ParaNames: A Massively Multilingual Entity Name Corpus

Jonne Sälevä and **Constantine Lignos**
Michtom School of Computer Science
Brandeis University
{jonnesaleva,lignos}@brandeis.edu

Abstract

This preprint describes work in progress on ParaNames, a multilingual parallel name resource consisting of names for approximately 14 million entities. The included names span over 400 languages, and almost all entities are mapped to standardized entity types (PER/LOC/ORG). Using Wikidata as a source, we create the largest resource of this type to-date. We describe our approach to filtering and standardizing the data to provide the best quality possible. ParaNames is useful for multilingual language processing, both in defining tasks for name translation/transliteration and as supplementary data for tasks such as named entity recognition and linking. Our resource is released at https://github.com/bltlab/paranames under a Creative Commons license (CC BY 4.0).

1 Introduction

Our goal for ParaNames is to introduce a massively multilingual entity name resource that provides names for entities of diverse types across the largest possible set of languages, is permissively-licensed, and can be kept up to date through an open-source and almost fully automated data filtering and preprocessing procedure.

In building massively multilingual language technology applications, it is often important to know how real-world entities are represented across various languages. These correspondences are not always easy to model, as they can involve a mix of transliteration and translation and often involve inconsistencies across languages or even between names in a given language. As a concrete example, some country names are translated in Finnish, so *United Kingdom* is written as *Yhdistynyt kuningaskunta*, a literal, word-by-word, translation. In contrast, smaller territories may or may not be translated: the U.S. states of *North Carolina* and *New York* are written as *Pohjois-Carolina* (with *North* translated) and *New York*, respectively. Moreover, Finnish versions of the U.S. states are often idiosyncratically translated, e.g. California is represented as *Kalifornia*, whereas Colorado is represented as *Colorado*.

The examples above demonstrate the complex choices that language speakers make in representing named entities—even when only dealing with Latin script—and underscore the need a large-scale, multilingual resources of named entity correspondences to effectively model these phenomena. Addressing this need is difficult. To create a practical and effective resource, we believe our resource should meet the following criteria.

Low manual annotation cost. The resource should not be costly to create, and should be easily updated to cover emerging entities and other additions.

Broad language and domain coverage. The resource should cover as broad a range of languages as possible, especially under-resourced ones.

Standardized scripts. The entities in each language should be written in scripts commonly used by speakers of the language.

As we discuss in Section 2, we are not aware of any existing datasets that meet all of these criteria. To achieve our goals, it makes sense to consider the use of publicly available multilingual resources and automated processing. Wikidata[1] is one such example, and is particularly suited for the task not only because of its broad language and domain coverage but also its nature as a perpetually updating collection which enables continuous improvement, especially for lower-resourced languages. In this paper, we present our approach to transforming a dump of the Wikidata knowledge graph into a dataset of

[1] https://www.wikidata.org

parallel names grouped into three high-level entity types, person, location, and organization.

However, our contribution lies not just in making this resource available. In addition to providing the reproducible process that we use to create the dataset, we also identify potential problems in the data, such as the lack of standardization of the script(s) used in each language, and provide a post-processing pipeline where for each language, we ensure that only the correct scripts are used for the names. In addition to filtering out undesirable scripts, we focus on making the names as parallel as possible by removing extraneous information that can accompany the names.

The following sections describe the characteristics of our dataset and our approach to constructing it. While our goal is to promote ParaNames as a useful resource, we examine the use of Wikidata from a skeptical perspective, pointing out properties that may limit its usefulness.

We plan to provide regular updates to this resource to include corrections and improvements to both Wikidata and our extraction process. The Wikidata names we use as a source are CC0 ("no rights reserved") licensed,[2] and our resource is licensed using the Creative Commons Attribution 4.0 International (CC BY 4.0).

2 Related work

While there is previous work in the construction of multilingual name resources, we are not aware of an *openly-accessible* resource containing the names of *modern* entities in many languages. Wu et al. (2018) create a translation matrix of 1,129 biblical names, with each English name containing translations into up to 591 languages.

