Overview of the CLEF 2022 JOKER Task 3: Pun Translation from English into French

Liana Ermakova[1], Fabio Regattin[3], Tristan Miller[4], Anne-Gwenn Bosser[5], Claudine Borg[6], Benoît Jeanjean[1], Elise Mathurin[1], Gaelle Le Corre[7], Radia Hannachi[8], Sílvia Araújo[9], Julien Boccou[1], Albin Digue[1] and Aurianne Damoy[1]

[1]Université de Bretagne Occidentale, HCTI, 29200 Brest, France
[2]Maison des sciences de l'homme en Bretagne, 35043 Rennes, France
[3]Dipartimento DILL, Università degli Studi di Udine, 33100 Udine, Italy
[4]Austrian Research Institute for Artificial Intelligence, Vienna, Austria
[5]École Nationale d'Ingénieurs de Brest, LabSTICC CNRS UMR 6285
[5]University of Malta, Msida MSD 2020, Malta
[7]Université de Bretagne Occidentale, CRBC, 29200 Brest, France
[8]Université de Bretagne Sud, HCTI, 56321 Lorient, France
[9]University of Minho, Portugal
[4]Austrian Research Institute for Artificial Intelligence, Vienna, Austria

Abstract
The translation of the pun is one of the most challenging issues for translators and for this reason has become an intensively studied phenomenon in the field of translation studies. Translation technology aims to partially or even totally automate the translation process, but relatively little attention has been paid to the use of computers for the translation of wordplay. The CLEF 2022 JOKER track aims to build a multilingual corpus of wordplay and evaluation metrics in order to advance the automation of creative-language translation. This paper provides an overview of the track's Pilot Task 3, where the goal is to translate entire phrases containing wordplay (particularly puns). We describe the data collection, the task setup, the evaluation procedure, and the participants' results. We also cover a side product of our project, a homogeneous monolingual corpus for wordplay detection in French.

Keywords
wordplay, computation humour, pun, machine translation, deep learning

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
liana.ermakova@univ-brest.fr (L. Ermakova)
https://www.joker-project.com/ (L. Ermakova)
0000-0002-7598-7474 (L. Ermakova); 0000-0003-3000-3360 (F. Regattin); 0000-0002-0749-1100 (T. Miller); 0000-0003-3858-5502 (C. Borg); 0000-0001-5157-1899 (B. Jeanjean); 0000-0002-7598-7474 (G. L. Corre); 0000-0003-4321-4511 (S. Araújo)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)
CEUR-WS.org/Vol-3180/paper-128.pdf

1. Introduction

Wordplay is ubiquitous in both speech and writing as a means to evoke humour. It can occur on or intersect with virtually any level of language, including the phonological, orthographical, morphological, lexical, syntactic, or textual [1]. Punning is a particular form of wordplay in which a word or phrase suggests two or more meanings by exploiting polysemy, homonymy, or phonological similarity to another word or phrase [2, 3]. Despite being a popular subject of research in translatology [4, 5], the translation of puns has received little attention in the fields of natural language processing (NLP) and machine translation (MT) [6]. With increasing global communication, the demand for translation grows ever faster, which has spurred rapid development of MT technology [7]. Recent developments in machine learning and artificial intelligence have greatly improved the quality of MT, but puns are often held to be untranslatable, particularly by statistical or neural MT [8, 9], which cannot robustly deal with texts that deliberately disregard or subvert linguistic conventions [6]. Among the main challenges in translating puns are linguistic and cultural differences [10, 11, 12], which can affect the target audience's comprehension of the joke and must therefore inform the translator's choice of strategy.

In 2022, the JOKER workshop at CLEF proposed three pilot tasks [13]: (1) classify and explain instances of wordplay, (2) translate single terms containing wordplay, and (3) translate entire phrases containing wordplay (puns) from English into French. This paper describes and discusses the third of these tasks, including the participating systems and their results. The goal of the workshop was to bring together translators and computer scientists to work on an evaluation framework for wordplay, including data and metric development, and to foster work on automatic methods for wordplay translation.

2. Related work

2.1. Wordplay translation strategies

Over the past few decades, the field of translation studies has devoted increasing interest to wordplay [14]. Various strategies for wordplay translation have been conceived and described over time, and, accordingly, some typologies have been produced. Two of them stand out for their quality and their universalist purpose. The first of these is the fourfold typology of Henry [15, pp. 176–192]:

1. *traduction isomorphe* (isomorphic translation)
2. *traduction homomorphe* (homomorphic translation)
3. *traduction hétéromorphe* (heteromorphic translation)
4. *traduction libre* (free translation)

The **isomorphic** strategy consists of translating a source-text (ST) wordplay with an identical wordplay (except for formal differences) in the target language (TL).

This is what happens, for example, when the German portmanteau adjective *famillionär* (amalgamating *Familie* + *Millionär*) is translated into English or French as *famillionaire*. As in this case, the isomorphic strategy is a borderline situation, which only happens due to fortuitous (or historical) similarities between languages.

The **homomorphic** strategy consists of translating an ST wordplay with a wordplay of the same typology, based on different linguistic material. This is what happens when we translate an anagram with an anagram, or a pun with a different pun (i.e., in the great majority of cases where we cannot lean on the isomorphic strategy).

The **heteromorphic** strategy involves translating an ST wordplay with a wordplay of a different typology in the TL. For instance, we could translate an anagram with a pun, or a portmanteau with assonance.

Free translation takes place when the ST wordplay is translated into something other than wordplay.

