Highly Parallel Texts Enriched with Highly Useful Metadata? A Wikipedia Case-Study Combining Machine Learning and Social Technology

Ahmad Aghaebrahimian, Andy Stauder and Michael Ustaszewski

Department of Translation Studies, University of Innsbruck, Innsbruck, Austria

This is an un-refereed preprint version of an article accepted for publication in *Digital Scholarship in the Humanities*. The refereed, revised and copyedited version of record "Ahmad Aghaebrahimian, Andy Stauder, Michael Ustaszewski: Automatically extracted parallel corpora enriched with highly useful metadata? A Wikipedia case study combining machine learning and social technology. In: *Digtal Scholarship in the Humanities*, DOI: fqaa002" is available online at: https://doi.org/10.1093/llc/fqaa002

Full correspondence details:

Dr. Michael Ustaszewski

Universität Innsbruck

Institut für Translationswissenschaft

Herzog-Siegmund-Ufer 15

A-6020 Innsbruck | Austria

Telefon +43 512 507 42482

E-Mail michael.ustaszewski@uibk.ac.at

Web www.uibk.ac.at/translation

Highly Parallel Texts Enriched with Highly Useful Metadata? A Wikipedia Case-Study Combining Machine Learning and Social Technology

Abstract

The extraction of large amounts of multilingual parallel text from web resources is a widely used technique in natural language processing. However, automatically collected parallel texts usually lack precise metadata, which are crucial to accurate data analysis and interpretation. The combination of automated extraction procedures and manual metadata enrichment may help address this issue. Wikipedia is a promising candidate for the exploration of the potential of said combination of methods, because it is a rich source of translations in a large number of language pairs and because its open and collaborative nature makes it possible to identify and contact the users who produce translations.

This article tests to what extent translated texts automatically extracted from Wikipedia by means of neural networks can be enriched with pertinent metadata through a self-submission-based user survey. Special emphasis is placed on data usefulness, defined in terms of a catalogue of previously established assessment criteria, most prominently metadata quality. The results suggest that from a quantitative perspective the proposed methodology is capable of capturing metadata otherwise not available. At the same time, the crowd-based collection of data and metadata may face important technical and social limitations.

1 Introduction

Parallel corpora are an indispensable resource for a wide range of language-related research and engineering problems. The availability of parallel corpora is limited, however, especially for less-common language pairs or certain subject domains. The extension of existing and compilation of new parallel corpora is therefore a prerequisite to fully reaping the benefits of the increasingly sophisticated and powerful data-driven approaches, be it in the humanities or in natural language processing (NLP).

Spurred by a high demand for multilingual training data, most notably in the field of machine translation, the automatic extraction of parallel text from existing multilingual websites has received considerable attention in NLP as a more time and cost-efficient alternative to the manual compilation of parallel corpora from the ground up. The extracted data are mainly used for domain adaptation of NLP tools and systems and to alleviate data bottlenecks faced by less-resourced languages. While automatic harvesting procedures usually yield large amounts of parallel data, they lack metadata that precisely describe the extracted parallel texts or text fragments, mainly because the mined websites do not make such metadata available. Metadata are a key component of knowledge discovery, prediction, and decision making based on (big) data analytics (Greenberg, 2017). Metadata are vital for accurate data interpretation by both humans and machines (ibid.), which was also confirmed from the perspective of digital humanities in an analysis of the importance and, conversely, the impact of a lack of metadata, highlighting that corpus size cannot make up for insufficient or inexistent metadata (Koplenig, 2017).

The present paper addresses the lack of metadata in automatically harvested data. Based on the use case of Wikipedia, we test to what extent translated texts automatically extracted using machine learning can be enriched with metadata of high usefulness, collected with the help of social technology. The proposed methodology leverages the wealth of openly available administrative metadata in a collaboratively built knowledge resource in combination with human intervention to obtain value-added data: highly parallel text data with pertinent highly useful metadata. The extracted data are analyzed as to whether they are compliant with the principles of what Greenberg (2017) calls *smart metadata* (see Section 3). The choice of using Wikipedia as a data source is motivated by its size, multilinguality, diverse topics, openness, dynamic growth, and transparency. It contains large amounts of CC-licensed material, unproblematic in terms of rights and data protection issues, and it enables interaction with data producers. Yet, Wikipedia translations have largely been neglected by research (Jones, 2018; Shuttleworth, 2018), at least in fields other than NLP (see Section 2). This may be due to the non-trivial assessment of the extent and exact location of translated material in Wikipedia, as well as to the

closely related difficulty of pinpointing fully parallel text pairs among interlanguage-linked articles that may evolve independently from each other over time (Shuttleworth, 2018). The present study addresses these difficulties in order to tap into Wikipedia as a multilingual translation resource. Unlike many automatic harvesting approaches documented in the literature, this study aims to extract fully parallel source-target text pairs at the document level rather than isolated sentence pairs. The methodology was tested within the scope of a collaborative, open-access, open-ended, open-domain corpus-building initiative in the area of translation that aims to make high-quality data freely available for reuse (Ustaszewski and Stauder, 2017), but the proposed combination of automatic extraction and enrichment by humans might be of relevance to other mono- and multilingual data collection projects as well.

2 Related Work

Multilingual websites hold great potential for parallel data harvesting, but it must be borne in mind that multilingual content can exhibit varying degrees of parallelity across languages. The following parallelity levels can be distinguished (Babych *et al.*, 2019; Sharoff *et al.*, 2013a):

- parallel: translations, i.e. source-target text pairs that can be aligned at sentence and phrase level
- strongly comparable: texts that can be aligned at the document level, such as heavily edited translations and their source texts, or independently produced but closely related texts about the same subject
- weakly comparable: texts from the same domain or genre but about different sub-topics; collections of such texts can usually only be aligned at sub-corpus rather than document level
- non-comparable, or unrelated: texts that cannot be aligned across languages