The Named Entity Workshop (NEWS) shared task has created parallel name resources across a series of shared tasks. In the 2018 version of the shared task (Chen et al., 2018a,b), participants were asked to transliterate between language pairs involving English, Thai, Persian, Chinese, Vietnamese, Hindi, Tamil, Kannada, Bangla, Hebrew, Japanese (Katakana / Kanji), and Korean (Hangul), although the task did not include transliteration between all pairs. The NEWS 2018 datasets are hand-crafted and much smaller than ours, at most 30k names per language pair. Unlike our resource, the datasets for these shared tasks are not fully publicly available; the test set is held back and the each of the five training sets is subject to different licensing restrictions.

We do not claim to be the first to harvest the parallel entity names available from Wikidata or Wikipedia. There is scattered prior work in this area, with one of the earliest explorations at scale being performed by Irvine et al. (2010). Recently, Benites et al. (2020) used Wikipedia as a data source and automatically extracted potential transliteration pairs, combining their outputs with several previously published corpora into an aggregate corpus of 1.6 million names.

Specifically for lower-resourced languages, many approaches to named entity recognition and linking for the LORELEI program (Strassel and Tracey, 2016) used Wikidata, Wikipedia, DBpedia, GeoNames, and other resources to provide name lists and other information relevant to the languages and regions for which systems were developed. However, while ad-hoc extractions of these resources were integrated into systems, we are unable to identify prior attempts to create a transparent, replicable extraction pipeline and to distribute the extracted resources with wide language coverage.

Merhav and Ash (2018) release bilingual name dictionaries for English and each of Russian, Hebrew, Arabic, and Japanese Katakana. However, their resource is limited to a few languages and only covers single token person names. In contrast, our dataset includes hundreds of languages, entities other than persons, and consists primarily of multi-token entity names.

3 Data extraction

To construct our dataset, we began by extracting all entity records from Wikidata and ingesting them into a MongoDB instance for fast processing. Each entity in Wikidata is associated with several types of metadata, including a set of one or more names that different languages use to refer to. For example, visiting the entity for Alan Turing at `https://www.wikidata.org/wiki/Q7251` will show his name written in over a hundred languages, including many that use non-Latin scripts. Given that we are working with such a large-scale dataset, there are important challenges that arise when working with the data, which we describe in this section.

[2] `https://www.wikidata.org/wiki/Wikidata:Copyright`

3.1 Language representation

Entities vary wildly with regards to how many languages they have labels in, and the Wikimedia language codes used in Wikidata do not correspond one-to-one with natural languages.[3] Often there are several Wikimedia codes for a given spoken language, varying in script or geography. For example, the Kazakh language is associated with the Wikimedia language codes `kk` (Kazakh), `kk-arab` (Kazakh in Arabic script), and `kk-latn` (Kazakh in Latin script).

These language codes can potentially be helpful in learning to transliterate between different scripts of the same language. At other times, the language codes are specific to geography rather than writing system. In the case of Kazakh, there are three main geography-specific language codes: `kk-cn` (Kazakh in China), `kk-kz` (Kazakh in Kazakhstan) and `kk-tr` (Kazakh in Turkey).

In our analysis and the resource itself, we do not include language codes for the small number of languages for which only one name is present among the entities we extracted from Wikidata, as we do not think this constitutes meaningful representation of the language.

3.2 Script usage

While language codes can identify a specific script for a language, unfortunately many Wikidata labels do not conform to the scripts used by each language. In many cases, this is simply a data quality issue, such as with Greek where approximately 8.9% of ORG entities are written in Latin script rather than the Greek alphabet.[4]

However, in other cases, the presence of several scripts can also reflect real world-usage depending on the language, as many languages commonly use several scripts. As an example, Kyrgyz uses both the Cyrillic and Arabic alphabets, thus multiple scripts are to be expected across a collection of names and our resource reflects this diversity.