Despite its allure (as well as its elegant terminological uniformity), Henry's taxonomy has a serious flaw: the fourth category, free translation, is a potentially very broad one, as it brings together many different strategies. The second wordplay translation typology, developed by Delabastita [16], dissects this fourth category in a much more precise way. This is the reason why we will rely on a combination of both typologies in the rest of this paper. While Henry's typology is mostly based on the author's experience as a translator, Delabastita's was developed on the basis of parallel corpus analysis and therefore reflects the real techniques used by human translators in their work. And while the typology was developed specifically for puns (a type of wordplay that exploits multiple meanings of a term or of similar-sounding words), many of the strategies it describes can be successfully applied to other types of wordplay that are not based on ambiguity. Delabastita lists the following options:

1. pun→pun: The ST pun is translated by a TL pun. This category can be further partitioned into three subtypes, using Henry's typology:
   - isomorphic translation
   - homomorphic translation
   - heteromorphic translation

   Strategies 2 to 8 below can all be related to Henry's fourth category, free translation:
2. pun→non-pun: The pun is translated by a non-punning phrase, which may reproduce all senses of the wordplay or just one of them, without trying to do this in an equally ambiguous way.
3. pun→related rhetorical device: The pun is replaced by some other, rhetorically charged, utterance (involving repetition, alliteration, rhyme, irony, paradox, etc.).
4. pun→zero: The part of text containing the pun is omitted altogether.
5. pun ST=pun TT: The punning text, and sometimes its immediate environment, is/are reproduced in the SL in the target text (TT), without attempting a TL rendering.

6. non-pun→pun: A pun is introduced in the TT where no wordplay was present in the ST.
7. zero→pun: New textual material involving wordplay is added to the TT, which bears no correspondence whatsoever in the ST.
8. editorial techniques: All the paratextual strategies involved in explaining, or presenting alternative renderings for, the pun of the ST (footnotes, prefaces, translator's notes, etc.).

Delabastita insists on one further point: these eight strategies are by no means exclusive. A translator could, for instance, suppress a pun somewhere in their TT (locally leading to a pun→non-pun solution), they could explain it in a footnote (editorial techniques), and finally try to compensate for the loss by adding another pun somewhere else in the text (non-pun→pun or zero→pun).

The very typology of translation strategies drawn by Delabastita directly points to the main reason for the difficulty of conceiving a working machine translation system for puns. How can we automate the omission of a pun, the introduction of wordplay somewhere else in a text, or the reproduction of a SL textual segment in the TT? One could say, then, that the typology developed by Henry could be more useful, because it (usually) only accounts for translations of a wordplay in the ST with a wordplay in the TT. Unfortunately, it cannot be stressed enough that this goes against most human translators' practice. Very often, the strategies used by human translators completely break any kind of textual relationship between the ST and the TT. This is the reason why wordplay translation is seen by many practitioners and theoreticians alike as something "other" than translation – say, as adaptation or as re-creation – and this is the reason why we believe that only Delabastita's typology should be the goal to achieve in the long term for a useful wordplay machine translation engine.

2.2. Computational humour

To date, there have been few studies on the MT of wordplay. Farwell and Helmreich [17] proposed a pragmatic-based approach to MT that accounts for the author's locutionary, illocutionary, and perlocutionary intents (that is, the "how", "what", and "why" of the text), and discuss how it might be applied to puns. However, no working system appears to have been implemented. Miller [18] proposed an interactive method for the computer-assisted translation of puns, an implementation (PunCAT) and evaluation of which was described by Kolb and Miller [19]. Their study was limited to a single language pair (English to German) and translation strategy (namely, the pun→pun strategy described previously). Furthermore, the tool's functionality is limited to facilitating exploration of the semantic fields corresponding to the two meanings of the pun; actually detecting and interpreting the ST pun, and devising a complete TL punning joke, is left to the user.

Numerous studies have been conducted for the related tasks of humour generation and detection. Pun generation systems have often been based on template approaches. Valitutti, Toivonen, Doucet, and Toivanen [20] used lexical constraints to generate adult humour by substituting one word in a pre-existing text. Hong and Ong [21] trained a system to extract automatically humorous templates which were then used for pun generation. Some current efforts to tackle this difficult problem more generally using neural approaches have been hindered by the lack of a sizable pun corpus [22]. Recent work [23] has tackled the issue for generating humourous puns in English based on the data provided at SemEval-2017 [2].

Meanwhile, the recent rise of conversational agents and the need to process large volumes of social media content point to the necessity of automatic humour recognition [24]. Humour and irony studies are now crucial when it comes to social listening [25, 26, 27, 28], dialogue systems (chatbots), recommender systems, reputation monitoring, and the detection of fake news [29] and hate speech [30]. However, the automatic detection, location, and interpretation of humorous wordplay in particular has so far been limited to punning. And while even the earliest such systems have achieved decent performance on the detection and location tasks [31], methods for actually interpreting the double meaning of the pun – a prerequisite for translation – have not been as intensively researched. Miller, Hempelmann, and Gurevych [31] report an accuracy of 16.0% and 7.7% accuracy for homographic and heterographic puns, respectively, and this baseline does not seem to have been improved upon in more recent work [32]. Again, indications point to the lack of sufficient training data as a stumbling block to further progress, especially for languages other than English.

A few monolingual humour corpora do exist, including the datasets created for shared tasks of the International Workshop on Semantic Evaluation (SemEval): #HashtagWars: Learning a Sense of Humor [33], Detection and Interpretation of English Puns [31], Assessing Humor in Edited News Headlines [34], and HaHackathon: Detecting and Rating Humor and Offense [35]. Mihalcea and Strapparava [36] collected 16 000 humorous sentences and an equal number of negative samples from news titles, proverbs, the British National Corpus, and the Open Mind Common Sense dataset, while another dataset contains 2400 puns and non-puns from news sources, Yahoo! Answers, and proverbs [37, 38]. Most datasets are in English, with some notable exceptions for Italian [39], Russian [40, 41], and Spanish [42]. To the best of our knowledge, no corpus exists for French.

To the best of our knowledge the only parallel corpus of wordplay was the one introduced in our research [13, 43]. We manually collected over a thousand translated examples of wordplay, in English and French, from video games, advertising slogans, literature, and other sources [13, 43]. Each example has been manually classified according to a multi-label inventory of wordplay types and structures, and annotated according to its lexical-semantic or morphosemantic components. However, the majority of the collected wordplay was *single-term* proper nouns or neologisms based on portmanteaux, the like of which are common in the Asterix and Harry Potter universes.

