

University of Naples 'L'Orientale'



Doctoral Course in the Department of Literary Linguistic and Comparative Studies

XXXVII° cycle

UNIOR NLP Research Group

Disciplinary Sector: L-LIN/02

Ph.D. Thesis

Addressing Machine Translation of Legal Terminology in Arabic Legislation:  
From Error Evaluation to Systems Optimisation.

**Supervisor:** Dr. Prof. Johanna Monti

**Coordinator:** Dr. Prof. Alberto Manco

**PhD Candidate:**

Khadija Ait ElFqih

Academic Year

2024/2025

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ،

الْحَمْدُ لِلَّهِ الَّذِي أَنْشَأَ وَبَرَى وَخَلَقَ الْمَاءَ وَالْأَرْضَ وَأَبْدَعَ كُلَّ شَيْءٍ وَدَرَى، الرَّحْمَانُ عَلَى الْعَرْشِ اسْتَوَى وَالصَّلَاةُ  
وَالسَّلَامُ عَلَى مَنْ بَكَى عَلَى أُمِّهِ الْمَبْعُوثِ فِي أُمَّ الْقُرَى صَلَّى اللَّهُ وَسَلَّمْ عَلَى الْحَبِيبِ الْمُصْطَفَى أَمَا بَعْدُ، فَإِنَّا  
طَلَبْنَا الْعِلْمَ لَوَجْهِ اللَّهِ تَعَالَى وَلَمْ نُرِدْ بِذَلِكَ إِلَّا وَجْهَ اللَّهِ وَرِضَاهُ وَذَلِكَ حُبُّ اللَّهِ يُؤْتِيهِ مَنْ يَشَاءُ. وَقَدْ قَالَ تَعَالَى  
﴿يَرْفَعُ اللَّهُ الَّذِينَ آمَنُوا مِنْكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ﴾، وَقَدْ رَفَعَنَا اللَّهُ بِذَلِكَ وَنَسَّأَلُهُ أَنْ يَرْفَعَ قِيَمَتَنَا وَمَقَامَنَا فِي  
الْآخِرَةِ مَعَ أَحِبَّتِنَا كَمَا رَفَعَهُمْ فِي الدُّنْيَا. إِنَّمَا طَلَبْنَا الْعِلْمَ لَوَجْهِ اللَّهِ ثُمَّ لِنُفْرِحَ وَالِدَيْنَا وَأَسَاتِدَتِنَا وَجَمِيعِ الْأَطْفَامِ التَّرْبَوِيَّةِ  
وَمَنْ لَهُ الْفَضْلُ بَعْدَ اللَّهِ جَاوَزَهُمُ اللَّهُ خَيْرًا يَا ذَنْيَ اللَّهِ تَعَالَى. سُبْحَانَكَ اللَّهُ بِحَمْدِكَ أَشْهَدُ أَنْ لَا إِلَهَ إِلَّا أَنْتَ  
أَسْتَغْفِرُكَ وَأَتُوبُ إِلَيْكَ، وَاللَّهُمَّ صَلِّ وَسَلِّمْ عَلَي سَيِّدِنَا مُحَمَّدٍ وَعَلَى آلِهِ وَأَصْحَابِهِ أَجْمَعِينَ، وَالسَّلَامُ عَلَيْكُمْ وَرَحْمَةُ  
اللَّهِ تَعَالَى وَبَرَكَاتُهُ.

﴿وَأَخِرُ دَعْوَاهُمْ أَنْ الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ﴾

## **Declaration**

I hereby certify that this material which is subject to submission for assessment leading to the award of PhD in Literary, Linguistics and Comparative Studies, where I am affiliated with the UNIOR NLP Research Group, from the University of Naples 'L'Orientale' is entirely my own work. I was careful to ensure that this work is original, and does not, to the best of my knowledge, breach any law of copyright. I also declare that I have not used any type of generative Artificial Intelligence tools for the writing of this manuscript nor for the creation of images, graphics, tables, or their corresponding captions.

PhD Candidate: Khadija Ait ElFqih

Matricola №: DLLC/00130

*Khadija Aitelfqih*

Date: 30 October 2024

## Dedication

To my Mother Lalla Karima, and my father Sidi Abdelmajid,

To my late big brother Si Mohamed, may you rest in peace,

To my little brother Si Yassine,

To Seyed Hirbod,

This work is dedicated to you, for everything you've done to help me reach this point.

إلى والدتي لالة كريمة، ووالدي سيدي عبد المجيد،

إلى أخي وسندي الراحل سي محمد، رحمه الله وأسكنه فسيح جناته،

إلى أخي الصغير سي ياسين،

إلى السيد هيربد،

أهدي هذا العمل لكم، تقديرًا لكل ما فعلتموه لمساعدتي في الوصول إلى هذه المرحلة.

## **Acknowledgment**

This work would not have been possible without the generous support of the University of Naples 'L'Orientale' funds. This support has allowed me to fully dedicate myself to my research, resulting in works that reflect the commitment and resources the university has invested in my growth as a researcher.

I am truly grateful to my supervisor Dr. Prof. Johanna Monti and co-supervisor Dr. Maria Pia Di Buono for their vital guidance through my PhD journey. Maria Pia, it is hard to put into words an acknowledgment for you, but thank you for being there during the countless challenges, and for your patience on days when I doubted myself and felt lost. You always knew the right thing to say, whether I needed motivation or perspective. In many ways, your encouragement gave me the strength I needed to keep pushing forward, and I truly couldn't have done this without God, you, and Prof. Monti. As supervisors, you have not only helped me grow as a researcher but also reminded me of the importance of hard work, resilience, and self-belief, I will carry every lesson into whatever lies ahead. I sincerely appreciate your contributions and wish you good health and further success.

I would like to extend my sincere gratitude to the legal experts Dr. Prof. Moulay Rachid Taqi and Dr. Reda ElMoumni involved in this project. Their expertise has been invaluable to this work, and I am sincerely grateful for their unconditional dedication and support.

Finally, special thanks go to my UNIOR NLP Research Group amazing and brilliant colleagues and to all my PhD colleagues of the XXXVII cycle for their constant support, camaraderie, and shared moments that made this journey unforgettable.

Above all, God, thank you for keeping me in good health because of you I am able to fulfil this.

