# DELIVERABLE D3.5
# Report on how to extend the features to another language

**H2020-DS-08-2017: SMOOTH**
**Project No. 786741**
**Start date of project: 01-05-2018**
**Duration: 30 months**

**Revision:** 02
**Date:** 30-04-2020

## Document Information

**Document Name:** D3.5 Report on how to extend the features to another language

**WP3 – Title: Analysis of informative documents on data protection and privacy (SMOOTEXT)**

**Task 3.2, 3.3, and 3.4**

**Author:** Matthias Gallé

Contributors: Jos Rozen

**Dissemination Level**

**Project co-funded by the EC within the H2020 Programme**

| | | |
|---|---|---|
| **PU** | Public | **X** |
| **PP** | Restricted to other programme participants (including the Commission Services) | |
| **RE** | Restricted to a group specified by the consortium (including the Commission Services) | |
| **CO** | Confidential, only for members of the consortium (including the Commission Services) | |

### Approvals

| | Name | Entity | Date | Visa |
|---|---|---|---|---|
| Author | Matthias Gallé | NAVER | 22/04/2020 | X |
| Author | Jos Rozen | NAVER | 28/04/2020 | X |
| WP Leader | Matthias Gallé | NAVER | 30/04/2020 | X |
| Coordinator | Rosa Araujo | Eurecat | 30/04/2020 | X |

### Document history

| Revision | Date | Modification |
|---|---|---|
| **Version 1** | 22/04/2020 | V1 |
| **Version 2** | 30/04/2020 | V2 |

## List of abbreviations and acronyms

| Abbreviation | Meaning |
|---|---|
| AEPD | Agencia Española de Protección de Datos |
| API | Application Programming Interface |
| CNIL | Commission nationale de l'informatique et des libertés |
| DPA | Data Protection Authority |
| DPIA | Data Protection Impact Assessment |
| DPO | Data Protection Officer |
| GDPR | General Data Protection Regulation |
| ICO | Information Commissioner's Office |
| MEnt(s) | Micro-enterprise(s) |
| SME(s) | Small and medium-sized enterprise(s) |
| TFEU | Treaty on the Functioning of the European Union |

| | |
|---|---|
| **Executive summary** | This deliverable summarizes the multi-lingual aspects involved in SMOOTEXT, as well as steps necessary to extend it to other languages. |

This deliverable reflects only the author’s view and the European Commission is not responsible for any use that may be made of the information it contains. (Art. 29.5)

SMOOTH project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement № 786741 (Art. 29.4)

## Index

# Table of Contents

# 1.- Introduction

With its 24 official languages of the European Union, any service which targets a wide spectrum of companies across the union has to have multi-linguality incorporated from the beginning. This applies particularly to SMOOTEXT, who focuses on the textual parts of the legal documents; and even more so in the context of micro-enterprises who might operate only in one local language without operational need of translating it into English for example. From a machine perspective, the ideal would be to have annotated data in each one of the target languages and to run mono-lingual models on each of them. However, this quickly becomes cumbersome to maintain and debug. More importantly, it makes the training phase of such models very expensive, both from a standpoint of money to obtain those annotations as that of time to roll it out towards a new language.

This is because the initial strategy was to develop cross-lingual system: as opposed to multi-lingual systems, those models do not maintain one model per language but only one which crosses across languages. This is made possible thanks to the existence of multi-lingual embedding, which map textual data from any of the pre-defined languages into a common real-valued space. It is then on point of that space that the machine learning algorithms operate. Obtaining that mapping from one language into the multi-lingual space is task-specific (in order to maximize performance of the downstream NLP task) and done mostly thanks to parallel or comparable data: different sentences or documents portion which convey the same or at least similar semantic meaning in different languages. The initial strategy was to adapt those methods to privacy policies using the policies of large companies which are produced in many different languages.

However, a request in March 2019 changed the priorities. The possibility of adding Italian as an official language of the SMOOTH project, in addition to English, Spanish and – at a lesser rate – Latvian created a situation where crawling parallel data for all those languages might become very time consuming. Instead, we experimented with an alternative strategy which became the solution finally chosen.

