An Annotation Scheme for Factuality and its Application to Parliamentary Proceedings

Gili Goldin
Dept. of Comp. Sci.
University of Haifa
gili.sommer@gmail.com

Shira Wigderson
Dept. of Comp. Sci.
University of Haifa

Ella Rabinovich
The Academic College
of Tel-Aviv Yaffo

Shuly Wintner
Dept. of Comp. Sci.
University of Haifa

Abstract

Factuality assesses the extent to which a language utterance relates to real-world information; it determines whether utterances correspond to facts, possibilities, or imaginary situations, and as such, it is instrumental for fact checking. Factuality is a complex notion that relies on multiple linguistic signals, and has been studied in various disciplines. We present a complex, multi-faceted annotation scheme of factuality that combines concepts from a variety of previous works. We developed the scheme for Hebrew, but we trust that it can be adapted to other languages. We also present a set of almost 5,000 sentences in the domain of parliamentary discourse that we manually annotated according to this scheme. We report on inter-annotator agreement, and experiment with various approaches to automatically predict (some features of) the scheme, in order to extend the annotation to a large corpus.

1 Introduction

With the abundance of information and the rise of generative AI, "fake" information, such as fake news or fake product reviews, is becoming increasingly ubiquitous. To evaluate the veracity of information, it is necessary to first identify which utterances are candidates for such verification. *Factuality* assesses the extent to which a language utterance relates to real-world information; it is a measure that determines whether utterances correspond to facts, possibilities, or imaginary situations. Factuality is a complex notion that has been studied in various disciplines, using varying domain-specific definitions and terminologies. The degree of factuality with which a speaker makes a claim amalgamates values for agency, ambiguity, authoritativeness, certainty, credibility, commitment, confidence, hedging, approximation, modality, perspective, stance, polarity, and more.

It is important not to confuse factuality with *veracity* or *truthfulness*. The factuality of a proposition does not align it with ground truth facts. Nevertheless, determining the factuality of sentences is a necessary step toward achieving this latter goal; specifically, factuality can help assess if a claim involves information that is potentially *fact-checkable* or *check-worthy*.

We describe a complex, multi-faceted annotation scheme of factuality that we apply to Hebrew texts (§4). The scheme amalgamates various linguistic and extra-linguistic cues that help identify factuality. Our ultimate goal is to annotate a sizable corpus of Hebrew parliamentary proceedings according to this scheme, thereby providing the infrastructure necessary for identifying fake information in Hebrew texts.

We manually annotated 4,987 Hebrew sentences from a corpus of parliamentary proceedings according to this scheme and assessed inter-coder agreement on this complex (and often subjective) task (§5). Next, we focused on one aspect of the scheme, namely the *check-worthiness* feature, and evaluated various models on predicting this feature (§6). We show that off-the-shelf SOTA GPT models perform rather poorly on this classification task, whereas fine-tuned Hebrew LLMs that use the annotated data are much more accurate. We use the best performing model to automatically annotate the entire parliamentary corpus for this feature.[1]

2 Related Work

Much of the information pertaining to factuality is encoded linguistically, and various computational works employ linguistic information to identify related patterns in text. To train classifiers that can predict (aspects of) factuality, annotated corpora are required, and several corpora include annotations that highlight information pertaining to factuality. Existing annotation schemes vary with

[1] All the resources and code we developed are available at our GitHub repository and are released under the Creative Commons Attribution-ShareAlike 4.0 International License.

Proceedings of Recent Advances in Natural Language Processing,pages 403–412
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-049

respect to the basic unit for annotation (claim, sentence, paragraph, text, etc.) and the target tag (what exactly is being coded). Classifiers that identify factuality again vary with respect to the basic unit for classification, the features used for representing each instance, the actual prediction and the classification model. We survey several examples below.

Annotation schemes were developed for research findings, hypotheses, and evidence-providing in scientific articles (Teufel, 2000; Shatkay et al., 2008). For example, FactBank (Saurí and Pustejovsky, 2009) is a corpus of newswire texts with annotations of perspective, polarity and factuality on a graded (12-point) scale. This annotation is text-based, "avoiding any judgment based on knowledge of how things are in the world" (Saurí, 2008, p. 137). FactBank was extended with subjective judgments: annotators indicated whether they believed the event described did (or will) happen (de Marneffe et al., 2012). Based on their annotated version of FactBank, de Marneffe et al. (2012) developed a classifier to predict what they call *veridicality*, a gradual property. The features used for classification were mostly linguistic, with some external knowledge.

