Are Language-Agnostic Sentence Representations actually Language-Agnostic?

Yu Chen **Tania Augustinova**
Department of Language Science and Technology
Saarland University, Germany
{yuchen,tania}@coli.uni-saarland.de

Abstract

With the emergence of pre-trained multilingual models, multilingual embeddings have been widely applied in various natural language processing tasks. Language-agnostic models provide a versatile way to convert linguistic units from different languages into a shared vector representation space. The relevant work on multilingual sentence embeddings has reportedly reached low error rate in cross-lingual similarity search tasks. In this paper, we apply the pre-trained embedding models and the cross-lingual similarity search task in diverse scenarios, and observed large discrepancy in results in comparison to the original paper. Our findings on cross-lingual similarity search with different newly constructed multilingual datasets show not only correlation with observable language similarities but also strong influence from factors such as translation paths, which limits the interpretation of the language-agnostic property of the LASER model.

1 Introduction

Multilingual joint embeddings map language units from different languages into the same embedding space in order to make them comparable, which facilitates cross-lingual transfer. Being essential for building NLP models of low resourced languages, such an integrated representation is also useful for cross-lingual tasks like machine translation, especially when multiple languages are involved or there is a lack of appropriate data.

While *word embeddings* are widely used in NLP tasks, sentence representations become quite important for capturing underlying semantic relations in texts across different languages. Hence, instead of simply pooling word representation together, various neural network methods have proposed to produce more coherent sentence representations. Recent advances in multilingual sentence embedding modeling (Schwenk and Douze, 2017; Feng et al., 2020; Hirota et al., 2020) begin to show strong performance on many multilingual NLP tasks, but it does not always work equally well for all languages. We repeat the cross-lingual similarity search task with LASER (Schwenk and Douze, 2017) with more challenging corpora in order to identify what affects the actual performance. Based on our findings, we propose directions for future development of such models.

2 LASER

LASER (Language-Agnostic SEntence Representations) (Schwenk and Douze, 2017; Artetxe and Schwenk, 2019) is contextualized language model based on a BiLSTM encoder trained using a translation objective on parallel data from Europarl (Koehn, 2005), United Nations (Ziemski et al., 2016), OpenSubtitles2018 (Lison et al., 2018), Global Voices (Prokopidis et al., 2016), Tanzil and Tatoeba mostly available on the OPUS website (Tiedemann, 2012). The LASER model is able to handle 93 different languages.

Also Schwenk and Douze (2017) has proposed a similarity-search-based framework in order to evaluate multilingual joint representations. With a collection of S parallel sentences for a given language pair, a multilingual similarity search is performed for the closest target sentence for each of the source sentences, and an error is counted if it is not the reference translation of that sentence in the target language. This approach requires calculating S^2 distance metrics. Duplicate sentences need to be removed from the experiment, otherwise the error rates are senseless. In order to have a meaningful comparison across N languages, a similarity search must in addition be performed on an N-way parallel sentence set. As the similarity search mainly evaluates the multilingual closeness prop-

Proceedings of Recent Advances in Natural Language Processing, pages 274–280
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_032

erty of multilingual joint sentence embeddings, the representations of the same sentence for different languages should be as similar as possible within the joint representation space.

	Target language					
Src	cs	de	en	es	fr	Avg
cs		0.70	0.90	0.67	0.77	0.76
de	0.83		1.17	0.90	1.03	0.98
en	0.93	1.27		0.83	1.07	1.02
es	0.53	0.77	0.97		0.57	0.71
fr	0.50	0.90	1.13	0.60		0.78
Avg	0.70	0.91	1.04	0.75	0.86	1.06

Table 1: Pairwise error rates (%) of similarity search of 5 languages (WMT2012).

Table 1 gives detailed similarity search error rates of LASER on the news test set from WMT2012[1]. The set consists of 3003 N-way parallel sentences in 5 European languages. Despite the significant differences between these languages, the error rates vary only slightly from the average of 1.06 with the highest error rate at 1.27. The results are consistent with those previously reported in (Artetxe and Schwenk, 2019), i.e. with the base for claiming that the model is language agnostic.

Being able to process 93 languages within a unified framework, supposedly without bias, LASER clearly has limitations. The evaluations have been executed with multilingual data that is either in-domain or fairly close to domains of training data. The applications have focused on a small subset of relatively high resource languages. Also the training data consist of more informational controlled translations, like official documents. It is unclear whether the agnostic property still holds for new domains, for different genres or for all language. After all, languages and translations in reality are much more diverse and robust than the available parallel corpora.

