New Datasets for Automatic Detection of Textual Entailment and of Contradictions between Sentences in French

Maximos Skandalis, Richard Moot, Christian Retoré, Simon Robillard

New Datasets for Automatic Detection of Textual Entailment and of Contradictions between Sentences in French

Maximos Skandalis, Richard Moot, Christian Retoré, Simon Robillard
LIRMM, CNRS, University of Montpellier
34095 Montpellier, France
{maximos.skandalis, richard.moot, christian.retore, simon.robillard}@lirmm.fr

Abstract

This paper introduces DACCORD, an original dataset in French for automatic detection of contradictions between sentences. It also presents new, manually translated versions of two datasets, namely the well known dataset RTE3 and the recent dataset GQNLI, from English to French, for the task of natural language inference / recognising textual entailment, which is a sentence-pair classification task. These datasets help increase the admittedly limited number of datasets in French available for these tasks. DACCORD consists of 1034 pairs of sentences and is the first dataset exclusively dedicated to this task and covering among others the topic of the Russian invasion in Ukraine. RTE3-FR contains 800 examples for each of its validation and test subsets, while GQNLI-FR is composed of 300 pairs of sentences and focuses specifically on the use of generalised quantifiers. Our experiments on these datasets show that they are more challenging than the two already existing datasets for the mainstream NLI task in French (XNLI, FraCaS). For languages other than English, most deep learning models for NLI tasks currently have only XNLI available as a training set. Additional datasets, such as ours for French, could permit different training and evaluation strategies, producing more robust results and reducing the inevitable biases present in any single dataset.

Keywords: datasets, recognising textual entailment, natural language inference, contradictions, sentence-pair classification tasks, French

1. Introduction

In Natural Language Processing (NLP), the Natural Language Inference (NLI) task, also known as Recognising Textual Entailment (RTE) is a sentence-pair classification task, with either three (entailment/0, neutral/unknown/1, contradiction/2) or two (entailment/0, no entailment/1) classes or labels. The corresponding label is attributed to each sentence pair according to whether the second sentence of the pair (usually called hypothesis) is a consequence of (entailment) or contradicts[1] the first sentence (premise) of the pair, or neither of the two (neutral).

In the previous 3-label case, the automatic detection of contradictions constitutes a part of the task, as far as predicting correctly the label of contradiction is concerned. The level of representation of contradicting pairs in NLI datasets varies from only 9% in FraCaS dataset (Cooper et al., 1996) to, at most, one third in other datasets like XNLI (Conneau et al., 2018).

On the other hand, the automatic detection of contradictions between sentences as an independent task is a binary sentence-pair classification task with two labels (contradiction/1, no contradiction/0). The interest in the latter one lies in the fact that the relation of contradiction is symmetric (which is not the case for entailment). It is also a more fundamental one. In fact-checking, for instance, we often find statements which cannot be directly inferred from other textual information, but which are nonetheless true.[2] On the contrary, for a piece of information to be true, it cannot contradict other confirmed or verified information.

On this ground, we consider that the three datasets introduced with the present article (2-class DACCORD, 3-class RTE3-FR and GQNLI-FR), and in particular the one dedicated to detecting contradictions (DACCORD), can be particularly useful and can have their place also in the less strictly defined and more diverse[3] task of misinformation

This work was supported and authorised by AID, French Defence Innovation Agency, and by ICO, Institut Cybersécurité Occitanie.

All the datasets introduced in this paper are available on `github` and on `huggingface`.

[1] The most common definition of contradiction in NLP, although not without its flaws, is that two statements are contradictory if it is impossible for them to be both true at the same time. We personally prefer the formulation "in the same situation" (i.e. in the same model in logic).

[2] For example, the sentences "the killer murdered 3 people" and "the killer was holding a knife" can both be true, even though neither phrase entails the other.

[3] Misinformation or fake news detection can be tackled in many different ways, more or less subjective, for example by detecting radical or abusive language in texts, by analysing reactions and comments replying to tweets, by detecting rumors (Gorrell et al., 2019) or propaganda techniques (Da San Martino et al., 2020), or by caracterising sources as reliable or unreliable (Guibon et al., 2019). Contradiction detection is, in our opinion, one of the most objective and neutral ways (if not the most) to

detection.

However, the interest in the aforementioned task is not limited solely to the detection of fake news. Being able to automatically detect contradictions and inferences in textual data is a crucial question of natural language understanding (NLU) and of automatic reasoning on natural language.

In summary, after an overview of the already available datasets for NLI in Section 2, this article introduces:

- DACCORD (Skandalis et al., 2023), a new dataset on the automatic detection of contradictions between sentences in French, shifting, for this time, the focus from the traditionally studied pair entailment / not entailment to the pair contradiction / not contradiction;
- RTE3-FR and GQNLI-FR, which are French manual translations of the original English datasets with the same names (RTE3 and GQNLI), for the task of Natural Language Inference;
- at the margin, a machine translation into 9 languages[4] of the linguist-in-the-loop (LitL) and LitL Chat protocols of the LingNLI dataset (Parrish et al., 2021), as well as a machine translation from English into French and Modern Greek of SICK dataset (Marelli et al., 2014).

The introduction of these new datasets is motivated by the fact that, at the time when the use of Transformers has become dominant within the NLP community, a great need for a large amount of datasets has emerged. Despite this, there is a clear lack of datasets for French for the two tasks concerned.

We then present in the Section 4 an evaluation of different Transformers on these three datasets, giving for comparison the results obtained on the already available XNLI dataset in French and, for RTE3 and GQNLI, on the original English versions of these datasets.

2. Related Work

2.1. Related Work in English

Numerous datasets exist in English for the task of NLI, namely FraCaS (Cooper et al., 1996), RTE (Dagan et al., 2006) (Dzikovska et al., 2013), SICK (Marelli et al., 2014), SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), XNLI (Conneau et al., 2018), ANLI (Nie et al., 2020), LingNLI (Parrish et al., 2021), GQNLI (Cui et al., 2022), WANLI (Liu et al., 2022), the GLUE benchmark (Wang et al., 2018)[5], SuperGLUE (Wang et al., 2019)[6].