Large pre-trained AI models, like Jurassic-1 [44], mT5 [45], BERT [46], and GPT [47, 48], have outperformed other state-of-the-art models on several NLP tasks, including MT [49]. Performance of such supervised MT systems depends on the quality and quantity of training data [50]. However, as mentioned above, there exist no large-scale, broad-coverage parallel corpora of wordplay. This corpus is a key prerequisite for the training and evaluation of MT models.

Humorous wordplay often exploits the confrontation of similar forms with different meanings, evoking incongruity between expected and presented stimuli. This makes it particularly important in NLP to study the strategies that human translators use for dealing with wordplay [51, 52]. On the one hand, this is because MT is generally ignorant of pragmatics and assumes that words in the source text are formed and used in a conventional manner. MT systems fail to recognise the deliberate ambiguity of puns or the unorthodox morphology of neologisms, leaving such terms untranslated or else translating them in ways that lose the humorous aspect [18].

3. Data

Our English corpus of puns is mainly based on that of the SemEval-2017 shared task on pun identification [31]. The original annotated dataset contains 3387 standalone English-language punning jokes, between 2 and 69 words in length, sourced from offline and online joke collections. Roughly half of the puns in the collection are "weakly" homographic (meaning that the lexical units corresponding to the two senses of the pun, disregarding inflections and particles, are spelled identically) while the other half are heterographic (that is, with lemmas spelled differently). The original annotation scheme is rather simple, indicating only the pun's location within the joke, whether it is homographic or heterographic, and the two meanings of the pun (with reference to senses in WordNet [53]).

In order to translate this subcorpus from English into French, we applied a gamification strategy. More precisely, we organised a translation contest.[1] The contest was open to students but we also received multiple translations out of official ranking from professional translators and academics in translation studies. The results were submitted via Google Forms. Forty-seven participants submitted 3950 translations of 500 puns from the SemEval-2017 dataset. We first took 250 puns in English from each of homographic and heterographic subsets. In the form, the homographic and heterographic puns were alternated. Each page of the form contained 100 puns.

Unfortunately, Google Forms does not allow questions to be shuffled for each participant. Thus, we observed a drastic drop in the number of translations per pun starting from the second page (see Figure 1). As we had two participants who translated almost all puns (see Figure 3), we have a conspicuous peak on the number of translations per query (Figure 2). However, this histogram does not provide a clear idea about the translation difficulty of puns as the vast majority of participants translated only the first page of the form. Figure 4, the number of translations per query on the first page only, perhaps better reflects the translation difficulty distribution.

[1] https://www.joker-project.com/pun-translation-contest/

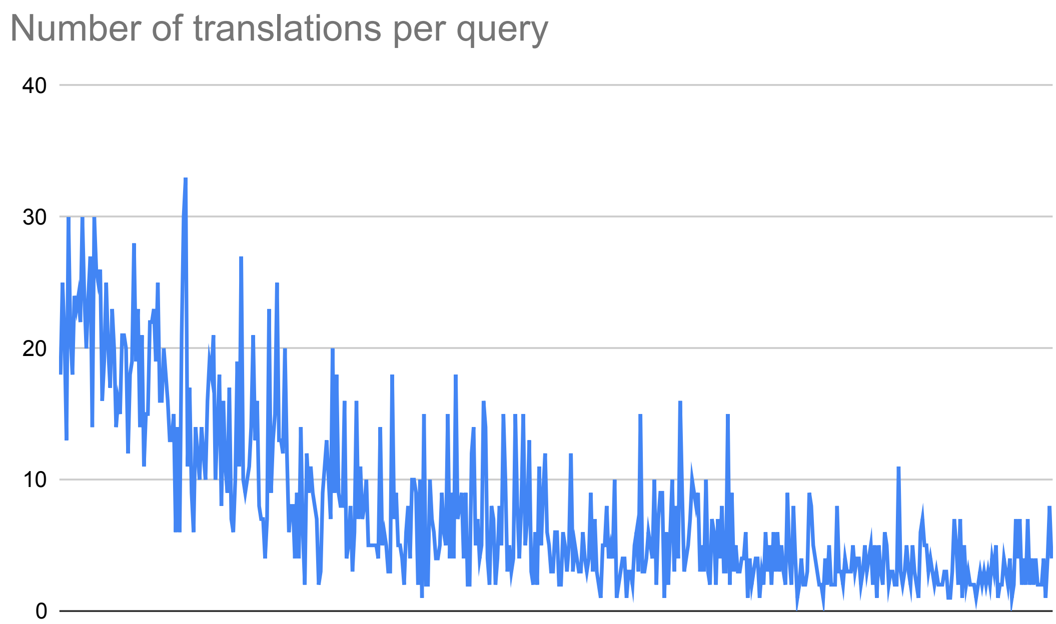

Figure 1: Number of translations per query

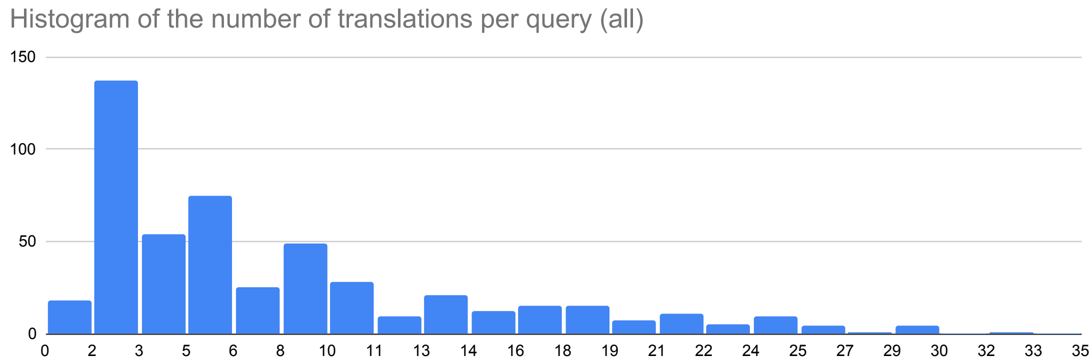

Figure 2: Histogram of the number of translations per query (all)

Besides this SemEval-derived data, we sourced further translation pairs from published literature and from puns translated by Master's students in translation.