[3]The relationship between Wikimedia language codes and other language codes is rather complex. Originally, the Wikimedia language codes were designed to comply with RFC3066, but the usage of second subtags (for example, the `us` in `en-us`), does not fully comply with the standard and standardization is unlikely to occur soon. Some, but not all, of the language codes are identical to modern BCP 47 codes (RFC5646). In this paper, we try to distinguish between the Wikimedia language codes—which may identify a language along with a script, geographical region, or dialect—and higher-level language identifiers which we identify using the first two letters of the language code. When we attempt to provide the total number of languages covered, we use the higher-level identifiers to prevent double-counting the same language written using multiple scripts.

[4]We confirmed with a Greek speaker that this represented a data issue and not common variation within the language about how names are written.

Entity type	Count	Percentage
PER	8,897,596	63.48%
LOC	3,098,276	22.10%
ORG	1,649,661	11.77%
LOC + ORG	367,499	2.62%
ORG + PER	3,880	0.03%
LOC + ORG + PER	230	<0.00%
LOC + PER	26	<0.00%
Total	14,017,168	100.0%

Table 1: Number of entities and percentage of all entities assigned to each combination of LOC, ORG and PER in ParaNames.

3.3 Providing entity types

Downstream tasks and analyis of the performance of systems across different types of entities requires that entities have high-level entity types—such as location (LOC), organization (ORG), and person (PER)—which Wikidata does not provide directly due to its complex type hierarchy.

We opted not to perform any type inference for entities at ingest time, but instead chose to extract entity types dynamically when constructing the final name resource. Specifically, we identified suitable high-level Wikidata types—Q5 (human) for PER, Q82794 (geographic region) for LOC, and Q43229 (organization) for ORG—and classified each Wikidata entity that is an instance of these types as the corresponding named entity type.

As shown in Table 1, a relatively small number of entities get assigned to multiple name categories due to multiple-inheritance in the entity type hierarchy of Wikidata. Our resource preserves this information, as assigning only a single type to complex entities could make our dataset less useful by ignoring the inherent uncertainty of entity typing.

A visualization of the name counts for the 100 languages with the most entity labels in Wikidata is shown in Figure 1. The number of entity names in Wikidata varies greatly across languages, and counts are distributed according to a Zipf-like power law distribution where a few languages contain most of the names. As expected, many of the