FraCaS contains 346 inference problems to be tackled (see 2.2 for more details, as it is one of the few datasets also available in French). SICK, meanwhile, contains 9840 examples, 14,47% of which are contradictions and 28,67% entailments. Their annotation was done by *crowdsourcing* and they are simplified sentences in English, similar to those in FraCaS.

For RTE, there are eight RTE datasets released as of now. RTE1 (Dagan et al., 2006) and RTE2 (Haim et al., 2006) were, at the time of their initial release, annotated only as a 2-way task (entailment, no entailment). RTE3 (Giampiccolo et al., 2007) was the first one in the series also proposed as a 3-way task (Entailment, Unknown, Contradiction), instead of only as a 2-way task (Entailment, No entailment) until then. RTE3 was annotated for this purpose in a 3-way by de Marneffe et al. (2008), who then also annotated the first 2 RTE datasets in a 3-way.[7][8] RTE4 (Giampiccolo et al., 2008) is the first one in the series annotated from the beginning as a 3-way task. This, and RTE5 (Bentivogli et al., 2009b), RTE6 (Bentivogli et al., 2009a), RTE7 (Bentivogli et al., 2011) are only freely available upon request. Finally, RTE8 (Dzikovska et al., 2013) was part of SemEval 2013.

Table 1 shows the partition of examples annotated as entailment, unknown, and contradiction, for validation and test subsets of each of the first three RTE datasets, which are publicly available.

MultiNLI, modeled on the SNLI corpus but covering a wider range of genres of spoken and written text than the latter one (MultiNLI for Multi-Genre NLI), is currently the largest NLI dataset in English,

tackle it.

[4]Bulgarian, Finnish, French, Greek, Italian, Korean, Lithuanian, Portuguese, Spanish.

[5]GLUE is just a collection of already existing NLI datasets, like MultiNLI, for the purpose of forming a benchmark for this task. Unfortunately, in the GLUE version of RTE, RTE datasets are divided into two classes (inference, no inference) and not into three (inference, contradiction, neutral), thus we do not privilege this version here.

[6]SuperGLUE is styled after GLUE. While the version of RTE found in SuperGLUE is, as in GLUE, a 2-label version, SuperGLUE also contains some 556 3-way labelled examples of the 1200 ones in total in the CommitmentBank corpus (de Marneffe et al., 2019), which focuses on embedded clauses and which was initially annotated using a 7-point Likert scale labelled at 3 points.

[7]The instructions on how they divided the "no-entailment" class in 2 classes, "contradiction" and "neutral", are given at https://nlp.stanford.edu/RTE3-pilot/contradictions.pdf. Looking at them, it can be seen that RTE datasets have a more probabilistic definition of the notion of contradiction.

[8]The 3-way annotated versions of RTE1, RTE2 and RTE3 are available at https://nlp.stanford.edu/projects/contradiction.

Dataset		Entailment	Unknown	Contradiction
RTE1	dev_1	142	97	48
	dev_2	140	85	55
	test	400	251	149
RTE2	dev	400	289	111
	test	400	296	104
RTE3	dev	412	308	80
	test	410	318	72

Table 1: Dataset breakdown by label for original RTE1, RTE2, RTE3

consisting of some 433 thousand sentence pairs annotated. XNLI is a subset of examples from MultiNLI that have been translated (by machine for its training set, manually for the development and test sets) into 14 different languages. These two are currently the largest and most used (especially for training) datasets in English.

Nie et al. (2020) transformed the English dataset FEVER, which was initially constructed by Thorne et al. (2018) with Wikipedia phrases modified by annotators and compared to the Introduction sections of a list of Wikipedia pages for a task of automatic fack-checking, to a dataset for textual inference (thus with a label "entailment", "contradiction" or "neutral" for each premise-hypothesis pair for this version).

GQNLI focuses on the use of generalised quantifiers like "more than", "less than", "half of", "some", "most", and consists of 30 premises and 300 hypotheses. Its premises were sampled from MultiNLI and ANLI.

Finally, when it comes to the application domain, some datasets attempt to deal specifically with the problem of misinformation with classification tasks influenced by NLI but going further than this. PHEME (Kochkina et al., 2018), for instance, has Twitter messages first classified by a journalist as "rumours" or "non-rumours", then those that are rumours as "true", "false" or "unverified". In addition, responses to these tweets were annotated as questioning, supporting, denying or commenting on the rumour.

2.2. Related Work in French

As already mentioned, there exists a recent French translation of FraCaS (Amblard et al., 2020). FraCaS classifies its examples in the following categories: Generalised Quantifiers, Plurals, (Nominal) Anaphora, Ellipsis, Adjectives, Comparatives, Temporal reference, Verbs, and Attitudes. The dataset is available in the form of question-answering and of premise(s)-hypothesis pairs. Contradictions make up 9% of the 346 inference problems of the entire FraCaS corpus, entailments form 52% of the corpus and neutral cases 27%[9].

[9]The remaining 12% of the examples require a more detailed answer.

Among the languages included, XNLI (Conneau et al., 2018) contains manual translations from English to French of the validation and test subsets of the original MultiNLI. In particular, its validation subset is composed of 2490 pairs of sentences (830 contradictions, 830 neutral cases, 830 inferences), while the test subset contains 5010 pairs (1670 contradictions, 1670 neutral cases, 1670 inferences). On the other hand, the training subset of XNLI in French (i.e. more than 98% of the sentence pairs of XNLI) is a mere machine translation of MNLI from English to French (by Conneau et al. (2018) and by Hu et al. (2020) independently).

Finally, MM-COVID (Li et al., 2020) is a multimodal and multi-lingual dataset, which also includes examples in French, all taken from social media messages. In terms of its French-language content, it includes 2821 tweets described as 'false' (and 4459 replies to them), compared with 166 tweets described as 'true' (and 5095 replies).

3. DACCORD, RTE3-FR, GQNLI-FR

3.1. Methodology

3.1.1. Construction of DACCORD

DACCORD (Skandalis et al., 2023) was constructed from articles on the `factuel.afp.com` website. AFP (Agence France-Presse) Factuel is a French fact-checking site. The sentence pairs were manually selected by carefully reading the articles on the aforementioned site and keeping the sentences of interest for the task under study. We paired the collected sentences either with each other or with sentences constructed from the collected sentences, so that the resulting pair satisfied the assigned label.

