**Received**: 11/02/2021
**Accepted**: 26/04/2021
**DOI:** 10.17533/udea.mut.v14n2a10

MUTATIS MUTANDIS
ISSN 2011-799X

# *Cadlaws* – An English–French Parallel Corpus of Legally Equivalent Documents[1]

*Francina Sole-Mauri*
francine@uoc.edu
https://orcid.org/0000-0003-4589-0580
Universitat Autònoma de Barcelona, Spain

*Pilar Sánchez-Gijon*
Pilar.Sanchez.Gijon@uab.cat
https://orcid.org/0000-0001-5919-4629
Universitat Autònoma de Barcelona, Spain

*Antoni Oliver*
aoliverg@uoc.edu
https://orcid.org/0000-0001-8399-3770
Universitat Oberta de Catalunya, Spain

**Abstract**

This article presents *Cadlaws*, a new English–French corpus built from Canadian legal documents, and describes the corpus construction process and preliminary statistics obtained from it. The corpus contains over 16 million words in each language and includes unique features since it is composed of documents that are legally equivalent in both languages but not the result of a translation. The corpus is built upon enactments co-drafted by two jurists to ensure legal equality of each version and to reflect the concepts, terms and institutions of two legal traditions. In this article the corpus definition as a parallel corpus instead of a comparable one is also discussed. *Cadlaws* has been pre-processed for machine translation and baseline Bilingual Evaluation Understudy (BLEU), a score for comparing a candidate *translation* of text to a *gold-standard* translation of a neural machine translation system. To the best of our knowledge, this is the largest parallel corpus of texts which convey the same meaning in this language pair and is freely available for non-commercial use.

**Keywords:** Corpus construction; parallel corpus; neural machine translation (NMT); English–French translation; *Cadlaws*.

1 This paper is part of the Ph. D. thesis "On the effect of human produced corpora on Neural Machine Translation" written by Francina Sole-Mauri in Translation and Intercultural studies at Universitat Autònoma de Barcelona, supervised by Pilar Sánchez-Gijon and Antoni Oliver.

## Cadlaws: un corpus paralelo de documentos jurídicos equivalentes inglés-francés

**Resumen**

Este artículo presenta *Cadlaws*, un nuevo corpus en los pares de lenguas inglés y francés, creado con base en documentos legales canadienses. Describe el proceso de construcción del corpus y las estadísticas preliminares que se obtuvieron de aquél. Este corpus contiene más de 16 millones de vocablos en cada idioma e incluye características únicas, pues está conformado por documentos equivalentes desde el punto de vista jurídico en ambos idiomas como lengua de partida. El corpus se basó en autos legales redactados de manera conjunta por dos juristas para garantizar la equivalencia jurídica de cada versión y reflejar los conceptos, términos e instituciones de dos tradiciones del derecho. En este artículo, también se estudia la definición de corpus como corpus paralelo en oposición al corpus comparable. *Cadlaws* se procesó previamente para traducción automática y el suplente de evaluación bilingüe de referencia (BLEU, por sus siglas en inglés), un puntaje que sirve para comparar un texto presentado como candidato para la *traducción* de un texto contra una traducción considerada *patrón de referencia* en un sistema de traducción automática neuronal. Hasta donde sabemos, este es el corpus paralelo de textos con el mismo significado en este par de lenguas más extenso que existe, y ofrece libre acceso para uso no comercial.

**Palabras claves:** construcción de corpus; corpus paralelo; traducción automática neuronal (TA); inglés-francés; *Cadlaws*.

## Cadlaws - Un corpus parallèle anglais-français de documents juridiquement équivalents

**Résumé**

Cet article présente Cadlaws, un nouveau corpus anglais-français construit à partir de documents juridiques canadiens. L'article décrit le processus de construction du corpus ainsi que les statistiques préliminaires obtenues à partir de celui-ci. Le corpus contient plus de 16 millions de mots dans chaque langue et présente des caractéristiques uniques puisqu'il est composé de documents juridiquement équivalents dans les deux langues mais qui ne sont pas le résultat d'une traduction. Le corpus est construit à partir de textes co-rédigés par deux juristes afin de garantir l'égalité juridique de chaque version et de refléter les concepts, termes et institutions de deux traditions juridiques. Dans cet article, la définition du corpus comme un corpus parallèle au lieu d'un corpus comparable est également discutée. Cadlaws a été prétraité pour la traduction automatique et offre la valeur d'évaluation BLEU (Bilingual Evaluation Understudy), un score permettant de comparer une traduction avec la norme d'un système de traduction automatique neuronal. À notre connaissance, il s'agit du plus grand corpus parallèle de textes véhiculant le même sens dans cette paire de langues et il est disponible gratuitement pour une utilisation non commerciale.

**Mots-clés :** construction de corpus ; corpus parallèle ; traduction automatique neuronale (NMT) ; anglais ; français ; Cadlaws.

## 1. Introduction

Machine learning has driven advances in many fields including translation. Way (2018) estimated that in 2012 Google was translating around 75 billion words per day while in 2016 its average daily volume was about 143 billion, representing nearly a doubling in just four years. Neural Machine Translation (NMT) integrates artificial neural networks that directly transform a source sentence to a target sentence. An introduction and review of Neural Machine Translation can be found in the work of Stahlberg (2020), where it is discussed how NMT has become the de facto standard for large-scale machine translation, going back to origin of NMT to word and sentence embeddings and neural language models. The most commonly used building blocks of NMT architectures are explained along with the advantages and disadvantages of several design choices with respect to decoding, training, and segmentation.

