On Shortcuts and Biases: How Finetuned Language Models Distinguish Audience-Specific Instructions in Italian and English

Nicola Fanton[1,2] **Michael Roth**[2]
University of Stuttgart
[1] Interchange Forum for Reflecting on Intelligent Systems
[2] Institute for Natural Language Processing
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

Instructional texts for different audience groups can help to address specific needs, but at the same time run the risk of perpetrating biases. In this paper, we extend previous findings on disparate social norms and subtle stereotypes in wikiHow in two directions: We explore the use of fine-tuned language models to determine how audience-specific instructional texts can be distinguished and we transfer the methodology to another language, Italian, to identify cross-linguistic patterns. We find that language models mostly rely on group terms, gender markings, and attributes reinforcing stereotypes.

Bias Statement

In this study, bias is defined as systematic differences in content and presentation of wikiHow articles that are tailored to different audiences, particularly in ways that can reinforce gender stereotypes or inequities. Such biases include the allocation of topics in a way that reinforces traditional gender stereotypes as well as the use of language that perpetuates hetero-normative gender roles.

Following Blodgett et al. (2020), we recognize that bias is not merely a technical issue but a deeply embedded social problem that reflects structural inequalities. This work analyzes social constructs, as described in collaboratively edited how-to guides, in which biases operate and which, when used as training data, can raise issues in NLP systems.

Potential harms of biased data, as defined above, include unequal access to information, exposure to content that can affect self-esteem and self-worth, as well as limiting individual aspirations. We identify sources of underlying biases in the data as a starting point for editors to create fairer content and for developers to foster more ethical AI systems. As such, our work aims to actively promote diversity and inclusion on a specific online platform and to generally contribute to a more nuanced understanding of origins of gender bias in NLP.

Flirtare Via SMS (Per Ragazze) "Flirting Via SMS (For Girls)" *Lascia che sia **lui** il primo a scrivere!* "Let **him** be the first one to write!"
Essere Figo alle Superiori (per Ragazzi) "Being Cool in High School (for Boys)" *Focalizza l'attenzione **sulle ragazze**.* "Focus attention **on the girls**."

Table 1: Examples from wikiHow in Italian.

1 Introduction

Instructional texts aim to convey the necessary knowledge for readers to accomplish specific tasks. On the collaboratively edited online platform wikiHow, hundreds of thousands of instructional texts are available on a variety of topics and in multiple languages. The goal, or mission, of this vast repository is to democratize access to knowledge and skills across diverse subject matters.[1] Among other works on wikiHow, prior research has explored in how far texts are formulated in linguistically inclusive terms and which adjustments are made for specific target audiences (Suhr and Roth, 2024; Fanton et al., 2023). However, these previous studies primarily relied on simple classifiers and focused exclusively on English texts, leaving a gap in understanding multilingual phenomena and if fine-tuning language models might exacerbate biases (see §2).

Acknowledging the limitations of prior research to English, we first compile a new dataset in a less resourced language, specifically Italian (see §3). Our initial research question investigates how texts for different target audiences in English and Italian vary in terms of the topics they address (see §4). This exploration directly contributes to the analysis of social biases in the data (see Table 1 for an example). To draw comparisons with previous

[1] https://www.wikihow.com/wikiHow:Mission

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 78–93

research, we then explore how articles for different target groups can be distinguished computationally and which characteristics are learned in this process (see §5). Unlike previous work, we employ fine-tuned language models and utilize a well-established interpretation method, integrated gradients (Sundararajan et al., 2017). This approach represents a recent advancement beyond simple classifiers to interpreting more sophisticated models that can provide deeper insights into language use and biases.

In short, we make the following contributions:

- We release a new data collection, **wikiHowAudIT** (short wHA-IT), and assess the audience-specific biases in how-to guides by a topic-based data analysis.
- We cross-lingually compare biases in wHA-IT and in an existing English dataset, wikiHowAudiences (short wHA-EN; Fanton et al., 2023), by fine-tuning and analyzing language models for audience classification.

2 Related Work

In this section we briefly review three related areas: Our work continues a series of recent contributions dealing with the collection of data sets for Italian. While there exists little work on instructional texts for Italian, data in English has been examined and tested from different angles and perspectives in the NLP community. Finally, work on model-based data interpretation has received increasing attention, but almost no work studied biases in audience-specific instructional texts.

Italian NLP datasets. Recent data collections for the Italian language include DIATOPIT (Ramponi and Casula, 2023a), a dataset representative in time and space on variations of non-Standard Italian. A new shared task for geo-locating the linguistic variation in Italy (Ramponi and Casula, 2023b) is based on this data collection. Another recent effort for the Italian language is IRMA, a data collection for studying misinformation (Carrella et al., 2023). In their paper, the authors curated a dataset from untrustworthy websites, and emphasized its significance for the less-represented language studied. Minnema et al. (2023) advance the task of responsibility perspective transfer, in the context of studying gender-based violence, and a dataset of sentences for Italian news about femicides. To the best of our knowledge, there are no previous studies on how-to guides in the Italian language.

Instructional texts. Anthonio et al. (2020) introduced wikiHowToImprove, a data collection of original and revised sentences based on wikiHow articles and their revision histories. Kojima et al. (2021) contribute with a continual approach for instruction generation. Fanton et al. (2023) examine audience-specific wikiHow guides in English. They find traces of subtle biases, using shallow classifiers and qualitative analyses. In this work, we extend their findings to fine-tuned language models in two languages.

Interpreting Language Models. A number of methods have been proposed recently for interpreting fine-tuned language models. Our work makes use of Integrated Gradients (Sundararajan et al., 2017), which computes the gradients of a model's output with respect to the input, based on (stepwise) back-propagation and summation as an approximation method. Falk and Lapesa (2022) employ a variant of Integrated Gradients for getting attributions and importance scores. They point to the capabilities of such method(s) "to uncover potential biases picked up by the model". In their case, the reveal of these biases concerns how the model's class prediction is influenced by sensitive words. Luu and Inoue (2023) propose the Counterfactual Adversarial Training (CAT) technique, with the broader goal of improving LMs' robustness. They make use of Integrated Gradients in CAT for calculating tokens' salience, before obtaining the counterfactual perturbations. This is then put into practice by changing the thus extracted important tokens. Other works that rely on Integrated Gradients include studies on irony detection in Dutch (Maladry et al., 2023) and gender-based violence in Italian (Minnema et al., 2023), among others.