Saurí and Pustejovsky (2012) proposed a linguistically-motivated computational model to distinguish facts from negated facts, qualifying them by certainty to label events as "possibly factual" or "possibly counterfactual". They identified lexical, morphosyntactic, and frame semantic markers relevant to two aspects of factuality: polarity and modality. They refer to an event as the atomic unit for factuality, and use 19 features for each *word* rather than for each sentence, as a single sentence may include more than one event. The features included linguistic cues, source or speaker information and event properties. These features ultimately generate a degree of factuality, which touches on both *polarity* and epistemic *modality* distinctions as encoded in factuality markers, and includes also the *source* assigning the factuality value to an event. *Modality* has three values: certain (CT), probable (PR), and possible (PS); polarity values are either positive (+) or negative (−). This leads to a total of six factuality values: CT+, PR+, PS+, CT-, PR-, PS-, plus a seventh value, Uu, for underspecified.

MAVEN-Fact (Li et al., 2024a) is a large-scale event factuality detection dataset. It includes factuality annotation of over 112K events, each labeled as one of 5 classes, based on the polarity and modality of an event (Saurí and Pustejovsky, 2009; Qian et al., 2018). All these datasets are in English, although datasets in other languages begin to emerge (e.g. Atanasova et al., 2018; Barrón-Cedeño et al., 2018; Hasanain et al., 2020; Nakov et al., 2021).

The scheme that we describe in §4 combines various facets of factuality that are drawn from all these works and more. We focus on linguistic markers that can be identified in the text, and adapt existing approaches to the special case of Hebrew, a language with complex morphology and deficient orthography (Itai and Wintner, 2008; Fabri et al., 2014; Wintner, 2014).

Automatic prediction of factuality is a well-established computational task. *ClaimBuster* (Hassan et al., 2017) classified claims in US presidential debates as *non-factual*, *unimportant factual*, and *check-worthy factual*. Gencheva et al. (2017) evaluated factuality for full sentences, but also considered longer texts. They adopted ClaimBuster's sentence level features, but modified them in various ways. They also used context level features: the *position* of the sentence in the segment, *segment size* including the size of the previous and next segments, and *metadata*.

Konstantinovskiy et al. (2021) annotated sentences into 7 categories: personal experience, quantity in the past or present, correlation or causation, current laws or rules of operation, prediction, other type of claim, or not a claim. Each claim was also given a binary value, whether it is *checkable* or not.

More recent approaches to detection of factuality mainly focus on the outputs of large language models (LLMs), in particular on identifying hallucinations; see Wang et al. (2024) for a review. Contemporary works use LLMs to predict factuality in various domains (Zhang et al., 2024; Azizov et al., 2024; Li et al., 2024b), as we do here (§6).

3 Data

Factuality is a core factor in all types of communication, but it is paramount in argumentative genres like political discourse, where implicit strategies may exist that help the authors avoid commitment, feign authoritativeness, or keep (checkable) facts vague. Therefore, we focus in this work on the Knesset Corpus (Goldin et al., 2025), a large-scale dataset of parliamentary proceedings, including both plenum sessions and committee deliberations, spanning several decades. This dataset is also annotated with rich meta-information, linguistic features

and named entities.

We manually annotated almost 5,000 sentences from this corpus for the full factuality scheme proposed in §4. The sentences were sampled from protocols of both plenary sessions and committee deliberations, balancing across various factors such as year, author's gender, political affiliation, native language, and more. This manually annotated subset is available at our GitHub repository.

4 An Annotation Scheme for Factuality

We present a complex, multi-faceted annotation scheme that characterizes various aspects that contribute to factuality. The scheme includes several features that can assist evaluators to verify the truth value of a proposition. The basic unit of annotation is a sentence, but sentences that include several claims can be annotated with sequences of features, one set per claim. We list and exemplify various elements of the scheme and the values they take (for convenience, the original sentences are translated to English in the examples). More examples are given in our GitHub repository. The scheme is organized by *layers*, each including several features.