This deliverable expands on the preliminary results provided in D3.3, and is structured in the following sections:

- Description of the chosen solution
- Benchmark study
- Limitations
- Extension to future languages

# 2.- Multi-lingual Aspects of SMOOTEXT

## 2.1.- Solution

We studied the possibility of relying on automatic translation to map the privacy policies into a chosen system language and perform all the processing in that system language.
The rationale behind was twofold. On one hand, modern translation systems have advanced significantly since the widespread use of neural machine translation. On the other hand, privacy policies have formulaic language, which is often repeated. In addition, due to the availability of some privacy policies in multiple languages we expect them to be part of the parallel training data used by major providers.
As system language we chose English, due to its widespread use and availabilities of resources in that language.

In the workflow (see D3.4), the translation comes immediately after the text extraction. The whole processing (GDPR element extraction, readability analysis, company extraction) is therefore done on the translated text. We will analyze the shortcomings of this in the last section.

Instead of translating the incoming text into English, an alternative approach would have been to translate the training data (English) into all source languages and train a language-specific model for each language. This however increases the maintenance cost for each language. Such an approach is interesting if at *inference* time the data-point is very short, for which translation might be poor. However, in our use-case (as opposed to, for example, classifying tweets) each data-point is a rather large document that additionally is self-contained.
Of course, automatic translation engines are not perfect, and it is questionable if the introduction of errors will not render the extraction problem un-usable. We performed a benchmark study to evaluate this.

## 2.2.- Performance Evaluation

To study the impact of translation, we used the OPP-115 dataset; and performed paragraph classification of those privacy policy into one of the 9 annotated classes. As evaluation measure, we used F1-score (min 0, max 1) which is a metric of accuracy that considers both the precision and the recall of the model to be evaluated i.e. High values of F-score will guarantee not only that the extracted paragraphs will be annotated correctly with the corresponding GDPR element but also that all mentioned GDPR elements in the paragraph to be extracted.

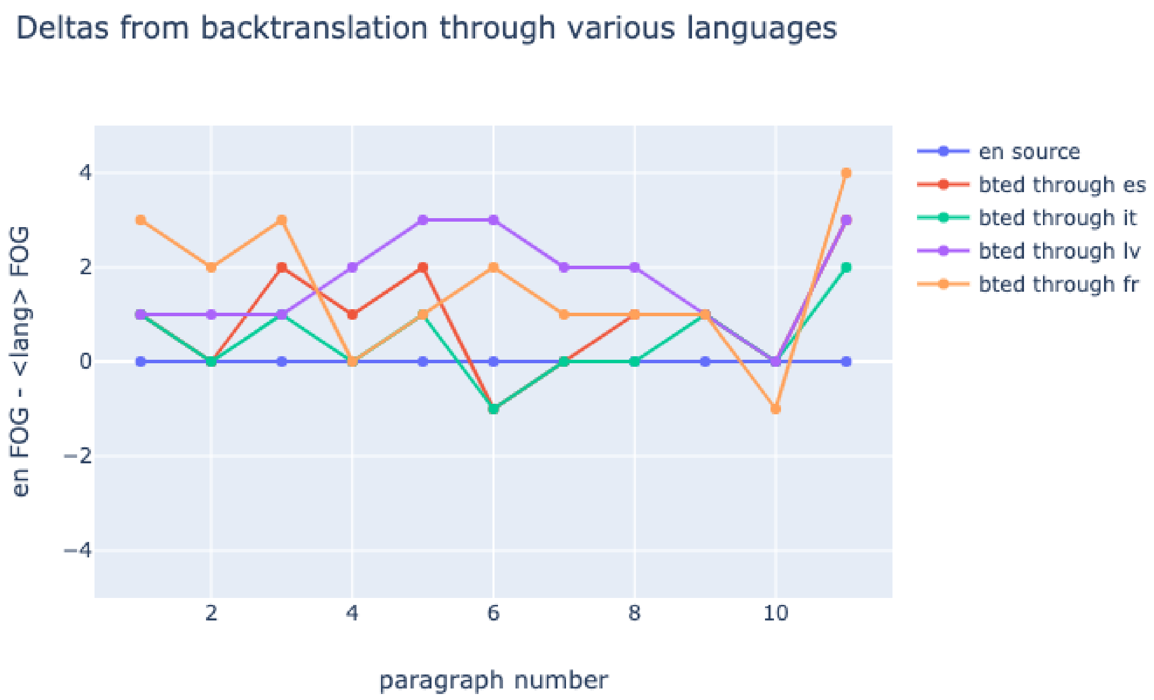
Insofar the GDPR extraction proceeds, we aim to measure the potential loss in performance created by the introduction of translation errors by following the process described previously.
However, due to the lack of annotated privacy policies in a second language, we proceeded as follows:
Original English policies were translated into a second language, Italian in this case. We then assumed that this translated policy constituted a request from a micro-enterprise to analyze its GDPR-element, and used this as entry for the data process desc It was therefore translated (back) into English, and then the model trained on English data run on top of it.
This was repeated twice, once with a commercial US-based translation engine; and one with a trained neural-machine translation model in house (details below).
In the following table we can see the impact of the translation model:

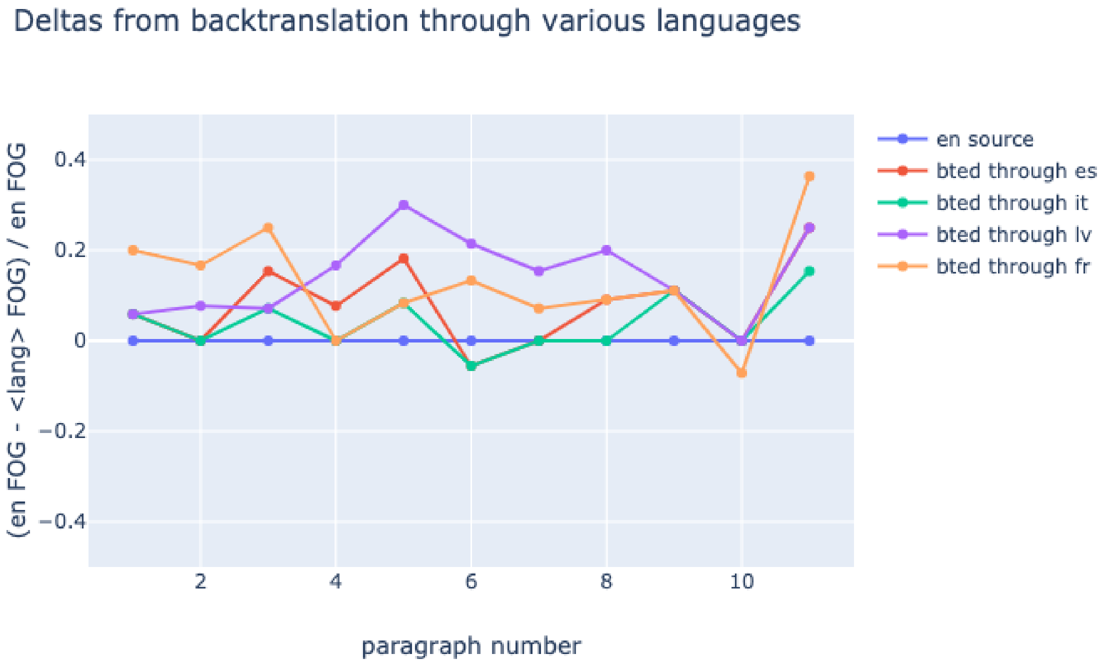
| | Original | Commercial MT system - backtranslated | EuroParl |
|---|---|---|---|
| First Party Collection/Use | 0.80 | 0.79 | 0.72 |
| Third Party Sharing/Collection | 0.78 | 0.77 | 0.68 |
| User Choice/Control | 0.60 | 0.57 | 0.47 |
| Data Security | 0.59 | 0.57 | 0.48 |
| International and Specific Audiences | 0.77 | 0.76 | 0.64 |
| User Access, Edit and Deletion | 0.49 | 0.48 | 0.18 |
| Policy Change | 0.69 | 0.66 | 0.54 |
| Data Retention | 0.18 | 0.16 | 0.08 |
| Do Not Track | 0.76 | 0.59 | 0.04 |
| Other | 0.70 | 0.70 | 0.64 |
| Average wo Other | **0.63** | **0.59** | **0.43** |
| Average | **0.64** | **0.60** | **0.45** |