Although NMT is a relatively new paradigm in the translation field, having been first explored towards the end of 2014, the main technology industries such as Google, Microsoft, and Yandex translation services have all switched to using NMT. One such NMT system, OpenNMT[2], was released in 2016 by the Harvard Natural Language Processing group (Klein et al., 2018).

So far, the focus has been on the technological development, improvement of the algorithms of the systems, and the combination of systems to achieve faster responses or more efficient resource utilization. However, the three biggest companies investing in the field of machine translation (MT), namely Google®, Facebook®, and Microsoft®, which also own the largest corpora, have started to focus the debate recently on a phenomenon called "translationese" (Freitag et al., 2019). Translationese is a feature detected in translated texts (Baker, 1993) which present a poorer use of language than originally written texts in the same language, due to the influence of both the source text and the source language. In comparison with originally written texts, translated texts tend to display lower rates of lexical and syntactical variety, poorer choice of linguistic structures, and a higher level of explicit vocabulary and/or concepts.

Parallel corpora are widely used for training NMT systems, which directly model the probability of a target-sentence given a source-language sentence. Search and assessment of candidate parallel texts is usually made based on the quantity rather than the quality due to the vast amount of data that NMT requires. Thus, most of the research done in NMT is built on corpora that, although in a more formal language and professionally translated, have already been produced using Computer Aided Translation (CAT) tools and by MT, to some extent at least. There are a number of English and French parallel corpora in the legal domain, such as the English-French *Hansard* Corpus (Germann, 2001), the Europarl Corpus[3] (Koehn, 2005), the JRC-Acquis (Steinberger et al., 2006), or the United Nations Corpus (Ziemski et al., 2016).

The lack of resources to train NMT has led to the use of comparable sources to provide parallel corpora to train NMT for low-density language pairs or to the use of sources that do not belong to the legal and political sphere. A review of the large body of research on mining parallel sentences in collections of monolingual texts from comparable corpora can be found at Schwenk

2 OpenNMT is an open-source ecosystem for neural machine translation and neural sequence learning available at https://opennmt.net/ (Accessed April 7th, 2021).

3 European Parliament Proceedings Parallel Corpus 1996-2011, available at: https://www.statmt.org/europarl/ (Accessed April 7th, 2021).

et al. (2019). The methodology used to extract a comparable corpus is different from that of a parallel one (see, for example, Hewavitharana & Vogel, 2016; Rapp et al., 2016).

This paper presents *Cadlaws*, a new English-French parallel corpus built from Canadian enactments, which is legally equivalent in both languages but not as the result of translation. The corpus compilation process is detailed in Section 2 and is followed by an analysis of the corpus features (Section 3). Then, the corpus evaluation in an NMT system is reported (Section 4). In Section 5, possible uses of the corpus are discussed. Finally, the conclusions and the references close the paper.

## 2. Corpus Compilation

The texts that together compose the *Cadlaws* corpus were downloaded from the Justice Department website of the Government of Canada. With respect to permission and access to the data, the website states the following:

> Anyone may, without charge or request for permission, reproduce enactments and consolidations of enactments of the Government of Canada, and decisions and reasons for decisions of federally constituted courts and administrative tribunals, provided due diligence is exercised in ensuring the accuracy of the materials reproduced and the reproduction is not represented as an official version. (Canada Government, Department of Justice, s.f.)

So, in accordance with the requirement, it must be noted that this work has been done using a copy of the official work, downloaded in January 2019, and that this reproduction has not been produced in affiliation with the Government of Canada or with the endorsement thereof.

The downloaded materials contain every enactment published in English and French from January 2001 to December 2018, broken into 5,609 files in each language and with a total of 28,870,027 words in English and 35,109,593 words in French.

English and French versions were collected in Extensible Markup Language (XML)[4] and were characterised by a complicated structural hierarchy. In order to preserve high quality texts, the tables, notes, figures, and the style markers were discarded as well as all the laws that had a different structure in each language. One special characteristic of the corpus is that each paragraph, while conveying the same meaning in each language, can have a number of sentences that may be different in English and French. Sentences were considered as the basic units and were automatically split by using the document's structure tags with one sentence per line. If the number of lines in a given law differed between the two languages, then the data was discarded for quality reasons since the equivalence in meaning and the alignment could not be guaranteed.

The text was tokenised with the *Moses* tokeniser (Koehn et al., 2007). The alignment was done using *Hunalign* (Varga et al., 2005), which takes tokenised sentence-segmented texts and outputs a sequence of bilingual sentence pairs. In a first step, *Hunalign* was used to match the pairs of sentences. Then, an automatic dictionary based on this first alignment was built. Finally, the algorithm realigned the sentences combining sentence-length information with the dictionary in a second iteration.

The sentence alignment was stored in one file per law with both languages. Then, the final sentence-aligned data was stored in one file

4 Extensible Markup Language is used to describe data. It is a programming language that enables information exchange between otherwise incompatible systems.

**Table 1.** Characterization of the New *Cadlaws* Corpus and the *Canadian Hansard*

| | *Cadlaws* | | *Hansard* | |
|---|---|---|---|---|
| | **English** | **French** | **English** | **French** |
| Lines | 741,071 | 741,071 | 1,070,259 | 1,070,259 |
| Sentences | 770,174 | 816,511 | 1,161,967 | 1,103,985 |
| Clauses | 1,721,601 | 1,750,732 | 1,790,110 | 1,966,843 |
| Words | 16,878,655 | 17,086,083 | 16,366,523 | 17,540,577 |
| Tokens | 17,016,518 | 18,243,835 | 16,460,385 | 18,701,654 |
| Words/Sentence | 21.9 | 20.9 | 14.1 | 15.9 |
| Unique words | 127,271 | 110,672 | 60,557 | 77,412 |

per language. The size of the sentence-aligned corpus is roughly 16 million words in around 740 thousand segments. This corpus contains only sentence pairs and even though the order of the sentences is the same as in the original for each legal document, the one-to-one, many-to-one, or many-to many alignments that were filtered out left some gaps.

