculty of Mathematics and Information Science at Warsaw University of Technology, working under Anna Wróblewska's guidance in Natural Language Processing course. They performed experiments on similar datasets and thus influenced our further research.

The research was funded by the Faculty of Mathematics and Information Science and the Centre for Priority Research Area Artificial Intelligence and Robotics of Warsaw University of Technology within the Excellence Initiative: Research University Program (grant no 1820/27/Z01/POB2/2021).

Competing interests. The authors declare none.

References

Abrusán M. (2016). Presupposition cancellation: explaining the 'soft–hard' trigger distinction. *Natural Language Semantics* **24**(2), 165–202.

Anand P. and Hacquard V. (2014). Factivity, belief and discourse. *The Art and Craft of Semantics: A Festschrift for Irene Heim* **1**, 69–90.

Beaver D. (2010). 3: Have you noticed that your belly button lint color is related to the color of your clothing? In *Presuppositions and Discourse: Essays Offered to Hans Kamp*. Leiden: Brill, pp. 65–100.

Breiman L. (2001). Random forests. *Machine Learning* **45**(1), 5–32.

Chierchia G. and McConnell-Ginet S. (2000). *Meaning and Grammar: An Introduction to Semantics*. Cambridge: MIT press.

Conneau A., Kiela D., Schwenk H., Barrault L. and Bordes A. (2017). *Supervised learning of universal sentence representations from natural language inference data*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen: Association for Computational Linguistics, pp. 670–680.

Cooper R., Crouch D., van Eijck J., Fox C., van Genabith J., Jaspars J., Kamp H., Milward D., Pinkal M., Poesio M. and Pulman S. (1996). *FraCaS: Using the Framework*. University of Edinburgh.

Dagan I., Glickman O. and Magnini B. (2005). The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*. Berlin/Heidelberg: Springer, pp. 177–190.

Dagan I., Roth D., Sammons M. and Zanzotto F.M. (2013). Recognizing textual entailment: models and applications. *Synthesis Lectures on Human Language Technologies* **6**(4), 1–220.

Dahlman R.C. (2016). Did people in the middle ages know that the earth was flat? *Acta Analytica* **31**(2), 139–152.

Dahlman R.C. and van de Weijer J. (2019). Testing factivity in Italian. Experimental evidence for the hypothesis that Italian sapere is ambiguous. *Language Sciences* **72**, 93–103.

Danielewiczowa M. (2002). Wiedza i niewiedza. In *Studium polskich czasowników epistemicznych*.

Dasgupta I., Guo D., Stuhlmüller A., Gershman S.J. and Goodman N.D. (2018). Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv: 1802.04302*.

Davies M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* **25**(4), 447–464. https://doi.org/10.1093/llc/fqq018

Davies M. and Gardner D. (2010). *Word Frequency List of American English*, vol. 10343885. London: Taylor & Francis.

de Marneffe M.-C., Simons M. and Tonhauser J. (2019). The commitmentbank: investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung* **23**(2), 107–124.

Delacruz E.B. (1976). Factives and proposition level constructions in Montague grammar. In *Montague Grammar*. Amsterdam: Elsevier, pp. 177–199.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2019). *BERT: pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis: Association for Computational Linguistics, pp. 4171–4186.

Dietz C.H. (2018). Reasons and factive emotions. *Philosophical Studies* **175**(7), 1681–1691.

Djärv K. (2019). *Factive and Assertive Attitude Reports*. Philadelphia: University of Pennsylvania.

Djärv K. and Bacovcin H.A. (2020). Prosodic effects on factive presupposition projection. *Journal of Pragmatics* **169**, 61–85.

Egré P. (2008). Question-embedding and factivity. *Grazer Philosophische Studien* **77**(1), 85–125.

Elliott D.E. (1974). Toward a grammar of exclamations. *Foundations of Language* **11**(2), 231–246.

Giannakidou A. (2006). Only, emotive factive verbs, and the dual nature of polarity dependency. *Language* **82**(3), 575–603.