3 Data

We first build a data collection to investigate our first research question, namely how texts for different target audiences in Italian vary in terms of the topics they address. As a starting point, we use how-to guides from publicly available wikiHow dumps[2] for Italian. Out of 34,801 guides, 1,031 feature an indicator between round parentheses at the end of the title (see Table 1). For each guide featur-

[2] https://ftp.fau.de/kiwix/zim/wikihow/, we refer to this file: wikihow_it_maxi_2023-02.

Audience	wHA-IT	wHA-EN
Women (W)	143	993
Men (M)	100	209
Kids (K)	22	499
Teens (T)	158	411
Total	423	2,112

Table 2: Distribution of articles across target groups.

C	Cluster Name	K	T
0	routines	20	13
1	attitudes	15	15
2	relationships and friendships	20	18
3	clothes and style	5	11
4	preparation and organization	20	15
5	self-care	5	18
6	school and work	15	8

Table 3: Cluster assignments (percentages) for the two audience groups pertaining the K-T task in wHA-IT.

ing a group indicator, we use wikiHow's *Esporta*[3] service to get the latest version. Following previous work (Fanton et al., 2023), we manually categorize the indicators into four target groups: Women (W), Men (M); Kids (K), Teens (T). Similar to previous work, we find that there is a lack of indicators for non-binary/other groups (see §A.1 for a complete list of common indicators), forcing us to consider only binary distinctions: Women–Men (W–M) and Kids–Tens (K–T).[4] Table 2 comprises the distribution over audience groups for the wikiHowAudIT (wHA-IT) corpus, which comprises a total of 423 how-to guides, and for the corpus from previous work (wHA-EN). For training, validation and testing, we create stratified experimental partitions for each task with a proportion of $8:1:1$ (see Table 12 in A.2 for details).

4 Data Analysis

We address our first research question, namely how texts for different target audiences in Italian vary in terms of the topics, by clustering articles according to their content. We describe the approach in §4.1 and findings in §4.2. For this part of our work, we focus exclusively on the TRAIN and DEV partitions of the data in Italian, so that the TEST part remains held-out for computational experiments (see §5).

[3] https://www.wikihow.it/Speciale:Esporta

[4] Note that an article may target two groups, meaning that some data points appear in both distinctions.

C	Cluster Name	W	M
0	organize activities	16	11
1	physical aspect and care	9	13
2	body-related (genitals)	9	10
3	body-related (care)	17	18
4	health	6	10
5	body-related (fat)	6	4
6	clothes and style	12	11
7	night-time	3	4
8	body-related (diet)	6	3
9	relationships and friendships	15	14

Table 4: Cluster assignments (percentages) for the two audience groups pertaining the W-M task in wHA-IT.

4.1 Clustering Approach

Our approach makes use of agglomerative clustering, using embeddings for capturing the contents of each article. First, we embed the articles with a sentence-transformer model[5] (Reimers and Gurevych, 2019). Second, we normalize the embeddings obtained. Third, we leverage the scikit-learn (Pedregosa et al., 2011) `AgglomerativeClustering` algorithm and default options to put into practice the clustering, with the distance threshold set to 1.5.[6] Finally, we review the titles of the guides assigned to each cluster in order to find an overarching topic.

Inspired by Montariol et al. (2021), we perform an additional validation for topics as cluster names. Specifically, we collect all word tokens within the articles of a cluster and sort them according to their tf-idf scores, providing us with the tokens that seem most relevant for the cluster. In order to select the

[5] The LM used here for wHA-IT is `nickprock/sentence-bert-base-italian-uncased` with input size 512 tokens, for wHA-EN `sentence-transformers/all-mpnet-base-v2` (384).

[6] The value of the distance threshold chosen is the default value implemented in the sentence-transformers library for the agglomerative clustering. For wHA-EN, we raised the threshold to 4 experimentally.

0	*stanza* "room"	*camera* "bedroom"	*tema* "theme"	*cose* "things"	*ta* "ta"	*genitori* "parents"	*cosa* "thing"
1	*ta* "ta"	*sopracciglia* "eyebrows"	*viso* "face"	*costume* "costume"	*lenti* "lenses"	*fascia* "band"	*capelli* "hair"
2	*cla* "cla"	*midi* "midi"	*pub* "pub"	*erta* "erta"	*infezione* "infection"	*vagina* "vagina"	*urina* "urine"
3	*capelli* "hair"	*pelle* "skin"	*peli* "hair"	*ila* "ila"	*viso* "face"	*lava* "washes"	*crema* "cream"
4	*ta* "ta"	*genitori* "parents"	*sito* "site"	*cosa* "thing"	*tosse* "cough"	*parlare* "speak"	*medico* "doctor"
5	*peso* "weight"	*calorie* "calories"	*perdere* "loose"	*esercizi* "exercises"	*im* "im"	*pesa* "weighs"	*pesi* "weights"
6	*vestiti* "clothes"	*indossa* "wears"	*pantaloni* "trousers"	*abbigliamento* "clothing"	*scarpe* "shoes"	*stile* "style"	*indossare* "wear"
7	*sveglio* "awake"	*00* "00"	*sveglia* "alarm"	*notte* "night"	*letto* "bed"	*restare* "remain"	*colazione* "breakfast"
8	*calorie* "calories"	*peso* "weight"	*mag* "may"	*dieta* "diet"	*pasti* "meals"	*mangiare* "eat"	*grasso* "fat"
9	*lui* "him"	*lei* "she"	*ragazzo* "boy"	*ragazza* "girl"	*cosa* "thing"	*gay* "gay"	*parlare* "speak"

Table 5: Highest scoring tokens (*Italian*, "translated") for each cluster in the TRAIN ∪ DEV parts of the **W–M** data.

most discerning tokens, for each cluster we leave out the tokens featured in all the other clusters.

We execute agglomerative clustering for each task separately: one time for the task W–M and once for K–T. For cross-lingual comparison, we perform the same steps for the wHA-EN corpus introduced by Fanton et al. (2023).