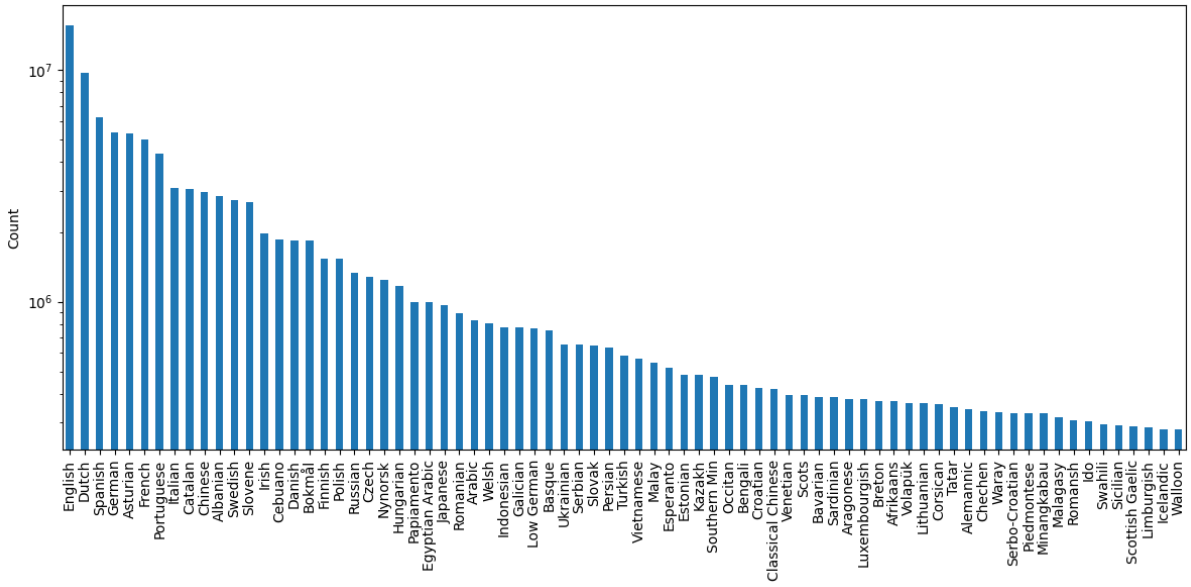

Figure 1: Name counts across the 75 languages with the most names (languages identified by first two letters of Wikimedia language code, $\log_{10}$ y-axis).

largest languages are also large in terms of number of speakers. However, there are notable exceptions, such as Asturian, which contains the fourth largest number of entity names, despite only having fewer than a million native speakers. We suspect that this is an artifact of non-human editing on Wikipedia, and many of these entities appear to be copies of the English name. We discuss this further in Section 5.

The relative proportions of entity types also seem to vary, with LOC entities comprising the bulk of names for most languages. There are exceptions, however. For instance, the number of PER names for Albanian seem to substantially outnumber the LOC entities. For many smaller languages with fewer names, entity types other than LOC are more evenly represented than among the larger languages.

4 Improving data quality

To ensure our resource is of the highest quality possible, we identified two properties that all languages in our corpus should adhere to for maximal usefulness.

First, for each language, all entities in a language should be written in script(s) that match real-world usage. Second, parallel names in our corpus would ideally have the same information on both sides; additional information like titles that appear in one language and not the other should be removed.

4.1 Script standardization

For the first property, we chose to normalize the names for each language by filtering out names that are not in the desired script(s) for the language. An example of this would be a Russian entity name like *Canada* which is not written in Cyrillic.

While we explored automated methods of doing this, ultimately we decided that manually constructing a list of allowed scripts for each language would yield the best results. For each language, we used Wikipedia as an authoritative source to look up which scripts are used to write the language, and filtered out all names whose most common Unicode script property is not among the allowed ones. We used the PyICU library[5] to identify the most frequent Unicode script tag in each name.

To quantify how much this filtering changed the entity names associated with each language, we attempted to measure script uniformity for each language. For each language, we aggregated the Unicode script tags produced by PyICU across names for each language and computed the entropy of this distribution, calling this quantity *script entropy* and used it as a proxy for script consistency within a language's names. Languages whose names are consistently written in a single script will have near-zero entropy.

After filtering, 468 Wikimedia language codes remained with a total of 124,343,697 names across