4.1 Check-worthiness

The first layer of the annotation consists of a check-worthiness estimation for each claim in the sentence. We provide several types of scores, based on previous works (see §2).

Check-worthiness score Following Hassan et al. (2017) we include an overall score of "check-worthiness". This feature can take the following possible values: *worth checking*, *not worth checking*, *not a factual proposition*. This score also subsumes the binary score used by Gencheva et al. (2017). Typical claims that are marked 'not a factual proposition' include questions and speech acts.

Claim type Following Konstantinovskiy et al. (2021), each claim in the sentence is assigned a claim-type. The possible values are: *personal experience*, *quantity in the past or present*, *correlation or causation*, *current laws or rules of operation*, *prediction*, *other type of claims*, *not a claim*, and *irrealis mood*.

Factuality profile Following Saurí and Pustejovsky (2012), we assign each claim a *factuality profile*: a pair that consists of the source and the factuality value that combines modality and polarity (see §2).

Examples

1. '*There are 700,000 pensioners in the State of Israel.*' check worthiness score='worth checking'; claim type='quantity in the past or present'; factuality profile='⟨speaker, CT+⟩'.

4.2 Event Selecting Predicates

Saurí and Pustejovsky (2009) propose that in many cases, the factuality of events is conveyed by *event selecting predicates (ESPs)*, which are predicates that select for an argument denoting an event. ESPs qualify the degree of factuality of their embedded event, which can be presented as a fact in the world, a counterfact, or a possibility. They distinguish between two kinds of ESPs: *source introducing predicates (SIPs)* (e.g., '*say*', '*know*') and *non-source introducing predicates (NSIPs)* (e.g., '*want*'). If an ESP is found in the sentence, we classify it as SIP or NSIP, based on whether or not the predicate introduces a new source. For SIPs we also mark the source they introduce, if present in the sentence, and mark a connection between the two.

ESP All ESPs present in the sentence are marked.

ESP type For each ESP, is it SIP or NSIP.

Examples

1. '*I don't remember that they ever approached me as a member of the finance committee and I didn't respond to their request.*' ESP="*remember*"; ESP type='SIP'; ESP source="*I*".

4.3 Agency

The existence of an agent and its characteristics impact the factuality of a proposition. The absence of an agent blurs the meaning of the sentence, and agent-less utterances are often used to avoid commitment. Even if an agent does exist, its referent may be vague (e.g., *we*). This component specifies the agent of the proposition, its position in the proposition, its animacy and its morphological properties, as well as the predicate it is an agent of and the relation between them.

Agent The (tokens making up the) agent in the proposition. If it is a pronoun, we indicate the referent, if known, separated by comma.

Experiencer The (tokens making up the) experiencer in the proposition. If it is a pronoun, we indicate the referent, if known. Unlike the agent,

the experiencer is not performing an action, but rather receives sensory or emotional input.

Agent-less If there is no agent in the proposition, we explain why; values of this feature include *passive w/o by-clause*, *middle_verb*, *infinitival clause*, *'there is (not)' + infinitive*, *impersonal modal verb*, *imperative*, *nominal clause*, *existential* and *unspecified* (we use 'unspecified' when the agent exists, but has no identifiable referent).

Position *subject, indirect object, embedded pronoun subject* or *embedded pronoun non subject*.

Animacy *human, animate, inanimate*.

Morphology *1/2/3, sg/pl*.

Examples

1. '*The Ministry of Transportation is promoting additional reforms, which include inter-city fare reform.*' agency agent="*The Ministry of Transportation*", position='*subject*', animacy='*inanimate*', morphology='*3sg*', agency-predicate="*promotes*".

4.4 Stance

This component characterizes the speaker's belief towards the proposition: the extent to which the speaker is certain or uncertain about what they claim. Does the speaker express a wish or a fact? Do they provide a reference (statistics, numbers, citations, etc.) to substantiate their proposition?

Previous works suggested various ways to annotate stance: Marín Arrese (2011) used the distinction between effective and epistemic stance, Pyatkin et al. (2021) suggested a modal classification that maps into similar categories of stance, which they called *priority and plausibility modality*, and Saurí and Pustejovsky (2012) and de Marneffe et al. (2012) characterized stance as a double-axis scale: polarity(binary) and modality(continuous). The combination of these two characteristics constitutes the factuality profile that is part of our checkworthiness score (see §4.1).