Therefore, we conduct a series of evaluations to examine the framework from different angles, to understand its disadvantages and to find paths for future development.

3 Similarity Search on Multilingual Corpora

We apply the LASER toolkit to two multilingual corpora that are not part of the training data used to build the pre-trained embedding models: the TED corpus (Cettolo et al., 2012) and the appropriate part of the Russian National Corpus (RNC) (Apresjan et al., 2006).

3.1 TED

We first perform a cross-lingual similarity search with the TED corpus (Cettolo et al., 2012), which contains about 17 thousand transcripts, corresponding to around 1000 English talks into 80 languages. As the distribution of translations over these 80 languages is not even and the similarity search requires N-way parallel corpus, we only consider a set of 23 languages (253 possible pairs). After excluding duplicates and limiting the sentence length to 50 tokens, we extract 10 thousand sentences that are 23-way parallel.

TED was not included for training LASER sentence encoders, while covering a large subset of languages that are supported by LASER. Unlike Europarl and UN corpora from documents mainly in the parliament and public office domain, TED involves larger varieties of domains and topics. Based on transcriptions of public presentations, TED corpus is style-wise closer to Europarl, but still differs from the parliament debates in many ways.

Table 2 displays the detailed search error scores (in percent) for 272 out of 506 language pairs in total. The results are quite different from those reported in Section 2 (Table 1) and in (Schwenk and Douze, 2017). The overall error rate for 23 langauges is 18.54, almost 9 times as much as the previous ones. Korean and Chinese turned out to be clear outliers among all languages, with search error rates as high as 43.59, while language pairs involving English tend to have less errors than the rest.

Overall, the search error rates in the table appear to correlate with language similarities between the languages. For example, among all language pairs from and to Ukrainian, which is a relatively more difficult language for the cross-lingual search task in general, the two pairs with Bulgarian and Russian perform significantly better than the rest of the language pairs. We observed similar results for Italian to French and Spanish.

It is quite clear that such multilingual sentence representation models are not equally applicable for all language pairs.

[1] https://github.com/facebookresearch/LASER/

Src	Target language															
	bg	cs	de	en	es	fr	hr	it	ko	pl	ro	ru	sr	uk	vi	zh
bg		11.62	12.62	6.83	9.53	8.15	8.28	10.71	35.95	17.16	9.44	14.62	11.09	14.20	13.32	24.76
cs	12.07		15.31	11.98	13.58	12.80	12.07	14.26	38.28	19.30	13.13	18.24	15.60	18.39	18.23	27.97
de	12.45	15.03		10.86	13.12	12.57	12.77	14.63	37.94	20.59	13.38	18.63	15.39	19.26	18.35	27.48
en	6.14	10.84	10.33		6.52	5.58	5.62	8.39	34.83	16.57	6.59	13.95	9.07	13.81	9.36	23.28
es	9.42	13.45	13.49	7.30		8.34	9.65	10.36	36.83	18.66	9.93	16.54	12.40	16.66	14.50	25.44
fr	8.54	12.47	12.94	6.41	8.72		8.19	10.01	35.86	17.20	8.55	15.49	11.64	15.46	13.26	23.97
hr	7.88	11.69	12.70	6.27	9.36	7.96		10.85	35.73	17.32	9.22	15.06	9.44	14.95	13.47	24.73
it	10.53	14.09	15.08	9.34	10.21	9.93	11.00		37.43	19.21	11.16	17.00	13.82	18.28	15.91	26.60
ko	37.01	39.19	38.71	36.04	37.58	37.01	36.92	38.62		43.07	37.92	40.97	39.37	42.53	41.09	43.03
pl	18.06	20.28	21.76	18.60	19.51	18.25	18.32	20.06	43.05		19.23	22.34	20.89	23.65	25.22	33.05
ro	9.53	13.27	14.32	8.11	10.00	9.11	9.05	11.49	36.89	18.55		16.53	12.57	16.58	14.79	26.13
ru	14.64	18.09	19.19	15.02	16.47	15.61	15.30	17.59	40.09	21.82	16.25		18.30	15.31	21.04	28.95
sr	10.73	14.87	15.37	9.59	12.05	11.49	9.24	13.50	38.35	19.94	12.18	17.78		17.80	17.01	27.56
uk	14.60	18.71	19.82	15.32	16.94	16.04	15.77	18.50	41.88	23.02	17.14	15.98	18.38		22.05	30.71
vi	14.31	18.95	19.74	11.72	14.93	13.63	14.80	16.74	40.55	24.62	15.69	21.89	18.08	21.93		28.68
zh	24.34	27.68	27.97	24.66	25.53	24.93	25.05	27.02	41.68	31.57	25.63	29.39	27.85	30.37	28.42	