A slightly different typology was suggested by Fung and Cheung (2004) who distinguish parallel, noisy parallel, comparable, and very-non-parallel corpora. The comparable corpus type widely used in contrastive and translation studies (Zanettin, 2012) mostly comprises *weakly comparable* materials according to the above typology. The object of this study, Wikipedia, makes multilingual articles on the

same topic easily accessible through so called *interlanguage links*[1]. The degree of parallelity of inter-language-linked articles may vary widely (Babych *et al.*, 2019), and the exact amount of parallel and comparable data in Wikipedia is unknown (O'Hagan, 2016). In a manual evaluation of 200 randomly sampled French-English article pairs from 2009, the proportions of *parallel*, *noisy parallel*, *comparable*, and *very non-parallel* texts were 14%, 11%, 29%, and 46%, respectively (Patry and Langlais, 2011), but the authors pointed out the difficulty of drawing clear-cut boundaries between the four degrees of comparability. In a more recent human evaluation study with eight language pairs, including seven less-resourced ones, 52.5% of interlanguage-linked document pairs were judged to be highly similar and 18.8% to be different on a five-point Likert scale, confirming that Wikipedia articles in different languages on the same topic are not necessarily (very) similar (Babych *et al.*, 2019). The study also elicited factors influencing the perceived parallelity of document pairs, indicating that pairs judged to be similar have similar structure, overlapping named entities, alignable fragments and contain translation equivalents. Altogether, these quantitative figures seem to confirm that while there are coordinated Wikipedia translation initiatives, translation is mainly self-motivated and only one of several mechanisms for the multilingual expansion of Wikipedia (Shuttleworth, 2018). Moreover, Wikipedia content is dynamic, which means that articles between which a translation relation obtained at a certain point in time may evolve independently from each other. Nevertheless, the sheer size of Wikipedia makes it a promising and valuable source of parallel text data, as will be discussed below.

Since comparable multilingual data is much more abundant on the web than strictly parallel data, the compilation and exploitation of comparable corpora as a means to compensate for the lack of parallel corpora has become a prolific line of research (Sharoff *et al.*, 2013b; Skadiņa *et al.*, 2019). Along this line, the identification and extraction of parallel sentences from comparable corpora has played an important role, most notably from Wikipedia and with the purpose of improving the performance of statistical machine translation systems (e.g. Barrón-Cedeño *et al.*, 2015; Labaka *et al.*, 2016; Smith

[1] https://en.wikipedia.org/w/index.php?title=Help:Interlanguage_links&oldid=885377785

et al., 2010). While most approaches are based on comparable corpora, Fung and Cheung (2004) exploited non-comparable corpora for this task. Ture and Lin (2012), on the other hand, aim to maximize the recall of extracted sentences; their system determines the degree of comparability between inter-language-linked Wikipedia article pairs using a more exhaustive and computationally more costly approach instead of relying on heuristics only, thus yielding 5.8 million English-German sentence pairs. More recently, neural networks have been used for parallel sentence extraction (e.g. Chu *et al.*, 2016; España-Bonet *et al.*, 2017; Grégoire and Langlais, 2018) – an approach that has been adopted in the present study, too (see Section 4). Web resources other than Wikipedia are also being used for parallel sentence extraction, for instance in a system that crawls the web to detect entry points to multilingual websites and subsequently uses intra-site crawlers and alignment procedures (Barbosa *et al.*, 2012).

However, there is also a good deal of research that focuses on identifying text data that exhibit high degrees of parallelity at the document level, as opposed to extracting isolated sentence pairs. Such approaches consist in finding interlingually corresponding text pairs from bi- or multilingual collections of candidate documents, for example Wikipedia (e.g. Enright and Kondrak, 2007; Etchegoyhen and Azpeitia, 2016; Mohammadi, 2016; Morin *et al.*, 2015; Patry and Langlais, 2011). They offer algorithmic alternatives to the manual identification of translated Wikipedia articles based on the information contained in Wikipedia's administrative pages and the structural mark-up assigned by editors to internally organize the content of the encyclopaedia (Shuttleworth, 2018). As mentioned in Section 1, the present study is in the vein of these works that aim at extracting fully parallel source-target text pairs, or bitexts, at the document level. However, it also aims to enrich the extracted bitexts with metadata provided by the producers of the respective texts.

3 Data Usefulness

The goal of the present study is to obtain parallel text data from Wikipedia that fulfil two criteria of usefulness. First, the extracted bitexts need to have the highest degree of interlingual document-level parallelity according to the typologies reviewed in Section 2. This means that text pairs are to share the same content, which, for the sake of practicality, is operationalized in terms of alignability at the

sentence or phrase levels (Babych *et al.*, 2019). For the purpose of this study, sentence alignment does not need to be in a bijective (i.e. one-to-one), monotonic (i.e. non-crossing) relation, which is the case when both sides of a bitext are structured strictly in the same order such that the n^{th} segment of the source side corresponds to the n^{th} segment on the target side (Tiedemann, 2011). Thus, cross-alignments (n^{th} source segment corresponds to m^{th} target segment, $n \neq m$) as well as one-to-many, many-to-one and many-to-many alignments (n source segments correspond to m target segments, $n \neq m$) are permitted. The decision to loosen the bijectivity and monotonicity constraints mirrors the fact that (conscious) alterations of text segmentation (e.g. splitting one long source sentence into several shorter target sentences) or structure (e.g. changing the order of semantic elements) are common types of so-called shifts (Gambier, 2010) in real-world translations. Similarly, as a result of translation shifts a given piece of information may have no counterpart in either the target text (omission of source text information) or source text (addition of explanatory information in the target text). Therefore, the presence of n-to-zero (omission) or zero-to-n (addition) relations is permitted in the extracted bitexts, provided the proportion of sentence pairs exhibiting such relations does not exceed an arbitrarily chosen threshold of 5%.

The second criterion is concerned with the usefulness of the metadata with which the extracted bitexts are enriched. Greenberg (2017) defines five principles of smart metadata that help provide context and meaning for data: good quality, accessibility, trust, actionability, and preservation. The collection of metadata within the present study aims to comply with the first three principles, whereas the remaining two principles are of more concern to the storage and (re-)use of the collected data and metadata. The following paragraphs describe how these three principles are reflected in this study's metadata collection, which complements the automatic extraction of parallel texts from Wikipedia by means of a dedicated self-submission web interface that invites the producers of their respective texts to interactively contribute pertinent metadata (see Section 4.2).

Principle 1: Good quality

Based on Bruce and Hillmann (2004), Greenberg (2017) lists five indicators of good quality metadata: accuracy, completeness, conformance to expectations, logical consistency and coherence, and timeliness. Each of these is discussed in the following.