We combine several of these suggestions to describe stance. Our stance marking includes an overall confidence level, whether it is effective or epistemic, the polarity, and a list of the lexical items that imply these values.

Confidence level The overall confidence level of the speaker towards their proposition, as evaluated by the annotator; the level of confidence is determined by the lexicon the speaker uses. The lexicon plays an important role in determining the confidence level of the speaker towards their statement, which is why we include a list of lexical items, mapped to each level of confidence. The list of lexical items is composed also of items from various sources (Netzer et al., 2007; Hooper, 1975; Gencheva et al., 2017). The values of this feature are *high, mid*, *low* and *irrelevant* (if the sentence does not contain a proposition).

Stance type We adopt the terminology of Marín Arrese (2011) and distinguish between *effective* and *epistemic* stance. When the speaker **explicitly** expresses *their attitude* towards the proposition it is marked as an effective stance; when the speaker expresses *knowledge* or *estimation* regarding the proposition and the possibility of its realization it will be marked as epistemic stance. When there is no proposition in the sentence, we mark it as 'irrelevant'. Some examples of effective stance markers include: '*must*', '*impossible*', and '*hoped*'. Examples of epistemic stance compatible items include '*I'm certain*', '*the fact is*', and '*it is evident*'.

Polarity Following Saurí and Pustejovsky (2012), we mark the polarity of the proposition. The values are *positive, negative*, or *underspecified*. We also indicate the lexical item(s) that helped the annotator decide the polarity of the proposition, according to a list we provide (available in the supplementary materials). The type and the lexical item are presented as pairs. Note that usually, for a sentence to be positive it does not necessarily need to have a positive lexical item, it just needs to have no negative lexical items.

Negative polarity lexical items include '*no/not*','*never*', '*no one*'; positive ones include '*there is*' and '*indeed*'. Underspecified is assigned when the source is not committed to the polarity of their statement.

Reference Whether the speaker mentions a source, and specifies the source's type and name (which can help determine how reliable that source is or how confident they are towards their proposition being true). Values are pairs of [*name*, *type*], where *name* is the name of the source that is being referred to and *type* can be one of the following: *article, book, research, survey, stats, numbers, laws, other*, quoting an expert/authority figure, etc.

Examples

1. '*Without a doubt, this issue of fare parity is an important one.*' stance confidence level='high', stance type='epistemic', polarity='(positive,)'.

4.5 Hedging

This component highlights lexical items used by speakers to express a low level of commitment towards their claim.

Hedge The lexical item used for hedging. A partial list of hedges includes '*many*', '*others*', '*often*', '*sometimes*', '*approximately*'.

Examples

1. '*These are not exactly unemployment benefits; a grant of up to NIS 4,000.*' hedge="*up to*"

4.6 Quantities

This component indicates whether the proposition contains quantitative expressions, and if so, how precise they are. The motivation is the assumption that the more accurate a quantity is, the more check worthy the proposition is. The following features characterize a quantified expression:

Exp The numerical expression in the sentence. When the expression is a percentage or a fraction, we indicate the whole part it refers to, if known, and mark the relation between them. The values of this feature are the literal numeral expressions.

Quantifier The quantifier in the sentence. Possible values include '*every*', '*there is*', and '*a few*'.

QuantifierType The type of the quantifier, if one exists. Values are *universal, existential*, and *partial*.

Accuracy Is the quantity an estimation, an accurate one, or obscured. Values are *accurate, estimate*, and *obscure*. This feature is only valid for numerical quantities, not for quantifiers.

Examples

1. '*There are hundreds and thousands of people who have retired, and as a fact, they have no pension.*' quantity exp="*hundreds*", accuracy='estimate', quantity exp="*thousands*", accuracy='estimate'.

4.7 Named entities

This component highlights named entities and specifies their type. As the Knesset Corpus is already annotated for named entities, we simply adopted the annotation, correcting it in rare cases.

Name The tokens of which the entity consists.

Type The type of the named entity, one of the types defined by the IAHLT NER guidelines.