The dataset currently covers three topics: the Russian invasion of Ukraine, the Covid-19 pandemic and the climate crisis. As far as we know, this is the very first NLP dataset covering the conflict between Russia and Ukraine.

DACCORD dataset can also be used for misinformation detection due to the topics covered. We have chosen AFP Factuel as our source for collecting sentences for the dataset, but we do not take any position on the truth of the statements chosen[10], as the topics dealt with are delicate and the notion of truth is, formally speaking, not trivial. That said, it is indicated in the corpus when a pair of sentences forms a contradiction or when the two sentences of an example are inter-compatible (non-

[10]Instructing (adult) people what to choose to believe can be viewed as intrusive by many. What's more, our perception of truth and knowledge of facts are not always complete, and can thus change over time. Note, finally, that contradiction can arise even between two false sentences.

contradictory), but it is not indicated which of the statements (if any) is true and which is not.

DACCORD dataset is composed of 1034 sentence pairs, of which 515 (49.18%) form contradictions. Of these 1034 pairs, 472 (214 contradictions) were collected from 106 articles on the Russian-Ukrainian war, dating from 24 February 2022 to 3 November 2022 (included). 450 pairs (252 contradictions) were selected from 164 articles published between 20 October 2021 and 23 November 2022 on the Covid-19 pandemic. Finally, the 112 pairs of sentences (49 contradictions) on global warming were taken from 33 articles published between 16 July 2021 and 24 October 2022.

It is available in both XML format and in TSV format, for greater ease of use and interoperability with FraCaS and RTE-EN (XML), and XNLI (TSV).

3.1.2. Translation of RTE3 and GQNLI

RTE3-FR and GQNLI-FR were manually translated from English to French by the authors of the present article.

For RTE3-FR, we also relied on the available German and Italian translations, to solve eventual ambiguities or to make choices on translation preferences. The fact that RTE3 is the latest RTE dataset freely available online (see 2.1) and that there are German and Italian translations avalaible, allowing for a combination resulting in a multi-lingual dataset or benchmark, was the reason for choosing RTE3 for a translation in French. This means that for RTE3, we now have at our disposal, along with the original English dataset, its German and Italian translations of 2013 and 2012 respectively, a multilingual 3-label[11] version in 4 major languages, also available on `huggingface.co`.

Both RTE3-FR and GQNLI-FR are available in TSV format.

3.2. Quantitative Analysis

Table 2 displays the number of sentence pairs per topic and per label for DACCORD.

Dataset		Contradiction	Compatible
DACCORD	Russian invasion	215	257
	Covid-19	251	199
	Climate change	49	63

Table 2: DACCORD breakdown by topic and label

Table 3 shows the number of sentence pairs per subset and per label for RTE3-FR and GQNLI-FR.

[11] Initial German and Italian releases of RTE3 were annotated with two labels (entailment, no entailment); we transformed them also to include three labels, all while respecting the annotation choices made in these two releases (which might sometimes be different from ours for RTE3-FR, see Appendix).

Dataset		Entailment	Neutral	Contradiction
RTE3-FR	dev	412	299	89
	test	410	318	72
GQNLI-FR	test	97	100	103

Table 3: Breakdown by label for RTE3 and GQNLI French translations

Table 4 gives details on the number of *tokens* per sentence per topic/subset (including punctuation), according to the NLTK tokeniser (Bird et al., 2009), for DACCORD, RTE3 (all languages), GQNLI-FR, as well as for XNLI and FraCaS, for comparison. According to it, DACCORD has on average the longest hypotheses, while the validation subset of RTE3-FR has the longest premises, on average.

Dataset		Total number of tokens of premises	Mean number of tokens per premise	Total number of tokens of hypotheses	Mean number of tokens per hypothesis
DACCORD	Rus-Ukr war	16455	34.86	12734	26.98
	Covid-19	13631	30.29	10014	22.25
	Climate	5614	50.13	4884	43.61
RTE3-EN	dev	30980	38.73	7577	9.47
	test	26438	33.05	7083	8.85
RTE3-FR	dev	34910	43.64	8761	10.95
	test	29778	37.22	8172	10.22
RTE3-ITA	dev	33262	41.58	8389	10.49
	test	27872	34.84	7634	9.54
RTE3-DE	dev	31703	39.63	7611	9.51
	test	26441	33.05	6957	8.70
GQNLI-FR	test	6450	21.57	3239	10.83
XNLI (test and val)		169092	22.55	87485	11.66
FraCaS		4805	9.13	3245	9.49

Table 4: Number of tokens for for all now available versions of RTE3

Finally, machine translations for two of the three subsets of LingNLI dataset ("LitL" and "LotS") were performed in March/April 2023 using the latest open-source neural machine translation OPUS-MT big models available for the respective languages. The LitL subset contains, for each language, 14995 and 2425 pairs for training and validation respectively, whereas the LitL Chat (or LotS) subset contains 14990 and 2468 pairs, respectively. The machine translations of SICK dataset from English into French and Modern Greek were carried out in November 2023 with the corresponding monolingual NMT models mentioned above. In all of the 9 languages[4] in which we machine-translated LingNLI (which include French and Modern Greek into which we also machine-translated SICK dataset), we had, until now, only the machine-translated training subset of XNLI available for use in training deep learning models. We think that differentiating in this way the resources available for training models, even if machine-translated, could help reduce potential biases during training, for example in the definition of contradictions, which might differ between different datasets (recall from Section 2.1 the more probabilistic definition of contradiction in the RTE annotation instructions).

3.3. Qualitative Analysis

The annex features a lot of examples taken from RTE3-FR and GQNLI-FR, hence this part will mostly cover examples from DACCORD, accompanied by English translations for broader comprehensibility.

The examples b048 and b049 are taken from the subset of DACCORD on Covid-19 and exhibit a case of contradiction and of non-contradiction based on a logical implication.