4.2 Cluster Findings

For the task W–M in wHA-IT, we found 10 clusters. For the task K–T, we found 7 clusters. An overview of the clusters for both tasks are shown in Table 3 and 4, including topic-based cluster names and counts for each target group. For W–M we find a prevalent presence of body-related clusters (labelled with 1, 2, 3, 5, 8), as well as socially coded occupations (labelled with 0, 4, 6, 7, 9). Interestingly, there are two clusters (labelled with 5 and 8) that focus not only on physical aspect, but also more in detail about being fit. Additional details can be seen based on the highest-scoring tokens ("weight", "calories", "fat"), as summarized for all clusters in Table 5. For K–T, unlike the previous task, we find more behavioral and social activities.

In summary, our analysis on wHA-IT shows how the examined articles are clusterable by topical information across audiences, indicating that topics are not specific for a target group. Considered these overlaps, we remark that there are less topical biases than we had assumed and it may be interesting to see which differences a computational model learns for distinguishing audiences in Italian.

In wHA-EN, we find 11 clusters for W-M, distributed over both target groups (see Table 7). For K-T, we find 8 clusters (Table 8). For W–M, we meta-group the clusters. The activities to perform in specific places, like in school or outside are labelled with 0, 5, 7, 8. Moreover, a further distinction is between activities in relation to others (labelled with 2, 10) and activities in relation to oneself (labelled with 1, 3, 6, 9). However, cluster 4 (`appear and act`) cannot unambiguously be allocated to activities in relation to others, nor to activities in relation to oneself, because it features subtly disparate guides. As examples, we show two titles per audience from that cluster:

W: "Be Drama Free",
"Eat Healthy Around Your Friends"
M: "Look Handsome", "Be More Socially Open"

The first example each might imply to work more on oneself rather than in direct relation to others, but it is not possible to conclude exactly so for the other two. That is to say, to eat in a certain way around other people, and to be more socially open, requires at least some relation to others.

C	W–M									
0	party	her	paint	could	bedroom	parents	furniture	games	play	bag
1	shoes	black	jeans	colors	shirts	makeup	shirt	pair	color	shorts
2	her	she	him	he	enemy	crush	relationship	his	could	girlfriend
3	erty	pub	ac	dry	ne	ving	sha	shave	her	razor
4	her	makeup	popular	smile	act	others	he	she	teeth	talking
5	alarm	wake	homework	breakfast	makeup	teeth	clock	class	routine	early
6	weight	fat	cal	ories	foods	diet	muscle	exercise	muscles	lose
7	dance	date	her	makeup	she	shoes	him	party	he	dancing
8	pack	bag	camp	swim	suit	suitcase	items	packing	pool	locker
9	comb	dry	hairs	oil	gel	ay	condition	tyle	scalp	pr
10	he	him	his	crush	guy	smile	flirt	kiss	conversation	guys

Table 6: High scoring tokens for each cluster in the TRAIN ∪ DEV partitions of the **W–M** data (wHA-EN).

C	Cluster Name	W	M
0	fun activities	15	5
1	clothes and style	22	23
2	relationships and friendships	5	18
3	personal care	10	13
4	appear and act	17	13
5	routines and school	7	2
6	body-related (weight)	3	4
7	going out	6	9
8	vacation	5	1
9	hairstyles	1	8
10	love relationships	8	4

Table 7: Cluster assignments (percentages) for the two audience groups pertaining the W-M task in wHA-EN.

C	Designation	K	T
0	routines	8	9
1	lifestyle	8	18
2	young people's issues (general)	25	27
3	parties	8	7
4	money (games)	7	5
5	relationships	7	13
6	holiday	10	12
7	crafting	28	9

Table 8: Cluster assignments (percentages) for the two audience groups pertaining the K-T task in wHA-EN.

Table 6 shows the most discerning tokens for the clusters induced from the English data for W–M. We note that pronouns appear among the highest scoring tokens for several clusters (e.g. clusters 0, 2 and 10), which are at the same time large clusters that contain disproportionally many articles for one of the two target groups (cf. Table 7).

Cross-lingually, we find for W–M that relations to other people (e.g. `relationships and friendships`), as well as self-centered activities (`self-care`, `personal care`) are similarly present in English and Italian. In contrast, body-related topics only seem narrowly present in wHA-EN, whereas they are highly pervasive in wHA-IT. The topics for K–T are largely overlapping across languages. For instance, we find `routines` for both languages. Similarly, we find `relationships and friendships`, `clothes and style` in wHA-IT and `relationships`, `lifestyle` in wHA-EN. However, `self-care` emerges only from the Italian data, while `crafting` and `money (games)` are specific to the English data, which may point at a need for financial education of the younger generations (Lusardi, 2019).

5 Experimental Setup

Given the data and analyses of the previous sections, we next investigate what features and biases computational models learn when they are trained to distinguish articles for different audiences.

5.1 Models

As discussed in Section 2, previous work attempted to distinguish texts by target groups using simple classifiers. However, we take the results from our data analysis as an indicator that lexical and off-the-shelf representations might not be fully sufficient

for this task. In order to find more nuanced biases, we propose to fine-tune language models. We test whether this leads to a higher distinction and which patterns are learned in the process. For comparability, we adopt the same setups for LMs fine-tuned on wHA-IT and wHA-EN.

We employ a set of LMs from Hugging Face (Wolf et al., 2020) and set up binary classification tasks based on the previously discussed data. Due to computational constraints, we use LMs with a maximum length of 512 tokens. For Italian, these include the monolingual language models Italian BERT cased/uncased (Schweter, 2020), UmBERTo cased/uncased (Parisi et al., 2020) as well the multilingual models mBERT cased/uncased (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). For English, we follow previous work and only tested BERT-cased/uncased (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). For hyperparameter selection, we maximize the macro F_1 on the DEV set. We perform 3 trials for each LM and for each task, W–M and K–T, using Optuna (Akiba et al., 2019) as the optimization framework. More details on the tested LMs and used hyperparameters are listed in Appendix A.4.