4.8 Time Expressions

The time at which the proposition is claimed to have happened (or will happen) helps determine how check worthy it is. Time expressions can be exact dates or parts thereof (e.g., *1994*); relative time expressions like *last month*, *next year*, or *the year the war started*; fuzzy time expressions like *(sometime) in the past five years*; etc.

Time expressions may describe a time-range, such as '*during last year*', *2018-2019*, etc. When a time range is expressed we mark the start point and the end point of the range. If one of them is missing from the data, we leave it empty; e.g., in '*in recent years*' we can assume the end point is the time of utterance, but we do not know for sure what is the start point of this time range.

When possible, we infer date and time information from the meta-data. E.g., expressions such as '*yesterday*' are interpreted relatively to the actual day of the session in which the text is included.

TimeExp The date and time expression, in the format `YY:MM:DD:HH:MM:SS`, where unspecified or missing components of the timeExp are left empty.

Token The tokens that make up the TimeExp.

Examples

1. '*On January 30, the Ministry of Health announced that the arrival of the corona virus in Israel is a matter of time.*' timeExp token="*30th January*", timeExp='2020:01:30:::'.

Sometimes the speaker refers to the current event at which they are speaking using time and location expressions such as '*here*' and '*now*'. We mark such expressions when we are certain that they refer to the current meeting, as this might help fact checkers verify the given statement.

5 Annotator Agreement Evaluation

Factuality is a complex notion and in particular, check-worthiness is highly subjective (Konstantinovskiy et al., 2021). To assess agreement across annotators we designated 100 sentences that were annotated by each of the three annotators. In some cases, annotators did not agree on the number of

claims in the sentence, leading to discrepancies in both the number of claims identified and which claims were tagged, complicating the evaluation of inter-coder agreement. We therefore focus on 65 sentences for which all annotators agreed on the set of claims, which serve as our evaluation set. We also measured agreement for a subset of the evaluation set, containing 56 single-claim sentences.

Another challenge in assessing agreement is the complexity of the scheme, which assigns a large, structured tag to each claim. Due to the high number of features, even a single feature with inconsistent annotation would count as disagreement, naturally resulting in a low agreement score that may not accurately reflect the true level of agreement. To mitigate this, we divided the features into *layers*, each containing between 4 and 10 features, and assessed agreement separately for each layer. Additionally, we specifically examined agreement on the primary feature, *check-worthiness score*.

The results for each evaluation set and each layer are presented in Table 1. The scores represent the level of agreement among the three annotators: A score of 3 indicates the percentage of sentences in the test set, where all three annotators agreed on the labels of *all* features in the layer. A score of 2 reflects cases where two out of three annotators reached full agreement and a score of 1 signifies that all annotators differed on *at least one* label.

Evidently, for the strictest agreement score (3), the results are relatively low for some of the layers, such as *Check-worthiness*, *Stance* and *Agency and ESP*, highlighting the difficulty of these tasks. However, majority agreement—where we count agreement among two or three annotators—is significantly higher. This is especially notable given the large number of features in each set and the fact that many features involve marking specific string spans within the text, where even a single-character difference would be counted as a disagreement.

We also calculated mean pairwise agreement, using Kappa (Cohen, 1960), among the annotators and between each annotator and the other two, on the *check-worthiness score* feature, to asses how consistently the annotators agreed on their judgments. Mean pairwise Kappa between each of the annotators and the other two ranged between 0.5 and 0.63; overall, among the three annotators, it was 0.58. The similar agreement scores across annotators indicate a consistent annotation process with reasonable reliability.

6 Annotation of the Knesset Corpus

Our ultimate goal is to automatically annotate all the sentences in the Knesset Corpus with the complex, multifaceted factuality tag, as detailed in §4. This is a large-scale task that will require multiple models; we are initially focusing on our primary feature, *check-worthiness score*. This feature has three possible values (§4): *worth checking*, *not worth checking* and *not a factual proposition*.

First, we used off-the-shelf GPT models to annotate the corpus for this feature (§6.1). Due to the limited success, we transitioned to more traditional model training approaches(§6.2), significantly improving the results.