The detailed statistics of the *Cadlaws* corpus are summarized in Table 1, which lists the words, tokens, average sentence length, and the vocabulary size of the corpus. All scores are provided for lowercased data and tokens were counted after processing with the Moses tokenizer. To better illustrate the special features of this corpus, the same data for the Canadian *Hansard* of the House of Commons for the period 1997-2000 is also provided (Germann, 2001).

Both corpora are similarly sized with approximately 17 million words in each language; but while the *Hansard* has only 60 to 77 thousand types (unique words), the new corpus doubles the number of types. In addition, with respect to sentence length, *Cadlaws* has 21 words per sentence while the *Hansard* has about 15. Interestingly, the *Hansard* French version has longer sentences than the English one, while in the *Cadlaws* corpus, it is the English version which has the higher number of words per sentence.

## 3. Corpus Features

The *Cadlaws* corpus is available for research and non-commercial use under a Creative Commons Attribution International Licence and can be fully accessed and downloaded.[5]

As mentioned, there are three characteristics that make *Cadlaws* unique. First, it is composed of legal documents that are equivalent in both languages but not as the result of translation; second, the sentences are very long compared to other corpora in the legal vocabulary field; and finally, it shows greater variety in terminology and expressions as it is based upon legal documents that have to respond to the requirements of two languages and two legal traditions.

Since 1978 Canadian federal laws have been co-drafted by two jurists, one francophone and one anglophone, who simultaneously draft the French and English versions of a legislative text to ensure equality of the two official languages (McLaren, 2014). In the beginning, as discussed by Regan et al. (2011, pp. 218-219), the process of co-drafting proved to be inappropriate since it focused on bilingualism more than on bijuralism and imposed common

5 *Cadlaws* corpus is available at https://ddd.uab.cat/record/238990?ln=en (Accessed April 7th, 2021).

law conventions on the French language text of federal legislation. The subsequent consolidation of terminology by both the provincial and federal governments and the enactment of a new civil code of Quebec in 1994 enhanced the policy of legislative bijuralism so that each language's version of any federal law reflected the concepts, terms, and institutions of the two legal systems in Canada. Thus, jurists co-drafting in both languages have to harmonize the concepts not only of two languages but of two legal traditions (see Regan et al., 2011, pp. 219-220).

In 1995, The Department of Justice Canada adopted the *Policy on Legislative Bijuralism*, which aimed to formally recognize the obligation to make federal legislative texts accessible to the various legal audiences of federal law in Canada, notably civil law and common law jurists, in both English and French. In 1999, legislative bijuralism in both linguistic versions was imposed as an obligatory drafting norm for all federal legislative texts as the *Cabinet Directive on Law-Making* was amended. Finally, in 2001, the *Federal Law-Civil Law Harmonization Act*, No. 1 (S. C. 2001, c. 4) enacted sections 8.1 and 8.2 which defined bijural rules of interpretation and the practice of revising draft bills and regulations was implemented. *Cadlaws* is therefore built from enactments published from 2001 onward to ensure that not only bilingualism but bijuralism is ensured in the language chosen to ensure that both versions are legally equivalent.

Neither version is a translation of the other, but they convey the same meaning in a given paragraph. Thus, while each segment has the same meaning, it can be formed by individual sentences that may not be considered a perfect translation in the other language, despite being legally equivalent. An extract (Example 1) from the *Controlled Drugs and Substances Act* (S.C. 1996, c. 19; last amended on 19/09/2019) exemplifies this fact (italics used in the original):

**Example 1**

| English | French |
|---|---|
| Her Majesty, by and with the advice and consent of the Senate and House of Commons of Canada, enacts as follows:<br>Short Title<br>This Act may be cited as the *Controlled Drugs and Substances Act*<br>Interpretation<br><br>2(1) In this Act, [...] | Sa Majesté, sur l'avis et avec le consentement du Sénat et de la Chambre des communes du Canada, édicte:<br>Titre Abrégé<br>*Loi réglementant certaines drogues et autres substances.*<br>Définitions et interprétation<br>2(1) Les définitions qui suivent s'appliquent à la présente loi. […] |

In this example, it can be seen how "[i]n this act" in the English version is equivalent to "[l]es définitions qui suivent s'appliquent à la présente loi" in French. Both sentences convey the same meaning and it cannot be considered as one language being the translation of the other one. As Justice P. Viau explained it, "deux langues, c'est d'abord deux styles, en matière de rédaction du moins. Et ailleurs aussi. Lois françaises et lois anglaises sont conçues différemment. Les mêmes idées ne se dissimulent pas de la même façon derrière des mots d'ont le sens et la portée sont parfois difficiles à cerner" (Gervais & Séguin, 2001).

In translation studies, a parallel corpus is usually considered as a corpus that contains source texts and their translations while a comparable corpus contains components that are collected using the same sampling frame and similar balance and representativeness (see McEnergy, 2003, p. 450, for example, for further discussion on the topic). *Cadlaws*, however, is formed by two source texts and, given that neither of the versions is a translation of the other language, it is considered to be a parallel corpus for several reasons. Firstly, because it is legally equivalent in both versions. Secondly, for a comparable corpus the sampling frame is essential to ensure both languages proportion, genre or domain.