To ensure accuracy of the collected metadata, metadata contributors are guided through the submission process by means of a lean, easy-to-use web interface. Instructions and tooltips in plain English are available throughout the interface. Instead of free text fields, dropdowns – each one with few, clear-cut choices – are used in the interface for the sake of consistency and manageability. In addition, trained staff are to verify and curate the submitted metadata.

Completeness is to be ensured with the help of a set of metadata labels that capture the relation between translated texts and their originals, including the circumstances under which they were produced. The label set strikes a middle ground between maximum detail and economic feasibility. The label set used in the web interface is limited to text-extrinsic information known to data producers only (e.g. translator's age at time of writing) and kept to a manageable size to minimize submission effort and maximize completion. Text-intrinsic metadata (e.g. subject matter) are to be complemented by trained project staff at a later stage in the project, of which this study is a part. To mitigate the legally motivated lack of mandatory fields, the interface at least encourages contributors to make complete metadata submissions with visual cues.

To meet the criterion of conformance to (user) expectations, the set of translation-specific metadata labels is, in part, a combination of labels used by existing translation corpora. Translation corpora are, as a rule, mostly narrow in focus, which makes necessary the aforementioned combination in order to cater to a wide variety of translation audiences and stakeholders (Ustaszewski and Stauder, 2017). So, the label set aims to make findable what a large community would reasonably *expect* to find in a translation corpus (c.f. the FAIR Data Principles, Wilkinson *et al.*, 2016).

The criterion of logical consistency and coherence has two facets: an internal one, which in our case boils down to the use of controlled vocabulary for the respective metadata fields, and an external one, which results from the fact that the developed metadata label set combines features from a number

of well-established repositories in the field of translation. Those postulating the latter facet (Bruce and Hillmann, 2004) seem to perceive communities (such as the translation community) as large systems which should use common standards, making their mode of operation coherent and consistent. The internal facet – keeping the labels used by *one* corpus internally consistent and coherent – can reasonably be assumed to be more easy to achieve due to the smaller number of people involved. Whether the latter facet can truly be regarded as coherence and consistency, or rather interdependence of several systems, depends on what one views as a whole.

Lastly, the criterion of timeliness consists, on the one hand, in having up-to-date metadata and, on the other, not publishing data that haven't been labelled with meaningful metadata yet. Due to the nature of the present study, which collects data of objects that cannot change (i.e. one-time snapshots of Wikipedia articles), the currency aspect cannot be applied here. As for the second criterion: publishing only data that already have metadata, this is to be satisfied by not making available the data as long as the pertinent metadata have not been collected, verified and complemented by trained staff.

Principle 2: Accessibility

On the technical side, the collected metadata are to be stored alongside the extracted parallel texts and made available through a still-under-construction faceted search interface which enables users to compile and download corpora on-demand, tailored to their specific needs. On the legal side, both the extracted texts and collected metadata are to be made freely available under either the CC-BY-SA license or a permissive license specially drafted for the purpose of this study by experts on legal aspects in the field of language data. The required permissions are obtained from the (meta-)data contributors through the metadata collection interface, where they can select licenses as part of the specially

drafted contributor agreement[2] and specify the desired degree of anonymity, which ensures conformity with the repository's privacy policy[3]. Hence, legal clearance and transparency as a prerequisite to accessibility is of paramount importance to metadata usefulness, especially in the context of open-access corpora.

Principle 3: Trust

Good quality metadata are trusted metadata and produced by reliable sources (Greenberg, 2017). In the present study, this means that the text producers themselves, i.e. translators of Wikipedia articles, contribute metadata through a self-submission interface. The underlying assumption is that the volunteer community of Wikipedians is open to the idea of sharing (meta-)data and contributing to a collaborative open-science initiative. No less important, collected metadata are to undergo revision by trained project staff prior to integration into the open-source repository.

4 Method

4.1 Automated Parallel Data Extraction

The extraction of bitexts from Wikipedia requires locating candidate text pairs, which was accomplished with the help of Wikipedia interlanguage links, which link corresponding articles on the same topic across languages, e.g. the English article on 'water' with the German article on 'Wasser'. When someone creates a version of an article in a different language, it is common practice to create an interlanguage link between the original article and the new text in the other language, by embedding a so-called *Interlanguage link template*[4] in the latter (Wikipedia 2019). All of the interlanguage links

[2] https://transbank.info/contributor-agreement [anonymized for review]

[3] https://transbank.info/privacy [anonymized for review]

[4] https://en.wikipedia.org/wiki/Template:Interlanguage_link

are stored in a dedicated SQL database[5], exported by Wikipedia at regular intervals. On Wikipedia, the interlanguage links appear in a sidebar on the left. We downloaded the SQL database from March 2018 to extract all interlanguage links. The links available in the database are unique identifiers, each pointing to two associated article revisions in different languages. As an example and to provide a better idea of the numbers involved, **Table 1** lists the respective numbers of articles in the top 13 languages that have been linked to respective English versions.

Table 1 Number of interlanguage-linked Wikipedia article pairs for English (March 2018)

Language Pair	# interlanguage-linked comparable article pairs
English-French	1,491,578
English-German	1,247,102
English-Italian	1,123,058
English-Spanish	1,096,328
English-Russian	975,983
English-Swedish	918,314
English-Dutch	906,950
English-Polish	906,105
English-Portuguese	900,508
English-Farsi	866,408
English-Chinese	703,217
English-Arabic	643,360
English-Japanese	631,855

The article revisions (=versions) to which the links in the aforementioned SQL database point are contained in database dumps for each language version of Wikipedia that are made available twice a year. These database dumps are the data source in which we searched for text pairs that constitute respective originals and translations. This is done by identifying pairs of article revisions with at least 95% of parallel sentences, as was described in Section 3. We used Wikipedia dump files (June 2018) of languages contained in the total set of language pairs from the SQL database. The text pairs could also be retrieved through the MediaWiki Action API[6], but for reasons of performance and ease of use we chose to download Wikipedia dumps and process them directly. Retrieving text pairs from data dumps

[5] https://dumps.wikimedia.org/wikidatawiki/latest/wikidatawiki-latest-langlinks.sql.gz

[6] https://www.mediawiki.org/wiki/API

makes it easier to browse different revisions of the same article, which is an essential step in filtering parallel data from comparable data. The reason for this is that due to the dynamic and collaborative nature of Wikipedia, individual articles may undergo editing by a number of different users. Therefore, although the interlanguage links provide information as to which pages are comparable, the degree of comparability of these pages may change over time due to changes made independently from other language versions.