## Publications

- **ElFqih, K. A** (2025). **Prompt Engineering Large Language Models for the Translation of Arabic Legislation.** In Proceedings of the 26th Annual Conference of the European Association for Machine Translation (**In Progress**). <https://eamt.org/eamt-annual-conference/> .
- **ElFqih, K. A.** & Monti, J. (2024, August). **Large Language Models as Legal Translators of Arabic Legislation: Do ChatGPT and Gemini Care for Context and Terminology?.** In Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024), Bangkok, Thailand. Association for Computational Linguistics. <https://aclanthology.org/2024.arabicnlp-1.10/> .
- **ElFqih, K. A.**, Di Buono, M.P., Monti, J. (2024, March). **Towards a Linguistic Annotation of Arabic Legal Texts: A Multilingual Electronic Dictionary for Arabic.** In: Bartulović, A., Mijić, L., Silberztein, M. (eds) Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. NooJ 2023. Communications in Computer and Information Science, vol 1816. Springer, Cham. [https://doi.org/10.1007/978-3-031-56646-2\\_5](https://doi.org/10.1007/978-3-031-56646-2_5) .
- **ElFqih, K. A.**, & Monti, J. (2023, September). **On the Evaluation of Terminology Translation Errors in NMT and PB-SMT In the Legal Domain: A Study on the Translation of Arabic Legal Documents into English and French.** In Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC) (pp. 26-35). <https://aclanthology.org/2023.contents-1.4> .
- **ElFqih, K. A.**, di Buono, M. P., & Monti, J. (2022, June). **Automatic Translation of Arabic Legal Terminology Using NooJ.** In 16th INTERNATIONAL NOOJ 2022 CONFERENCE (p. 26). <https://rephip.unr.edu.ar/server/api/core/bitstreams/2b763b3e-8238-4ce8-9990-83c088f5d116/content> .

# Table Of Content

Declaration.....	
Dedication.....	
Acknowledgment.....	
Publications.....	
Abstract.....	3
Chapter One: Introduction .....	4
1. Background of the Study .....	4
2. Statement of the Research Problem .....	5
3. Research Questions and Hypotheses .....	7
4. Research Objectives.....	7
5. Contributions.....	7
Chapter Two: Review of Relevant Literature .....	10
1. Arabic Terminology Translation in the Legal Domain .....	10
1.1. Nature of Arabic Legal Discourse.....	11
1.2. Translation of Arabic Legal Terminology into English and French.....	15
1.3. Resources for Translators in the Legal domain for Arabic: Terminological, Linguistic, and others.....	22
2. Arabic Machine Translation in the Legal Domain.....	26
2.1. Machine Translation of Arabic Legal Terminology .....	28
2.2. Quality Evaluation of Machine-Translated Terminology in the Legal Domain for Arabic.....	31
3. Linguistic and Legal Terminology Resources for Arabic Machine Translation Applications .....	34
Chapter Three: Research Design, Methodology, and Experimental Set-up .....	40
2. Methodology .....	40
3. Data Collection and Resources creation .....	41
4. Transformer Models as Legal Translators of Arabic Legislation.....	48
4.1. Introduction.....	48
4.2. Experimental Set-up and Methodology .....	50
4.3. Results and Evaluation.....	58
4.4. Conclusion .....	69

5. Towards a Linguistic Annotation of Arabic Legal Texts: a Multilingual Electronic Dictionary for Arabic ‘ULTRA’.....	70
5.1. Introduction.....	70
5.2. Experimental Set-up and Methodology .....	71
5.3. Evaluation and Results.....	80
5.4. Conclusion .....	82
6. Prompt Engineering Large Language Models for the Translation of Arabic Legislation	82
6.1. Introduction.....	82
6.2. Experimental set-up and Methodology .....	84
6.3. Evaluation and Results.....	91
6.4. Conclusion .....	101
Chapter Four: Discussion and Research Limitations.....	101
Chapter Five: Conclusion and Future Work .....	105
Bibliography .....	107
Annexes A.....	127
<b>A1. Properties definitions dictionary « Properties.def » .....</b>	<b>127</b>
<b>A2. Extract from the inflectional and derivational dictionary «ARLD2.nof» .....</b>	<b>129</b>
<b>A3. An Extract from the UNIOR Legal TeRmbase for Arabic (ULTRA).....</b>	<b>137</b>

## List of Tables

Table 1: The linguistic Data Consortium Resources.....	36
Table 2: Arabic Legislation Documents.....	43
Table 3: Extract from Out-of-Context Reference Dataset. ....	45
Table 4: Extract from In-Context Reference Dataset.....	47
Table 5: Comparison of ChatGPT-4, Gemini 1.5 Pro, and GT MT of Legal Terms from AR→EN / AR→FR in Out-of-Context and In-Context Scenarios, Based on Accuracy Criterion. .....	52
Table 6: Errors Count and Classification of Machine-translated Legal Terms by Google Translate for AR→EN / AR→FR in Out-of-Context and In-Context Scenarios.....	53
Table 7: Errors Count and Classification of Machine-translated Legal Terms by ChatGPT-4 for AR→EN / AR→FR in Out-of-Context and In-Context Scenarios.....	54

Table 8: Errors Count and Classification of Machine-translated Legal Terms by Gemini 1.5 Pro for AR→EN / AR→FR in Out-of-Context and In-Context Scenarios. ....	55
Table 9: MQM Evaluation Scorecard of Google Translate Performance in Out-of-Context AR→EN.....	59
Table 10: MQM Evaluation Scorecard of Google Translate Performance in In-Context AR→EN.....	59
Table 11: MQM Evaluation Scorecard of Google Translate Performance in In-Context AR→FR. ....	60
Table 12: MQM Evaluation Scorecard of ChatGPT-4 Performance in Out-of-Context AR→EN. ....	61
Table 13:MQM Evaluation Scorecard of ChatGPT-4 Performance in In -Context AR→EN. ....	62
Table 14: MQM Evaluation Scorecard of ChatGPT-4 Performance in Out-of-Context AR→FR. ....	62
Table 15: MQM Evaluation Scorecard of Gemini Performance in Out-of-Context AR→FR. ....	63
Table 16: MQM Evaluation Scorecard of Gemini Performance in Out-of-Context AR→EN. ....	66
Table 17:MQM Evaluation Scorecard of Gemini performance in In-Context AR→EN.....	67
Table 18:MQM Evaluation Scorecard of Gemini Performance in Out-of-Context AR→FR..	67
Table 19: MQM Evaluation Scorecard of Gemini Performance in Out-of-Context AR→FR. ....	68
Table 20: Terminology Annotation Table.....	73
Table 21: Annotation of the Entry 'خَبِيرٌ' {Expert}.....	74
Table 22: Annotation of the Entry 'مَسْئُولٌ' {Responsible}. ....	75
Table 23: Annotation of the Terms 'عَقْدٌ' {Contract} and 'وَيْثِقَةٌ' {Document}.....	77
Table 24: Annotation of the Term 'مِسْطَرَّةٌ جِنَائِيَّةٌ' {Criminal Procedure Law}.....	78
Table 25: Annotation of Some fixed Terms. ....	80
Table 26: Zero-Shot Prompts for Out-of-Context MT of Arabic Legal Terminology. ....	85
Table 27: Zero-Shot Prompts for In-Context MT of Arabic Legal Terminology.....	87
Table 28. Prompting Results in Out-of-Context scenario.....	91
Table 29. Prompting Results in In-Context Scenario.....	93