Givón T. (1973). The time-axis phenomenon. *Language* **49**(4), 890–925.

Grant L.E. (2005). Frequency of 'core idioms' in the british national corpus (bnc). *International Journal of Corpus Linguistics* **10**(4), 429–451.

Grice H.P. (1975). Logic and conversation. In *Speech Acts*. Leiden: Brill, pp. 41–58.

Grigore N. (2016). Factive verbs and presuppositions for 'regret' and 'know'. *Revista Română de Filosofie Analitică* **10**(2), 19–34.

Gururangan S., Swayamdipta S., Levy O., Schwartz R., Bowman S.R. and Smith N.A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv: 1803.02324*.

Hanink E. and Bochnak M.R. (2017). *Factivity and two types of embedded clauses in Washo*. In *Proceedings of the 47th Annual Meeting of North-East Linguistic Society (NELS 47)*, pp. 65–78.
Hazlett A. (2010). The myth of factive verbs. *Philosophy and Phenomenological Research* **80**(3), 497–522.
Hazlett A. (2012). Factive presupposition and the truth condition on knowledge. *Acta Analytica* **27**(4), 461–478.
He H., Zha S. and Wang H. (2019). Unlearn dataset bias in natural language inference by fitting the residual. *arXiv preprint arXiv: 1908.10763*.
Hooper J.B. (1975). On assertive predicates. In *Syntax and Semantics*, vol. 4. Leiden: Brill, pp. 91–124.
Hsueh P.-Y., Melville P. and Sindhwani V. (2009). *Data quality from crowdsourcing: a study of annotation selection criteria*. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pp. 27–35.
Huang Y. (2011). 14. Types of inference: entailment, presupposition, and implicature. In *Foundations of Pragmatics*. Berlin: De Gruyter Mouton, pp. 397–422.
Ipeirotis P.G., Provost F. and Wang J. (2010). *Quality management on Amazon mechanical turk*. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pp. 64–67.
Jarrah M. (2019). Factivity and subject extraction in Jordanian Arabic. *Lingua* **219**, 106–126.
Jeretic P., Warstadt A., Bhooshan S. and Williams A. (2020). *Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 8690–8705.
Jiang N. and de Marneffe M.-C. (2019). *Evaluating BERT for natural language inference: a case study on the CommitmentBank*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, pp. 6085–6090.
Karttunen L. (1971a). Implicative verbs. *Language* **47**(2), 340–358.
Karttunen L. (1971b). Some observations on factivity. *Research on Language & Social Interaction* **4**(1), 55–69.
Karttunen L. (2016). Presupposition: What went wrong? In *Semantics and Linguistic Theory*, vol. 26, pp. 705–731.
Kastner I. (2015). Factivity mirrors interpretation: the selectional requirements of presuppositional verbs. *Lingua* **164**, 156–188.
Kiparsky P. and Kiparsky C.P. (1971). Fact. *Semantics* **1**, 345–369.
Kłeczek D. (2020). PolBERT: attacking Polish nlp tasks with transformers. In M. Ogrodniczuk, L. Kobyliński (eds), *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences.
Lee K., Artzi Y., Choi Y. and Zettlemoyer L. (2015). *Event detection and factuality assessment with non-expert supervision*. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon: Association for Computational Linguistics, pp. 1643–1648.
Levinson S.C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
Levinson S.C. (2001). Pragmatics. In *International Encyclopedia of Social and Behavioral Sciences*, vol. 17. Oxford: Pergamon, pp. 11948–11954.
Lotan A., Stern A. and Dagan I. (2013). *Truthteller: annotating predicate truth*. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 752–757.
McCoy R.T., Pavlick E. and Linzen T. (2019). Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv: 1902.01007*.
Minard A.-L., Speranza M., Urizar R., Altuna B., van Erp M., Schoen A. and van Son C. (2016a). *MEANTIME, the NewsReader multilingual event and time corpus*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association (ELRA), pp. 4417–4422.
Minard A.-L., Speranza M., Urizar R., Altuna B., Van Erp M., Schoen A. and Van Son C. (2016b). *Meantime, the newsreader multilingual event and time corpus*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4417–4422.
Mroczkowski R., Rybak P., Wróblewska A. and Gawlik I. (2021). *HerBERT: efficiently pretrained transformer-based language model for Polish*. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, Kiyv: Association for Computational Linguistics, pp. 1–10.
Nairn R., Condoravdi C. and Karttunen L. (2006). *Computing relative polarity for textual inference*. In *Proceedings of the Fifth International Workshop on Inference in Computational Semantics (icos-5)*.
Ogrodniczuk M., Głowińska K., Kopeć M., Savary A. and Zawisławska M. (2014). *Coreference: Annotation, Resolution and Evaluation in Polish*. Berlin: Walter de Gruyter GmbH & Co KG.
Oren Y., Sagawa S., Hashimoto T.B. and Liang P. (2019). Distributionally robust language modeling. *arXiv preprint arXiv: 1909.02060*.
Öztürk E.Ö. (2017). A corpus-based study on 'regret' as a factive verb and its complements. *European Journal of Foreign Language Teaching* **2**, 88–108.
Özyıldız D. (2017). *Factivity and Prosody in Turkish Attitude Reports*. Cambridge: MIT Ling Lunch.
Parrish A., Schuster S., Warstadt A., Agha O., Lee S.-H., Zhao Z., Bowman S.R. and Linzen T. (2021). *NOPE: a corpus of naturally-occurring presuppositions in English*. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 349–366.
Partee B. (2014). Topic, focus and quantification. In *Semantics and Linguistic Theory*, pp. 159–188.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Przepiórkowski A., Bańko M., Górski R.L. and Lewandowska-Tomaszczyk B. (eds.) (2012). *Narodowy Korpus Języka Polskiego (National Corpus of Polish Language)*. Warsaw: Wydawnictwo Naukowe PWN.