5.2 Attributions

Based on the F_1 scores obtained for each task on the DEV sets, we select the best-performing LM for further analysis. We leverage the Transformers Interpret[7] library to inspect which are the parts of the articles that are relevant in distinguishing the audience-specific guides. Specifically, we pass the fine-tuned LM, their tokenizer and the (truncated) articles as inputs to the `SequenceClassificationExplainer`. The output of each pass is a list of attributions: tokens with respective scores. For W–M, each text is explained with respect to the class label W. For K–T, explanations are taken with respect to the label K. For each task, we first collect attributes for each article and then summarize them for the full task data by averaging the scores found for each article.

6 Results

We first discuss results in terms of model performance for the classification task itself (§6.1) and then analyze the attributions of the models that perform best at distinguishing audiences (§6.2).

[7] https://github.com/cdpierse/transformers-interpret

Task	wHA-IT	wHA-EN	Fanton et al.
W–M	0.83	0.86	0.71
K–T	0.60	0.82	0.78

Table 9: Macro F_1 on the TEST sets.

6.1 Performance

In Table 9, we report solely the performance of our best configuration (as determined on the development set) and comparison numbers from Fanton et al. (2023) on wHA-EN. Specifically, we use `bert-base-italian-cased` for both tasks on wHA-IT, and `roberta-base` and `bert-base-uncased` for W-M and K-T, respectively, on wHA-EN. More details on the experiments, i.e. the scores on the three experimental partitions for each corpus, can be found in Table 18 and Table 17 in A.4.

Cross-task comparisons. Considering the wHA-IT column, the F_1 score is higher for the W–M task than the one obtained for the K–T task. The same finding can be observed for the wHA-EN column. Intuitively, this result could be explained in that the categories of men and women are typically viewed by editors as more discrete than the categories of kids and teens, whose boundaries are continuous in general. This finding represents the opposite of previous work, where a lower score was obtained for the W–M task than for K–T (0.71 vs 0.78; Fanton et al., 2023). We note, however, that results are only partially comparable as Fanton et al. did not apply fine-tuned language models and their experimental setup did not account for stratified partitions.

Cross-language comparisons. We focus now on the W–M row. What emerges is that the performance of the LM finetuned for wHA-EN is slightly higher than the performance of the LM finetuned for wHA-IT (with a difference of about 3 percentage points). We observe a much larger difference for K–T, with a decrease in F_1 of around 22 percentage points. Both differences could be explained by the data scarcity for Italian (see Table 2), which seems particularly problematic for the K–T task.

It is further worth pointing out that multilingual models performed consistently worse in our experiments than monolingual models, suggesting that cross-lingual training might not be promising (see also Table 17 in A.4). This finding is in line with findings on the task of responsibility perception prediction for gender-based violence in Italian

	wHA-IT
W	*ragazze* (“girls”); *Se* (“If”); *donne*(“women”); *ragazza* (“girl”); *una* (“one”, f.) *sicura* (“sure”, f.); *non* (“not”): *la* (“the/her”, f.); *amica* (“friend”, f.); *amiche* (“friends”, f.);
M	*amici* (“friends”, m.); *uomini* (“men”); *stesso* (“same”, m.); *ragazzo* (“boy”); *uomo* (“man”); *amico* (“friend”, m.); *pronto* (“ready”, m.); *sicuro* (“sure”, m.); *modo* (“way”, m.); *quello* (“that”, m. sing.);
K	*in* (“in”); *da* (“from”); *a* (“to”); *se* (“if”); *il, m.* (“the”); *del* (“of the”, m. sing.); *per* (“for”); *o* (“or”); *prima* (“before”); *dei* (“of the”, m. plur.);
T	*non* (“not”); *articolo* (“article”); *Non* (“Not”); *è* (“is”); *le* (“her”, f. sing. / “the/them”, f. plur.); *troppo* (“too much/many”); *sono* (“am/are”); *capelli* (“hair”); *bella* (“beautiful/nice” , f.); *di* (“of”);

Table 10: Top-ranked tokens for each audience in wHA-IT. Highlighted tokens indicate feminine (f.) and masculine (m.) grammatical gender. A more comprehensive list with scores is provided in the Appendix.

	wHA-EN
W	girl; girls; your; Girls; you; she; women; You; her; makeup;
M	men; guy; him; boy; man; boys; He; he; guys; his;
K	kids; kid; children; middle; school; toys; people; pre; mom; use;
T	teen; the; and; are; if; a; your; is; teenage; for;

Table 11: Top-ranked tokens for each target group in wHA-EN. A full list of attributions with scores, including punctuation and sub-word tokens not reported here, are available in the Appendix.

(Minnema et al., 2022), where better performance was also observed by monolingual models.

6.2 Attributions

Our final analysis concerns the attributions by the language models with the highest results on each task, which provide us with insights on generalizable patterns learned from the training data. Table 10 and Table 11 show the top-10 tokens, after filtering of punctuation and sub-word tokens, for each audience in wHA-IT and wHA-EN, respectively.

“Group terms”. We observe that many of the top features to be direct addresses of the reader in terms of their group membership (“even if you’re a kid”). The presence of such “group terms” was also found in the analysis of word-based logistic classification models by Fanton et al. (2023).

For all audiences, our model analysis consistently shows fewer group terms among the top-ranked and filtered tokens in Italian, as compared to English. For example, 6 out of 10 top tokens for M in wHA-EN are group terms (‘men’, ‘guy’, ‘boy’, ‘man’ ‘boys’, ‘guys’), whereas for wHA-IT we only find *uomini* (“men”), *uomo* (“man”) and *ragazzo* (“boy”). Although we also observe such group terms for K–T in wHA-EN experiments (e.g. ‘kids’, ‘teen’), this is not the case for the experiments conducted with wHA-IT. If classifiers rely to a high degree on such “group terms” for classification, this finding might explain the low model performance for the Italian K–T data.

Negation. Another feature discussed in previous work concerns the presence of negations. Like in the case of English, we also find for wHA-IT that *non* (“not”) is among the 10 top-ranked features exactly for the audience W. As highlighted by Fanton et al. (2023), this might raise concerns as negations have been shown to be used in stereotype-maintaining function (Beukeboom et al., 2010, 2020). Consider the following example:

> *Se stai cercando di farti notare da qualcuno di cui ti sei infatuata o ti trovi al primo appuntamento con lui, non concederti troppo facilmente.*
> “If you’re trying to get noticed by someone you’re infatuated with or you’re on a first date with him, don’t give in too easily.”