The next step was retrieving parallel texts from the set of texts that had been identified as comparable with the help of interlanguage links. For this purpose we developed a deep neural network architecture to identify fully parallel texts among the candidate pairs. This approach offers a language-independent, robust and highly scalable state-of-the-art solution to parallel sentence extraction (Aghaebrahimian, 2018). As has been mentioned in Section 3, the criterion for identifying parallel texts was that they contained at least 95% of sentences that had been identified as parallel. The threshold used by the neural network methodology for deciding whether two sentences were parallel or not was determined with the help of human evaluation: human raters were presented sentence pairs from the Europarl corpus (Koehn 2005) and had to decide whether the two sentences of each pair were parallel or not (Aghaebrahimian, 2018).

4.2 Metadata Enrichment

4.2.1 Identification of Translators

For each extracted Wikipedia translation, we identified the user who had written it with the help of the pointers from the SQL database article's revision history. In this way we could retrieve the name of each user who has used an interlanguage link template[7] for linking a revision written by him/her to the corresponding Wikipedia page in another language. Material of users that had the email feature of Wikipedia turned off, and thus could not be contacted, was left out. Also, "translations" that were

[7] https://en.wikipedia.org/wiki/Template:Interlanguage_link

obviously unusable, e.g. because they consisted only of a title and had no text body[8], were filtered out. Each one of the users has a unique identifier through which they can be contacted via Wikipedia's internal mailing system, as long as they have activated this feature for their respective Wikipedia account. When users create a translation of a Wikipedia article, and if they follow Wikipedia's recommendation, they add an interlanguage link template containing their own username to the newly created revision. If someone else decides to modify this translation or add another translation for the same source page, another revision is created which contains another interlanguage link template. So in this way extracting the name of the translators is deterministically feasible.

Users identified as translators of the extracted Wikipedia articles were contacted manually via Wikipedia's built-in email facility in order to invite them to provide metadata about themselves and their translations using an on-line interface designed by us for the purpose (see Section 4.2.2). We compiled a list of translators' user IDs, taken from the interlanguage-link SQL database mentioned in Section 4.1. The list was compiled as a one-to-many list, meaning that it maps each user ID to the titles of the revisions constituting their translations. To make sure that the users could only access their own texts, we generated user-specific log-in names and passwords using which they could log into their personalized page in our interface.

Contacting the users via email is a time-consuming process since the Wikipedia email system does not allow sending more than ten emails per 24 hours per IP address to prevent spamming through the system. Moreover, as has been mentioned, sending emails is only possible to users who activated this for their account. Due to all these limitations, the invitation emails containing the credentials for the metadata submission interface were sent manually from five Wikipedia accounts. 934 Wikipedia users were contacted in this way. A more detailed account of the process follows in the next subsection.

---

[8] Some users apparently create interlanguage-linked pages and only translate the respective title of an article, maybe to get the translation process going. We did not research this phenomenon. It should only be noted that we filtered out such empty articles, because they could easily be identified as non-translations outright.

4.2.2 Metadata Submission Interface

The contacted users were invited to contribute metadata about themselves and the extracted texts through the metadata submission interface, for which they were given personalized user accounts, each of which was associated with all the extracted texts produced by the respective user. The interface consisted of four components, through which contributing users were taken sequentially: (1) a login page for Wikipedia translators; (2) an instructions page, including opt-in checkboxes for reading and accepting the project's terms of service and privacy policy; (3) a form for entering basic personal information, including choices of the preferred degree of anonymity; (4) a form for entering text-related metadata, displayed on separate pages for each text produced by a user. While Component 3 collected immutable personal data (native language, gender), Component 4 collected personal information relative to a given text (e.g. age or education level at the time of text production) and information on the text production circumstances (e.g. type of translation tools used), as shown in **Fig. 1**.

The contribution of metadata was not remunerated; instead, the contacted Wikipedians were encouraged to contribute to a collaborative open science initiative. The contributors were given control as to their choice of license, and full transparency regarding the project's legal framework was a major concern in the design of the interface.