(b048) P-FR: Si l'ACIP vote en faveur de l'ajout du vaccin anti-Covid au calendrier vaccinal des enfants, les écoliers devront se faire vacciner pour être scolarisés dans une école publique aux États-Unis. Le 20 octobre, l'ACIP a approuvé à l'unanimité l'ajout de la vaccination contre le Covid-19 au calendrier des vaccinations recommandées.
H-FR: Les écoliers devront se faire vacciner pour être scolarisés dans une école publique aux États-Unis.
P-EN: If ACIP votes in favour of adding the Covid vaccine to the childhood immunisation schedule, schoolchildren will have to be vaccinated in order to attend a public school in the United States. On 20 October, ACIP unanimously approved the addition of the Covid-19 vaccine to the recommended vaccination schedule.
H-EN: Schoolchildren will have to be vaccinated to attend public schools in the United States.
Label: Compatibles

(b049) P-FR: Si l'ACIP vote en faveur de l'ajout du vaccin anti-Covid au calendrier vaccinal des enfants, les écoliers devront se faire vacciner pour être scolarisés dans une école publique aux États-Unis. Le 20 octobre, l'ACIP a approuvé à l'unanimité l'ajout de la vaccination contre le Covid-19 au calendrier des vaccinations recommandées.
H-FR: Les écoliers pourront être scolarisés dans une école publique aux États-Unis sans se faire vacciner.
P-EN: If ACIP votes in favour of adding the Covid vaccine to the childhood immunisation schedule, schoolchildren will have to be vaccinated in order to attend a public school in the United States. On 20 October, ACIP unanimously approved the addition of the Covid-19 vaccine to the recommended vaccination schedule.
H-EN: Schoolchildren will be able to attend public schools in the United States without being vaccinated.
Label: Contradiction

The corpus contains real-world sentences, which means that some examples require the assumption of hidden premises for the contradiction. This is of particular importance should symbolic or logical approaches be attempted on the dataset. The example a097 from the subset of DACCORD on Russo-Ukrainian war gives two contradictory sentences. The tills in a shop are in use and produce a non-zero turnover when the shop is open and, surely, not empty. The contradiction here does not simply arise from the introduction of a negation in one of the two sentences. This is a case that is at the present difficult for a machine to detect, and it also illustrates the importance of semantic analysis of sentences for this task.

(a097) P-FR: 58 caisses étaient en service dans le centre commercial Amstor le jour de l'attaque, enregistrant ce jour-là un chiffre d'affaires de 2,9 millions de hryvnia ukrainiennes, soit environ 97.000 euros. Des employés du centre commercial, blessés le 27 juin, ont témoigné auprès de l'AFP après l'attaque.
H-FR: Amstor était fermé et vide au moment des frappes par les missiles russes.
P-EN: 58 checkouts were in operation in the Amstor shopping centre on the day of the attack, recording sales of 2.9 million Ukrainian hryvnia that day, equivalent to around €97,000. Employees of the shopping centre, who were injured on 27 June, spoke to AFP after the attack.
H-EN: Amstor was closed and empty at the time of the Russian missile strikes.
Label: Contradiction

Finally, the pair b398 given below could be considered as a meta-referential example of false information.

(b398) P-FR: Interrogée par l'AFP, l'Autorité régionale de santé (ARS) de Guadeloupe déplore une fausse information circulant et précise que ce n'est jamais elle qui passe les commandes de médicaments.
H-FR: C'est une fausse information que ce n'est pas l'Autorité régionale de

santé (ARS) de Guadeloupe qui passe les commandes des médicaments.

P-EN: When questioned by AFP, the Guadeloupe Regional Health Authority (ARS) deplored the false information circulating and stated that it is never the ARS that places the orders for the drugs.

H-EN: It is false information that it is not the Guadeloupe Regional Health Authority (ARS) that places the orders for the drugs.

Label: Contradiction

When it comes to RTE3, the original RTE3 dataset was collected by human annotators and consists of four subsets which correspond to different application settings from which the pair has been generated: Information Extraction (IE), Information Retrieval (IR)[12], Question Answering (QA), and Multi-Document Summarization (SUM). The premise may be made up of one or more sentences while the hypothesis usually contains one simple sentence.

Contrariwise, the sentences in c023 are one example from the subset on climate change of DACCORD dataset where the hypothesis is longer than the premise.

(c023) P-FR: D'après le dernier rapport du GIEC, les océans pourraient encore gagner plusieurs centimètres d'ici à la fin du siècle, selon les scénarios envisagés.

H-FR: Le "résumé technique" indique qu'à horizon 2100, le niveau global de la mer est projeté à une augmentation de 28 à 55 cm, dans le cas d'une réduction significative des émissions de gaz à effet de serre. Une augmentation de 60 cm à 1 m pourrait advenir dans le cas de figure le plus catastrophique.

P-EN: According to the latest IPCC report, the oceans could rise by several centimetres by the end of the century, depending on the scenarios considered.

H-EN: The "technical summary" indicates that by 2100, global sea level is projected to rise by between 28 and 55 cm, assuming a significant reduction in greenhouse gas emissions. A rise of 60 cm to 1 m could occur in the most catastrophic scenario.

Label: Compatibles

For GQNLI (Cui et al., 2022), Table 5 displays the way in which we translated determiners, which is in accordance with the translation choices also made by Amblard et al. (2020). Note that syntactic differences can arise from one language version to another: e.g. in English, *most* is a determiner, whereas French expresses the same concept through the common nouns *la plupart de* or *la majorité de*.

each	chaque
every	tout/tous les
most	la plupart de
few	peu de
a few	quelques
neither	aucun des deux
many	beaucoup
no	aucun
some+singular	un
some+plural	certains
several	plusieurs

Table 5: Translation of determiners for GQNLI-FR

4. Experiments and Results

4.1. Test Protocol

In order to evaluate the performance of the state-of-the-art models on these three new datasets for French, we have chosen to use deep learning models based on the transformer architecture (Vaswani et al., 2017). The models selected for evaluation on DACCORD, RT3-FR and GQNLI-FR are DistilmBERT (Sanh et al., 2019), XLM-R (Conneau et al., 2020), mDeBERTa-v3 (He et al., 2021), and CamemBERT (Martin et al., 2020). They are all either partially (DistilmBERT, XLM-R, and mDeBERTa) or entirely (CamemBERT) trained on French data. For their evaluation, we used versions adjusted on XNLI, available on huggingface.co.