This extract is from the guide titled *Apparire Bella Davanti al Tuo Ex Ragazzo (Solo Ragazze Adolescenti)* (“Looking Beautiful In Front Of Your Ex Boyfriend (Teenage Girls Only)”). It reinforces gender-roles, as the targeted audience (Teenage Girls) is not at all encouraged to make the first

move according to their feelings, but rather to stay passive, and to conform to the stereotype about men's agency (Ellemers, 2018). Moreover, instead of information about what to do, the instruction explicitly points out what "not" to do.

Grammatical gender. What is also interesting in the aforementioned example is the presence of heteronormativity, defined as heterosexuality as the norm (see Warner, 1991, and Vásquez et al., 2022). While this can already be inferred from the title, the explicit use of the masculine pronoun *lui* ("him") in the excerpt leaves no space for ambiguity in the interpretation of the assumed gender of the referent.

We can argue that *qualcuno* ("someone", m.), is encapsulating generic masculine (Silveira, 1980), as it is not *qualcuna* ("someone", f.). Unlike English, Italian features grammatical gender, in terms of which we find a polarising situation: feminine tokens (80%) for W and of masculine tokens (100%) for M (data: wHA-IT). This might provide a shortcut for classifiers to distinguish the instances in the (Italian) W–M task. For K–T, in contrast, we only find traces of masculine gender for kids (30%). Nonetheless, it is worth noting that the usage of generic masculine in Italian, especially, from Table 10, *dei* ("of the", m.) could capture cases of collective plurals, for which it is used a masculine plural to refer to groups of unknown genders (also to heterogeneous group in terms of gender).

> *Taglia i prati. Devi stabilire diverse tariffe in base alla dimensione del giardino. Fatti pubblicità nel quartiere attaccando qualche volantino alle porte dei vicini, ma cerca di essere discreto.*
> "Cut lawns. You need to set different rates based on the size of the yard. Advertise in the neighborhood by sticking a few flyers on neighbors' doors, but try to be discreet."

The sentences above are extracted from *Guadagnare dei Soldi (per Ragazzini)* ("Earn Money (for Kids)"). From those, *dei vicini* ("of the neighbors", m.) exemplifies masculine generics.

In summary, we find that grammatical gender in wHA-IT provides a shortcut for language models to distinguish instructions for different audiences. We provide additional attributions in a longer list in the appendix (see Table 21), containing also tokens that correspond to the same lemma: for example, *sicura* ("sure", f.) for W versus *sicuro* ("sure", m.) for M. In comparison, the longer list of top attributions for wHA-EN (see Table 19 in A.5), features tokens that represent rather stereotypical attributes such as "makeup", "pretty", "pink" for W, and "gentleman", "nerd", "handsome" for M.

7 Conclusion

We introduced wikiHowAudIT, a dataset of instructional texts from wikiHow for different audiences in Italian. Our data analysis has shown that wikiHowAudIT contains different topics across audiences, which makes computational modeling difficult. In order to still learn what biases can be found in texts for different audiences, we fine-tuned language models and investigated which attributes rank highest for each target group. As a result, we found that models perform very well even with training on only 100 data points and that they capture more fine-grained differences in English than simpler models from previous work.

However, our analysis of the attributes also confirmed trends already observed with simpler methods: Regardless of language, models consistently learn that texts for different audiences can be distinguished with high effectiveness based on group terms, grammatical gender, negations and stereotype reinforcing references. Several of these points may represent critical issues, particularly given that wikiHow is one of the most visited websites on the internet.[8] Our results further support existing findings on gender roles in other domains, such as in stories for children and educational resources for young age groups, where females are also associated with gender stereotypes (Adukia et al., 2022).

One reason for us to analyze texts regarding biases is that we want to understand assumptions structurally made about the readers and to what extent these potentially reflect actual characteristics. Future work should accordingly focus on how to identify and remove those biases that are inadequate (e.g. stereotypes) while maintaining adaptations that appeal to an audience (e.g. group terms). Future work could also include different languages and varieties in order to provide a wider understanding of the shortcuts and biases hereby highlighted. For deeper insights on the biases, we encourage future research that could, for example, mask shortcuts by LMs as identified in our study.

We believe that wikiHow is an ideal resource for

[8] https://www.wikihow.com/wikiHow:About-wikiHow

such work because its collaborative nature makes it possible to put changes directly into practice and instructional texts in general would strongly benefit being easier accessible and more inclusive.

Acknowledgements

This work is supported by the Ministry of Science, Research, and the Arts, Baden-Württemberg through the project IRIS3D (Reflecting Intelligent Systems for Diversity, Demography and Democracy, Az. 33-7533-9-19/54/5). Work by the second author was funded by the DFG Emmy Noether program (RO 4848/2-1).

Limitations

While the present work focuses on a less frequently studied language, namely Italian, in addition to English, the work is still limited culturally (i.e., to "western culture"). Critically, the considered audience attributes, gender and age, are subjected to a simplification that is for now lacking, in particular, intersectional perspectives (Crenshaw, 1991). Another limitation of this work lies in the focus on a single data source. For better generalizations over the instructional scenarios, it is important to contemplate other, different, data sources. The present work is by no means aimed at reinforcing representational bias. We conceive our research efforts as first steps towards inclusion, especially for queer identities, who can be audiences of instructions but are insufficiently accounted as such. With the present work, our hope is also to stimulate future work on instructions in other, especially under-represented, languages and cultures.

References

Anjali Adukia, Patricia Chiril, Callista Christ, Anjali Das, Alex Eble, Emileigh Harrison, and Hakizumwami Birali Runesha. 2022. Tales and tropes: Gender roles from word embeddings in a century of children's books. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3086–3097, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.

Camiel J. Beukeboom, Christian Burgers, Zsolt P. Szabó, Slavica Cvejic, Jan-Erik M. Lönnqvist, and Kasper Welbers. 2020. The negation bias in stereotype maintenance: A replication in five languages. *Journal of Language and Social Psychology*, 39(2):219–236.