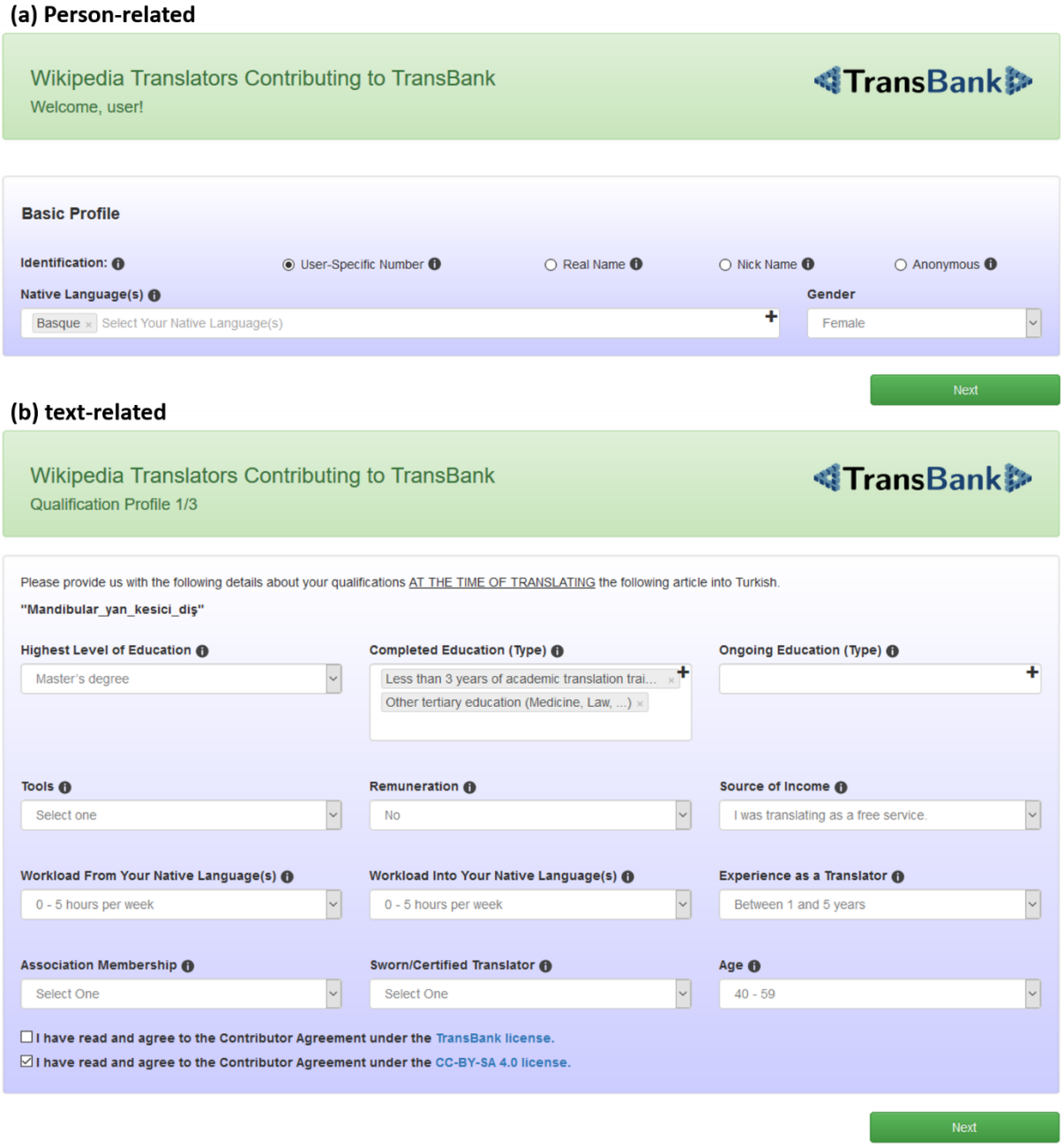

Fig. 1 Screenshot of person-related (a) and text-related metadata (b) submission form for Wikipedia translators

5 Results and discussion

We extracted an arbitrary quantity of bitexts containing about 50,000,000 tokens from Wikipedia, based on the information on which text was connected to which via an interlanguage link. From this material, about 4,800 parallel bitexts texts could be extracted using the described neural-network based translation identification methodology. The translators of 3,104 of these bitexts had the email functionality of their Wikipedia accounts enabled and could thus be contacted, in principle. In total,

invitations to contribute metadata were sent to 934 translators – some translators had produced more than one text, therefore the numbers of bitexts and translators are not the same. Due to limitations we already mentioned about the number of emails that can be sent per 24 hours per IP address, we used five Wikipedia accounts to send the emails containing the log-in information to translators. The procedure described in Section 4.2 took almost a month but made the email distribution possible.

Of the 934 translators that were contacted, a total of 216 logged into our interface and provided us with the metadata for each of their articles. As they had been identified as translators, every one of them had translated at least one article, the most productive user, however, had written 136 article translations, with more than 38,000 contributions to Wikipedia in general. On average, users that had provided metadata through the interface had translated 4.1 articles each (Mdn = 1.0, SD = 10.5).

We received metadata for 881 different bitexts and 44 different language pairs, with Russian-Ukrainian being the most numerous one (191 bitexts), 15 language-pairs yielding only one article pair each (see **Table 2**). It must be stated at this point that we do not know if there are language pairs with even more texts as this is not entirely clear from the SQL database, because, firstly, the database contains this information in an unstructured way and we stopped parsing it after we had reached our arbitrarily chosen number of texts, and, secondly it is not clear how many users follow the recommendation of including an interlanguage-link template when creating a translation. As far as our data are concerned, on average, 20.0 bitexts (SD = 39.4, Mdn = 3.5) were observed for each language pair. On average there were 5.8 different translators per language pair (Mdn = 2.0, SD = 8.8). The average number of texts per translator varied greatly across language pairs: the most productive translators were observed for Spanish-Asturian (35.5 texts per translator), French-Portuguese (8.0) and Russian-Ukrainian (6.4), while the mean number of translations per translator averaged across all language pairs was 2.8 (Mdn = 1.4, SD = 5.3).

Table 2 Quantitative summary of extracted and metadata-enriched bitexts by language pair

Language Pair	# bitexts	# translators	Word count source language	Word count target language	Language Pair	# bitexts	# translators	Word count source language	Word count target language
Russian – Ukrainian	191	30	182682	181521	English – Russian	3	1	1968	1721
Spanish – Asturian	142	4	214420	200875	French – Romanian	3	3	1573	1667
English – Spanish	121	44	171063	187488	Spanish – Italian	2	1	4061	3791
English – French	64	24	98232	108640	Italian – French	2	1	1552	1618
Spanish – Catalan	41	18	53420	53307	French – Catalan	2	2	2984	3185
English - Portuguese	41	21	68489	72400	Belarusian – Ukrainian	2	2	1269	1265
Spanish – Galician	36	9	49751	46975	German – Portuguese	1	2	1308	1405
English – Romanian	35	6	67275	70974	Portuguese – Galician	1	1	3509	3391
English – Greek	26	5	27771	28758	English – Norwegian	1	1	473	415
Ukrainian – Russian	24	9	13804	13817	English – Asturian	1	1	2701	2696
French – Portuguese	24	3	20839	20252	French – Italian	1	1	470	405
English – Catalan	23	12	49968	54294	Portuguese – French	1	1	2137	2558
English – Italian	15	6	17128	17658	Spanish – French	1	1	3769	3379
Russian – Belarusian	15	4	10864	10911	Italian – English	1	1	681	685
Catalan – Spanish	11	6	9853	9854	German – Romanian	1	1	1678	1832
French – Spanish	11	8	13346	13476	Bulgarian – Catalan	1	1	1779	2109
English – Ukrainian	7	5	8921	7283	German – French	1	1	2823	3265
Spanish – Portuguese	6	4	5841	5535	English – Turkish	1	1	122	110
French – English	4	3	5673	5247	Romanian – Catalan	1	1	3599	4112
German – English	4	2	3460	3472	French – Dutch	1	1	404	459
Russian – English	4	1	3985	4760	Belarusian – Russian	1	1	1357	1347
Portuguese – Spanish	4	3	8464	8892	TOTAL	881	-	1,147,394	1,169,826
Swedish – Norwegian	3	1	1928	2022	MEAN/PAIR	20.0	5.8	26,077.1	26,587.0

Through the metadata submission interface described in Section 4.2 we collected two items of person-related and eleven items of text-related metadata. The former were considered immutable and therefore users were prompted only once to submit them via Component 3 of the interface, whereas Component 4 for the collection of text-related metadata was displayed for every single text produced by a given user. Since in Section 3 completeness and accuracy were identified as two fundamental

criteria of metadata quality and hence of data usefulness, the viability of the proposed data collection methodology needs to be assessed in terms of these two criteria.