We annotated the dataset according to the classification used for Pilot Task 1 of our workshop [54].

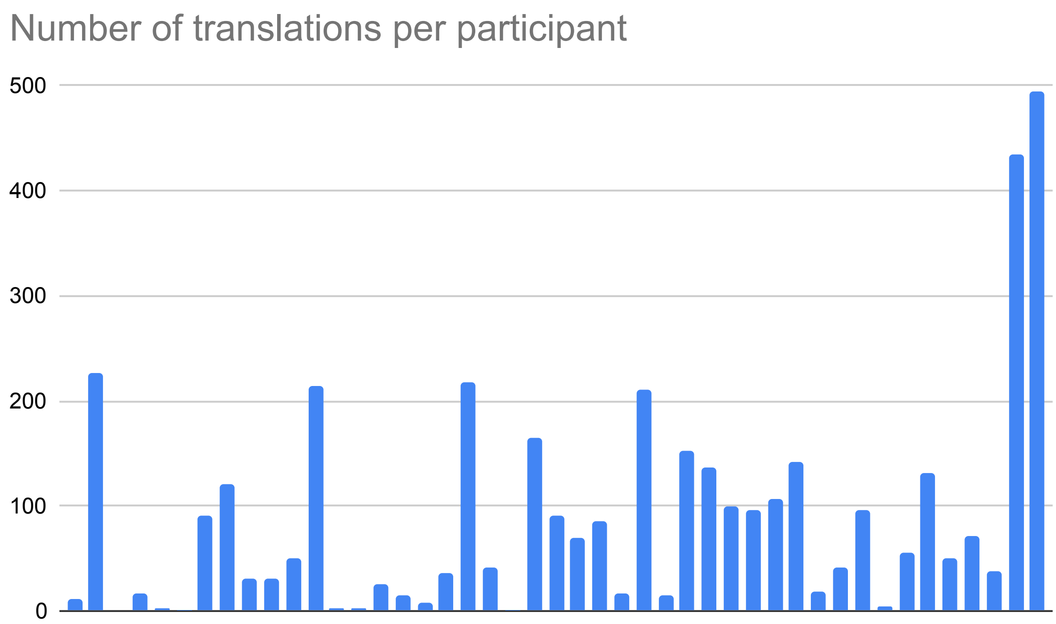

Figure 3: Number of translations per participant

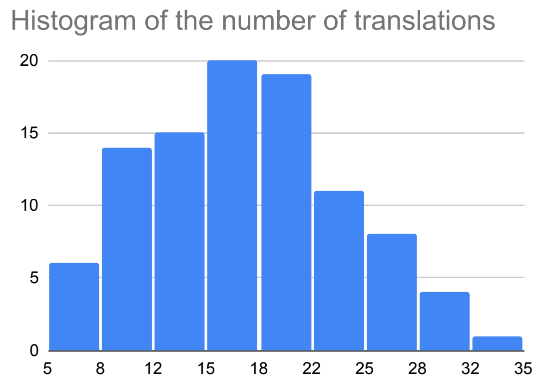

Figure 4: Histogram of the number of translations per query (first page)

3.1. Training data

In total, the final annotated training set in English contained 1772 instances. The French collection contained 4753 annotated instances. The data was provided to participants as a JSON file (or a CSV file for manual runs) with fields denoting the instance's unique ID (`id`), the source text in English (`en`), and a target text in French (`fr`). For example:

```
[
  {
    "id": "pun_724_1",
    "en": "My name is Wade and I'm in swimming pool maintenance.",
    "fr": "Je m'appelle Jacques Ouzy, je m'occupe de l'entretien des
       piscines."
  }
]
```

3.1.1. Test data

The test set contains 2378 instances in English from the SemEval-2017 pun task [31]. The data format was identical to that of the training data, except that the field for the target text was omitted. Example:

```
[
  {
    "id": "het_713",
    "en": "Ever since my mineral extraction facility was converted to
       parking, I've had a lot on my mine.",
  }
]
```

The expected output format was identical to that of the training data, but with the addition of fields `RUN_ID` and `MANUAL`. The `RUN_ID` field value uniquely identifies a given run and is formed of the team ID (as registered on the CLEF website) followed by the task ID (in this pilot task, always `task_3`), followed by the run number. The `MANUAL` field value can be either a 1 (indicating a manual translation run) or a 0 (indicating a machine translation run). Example:

```
[
  {
    "RUN_ID": "JCM_task_3_run1",
    "MANUAL": 1,
    "id": "pun_724_1",
    "en": "My name is Wade and I'm in swimming pool maintenance.",
    "fr": "Je m'appelle Jacques Ouzy, je m'occupe de l'entretien des
       piscines."
  }
]
```

4. Evaluation metrics

As we have previously argued [13], the prevailing BLEU metric for machine translation is clearly inappropriate for use with wordplay, where a wide variety of translation strategies (and solutions implementing those strategies) are permissible. Many of these strategies require metalexical awareness and preservation of features such as lexical ambiguity and phonetic similarity.

For our evaluation, participants' runs were pooled together. We filtered out all translations that did not match the regular expression `.+[?.!"]\s*$` as we considered these translations to be truncated. Indeed, in some runs (e.g., Cecilia's run 3) the majority of generated translations were too short with regard to the source wordplay and truncated in the middle of the sentence. We refer further the retained translations as *valid*.

We then filtered out French translations identical to the original wordplay in English, as we considered these wordplay instances to be *not translated*.

