Manifesto Project Corpus

Manifesto Corpus Translation

manifesto-communication@wzb.eu

Website: https://manifesto-project.wzb.eu/

The handbook specifies how the Manifesto Corpus is translated into English.

Manifesto Project's Handbook Series
1st edition
from April 23, 2024

1 Introduction & Citation

The following technical report provides an introduction to the English language edition of the Manifesto Corpus (Lehmann et al., 2024). Overall, 1,593,932 quasi-sentences contained in the Manifesto Corpus were translated into English with the DeepL Translator (*DeepL Translator*, 2017). Together with the electoral programmes, which have English as original language, the English language edition of the Manifesto Corpus provides users with a resource that contains 1,626 machine-readable electoral programmes from more than 50 countries between 1946 and 2022 in English. The full multilingual Manifesto Corpus (Lehmann et al., 2024) contains 2,032 machine-readable manifestos between 1946 and 2023. Thus, the English language edition of the corpus currently covers circa 80% of all manifestos contained in the full Manifesto Corpus.

The following sections provide an introduction to the translation of the corpus. First, a description of the Manifesto Corpus (Lehmann et al., 2024) and the DeepL Translator (*DeepL Translator*, 2017) used for the translation is provided. Then, the translation procedure is introduced in detail. Finally, the validation steps taken to ensure the quality of the translation are described.

Please check our website for the latest version of the Manifesto Corpus and its translation:

`https://manifesto-project.wzb.eu/`

When citing this document, please refer to:

> Ivanusch, Christoph / Regel, Sven (2024): Manifesto Corpus Translation. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB) / Göttingen: Institut für Demokratieforschung (IfDem).

2 Translation

2.1 Manifesto Project Corpus

The Manifesto Corpus is a free, digital, multilingual, and annotated collection of electoral programmes (Lehmann et al., 2024). It is based on the collection of the Manifesto Project (MRG/CMP/MARPOR), comprising the currently largest collection of annotated electoral programmes. The Manifesto Corpus provides the electoral programmes, document meta data (e.g. party name, election date, language) and the codings of single quasi-sentences in machine-readable format. Currently, the full corpus contains 2,032 machine-readable electoral programmes from more than 60 different countries in 40 languages. The annotated part of the corpus covers more than 2,000,000 quasi-sentences coded according to the Manifesto Project coding scheme (Werner et al., 2021). The corpus is regularly updated, corrected and extended. It can be accessed through different avenues, such as explored online, downloaded as csv documents, acessed through an API, the R package manifestoR or the stata add-on manifestata. More information on the corpus is available on the website of the Manifesto Project (`https://manifesto-project.wzb.eu/information/documents/corpus`).

For the translation of the Manifesto Corpus into English, we have focused on the languages available for translation with the DeepL Translator (*DeepL Translator*, 2017), described in more detail below. This allowed us to translate manifestos into English in 24 of the 39 languages (exclusive of English) covered by the Manifesto Corpus.[1] Overall, we translated 1,593,932 quasi-sentences contained in the corpus into English, while further 254,402 quasi-sentences already have English as the original language. Table 1 provides an overview of the number of quasi-sentences per language that were relevant for the translation into English.

2.2 DeepL Translator

DeepL Translator was launched in 2017 and uses neural networks to provide machine translation services (*DeepL Translator*, 2017). DeepL Translator currently supports more than 30 languages and can be used through different avenues. Users can translate (short) texts directly via a web-interface, upload full documents (e.g. pdf, word) to a web-interface and receive the documents back in the same format[2], or leverage the DeepL API to translate texts.

For larger tasks such as the corpus translation, the full document translation and the API translation are both feasible avenues. However, both options differ in workflows and costs. While leveraging the API would allow (nearly) complete automatisation of the translation process, the full document translation via the web-interface is much more cost-efficient. Furthermore, the full document translation is better suited to our translation approach, described in more detail below. Based on these considerations, we make use of the document translation option of the DeepL Translator to save costs, keep to our budget and fit our translation approach.