## List of Figures

Figure 1. General Language and Special Language Constant Exchange (Cabré, 1999: p. 66). ....	11
--	----

Figure 2. Evolution and Classification of Machine Translation Approaches. ....	27
Figure 3: The General Methodology Pipeline.....	40
Figure 4. Dataset. ....	42
Figure 5: Syntactic Grammar for Legal Terminology Extraction.....	43
Figure 6: NooJ Syntactic Grammar for Concordances Extraction. ....	44
Figure 7: Concordance Sequences for Capturing the Context.....	45
Figure 8: Our Gold Standard Dataset in the Legal Domain for Arabic. ....	50
Figure 9: Resource Creation Workflow .....	71
Figure 10: Grammar of Recognition for MWEs and Single Legal Entries. ....	81
Figure 11: An Extract of the Parsing Result from NooJ Grammar of Recognition. ....	82

## List of Abbreviations

AI: Artificial Intelligence.

NLP: Natural Language Processing.

MT: Machine Translation.

NMT: Neural Machine Translation.

PB-SMT: Phrase-based Machine Translation.

RBMT: Rule-based Machine Translation.

SMT: Statistical Machine Translation.

TQE: Translation Quality Evaluation.

SL: Source Language.

TL: Target Language.

TA-NMT: Triangular Training Architecture for Neural Machine Translation.

GT: Google Translate.

HT: Human Translation.

ULMFiT: Universal Language Model Fine-tuning.

LLMs: Large Language Models.

BERT: Bidirectional Encoder Representations from Transformers.

BLEU: BiLingual Evaluation Understudy.

GPT: Generative Pre-trained Transformer.

RNNs: Recurrent Neural Networks.

AL-BLEU: Arabic Language-BiLingual Evaluation Understudy metric.

BAMA: Buckwalter Arabic Morphological Analyzer.

SAMA: Standard Arabic Morphological Analyzer.

RAG: Retrieval-Augmented Generation.

MQM: Multidimensional Quality Metrics.

EC: Error Count.

ESM: Error Severity Multiplier.

SPM: Severity Penalty Multiplier.

ETD: Error Types Dimension.

ETW: Error Type Weight.

ETPT: Error Type Penalty Totals.

APT: Absolute Penalty Total.

WER: Word Error Rate.

LDC: Linguistic Data Consortium.

LSP: Language for Specific Purposes.

MAC: Most Arab Countries.

MWEs: Multiword Expressions.

POS: Part of Speech.

JRC-Acquis: Joint Research Centre Acquis Communautaire corpus.

CCRE: Conseil des communes et régions d'Europe (Council of European Municipalities and Regions).

NIST: National Institute of Standards and Technology.

NooJ: A linguistic development software for processing natural language.

IATE: Interactive Terminology for Europe.

ILOTERM - International Labour Organisation Terminology Database.

IMF TermBase: International Monetary Funds Multilingual Terminology Database.

ISO: International Organization for Standardization..

UNTERM: United Nations Terminology Database.



## Abstract

Legal translatability is a space of possibilities, an autonomous realm of ‘cross-cultural events’, within which the ‘system-bound’ of legal concepts and notions deeply rooted in language, history, culture, and societal evolution of one country are transformed and integrated into the language of another. Arabic-English / Arabic-French Legal translation is considered an arduous task because of the wide gap between the source and target languages linguistic and legal systems. Thus, translators from and into Arabic face difficulties on different linguistic levels, be they syntactic, semantic, or terminological (i.e., Sharī‘ah Law vs Common and Civil Law terms). The forementioned legal systems are strikingly different and each of them is embedded in the cultural background of each system. Accordingly, the main complexity in legal translatability from Arabic will be to gather terminology of multiple origins and to transfer it into another linguistic and cultural framework. These frameworks are originally a binary code and comprises two semantic spaces: the source space Arabic and the target space English and French.

To translate Arabic legislation that holds technical yet complex legal terminology, translators need access to extensive and reliable resources that serve as an ashgate guide to avoid literal translation, which may lead to serious legal consequences due to the misinterpretation of a legal term in a given legal context. Unfortunately, unlike for the Indo-European languages, there is a significant lack in the literature of Arabic legal terminology in terms of efforts to develop legal terminology resources and other resources that can be adopted to the needs of the legal translators. This absence has a direct effect on many Natural Language Processing applications namely the optimisation of Machine Translation systems to meet the accuracy requirement in the translation of Arabic legislation. Therefore, this thesis work comes as a reaction to fill this research gap by fostering collaborative initiatives between legal experts to develop a robust infrastructure of multilingual terminology for Arabic as a reference for translators, and as a foundational knowledge-base necessary for enhancing the accuracy criteria of Machine Translation systems required in the legal domain for Arabic. To do this, the study attempts to answer three main relevant questions: i) why is the translation of Arabic terminology in the legal domain challenging? ii) how accurate is the automatic translation of Arabic legal terms? iii) how can we reinforce the Machine Translation systems performance in the legal domain for Arabic?

## **Chapter One: Introduction**

### **1. Background of the Study**

Terminology translation is the process of translating specialised terms from one language to another while maintaining accuracy, consistency, and domain-specific connotations (Munday, 2010; Gambier & Doorslaer, 2010; Cabré, 2023). It involves identifying and translating terms with specific meanings within a particular field, industry, or profession (Cabré, 1999). It plays a significant role in domain-specific Machine Translation (MT) (Štajner et al., 2016). Nevertheless, some languages and knowledge domains, such as legal, medical, military, and aerospace, still suffer from the lack of high-quality MT results due to the mistranslation of terminology (Mediouni, 2016). This is the case, especially in the legal domain and the translation of Arabic legal terms. Most MT systems namely Neural Machine Translation (NMT) systems and Large Language Models (LLMs) are considered state-of-the-art for language pairs for which large amounts of parallel data for training are available (Cho et al. 2014; Štajner et al. 2016), However, the available corpora are usually limited to one or two domains. These MT systems present numerous issues in their results, as proven by several scholars evaluating MT outputs and errors (Koehn & Knowles, 2017; Müller et al., 2019; Zakraoui et al., 2020; Moslem et al., 2023). In other words, current MT outputs do not adhere to the terminological constraints posed by translating legal terms in context, especially when addressing legal translation from Arabic into English and French. The issue of the complexities of translating legal terms, particularly between originally unrelated languages, like legal Arabic and legal English, and legal French, has long been a real challenge in legal translation (El-Farahaty, 2015; Moslem et al., 2023). It stems from the conceptual incongruence between legal terms of different legal languages derived from different legal cultures, legal systems, and religions (Wagner, 2003; Wagner et al. 2016; Matulewska, 2016). Such abstract asymmetry is because law has no universal reference, and that legal language is what determines the degree of difference in conceptual correspondence (Catford, 1965).