The pool of *valid* distinct translations into French contains 9513 instances. Three Master's students in translation, French native speakers, manually evaluated each valid translation as follows. We evaluated the following errors:

- *nonsense*: This metric is *true* when the translation contains a nonsensical passage.
- *syntax problem*: This metric is *true* when the translation contains a passage with errors in syntax.
- *lexical problem*: This metric is *true* when the translation contains a passage with errors in word choice/use.

An instance was not evaluated for subsequent metrics if one of the above errors was identified. For translations without these errors, we evaluated:

- *lexical field preservation, sense preservation, comprehensible terms, wordplay form*: These four metrics are evaluated as in Task 2.
- *identifiable wordplay*: A value of *true* is assigned to translations that are wordplay and are understandable for general audience. For example, the wordplay *"Je n'abandonnerai jamais mes chiens!" dit Tom cyniquement.* (meaning " 'I'll never abandon my dogs!' Tom said cynically") requires etymological knowledge that is beyond most readers.
- *over-translation*: A value of *true* is assigned to translations that have useless multiple wordplay instances when the source text has just one.
- *style shift*: A value of *true* is assigned to translations that have style shift (e.g., where a vulgarism is present either in the source text or the translation but not in both).
- *humorousness shift*: A value of *true* is assigned to translations that were judged to be much more or much less funnier than the source wordplay.

Note that the categories *over-translation*, *style shift* and *humorousness shift* are necessarily subjective.

Table 1
Scores of participants' runs for Pilot Task 3

	LJGG DeepL	FAST_MT	LJGG auto	Cecilia run 1	Humorless	Cecilia run 3
total	2378	2378	2378	2378	2378	2378
valid	2324	2120	2264	2343	384	7
not translated	39	103	206	49	22	2
nonsense	59	220	349	51	297	3
syntax problem	17	58	46	41	6	0
lexical problem	25	79	78	52	10	0
lexical field preservation	2184	1739	1595	2155	118	6
sense preservation	1938	1453	1327	1803	100	6
comprehensible terms	1188	867	827	744	56	5
wordplay form	373	345	261	251	19	1
identifiable wordplay	342	318	240	243	16	1
over-translation	3	1	9	13	0	0
style shift	9	12	4	4	0	0
humorousness shift	930	765	838	1427	68	4

5. Methods used by the participants

Four teams participated in Pilot Task 3: FAST_MT [55], Cecilia [56], Humorless (no paper submitted), and LJGG [57]. Cecilia updated their run, and LJGG submitted two runs, one of which was produced with DeepL.[2] LJGG's other run, and that of Cecilia, were generated using the SimpleT5 library[3] for the Google T5 (Text-To-Text Transfer Transformer) model, which is based on the transfer learning with a unified text-to-text transformer [58].

FAST_MT also applied transformers but decided not to do fine-tuning; more precisely, the team used the Helsinki/NLP/opus-mt-en-fr model [59] from the Hugging Face[4] repository.

6. Results

Table 1 presents the results of submitted runs for Task 3. We observe that in many cases the successful translations are due to the existence of the same lexical ambiguity (homonymy) in both languages:

Example 6.1. A train load of paint derailed. Nearby businesses were put in the red.

Un train de peinture a déraillé. Les entreprises voisines ont été mises dans le rouge.

[2] https://www.deepl.com/
[3] https://github.com/Shivanandroy/simpleT5
[4] https://huggingface.co/

Example 6.2. An undertaker can be one of your best friends, he is always the last one to let you down.

Un entrepreneur peut être l'un de vos meilleurs amis, il est toujours le dernier à vous laisser tomber.

We also noticed some surprisingly successful translations:[5]

Example 6.3. Success comes in cans, failure comes in cant's.
Le succès c'est dans les canons, le pétrin c'est dans les canettes.

Example 6.4. Wal-Mart Is Not the Only Saving Place. Come On In.
Le clerc n'est pas le seul à faire des économies.

Notably, a few successful translations used anglicisms:

Example 6.5. I used to be addicted to soap, but I'm clean now.
Avant, j'étais accro au savon, mais je suis clean maintenant.

Example 6.6. When the beekeeper moved into town he created quite a buzz.
Lorsque l'apiculteur s'est installé en ville, il a créé un véritable buzz.

Out of over 1155 translations containing wordplay, only 311 were translations of heterographic puns. This suggests that the state-of-the art machine translation is still unsuitable for translating wordplay, even with a manually annotated training set. The successful machine translations are seemingly accidental, owing to the existences of the same word ambiguity in both languages.

In total only 13% of automatically translated plays on words were successful, compared to the 90% success rate for instances translated by the human participants of our contest.

7. French corpus for wordplay detection

A side product of our project is a creation of homogeneous monolingual corpus for wordplay detection in French.

As stated previously, our parallel wordplay corpus is primarily constructed by the translation of the corpus of English puns introduced at SemEval-2017 Task 7: Detection and Interpretation of English Puns [2]. This corpus contains 2250 homographic and 1780 heterographic puns. All puns were translated during the translation contest described in §3 and 90% of these translations were successful. These facts provide evidence that pun translation is possible. On the other hand, machine translations succeeded only in 13% of cases. We manually annotated all 9513 machine translations submitted by our participants. Note that the translations of the same sentence are close to each other in terms of length and lexical field. Given successful and

[5]On closer inspection, we determined that Example 6.4 was very close to an example from a training set.

Table 2
Confusion matrix of T5 on all SemEval-2017 Task 7 data

	Pun (ground truth)	Not pun (ground truth)
	1607 Homographic: 11.64 avg len 1271 Heterographic: 11.6 avg len	643 Homographic: 8.7 avg len 509 Heterographic: 8.6 avg len
Pun (predicted)	1564 Homographic: 11.7 avg len 1238 Heterographic: 11.7 avg len	25 Homographic: 9.1 avg len 18 Heterographic: 9.5 avg len
Not pun (predicted)	43 Homographic: 10.7 avg len 33 Heterographic: 7.7 avg len	618 Homographic: 8.7 avg len 491 Heterographic: 8.5 avg len

unsuccessful human and machine translations, we obtained a homogeneous corpus in French containing wordplay and non-wordplay with similar characteristics. This similarity in terms of length and lexicon is crucial to build a corpus for wordplay detection, as the vast majority of state-of-the-art NLP approaches are neural ones [60]. Thus, these models might learn the difference in lexicon or sentence length instead of the ambiguity in a pun.