In *Cadlaws*, the sampling frame is irrelevant because the corpus components at segment level ensure the exact same meaning. Lastly, a parallel corpus compilation process has been followed.

The nature of the legal documents used to build the corpus provides *Cadlaws* with another special feature: the average sentence length is, in both languages, larger than in the most common corpora usually available for this language pair. Example 2 from the *Good Samaritan Drug Overdose Act* (S. C. 2017, c. 4) shows this fact, as the segments have a length of 87 and 65 words in English and 101 and 77 words in French.

These long sentences are not abnormal and represent the usual use of the languages in this specialized field. Normally such long sentences are removed during corpus preparation to train NMT systems. This means that the segment length, since it may be a sentence or a combination of them to maintain the meaning in both languages, is longer than that of other parallel corpora in the legal or political domain.

Another unique characteristic of Cadlaws is that it is based upon legal texts that have been written with special attention to the vocabulary and meaning, and as a result, it shows greater variety in terminology and expressions. Moreover, the drafting of enactments has to respond to Canadian bijuralism and respect the coexistence of the civil law and common law legal traditions in Canada. Allard (2001) discussed how this bijural nature of the Canadian legal system, along with the obligations that derive from bilingualism, has an unquestionable impact on the drafting of federal legislation to ensure that in both English and French, the common law and civil law legal traditions are acknowledged.

In Example 3, the extracts from the *Bank Act* (S.C. 1991, c. 46) show how drafting is adapted to take into account at least four legal audiences

**Example 2**

| English | French |
|---|---|
| No one who seeks emergency medical or law enforcement assistance because that person, or another person, is suffering from an overdose, or who is at the scene upon the arrival of the assistance, is to be charged with an offence concerning a violation of a pre-trial release, probation order, conditional sentence or parole relating to an offence under subsection 4(1) if the evidence in support of that offence was obtained or discovered as a result of that person having sought assistance or having remained at the scene. [87 words] | La personne qui demande, de toute urgence, l'intervention de professionnels de la santé ou d'agents d'application de la loi parce qu'elle-même ou une autre personne est victime d'une surdose ou qui se trouve sur les lieux à l'arrivée des secours ne peut être accusée d'une infraction en lien avec la violation de conditions de mise en liberté provisoire, d'une ordonnance de probation, d'une ordonnance de sursis ou des modalités d'une libération conditionnelle relativement à une infraction prévue au paragraphe 4(1) si la preuve à l'appui de cette infraction a été obtenue ou révélée parce que cette personne a demandé du secours ou est restée sur les lieux. [101 words] |
| Any condition of a person's pre-trial release, probation order, conditional sentence or parole relating to an offence under subsection 4(1) that may be violated as a result of the person seeking emergency medical or law enforcement assistance for their, or another person's, overdose, or as a result of having been at the scene upon the arrival of the assistance, is deemed not to be violated. [65 words] | Est réputée n'avoir jamais eu lieu la violation, relativement à une infraction visée au paragraphe 4(1), de conditions de mise en liberté provisoire, d'une ordonnance de probation, d'une ordonnance de sursis ou des modalités d'une libération conditionnelle qui résulte du fait que la personne a demandé, de toute urgence, l'intervention de professionnels de la santé ou d'agents d'application de la loi parce qu'elle-même ou une autre personne était victime d'une surdose ou est restée sur les lieux à l'arrivée des secours. [77 words] |

as the civil law and the common law in both English and French.

Federal statutes and regulations could be addressed as "[t]he simple co-existence of two legal traditions, the interaction between two traditions, the formal integration of two traditions within a given context" or "the recognition of and respect for the cultures and identities of two legal traditions" (Allard, 2001). The extracts from the Bank Act show how in each parallel sentence the same meaning is imposed.

**Example 3**

| English | French |
|---|---|
| Deemed control<br>(3) A person is deemed to control, within the meaning of paragraph (1)(a) or (b), an entity if the aggregate of<br>(a) any securities of the entity that are beneficially owned by that person, and<br>(b) any securities of the entity that are beneficially owned by any entity controlled by that person is such that, if that person and all of the entities referred to in paragraph (b) that beneficially own securities of the entity were one person, that person would control the entity. | Présomption de contrôle<br>(3) Pour l'application des alinéas (1)a) ou b), une personne est réputée avoir le contrôle d'une entité quand elle-même et les entités qu'elle contrôle détiennent la propriété effective d'un nombre de titres de la première tel que, si elle-même et les entités contrôlées étaient une seule personne, elle contrôlerait l'entité en question au sens de ces alinéas. |
| Powers restricted<br>(2) A bank shall not carry on any business or exercise any power that it is restricted by this Act from carrying on or exercising, or exercise any of its powers in a manner contrary to this Act. | Réserve<br>(2) La banque ne peut exercer ses pouvoirs ou son activité commerciale en violation de la présente loi. |
| No invalidity<br>16 No act of a bank or authorized foreign bank, including any transfer of property to or by a bank or authorized foreign bank, is invalid by reason only that the act or transfer is contrary to<br>(a) in the case of a bank, the bank's incorporating instrument or this Act; or<br>(b) in the case of an authorized foreign bank, this Act. | Survie des droits<br>16 Les faits de la banque ou de la banque étrangère autorisée, notamment en matière de transfert de biens, ne sont pas nuls au seul motif qu'ils sont contraires, dans le cas d'une banque, à la présente loi ou à son acte constitutif ou, dans le cas d'une banque étrangère autorisée, à la présente loi. |
| Idem<br>(3) Notwithstanding the existence of a pre-emptive right, a shareholder of a bank has no pre-emptive right in respect of shares to be issued<br>(a) where the issue of shares to the shareholder is prohibited by this Act; or<br>(b) where, to the knowledge of the directors of the bank, the offer of shares to a shareholder whose recorded address is in a country other than Canada ought not to be made unless the appropriate authority in that country is provided with information in addition to that submitted to the shareholders at the last annual meeting. | Idem<br>(3) Le droit de préemption ne s'applique pas, non plus, aux actions :<br>a) dont l'émission est interdite par la présente loi;<br>b) qui, à la connaissance des administrateurs de la banque, ne devraient pas être offertes à un actionnaire dont l'adresse enregistrée est dans un pays étranger, sauf s'il est fourni aux autorités compétentes de ce pays des renseignements autres que ceux présentés aux actionnaires à la dernière assemblée annuelle. |
| Donated shares and membership shares (3) A bank may accept from any shareholder or member a share or membership share, as the case may be, of the bank surrendered to it as a gift, but may not extinguish or reduce a liability in respect of an amount unpaid on any such share or membership share except in accordance with section 75. | Donation d'actions et de parts sociales (3) La banque peut accepter toute donation d'actions ou de parts sociales, mais ne peut limiter ni supprimer l'obligation de les libérer autrement qu'en conformité avec l'article 75. |