Arabic legislation, like those in any other language, possesses unique linguistic, discursive, and terminological features. The language used in these documents is formal, precise, and technical, drawing heavily from classical Arabic, Islamic legal traditions, and jurisprudence. As a result, translating such documents requires a deep understanding of both the Arabic language and the relevant legal systems. Moreover, in legal translation, achieving functional equivalence is crucial. This involves ensuring that legal terms in different languages, which may refer to distinct legal concepts or institutions, fulfil the same or similar functions within

their respective legal systems (e.g., Islamic jurisprudence vs. common or civil law) (Šarčević, 2016). Law is the enforceable body of rules that governs any society (Rei et al., 2022), these systemic differences between different legal systems make rendering a legal instrument from one language into another a difficult endeavour (Šarčević, 1997).

## 2. Statement of the Research Problem

With regards to the translation of legal terms from Arabic into English and French, Cao (2007) in her book ‘Translating Law’ states:

*“Finding the meaning of these texts is often quite difficult, even when one is working entirely within a familiar legal paradigm, with a language learned at one’s mother’s knee and with concepts that are known and accepted” (2007: 33).*

In other words, even when someone is deeply entrenched in a familiar legal system, using a language they have been exposed to since childhood, and dealing with concepts and terminology they are accustomed to, it can still be challenging to interpret the meaning of certain texts. This suggests that the complexity of understanding textual content transcends mere familiarity with the language or legal framework, requiring careful analysis and interpretation. Based on an evaluation of MT results of legal terms in and out of context from Arabic into English and French using a corpus built out of legal documents, namely contracts, provisions, and codes, we stated that the automatic translation of Arabic legal terminology predominantly encounters significant obstacles leading to its failure. These challenges can be attributed to various factors (ElFqih & Monti., 2023):

- The context and culture-dependent nature of this type of terminology,
- The linguistic characteristics of this type of discourse,
- MT systems have lower quality for out-of-domain, to the point that they completely sacrifice adequacy for the sake of fluency.
- MT systems have a steeper learning curve with respect to the amount of training data, resulting in worse quality in low-resource settings but better performance in high-resource settings.

Many authors including El-Farahaty (2010); Alwazna (2013); El-Farahaty (2015); El-Farahaty (2016); Mahraj & Hdouch (2020); analyse and address the untranslatability issues of Arabic terminology in the legal domain from a theoretical point of view, and they present an extensive analysis of the lexical features of legal discourse in Arabic. They demonstrate

solutions for coping with legal terminology translation in Arabic legislation by adjusting translation techniques, namely borrowing, transliteration, adaptation, expansion, etc. In fact, it is true that theories are highly relevant in solving translatability issues of legal terminology, because they provide valuable frameworks, insights, and guidelines for translators, helping them to manually navigate linguistic, cultural, and contextual challenges. Nevertheless, overcoming the challenges of legal terminology translation from Arabic can be achieved by leveraging quantitative methods, such as subjective evaluation frameworks, which will not only define the related issues but also propose optimal solutions towards the optimisation of the MT systems performance in the legal domain for Arabic.

Translating legal documents is not only rendering a source text into a target text, but it primarily concerns the conceptual and cultural transfer that require human interpretation, judgment, and a certain degree of acquaintance with the domain in use in addition to a comprehensive and valid legal term base in the case of Arabic as the source language and English or French as the target language (Asensio, 2007; 2014). Despite the great progress that computational linguistic approaches have achieved, MT is not perfect in low-resource languages and domain-specific texts, and it is still far from being perfect. However, the advent of LLMs presents new possibilities for enhancing translation accuracy. They have been developed using deep learning techniques, particularly variants of Recurrent Neural Networks (RNNs) or transformer architectures, examples include OpenAI's Generative Pre-trained Transformer series (Brown et al., 2020) or Google's Bidirectional Encoder Representations (BERT) (Devlin et al., 2018) and Gemini (Rohan et al., 2023) series. They are trained on vast amounts of data, learning the statistical patterns and structures of language to generate coherent and contextually relevant responses to prompts. This grants them the capacity to excel in various Natural Language Processing (NLP) tasks. Current Arabic MT research has primarily focused on the translation of few domains such as news and official texts, whilst few attempts focus on domain-specific translation such as medical or legal domains. Specifically, most of the parallel data available to the researcher are limited to general-purpose texts or produced by international organizations such United Nations Parallel Corpus v1.0 or used in parliamentary debates. As a result, when translating domain-specific texts, these models may struggle to accurately interpret and translate specialised terminology, and contextually rich content. This leads to mistranslations, inaccuracies, and loss of important nuances in the translated output.

Therefore, the aim of this thesis is to experiment if developing and incorporating training data specific to Arabic legal terminology may improve the quality of MT outputs with reference to specialised legal documents.

### **3. Research Questions and Hypotheses**

In this thesis, it is hypothesised that the MT of Arabic legislation is challenging and has a long way to go with the development and training of more unified and reliable datasets. The research hypothesis proposes that MT is imperfect in low-resourced languages setting, namely Arabic, and is still far from being accurate in domain-knowledge scenarios, and the legal domain is a distinguishable case phenomenon.

To test and verify our research hypotheses, this study will attempt to answer the following research questions:

1. Why is the translation of Arabic terminology in the legal domain challenging?
2. How accurate is the automatic translation of Arabic legal terms?
3. How can we improve the MT systems performance in the legal domain for Arabic?

### **4. Research Objectives**

The objectives of this study are to explore why translating Arabic legal terminology is challenging by analysing structural differences between Arabic legal discourse, English, and French legal discourses. The research aims at investigating the current state-of-the-art in MT systems, evaluating their effectiveness in translating Arabic legal terminology and identifying their challenges and limitations in capturing complex legal terms in and out of context. Finally, the study aims to identify potential improvements through experimenting MT systems optimisation techniques to enhance their accuracy state in the legal domain for Arabic.