Metadata completeness

The mean submission rates for the eleven text-related and two person-related metadata fields of the submission interface are reported in **Table 3**. Due to technical issues, the data for one of the fields (the tools field, where participants could state whether they used electronic translation tools such as translation memory systems) were not collected correctly and were therefore excluded from the analysis. The two items of immutable person-related metadata, native language and gender, which can arguably be considered more prone to contributors' data privacy concerns, were provided by 100% and 88.4% of users, respectively. For 287 out of 881 texts (35.6%), users provided all of the requested metadata. Summing up, 216 out of 934 contacted Wikipedia translators (23.1%) submitted metadata, thus providing metadata-enrichment for 881 of the 3104 (28.4%) extracted parallel texts for which we were able to request metadata from their translators. 287 out of these 3104 texts (9.2%) were fully annotated by the translators with both person-related and text-related metadata. At the text level, the overall completeness rate for the 881 metadata-enriched texts is 92.2%.

Table 3 Degree of completeness of filling in metadata fields: e.g., "83.4%" means a field has been filled in 83.4% of the time.

Category	# fields	N	Min (least filled-in field)	Mdn	Mean	SD	Max (most filled-in field)
text-related	11	881 texts	83.4%	88.1%	91.1%	7.3	100% (4 fields)
person-related	2	216 users	88.4%	94.2%	94.2%	8.2	100% (1 field)

The reported completeness statistics have to be seen in the light of the voluntary nature of user submissions and the legally motivated absence of compulsory metadata fields in the submission interface. For the users who contributed less than 10 texts, the submission rate was 91.4%. Among the 19 users who contributed ten texts or more – their contributions account for 461 (52.3%) of all texts – the

submission rate for text-related metadata was 90.9%, which can be seen as an indicator of commitment. However, since we cannot entirely rule out the possibility that parts of our invitations were not correctly received by the addressees due to the anti-spam filters in the Wikipedia messaging system, we cannot be sure whether the response rates and completeness statistics adequately mirror Wikipedians' willingness to contribute to open science initiatives.

Metadata accuracy

In the context of the present study metadata accuracy is closely related to the quality criterion of trust, because the metadata collection was limited to text-extrinsic information known to data producers only. We used information publicly available on Wikipedia user pages[9] as a proxy to assess the reliability of submitted metadata. To this end, we randomly sampled 32 from the 216 contributing users (14.8%) and manually compared the metadata submitted by them with their user pages, focusing on gender and native language. Users who left either field empty (25 out of 216 users = 11.6% left the *gender* field empty; the *native language* field, on the other hand, was filled in by 100% of users) were not considered for the random sample, because, firstly, not providing certain metadata is a conscious decision taken by users and in line with the study's legal framework, and, secondly, missing data is a concern of data completeness rather than accuracy and reliability. Only explicit information from the page text, highly likely to have been written by the users themselves (e.g. gender-specific occupations, such as Spanish *escritor* 'male writer' vs. *escritora* 'female writer') or from the so-called userboxes[10] was used to verify users' gender and native language. Less reliable information was ignored, e.g. the grammatical gender of usernames or gender-specific forms of the word *user* in the page header (e.g. Spanish *usuario vs. usuaria* or German *Benutzer* vs. *Benutzerin*) – it is not clear if the grammatical gender actually reflects the gender of the respective user themselves in the case of usernames; also, the

[9] https://en.wikipedia.org/wiki/Wikipedia:User_pages

[10] https://en.wikipedia.org/wiki/Wikipedia:Userboxes

male form of the word meaning *user* is often used generically in many languages. A major limitation to this analysis is that Wikipedia user pages are not standardized and users are entirely free to choose whether, what and how to publish personal data; however, they are the *only* source for cross-checking submitted metadata.

Table 4 shows the confusion matrix for user-submitted vs. manually cross-checked gender metadata. In 18 cases, no gender information was found on the user pages, which means that in most cases user submissions are the only way to collect pertinent metadata. However, in the remaining 14 cases for which gender data was available, the accuracy of submitted vs. cross-checked gender information was merely 0.36 (Kappa agreement = 0.06, $p > 0.05$), suggesting that user-submitted metadata was less reliable than initially assumed. Further evidence for the lacking reliability of submitted metadata is that women account for approximately 34% of all users in both the sample and full dataset, while estimates suggest that less than 10% of Wikipedia editors are women (Ford and Wajcman, 2017). Contradictory gender information was also observed in spot checks for several other users not part of the random sample, including the most productive of all users. It is surprising that nine of the sampled users submitted gender information that contradicted their user pages, since users were free to leave the gender field empty in the submission interface. We can only speculate about the reasons for these contradictions: a lack of intrinsic motivation to contribute to an open science project, or data privacy concerns despite a transparent data privacy policy that contains the option to leave fields empty and to freely choose from various degrees on anonymity. A further source of contradiction may be ascribed to the availability of three rather than two gender options in the metadata submission field (male, female, other), which was to increase the social inclusiveness of the metadata survey. The option 'other', selected by four of the sampled users and by 12.5% of all users, may have been misinterpreted as 'not specified', especially among non-native speakers of English. The reliability of these cases can hardly be assessed, because it is unclear whether users who do not identify themselves with the traditional binary distinction would openly specify their gender identity on their user pages. Note, however,

that there are numerous Wikipedia userbox templates[11] that allow specifying gender identities other than male and female.