4.2. Results

Table 6 presents the evaluation (measuring accuracy and F1 score) of four multilingual models on all now available versions of RTE3 (English/original, Italian, German, French/ours) on a 3-label NLI task. Table 7 reports the same metrics on the same NLI task in English and French, also on GQNLI and including French monolingual models this time, for the French versions of the datasets under study. They both show that results obtained on the French versions of the dataset examined are consistent with the results obtained on all other language versions of the two datasets, with the same deep learning models.

[12]RTE3 also contains some examples where a premise is compared to a title (examples 257, 280, 332 of the RTE3 dev set).

Modèles	RTE3-EN		RTE3-FR		RTE-DE		RTE-IT	
	Accuracy	F1 Score	*Accuracy*	F1 Score	*Accuracy*	F1 Score	*Accuracy*	F1 Score
DistilmBERT$_{\text{Base-cased}}$	60,75	47,92	61,13	46,65	58,00	46,15	59,75	46,20
mDeBERTa-v3$_{\text{Base, XNLI}}$	67,13	56,26	67,13	55,01	65,63	52,30	66,50	55,16
mDeBERTa-v3$_{\text{Base, NLI-2mil7}}$	71,25	61,33	69,63	60,57	69,75	59,06	71,00	61,61
XLM-R$_{\text{Large}}$	**72,88**	**63,62**	**71,25**	**62,47**	**71,13**	**60,22**	**73,38**	**64,22**

Table 6: Label prediction results by transformers in all available languages for RTE3

Modèles	RTE3-EN		RTE3-FR		GQNLI		GQNLI-FR	
	Accuracy	F1	*Accuracy*	F1	*Accuracy*	F1	*Accuracy*	F1
DistilmBERT$_{\text{Base-cased}}$	60,75	47,92	61,13	46,65	26,00	26,03	27,67	26,88
XLM-R$_{\text{Base}}$	-	-	60,50	49,61	-	-	31,67	31,46
CamemBERT$_{\text{Base, 3-class}}$	-	-	63,13	51,52	-	-	33,67	33,44
mDeBERTa-v3$_{\text{Base, XNLI}}$	67,13	56,26	67,13	55,01	28,33	27,73	28,67	27,94
mDeBERTa-v3$_{\text{Base, NLI-2mil7}}$	71,25	61,33	69,63	60,57	**36,67**	**37,04**	**38,33**	**38,58**
XLM-R$_{\text{Large}}$	**72,88**	**63,62**	**71,25**	**62,47**	35,33	35,02	36,34	35,94
CamemBERT$_{\text{Large, 3-class}}$	-	-	71,13	61,97	-	-	33,33	31,62

Table 7: Label prediction results by transformers on RTE3-FR and GQNLI-FR

Since our research project focuses in particular on the automatic detection of contradictions (for the reasons explained in Section 1), we calculated the accuracy and the F1 score[13] on the probability predicted by the models that the "contradiction" label is true, even though the models were trained on XNLI to be able to predict one of three labels (entailment, neutral, contradiction). Table 8 gives the results of this evaluation. Table 9 depicts, for the same F1 score of Table 8 on our datasets, the detailed precision and recall scores.

4.3. Discussion

Looking at the results, we note, firstly, a progressive and constant evolution in the performance of the multi-lingual models even on uni-lingual tasks (for example, mDeBERTa as opposed to DistilmBERT).

Furthermore, the models trained on XNLI show, without exception, inferior performance on all three new datasets than on XNLI. This comes as no surprise, since both DACCORD and GQNLI have been designed so as to test the capabilities of existing models.

For the record, all the models evaluated in the article failed to detect the contradiction in the example b398. The label for the example a097 was not correctly predicted by DistilmBERT and CamemBERT$_{\text{Large, 3-class}}$.

Another point is that we can observe in Table 8 some consistency between the performance on XNLI of the models studied and their performance on DACCORD or even RTE3-FR and GQNLI-FR (only for accuracy for the latter one), the two best models for XNLI (XLM-R and CamemBERT) also being the best models for DACCORD and RTE3-FR, even if the results remain inferior to those obtained on XNLI.

We should, nonetheless, keep in mind that these models are almost all (with the exception of mDeBERTa-v3$_{\text{Base, NLI-2mil7}}$, which is partially trained on XNLI and on other machine-translated sources) trained on the training subset of XNLI, as it is the only set large enough to train neural models. Besides, the three datasets introduced (DACCORD, RTE3-FR, GQNLI-FR) possess validation and test subsets, but not training subsets. This does not prevent someone from using a part or the whole of these datasets as training input, if the trained model is meant to be used for zero-shot classification afterwards, for instance. Adding the machine translations of LingNLI and SICK datasets, made available here, in the training, along with XNLI, could possibly enhance the performance for those models.

Finally, performance on GQNLI is particularly low. We assume that this is due to the fact that deep learning models are trained to predict NLI, but, in reality, what they learn is some kind of semantic similarity between phrases. However, two affirmative sentences, one containing a numeral and the other containing a quantifier contradicting the former numeral, cannot be effectively reduced to computing an embedding or vector similarity.

5. Conclusion and Perspectives

In this article, we presented DACCORD, a new dataset for the task of automatically detecting con-

[13] As a reminder, the F1 measure is the harmonic mean of precision and recall.