However, as discussed by Justice L. P. Pigeon, English drafting "subordinates every consideration to the search for precision. It attemps to say all, define all, to intimate nothing, and to never assume the intelligence of the reader." In the French drafting, "one tries to find the precise word, and to formulate a general rule" (Gervais and Séguin, 2001).

Each concept or matter expressed in each version needs to be compatible with the legal system and it has an impact on the terminology used. Thus, *Cadlaws* reflects this bijural terminology depending on the language and legal tradition (Table 2) and this wider use of terminology is clearly seen by the number of unique words used in *Cadlaws* in comparison to the *Hansard*, which in English is almost doubled and in French increases its number by over forty per cent.

## 4. Corpus Evaluation

The evaluation of the corpus was done using OpenNMT. For insight into the field of NMT, we refer the reader to overview papers such as Neubig (2017), Cromieres et al. (2017), and Popescu-Bels (2019). The architecture of the NMT was not optimized, since the aim of this work was to evaluate the corpus, not to develop the best performing NMT system. OpenNMT, like most of the models, follows a common sequence-to-sequence (seq2seq) learning framework (Sutskever et al. 2014), with attention as described by Bahdanau et al. (2014).

Neural Networks can be understood as a set of algorithms that are designed to recognize patterns. In a sequence-to-sequence model, word embedding is used to transform the source language, while a context vector is also maintained. It has three components: an encoder that transforms a source sentence into a list of vectors, one for each input symbol; a decoder that produces one symbol at a time until the end-of-sentence symbol is found; and an attention model that connects the encoder and decoder so that the decoder focuses on different parts of the source sentence during the decoding process. The complete model has to be trained end-to-end to minimize the negative log-likelihood of the training corpus.

Once aligned, the corpus was then divided into three sets: training (160,434 lines; English: 11,696,157 words; French: 11,678,091 words), testing (30,717 lines; English: 2,269,919 words; French: 2,241,899 words), and validation (11,041 lines; English: 701,741 words; French: 670,650 words). Tokenization for both languages was done using Moses (Koehn et al., 2007), which includes markers that allow simple deterministic detokenization and has simple language-independent tokenization rules. To better understand and interpret the behaviour of this corpus, a model was trained using the OpenNMT system (Klein et al., 2018) for the English-French pair.

Translation error analysis and quality estimation has been widely researched in the field

**Table 2** Bijural Terminology: English Civil Law, French Civil Law, English Common Law, French Common Law

| Civil law | | Common law | |
|---|---|---|---|
| **English** | **French** | **English** | **French** |
| Immovable | Immeuble | Real property | Bien réel |
| Hypothec | Hypothèque | Mortgage | Hypothèque |
| Lease | Bail | Lease | Bail |

of MT. Fully manual error analysis, as used for example by Farrús et al. (2011), Comparin (2017), and Daems et al. (2017a) is slow, expensive, and sometimes inconsistent. Automatic evaluation uses metrics such as bleu (Papineni et al., 2001), which compute the similarity between a human supplied "gold standard" reference and the MT output based mostly on n-grams occurrence. One known issue with BLEU is that it is limited to working at sentence level. According to Way (2018), despite the well-known problems with BLEU, it has been by significant margin the most reported metric in papers involving MT experiments. This was the main reason for its use to evaluate the model's predictions although it might not be totally adequate for such a heterogeneous corpus as *Cadlaws*, and it might provide lower scores than other metrics.

Thus, *Cadlaws* corpus scored more than 6 points higher than the one obtained with a corpus that, although it is very similar in size, has a sentence length closer to other available corpora and has a narrower variety of vocabulary. It cannot at this point be definitively stated that the larger number of types is responsible for the better metrics that were observed. However, the similarity of both corpora suggests this may be the case.