### **5. Contributions**

In Chapter Two: Review of Relevant Literature, we review and analyse related work on terminology translation in the legal domain for Arabic. We highlight several authors who first analyse and understand the language of law in Arabic and its discourse components. We understand that Arabic legal discourse stands as a testament to the rich and complex heritage of the Arab world, encompassing a diverse array of legal traditions and principles deeply rooted in history, religion, and culture. They also find that Arabic legal discourse boasts a rich and intricate linguistic landscape. In other words, its complexity arises from a blend of classical Arabic rooted in Islamic jurisprudence, modern legal terminology, and regional dialectical

variations, each infused with nuanced semantics and rhetorical devices. This complexity extends to syntax, morphology, and lexicon, where precise terminology, formulaic expressions, and intricate grammatical structures play pivotal roles in conveying legal concepts with utmost clarity and authority. Moreover, the influence of religious texts, scholarly tradition, and cross-cultural legal exchanges further enriches Arabic legal discourse, rendering it a dynamic and multifaceted linguistic domain that requires specialised expertise for comprehensive understanding, and accurate translation across languages. It is at this point that we land into reviewing the efforts and methods used to ease the translation of Arabic legal terminology given the previously mentioned challenges. Scholars tend to employ traditional translation techniques while others believe that the key solution to overcome the translation complexity of Arabic legal terminology in legal documents into English and/or French is not only the translation techniques but also the translator's background knowledge of the legal connotation in hand as well as the source and target language. Others tend to use MT systems followed by post-editing, as most MT systems lack accuracy due to the absence of specialised datasets and terminology, as well as the inconsistency and inaccuracy of existing in-domain translations. Additionally, the limited availability of high-quality parallel corpora and the challenges in adapting MT systems to the nuances of specific legal contexts further contribute to the need for post-editing to ensure precise and reliable translations. Consequently, this review of literature also explores the significant advancements made in MT technologies (e.g., NMT, LLMs) for Arabic and particularly in the translation of legal terminology into English and French, the least explored study in the MT field. We review the methodologies and approaches followed by several leading authors in the MT field to solve the untranslatability of legal terminology not only from Arabic into English and French but also from and into other language pairs. In light of the research questions and related work, the contribution of this research can be outlined as follows:

- Develop a monolingual corpus for Arabic in the legal domain, and leverage it to generate a real-world multilingual terminological dataset from AR→EN / AR→FR. (Chapter Three)
- Leverage the application of reliable frameworks for analytic Translation Quality Evaluation (TQE) of MT systems to evaluate their correctness, consistency, and completeness using our gold standard dataset developed specifically for assessing the quality and accuracy of machine-translated legal terms from AR→EN / AR→FR (Chapter Three, Section 4).

- Explore the gains that can be achieved through the development of a multilingual AR→EN / AR→FR legal dictionary to demonstrate how feeding MT system with in-domain resources can address context-dependent issues, automate the process of annotating Arabic legal texts, and facilitate the automatic translation of technical legal terms from AR→EN / AR→FR. (Chapter Three, Section 5)
- Investigate approaches for optimising foundation models' ability in achieving desired MT results in the legal domain from Arabic (Chapter Three, Section 6).

This work architecture is divided into five fundamental chapters. Chapter One introduces the study by first providing the background which contextualises the research within existing knowledge, highlighting challenges and knowledge gaps that underscore the study's relevance. The statement of the research problem defines the specific issue the study addresses, emphasising its importance and the need for investigation. It also outlines the research questions and hypotheses that the study seeks to experiment and validate. The research objectives are then presented, describing the study's primary aims and intended outcomes. Finally, the chapter concludes by detailing the study contributions. Chapter two reviews relevant literature, examining studies and resources that underpin the translation of terminology in Arabic legislation into English and French, discussing the nature of this discourse and its unique linguistic features. It also addresses how MT handles Arabic legal terms and evaluation frameworks applied to assess accuracy and reliability for legal use, in addition to covering the linguistic and terminology resources that supports the optimisation of Arabic MT. This literature review provides a comprehensive foundation for understanding the challenges, solutions, and resources associated with machine-translated legal terms for Arabic.

Chapter Three outlines the research design, methodology and experimental set-up for the study, detailing the approach and tools used to address the research questions. It begins with an overview of the general methodology about the research framework and key methods employed to ensure robust and valid results. It also discusses the data collection process relevant to the study. This chapter highlights actionable feedback suggested by the evaluation conducted and the solutions it proposes towards the optimisation of MT systems in responding to the accuracy criteria needed in Arabic legalese. Chapter Four presents a discussion of the study findings analysing how the results contribute to answering the research question and objectives and how they compare to prior research. It also mentions the research limitations encountered and their impact on the findings. Chapter Five offers a conclusion by summarising

the study's contributions and their significance to the field. It further outlines avenues for future work, suggesting enhancements and potential applications for optimising the MT performance in the legal domain for Arabic. Finally, the bibliography provides a comprehensive list of references, documenting all sources consulted throughout the research.

## **Chapter Two: Review of Relevant Literature**

### **1. Arabic Terminology Translation in the Legal Domain**

Terminology translation plays a crucial role within the framework of Language for Specific Purposes (LSP) by facilitating accurate and contextually relevant communication within specialised fields (Balboni, 1986). Sager et al. (1980) state that general languages are semi-autonomous, complex semiotic systems based on and derived from general language; their use presupposes special education and is restricted to communication among specialists in the same or closely related fields. Pitch and Draskau (1985) also support this view stating that LSP is a formalised and codified variety of language, used for special purposes and in a legitimate context, that is to say, with the function of communicating information of a specialised nature at any level, at the highest level of complexity, between initiated experts, and at lower levels of complexity, with the aim of informing or initiating other interested parties in the most economic, precise, and unambiguous terms possible. In other words, according to Sager et al. (1980), special languages are limited to communication between specialists, they add:

*“Special languages, or more precisely special subject languages are usually thought of as the means of expression of highly qualified subject specialists like engineers, physicians, lawyers, etc. and are often derogatively referred to as ‘jargon’. (1980: 25)*

Cabré (1999) assumes that special languages are a subset of the language as a whole. They intersect with the general-purpose language, with which it not only shares features but also maintains constant exchange of units and conventions. She visualises her assumption in Figure 1.

