A probabilistic approach to language structure

1. Introduction

The translation of international legal instruments requires a high degree of accuracy and consistency. With the increasing demand for multilingual texts, translation memory tools and research on parallel corpora have proved to be particularly useful for the translation of repetitive documents, as well as for those subject to an evolutive drafting process and production. Moving from this assumption, the present study aims at predicting translation equivalents with the help of a probabilistic approach. Data (1,404,723 words) consists of a multilingual parallel corpus in four languages: English, French, German and Italian. All the documents have been taken from the EU secondary legislation and include Regulations, Decisions, Directives and Recommendations, chosen between the years 2001-04. Texts are all strictly ‘normative’ and discourse is expected to be precise with minimum scope for ambiguity. The main focus is prescriptive statements, namely deontic norms (permission, obligation, prohibition) and constitutive performatives. Their formulation is highly standardized in English both within and outside the EU context. On the other hand, their expression in other languages is more vague and extensive, with potential consequences on the translation of norms. Bearing these remarks in mind, our objective is: 1) to evaluate the degree of prescriptive standardization with reference to English and the other three languages, and 2) to predict translation equivalents in the other languages under the condition that (i) English legal drafting is highly standardized, (ii) the EU and the main English drafting guidelines tend to use modal verbs in prescriptive statements (iii) text types under examination are repetitive and reusable (iv) the four EU instruments can be more or less binding. English modals are used as the main entry point and entropy analysis is exploited to measure the number of alternatives (degree of uncertainty) occurring in the other three languages. By adding knowledge to a system (e.g. a more standardized formulation), one reduces the number of alternatives (uncertainty), which leads to a decrease of entropy and to a gain of information in the expression of the norm. Although language phenomena cannot be fully described, the results of this analysis have empirically proved that given a set of conditions, certain linguistic structures are more easily predictable than other when comparing several languages. These types of analysis can foster research in language testing, evaluation, and in the development of automated translation’s tools.

2. Theoretical background and probabilistic variables

Normative sentences can take different grammatical and lexical forms. The main verb usually determines the type of norm that is to be expressed (e.g. obligation, permission, empowering, prohibition) and following Austin (1962) can be ‘explicit’ (order, permit, forbid) or ‘implicit’ (shall, may, must). To this point, English, legal drafting appears to be highly standardized and encompasses specific rules for the use of modal verbs.
(Coode 1843, Driedger 1976, Thornton 1996, The EU Inter-institutional Style Guide). For the purpose of this specific analysis, we chose a EU Regulation sub-corpus in 4 languages (334,425 words in total). With the help of Paracon we initially retrieved all the English modal verbs inherent to the expression of norms, together with their translation equivalents in French, German and Italian. Corpus findings confirmed the predominant occurrence of the English modals shall, must, may and to a lesser extent can and should. The other three languages showed a variety of linguistic forms (alternatives) that are grouped as follows: (a) indicative (mostly present), (b) modal verbs, (c) verbal periphrases, (d) lexicalized modal expressions, (e) ellipsis or zero correspondence. In order to apply probabilistic treatment, we selected 5 categories of expressions corresponding to each modal verb or language alternative. They include: (a) constitutive norms and obligations, (b) logical necessity, (c) permission and authorization, (d) capability and (e) non-binding norms. The probabilistic approach starts with determining the frequency of occurrence \( n_i \) of each linguistic form (modals and other linguistic alternatives) associated with a category. A probability variable \( p_i \) is then derived from the estimated proportion of occurrence of a particular modal verb in the corpus. This is given by \( p_i = n_i / n \), where \( n \) is the total number of modals or their equivalents.

In French, German and Italian texts of the same document it is expressed as:

\[ p_1 = P_{\text{indicative}} + P_{\text{mv}} + P_{\text{vp}} + P_{\text{me}} + P_{\text{ellipses}}; \quad p_2 = P_{\text{indicative}} + P_{\text{mv}} + P_{\text{vp}} + P_{\text{me}} + P_{\text{ellipses}} \text{ etc.} \]

The modal shall is the most frequent auxiliary to impose obligations and binding norms while may is used to express permissions and authorizations. From a statistical point of view, variations in the linguistic forms of expression are possible due to the number of alternatives inherent in a language.