The classification model in this case was a linear one (logistic regression, with l-2 regularization) - reason for which the results differ slightly from the previous table. We report the F1 score per category. “Average” is the macro-average, while “Average wo Other” is the macro-average excluding the category *Other*.
As can be seen, using the commercial, state-of-the-art model, results in a very low performance drop. The particularly high drop in “Do Not Track” is probably due to the specific naming that is used to refer to that category: the classification model probably learnt to over fit to this naming, and it might get miss-translated in the double translation loop.
Note that the fact there is this double translation could indicate that the actual performance drop will be actually lower, as in our evaluation setting there are two translations which could introduce errors. On the other hand, not all results of translating from already translated content can carry over to the translation of original content [Popel2018 ,Dowmut2019].

Using an in-house model however performs much worse, despite using a state-of-the-art architecture [Transformer]. We believe the gap could be crossed making a better usage of data, hyper-parameter searches and fine-tuning: for this model we used the EuroParl corpus [Kohn2005], the multi-lingual proceedings of the European Parliament. Performance could be improved by cleaning and enriching that data-set, as for the moment we only used 1.7M parallel sentences.
Moreover, the performance on test data of this same type is lower than the state-of-the-art (our model obtains 20.4 BLEU points for EN->IT, compared to > 25 points for fine-tuned systems). Regarding the good performance obtained by commercial system, and the little room for improvement left between that and in-language policies (a proxy for the results obtained with language-specific annotations), we decided however not to focus on improving an in-house translation engine and rely on external ones.

We performed an equivalent study for readability.
Again, the documents were pivot-translated through a second language, and we measured the difference between the FOG score of the original document and the FOG score of the pivot-translated documented. The following graph is an representative example, in this case the first paragraphs of a literary book:





The figure shoes the difference of the FOG score of the original text and the pivot-translated, and the normalized difference with the original FOG score. As can be appreciated, the difference is rather small, and the double translation has an apparent effect of simplifying slightly the language.

Note that – as with the GDPR extraction – the added noise by translation is amplified here by double translating. In practice, when a single translation is done, we would expect a smaller error.

## 2.3.- Adding additional languages

One advantage of our proposed approach is that adding another language is rather straightforward. The service which is currently used (Google Translate) supports 109 languages.

The current service performs language identification only for a subset of those, namely:

- English
- Spanish
- Italian
- Latvian
- Portuguese
- French

Adding more languages is rather straightforward and just consists on adding it in the corresponding list

## 2.4.- Limitations

Working directly on the translated text have other impacts as well. We studied in the preceding section the impact on what we consider the most important aspect, the one of classifying paragraphs into a GDPR aspect, and concluded that the impact is minor.

We list here two other possible consequences:

1. Finer-grained extraction. In SMOOTEXT, there are two classes of finer-grained information that are extracted: personal data and company names. In general, none of them should be hugely impacted by the translation: we expect the personal data item to be translated correctly, and major companies (the one most probably be present) as well. However, we have observed wrong translations with lesser-known brands
2. The use of a third service adds additional costs. For the moment, those are rather low ($20 for 1 million characters), but have to be factored in if SMOOTH is to be commercialized

## 3.- CONCLUSION

This deliverable describes the multi-lingual approach taken by SMOOTEXT. In it, we describe the use of a third-party translation service, and the results of benchmarking studies. Those let us to conclude that the loss in performance is negligible compared to the additional costs of developing true multi-lingual solutions, and the additional investment such a solution would require to extend it to more languages.
The current approach has the advantage that it can be extended easily and without much additional overhead and without any specialized knowledge.