Modèles	DACCORD		XNLI		RTE3-FR		GQNLI-FR	
	Accuracy	F1 Score	*Accuracy*	F1 Score	*Accuracy*	F1 Score	*Accuracy*	F1 Score
DistilmBERT$_{\text{Base-cased}}$	63,73	52,59	79,98	68,01	79,63	11,89	51,67	19,89
XLM-R$_{\text{Base}}$	71,57	67,62	87,17	81,14	77,75	21,93	49,33	23,23
CamemBERT$_{\text{Base, 3-class}}$	77,76	76,19	89,64	85,09	80,36	26,29	50,33	**36,05**
mDeBERTa-v3$_{\text{Base, XNLI}}$	80,75	78,30	90,98	86,39	85,75	30,49	52,00	20,88
mDeBERTa-v3$_{\text{Base, NLI-2mil7}}$	80,95	78,47	90,76	85,89	87,00	38,82	50,67	34,51
XLM-R$_{\text{Large}}$	82,01	80,00	**96,49**	**94,74**	86,75	41,11	**53,33**	25,53
CamemBERT$_{\text{Large, 3-class}}$	83,27	81,01	92,30	88,12	**87,63**	**41,42**	52,67	31,07
CamemBERT$_{\text{Large, 2-class}}$	**84,24**	**82,49**	91,70	87,66	85,75	37,36	48,00	19,59

Table 8: Results of detection of contradiction label by Transformers on DACCORD, XNLI, RTE3-FR and GQNLI-FR

Modèles	DACCORD			RTE3-FR			GQNLI-FR		
	Precision	*Recall*	F1	*Precision*	*Recall*	F1	*Precision*	*Recall*	F1
DistilmBERT$_{\text{Base-cased}}$	75,36	40,39	52,59	9,82	15,07	11,89	23,08	17,48	19,89
XLM-R$_{\text{Base}}$	78,12	59,61	67,62	16,13	34,25	21,93	24,21	22,33	23,23
CamemBERT$_{\text{Base, 3-class}}$	81,60	71,46	76,19	20,00	38,36	26,29	**32,31**	**40,78**	**36,05**
mDeBERTa-v3$_{\text{Base, XNLI}}$	89,30	69,71	78,30	27,47	34,25	30,49	24,05	18,45	20,88
mDeBERTa-v3$_{\text{Base, NLI-2mil7}}$	89,75	69,71	78,47	34,02	45,21	38,82	31,71	37,86	34,51
XLM-R$_{\text{Large}}$	89,64	72,23	80,00	34,58	**50,68**	41,11	28,24	23,30	25,53
CamemBERT$_{\text{Large, 3-class}}$	93,18	71,65	81,01	**36,46**	47,95	**41,42**	31,07	31,07	31,07
CamemBERT$_{\text{Large, 2-class}}$	**92,31**	**74,56**	**82,49**	31,20	46,58	37,36	20,88	18,45	19,59

Table 9: Detailed F1 score for detection of contradiction label by transformers on DACCORD, RTE3-FR and GQNLI-FR

tradictory statements in French. It consists of 1034 pairs of sentences, all selected manually, including 515 contradictions. To our knowledge, this is the first corpus in French dedicated exclusively to the contradiction detection task and covering the themes of Covid-19 and the war between Russia and Ukraine. DACCORD is the most complicated dataset for this task given the length of the premises and hypotheses included and the nature of the sources used (press articles rather than social media messages). We also introduced two manual translations from English to French of two other datasets for the NLI task, namely RTE3-FR and GQNLI-FR. Since most deep learning models in languages other than English currently do not have any option to be trained on besides XNLI for the NLI task, the new datasets introduced in the present paper should help towards amplifying the offer of available datasets other than XNLI for French.

When evaluating the models examined, DACCORD proved to be more challenging for them than the already existing XNLI dataset when it comes to detecting the label “contradiction”. Finally, results show that both RTE3-FR and especially GQNLI-FR are more difficult than XNLI for current deep learning models both in the NLI 3-label task, in general, and in the prediction of the contradiction label, in particular.

We now plan to experiment with neuro-symbolic and logical methods on the corpora constructed. We also want to investigate few-shot learning possibilities on them. Finally, we would like to enrich DACCORD with sentences covering other topics not sufficiently incorporated in the currently available datasets.

6. Acknowledgements

The research hereby presented was carried out with the financial support of the French Ministry of Defence - Defence Innovation Agency (AID - DGA), to which we express our gratitude. This work was likewise supported by ICO, Institut Cybersécurité d’Occitanie, funded by Région Occitanie, France, which we would also like to thank.

7. Bibliographical References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O’Reilly Media, Inc.".

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.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio. Association for Computational Linguistics.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Gaël Guibon, Liana Ermakova, Hosni Seffih, Anton Firsov, and Guillaume Le Noé-Bienvenu. 2019. Multilingual Fake News Detection with Satire. In *CICLing: International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

8. Language Resource References

Maxime Amblard, Clément Beysson, Philippe de Groote, Bruno Guillaume, and Sylvain Pogodalla. 2020. A French version of the FraCaS test suite. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5887–5895, Marseille, France. European Language Resources Association.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009a. The sixth PASCAL recognizing textual entailment challenge. In *Text Analysis Conference*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The seventh PASCAL recognizing textual entailment challenge. *Theory and Applications of Categories*.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009b. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*. NIST.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Robin Cooper, Richard Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspars, Hans Kamp, Manfred Pinkal, David Milward, Massimo Poesio, Stephen Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework. Technical report, FraCaS: A Framework for Computational Semantics. FraCaS deliverable D16, 136 pages, also available by anonymous ftp from `ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz`.

Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. 2022. Generalized quantifiers as a source of error in multilingual NLU benchmarks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4875–4893, Seattle, United States. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. 2008. The fourth PASCAL recognizing textual entailment challenge. In *Text Analysis Conference (TAC)*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. MM-COVID: A multilingual and multimodal data repository for combating covid-19 disinformation.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maximos Skandalis, Richard Moot, and Simon Robillard. 2023. DACCORD : un jeu de données pour la détection automatique d'énonCés COntRaDictoires en français. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le*

Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 285–297, Paris, France. ATALA.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Curran Associates Inc., Red Hook, NY, USA.