6.1 Annotation with GPT Models

As an initial approach, we experimented with GPT models to predict the *check-worthiness score* feature. Our goal was to evaluate whether these models could provide accurate predictions, potentially reducing the need for extensive manual labeling. We conducted several experiments using GPT-4 (OpenAI et al., 2024) and GPT-4o, which are considered state-of-the-art for similar tasks. We set the temperature of the models to zero in order to ensure deterministic outputs. We evaluated these models on the one-claim sentences evaluation set described in §5, which contains 56 sentences. We chose this set because it is both reliable, having been annotated by all three annotators, and relatively simple, as each sentence contains only one claim, making it suitable for establishing a baseline for these GPT models. The predictions generated by the models were then compared to the majority vote of the human annotators. We now describe the different experiments we conducted with these models in an attempt to achieve the highest performance. The full results of all the experiments are compiled in Table 2.[2]

Zero-Shot Prompting We first established a baseline for the models by evaluating their performance without providing explanations or additional examples. The prompts used for this experiment are listed on GitHub.

Table 2 depicts the results: accuracy of 37.5% with GPT-4o and 30.4% with GPT-4. To better in-

[2]The evaluation set is unusually small, but it is high-quality, as these sentences were annotated by three experts. Given that the GPT models perform so poorly on the relatively easy task of annotating a single-claim sentence, there was not much point testing them on a more complex test set. We do test our fine-tuned models more rigorously (§6.2).

Test Set	Full (65 sents.)			Single-claim (56 sents.)		
Layer	3	2	1	3	2	1
check-worthiness score	56.92	43.08	0	58.93	41.07	0
Check-worthiness	30.77	47.69	21.54	35.71	48.21	16.07
Agency and ESP	35.38	36.92	27.69	37.50	29.29	23.21
Stance	35.38	41.54	23.08	39.29	44.64	16.07
Quantities	67.69	32.31	0	69.64	30.36	0
Time Expressions	92.31	4.62	3.08	94.64	3.57	1.79

Table 1: Annotator agreement score (% of sentences annotated identically) by layer and test set.

	Accuracy		Kappa	
Prompting Technique	**GPT 4**	**GPT 4o**	**GPT 4**	**GPT 4o**
Zero-shot	30.4	37.5	0.09	0.16
Hebrew prompts, English labels	35.7	37.5	0.14	0.13
Hebrew prompts, Hebrew labels	28.6	37.5	0.06	0.12
Instruction-Based	32.1	48.2	0.10	0.28
Few-shot	60.7	58.9	0.25	0.25

Table 2: GPT models accuracy, as evaluated on annotators' majority vote, and mean pairwise Kappa agreement between models and annotators. Recall that Kappa for human annotators ranged between 0.5 and 0.63.

terpret these results, we also calculated the mean pairwise Kappa agreement between the model and each of the annotators. This allowed us to assess the annotation quality in comparison to human annotation. The results, presented in Table 2, indicate that the models performed significantly worse compared to human annotators.

Hebrew Prompting The sentences we annotate are in Hebrew; we hypothesized that using Hebrew prompts might improve model performance, following Behzad et al. (2024) who suggested that prompting in a different language may yield better results. To test this, we conducted experiments using two different setups: 1. A Hebrew prompt with labels remaining in English. 2. A Hebrew prompt with labels also translated into Hebrew. The prompts used in these experiments are similar to those in the zero-shot experiment, but are translated to Hebrew. The results of these experiments are presented in Table 2. In most cases Hebrew prompts did not improve the results compared to the English prompts, suggesting that the models' performance is not significantly influenced by the prompt language in this case.

Instruction-based Prompting We conducted an additional experiment where we provided the model with explicit definitions for each label, along with simple examples, before asking it to classify sentences. The accuracy with this approach was 48.2% for GPT-4o and 32.1% for GPT-4, indicating a slight improvement in performance. A minor improvement is evident also in the mean pairwise agreement scores.

Few-shot Prompting In this experiment we also included four real examples from the annotated dataset (disjoint from the evaluation set of course) for each label, in addition to the explanations provided in the previous experiment. This approach led to a significant improvement in results, with GPT-4o achieving 58.9% accuracy and GPT-4 achieving 60.7% accuracy. Kappa scores are 0.25, better than the previous approaches but still far below human agreement. We conducted additional experiments with varying numbers of examples from the dataset, but these variations did not lead to a significant change in results.