2.3 Corpus Preparation

The main task in preparing the corpus for the translation is to convert it into a suitable format for the document translation with the DeepL Translator. We opt for the word format (.docx). Next, we needed to take a number of steps in text preparation that are related to properties of the Manifesto Corpus.

As introduced above, the Manifesto Corpus is an annotated text corpus. The annotation is based on the quasi-sentence level. Quasi-sentences are distinct from natural sentences. While natural sentences can contain multiple policy statements, goals or messages, quasi-sentences 'contain exactly one statement or message' (Werner et al., 2011). Thus, coders in the Manifesto Project split natural sentences into quasi-sentences 'if they contain unrelated statements, possibly indicated by semi-colons, or if it is possible to allocate different codes to different parts of the natural sentence' (Merz et al., 2016), or if they contain

[1] Manifestos in the following languages are not yet covered by the translation: Armenian, Bosnian, Bosnian-cyrillic, Catalan, Croatian, Galician, Georgian, Hebrew, Icelandic, Japanese, Korean, Macedonian, Montenegrin, Serbian-cyrillic and Serbian-latin. The DeepL Translator offers the possibility of translating Japanese and Korean texts. However, our validation procedures showed that our translation approach did not deliver satisfactory results in the case of Japanese and Korean. Therefore, these two languages are not (yet) included in the English language edition of the Manifesto Corpus.

[2] In the case of paid accounts, users can translate documents in the following formats via the web-interface: .doc, .docx, .html, .pdf, .pptx, .txt, .xlf, .xliff and .xlsx.

Table 1: Number of quasi-sentences per language in Manifesto Corpus relevant for translation.

Language	N
Bulgarian	12,362
Czech	26,224
Danish	18,910
Dutch	215,776
Estonian	16,813
Finnish	22,280
French	132,085
German	178,937
Greek	47,418
Hungarian	47,666
Italian	22,177
Latvian	2,031
Lithuanian	44,980
Norwegian	85,170
Polish	28,545
Portuguese	94,202
Romanian	15,465
Russian	7,480
Slovak	30,257
Slovenian	40,029
Spanish	418,916
Swedish	22,286
Turkish	57,244
Ukrainian	6,679
All	1,593,932

several but separate statements that are allocated the same code. In some languages, natural sentences can be very long and include multiple statements. Furthermore, several manifestos contain lists or bullet points. In such instances, natural sentences are split into several quasi-sentences. This feature has implications for the translation.

In principle, two options for dealing with this feature are available to us. First, we could translate each quasi-sentence individually. This would provide an easy solution as each translation could be matched back to the original text based on an ID or the like. However, this approach also has a major drawback. If the translation is based exclusively on individual quasi-sentences, the quality of the translation is reduced. A meaningful translation of individual quasi-sentences often requires context information from other quasi-sentences that belong to the same natural sentence. Thus, higher translation quality requires translation of natural sentences that than need to be split into quasi-sentences again to be merged back to the original text and the corresponding annotation. This provides some challenges as words or text passages belonging to certain quasi-sentences might occur in different order within a natural sentence in the original text and in the translation. Thus, merging the translated quasi-sentences back to original texts is not a trivial task anymore. Therefore, we have developed a procedure that allows to identify the corresponding quasi-sentences in the original language text and the English translation. We have concatenated individual quasi-sentences into natural sentences (i.e. 'chunks') and then assigned different text styles (e.g. colour, bold) to each quasi-sentence within each chunk. These styles remain unchanged when being processed by the document translation of the DeepL Translator, allowing us to merge back the translated texts to the original text on the quasi-sentence level after getting the translated documents back from the DeepL Translator.

Overall, we have created more than 200 word documents for upload to DeepL Translator. Each word document is dedicated to one original language (e.g. bulgarian, czech, danish) and contains a maximum of 1 million characters as DeepL does not allow the translation of larger documents.