In the information theory, the metric used to measure information is known as entropy \( h \) and corresponds to a degree of uncertainty (a shortage of information due to the large numbers of alternatives) in a message. According to Shannon (1949), the information value or content \( h(p) \) is dependent on the probability of occurrence \( p \) of an event. This dependence is described by the formula: \( h(p) = - \log (p) = \log (1/p) \). Different languages in their repertoire have different linguistic forms, and therefore each mode carries different probability values. The more precise or standardized the system is, the less its entropy (e.g. the number of alternatives) is. Considering the EU Regulation document, the probability \( p_i \) of occurrence of each individual form (e.g. ind, mv, vp, me and ellipses) belonging to the 5 categories of norms is linked to a certain information value. The sum of these probabilities over all the distinct forms produces different results and hence different information values. The expected information content of a system is the sum of the information contents weighted by the probabilities of the respective constituent attributes. This sum is expressed as follows:

\[ H = \sum p_i \cdot h(p_i) \]

---

1 The whole study includes the four EU secondary legislation text type. For reason of space, we are presenting only data concerning the Regulation that is the most binding text out of the four.
where $p_i$ is the probability belonging to a certain category of expression $i$ (e.g. (a) constitutive norms and obligations, (b) logical necessity, (c) permission and authorization etc).

3. Entropy results

Entropy evaluation has been carried out at two levels: (1) entropy measures with respect to the 5 categories of expression (constitutive and performative norms, logical necessity, permission and authorization, capability and non-binding norms) as in Figure 1; and (2) overall entropy in the EU Regulation as well as other Secondary Legislation corpus according to the four text types.

Figure 1. Entropy measures in the EU Regulation (5 category of expressions)

The different heights for French, German and Italian reflect in each cluster the use of alternatives in these languages compared to the English system, where only modal verbs have been considered. In the material examined, the overall entropy of a language is the sum of the separate entropy measures with respect to the modal verbs as found in the different linguistic versions of the corpus. Extending this approach to norm formulation in the 4 types of EU documents, we were able to determine the overall entropy for each text type in the 4 languages and compare their linguistic alternatives.

Figure 2. Overall entropy in the EU Secondary Legislation Corpus
In this case entropy results provide measures of particular EU text types and can confirm inference on their degree of mandatory force.

4. Discussion

By applying entropy analysis, four language systems have been compared in the attempt to ascertain the degree of prescriptive standardization occurring in a relative small corpus of EU normative texts. English modal verbs serve here as a parameter for their consolidated position in the international legal drafting and also for English being the main working language of the EU. Its entropy results are therefore lower when compared to the other three languages and do not constitute a relevant asset to the goal of this analysis. From Figure 1, it is possible to remark that the formulation of logical necessity, permissions and authorization and capability is quite standardized in the four languages. For each English modal, it is to be expected an equivalent modal verb in French, German and Italian. This is not the case of the constitutive and performative norms where a hypothetical translator or translation tool is exposed to a considerable variation. The three languages account for a larger number of alternatives against the English shall, with an overall preference for the present indicative. This is partly due to the widespread use of shall in the English EU drafting, but also to the inherent complexity of these norm types, which can indicate definitions, constitutive performatives, obligations and prohibitions. It is also interesting to remark that although French and Italian boast a similar semantic and grammatical language system, entropy results are not as close as in the other categories. This is probably due to the more prominent role of French in the EU context and, hence to an increased standardization. The lower entropy results in the German non-binding norms are due instead to the established position of the conditional form of the modal sollen when formulating general guidelines and recommendations. Figure 2 gives entropy results on the basis of the four EU legal instruments and text types. Regulations and Decisions present in the four languages lower entropy because the direct applicability of norms requires more precision and a more standardized formulation. Again, French Regulations and Decisions account for slightly minor entropy than Italian and German. On the other hand, the Recommendations text type highlights several alternatives above all in French and Italian, whose figures look closer in these respects. In conclusion, the application of probabilistic theories has proved that given certain conditions, it is possible to predict with some degree of certainty the occurrence of a particular factor. When applied to parallel texts, entropy analysis can delve into theoretical issues about language structure, but can also provide a resourceful ground of applications for language testing and evaluation of machine translations and other automated translation tools.

References
Coode, G. 1843. Legislative Expressions. Appendix to the Report of the Poor Law Commissioners on Local Taxation. Published separately 1845, 2nd Ed. 1852.