Pustejovsky J., Verhagen M., Sauri R., Littman J., Gaizauskas R., Katz G., Mani I., Knippen R. and Setzer A. (2006). *TimeBank 1.2. LDC2006T08*. Philadelphia: Linguistic Data Consortium.

Richardson K., Hu H., Moss L. and Sabharwal A. (2020). *Probing natural language inference models through semantic fragments*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8713–8721.

Ross A. and Pavlick E. (2019a). *How well do NLI models capture verb veridicality?* In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, pp. 2230–2240.

Ross A. and Pavlick E. (2019b). *How well do nli models capture verb veridicality?* In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2230–2240.

Rudinger R., White A.S. and Van Durme B. (2018). *Neural models of factuality*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans: Association for Computational Linguistics, pp. 731–744.

Sauerland U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy* **27**(3), 367–391.

Saurí R. and Pustejovsky J. (2009). FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation* **43**(3), 227–268.

Saurí R. and Pustejovsky J. (2012). Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics* **38**(2), 261–299.

Shavrina T., Fenogenova A., Anton E., Shevelev D., Artemova E., Malykh V., Mikhailov V., Tikhonova M., Chertok A., Evlampiev A. (2020). *RussianSuperGLUE: a Russian language understanding evaluation benchmark*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 4717–4726.

Simons M., Beaver D., Roberts C. and Tonhauser J. (2017). The best question: explaining the projection behavior of factives. *Discourse Processes* **54**(3), 187–206.

Speaks J. (2021). Theories of meaning. In Zalta E.N. (ed), *The Stanford Encyclopedia of Philosophy*, Spring 2021 edn. Stanford: Metaphysics Research Lab, Stanford University.

Stalnaker R., Munitz M.K. and Unger P. (1977). *Pragmatic presuppositions*. In *Proceedings of the Texas Conference on Per Formatives, Presuppositions, and Implicatures*. Arlington: Center for Applied Linguistics, ERIC, pp. 135–148.