## 5. Possible Uses

The *Cadlaws* corpus is a new tool which may be used in the future to deepen in the analysis of "translationese" and the differences between original and translated texts. Machine learning strategies have also been used to identify differences between translated and original texts, which challenges the notion of translation universals. "Translationese" is used to indicate the difference between original text and translated text, and a lot of research has been done to analyse whether "translationese" actually exists. Some authors believe that translated texts tend to be more explicit, more conservative, and less lexically dense than the original (Laviosa, 1998; Olohan, 2003; Puurtinen, 2003; Hansen-Schirra, 2011), or that translations tend to underrepresent the typical linguistic features of the target language if there is not an obvious equivalent in the source language (Tirkkonen-Condit, 2002; Mauranen, 2004). However, Baroni and Bernardini (2005) pointed out that these differences between original and translations might be due to confounding factors, such as gender-based difference between languages, where corpus construction might introduce confounding variables. Daems et al. (2017b) could not find user perception to identify post-edited texts from human translations, either perceived or measured. From these studies, it seems that readers are not capable of identifying the difference between an original text and a translated text (Tirkkonen-Condit, 2002; Baroni & Bernardini, 2005).

Interestingly enough, computers have successfully been trained to detect these differences by taking lexical and grammatical information into account (Baroni & Bernardini, 2005; Ilisei et al., 2010, Koppel & Ordan, 2011, Carter & Inkpen, 2012, Volanski et al., 2015, Laippala et al., 2015, Jiang & Tao, 2017, Rabinovich et al., 2017). Van Halteren (2008) identified the source language of medium-length speeches in the *Europarl Corpus* on the basis of frequency counts of word n-grams and identified source language markers. Kurokawa et al. (2009) used the *Hansard Corpus* in to investigate whether a text was an original or a translation. The authors went one step further to measure its impact on statistical machine translation, and they found that a model trained on five times fewer data yielded a similar performance because it was trained on the "right kind of data" (Kurokawa et al., 2009).

The cultural dimension in translation is not a new subject in translation, but it has been brought back with the wider implementation of artificial intelligence and the use of data set that do not necessarily take into account culture differences. *Cadlaws* could be an example of the sort of data needed for (N)MT to ensure that there is not only a language transfer but also another culture in the translations (Sánchez-Gijón & Piqué, 2020) and, in this case, to take into consideration the coexisting two legal cultures.

The *Cadlaws* corpus could also be used as valuable resource for linguistic research and natural language processing applications such as parallel term extraction, information retrieval, question answering, or word sense disambiguation in the legal domain. To the best of our knowledge, there is no other corpus formed by two source texts that are equivalent in meaning to the point of being legally equivalent.

## 7. Conclusions

*Cadlaws*, a new English-French parallel corpus, is publicly available for research and non-commercial use under a Creative Commons Attribution International Licence. The sentence-aligned corpus is roughly 16 million words in each language and contains around 740 thousand segments.

It has been compiled from Canadian legal documents and it has three unique features. First, it has been compiled from enactments that are legally equivalent in both languages but not the result of a translation. Canadian federal laws are co-drafted simultaneously in English and French, to not only guard the equality of both official languages, but to assure the legislative bijuralism. Thus, each language version has to reflect the concepts, terms, and institutions of the common law and civil law systems coexisting in Canada. Among other examples, it has been examined how for example, "[i]n this act" in the English version is equivalent to "[l]es définitions qui suivent s'appliquent à la présente loi" in French.

Second, the sentences are very long compared to other corpora in the legal vocabulary field for this language pair. Neither version is a translation of the other, but they convey the same meaning in a given paragraph. Hence, while each segment has the same meaning, it can be formed by individual sentences that may not be considered a perfect translation in the other language despite being legally equivalent. Normally such long sentences are removed during corpus preparation to train NMT systems, but *Cadlaws* is built precisely upon these long segments to provide the legally equivalent meaning in English and French.

Finally, *Cadlaws* is based upon legal texts that have been written with special attention to the vocabulary and meaning, and as a result, it shows greater variety in terminology and expressions. Moreover, each concept or matter expressed in each version needs to be compatible with the legal system and has an impact on the terminology used. *Cadlaws* reflects this bijural terminology depending on the language and legal tradition and this wider use of terminology is clearly seen by the number of unique words used. Jurists co-drafting in both languages are required to harmonize the concepts not only of two languages but of two legal traditions.

It has been discussed that *Cadlaws* is considered a parallel corpus. In translation studies, a parallel corpus is usually regarded as a corpus that encompass source texts and their translations while a comparable corpus consists of texts that are collected using the same sampling frame and similar balance and represent-

ativeness. *Cadlaws* is a parallel corpus because essentially its meaning is legally equivalent in both versions. It is built upon two equivalent source texts and neither of the versions is a translation of the other language. Furthermore, for a comparable corpus the sampling frame is essential to ensure both languages proportion, genre, or domain. In *Cadlaws*, the sampling frame is irrelevant because the corpus components at segment level ensure the exact same meaning.

Lastly, a parallel corpus compilation process has been followed. The enactments were downloaded in XML from January 2001 to December 2018. Sentences were considered as the basic units and were automatically split by using the document's structure tags with one sentence per line. Moses was used to tokenize, and the alignment was done using *Hunalign*. The order of the sentences is the same as in the original for each legal document, the one-to-one, many-to-one, or many-to-many alignments that were filtered out left some gaps.

The *Canadian Hansard* for the period 1997-2000 has been used to expose the special characteristics of *Cadlaws* since both corpora are equivalently sized and belong to the same language domain. It stands out that the number of unique words of this new corpus almost doubles that of the *Hansard* despite both being very similar in size. Concerning the sentence length, *Cadlaws* has 21 words per sentence while the *Hansard* has about 15. It has been exposed that the *Hansard* French version has longer sentences than the English one, while in the *Cadlaws* corpus, it is the English version the one with higher number of words per sentence, attributed to the different legal tradition.