2.4 Translation Procedure

For the translation, we proceeded in three major steps. As discussed above, the aim of the translation was to prepare word documents containing all quasi-sentences supposed for translation as natural sentences (i.e. 'chunks'), then upload these word documents to a web-interface of DeepL Translator and finally, process the translation output to merge it back with the quasi-sentences in the original language and the corresponding annotations contained in the Manifesto Corpus.

First, we uploaded each of the more than 200 word documents to the web-interface of the DeepL Translator. As discussed above, these documents contained multiple chunks of natural sentences (i.e. one or more related quasi-sentences) with each individual quasi-sentence marked with a specific text style (e.g. colour, bold). Each document was devoted to one language covered in the Manifesto Corpus and had a maximum of 1 million words (limit for DeepL document translation). Thus, we regularly had to translate multiple documents per language. The web-interface used for this task allows to upload a word document (.docx), then allows to specify the language of the text in the document, performs the translation into the desired language, i.e. English in our case, and finally provides the translated document as a word file again (.docx). Thus, we uploaded each word document that we prepared for translation to the web-interface, performed the translation and then downloaded the output again to our file store.

Second, we loaded all texts from the translated word documents into *R* (R Core Team, 2023). Here, we leveraged the styles that we assigned to each quasi-sentence to merge the translated text with the original corpus. Together with our knowledge about the number and ordering of quasi-sentences within each chunk in the original corpus, we were able to correctly merge large parts of the translated quasi-sentences to the original quasi-sentences.

Finally, we dealt with those quasi-sentences where the merge with the original corpus did not work. There are several reasons why a merge between the original corpus and translated quasi-sentences was not possible or potentially problematic. Examples of such reasons include unclean text passages in the word document used as input for the translation in the first place, a very different order of words or even quasi-sentences in the translation output compared to the original text, strange results for the translation or simple technical difficulties. Thus, the chosen translation approach based on natural sentences (i.e. chunks of one or more related quasi-sentences) failed in these cases. To fill this gap, we translated the affected text chunks again, but this time quasi-sentence by quasi-sentence individually. Overall, circa 13% of the quasi-sentences in the translated corpus were processed in this way.

3 Validation

To validate our translation procedure, we performed two separate validation tasks. As mentioned earlier, we translated more than 1.5 million quasi-sentences from 24 languages into English. Providing a detailed validation for each language, would exceed our resources. We therefore opted for two separate tasks, where the first task takes a broader and the second task a more detailed approach. The two validation tasks and the corresponding results are reported in the following sections.

3.1 Validation Task #1

The first validation task aims to provide a broad, indicative overview of how well the translation procedure performed. To achieve this goal, we opted to compare our translations with other translations of the individual quasi-sentences contained in the Manifesto Corpus. Thus, the first validation task is not a classic validation procedure for machine translation, but rather gives an indication of how well our translation approach aligns with other translation approaches and tools.

For this task, we provide two coders with a sample of quasi-sentences in the original language, our translation into English as well as further translations into English. In contrast to our translation described above, the other translations were performed at the simple quasi-sentence level, one again with the DeepL Translator and others using the open-source machine translation tools Argos Translate (Finlay, 2021) and OPUS-MT (Tiedemann and Thottingal, 2020). While DeepL is clearly superior to Argos Translate and OPUS-MT, they achieve satisfying results (e.g. Licht et al, 2024) and are therefore well-suited tools for this validation task. Table 2 shows examples of sampled quasi-sentences and the corresponding translations. The column 'Text (original)' contains the quasi-sentences in the original language, the

column 'Translation' is dedicated to our translation resulting from the procedure described above and the remaining columns contain the additional translations on the quasi-sentence level conducted for this validation task ('QS Translation (DeepL)', 'QS Translation (Argos)' and 'QS Translation (OPUS-MT)'). Table 3 provides an overview of the number of quasi-sentences sampled for validation per language and coder. Overall, we sampled 639 quasi-sentences for this task, with a subset of these included in the coding samples for both coders to check intercoder reliability. For this sample, we randonly sampled 20 chunks (i.e. one or more related quasi-sentences) per language. This strategy resulted in a sample of 27 quasi-sentences on average per language, with a minimum of 21 quasi-sentences and a maximum of 35 quasi-sentences per language contained in the sample.