Camiel J. Beukeboom, Catrin Finkenauer, and Daniël H. J. Wigboldus. 2010. The negation bias: When negations signal stereotypic expectancies. *Journal of Personality and Social Psychology*, 99(6):978–992.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Fabio Carrella, Alessandro Miani, and Stephan Lewandowsky. 2023. IRMA: the 335-million-word Italian coRpus for studying MisinformAtion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2339–2349, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Kimberle Crenshaw. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Naomi Ellemers. 2018. Gender stereotypes. *Annual Review of Psychology*, 69(1):275–298.

Neele Falk and Gabriella Lapesa. 2022. Scaling up discourse quality annotation for political science. In *Proceedings of the Thirteenth Language Resources*

and Evaluation Conference, pages 3301–3318, Marseille, France. European Language Resources Association.

Nicola Fanton, Agnieszka Falenska, and Michael Roth. 2023. How-to guides for specific audiences: A corpus and initial findings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 321–333, Toronto, Canada. Association for Computational Linguistics.

Noriyuki Kojima, Alane Suhr, and Yoav Artzi. 2021. Continual learning for grounded instruction generation by observing human following behavior. *Transactions of the Association for Computational Linguistics*, 9:1303–1319.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Annamaria Lusardi. 2019. Financial literacy and the need for financial education: evidence and implications. *Swiss Journal of Economics and Statistics*, 155(1).

Hoai Linh Luu and Naoya Inoue. 2023. Counterfactual adversarial training for improving robustness of pre-trained language models. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 881–888, Hong Kong, China. Association for Computational Linguistics.

Aaron Maladry, Els Lefever, Cynthia Van Hee, and Veronique Hoste. 2023. A fine line between irony and sincerity: Identifying bias in transformer models for irony detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 315–324, Toronto, Canada. Association for Computational Linguistics.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022. Dead or murdered? predicting responsibility perception in femicide news reports. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1078–1090, Online only. Association for Computational Linguistics.

Gosse Minnema, Huiyuan Lai, Benedetta Muscato, and Malvina Nissim. 2023. Responsibility perspective transfer for Italian femicide news. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7907–7918, Toronto, Canada. Association for Computational Linguistics.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alan Ramponi and Camilla Casula. 2023a. DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 187–199, Dubrovnik, Croatia. Association for Computational Linguistics.

Alan Ramponi and Camilla Casula. 2023b. GeoLingIt at EVALITA 2023: Overview of the geolocation of linguistic variation in Italy task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, Parma, Italy. CEUR.org.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stefan Schweter. 2020. Italian bert and electra models.

Jeanette Silveira. 1980. Generic masculine words and thinking. *Women's Studies International Quarterly*, 3(2-3):165–178.

Katharina Suhr and Michael Roth. 2024. A diachronic analysis of gender-neutral language on wikiHow. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 118–123, St. Julian's, Malta. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks.

Juan Vásquez, Gemma Bel-Enguix, Scott Thomas Andersen, and Sergio-Luis Ojeda-Trueba. 2022. HeteroCorpus: A corpus for heteronormative language detection. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 225–234, Seattle, Washington. Association for Computational Linguistics.

Michael Warner. 1991. Introduction: Fear of a queer planet. *Social Text*, (29):3–17.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

A.1 Indicators (Italian)

[(‘Android’, 135), (‘PC-o-Mac’, 115), (‘iPhone-o-iPad’, 103), (‘per-Ragazze’, 37), (‘Ragazze’, 22), (‘per-Donne’, 18), (‘Uomini’, 14), (‘Windows-e-Mac’, 14), (‘per-Ragazzi’, 13), (‘Per-Ragazze’, 13), (‘Adolescenti’, 12), (‘per-Adolescenti’, 12), (‘Windows’, 11), (‘Ragazzi’, 11), (‘PC-e-Mac’, 8), (‘per-Principianti’, 8), (‘per-Uomini’, 8), (‘Ragazze-Adolescenti’, 7), (‘per-Bambini’, 7), (‘iPhone’, 7), (‘Cristianesimo’, 6), (‘per-Ragazze-Adolescenti’, 6), (‘per-ragazze’, 5), (‘Donne’, 4), (‘Windows-10’, 4), (‘USA’, 3), (‘Principianti’, 3), (‘per-Cristiani’, 3), (‘PC’, 3), (‘Per-Uomini’, 3), (‘SEO’, 2), (‘Per-Ragazze-Adolescenti’, 2), (‘per-Preadolescenti’, 2), (‘Per-Ragazzi’, 2), (‘Jicama’, 2), (‘Windows-7’, 2), (‘MRI’, 2), (‘Per-gli-Uomini’, 2), (‘per-Ragazzini’, 2), (‘Per-Adolescenti’, 2), (‘RCP’, 2), (‘MRSA’, 2), (‘per-le-Donne’, 2), (‘Yoga’, 2), (‘Per-Ragazze-Teenager’, 2), (‘per-le-Adolescenti’, 2), (‘Negli-Stati-Uniti’, 2), (‘per-i-Ragazzi’, 2), (‘LAN’, 2), (‘per-Bambine’, 2), (‘DOC’, 2), (‘Scuola-Media’, 2), (‘Teenager’, 2), (‘Per-Uomini-Gay’, 2), (‘Atletica-Leggera’, 2), (‘Bambini’, 2), (‘DPTS’, 2), (‘Per-Bambini’, 2), (‘iOS’, 2), ...]

A.2 Experimental Partitions

Partition	Aud.	wHA-IT	wHA-EN
TRAIN	W	114	794
	M	80	167
	K	18	399
	T	126	329
DEV	W	14	99
	M	10	21
	K	2	50
	T	16	41
TEST	W	15	100
	M	10	21
	K	2	50
	T	16	41

Table 12: Breakdown of the partitions by audience.