[5] https://gitlab.pyicu.org/main/pyicu

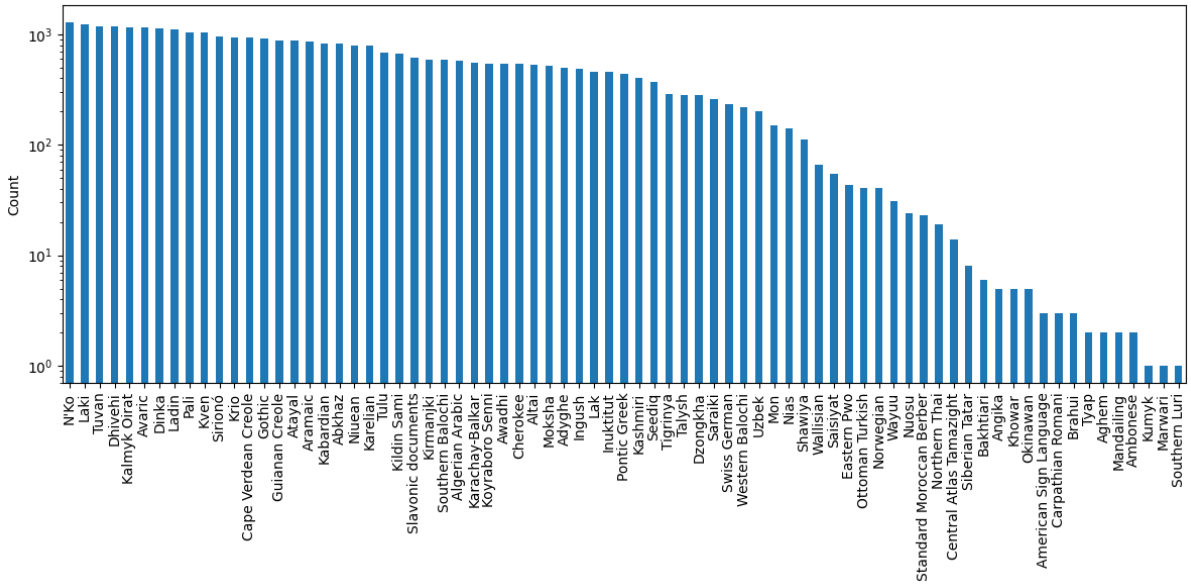

Figure 2: Name counts across the 75 higher-level languages with the fewest names (languages identified by first two letters of Wikimedia language code, $\log_{10}$ y-axis).

14,017,168 entities. The filtering process decreased the average script entropy from 0.139 to 0.025.

4.2 Matching information across languages

We noticed that some names contain additional information in parentheses following the actual tokens of the entity name, intended to help disambiguate the name from other similar-looking entities. For instance, the entity with the English label *Wang Lina (boxer)* (Q60834172) has a Russian label which contains the translation of word *boxer* in parentheses. However, this is not the case for all languages: for example, the Spanish name for the entity is simply *Wang Lina*.

To standardize the amount of information per name across languages, we remove all parentheses and tokens inside them using a regular expression.

5 Limitations

5.1 Single name per language code

Our dataset only uses the "label" property in Wikidata to identify names for entities. One of the potential limitations of this approach is that a given entity can only have a single label within a single Wikimedia language code, even though there may be multiple possible transliterations of an entity name for that language code. This can be especially problematic for languages that use more than one script but for which a finer-grained language code the specifies the script, such as `sr-cyrl`, is not available. For example, Bosnian only has the language code `bs` but is commonly written in Cyrillic and Latin scripts.

There is a possible solution in Wikidata for this limitation. There is an "also-known-as" (AKA) property, which for many entities contains useful examples of real-world names used to refer to it and can include alternative transliterations. Unfortunately, it often includes names that only loosely correspond to the canonical name of the entity. For example, AKAs for the late U.S. Supreme Court justice Ruth Bader Ginsburg (Q11116) contain not only her full name, *Ruth Joan Bader Ginsburg*, but also common aliases from popular culture, such as *Notorious RBG*. In the case of Donald Trump (Q22686) the AKAs contain other variations of his name (*Donald John Trump*, *Donald J. Trump*, etc.), but also pseudonyms that he has used that do not correspond to his actual name (*John Barron*, *John Miller*, *David Dennison*, etc.). While this information could be argued to be useful for downstream tasks such as entity linking, we felt that these alternative names introduced potentially unwanted variation in the names across languages. For this reason, we chose not to include the also-known-as fields in our dataset at this time.

There are other datasets that do not share the limitation of only having one name for an entity per language. For example, the NEWS 2018 shared task dataset (Chen et al., 2018a,b) allows for multi-

ple correct reference transliterations. Participants in that shared task also produced a ranked list of candidate translations, which can help handle the arbitrary nature of picking from an otherwise synonymous list of candidates.

5.2 Wikidata quality issues

Another limitation of our resource is our limited ability to address cases where Wikidata contains labels that may have been copied from one language to another without scrutiny. While our preprocessing pipeline removes names that appear in an incorrect script for a given language—for example, a Latin-script named copied into a language that does not use the Latin script—names blindly copied from one language into another that are in the correct script cannot reliably be detected.