The evaluation of the corpus was done using OpenNMT. The architecture of the NMT was not optimized since the aim of this work was to evaluate the corpus, not to develop the best performing NMT system. Benchmark baseline NMT results on this corpus scored a BLEU metrics of 44.82, an overall score more than 6 points higher than the one obtained with the *Canadian Hansard*. Further work might help determine whether the larger number of types or the sentence length is responsible for the better metrics. However, the similarity of both corpora suggests this may be the case.

Possible uses of the corpus have been considered. The special features of *Cadlaws* might be used to deepen the study of the differences between original and translated texts for linguistic research and natural language processing applications such as parallel term extraction, information retrieval, question answering, or word sense disambiguation in the legal domain. To the best of our knowledge, there is no other corpus formed by two source texts that are equivalent in meaning to the point of being legally equivalent.

## Acknowledgments

The work in this paper has been partially supported by the MINECO project Despite-MT (PID2019-10865RB-I00).

## References

Allard, F. (2001). The Supreme Court of Canada and its impact on the expression of bijuralism. *The Harmonization of federal legislation with the civil law of the province of Quebec and Canadian bijuralism* (Second Publication), Booklet 3, Ottawa, Department of Justice Canada.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. Conference paper presented at the 2015 International Conference on Learning Representations —ICRL—. https://arxiv.org/abs/1409.0473

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonel-

li (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233-252). John Benjamins. https://doi.org/10.1075/z.64.15bak

Baroni, M. & Bernardini, S. (2005). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), 259–274. https://doi.org/10.1093/llc/fqi039

Canada Government, Department of Justice. (s. f.). Justice Laws [Website]. https://laws-lois.justice.gc.ca/eng/ (Accessed April 7th, 2021).

Carter, D. & Inkpen, D. (2012). Searching for poor quality machine translated text: Learning the difference between human writing and machine translations. In L. Kosseim & D. Inkpen (Eds.), *Advances in artificial intelligence* (pp. 49–60). Springer. https://doi.org/10.1007/978-3-642-30353-1_5

Comparin, L.. (2017). *Quality in machine translation and human post-editing: Error annotation and specifications*. [M. A. thesis], Universidade de Lisboa, Lisbon. https://repositorio.ul.pt/handle/10451/27969?mode=full

Cromieres, F., Toshiaki, N., & Raj, D. (2017). Neural machine translation: Basics, practical aspects and recent trends. *Proceedings of the IJCNLP 2017, Tutorial Abstracts*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 11-13.

Daems, J., De Clercq, O., & Macken, L. (2017). Translationese and post-editese: How comparable is comparable quality? *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 89–103.

Daems, J., Vandepitte, S., Hartsuiker, R. J., & Macken, L. (2017). Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology*, *8*, 1282. https://doi.org/10.3389/fpsyg.2017.01282

Farrús, M., Costa-Jussà, M. R., Mariño, J. B., Poch, M., Hernández, A., Henríquez, C., & Fonollosa, J. A. (2011). Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan–Spanish language pair. *Language Resources and Evaluation*, *45*, 181-208. https://doi.org/10.1007/s10579-011-9137-0

Federal Law-Civil Law Act, N.°1. Royal Assent. Government of Canada https://www.parl.ca/DocumentViewer/en/37-1/bill/S-4/royal-assent (Accessed April 7th, 2021).

Freitag, M., Caswell, I., & Roy, S. (2019). APE at scale and its implications on MT evaluation biases. *Fourth Conference on Machine Translation* (WMT) (vol. 2, pp. 34–44). Florence. https://doi.org/10.18653/v1/W19-5204

Germann, U. (2001). *Aligned Hansard of the 36th Parliament of Canada*. Natural Language Group of the USC Information Sciences Institute. https://www.isi.edu/natural-language/download/Hansard/ (15th December, 2019).

Gervais, M.-F. & Séguin, M.-C. (2001). Some thoughts on bijuralism in Canada and the world. In Canada, Department of Justice, *The harmonization of federal legislation with the civil law of the province of Quebec and Canadian bijuralism*. Ottawa, Department of Justice. https://www.justice.gc.ca/eng/rp-pr/csj-sjc/harmonization/hfl-hlf/b2-f2/bf2.pdf (Accessed April 7th, 2021).

Hansen-Schirra, S. (2011). Between normalization and shining-through: Specific properties of English-German translations and their influence on the target language. In S. Kranich, V. Becher, S. Höder, & J. House (Eds.), *Hamburg Studies on Multilingualism* (pp. 133–162). John Benjamins Publishing Company. https://doi.org/10.1075/hsm.12.07han

Hewavitharana, S. & Vogel, S. (2016). Extracting parallel phrases from comparable data for machine translation. *Natural Language Engineering*, *22*(4), 549–573. https://doi.org/10.1017/S1351324916000139

Ilisei, I., Inkpen, D., Corpas Pastor, G., & Mitkov, R. (2010). Identification of translationese: A machine learning approach. In A. Gelbukh (Ed.), C*omputational linguistics and intelligent text processing*. 11th International Conference, CICLing 2010, Iaşi, Romania, 21–27 March. Proceedings (pp. 503–511). Springer. https://doi.org/10.1007/978-3-642-12116-6_43

Jiang, Z. & Tao, Y. (2017). Translation universals of discourse markers in Russian-to-Chinese academic texts: A corpus-based approach.