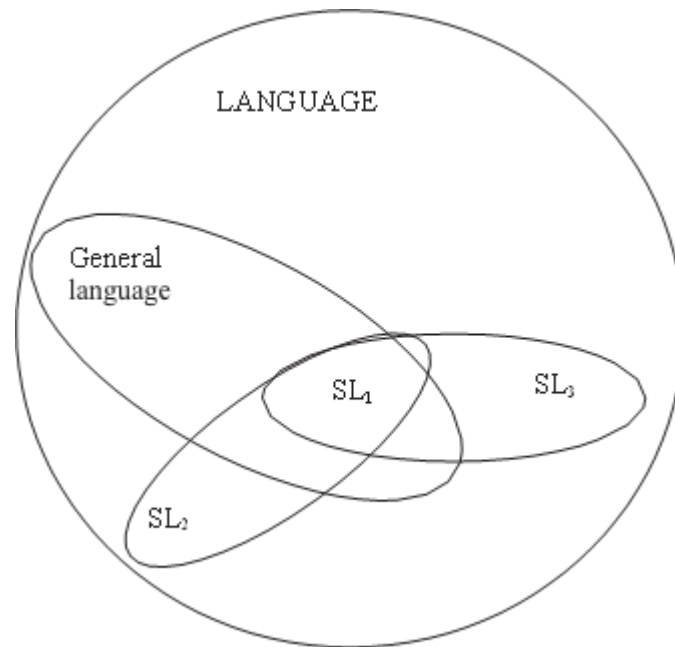


Figure 1. General Language and Special Language Constant Exchange (Cabr , 1999: p. 66).

The instructions in LSPs are situational and based on topics related to some specific fields like science, computer, architecture, business, law, engineering, and finance, etc (Hayati, 2008; Hanford, 2011), which is what Anderman & Rogers (2005) assert when saying that special language is coupled with or tied to knowledge of specific subject fields. In this regard, Halliday (1994) sets out three basic features that he gives credit to in the development of the domain of LSP, stating that "field" (subject matter), "tenor" (related to the level of formality), and "mode" (involving various aspects of textuality such as cohesion) contribute all together to make up the register membership of a text. In the realm of LSP, terminology holds a central position in facilitating accurate and contextually relevant communication, ensuring that professionals can convey precise and domain-specific concepts and information within their respective fields. In the following we will scrutinise how the Arabic language adapts and become specialised to meet the unique needs to serve the legal jargon and/or discourse.

### 1.1. Nature of Arabic Legal Discourse

Arabic legal discourse stands as a testament to the rich and complex heritage of the Arab world, encompassing a diverse array of legal traditions and principles deeply rooted in history, religion, and culture. The nature of this discourse is characterised by a synthesis of Islamic jurisprudence, customary law, and modern legal systems, reflecting a dynamic interplay between tradition and modernity (El-Farahaty, 2015).

At its core, Arabic legal discourse is deeply influenced by the Islamic legal tradition, which has played a pivotal role in shaping the legal landscape of the Arab world. Islamic law, or Sharī'ah, serves as a foundational source of guidance for matters ranging from family law to commercial transactions. The Quran (the central holy book of Islam) and the Hadith (the teachings and sayings of the Prophet *Muhammad Sallallahu Alayhi Wasalaam* (SAW)), serve as primary sources, guiding legal scholars in the interpretation and application of Islamic law. This emphasis on divine revelation and jurisprudential interpretation, infuses the Arabic legal discourse with a strong moral and ethical compass, emphasising justice, equity, and the well-being of the community.

Arabic is said not to have its legal register as believed by Hatim (2009). El-Farahaty (2015) took Hatim's hypothesis as a point of departure, despite the scarcity of references on the issue and analysed Arabic legal discourse in terms of lexical, syntactic, and textual features. She states that lexical features of Arabic legal discourse are affected by the religious, culture-specific and system-based terms and expressions used therein. A marriage contract for example is full of Sharī'ah Law and religious expressions. For instance, {الأهلية} can be translated into 'legal capacity' in English. {على كتاب الله و سنة رسوله} can be rendered into English as 'according to the Holy Book of Allah and the Sunnah of His Messenger'. {على مذهب أبو حنيفة} is rendered into English as (according to the doctrine of Abu-Hanifa). {لا قدر الله} is literally translated as 'Allah forbids'. {نفقة} is translated into English as 'alimony'. Religious greetings are frequent formulaic expressions in legal documents such as {باسم الله الرحمن الرحيم و الصلاة و السلام على سيدنا محمد} 'after praising God and prayer and peace be upon His Messenger'.

In addition to religious terms, Arabic legal discourse is characterised by formality and expressions of courtesy. For instance, certain treaties address the King as his Excellency or his Majesty. Besides, gender-biased terms are the ones that favour the male or the female in their description. By default, words like {المالك}, {المستأجر}, {الشاهد}, {المتعاقد} translated in English as 'the landlord, the tenant, the contracted party and the witness' are masculine in Arabic and gender-neutral in English. The masculine is favoured over the feminine in most Arabic legal discourse cases, which violates the very basic principles of justice and quality on which the legal discourse is based (El-Farahaty, 2015).

In addition to the lexical features, the Arabic legal discourse is characterised by several syntactic features, examples include nominalization, passivation, and modals along with complex sentence structures, participle, as well as doublets and triplets. The first linguistic

aspect of the Arabic legal discourse is nominalization. It changes a verb into a noun (a.k.a a verbal noun) which serves to hide the agent and function more or less as the passive voice as in ‘respecting the law’ and ‘slow driving’. The second feature of Arabic legal discourse is the use of active and passive voice. The question remains as to which one is more frequent and in what legal text types. Rosenhouse (1988), as cited in (El-Farahaty 2015) argues that “Arabic is known to avoid passive verb forms and not to favour much use of them in its sentences”. However, El-Farahaty (2015) believes that there is a tendency to use the passive voice in Modern legal documents. Besides, Arabic legal discourse contains numerous cases of modal expressions, such as:

- Those initiated by the preposition ‘min’ {من الازم} ‘it is necessary’, ‘may’, {يجب} ‘must’;
- Prepositions and particles are also frequent such as {أن . . .} ‘for . . . to’ which is a short version of: {أن يجوز} ‘it is allowed for somebody to’ {أن . . . على} ‘on . . . to’ which is an abbreviated version of {... يجب على} ‘it is incumbent on somebody to...’.

Complex sentence structures are observed in Arabic legal discourse given that Arabic contains an inconsistent punctuation system and a lot of embedded coordinating and subordinating clauses in long sentences (Alcaraz Varo & Hughes, 2014).

Doublets and triplets, which are also known as binomials and polynominals, are two or more synonymous or near synonymous words conjoined together with ‘and’ or ‘or’. The participle is one of the grammatical features of Arabic legal discourse. In Arabic, there is ‘إسم المفعول’: the passive participle, ‘المفعول المطلق’: the absolute object or cognate accusative. El-Farahaty (2015) cited an example from the 2011 Moroccan Constitution:

- **Arabic:** الملك، أمير المؤمنين وحامي حمى الملة والدين، والضامن لحرية ممارسة الشؤون الدينية.
- **El-Farahaty English translation (2015):** The king, Amir al-Mu'minin (commander of the faithful), **protector** of the faith and religion and the **guarantor** of the free exercise of religious affairs.