Table 2: Similarity search error rates (%) on 17-way parallel WIT^3 (TED talks)

3.2 Russian National Corpus

To further explore LASER models, we apply them to the Russian National Corpus (RNC) (Apresjan et al., 2006), the multilingual section of which includes literary translations of several classical novels into different languages, which makes the RNC fundamentally different from many other parallel corpora. This is because in addition to rendering the information for the reader, a literary translation also needs to recreate the artistic imagery of the respective original work. The translator must produce a rendition in the target language, taking into account various specific features of the text, sometimes even rewriting it completely. The translations in the RNC are mostly in Slavic languages, some of which are considered low resourced for multilingual NLP tasks. Another distinctive feature of the multilingual section of the RNC is that for some of the novels there are multiple translation into the same language.

The texts in the multilingual RNC are all paragraph-aligned. We segmented the paragraphs into sentences and then align them pairwise with Hunalign (Varga et al., 2007). Then, the pairwise alignments are intersected with a relatively high alignment confidence threshold to produce a N-way parallel set. Unaligned sentences and duplicates are removed. We describe a few experiments with the multilingual RNC in the following subsections.

3.2.1 *"The Little Prince"*

Error rates for cross-lingual similarity search performed on the French novel "*The Little Prince*" in 12 languages are listed in Table 3. There are 867 sentences for each of the languages.

Similar to English in Table 2, French, the language of the original work, corresponds to significantly lower error rates except for between French and Russian. As a matter of fact, Russian tends to have higher error rates in this experiment. *It is possibly due to the fact the Russian translator Nora Gal, a primarily English-Russian translator, could well have been influenced by other versions of the novel.*

Furthermore, for each source language, the highest error rate can be more than 10 higher than the lowest and the differences do not seem to be random. The lowest search errors usually happen between languages that are more similar to each other. For instance, with Czech sentence embedding inputs, more Slovak sentences than in any other language are correctly retrieved. Likewise, the Russian-Ukrainian pair exhibits a similar property. The same pattern also exists for Bulgarian, Coatian, Macedonian and Serbian. That is, using this matrix of search error rates, we are able to divide the investigated set of Slavic languages into 3 groups:

- Eastern Slavics: Czech, Slovak;
- Western Slavics: Russian, Ukrainian;
- South Slavics: Bulgarian, Coatian, Macedonian and Serbian.

In other words, the cross-lingual similarity search task tends to be easier in case of closely related languages. As for the multilingual sentence embedding models, the distances between vectors representing the same sentence in different languages are clearly affected by language similarities. When applying available pre-trained models to other cross-lingual tasks, it is necessary to take into considerations that linguistic distances could affect the performance.

Src.	Target language bg	cs	fr	hr	mk	ru	sk	sl	sr	uk	avg
bg		11.53	9.23	10.38	10.15	15.92	11.19	11.88	10.96	12.34	11.51
cs	10.73		11.3	11.76	12.57	18.8	9.69	12	13.49	14.42	12.75
fr	8.65	9.8		9.46	8.3	16.49	10.96	11.07	9.34	12.8	10.76
hr	9.23	11.19	9.34		9.57	17.88	10.61	13.03	8.42	13.26	11.39
mk	9.92	11.88	9.69	11.07		19.03	13.15	13.26	11.88	13.38	12.58
ru	15.57	17.88	17.42	17.99	19.72		19.38	18.8	19.49	13.49	17.75
sk	10.61	8.77	10.5	10.73	12.11	18.57		13.03	12.46	14.53	12.37
sl	12.34	12.57	12.23	14.76	12.8	20.07	15.8		15.46	16.61	14.74
sr	10.73	13.26	10.15	8.77	11.65	19.72	12.69	14.53		14.65	12.91
uk	12.11	15.34	13.03	15.34	13.96	13.73	14.99	16.61	16.15		14.58
Avg	11.10	12.47	11.43	12.25	12.31	17.80	13.16	13.80	13.07	13.94	13.13

Table 3: Similarity search error rates (%) on "The Little Prince"

3.2.2 "*Alice in Wonderland*"

For some of the classical novels, the RNC includes more than one edition for the same language. "*Alice in Wonderland*" is one of them: there are 3 Russian translations and 2 Polish translations. The respective publishing dates and translators are not provided in the corpus. From the multilingual section of the RNC, we extracted 356 sentences that are parallel in 11 languages and repeated the cross-lingual similarity search on these sentences. Table 4 shows partial results from the experiment.