Tonhauser J. (2016). Prosodic cues to presupposition projection. In *Semantics and Linguistic Theory*, vol. 26, pp. 934–960.

Tsohatzidis S.L. (2012). How to forget that "know" is factive. *Acta Analytica* **27**(4), 449–459.

Tu L., Lalwani G., Gella S. and He H. (2020). An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics* **8**, 621–633.

Turri J. (2011). Mythology of the factive. *Logos & Episteme* **2**(1), 141–150.

White A.S., Reisinger D., Sakaguchi K., Vieira T., Zhang S., Rudinger R., Rawlins K. and Van Durme B. (2016). *Universal decompositional semantics on Universal Dependencies*. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin: Association for Computational Linguistics, pp. 1713–1723.

Williams A., Nangia N. and Bowman S. (2018). *A broad-coverage challenge corpus for sentence understanding through inference*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans: Association for Computational Linguistics, pp. 1112–1122.

Zhang Z., Wu Y., Zhao H., Li Z., Zhang S., Zhou X. and Zhou X. (2020). *Semantics-aware BERT for language understanding*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 9628–9635.

Ziembicki D. (2022). *Lingwistyczna analiza zjawiska faktywności (na materiale współczesnej polszczyzny)*. Warsaw: Warsaw University.

Appendix A. Examples from our dataset

In the following, there are a few examples from the *LingFeatured NLI* dataset and their descriptions.

<T, H> Example 30.
***T**: [PL] Jeżeli ksiadz odmówi mszy za zmarłego, lub pogrzebu z powodu zbyt niskiej zapłaty wiedzac, że proszacy jest bardzo biedny to rzeczywiście nie jest w porzadku, ale to inna sprawa.*

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

T: *[ENG] If a priest refuses Mass for a deceased person or a funeral because he received too little money, knowing that the payer is very poor, it is of course not right, but it is another matter.*
H: *[PL] Proszacy jest bardzo biedny.*
H: *[ENG] The payer is very poor.*
Sentence type—conditional
Verb labels: wiedzieć, że/know that, present, epistemic, factive, negation does not occur
Complement tense—present
GOLD—Neutral

Example 30: Despite the factive verb *wiedzieć, że/know that*, GOLD label is neutral. This is because the whole utterance is conditional. An additional feature not included in the example is that the whole sentence does not refer to specific objects but is general.

<T, H> Example 31.
T: *[PL] Nie spodziewałeś się, że kiedykolwiek to ode mnie usłyszysz, co?*
T: *[ENG] You never expected to hear that from me, did you?*
H: *[PL] Kiedykolwiek to ode mnie usłyszysz.*
H: *[ENG] You heard it from me.*
Sentence type—interrogative
Verb labels: spodziewać się, że/expect that, past, epistemic, non-factive, negation occurs
Complement tense—future
GOLD—Entailment

Example 31: Despite the non-factive verb *spodziewać się, że/expect that*, GOLD label is entailment. In this pair, non-lexical mechanisms are the basis of the entailment relation. Proper judgment of this example requires consideration of the prosodic structure of T's utterance.

It is worth noting that the sentence **H** is incorrectly written—strictly speaking, it should be "**H'**:*You have heard it from me./Usłyszałeś to ode mnie.*" So it is since **H** sentences were extracted semi-automatically. However, we did not want to change the linguistic features of the complement. The annotators were informed that in such situations, they should take into account not the **H** sentence, but its proper form—in the above case, it is **H'**. From the perspective of bilingualism of the set, it is also vital that the information provided by the expression "*never*" forms part of the main clause. In the Polish language, this content conveys the expression "*kiedykolwiek*" and is part of the complement clause.

<T, H> Example 32.
T: *[PL] może się bał że się wygadam. . .*
T: *[ENG] maybe he was afraid that I would spill the beans. . .*
H: *[PL] Się wygadam.*
H: *[ENG] I would spill the beans.*
Sentence type—indicative
Verb labels: bać się, że/afraid that, past, emotive, non-factive, negation does not occur
GOLD—?