A.3 Clustering results for K–T data

0	*sveglio*	*00*	*sveglia*	*notte*	*sonno*	*giornata*	*dormire*	*letto*
	"awake"	"00"	"alarm"	"night"	"sleep"	"day"	"sleep"	"bed"
1	*grin*	*capelli*	*ragazza*	*tsu*	*nam*	*ragazzo*	*stile*	*suo*
	"grin"	"hair"	"girl"	"tsu"	"nam"	"boy"	"style"	"her"
2	*lui*	*lei*	*ragazzo*	*ragazza*	*gay*	*baciare*	*relazione*	*bacio*
	"him"	"her"	"boy"	"girl"	"gay"	"to kiss"	"relation"	"kiss"
3	*pantaloni*	*jeans*	*camicia*	*stile*	*nerd*	*indossa*	*paio*	*abbigliamento*
	"pants"	"jeans"	"shirt"	"style"	"nerd"	"wears"	"pair"	"clothing"
4	*stanza*	*tema*	*camera*	*borsa*	*letto*	*carta*	*dip*	*gatto*
	"room"	"theme"	"bedroom"	"bag"	"bed"	"paper"	"dip"	"cat"
5	*capelli*	*viso*	*ila*	*doccia*	*idra*	*crema*	*sopracciglia*	*dep*
	"hair"	"face"	"ila"	"shower"	"hydra"	"cream"	"eyebrows"	"dep"
6	*sito*	*spia*	*studia*	*squadra*	*libri*	*appunti*	*leggere*	*estate*
	"site"	"spy"	"studies"	"squad"	"books"	"notes"	"light"	"summer"

Table 13: Highest scoring tokens (*Italian*, "translated") for each cluster in the TRAIN ∪ DEV parts of the **K–T** data.

C	**K–T**									
0	bed	night	sleep	bedroom	furniture	desk	alarm	morning	wake	clock
1	skin	girl	makeup	wash	style	ne	ac	jeans	moist	uri
2	learn	weight	healthy	phone	him	bully	stress	adult	he	eating
3	christmas	sleep	guests	snow	tree	santa	theme	gift	halloween	night
4	sell	business	car	lawn	items	bank	store	selling	pet	chores
5	him	he	she	guy	crush	girl	guys	boy	kiss	flirt
6	pack	plane	car	trip	items	horse	packing	books	phone	vacation
7	club	glue	blog	books	members	notebook	barbie	makeup	color	nail

Table 14: High scoring tokens for each cluster in the TRAIN ∪ DEV partitions of the **K–T** data (wHA-EN).

A.4 Modeling

Hyperparameter	**Set**
Seed	[22, 17, 4]
Learning rate	[2e-5, 2e-6]
Batch size	[4, 8]
Epochs	[5]

Table 15: Hyperparameters.

https://huggingface.co/model-name	Param.	Reference
google-bert/bert-base-uncased	1.10e+08	Devlin et al. (2019)
google-bert/bert-base-cased	1.10e+08	Devlin et al. (2019)
FacebookAI/roberta-base	1.25e+08	Liu et al. (2019)
dbmdz/bert-base-italian-uncased	1.10e+08	Schweter (2020)
dbmdz/bert-base-italian-cased	1.10e+08	Schweter (2020)
Musixmatch/umberto-wikipedia-uncased-v1	1.11e+08	Parisi et al. (2020)
Musixmatch/umberto-commoncrawl-cased-v1	1.11e+08	Parisi et al. (2020)
google-bert/bert-base-multilingual-uncased	1.67e+08	Devlin et al. (2019)
google-bert/bert-base-multilingual-cased	1.78e+08	Devlin et al. (2019)
FacebookAI/xlm-roberta-base	2.78e+08	Conneau et al. (2020)

Table 16: The names of the LMs used from the HuggingFace Hub and their size in terms of number of parameters.

W–M	TRAIN	DEV	TEST
bert-base-italian-uncased	1.00	0.96	0.87
bert-base-italian-cased	0.98	1.00	0.83
umberto-wikipedia-uncased-v1	0.97	0.92	0.92
umberto-commoncrawl-cased-v1	0.99	0.96	0.92
bert-base-multilingual-uncased	0.99	0.86	0.70
bert-base-multilingual-cased	0.97	1.00	0.76
xlm-roberta-base	0.90	0.96	0.80
K–T			
bert-base-italian-uncased	0.72	0.47	0.47
bert-base-italian-cased	0.96	0.48	0.60
umberto-wikipedia-uncased-v1	0.46	0.47	0.47
umberto-commoncrawl-cased-v1	0.46	0.47	0.47
bert-base-multilingual-uncased	0.52	0.47	0.47
bert-base-multilingual-cased	0.47	0.47	0.47
xlm-roberta-base	0.47	0.47	0.47

Table 17: The performance of the LMs in terms of macro F_1 for the LMs fine-tuned with wHA-IT.

W–M	TRAIN	DEV	TEST
bert-base-uncased	0.99	0.80	0.84
bert-base-cased	0.97	0.83	0.84
roberta-base	0.99	0.85	0.86
bert-base-multilingual-uncased	0.83	0.78	0.85
bert-base-multilingual-cased	0.85	0.79	0.82
xlm-roberta-base	0.82	0.75	0.76
K–T			
bert-base-uncased	0.99	0.84	0.82
bert-base-cased	0.98	0.76	0.79
roberta-base	0.96	0.78	0.81
bert-base-multilingual-uncased	0.92	0.81	0.72
bert-base-multilingual-cased	0.94	0.70	0.81
xlm-roberta-base	0.89	0.72	0.78

Table 18: The performance of the LMs in terms of macro F_1 for the LMs fine-tuned with wHA-EN.

A.5 Attributions

girl	0.111460	men	-0.030749
girls	0.104457	guy	-0.021744
your	0.072209	him	-0.015896
Girls	0.045587	boy	-0.015180
you	0.043417	man	-0.012263
she	0.029879	boys	-0.008853
!	0.029790	He	-0.007704
women	0.029614	he	-0.007353
You	0.024623	guys	-0.006787
her	0.023684	his	-0.004472
makeup	0.023520	male	-0.004046
Girl	0.022483	gentleman	-0.003008
school	0.020280	kid	-0.002303
</s>	0.019971	Guy	-0.002108
Make	0.019166	Men	-0.001930
pretty	0.018957	partner	-0.001316
it	0.017495	teenager	-0.001207
the	0.017092	Boy	-0.001193
.	0.015737	professional	-0.001101
pink	0.015242	nerd	-0.001016
skirts	0.015235	into	-0.000949
skirt	0.014393	ologne	-0.000911
,	0.014049	Ever	-0.000837
dress	0.012912	Male	-0.000826
a	0.012606	penis	-0.000816
yourself	0.012183	geek	-0.000739
dresses	0.011823	dude	-0.000683
She	0.011325	handsome	-0.000655
It	0.011283	masculine	-0.000611
them	0.011071	date	-0.000570
make	0.010421	Gu	-0.000564
If	0.010115	bar	-0.000510
that	0.009836	kitchen	-0.000495
some	0.009506	grown	-0.000492
This	0.009213	puberty	-0.000455
beautiful	0.009158	ican	-0.000454
all	0.008865	off	-0.000420
want	0.008753	dating	-0.000419
this	0.008707	between	-0.000415
Your	0.008293	himself	-0.000411