Indeed, when we tested the Google T5 model [58] via the SimpleT5 library on the shuffled SemEval-2017 data, we obtained 92.8% on the test set (403 shuffled instances). The split was 70% train, 20% validation, and 10% test. However, a closer look at the confusion matrix (see Table 2) provides evidence that the non-puns are much shorter than puns in the corpus in average and the model fails when it is not the case. Thus, the homogeneity of the corpus for wordplay detection is important.

To the best of our knowledge, this is the first corpus for wordplay detection in French.

This corpus has been already used for a five-step wordplay generation, aiming to transform a non-humorous text into wordplay [61]. This source corpus without wordplay has the potential to be transformed into a corpus of wordplays. Only the machine translations that were annotated not to contain wordplay were used for this generation (6780 texts in total).

8. Conclusion

The goal of the JOKER project is to advance the automation of creative-language translation by developing the requisite parallel data and evaluation metrics for translating wordplay. To this end, we organised the JOKER track at CLEF 2022, consisting of a workshop and associated pilot tasks on automatic wordplay analysis and translation. We collected a unique English–French parallel wordplay corpus.

Successful translations of puns in Pilot Task 3 are usually accidental, as they exploit the ambiguity of the literal translation of the target wordplay term both in English and French. However, some translations are successful due to the correct use of anglicisms in French.

A side product of our project is a creation of homogeneous monolingual corpus for wordplay detection in French. To the best of our knowledge, this is the first corpus for wordplay detection in French.

Further details on the other pilot tasks and the submitted runs can be found in the CLEF CEUR proceedings [62]. The overview of the entire JOKER track can be found in the LNCS proceedings [43]. Additional information on the track is available on the JOKER website: http://www.joker-project.com/

9. Authors' contribution

The general framework was proposed by L. Ermakova with the participation of T. Miller and A.-G. Bosser. L. Ermakova, F. Regattin, S. Araújo, B. Jeanjean, C. Borg, G. Le Corre, E. Mathurin, R. Hannachi, and T. Miller worked on the translation contest organisation. J. Boccou, A. Digue, and A. Damoy participated in data creation and worked on the result evaluation under supervision of L. Ermakova. F. Regattin wrote the state-of-the-art on wordplay translation strategies.

Acknowledgments

This work has been funded in part by the National Research Agency under the program *Investissements d'avenir* (Reference ANR-19-GURE-0001) and by the Austrian Science Fund under project M 2625-N31. JOKER is supported by *La Maison des sciences de l'homme en Bretagne*.

We thank other members of the jury of the pun translation contest: Caroline Comacle, Mohamed Saki, Helen McCombie, and Catherine Davis, as well as the translation contest participants. We thank Alain Kerhervé for the financial support of the translation contest and Eric Sanjuan who provided resources for data management.

We would like also thank other JOKER organisers: Monika Bokiniec, Ġorġ Mallia, Gordan Matas, Mohamed Saki, Benoît Jeanjean, Radia Hannachi, Danica Škara, and other PC members: Grigori Sidorov, Victor Manuel Palma Preciado, and Fabrice Antoine.

References

[1] S. Laviosa, Wordplay in advertising: Form, meaning and function, Scripta Manent 1 (2015) 25–34.

[2] T. Miller, C. F. Hempelmann, I. Gurevych, SemEval-2017 Task 7: Detection and interpretation of English puns, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 58–68. doi:10.18653/v1/S17-2005.

[3] L. Bobchynets, Lexico-semantic means of creation of pun in spanish and portuguese jokes, Nova fìlologìâ (2020). doi:10.26661/2414-1135-2020-80-1-11.

[4] E. S. Rudenko, R. I. Bachieva, Wordplay as a translation problem, Bulletin of the Moscow State Regional University (Linguistics) (2020). doi:10.18384/2310-712x-2020-2-78-85.
[5] R. Tuzzikriah, H. Ardi, Students' perception on the problem in translating humor text, Proceedings of the Eighth International Conference on English Language and Teaching (ICOELT-8 2020) (2021). doi:10.2991/assehr.k.210914.061.
[6] T. Miller, The punster's amanuensis: The proper place of humans and machines in the translation of wordplay, in: Proceedings of the Second Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019), 2019, pp. 57–64. doi:10.26615/issn.2683-0078.2019_007.
[7] Y. He, Challenges and countermeasures of translation teaching in the era of artificial intelligence, Journal of Physics: Conference Series 1881 (2021). doi:10.1088/1742-6596/1881/2/022086.
[8] H. Ardi, M. A. Hafizh, I. Rezqi, R. Tuzzikriah, Can machine translations translate humorous texts?, Humanus (2022). doi:10.24036/humanus.v21i1.115698.
[9] F. Regattin, Traduction automatique et jeux de mots : l'incursion (ludique) d'un inculte, 2021. URL: https://motsmachines.github.io/2021/en/submissions/Mots-Machines-2021_paper_5.pdf.
[10] F. R. B. Kembaren, The challenges and solutions of translating puns and jokes from english to indonesian, VISION (2020). doi:10.30829/vis.v16i2.807.
[11] O. G. Hniedkova, Z. O. Karpenko, Peculiarities of pun formation and translation of pun as a type of wordplay, "Scientific notes of V. I. Vernadsky Taurida National University", Series: "Philology. Journalism" 2 (2021) 254–261. doi:10.32838/2710-4656/2021.1-2/44.
[12] G. Kovács, Translating humour – a didactic perspective, Acta Universitatis Sapientiae, Philologica 12 (2020) 68–83.
[13] L. Ermakova, T. Miller, O. Puchalski, F. Regattin, É. Mathurin, S. Araújo, A.-G. Bosser, C. Borg, M. Bokiniec, G. L. Corre, B. Jeanjean, R. Hannachi, Ġ. Mallia, G. Matas, M. Saki, CLEF Workshop JOKER: Automatic Wordplay and Humour Translation, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, volume 13186 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2022, pp. 355–363. doi:10.1007/978-3-030-99739-7_45.
[14] F. Regattin, Traduire les jeux de mots : une approche intégrée, Atelier de traduction (2015) 129–151. URL: http://www.diacronia.ro/ro/indexing/details/A19521/pdf.
[15] J. Henry, La Traduction Des Jeux De Mots, Presses de la Sorbonne Nouvelle, Paris, 2003.
[16] D. Delabastita, Wordplay as a translation problem: a linguistic perspective, in: Ein internationales Handbuch zur Übersetzungsforschung, volume 1, De Gruyter Mouton, 2008, pp. 600–606. doi:10.1515/9783110137088.1.6.600.
[17] D. Farwell, S. Helmreich, Pragmatics-based MT and the translation of puns, in: Proceedings of the 11th Annual Conference of the European Association for Machine Translation, 2006, pp. 187–194. URL: http://www.mt-archive.info/