Example 32: Example in which linguists decided to label a "?".[y] Whether the state of affairs reflected by the complement clause was realized, it belongs to the common knowledge of the interlocutors. Without context, we are not able to say whether the sender spilled the beans or not. It is also worth noting that in the English translation, the modal verb is present. This element is absent in the Polish complement clause. We can also see that the lack of context does not make it possible to determine the **H** sentence.

[y]These utterances were removed from the dataset for training and testing our benchmarks.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

<T, H> Example 33.
T: *[PL] I dlatego nie starałem się przypomnieć, myślałem, że nikt tu o mnie nie pamięta.*
T: *[ENG] And that's why I made no effort to remind anyone of myself, I thought nobody here would remember me.*
H: *[PL] nikt tu o mnie nie pamięta*
H: *[ENG] nobody here would remember me.*
Sentence type—indicative
Verb labels: myśleć, że/think that, past, epistemic, non-factive, negation does not occur
GOLD—Contradiction

Example 33: The main verb is non-factive, and the relation between the whole sentence and its complement is a contradiction. The grounding of this relation has a pragmatic nature.

Appendix B. Polish-English translation

Our dataset is two-lingual. We translated its original Polish version into English. In the following, we described methodological challenges and doubts related to the creation of the second language version and the solutions we have adopted.

We translated the whole dataset into English. First, we used the deepL translator, and then, a professional translator corrected the automatic translation.[z] The human translator acted following the guidelines: (a) not to change the structure "[verb] "to"—"that" [complement clause]," provided the sentence in English remained correct and natural, (b) to keep various types of linguistic errors in translation.

We believe that the decision on whether the translator knows how to use the dataset is important from the methodological point of view. Therefore, we decided to inform the translator that it is crucial for us that the translated sentence retains its set of logical consequences, especially the relation between the whole sentence and its complement clause. However, we did not provide the translator with a GOLD column (annotations of specialist linguists). The translator was aware that, in her task, this aspect is essential. On the other hand, during her work on each sentence, she had to assess the Polish relation and try to keep it in translation.

The English version differs from the Polish in several important issues. Each Polish sentence contains a complementizer *że*/*that*. In English, we can observe more complementizers, especially *that to* and other, for example *about*, *for*, *into*. There are also sentences without a complementizer. In Polish, a complementizer cannot be elliptical, in contrast to English (e.g., *Nie planowali, że będa mieli wielka, babska rodzinę./They didn't plan they will have a big, girl family*) In some English sentences, an adjective, a noun, or a verb phrase has appeared instead of a verb; for example, *The English will appear in a weakened line-up for these meetings.* (in Polish: *okazuje się, że Anglicy*. . .)

It happens that depending on the sentence, the Polish verb has more than one English equivalent, for example *cieszyć się—glad that* or *happy to*; *realize that—zdawać sobie sprawę* or *zrozumieć*. (In the dictionaries, *zrozumieć* is closest to *understand*). For this reason, the frequency of verbs is different in respective sets. Different language versions also pose problems related to verb signatures. First of all, the signatures developed by us are for Polish verbs. Therefore, we do not know how many pairs <V(pl); V(eng)> there are, where verbs have identical signatures (factive or non-factive). Secondly, a verb in language L1 may not have its equivalent in language L2 and vice versa.

Given these problems, it should be noted that the translated dataset is in a way artificial. In particular, we do not know whether the distribution between the factive/non-factive verbs and ECN relation in an English corpus (for instance BNC) would be similar, the more so interdependencies between them.

[z] *https://www.deepl.com/translator*

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Appendix C. Test annotations by non-experts

Table 14 shows examples of annotations made by non-experts in our study.