Table 2: Examples of coding task #1 for validation.

Manifesto ID	Language	Text (original)	Translation	QS Translation (DeepL)	QS Translation (Argos)	QS Translation (OPUS-MT)
82430_201710	Czech	Personálně posílíme regulační orgány (zejména Český telekomunikační úřad a Energetický regulační úřad) o pozice pro osobnost, jejíž zkušeností a náplní bude posílení postavení spotřebitele.	We will strengthen the regulatory bodies (in particular the Czech Telecommunications Office and the Energy Regulatory Office) with positions for a person whose experience and focus will be consumer empowerment.	We will strengthen the regulatory bodies (in particular the Czech Telecommunications Office and the Energy Regulatory Office) with positions for a person whose experience and focus will be consumer empowerment.	We will personally strengthen regulatory bodies (especially the Czech Telecommunications Office and the Energy Regulatory Office) by a position for a person whose experience and purpose will be to strengthen consumer status.	We will personally strengthen regulatory bodies (especially the Czech Telecommunications Office and the Energy Regulatory Office) by a position for a person whose experience and purpose will be to strengthen consumer status.
82430_201710	Czech	Je třeba obnovit život na venkově,	It is necessary to rebuild rural life,	Rural life needs to be restored,	It is necessary to restore life in the countryside,	It is necessary to restore life in the countryside,
82430_201710	Czech	vychovávat mladou generaci	educate the young generation	educate the young generation	Raising a Young Generation	Raising a Young Generation
82430_201710	Czech	a cílevědomě vytvářet podmínky pro život a práci na venkově,	and purposefully create conditions for living and working in the countryside, to	and purposefully create the conditions for living and working in the countryside,	and purposefully create conditions for living and working in the countryside,	and purposefully create conditions for living and working in the countryside,
82430_201710	Czech	vrátit zemědělcům prestiž a zázemí.	restore prestige and facilities to farmers.	to give farmers back their prestige and background.	return to farmers prestige and background.	return to farmers prestige and background.
42320_199910	German	Eine gute Verknüpfung von Akutbehandlung und Nachsorge für pflegebedürftige oder chronisch kranke Menschen muß sichergestellt sein.	A good link between acute treatment and aftercare for people in need of care or chronically ill must be ensured.	A good link between acute treatment and aftercare for people in need of care or with chronic illnesses must be ensured.	A good link between acute treatment and care for people in need of care or chronically ill must be ensured.	A good link between acute treatment and aftercare for people in need of care or chronically ill must be ensured.
42320_199910	German	In Österreich selbst wollen wir darauf hinarbeiten, daß strategisch wichtige Industrien und Unternehmen in österreichischer Hand bleiben,	In Austria itself, we want to work to ensure that strategically important industries and companies remain in Austrian hands, with	In Austria itself, we want to work towards ensuring that strategically important industries and companies remain in Austrian hands,	In Austria, we want to work towards the fact that strategically important industries and businesses remain in Austria,	In Austria itself, we want to work to ensure that strategically important industries and companies remain in Austria's hands.
42320_199910	German	wobei die ÖIAG als kompetenter Kernaktionär der Republik Österreich eine zentrale Rolle spielt.	ÖIAG playing a central role as a competent core shareholder of the Republic of Austria.	ÖIAG plays a central role as a competent core shareholder of the Republic of Austria.	where the ÖIAG plays a central role as a competent core shareholder of the Republic of Austria.	ÖIAG plays a central role as the competent core shareholder of the Republic of Austria.
32902_200804	Italian	Diffusione capillare di asili e asili nido con orari di apertura identici a quelli di lavoro	Widespread spread of daycare centers and kindergartens with opening hours identical to work hours	Widespread dissemination of daycare centers and kindergartens with identical opening hours to work hours	Dissemination of nurseries and nurseries with opening hours identical to working hours	Widespread distribution of kindergartens and kindergartens with opening hours identical to working hours
32902_200804	Italian	e creazione di un sistema di tutele per le lavoratrici madri.	and creation of a system of protections for working mothers.	and creation of a system of protections for working mothers.	and creation of a brace system for mother workers.	and the creation of a protection system for mother workers.