Legal texts in Arabic are characterised by the use of appositives in their structures. The following extract illustrates this:

- **Arabic:** لا يسوغ مواخذة أحد على فعل لم يعد يعتبر جريمة بمقتضى قانون صدر بعد ارتكابه. فإن كان قد صدر حكم بالإدانة، فإن العقوبات المحكوم بها، أصلية كانت أو إضافية، يجعل حد لتنفيذها.

- **English translation (Benhima, 2023):** No one shall be convicted of an act that is no longer considered a crime in pursuance of a law issued after its commission. If a conviction were issued, the implementation of penalties, **principal** or **supplementary** ones, should be put to an end.

Appositives in Arabic are similar to English and French in that they are enclosed between commas in the mentioned languages. They serve to give extra information on a noun. The words written in bold are examples of appositives that describe the types of penalties or sanctions.

Arabic legal discourse has common textual features summarised in lexical repetition, reference, and conjunctions as well as punctuation. Lexical repetition is also known as “semantic redundancy, verbosity or wordiness” (Shunnaq, 1994) as cited in El- Farahaty, 2015, and it is used for the sake of accuracy in legal documents. It takes different forms, examples include root or word repetition, expression repetition, lexical density and parallel structures. The first subtype is manifested in the recurrence of a root or a word more than once in or out of the sentence boundary. Examples include the absolute object { **طَلَبًا** – **طَلَبًا** } which means ‘to request - a request’ or word derivation { **التَّعَهُدُ** – **يَتَعَهُدُ** } which means ‘to undertake -undertaking’, respectively. The second subtype of lexical repetition is expression repetition as in the repetition of ‘the second party’ { **الطَّرْفُ الثَّانِي** } in a contract. Lexical density is classified as syndetic with connectors (or: **أَوْ** / and: **وَ**) and asyndetic with no connector. Arabic has many syndetic structures as in the following: { **ضِدَّ الْمَجْنِيِّ عَلَيْهِ أَوْ زَوْجِهِ أَوْ أَحَدِ أَصُولِهِ أَوْ فُرُوعِهِ** } translated into English as ‘against the victim or against his/her spouse or one of his/her ascendants or descendants’. Parallel structures can be exemplified by { **بَيْعًا صَاحِبًا شَرْعِيًّا قَاطِعًا مَاضِيًّا مُمَضِيًّا** } translated into English as ‘a valid, lawful, definitive, effectual, ratified sale’. Parallelism in Arabic can be phonological, morphological, lexical, and syntactic depending on the repeated units.

One of the features which should be mentioned with regard to Arabic discourse is the special use of punctuation and conjunction. In contrast to English which has a developed punctuation system, Arabic still relies only on the conjunctions ‘and: wa / **وَ**’, and ‘so: fa / **فَ**’ along with the frequent use of the comma and the full stop. Capital letters do not exist in Arabic because all the Arabic letters are written in the same form. They differ only in the script, especially in religious texts, such as the Quran. A simple script, absence of capital letters, and many commas and dots characterise Arabic legal discourse.

Many are features of English, French and Arabic legal discourse (Douglas, 2001). Concerning the English and French legal discourse features, they are summarised in lexical, syntactic, and textual features. The lexical features are archaic words along with Latin and French terms. Syntactic features are composed of passivisation, wh-deletion, conditionals, complex sentences and negation and Textual features are grouped into conjunction, reference and lexical repetition (Engberg, 2013). Similarly, the features of the Arabic legal discourse are classified into lexical, syntactic and textual features. Lexical features are highly affected by the culture, religious terms, formality, gender-biassed terms, and template terms. Syntactic features of Arabic legal discourse encompass nominalization, passivation, modals, complex sentence structures, and doublets and triplets (Abulhassan, 2014). The textual features of Arabic legal discourse are lexical repetition, anaphoric and cataphoric references, and punctuation (Abulhassan, 2014). The main critical feature that highlights the nature of the Arabic legal discourse is the culture and the variations of terms used to express a particular legal practice among the Arab countries (Jopek-Bosiacka, 2018). In other words, Arabic legal terminology is deeply rooted in cultural nuances and historical connotations. It often encompasses terms that have evolved over centuries and have specific connotations that are difficult to capture without a profound understanding of the cultural and historical context. The upsurge of references related to the discussion of the Arabic legal discourse are extremely limited. This significantly hinders the progress and development of the field of legal studies and legal translation studies. This scarcity limits researchers' access to crucial scholarly materials and resources necessary for a comprehensive understanding of the intricacies of the Arabic legal system. Therefore, in this thesis work we will focus on filling the gap that has been neglected theoretically and technically. In what follows we will address the related work of authors dealing with the translation of Arabic legal terminology.

## 1.2. Translation of Arabic Legal Terminology into English and French

The term “terminology” conveys three different concepts: the methods and practices used for presenting terms; the arguments and theories used for clarifying the connection between concepts and terms that are deemed crucial for a well-structured activity (Sager, 1990; Chromá, 2011). Sager (1990) gives three definitions of Terminology:

1. *The set of practices and methods used for the collection, description, and presentation of terms.*

2. *A theory, i.e., the set of premises, arguments and conclusions required for explaining the relationships between concepts and terms which are fundamental for a coherent activity under (1).*
3. *A vocabulary of a special subject field.*

“Presentation” of terms (definition (1)) refers to visually representing terms in structures expressing terminological relations (equivalence, generic hierarchy, part-whole hierarchy and complex or associative relational structures). Definition (2) – i.e., theory – is the basis for definitions (1) and (3), where the third definition – i.e., the vocabulary – is the result of definition (1) – i.e., the set of practices and methods. Clearly, definition (2) is not appropriate for a professional translators: theory and theory-building would simply take up too much time, although they may use its principles as “(heuristic) discovery procedures” (Thelen 2002b, 2012)<sup>1</sup> or as a guidance in the direction of translation solutions. Definition (3), the vocabulary, is the resource that the translators consult when translating, or the one they may supplement or correct, on the basis of issues that they have to resolve while translating. The translators are not concerned with a systematic and complete mapping of a special vocabulary. Definition (1), the set of practices and methods, fits best in the actual practice of a professional translator, but the recourse to terminology in definition (2) (theory) can be of great help for the solution of terminology-related translation issues (Sager, 1990).