Table 19: wHA-EN, W-M, `roberta-base` (TRAIN 0.99, DEV 0.85, TEST 0.86)

kids	0.058039	[SEP]	-0.579180
[CLS]	0.039831	.	-0.014464
kid	0.028616	,	-0.010539
##n	0.010980	teen	-0.010233
##wee	0.010135	the	-0.008755
children	0.009953	and	-0.008112
middle	0.008926	are	-0.007597
school	0.006676	if	-0.007460
toys	0.006589	’	-0.006806
people	0.005054	a	-0.006057
pre	0.003502	your	-0.005827
mom	0.003131	?	-0.005743
use	0.003089	is	-0.005117
t	0.002970	teenage	-0.004947
##s	0.002949	for	-0.004597
child	0.002645	you	-0.004572
toy	0.002466	in	-0.004447
/	0.002186	as	-0.004394
time	0.002168	up	-0.004158
animals	0.002096	from	-0.003803
young	0.002094	teens	-0.003675
they	0.001994	at	-0.003446
##ns	0.001882	don	-0.003345
learn	0.001882	when	-0.003291
parents	0.001709	an	-0.003243
example	0.001646	)	-0.003235
remember	0.001641	with	-0.003219
age	0.001640	over	-0.003079
movie	0.001559	will	-0.003077
might	0.001540	good	-0.003067
how	0.001527	to	-0.002957
music	0.001497	can	-0.002774
playing	0.001472	about	-0.002575
food	0.001468	have	-0.002426
dad	0.001454	out	-0.002346
guys	0.001431	all	-0.002181
little	0.001426	get	-0.002179
old	0.001424	just	-0.002161
girls	0.001391	(	-0.002052
light	0.001363	teenagers	-0.001999

Table 20: wHA-EN, K-T, `bert-base-uncased` (TRAIN 0.99, DEV 0.84, TEST 0.82)

[CLS]	0.096002	[SEP]	-0.340334
ragazze	0.063860	amici	-0.030969
.	0.052733	uomini	-0.024488
Se	0.037144	stesso	-0.023203
donne	0.035291	ragazzo	-0.020641
ragazza	0.033087	uomo	-0.017189
una	0.026854	amico	-0.015206
sicura	0.024647	pronto	-0.014043
##ta	0.019987	sicuro	-0.011919
Non	0.019283	modo	-0.011276
la	0.019149	quello	-0.010895
amica	0.019090	soggetto	-0.009081
amiche	0.018528	stanco	-0.008884
:	0.018383	##to	-0.007624
le	0.017927	articolo	-0.007204
Fai	0.017582	uno	-0.007169
stessa	0.017464	all	-0.006952
tutte	0.016846	più	-0.006778
donna	0.016164	comodo	-0.006113
Puoi	0.015974	questo	-0.006060
Scegli	0.015732	orgoglioso	-0.005598
tua	0.015248	fortunato	-0.005424
Per	0.015175	preoccupato	-0.005362
Le	0.015159	##ato	-0.005178
!	0.014648	costretto	-0.005104
delle	0.013800	gli	-0.004869
di	0.013678	stessi	-0.004718
Una	0.012642	senti	-0.004656
La	0.012504	##ro	-0.004652
Cerca	0.012358	##ino	-0.004572
)	0.012122	invitato	-0.004422
della	0.012100	riuscito	-0.004247
essere	0.012071	##vo	-0.004244
Prova	0.011706	bloccato	-0.004043
persone	0.011059	##tatore	-0.004016
##te	0.010820	##mo	-0.004009
,	0.010180	cui	-0.003994
per	0.010145	##gro	-0.003917
ogni	0.009495	sveglio	-0.003636
tue	0.009448	##gato	-0.003608

Table 21: wHA-IT, W-M, `bert-base-italian-cased` (TRAIN 0.98, DEV 1.00, TEST 0.83)

[CLS]	0.245312	[SEP]	-0.165349
:	0.080714	.	-0.071660
!	0.053150	,	-0.047196
in	0.035297	;	-0.043113
da	0.027325	”	-0.041416
?	0.027242	’	-0.036960
a	0.026157	’	-0.030604
Se	0.024843	non	-0.016650
il	0.022034	articolo	-0.015161
del	0.018384	Non	-0.014897
per	0.016414	è	-0.014392
o	0.015331	le	-0.012302
prima	0.015157	troppo	-0.010484
dei	0.015040	sono	-0.010212
giorno	0.013294	–	-0.009822
un	0.012921	capelli	-0.008484
al	0.012534	bella	-0.006242
l	0.012455	di	-0.005976
##re	0.012187	elegante	-0.005837
dopo	0.011774	(	-0.005455
con	0.011254	colore	-0.005248
della	0.011226	si	-0.005077
Puoi	0.010995	Scopri	-0.004918
questo	0.010358	look	-0.004853
i	0.009834	Una	-0.004806
)	0.009756	può	-0.004751
cosa	0.009436	vesti	-0.004506
qualcosa	0.009397	una	-0.004433
##rlo	0.009286	odore	-0.004418
Dopo	0.009121	stile	-0.004360
e	0.008924	Le	-0.004137
lavoro	0.008719	colori	-0.003829
Assicurati	0.008706	Un	-0.003818
andare	0.008485	agio	-0.003690
##ndo	0.008206	ma	-0.003689
perché	0.008186	“	-0.003666
quando	0.008076	La	-0.003652
vuoi	0.008004	profumo	-0.003592
su	0.007878	carina	-0.003561
te	0.007875	tue	-0.003428

Table 22: wHA-IT, K-T, `bert-base-italian-cased` (TRAIN 0.96, DEV 0.48, TEST 0.60)