Table 4 Confusion matrix of user-submitted vs. cross-checked gender metadata (F = female, M = male, O = other)

Submitted \ Cross-checked	F	M	O	n.a.	Total
F	2	4	0	5	11
M	1	3	0	13	17
O	0	4	0	0	4
Total	3	11	0	18	32

By contrast, the submitted metadata about users' native languages from the random sample was in perfect agreement (accuracy = 1.0, Kappa = 1.0, $p \leq 0.01$) with the information from user pages. Ten of the sampled user pages did not contain native language information, whereas all users indicated their native language in the submission interface. This, again, shows that the survey was the only way to retrieve the metadata. Given the perfect response and agreement rates, it seems that information about the native language is perceived to be less sensitive by users. Altogether, the discrepancies in completeness and accuracy of the two metadata items under investigation suggest that the willingness to provide metadata is dependent on the type of the requested information.

Spot checks revealed that there might be accuracy issues with regard to other metadata fields as well, for instance age. However, these instances were not investigated systematically, because the presented analyses suffice to show that user-submitted data may be inaccurate to a certain extent.

6 Conclusion and Outlook

Large amounts of parallel text data covering a wide range of different languages, including less common language pairs, can be extracted in a fairly straightforward manner from Wikipedia using state-of-the-art neural network approaches. Our methodology could have yielded many more parallel texts, but since the aim of the present study was to evaluate in how far the automatically extracted data can

[11] https://en.wikipedia.org/wiki/Category:Gender_user_templates

be manually enriched, the corpus had to be limited to a manageable size. Also, the availability of disk space and memory to process and analyse the Wikipedia dumps might constitute limiting factors, depending on the available infrastructure. Noteworthy, our study was not restricted to particular language pairs, because we aimed to explore the multilingual potential of parallel text extraction from Wikipedia and to capture the "dark matter" (Shuttleworth, 2017) of Wikipedia translations, i.e. texts whose translational status is not explicitly declared. That said, our approach depends on certain administrative and structural information to select a set of candidate text pairs – information contained in the discussed interlanguage-link database. This also means that the language pairs that ended up in our sample were not arbitrarily chosen by us, but are due to the structure of this database and might therefore not be representative of the characteristics of Wikipedia when it comes to translation. Theoretically, the neural network architecture could mine the entirety of Wikipedia for parallel texts, but such an approach would be too time consuming, computationally too costly and thus impractical. Even with those restrictions in mind, Wikipedia is a rich source of parallel texts in constant growth.

Wikipedia's structural and administrative information not only makes parallel text extraction feasible, it also allows identifying translators. The identification of source text authors was deliberately not tackled in the present study, because the collaborative nature of text production in Wikipedia challenges the traditional understanding of authorship. Attempts to pinpoint individual users as authors of particular texts are thus inevitably mere approximations. Similarly, metadata labelsets traditionally used in corpus linguistics are of limited value to capture collaborative text production in all its complexity.

As far as the metadata enrichment procedure is concerned, important technical and social conclusions can be drawn. On the technical side, Wikipedia's limitations on the messaging system and its anti-spam filter have repercussions on the scalability of metadata enrichment via self-submission, because users can only be contacted manually, which is very time-consuming in view of the daily messaging limits. Alternative ways to engage with the community of Wikipedia translators might circumvent this technical limitation, for instance through a closer collaboration with one of the numerous

translation initiatives and task forces in Wikipedia. On the social side, the usefulness of the collected data did not fully meet the established quality criteria. Although the assessment of the collected metadata was based on a small sample and focused only on a small subset of the metadata fields of interest, it highlighted the need for metadata verification and curation. No less important, solutions to improve metadata completeness and accuracy are crucial to crowd-based data collection initiatives. Social technology has a great potential to foster user motivation and involvement, for instance through the already mentioned closer collaboration with the user community in question, through gamification or user rating systems.

Despite the identified quality issues, the study has shown that considerable quantities of metadata otherwise not available can be collected with the help of social technology. The presented approach tries to bridge the gap between well-established NLP techniques and the needs of research that requires high-quality metadata, including, but not limited to, the digital humanities. In the future, automatically extracted and user-enriched translation data may contribute to work on numerous research questions, such as the improvement of multilingual NLP applications, the study of the multilingual expansion of Wikipedia and the dynamics of cross-lingual and cross-cultural knowledge production, or shed light on peripheral and thus underresearched translation practices, most prominently the novel phenomenon of *massively open translation* (O'Hagan, 2016).

The findings of our study suggest that the combination of automated and manual procedures is a viable approach to collecting humanities data, part of which are translation data. Crowd-based self-submission makes it possible to expand data repositories and thus to extend the scope of research to previously untapped resources. However, such approaches are not without risk and require the careful consideration of questions such as how to ensure user involvement, quality control, and data protection.

Funding

This research is part of the project “TransBank: A Meta-Corpus for Translation Research”, funded by the *go!digital 2.0* programme of the Austrian Academy of Sciences (grant number GD 2016/56).

Acknowledgments

We thank the Wikipedia who voluntarily participated in the metadata collection. The computational results presented have been achieved (in part) using the HPC infrastructure LEO of the University of Innsbruck.

References

Aghaebrahimian, A. (2018). Deep Neural Networks at the Service of Multilingual Parallel Sentence Extraction. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1372–1383). Santa Fe, New Mexico: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/C18-1116

Babych, B., Su, F., Hartley, A., Aker, A., Paramita, M. L., Clough, P., & Gaizauskas, R. (2019). Cross-Language Comparability and Its Applications for MT. In I. Skadiņa, R. Gaizauskas, B. Babych, N. Ljubešić, D. Tufiş, & A. Vasiļjevs (Eds.), *Theory and Applications of Natural Language Processing. Using Comparable Corpora for Under-Resourced Areas of Machine Translation* (Vol. 1866, pp. 13–53). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-99004-0_2

Barbosa, L., Rangarajan Sridhar, V. K., Yarmohammadi, M., & Bangalore, S. (2012). Harvesting Parallel Text in Multiple Languages with Limited Supervision. In M. Kay & C. Boitet (Eds.), *Proceedings of COLING 2012: Technical Papers* (pp. 201–214). Mumbai: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/C12-1013

Barrón-Cedeño, A., España-Bonet, C., Boldoba, J., & Màrquez, L. (2015). A Factory of Comparable Corpora from Wikipedia. In P. Zweigenbaum, S. Sharoff, & R. Rapp (Eds.), *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora* (pp. 3–13). Beijing: Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3402

Bruce, T. R., & Hillmann, D. I. (2004). The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In D. I. Hillmann & E. L. Westbrooks (Eds.), *Metadata in practice* (pp. 238–256). Chicago: American Library Association.