The novel is originally written in English. While the scores concerning English are frequently lower than expected, there are clear exceptions. When using sentence embeddings generated from two alternative Polish translations to retrieve original English sentences, the error rates range from 5.06% to 23.88%, almost a 400% increase. Apparently, one version (pl_2) consists of sentences that are closer to the English original than the other version when mapped to the joint multilingual sentence representation space. The same holds for the three Russian translations. In the English target column, we can see that the three scores related to Russian alternative translations are about 10% apart. Interestingly, the version (ru_3) has lower error rates if we search for Ukrainian sentences rather than for English sentences. This brings us to speculate that this Russian version might have served as a source for producing the Ukrainian translation instead of English.

Obviously, there are further factors affecting the language independence of the multilingual sentence representations. The ontological differences from original untranslated texts and translation path appear to keep translations distinguishable from the originals. The literacy translators' freedom to recreate the work in a new language somehow amplifies the issue. Different translations may differentiate from the original source in so many different ways, This is clearly an interest field to explore with multilingual representations.

3.2.3 Belorussian and Ukrainian

After carrying out the similarity search experiments with all 9 novels in RNC, we summarize the error rates by language in Table 5. The columns "A"-"I" represent the nine novels. Most of the languages have an average error rate around 20%, but Ukrainian and Belorussian has much higher error. In particular, the search error rates of Belorussian experiments rice up to 73.60%. In addition to their distinctive linguistic features, it is likely due to the relatively smaller amount of parallel data that is available for training the LASER models, which are not effective for all languages.

4 Similarity search on different translation paths

As we have discussed in previous sections, the majority of multilingual parallel corpora are collections of translations of the same source documents into different target languages, between which cross-lingual similarity search appears to be more difficult. To investigate the underlying factors that affect similarity search by sentence embeddings, we construct new 4-way parallel data by adding translations from different paths to an existing 3 way parallel set in the following manner: we first select 8 TED talks from the recent online release so as to minimize the translator's prior knowledge of the talks. All the talks have been transcribed in English and translated into both German and Chinese. There are 629 sentences for each of the 3 languages. The texts are sentence aligned across all 3 languages manually. We sent the sentence segmented German translations as source documents to 4 German-Chinese profes-

Src.	Target language en	pl_1	pl_2	ru_1	ru_2	ru_3	uk	Avg.
en		25.56	6.18	25.84	10.96	31.18	23.60	20.55
pl_1	23.88		18.26	33.99	28.37	38.20	32.02	29.12
pl_2	5.06	20.22		22.19	14.33	32.87	23.31	19.66
ru_1	21.07	34.83	20.79		21.63	33.99	23.88	26.03
ru_2	10.39	28.93	12.64	21.35		27.25	21.91	20.41
ru_3	28.93	38.20	28.37	35.67	29.78		24.72	30.95
uk	20.22	30.34	18.54	23.31	21.07	24.16		22.94
Avg.	18.26	29.68	17.46	27.06	21.02	31.28	24.91	24.24

Table 4: Similarity search error rates (%) on "Alice in Wonderland"

	A	B	C	D	E	F	G	H	I	Avg.
be	63.40	44.75	50.05			73.60				57.95
bg	24.29	8.65	14.24	20.33	30.05	32.46		6.94	14.49	18.93
cs	25.41	16.63	15.76	16.65		30.05	14.31	7.64		18.06
hr	28.34	8.01	14.35	14.37		30.77	13.01			18.14
mk	22.24	8.43	15.80	14.93		30.04	13.36			17.47
pl	27.98	12.33	18.37	19.05	32.44	33.06	15.21	9.63	20.03	20.90
ru	30.97	7.35	19.72	23.33	41.54	33.85	16.50	14.61	20.66	23.17
sk	23.49	11.94	15.27	17.52		28.91		6.76		17.32
sl	21.75	9.97	17.90	16.86	33.70	30.18				21.73
sr	25.65	10.75	15.80	14.28		29.90		6.13		17.09
uk	28.61	8.74	16.91	20.84	38.81	36.41				25.05

Table 5: Average search error rates by language on Russian National Corpus

Src	Target Language (en-)de	(de-)zh	en org.	(en-)zh	Avg.
(en-)de		2.70	8.74	10.81	7.42
(de-)zh	2.38		11.45	12.72	8.85
en	9.86	11.92		8.74	10.17
(en-)zh	13.04	14.31	7.15		11.50
Avg.	8.43	9.64	9.11	10.76	9.49

Table 6: Cross-lingual similarity search error rates (%) on a 4-way parallel corpus that contains translations produced from different paths

sional translators and asking them to produce Chinese translations and to stay close to the German texts semantically as much as possible. Accordingly, the translators might not have as much creative space to improvise in the process as a normal freelance translator. Since the German texts have been sentence segmented before the translation, the resulting Chinese texts can be easily aligned back to the 3-way parallel set after review. The newly translated Chinese texts are then added into the set as yet another separate copy, and we perform cross-lingual similarity search on the updated set using LASER as shown in Table 6.