Table 3: Number of quasi-sentences per language in validation sample (task #1) per coder.

Language	Coder #1	Coder #2
Bulgarian	12	17
Czech	17	18
Danish	12	16
Dutch	13	16
Estonian	15	13
Finnish	14	14
French	18	14
German	13	16
Greek	12	25
Hungarian	22	17
Italian	19	26
Latvian	15	24
Lithuanian	14	16
Norwegian	13	13
Polish	12	13
Portuguese	14	19
Romanian	23	15
Russian	15	12
Slovak	16	15
Slovenian	21	12
Spanish	19	16
Swedish	16	13
Turkish	12	14
Ukrainian	20	13
All	377	387

The coders were then asked to compare our translation with the other translations and mark the translation with the codes *Agree* or *Disagree*. The exact task description reads as follows:

- *Agree*: the translation in the column 'Translation' needs to substantially overlap with the translation in the column 'QS Translation (DeepL)' AND with at least one of the translations from the columns 'QS Translation (Argos)' or 'QS Translation (OPUS-MT)'. The code *Agree* also applies when different translations use slightly different terms (e.g. synonyms), as long as this would not result in a different annotation based on the Manifesto Project coding scheme.
- *Disagree*: less than three translations from the columns 'Translation', 'QS Translation (DeepL)', 'QS Translation (Argos)' and 'QS Translation (OPUS-MT)' overlap substantially or abnormalities are present in the column 'Translation'.

Overall, the two coders identified 89.66% of the translations contained in the validation sample to substantially overlap, i.e. were assigned the code *Agree*. The two coders thereby agreed in 92% of their coding decisions, based on the 125 quasi-sentences that were included in both coding samples to check intercoder reliability. Detailed results for the first validation task per language are given in Table 4. While the translation works well for most languages, some seem to work less well (e.g. slovak, turkish). Overall, the results indicate that the approach chosen for our translation seems to largely agree with other translation approaches and tools. Furthermore, the qualitative evidence from our coders suggests that most cases of disagreement between the different translations appear not to be due to our chosen translation approach, but to irregularities in the simple quasi-sentence translations performed specifically for the validation task (i.e. columns 'QS Translation (DeepL)', 'QS Translation (Argos)' and 'QS Translation (OPUS-MT)').

Table 4: Results for validation task #1 per language.

Language	Agree (N)	Disagree (N)	Agree (%)
Bulgarian	28	1	96.55
Czech	32	3	91.43
Danish	27	1	96.43
Dutch	26	3	89.66
Estonian	27	1	96.43
Finnish	27	1	96.43
French	30	2	93.75
German	29	0	100.00
Greek	32	5	86.49
Hungarian	34	5	87.18
Italian	41	4	91.11
Latvian	36	3	92.31
Lithuanian	29	1	96.67
Norwegian	22	4	84.62
Polish	24	1	96.00
Portuguese	28	5	84.85
Romanian	34	4	89.47
Russian	23	4	85.19
Slovak	22	9	70.97
Slovenian	33	0	100.00
Spanish	33	2	94.29
Swedish	26	3	89.66
Turkish	16	10	61.54
Ukrainian	26	7	78.79
All	685	79	89.66

3.2 Validation Task #2

The second validation task delivers a more detailed investigation of how well the translation has worked for the manifesto quasi-sentences. Therefore, we provided speakers of certain languages with a sample of quasi-sentences in the original language and the translation. These coders than had to compare the original text with the translation into English and judge whether the translation accurately reflects the content of the original sentence.