Chu, C., Dabre, R., & Kurohashi, S. (2016). Parallel Sentence Extraction from Comparable Corpora with Neural Network Features. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, . . . S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 2931–2935). Paris: European Language Resources Association.

Enright, J., & Kondrak, G. (2007). A Fast Method for Parallel Document Identification. In C. Sidner, T. Schultz, M. Stone, & C. Zhai (Eds.), *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers* (pp. 29–32). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N07-2008

España-Bonet, C., Varga, Á. C., Barrón-Cedeño, A., & van Genabith, J. (2017). An Empirical Analysis of NMT-Derived Interlingual Embeddings and Their Use in Parallel Sentence Identification. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1340–1350. https://doi.org/10.1109/JSTSP.2017.2764273

Etchegoyhen, T., & Azpeitia, A. (2016). A Portable Method for Parallel and Comparable Document Alignment. *Baltic Journal of Modern Computing*, *4*(2), 243–255. Retrieved from https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/4_2_15_Etchegoyhen.pdf

Ford, H., & Wajcman, J. (2017). 'Anyone can edit', not everyone does: Wikipedia's infrastructure and the gender gap. *Social Studies of Science*, *47*(4), 511–527. https://doi.org/10.1177/0306312717692172

Fung, P., & Cheung, P. (2004). Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *Proceedings of the 2004 Conference on Empirical Methods*

in Natural Language Processing: 25-26 July2004, Barcelona (pp. 57–63). Barcelona: Association for Computational Linguistics.

Gambier, Y. (2010). Translation strategies and tactics. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of Translation Studies. Handbook of Translation Studies: Volume 1* (Vol. 1, pp. 412–418). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/hts.1.tra7

Greenberg, J. (2017). Big Metadata, Smart Metadata, and Metadata Capital: Toward Greater Synergy Between Data Science and Metadata. *Journal of Data and Information Science*, *2*(3), 19–36. https://doi.org/10.1515/jdis-2017-0012

Grégoire, F., & Langlais, P. (2018). Extracting Parallel Sentences with Bidirectional Recurrent Neural Networks to Improve Machine Translation. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1442–1453). Santa Fe, New Mexico: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/C18-1122

Jones, H. (2018). Wikipedia, Translation, and the Collaborative Production of Spatial Knowledge. *Alif: Journal of Comparative Poetics*. (38), 264–297.

Koplenig, A. (2017). The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets—Reconstructing the Composition of the German Corpus in Times of WWII. *Digital Scholarship in the Humanities*, *32*(1), 169-188. https://doi.org/10.1093/llc/fqv037

Labaka, G., Alegria, I., & Sarasola, K. (2016). Domain Adaptation in MT Using Titles in Wikipedia as a Parallel Corpus: Resources and Evaluation. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, . . . S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 2209–2213). Paris: European Language Resources Association. Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/pdf/632_Paper.pdf

Mohammadi, M. (2016). Parallel Document Identification using Zipf's Law. In R. Rapp, P. Zweigenbaum, & S. Sharoff (Eds.), *Proceedings of the 9th Workshop on Building and Using Comparable Corpora* (pp. 21–25). Portorož: Evaluations and Language resources Distribution Agency.

Morin, E., Hazem, A., Boudin, F., & Loginova-Clouet, E. (2015). LINA: Identifying Comparable Documents from Wikipedia. In P. Zweigenbaum, S. Sharoff, & R. Rapp (Eds.), *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora* (pp. 88–91). Beijing: Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3413

O'Hagan, M. (2016). Massively Open Translation: Unpacking the Relationship Between Technology and Translation in the 21st Century. *International Journal of Communication*, *10*, 929–946.

Patry, A., & Langlais, P. (2011). Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia. In P. Zweigenbaum, R. Rapp, & S. Sharoff (Eds.), *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web* (pp. 87–95). Portland: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W11-1212

Sharoff, S., Rapp, R., & Zweigenbaum, P. (2013). Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora. In S. Sharoff, R. Rapp, P. Zweigenbaum, & P. Fung (Eds.), *Building and Using Comparable Corpora* (pp. 1–17). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20128-8_1

Sharoff, S., Rapp, R., Zweigenbaum, P., & Fung, P. (Eds.). (2013). *Building and Using Comparable Corpora*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20128-8

Shuttleworth, M. (2017). Locating foci of translation on Wikipedia. *Translation Spaces*, *6*(2), 310–332. https://doi.org/10.1075/ts.6.2.07shu

Shuttleworth, M. (2018). Translation and the Production of Knowledge in Wikipedia: Chronicling the Assassination of Boris Nemtsov. (38), 231–263.

Skadiņa, I., Gaizauskas, R., Vasiļjevs, A., & Paramita, M. L. (2019). Introduction. In I. Skadiņa, R. Gaizauskas, B. Babych, N. Ljubešić, D. Tufiş, & A. Vasiļjevs (Eds.), *Theory and Applications of Natural Language Processing. Using Comparable Corpora for Under-Resourced Areas of Machine Translation* (Vol. 1866, pp. 1–11). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-99004-0_1

Smith, J. R., Quirk, C., & Toutanova, K. (2010). Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In R. Kaplan, J. Burstein, M. Harper, & G. Penn (Eds.), *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 403–411). Los Angeles: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N10-1063

Tiedemann, J. (2011). *Bitext Alignment*. https://doi.org/10.2200/S00367ED1V01Y201106HLT014. Morgan & Claypool.

Ture, F., & Lin, J. (2012). Why Not Grab a Free Lunch? Mining Large Corpora for Parallel Sentences to Improve Translation Modeling. In E. Fosler-Lussier, E. Riloff, & S. Bangalore (Eds.), *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 626–630). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N12-1079

Ustaszewski, M., & Stauder, A. (2017). TransBank: Metadata as the Missing Link between NLP and Traditional Translation Studies. In I. Temnikova, C. Orasan, G. Corpas Pastor, & S. Vogel (Eds.), *Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology* (pp. 29–35).

Wikipedia. (2019). *Wikipedia:Translation*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:Translation&oldid=888728057. Accessed 24 June 2019

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

603 **Zanettin, F.** (2012). *Translation-driven corpora corpus resources for descriptive and applied transla-*
604 *tion studies*. *Translation practices explained*. Manchester, UK, Kinderhook, NY: St. Jerome Pub.