This data set includes two Chinese versions of of the same English presentations that may be considered paraphrases to each other: one is directly translated from the original English transcription and the other is pivotally translated through German. In this setup, a perfect language agnostic embedding model should be able to map the sentences into vector clusters, each of the clusters representing an English sentence with its German translations together with two variants of Chinese translations. However, our results contradict this assumption.

The distance between the English original and the English-Chinese translations is not far from that between the English and the German ones. The differences in error rate are within 1, around 6 sentences out of 629. The German-Chinese translations did turn out to be much closer to the German texts in the sentence representations. We believe it is due to the strict requirements given to the translator. Despite being translations into the same language, the two Chinese texts lead to the highest search error rates.

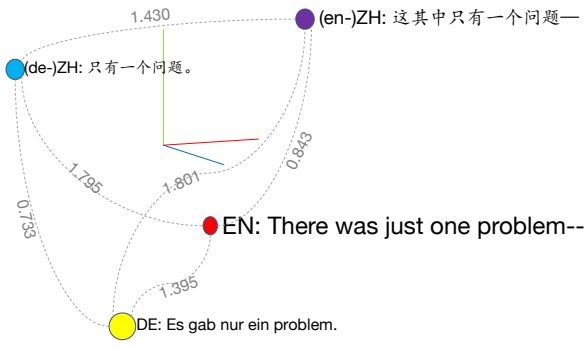

Figure 1: Vector space visualization of German and Chinese translations and the English source sentence

Figure 1 illustrates an example from this data set. The dots represent vectors generated for each of the sentences with the LASER toolkit in a joint sentence embedding space and the edges connecting the dots are labeled with distances. Notably, the German translation uses period instead of "–" at the end of the sentence, which is clearly not possible to recover in the German-Chinese translation. This is reflected in the distances between these representation vectors. The German-Chinese translation is closer to the German sentence rather than to the English or the Chinese sentence. Similar examples are fairly common in this set, which also explains the distinctive performance of cross-lingual similarity search for the translations into the same language we discussed in Section 3.2. It is inevitable that translation introduces distortions into texts. Even though the ultimate goal of building up a multilingual sentence representation model is to allocate sentences with the same meanings regardless of their languages as close to each other as possible, translation distortions are still visible in the state-of-the-art multilingual sentence representations. Potentially, the joint sentence embeddings may be one way to identify translation paths or even to quantify translation distortions.

5 Conclusion

Neural embeddings have been widely applied in all fields of natural language processing. Multilingual embeddings with shared representation space enable few-shot and zero-shot transfer from one language to another with minimum additional training or data requirement. Recent developments on multilingual sentence representations such as LASER have opened up new path towards competitive NLP performance across high- and low-resource languages.

Yet, our evaluation of LASER reveals many constraints when applied in realistic and challenging scenarios. The performance of the framework is largely influenced by the similarity between languages in the multilingual application. Not all languages work equally well currently. Low performance on specific languages is attributed to the small training data size.

These observations caution the interpretation of language-agnostic property of such cross-lingual sentence representations and their application in multi-lingual NLP applications. The newly constructed multilingual corpus in this paper can be used as a new evaluation benchmark for future cross-lingual representation learning research. We

plan to release the data set to the research community.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Juri D. Apresjan, Igor Boguslavsky, Boris Iomdin, Leonid L. Iomdin, Andrei Sannikov, and Victor G. Sizov. 2006. A syntactically and semantically tagged corpus of russian: State of the art and prospects. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*, pages 1378–1381. European Language Resources Association (ELRA).

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Wataru Hirota, Yoshihiko Suhara, Behzad Golshan, and Wang-Chiew Tan. 2020. Emu: Enhancing multilingual sentence embeddings with semantic specialization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7935–7943.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel Global Voices: a collection of multilingual corpora with citizen media stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905, Portorož, Slovenia. European Language Resources Association (ELRA).

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).