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

Appendix A. Translation Choices for RTE3-FR

Development Set

Example ID	Premise	Hypothesis	Label	Task	Length	Comment
11	EN: I recently took a round trip from Abuja to Yola, the capital of Adamawa State and back to Abuja, with a fourteen-seater bus. FR: J'ai récemment fait un voyage aller-retour d'Abuja à Yola, la capitale de l'État d'Adamawa et retour à Abuja, avec un bus de quatorze places.	EN: Abuja is located in Adamawa State. FR: Abuja est située dans l'État d'Adamawa.	UNKNOWN	IE	Short	Round trip translates to "voyage aller-retour" in French, so no need for "et retour à Abuja".
26	EN: The Prime Minister of Spain Zapatero visited Brazil, Argentina, Chile and Uruguay recently, in a effort to build a left axis in South America. The cited countries' South American Presidents agreed to collaborate at international level, particularly in the United Nations , European Union and with Paris, Berlin and Madrid. FR: Le Premier ministre espagnol Zapatero s'est rendu récemment au Brésil, en Argentine, au Chili et en Uruguay, dans le but de construire un axe de gauche en Amérique du Sud. Les présidents sud-américains des pays cités ont convenu de collaborer au niveau international, notamment aux Nations unies, l'Union européenne et avec Paris, Berlin et Madrid.	EN: Brazil is part of the United Nations. FR: Le Brésil fait partie des Nations unies.	YES	IE	Long	According to the original label, should Brasil be member to the UN, according to the premise given, it should also be member of the EU.
35	FR: Un important groupe de défense des droits de l'homme a identifié mercredi la Pologne et la Roumanie comme les lieux probables en Europe de l'Est des prisons secrètes où les suspects d'Al-Qaïda sont interrogés par la Central Intelligence Agency.	FR: Les prisons secrètes de la CIA étaient situées en Europe de l'Est.	YES	IE	Short	We chose to keep CIA as Central Intelligence Agency in France, due to its well-known name as it stands.
50	EN: Edison decided to call "his" invention the Kinetoscope, combining the Greek root words "kineto" (movement), and "scopos" ("to view"). FR: Edison décida d'appeler "son" invention le Kinetoscope, combinant les mots racines grecques "kineto" (mobile), et "scopos" ("regarder").	EN: Edison invented the Kinetoscope. FR: Edison a inventé le Kinétoscope.	YES	IE	Short	"Kineto" means "mobile" in Ancient Greek.
107	EN: Since joining the Key to the Cure campaign three years ago, Mercedes-Benz has donated over $2 million toward finding new detection methods, treatments and cures for women's cancers. FR: Depuis qu'elle a rejoint la campagne "Key to the Cure" il y a trois ans, Mercedes-Benz a fait don de plus de 2 millions de dollars pour trouver de nouvelles méthodes de détection, des traitements et des remèdes pour les cancers féminins.	EN: Mercedez-Benz supports the Key to the Cure campaign. FR: Mercedez-Benz soutient la campagne "Key to the Cure".	YES	IE	Short	We have followed German translation practice, and have not translated phrases such as the "Key to the Cure" campaign.
178	EN: Anglo/Dutch Royal Dutch Shell, Total, of France, and Spain's Repsol were all named as examples of established oil companies involved in the oil-for-food programme before surcharges began in 2001. FR: L'anglo-néerlandaise Royal Dutch Shell, la française Total et l'espagnole Repsol ont toutes été citées comme exemples de compagnies pétrolières établies participant au programme Oil-for-food avant le début des surtaxes en 2001.	EN: Total participated in the oil-for-food programme. FR: Total a participé au programme "pétrole contre nourriture".	YES	IE	Short	Same as example 107.
308	EN: On 29 June the Dutch right-wing coalition government collapsed. It was made up of the Christian-democrats (CDA) led by Prime Minister Jan Peter Balkenende, the right wing liberal party (VVD) and the so-called 'left-liberal' D66. FR: Le 29 juin, le gouvernement de coalition de droite néerlandais s'est effondré. Il était composé des chrétiens-démocrates (CDA) dirigés par le Premier ministre Jan Peter Balkenende, du parti libéral de droite (VVD) et du parti dit "libéral de gauche" D66.	EN: Three parties form a Dutch coalition government. FR: Trois partis avaient formé un gouvernement de coalition néerlandais.	YES	IR	Short	We changed the tense of the verb in the hypothesis.

Table 10: Remarks on the translation of the RTE3 dev set.

Test Set

Example ID	Premise	Hypothesis	Label	Task	Length	Comment
76	EN: Edison, Dickson and the other employees of Edison's laboratory made progress on the design to a point. FR: Edison, Dickson et les autres employés du laboratoire d'Edison ont progressé sur la conception jusqu'à un certain point.	EN: Dickson worked for Edison. FR: Dickson travaillait pour Edison.	YES	IE	Short	We followed the Italian tranlsation ("lavorava") as to whether translate "worked" as "travaillait" or "a travaillé".
80	EN: Buckley's Mixture is a cough syrup invented in 1919 (and still produced today) noted for its extremely bitter taste. FR: Le mélange de Buckley est un sirop contre la toux inventé en 1919 (et encore produit aujourd'hui) noté pour son goût extrêmement amer. IT: Il Buckley's Mixture è uno sciroppo per la tosse inventato nel 1919 (e prodotto ancora oggi) noto per il suo sapore estremamente amaro. DE: Die Buckley-Mischung ist ein Hustensaft, erfunden im Jahre 1919 (und noch heute produziert), der für seinen extrem bitteren Geschmack bekannt ist.	EN: Buckley's Mixture is a remedy against cough. FR: Le mélange de Buckley est un remède contre la toux. IT: Buckley's Mixture è un rimedio contro la tosse. DE: Die Buckley-Mischung ist ein Mittel gegen Husten.	YES	IE	Short	"Mixture", we translated "mixture", as they did in the German version, and unlike the Italian version.
280	EN: Setting national goals and developing national standards to meet them are recent strategies in the our education reform policy. Support for national education standards by state governments originated in 1989, when the National Governors Association endorsed national education goals. President George Bush immediately added his support by forming the National Education Goals Panel. FR: La fixation d'objectifs nationaux et l'élaboration de normes nationales pour les atteindre sont des stratégies récentes dans notre politique de réforme de l'éducation. Le soutien des gouvernements des États aux normes nationales d'éducation a vu le jour en 1989, lorsque la National Governors Association a approuvé les objectifs nationaux d'éducation. Le président George Bush a immédiatement ajouté son soutien en formant le groupe d'experts sur les objectifs de l'éducation nationale. IT: Stabilire obiettivi nazionali e sviluppare standard nazionali per raggiungerli sono strategie recenti della nostra politica di riforma dell'educazione. Il sostegno agli standard d'educazione nazionali da parte dei governi statali cominciò nel 1989, quando la National Governors Association approvò obiettivi educativi nazionali. Il Presidente George Bush aggiunse immediatamente il suo sostegno formando il National Education Goals Panel. DE: Für die Festlegung nationaler Ziele und die Entwicklung nationaler Standards sind neue Strategien in unserer Bildungsreformpolitik nötig, , um sie zu erreichen. Die Unterstützung nationaler Bildungsstandards durch die Landesregierungen entstand im Jahre 1989, als die National Governors Association die nationalen Bildungsziele befürwortete. Präsident George Bush hat sofort seine Unterstützung beigesteuert, indem er das Forum für Nationale Bildungsziele bildete.	EN: U.S. sets new educational standards. FR: Les États-Unis établissent de nouvelles normes éducatives. IT: Gli USA stabiliscono nuovi standard educativi. DE: Die USA legen neue Bildungsstandards fest.	YES	IR	Long	We translated the National Education Goals Panel to French, but not the National Governors Association, following the German translators and not the Italian ones, who kept both in English as in the original. There was a time when we would tend to translate "national education goals" as "objectifs de l'éducation nationale" instead of "objectifs nationaux de l'éducation".
286	EN: A Russian court has dismissed a criminal case against a village school teacher accused of using pirated Microsoft software. FR: Un tribunal russe a rejeté une affaire pénale contre un enseignant d'école de village accusé d'avoir utilisé des logiciels Microsoft piratés.	EN: Russian court throws out Microsoft piracy case. FR: Un tribunal russe rejette l'affaire de piratage de Microsoft.	YES	IR	Short	We used the same word ("rejeter") for "dismiss" and "throw out", as it was done in the German and Italian version.
326	EN: European officials have commented on the slowdown in Turkish reforms which, combined with the Cyprus problem, has led the EU's enlargement commissioner to warn of an impeding 'train crash' in negotiations with Turkey. Despite these setbacks, Turkey has closed its first chapter in negotiations in June 2006. FR: Les responsables européens ont commenté le ralentissement des réformes turques qui, combiné au problème chypriote, a conduit le commissaire européen à l'élargissement à mettre en garde contre un "naufrage" imminent dans les négociations avec la Turquie. Malgré ces revers, la Turquie a clos son premier chapitre de négociations en juin 2006.	EN: Turkey negotiates to join the EU. FR: La Turquie négocie son adhésion à l'UE.	YES	IR	Short	We replaced the train crash by a shipwreck ("naufrage").