Table 14. Annotation examples for non-experts. "Annot." indicate non-expert annotations

Labelname	Value
T	[PL] *Wiem, że zimno degraduje umysł i wiedzie go do ospałości.* [ENG] *I know the cold degrades the mind and makes it sluggish.*
H	[PL] *Zimno degraduje umysł i wiedzie go do ospałości* [ENG] *The cold degrades the mind and makes it sluggish.*
Task	GOLD—E, Annot.—E E E E
T	[PL] *Etatyści wierza, że ludzie moga wymagać od państwa.* [ENG] *Statists believe that people can be demanding towards the state.*
H	[PL] *Ludzie moga wymagać od państwa.* [ENG] *People can be demanding towards the state.*
Task	GOLD—N, Annot.—N N N N
T	[PL] *Myślałam, że jesteś kawalerem.* [ENG] *I thought you were a bachelor.*
H	[PL] *Jesteś kawalerem.* [ENG] *You were a bachelor*
Task	GOLD—N, Annot.—C C N N
T	[PL] *Wyobrażałem sobie, że jak jest wina, to jest i kara.* [ENG] *I imagined that if there was guilt, then there was punishment.*
H	[PL] *Jak jest wina, to jest i kara.* [ENG] *If there was guilt, then there was punishment.*
Task	GOLD—C, Annot.—N C N C
T	[PL] *Być może zamawiała pochopnie, ale naprawdę liczyła, że stadion pozostanie miejscem handlu, nie sportu.* [ENG] *She may have ordered hastily, but she really hopedthat the stadium would remain a place of trade, not sport.*
H	[PL] *Stadion pozostanie miejscem handlu.* [ENG] *The stadium would remain a place of trade, not sport.*
Task	GOLD—C, Annot.—N C N N
T	[PL] *Zastanawiałam się, kiedy można mówić o tym, że życie kobiety miało sens. . .* [ENG] *I was wondering when you could talk about a woman's life making sense. . .*
H	[PL] *Życie kobiety miało sens. . .* [ENG] *A woman's life making sense. . .*
Task	GOLD—N, Annot.—N ? N ?

Appendix D. Factive and non-factive verbs

Factive Verbs: [e] czuć, że_; [e] poczuć, że_; [e] spostrzec, że_; [e] widać, że_; [e] widzieć, że_; [e] zauważać, że_; [e] zauważyć, że_; [e] zobaczyć, że_; [p] czuć, że_; [p] dostrzec, że_; [p] poczuć, że_; [p] słychać, że_; [p] słyszeć, że_; [p] widać, że_; [p] widzieć, że_; [p] zauważyć, że_; [p] zobaczyć, że_; brać pod uwagę, że_; być dumnym, że_; być świadomym, że_; cieszyć się, że_; dać poznać po sobie, że_; domyślać się, że_; domyślić się, że_; dowiadywać się, że_; dowiedzieć się, że_; dziwić się, że_; mieć świadomość, że_; odkrywać, że_; odkryć, że_; olśniło kogoś, że_; orientować się, że_; pamiętać, że_; podejrzewać, że_; pojmować, że_; pokazać2, że_; pokazywać2, że_; pomyśleć, że_;

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

poznać, że_; przekonać się, że_; przekonywać się, że_; przeoczyć, że_; przewidzieć, że_; przypominać [komuś], że_; przypominać [sobie], że_; przypomnieć [komuś], że_; przypomnieć [sobie], że_; przyznawać, że_; przyznać, że_; rozumieć, że_; ukazać, że_; unaocznić, że_; uprzedzić, że_; uprzytomnić sobie, że_; uwzględnić, że_; uzmysławiać sobie, że_; uznać, że_; uświadamiać komuś, że_; uświadamiać sobie, że_; uświadomić komuś, że_; uświadomić sobie, że_; wiadomo [komuś], że_; wiedzieć, że_; wspominać1, że_; wyjść na jaw, że_; wypominać [komuś], że_; wytykać, że_; wyznać, że_; zaobserwować, że_; zapamiętać, że_; zapominać, że_; zapomnieć, że_; zdawać sobie sprawę, że_; zdać sobie sprawę, że_; zdziwić się, że_; zorientować się, że_; zrozumieć, że_; zważyć, że_; zwrócić uwagę na to, że_; zwąchać, że_; śmiać się [z tego], że_.