*Zeitschrift für Slawistik*, 62(1), 583–605. https://doi.org/10.1515/slaw-2017-0037

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., & Rush, A. (2018). OpenNMT: Neural Machine Translation Toolkit. *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas* (pp. 177–184, vol. 1, Research Papers). AMTA.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Conference Proceedings: The Tenth Machine Translation Summit* (pp. 79–86). Phuket, Thailand, AAMT.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion* (177–180). Prague, ACL. https://doi.org/10.3115/1557769.1557821

Koppel, M., & Ordan, N. (2011). Translationese and its dialects. Paper presented at the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland.

Kurokawa, D., Goutte, C., Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. *Proceedings of The Twelfth Machine Translation Summit International Association for Machine Translation* (pp. 81–88). Ottawa: Association for Machine Translation in the Americas.

Laippala, V., Kanerva, J., Missil, A., Missilä, A., Pyysalo, S., Salakoski, T., & Ginter, F. (2015). Towards the classification of the Finnish Internet Parsebank: Detecting translations and informality. In *Nodalida*. Linköping University Electronic Press.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, *43*(4), 557–570. https://doi.org/10.7202/003425ar

Lin, C.-Y., Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistic. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 605–612). Barcelona: ACL. https://doi.org/10.3115/1218955.1219032

Mauranen, A. (2004). Corpora, universals and interference. In A. Mauranen, P. Kujamäki. (Eds.), *Benjamins translation library* (pp. 65–82). John Benjamins Publishing Company. https://doi.org/10.1075/btl.48.07mau

McEnergy, A. (2003). Corpus linguistics. In R. Mitkov (Ed.), *Oxford handbook of computational lingustics*. Oxford University Press.

McLaren, K. (2014). Bilinguisme législatif : regard sur l'interprétation et la rédaction des lois bilingues au Canada. *Ottawa Law Review, 45*(1), 21–37.

Neubig, G. (2017). *Neural machine translation and sequence-to-sequence models: A tutorial*. https://arxiv.org/abs/1703.01619 (15th December, 2019).

Olohan, M. (2003). How frequent are the contractions?: A study of contracted forms in the Translational English Corpus. *Target*, *15*, 59–89. https://doi.org/10.1075/target.15.1.04olo

Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311-318). Philadelphia, ACL. https://doi.org/10.3115/1073083.1073135

Policy on legislative bijuralism. (1995). https://www.justice.gc.ca/eng/csj-sjc/harmonization/bijurilex/policy-politique.html (Accessed April 7th, 2021).

Popescu-Bels, A. (2019). *Context in neural machine translation: A review of models and evaluations*. https://arxiv.org/abs/1901.09115 (Access: 15th December, 2019).

Puurtinen, T. (2003). Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and Linguistic Computing*, *18*, 389-406. https://doi.org/10.1093/llc/18.4.389

Rabinovich, E., Ordan, N., & Wintner, S. (2017). *Found in translation: Reconstructing phylogenetic language trees from translations* (pp. 530–540). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1049

Rapp, R., Sharoff, S., & Zweigenbaum, P. (2016). Recent advances in machine translation using comparable corpora. *Natural Language Engineering*, *22*(4), 501–516. https://doi.org/10.1017/S1351324916000115

Regan, V., Lemée, I., & Conrick, M. (2011). *Multiculturalism and integration: Canadian and Irish experiences*. University of Ottawa Press.

Sánchez-Gijon, Pilar, Piqué, Ramon. (2020). *NMT and the indivisibility of culture and language*. CIUTI Conference 2020. Artificial Intelligence & Intercultural Intelligence. Paris, 9–11 December. CIUTI-ISIT.

Schwenk, H., Wenzek, G., Edunov, S., Grave, E., Joulin, A. (2019). *CCMatrix: Mining billions of high-quality parallel sentences on the web*. https://arxiv.org/pdf/1911.04944.pdf (December 15th, 2019).

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation*. ELRA, 2142-2147.

Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, *69*, 343–418. https://doi.org/10.1613/jair.1.12007

Sutskever, I., Vinyals, O., Le Quoc, V. (2014). Sequence to sequence learning with neural networks. *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*. (vol. 2, pp. 3104-3112).

Tirkkonen-Condit, S. (2002). Translationese —A myth or an empirical fact?: A study into the linguistic identifiability of translated language. *Target*, *14*, 207-220. https://doi.org/10.1075/target.14.2.02tir

van Halteren, H. (2008). Source language markers in EUROPARL translations. *Proceedings of the 22nd International Conference on Computational Linguistics —COLING '08* (vol. 1, pp. 937–944). Stroudsburg, Association for Computational Linguistics. https://doi.org/10.3115/1599081.1599199

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. (2005). Parallel corpora for medium density languages. *Recent Advances in Natural Language Processing IV* (pp. 247-258), Selected papers from RANLP. https://doi.org/10.1075/cilt.292.32var

Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, *30*(1), 98-118. https://doi.org/10.1093/llc/fqt031

Way, A. (2018). Quality expectations of machine translation. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment. Machine translation: Technologies and applications* (vol. 1). Springer, Cham.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations parallel corpus, language resources and evaluation. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 3530-3534). Portorož, Slovenia.

**How to cite this article:** Sole-Mauri, F., Sánchez-Gijon, P., & Oliver, A. (2021). *Cadlaws*: An English-French parallel corpus of legally equivalent documents. *Mutatis Mutandis. Revista Latinoamericana de Traducción*, *14*(2), 494-508. https://doi.org/10.17533/udea.mut.v14n2a10