Table 11: Remarks on the translation of the RTE3 test set.

Appendix B. Corrections Made

RTE3

Development Set

Example ID	Premise	Hypothesis	Initial label	Task	Length	Comment
169	EN: The pet passport is a pink A4 sheet which contains the microchip number and certification that the dog has a rabies vaccination, and needs to be signed by a veterinary surgeon who has LVI status. The passport is not to be confused with the much smaller purple folder routinely issued by vets which records the complete vaccination history of the pet. FR: Le passeport pour animaux de compagnie est une feuille A4 rose qui contient le numéro de la micropuce et la certification que le chien a été vacciné contre la rage, et qui doit être signée par un vétérinaire ayant le statut LVI. Le passeport ne doit pas être confondu avec la pochette violette beaucoup plus petite délivrée couramment par les vétérinaires qui enregistre l'historique complet des vaccinations de l'animal.	EN: The pet passport contains the complete vaccination history of pets. FR: Le passeport pour animaux de compagnie contient l'historique complet des vaccinations des animaux de compagnie.	UNKNOWN	IE	Long	Changed from UNKNOWN to NO.

Table 12: Corrections on the label of the translation of the RTE3 dev set.

Test Set

Example ID	Premise	Hypothesis	Initial label	Task	Length	Comment
11	EN: In the Super Nintendo Entertainment System release of the game as Final Fantasy III , Biggs' name was Vicks. FR: Dans la version sortie sur la Super Nintendo Entertainment System du jeu sous le nom de Final Fantasy III, le nom de Biggs était Vicks.	EN: Final Fantasy III is produced by the Super Nintendo Entertainment System. FR: Final Fantasy III est produit par la Super Nintendo Entertainment System.	YES	IE	Short	Changed from YES to UNKNOWN.
775	EN: Wal-Mart's belated decision is seen as an important victory at a time when women's organisations fear abortion rights are under threat. FR: La décision tardive de Wal-Mart est considérée comme une victoire importante à un moment où les organisations de femmes craignent que le droit à l'avortement soit menacé.	EN: Women's organisations are fighting against abortion rights. FR: Des organisations de femmes luttent contre le droit à l'avortement.	UNKNOWN	SUM	Short	Changed from UNKNOWN to NO.
776	EN: Last month, South Dakota moved to ban all abortion in the state, even in cases of rape and incest. Mississippi is considering similar legislation. FR: Le mois dernier, le Dakota du Sud a décidé d'interdire tout avortement dans l'État, même en cas de viol et d'inceste. Le Mississippi envisage une législation similaire.	EN: In South Dakota abortion is allowed in cases of rape and incest. FR: Dans le Dakota du Sud, l'avortement est autorisé en cas de viol et d'inceste.	UNKNOWN	SUM	Short	Changed from UNKNOWN to NO.
778	EN: Firefighters were called to The Memorial Baptist Church shortly after 11 p.m. and found smoke billowing from its educational wing. IT: I pompieri vennero chiamati alla Chiesa Battista The Memorial subito dopo le 23.00 e trovarono il fumo che si levava dalla sua ala per l'istruzione.	EN: A fire were signaled at The Memorial Baptist Church. IT: Un incendio fu segnalato alla Chiesa Battista Unity Free Will.	YES	SUM	Short	Following the original pair, Chiesa Battista The Memorial in the hypothesis in Italian, not Chiesa Battista Unity Free Will.

Table 13: Corrections on RTE3 test set.

GQNLI-FR

Example ID	Premise	Hypothesis	Initial label	Comment
221	EN: There are 100 footballers and 100 swimmers. Most footballers and most swimmers hate each other. FR: Il y a 100 footballeurs et 100 nageurs. La plupart des footballeurs et la plupart des nageurs se détestent.	EN: One villager and one townsman hate each other. FR: Un villageois et un citadin se détestent.	Entailment	Changed from Entailment to Neutral.

Table 14: Corrections on GQNLI dataset.