EAMT-2006-Farwell.pdf.
[18] T. Miller, The punster's amanuensis: The proper place of humans and machines in the translation of wordplay, in: Proceedings of the Second Workshop on Human-Informed Translation and Interpreting Technology, 2019, pp. 57–64. doi:10.26615/issn.2683-0078.2019_007.
[19] W. Kolb, T. Miller, Human–computer interaction in pun translation, in: J. Hadley, K. Taivalkoski-Shilov, C. S. C. Teixeira, A. Toral (Eds.), Using Technologies for Creative-Text Translation, Routledge, 2022. To appear.
[20] A. Valitutti, H. Toivonen, A. Doucet, J. M. Toivanen, "let everything turn well in your wife": Generation of adult humor using lexical constraints, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, volume 2, Association for Computational Linguistics, 2013, p. 243–248. URL: https://aclanthology.org/P13-2044.
[21] B. A. Hong, E. Ong, Automatically extracting word relationships as templates for pun generation, in: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity, Association for Computational Linguistics, Boulder, Colorado, 2009, pp. 24–31. URL: https://aclanthology.org/W09-2004.
[22] Z. Yu, J. Tan, X. Wan, A neural approach to pun generation, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, volume 1, Association for Computational Linguistics, 2018, p. 1650–1660. URL: https://aclanthology.org/P18-1153. doi:10.18653/v1/P18-1153.
[23] H. He, N. Peng, P. Liang, Pun Generation with Surprise, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1734–1744. URL: https://aclanthology.org/N19-1172. doi:10.18653/v1/N19-1172.
[24] A. Nijholt, A. Niculescu, A. Valitutti, R. E. Banchs, Humor in human-computer interaction: a short survey, in: A. Joshi, D. K. Balkrishan, G. Dalvi, M. Winckler (Eds.), Adjunct Proceedings: INTERACT 2017 Mumbai, Industrial Design Centre, Indian Institute of Technology Bombay, 2017, pp. 199–220. URL: https://www.interact2017.org/downloads/INTERACT_2017_Adjunct_v4_final_24jan.pdf.
[25] B. Ghanem, J. Karoui, F. Benamara, V. Moriceau, P. Rosso, IDAT@FIRE2019: Overview of the track on irony detection in Arabic tweets, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, Association for Computing Machinery, 2019, p. 10–13. doi:10.1145/3368567.3368585.
[26] J. Karoui, F. Benamara, V. Moriceau, V. Patti, C. Bosco, N. Aussenac-Gilles, Exploring the impact of pragmatic phenomena on irony detection in tweets: a multilingual corpus study, in: 15th Conference of the European Chapter of the Association for Computational Linguistics, volume 1, Association for Computational Linguistics, 2017, p. 262–272. URL: https://oatao.univ-toulouse.fr/18921/.
[27] J. Karoui, B. Farah, V. Moriceau, N. Aussenac-Gilles, L. Hadrich-Belguith, Towards a contextual pragmatic model to detect irony in tweets, in: Proceedings

of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, volume 2, Association for Computational Linguistics, 2015, pp. 644–650. URL: http://aclweb.org/anthology/P15-2106. doi:10.3115/v1/P15-2106.

[28] A. Reyes, P. Rosso, D. Buscaldi, From humor recognition to irony detection: the figurative language of social media, Data & Knowledge Engineering 74 (2012) 1–12. doi:10.1016/j.datak.2012.02.005.

[29] G. Guibon, L. Ermakova, H. Seffih, A. Firsov, G. Le Noé-Bienvenu, Multilingual Fake News Detection with Satire, in: CICLing: International Conference on Computational Linguistics and Intelligent Text Processing, La Rochelle, France, 2019. URL: https://halshs.archives-ouvertes.fr/halshs-02391141.

[30] C. Francesconi, C. Bosco, F. Poletto, M. Sanguinetti, Error Analysis in a Hate Speech Detection Task: the Case of HaSpeeDe-TW at EVALITA 2018, in: R. Bernardi, R. Navigli, G. Semeraro (Eds.), Proceedings of the 6th Italian Conference on Computational Linguistics, 2018. URL: http://ceur-ws.org/Vol-2481/paper32.pdf.

[31] T. Miller, C. F. Hempelmann, I. Gurevych, SemEval-2017 Task 7: Detection and interpretation of English puns, in: Proceedings of the 11th International Workshop on Semantic Evaluation, 2017, pp. 58–68. doi:10.18653/v1/S17-2005.

[32] A. Jain, P. Yadav, H. Javed, Equivoque: Detection and interpretation of English puns, in: Proceedigns of the 8th International Conference System Modeling and Advancement in Research Trends, 2019, pp. 262–265. doi:10.1109/SMART46866.2019.9117433.

[33] P. Potash, A. Romanov, A. Rumshisky, SemEval-2017 Task 6: #HashtagWars: Learning a sense of humor, in: Proceedings of the 11th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2017, pp. 49–57. doi:10.18653/v1/S17-2004.