Non-Factive Verbs: [e] czuć, że_; [m] wskazywać, że_; [m] zauważyć, że_; [p] obserwować, że_; alarmować, że_; argumentować, że_; bać się, że_; bywać, że_; być pewnym, że_; być przekonanym, że_; być zdania, że_; być zgodnym, że_; coś pokazało, że_; coś pokazuje, że_; czepiać się, że_; czytać, że_; dać do zrozumienia, że_; dać słowo honoru, że_; dać znać, że_; decydować, że_; deklarować, że_; dodawać, że_; dodać, że_; domyślać się, że_; dopowiedzieć, że_; dowieść, że_; dowodzić, że_; grozić, że_; grzmieć, że_; gwarantować, że_; głosić, że_; informować, że_; jęczeć, że_; krakać, że_; ktoś nie ma wątpliwości, że_; lamentować, że_; liczyć się z czymś, że_; liczyć, że_; martwić się, że_; meldować, że_; mieć nadzieję, że_; mieć wrażenie, że_; mniemać, że_; myśleć, że_; móc się domyślać, że_; mówić, że_; nadmienić, że_; napisać, że_; napomknąć, że_; narzekać, że_; nie mieć wątpliwości, że_; nie ulega watpliwości, że_; niepokoić się, że_; obawiam się, że_; obawiać się, że_; obić się o uszy, że_; obliczać, że_; obliczyć, że_; oceniać, że_; ocenić, że_; oczekiwać, że_; odczuwać, że_; odkryć, że_; odnieść wrażenie, że_; odnosić wrażenie, że_; odpisać, że_; odpowiadać, że_; odpowiedzieć, że_; ogłosić, że_; okazać się, że_; okazywać się, że_; opowiadać, że_; orzec, że_; oszacować, że_; oznajmić, że_; oświadczać, że_; oświadczyć, że_; pisać, że_; planować, że_; podawać, że_; podać, że_; podejrzewać, że_; podkreślać, że_; podkreślić, że_; podpisywać, że_; podpowiadać, że_; poinformować, że_; pokazać1, że_; policzyć, że_; pomyśleć, że_; postanowić, że_; potwierdzać, że_; potwierdzić, że_; powiadać, że_; powiedzieć, że_; powtarzać, że_; przeczuwać, że_; przeczytać, że_; przekazywać, że_; przekonać [kogoś], że_; przekonywać [kogoś], że_; przekonywać, że_; przepowiadać, że_; przerazić się, że_; przewidywać, że_; przyjmować, że_; przyjąć, że_; przypuszczać, że_; przyrzekać, że_; przysiac, że_; przysięgać, że_; przyznawać się, że_; przyznać się, że_; rechotać, że_; rozumieć, że_; skłamać, że_; spodziewać się, że_; sprawiać wrażenie, że_; sprawiać, że_; sprawić, że_; stwierdzam, że_; stwierdzać, że_; stwierdzić, że_; sugerować, że_; sygnalizować, że_; szacować, że_; szepnać, że_; szeptać, że_; sądzić, że_; słyszeć, że_;[mowa] twierdzić, że_; tłumaczyć się, że_; tłumaczyć, że_; uczyć, że_; udawać, że_; udać, że_; udokumentować, że_; udowadniać, że_; udowodnić, że_; umówić się, że_; upewnić się, że_; upierać się, że_; uroić sobie, że_; uspokajać, że_; usprawiedliwiać się, że_; ustalać, że_; ustalić [na podstawie czegoś], że_; ustalić, że_; usłyszeć, że_;[mowa] utrzymywać, że_; uważać, że_; uwierzyć, że_; uzasadniać, że_; uznawać, że_; uznać, że_; wciskać, że_; wierzyć, że_; wmawiać, że_; wmówić, że_; wnioskować, że_; wnosić z czegoś, że_; wskazywać, że_; wspomnieć2, że_; wybrzydzać, że_; wychodzić komuś, że_; wychodzić z założenia, że_; wyczytać, że_; wydaje się [komuś], że_; wydaje się, że_; wygląda na to, że_; wyjaśniać, że_; wyjaśnić, że_; wykazać, że_; wykazywać, że_; wykluczać, że_; wyliczyć, że_; wynikać, że_; wyobrazić sobie, że_;1 wyobrazić sobie_;2 wyobrażać sobie, że_;1 wyobrażać sobie, że_;2 wypisywać, że_; wyszło komuś, że_; wytłumaczyć, że_; wątpić, że_; zadeklarować, że_; zakładać, że_; zapewniać, że_; zapewnić, że_; zapisać, że_; zaplanować, że_; zapowiada się, że_; zapowiadać, że_; zapowiedzieć, że_; zaproponować, że_; zaprzeczyć, że_; zarzekać się, że_; zarzucać, że_; zarzucić, że_; zaręczać, że_; zastrzec, że_; zastrzegać, że_; zasugerować, że_; zawiadamiać, że_; zaznaczać, że_; zaznaczyć, że_; założyć się, że_; założyć, że_; zaświadczać, że_; zdaje się [komuś], że_; zdarzać się, że_; zdarzyć się, że_; zdawać się, że_; zdecydować się, że_; zdecydować, że_; zeznać, że_; zgadzać się, że_; zgodzić się, że_; zgłaszać, że_; zobowiązać się, że_; zreflektować się, że_; zwracać komuś uwagę, że_; łudzić się, że_; śmiać się, że_; śnić się, że_; świadczyć, że_; żalić się, że_; żartować, że_; żałować, że_.