Thus, a Latin-script language like Asturian which contains many names on Wikidata but has few speakers—raising the question of whether those names were added by actual speakers of the language—may have many names in our resource that were copied from English without any human review. We cannot automatically filter out these names, and collecting native speaker judgments on each one would be cost-prohibitive. While heuristic approaches like computing the percentage of names exactly equal to English could be employed, as many names are identical across languages, this may not be a meaningful heuristic.

5.3 Nicknames

Another source of variation not addressed in this work is nicknames, which can create non-parallelism. For example, while the English Wikidata label for *Joe Biden* uses the nickname *Joe*, a minority of the labels in other languages use forms of *Joseph*.

However, it can be difficult to differentiate the use of nicknames from ordinary transliteration of the full name, which may show the affects of phonological adaptation or morphological simplification. For example, the first name of *Konstantinos Ypsilantis* (Q2272090) may be written with the nominative *-os* suffix of the original Greek in some languages but appear without it in others (Polish: *Konstantyn*, Slovenian: *Konstantin*, etc.).

Unlike removing undesirable nickname variation like *Joe/Joseph* and *Will/William*, normalizing the dataset to always include or remove the *-os* suffix in the name of cross-language consistency would overly simplify the translation task.

6 Future work

In addition to fixing data quality issues, in future work we wish to empirically validate the usefulness of our resource by using it as training data for neural transliteration models. Examples of such models include the recurrent architectures of Moran and Lignos (2020) and the Transformer-based architecture of Wu et al. (2021) which holds promise for several character-level transduction tasks.

In our experiments, we seek to quantify to what extent models can benefit from being trained on massively multilingual collections of potentially noisy entity names, as opposed to more curated, smaller datasets used in past work. While this process will expose the model to more noise, and it is impossible to fully address all sources of noise and non-parallelism, we believe that using a larger dataset like ParaNames can help us understand how robust neural transliteration models are to such variations.

7 Conclusion

ParaNames enables the modeling of names cross-linguistically for millions of entities in over 400 languages. While we use Wikidata as our source, we have not simply taken its data as-is. Through careful analysis of the source data, we have developed an approach to processing Wikidata labels to create a massively multilingual name corpus where names are in the expected scripts and all entities have usable entity type information.

As it is derived from volunteer editing, we do not claim that this resource will provide perfect data. However, it does provide the broadest coverage of entities and languages available of any resource to date, to the best of our knowledge. The release of this resource can help enable multi-faceted research in names, including name translation/transliteration and further research in named entity recognition and linking, especially in lower-resourced languages.

We believe that the broader impact of this work will be the improvements to applications that it enables and the scrutiny it can place on the contents of Wikidata so that the quantity and quality of entity names across languages can be improved.

References

Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, and Mark Cieliebak. 2020.

TRANSLIT: A large-scale name transliteration resource. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3265–3271, Marseille, France. European Language Resources Association.

Nancy Chen, Rafael E. Banchs, Min Zhang, Xiangyu Duan, and Haizhou Li. 2018a. Report of NEWS 2018 named entity transliteration shared task. In *Proceedings of the Seventh Named Entities Workshop*, pages 55–73, Melbourne, Australia. Association for Computational Linguistics.

Nancy Chen, Xiangyu Duan, Min Zhang, Rafael E. Banchs, and Haizhou Li. 2018b. NEWS 2018 whitepaper. In *Proceedings of the Seventh Named Entities Workshop*, pages 47–54, Melbourne, Australia. Association for Computational Linguistics.

Ann Irvine, Chris Callison-Burch, and Alexandre Klementiev. 2010. Transliterating from all languages. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Yuval Merhav and Stephen Ash. 2018. Design challenges in named entity transliteration. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 630–640, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Molly Moran and Constantine Lignos. 2020. Effective architectures for low resource multilingual named entity transliteration. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 79–86, Suzhou, China. Association for Computational Linguistics.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Winston Wu, Nidhi Vyas, and David Yarowsky. 2018. Creating a translation matrix of the Bible's names across 591 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).