[34] N. Hossain, J. Krumm, M. Gamon, H. Kautz, SemEval-2020 task 7: Assessing humor in edited news headlines, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 746–758. URL: https://aclanthology.org/2020.semeval-1.98. doi:10.18653/v1/2020.semeval-1.98.

[35] J. A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, W. Magdy, Semeval-2021 task 7: Hahackathon, detecting and rating humor and offense, in: Proceedings of the 15th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2021, p. 105–119. URL: https://aclanthology.org/2021.semeval-1.9. doi:10.18653/v1/2021.semeval-1.9.

[36] R. Mihalcea, C. Strapparava, Making computers laugh: Investigations in automatic humor recognition, in: Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing: Proceedings of the Conference, Association for Computational Linguistics, Stroudsburg, PA, 2005, pp. 531–538. URL: http://www.aclweb.org/anthology/H/H05/H05-1067. doi:10.3115/1220575.1220642.

[37] A. Cattle, X. Ma, Recognizing humour using word associations and humour anchor extraction, in: Proceedings of the 27th International Conference on

Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, p. 1849–1858. URL: https://www.aclweb.org/anthology/C18-1157.

[38] D. Yang, A. Lavie, C. Dyer, E. Hovy, Humor recognition and humor anchor extraction, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, p. 2367–2376. URL: https://www.aclweb.org/anthology/D15-1284. doi:10.18653/v1/D15-1284.

[39] A. Reyes, D. Buscaldi, P. Rosso, An analysis of the impact of ambiguity on automatic humour recognition, in: V. Matoušek, P. Mautner (Eds.), Text, Speech and Dialogue, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2009, pp. 162–169. doi:10.1007/978-3-642-04208-9_25.

[40] V. Blinov, V. Bolotova-Baranova, P. Braslavski, Large dataset and language model fun-tuning for humor recognition, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 4027–4032. doi:10.18653/v1/P19-1394.

[41] A. Ermilov, N. Murashkina, V. Goryacheva, P. Braslavski, Stierlitz meets SVM: Humor detection in Russian, in: D. Ustalov, A. Filchenkov, L. Pivovarova, J. Žižka (Eds.), Artificial Intelligence and Natural Language: 7th International Conference, AINL 2018, volume 930 of *Communications in Computer and Information Science*, Springer, Cham, Switzerland, 2018, pp. 178–184. doi:10.1007/978-3-030-01204-5_17.

[42] S. Castro, L. Chiruzzo, A. Rosá, D. Garat, G. Moncecchi, A crowd-annotated Spanish corpus for humor analysis, in: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, 2018, p. 7–11. URL: https://www.aclweb.org/anthology/W18-3502. doi:10.18653/v1/W18-3502.

[43] L. Ermakova, T. Miller, F. Regattin, A.-G. Bosser, E. Mathurin, G. L. Corre, S. Araújo, J. Boccou, A. Digue, A. Damoy, B. Jeanjean, Overview of JOKER@CLEF 2022: Automatic Wordplay and Humour Translation workshop, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), volume 13390 of *LNCS*, 2022.

[44] O. Lieber, O. Sharir, B. Lentz, Y. Shoham, Jurassic-1: Technical Details and Evaluation, White paper, AI21 Labs, 2021. URL: https://uploads-ssl.webflow.com/60fd4503684b466578c0d307/61138924626a6981ee09caf6_jurassic_tech_paper.pdf.

[45] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2021, p. 483–498. URL: https://aclanthology.org/2021.naacl-main.

41. doi:10.18653/v1/2021.naacl-main.41.
[46] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, volume 1, Association for Computational Linguistics, 2019, p. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.
[47] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.
[48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models Are Unsupervised Multitask Learners, Technical report, OpenAI, 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv:1706.03762 [cs] (2017). URL: http://arxiv.org/abs/1706.03762.
[50] C. Jiang, M. Maddela, W. Lan, Y. Zhong, W. Xu, Neural crf model for sentence alignment in text simplification, arXiv:2005.02324 [cs] (2020). URL: http://arxiv.org/abs/2005.02324.
[51] D. Delabastita, There's a Double Tongue: an Investigation into the Translation of Shakespeare's Wordplay, with Special Reference to Hamlet, Rodopi, Amsterdam, 1993.
[52] P. Vrticka, J. M. Black, A. L. Reiss, The neural basis of humour processing, Nature Reviews Neuroscience 14 (2013) 860–868. doi:10.1038/nrn3566.
[53] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998.
[54] L. Ermakova, F. Regattin, T. Miller, A.-G. Bosser, S. Araújo, C. Borg, G. L. Corre, J. Boccou, A. Digue, A. Damoy, P. Campen, O. Puchalski, Overview of the CLEF 2022 JOKER Task 1: Classify and explain instances of wordplay, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2022.
[55] F. Dhanani, M. Rafi, M. A. Tahir, FAST_MT participation for the JOKER CLEF-2022 automatic pun and human translation tasks, in: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022, p. 14.
[56] L. Glemarec, Use of SimpleT5 for the CLEF workshop JokeR: Automatic Pun and Humor Translation, in: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to

8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022, p. 11.

[57] L. J. G. Galeano, LJGG @ CLEF JOKER Task 3: An improved solution joining with dataset from task, in: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022, p. 7.

[58] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[59] J. Tiedemann, S. Thottingal, OPUS-MT – building open translation services for the world, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, 2020.

[60] S. Zhao, R. Meng, D. He, A. Saptono, B. Parmanto, Integrating Transformer and Paraphrase Rules for Sentence Simplification, in: Proc. of EMNLP 2018, ACL, Brussels, Belgium, 2018, pp. 3164–3173. URL: https://www.aclweb.org/anthology/D18-1355.

[61] L. Glémarec, A.-G. Bosser, L. Ermakova, Generating Humourous Puns in French, in: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022, p. 8.

[62] G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2022.