6.2 Experiments with Fine-tuned Models

Since prompt engineering for GPT models did not yield satisfactory results, we turned to pre-trained encoder-based Hebrew models and fine-tuned them for our task. Specifically, we experimented with AlephBertGimmel (Gueta et al., 2022), DictaBERT (Shmidman et al., 2023) and Knesset-DictaBERT (Goldin and Wintner, 2024) models. To train and evaluate these models, we used the 4,987 sentences that had been manually annotated by human annotators as described in §3, excluding the sentences that were used for annotation agreement evaluation.

	Accuracy		Kappa	
Model	**Test set**	**Single-claim Set**	**Test set**	**Single-claim Set**
AlephBertGimmel	74.61	76.79	0.59	0.61
DictaBERT	75.50	76.79	0.60	0.62
Knesset-DictaBERT	77.26	78.57	0.63	0.63

Table 3: Accuracy and Kappa agreement of fine-tuned models on the test set and on the single-claim evaluation set, according to annotators' majority vote in the case of the single-claim set, and a single annotation in the case of the larger test set. Recall that Kappa for human annotators ranged between 0.5 and 0.63.

Since the *check-worthiness score* feature is annotated for each claim in a sentence, while we wanted to annotate each sentence with a single value, we adjusted the labels for each sentence as follows: if at least one claim in the sentence was labeled as *worth checking*, the label given to the sentence was *worth checking*. If neither one of the claims is check worthy, but at least one of the claims was labeled as *not a factual proposition*, then this was the label given to the sentence. Otherwise, the label was *not worth checking*.

The models were trained on 80% of the dataset, while 10% was used for tuning and the remaining 10% (488 sentences) constitute the test set. We also evaluate our models on the evaluation set of 56 single-claim sentences (§5). The full details of the training process, along with a set of fully-annotated examples, are available on our GitHub repository.

The classification results for these datasets are presented in Table 3. They indicate that fine-tuned models achieved significantly better results compared to GPT models. This suggests that, despite GPT's proven strengths across many NLP tasks, for this challenging and non-trivial classification task, fine-tuning models on labeled data provides a clear advantage. Among the models tested, Knesset-DictaBERT achieved the highest accuracy. This is unsurprising, as this model underwent domain adaptation specifically tailored to Knesset data. Given its superior performance, we will use this model to annotate the entire Knesset Corpus for check-worthiness.

7 Conclusion

We presented a complex annotation scheme for factuality and a set of 4,987 manually-annotated sentences, of which 100 are annotated thrice. These resources are open and can be used to train fact-checking applications. We release the full Knesset Corpus automatically annotated for check-worthiness and the fine-tuned Knesset-DictaBERT model. We also release a set of annotated examples and the prompts used to train the models.

This is an ongoing project, and our current effort centers on developing methods for predicting the complex annotation tags that our scheme defines on a large corpus of parliamentary proceedings texts. The initial experiments that we reported on here are promising, but they are limited to a single feature of the scheme (albeit a critical one), and in future work we would like to predict the full factuality structures. Other plans for future work include adaptation of our scheme to other languages, with a focus on morphologically-rich languages. Finally, we plan to explore how factuality is manifested in different groups of the parliament (e.g., government vs. opposition).

Limitations Our study provides a valuable schema for factuality annotation and preliminary experimental results, yet it has several limitations worth discussing. First, we automatically annotated the Knesset Corpus for only one aspect of the schema, the check-worthiness score. Future work will focus on developing and refining models for the automatic annotation of other schema elements. Second, our annotation experiments with contemporary LLMs have been limited to GPT-4o. Finally, while we believe the schema could be applied to other domains and languages, this has yet to be demonstrated.

Ethical Considerations Our main dataset is the Knesset Corpus which is open and publicly available. Adding annotation of factuality cannot, in our eyes, be abused or dual-used.

We employed three linguists (two women, one man, all native Hebrew speakers residing in Israel) as annotators. They were recruited directly (i.e., not via any crowd-sourcing platform) and were paid an hourly wage that is approximately 2.5 times the minimum wage in Israel. No human participants were required for this project.