https://doi.org/10.1017/S1351324923000220 Published online by Cambridge University Press

Appendix E. Undersampling experiments

As the dataset is unbalanced, additional experiments were conducted on the balanced dataset *LingFeatured NLI*. For this purpose, 107 (number of observations in the smallest "C" class) observations were randomly selected from each class so that each class had the same contribution to training and evaluation. Table 15 shows the results of these experiments. Random Forest achieves the highest performance among the models tested on the balanced dataset. It surpasses the best transformer-based model by 9.18 percentage points in terms of F1 score.

Table 15. Classification results of entailment (E), contradiction (C), and neutral (N) using the ***LingFeatured NLI*** dataset in balanced settings. Linguistic features comprise the verb, the grammatical tense of the verb, the occurrence of internal negation, the grammatical tense of the complement clause, utterance type, verb semantic class, and verb type (factive/non-factive). F1 score depicts the weighted F1 score. LF denotes linguistic features. The results for the baselines are calculated over the entire input dataset

		F1 score (%)				
Model	Input	All	C	E	N	Accuracy (%)
RandomForest	LF	**82.41 ± 6.69**	**79.36 ± 8.35**	**93.79 ± 3.69**	**74.12 ± 13.24**	**82.67 ± 6.69**
HerBERT	verb	68.92 ± 9.48	69.26 ± 12.27	82.66 ± 8.59	55.50 ± 16.44	69.43 ± 8.89
HerBERT	sentence	72.27 ± 8.84	73.88 ± 8.84	79.47 ± 10.83	63.18 ± 14.19	72.87 ± 8.10
PolBERT	verb	67.89 ± 5.44	67.99 ± 6.81	85.39 ± 8.02	49.97 ± 10.72	68.84 ± 5.24
PolBERT	sentence	73.23 ± 6.85	73.06 ± 11.74	74.22 ± 6.66	72.68 ± 8.56	73.52 ± 6.43

Cite this article: Ziembicki D, Seweryn K and Wróblewska A (2024). Polish natural language inference and factivity: An expert-based dataset and benchmarks. *Natural Language Engineering* **30**, 385–416. https://doi.org/10.1017/S1351324923000220