For this task, we randomly sampled chunks of quasi-sentences (i.e. one or more related quasi-sentences) for certain languages. We selected the languages based on their importance in the Manifesto Corpus (i.e. number of quasi-sentences per language in the corpus) and the available language skills in our team, and also examined some languages for which validation task #1 delivered comparatively weak results. We therefore sampled quasi-sentences for this validation task in the languages Danish, Dutch, French, German, Slovak, Spanish, Swedish, Turkish and Ukrainian. In general, we randomly sampled 30 chunks per language. This strategy resulted in a sample of 37.5 quasi-sentences on average per language, with a minimum of 30 quasi-sentences and a maximum of 48 quasi-sentences per language contained in the sample.

The coders were then asked to compare the quasi-sentence in the original language with the translation into English. The coders then had to assign the codes TRUE or FALSE depending on the correctness of the translation. The exact task description reads as follows: 'Please compare the original quasi-sentence and the translation into English for correctness of the translation. Correct translations accurately reflect the content of the original sentence and do not convey any other meaning. In addition, certain terms (e.g. party names, political institutions, laws) are stated correctly and are not mistranslated. Correct sentences should be marked with TRUE in the column code; incorrect sentences should be marked with FALSE in the columns code; the column note provides space to report reasons for certain judgements, report any abnormalities etc.'

Overall, the coders marked 322 out of 337 observations as correct (95.55%). A detailed overview of the results for the second validation task is reported in Table 5. Across all languages at least 85% of

all quasi-sentences were translated correctly, in most cases the correct translations exceed 95%. These results show that the translation into English with the DeepL Translator largely seems to have worked well for the manifesto quasi-sentences.

Table 5: Results for validation task #2 per language.

Language	TRUE (N)	FALSE (N)	TRUE (%)
Danish	31	0	100.00
Dutch	30	0	100.00
French	47	1	97.92
German	33	1	97.06
Slovak	32	0	100.00
Spanish	35	4	89.74
Swedish	38	1	97.44
Turkish	36	6	85.71
Ukrainian	40	2	95.24
All	322	15	95.55

4 References

DeepL Translator. (2017). https://www.deepl.com/translator

Finlay, P. J. (2021). *Argos Translate*. https://github.com/argosopentech/argos-translate

Lehmann, P., Franzmann, S., Al-Gaddooa, D., Burst, T., Ivanusch, C., Lewandowski, J., Regel, S., Riethmüller, F., and Zehnter, L. (2024). *Manifesto Corpus. Version: 2024-1*. Berlin: WZB Berlin Social Science Center/Göttingen: Institute for Democracy Research (IfDem).

Licht, H., Sczepanski, R., Laurer, M. and Bekmuratovna, A. (2024). *No more cost in translation: Validating open-source machine translation for quantitative text analysis*. https://EconPapers.repec.org/RePEc:ajk:ajkdps:276

Merz, N., Regel, S., and Lewandowski, J. (2016). The Manifesto Corpus: A New Resource For Research On Political Parties And Quantitative Text Analysis. *Research & Politics*, *3*(2), 2053168016643346. https://doi.org/10.1177/2053168016643346

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Tiedemann, J., and Thottingal, S. (2020). OPUS-MT — Building Open Translation Services for the World. *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*. https://aclanthology.org/2020.eamt-1.61.pdf

Werner, A., Lacewell, O., and Volkens, A. (2011). *Manifesto Coding Instructions. 4th Fully Revised Edition*. https://manifestoproject.wzb.eu/down/papers/handbook_2011_version_4.pdf

Werner, A., Lacewell, O., Volkens, A., Matthieß, T., Zehnter, L., and van Rinsum, L. (2021). *Manifesto Coding Instructions. 5th Fully Revised Edition*. https://manifestoproject.wzb.eu/down/papers/handbook_2021_version_5.pdf