Acknowledgments

We are indebted to Piroska Lendvai for suggesting the topic of this research and for extensive collaboration during the initial phases of the project. Many of the ideas we implemented are hers. We are grateful to Israel Landau and Avia Vaknin for their meticulous annotation efforts. We thank the Idit PhD Fellowship Program at the University of Haifa for supporting the first author. This research was supported by the Ministry of Science & Technology, Israel under grant no. 3-17990.

References

Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Dilshod Azizov, Zain Muhammad Mujahid, Hilal AlQuabeh, Preslav Nakov, and Shangsong Liang. 2024. SAFARI: Cross-lingual bias and factuality detection in news media and news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12217–12231, Miami, Florida, USA. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. task 2: Factuality. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2024. To ask LLMs about English grammaticality, prompt them in a different language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15622–15634, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Ray Fabri, Michael Gasser, Nizar Habash, George Kiraz, and Shuly Wintner. 2014. Linguistic introduction: The orthography, morphology and syntax of Semitic languages. In Imed Zitouni, editor, *Natural Language Processing of Semitic Languages*, Theory and Applications of Natural Language Processing, pages 3–41. Springer, Berlin Heidelberg.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276, Varna, Bulgaria. INCOMA Ltd.

Gili Goldin, Nick Howell, Noam Ordan, Ella Rabinovich, and Shuly Wintner. 2025. The Knesset corpus: an annotated corpus of Hebrew parliamentary proceedings. *Language Resources and Evaluation*, 59(3):2973–3004.

Gili Goldin and Shuly Wintner. 2024. Knesset-DictaBERT: A Hebrew language model for parliamentary proceedings. *Preprint*, arXiv:2407.20581.

Eylon Gueta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2022. Large pre-trained models with extra-large vocabularies: A contrastive analysis of Hebrew BERT models and a new one to outperform them all. *Preprint*, arXiv:2211.15199.

Maram Hasanain, Fatima Haouari, Reem Suwaileh, Zien Sheikh Ali, Bayan Hamdan, Tamer Elsayed, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. 2020. Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1803–1812, New York, NY, USA. ACM.

Joan B. Hooper. 1975. *On Assertive Predicates*, pages 91 – 124. Brill, Leiden, The Netherlands.

Alon Itai and Shuly Wintner. 2008. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, 2(2).

Chunyang Li, Hao Peng, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li. 2024a. Maven-fact: A large-scale event factuality detection dataset. *Preprint*, arXiv:2407.15352.

Chunyang Li, Hao Peng, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li. 2024b. MAVEN-FACT: A large-scale event factuality detection dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11140–11158, Miami, Florida, USA. Association for Computational Linguistics.

Juana Marín Arrese. 2011. Effective vs. epistemic stance and subjectivity in political discourse: Legitimising strategies and mystification of responsibility. In Chris Hart, editor, *Critical discourse studies in context and cognition*, pages 193–224. John Benjamins.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021. COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 997–1009, Held Online. INCOMA Ltd.

Yael Netzer, Meni Adler, David Gabay, and Michael Elhadad. 2007. Can you tag the modal? you should. In *Proceedings of the ACL-2007 Workshop on Computational Approaches to Semitic Languages*, pages 57–64.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Valentina Pyatkin, Shoval Sadde, Aynat Rubinstein, Paul Portner, and Reut Tsarfaty. 2021. The possible, the plausible, and the desirable: Event-based modality detection for language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 953–965, Online. Association for Computational Linguistics.

Zhong Qian, Peifeng Li, Yue Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Event factuality identification via generative adversarial networks with auxiliary classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4293–4300. International Joint Conferences on Artificial Intelligence Organization.

Roser Saurí. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Computer Science Department, Brandeis University.

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

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

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. DictaBERT: A state-of-the-art BERT suite for Modern Hebrew. *Preprint*, arXiv:2308.16687.

Simone Teufel. 2000. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

Shuly Wintner. 2014. Morphological processing of Semitic languages. In Imed Zitouni, editor, *Natural Language Processing of Semitic Languages*, Theory and Applications of Natural Language Processing, pages 43–66. Springer, Berlin Heidelberg.

Xiaokang Zhang, Zijun Yao, Jing Zhang, Kaifeng Yun, Jifan Yu, Juanzi Li, and Jie Tang. 2024. Transferable and efficient non-factual content detection via probe training with offline consistency checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12348–12364, Bangkok, Thailand. Association for Computational Linguistics.