Thelen (2015) distinguished between two types of terminology:

1. **Theory-oriented terminology.** The author understands that it is the type of terminology work done by terminologists who are essentially concerned with the relationship between terms and concepts, concept formation, term formation, and standardisation (Thelen 2012; 2008). Thelen (2015) mentions that this pertains to all the three definitions of terminology in (Sager, 1990), but in particular to the 2<sup>nd</sup> definition. For definition 1 the terminologists apply a well-defined set of practices and methods so as to contribute to the special vocabulary of a given domain (definition 3).
2. **Translation-oriented Terminology.** It refers to the type of Terminology carried out by translators for use in translations. Thelen (2015) states that translation-oriented Terminology is:

---

<sup>1</sup> Heuristic discovery procedures (Thelen 2002b), or simply “discovery procedures” (Thelen 2012) are defined as “procedures or structures to help [...] discover the meaning of the source text” (Thelen 2012) but may be generalised as the application of theoretical principles, procedures, or structures to help solve a problem.

*[...] the kind of terminology work done by translators, either monolingually (in order to analyse the meaning of a term in the source language and/or the meaning of an equivalent term in the target language) or bilingually or multilingually (in order to compare the results of the monolingual analyses to see if there is equivalence between them), but always with a view to translation, where effectiveness and efficiency of the translation process and speed are most important. (2015: 2)*

Since legal translation is bound by each language's culture, it is not merely the transcoding between the SL and the TL or as Cao (2007) puts it '*rendering of legal texts from the SL into the TL*'. It is rather '*a translation from one legal system into another – from the source legal system into the target legal system*' (Šarčević, 1997). It involves all the legal sub-text types which are used in various legal settings, whether a court, a national or international organisation, a law book, a legal report, a birth certificate, a contract, among many others.

In this regard, translating the legal terminology used in Arabic legislation into English and French has been a real challenge because of the nuances that are deeply embedded within the legal systems of these languages. El-Farahaty (2016) conducted research where the aim is to discuss ways of translating terms between Arabic and English. To study this issue, she uses data extracted from authentic legal texts from Arabic and English and excerpts from her Arabic legal corpus. She classifies the major problems faced by legal translators of Arabic legal texts into English into archaic and Latin terms, culture-specific and system-based terms. Her methodology involves translating Arabic legal documents into English by translators. She concludes that the translators should understand the different types of vocabulary they are dealing with in the text, whether they are common, specialised, archaic, abstract, or functional. Each type of these lexical items requires caution, investigation, and knowledge on the part of the translator. For example, with common words, translators are entitled to distinguish the exact meaning of these words in the legal context. They should consult specialists and analyse similar texts for reaching the best solution. When translating archaic expressions, translators should find an approximate corresponding expression in the TL or use paraphrasing. Abstract words are very sensitive, and they are subject to many legal interpretations in the legal context. Thus, legal translators should translate them literally and should not try to disambiguate them even if this translation will result in a vague text. In translating religious and culture-based terms in the documents, translators could employ different techniques such as 'expansion, adaptation, transposition, and structure shift' (El-Farahaty, 2016). In other words, El-Farahaty (2016) suggests that the translation techniques and the translators background knowledge about legal

documents or systems in hand are the key solution to overcome the translation complexities of Arabic legal terminology. Similarly, Al Aqad (2014) study investigates the quality of the translated message from Arabic into English and on the change of the message in the translation process that is attributable to functional and verbal equivalence through the translation of terminology in Arabic and English. The study uses five different forms of marriage contracts translated by different native translators in Arabic as a source language and their correspondence into English as a target language. It relies on Baker's theory (1992) to examine whether the semantic changes affect the quality of the translated message in terms of equivalence, along with Newmark (1988) methods in translation. The study reveals that translating legal terms between different legal systems often results in ambiguous and indirect translations. Each language has its own terminology and achieving exact translations that fully capture the intended legal concepts is sometimes impossible. For Al Aqad (2014) a perfectly qualified legal translator may not always be able to provide a correct translation for such terms adopting some translation techniques. For example, translating a term such as 'examining magistrate' into Arabic, which is used in more than one legal system, is complicated. For instance, the word 'magistrate' in all European languages means 'judiciary man' {رجل القضاء}. The word is a relatively common use in the judicial system. But in the English legal system it is a (Judicial officer) who is authorised to investigate research and examine the legal matters. On the other hand, the Arabic legal system such as the Moroccan legal system calls it 'Investigation Judge' {قاضي التحقيق}. In the Tunisian legal system, it is called 'Magistrate of Investigation' {حاكم التحقيق}, and in the Lebanese legal system they call it 'Forensic investigator' {حاكم التحقيق}. However, in the Kingdom of Saudi Arabia the synonym of this term is 'Bureau of Investigation' {هيئة التحقيق} due to the fact that the Saudi Arabia judiciary system is voided from the position of examining magistrate, so the translator is supposed to search tirelessly for the appropriate term of functional equivalent. Al Aqad (2014) concludes that legal translation differs from other types of translation in two basic components: the legal system and the term associated with that system, thus understanding the legal term and its translation into another language depends on the understanding of its locus in the legal system to which it belongs. For instance, the Arabic language has some words which have the same lexical structure, same phoneme and morpheme but different meaning due to the legal system and the terms followed by that system. This means that the translator can only translate the legal term into the TL by understanding its position in the legal system of the SL. Legal translators should add one more skill to their linguistics and translation skills which is the comparative skill between two legal systems: the legal system of SL and the legal system of TL. Id-Youss et al. (2017) also agree

that differences in legal terminology across Arab countries affect legal translation, the choice of legal translators for achieving functionally equivalent terminology. This is owing to the fact that legal concepts emerge and evolve within national borders independently of their counterparts in the rest of the world, and this realism has serious consequences for communication across legal systems. Halimi & Alwazna (2023) in an attempt to also investigate the impact of Arabic legal terminology's lack of uniformity on producing functional equivalent in the legal domain, agrees with Id-Youss et al. (2017). They add that legal translators should distinguish between two types of issues, namely, the problems due to legal considerations per se and those due to the legal translators inadequate terminological choices.

Nevertheless, Shiyab (2019) in his work on the dilemma of legal terminology in the Arab world states that many academics teaching translation at public or private institutions lack sufficient knowledge and pedagogical experience in legal translation to teach and train students to become proficient legal translators. This continues to mislead students and individuals who work in this profession to believe that legal translation is not a highly specialised form of translation that requires specific experience in the legal field, good knowledge of the legal systems concerned, and understanding of the cultures of the two languages. The absence of qualified translation teachers coupled with the lack of political will to organise and regulate the profession has given rise to poor quality terminology and translations in all legal fields.

Alwazna (2022) hypothesises that there has been a lack of translation methodology among practitioners and trainees and a lack of practical operational model that comprises all the primary measures and criteria for analysing legal terminology translation in Arabic legislation. Therefore, he attempts to test the efficacy of the integrative model proposed by Prieto Ramos (2011) in surmounting terminological problems of Arabic-English legal translation through applying the said model to translate three Arabic legal terms 'بيع التولية', 'بيع الوفاء و الأمانة', 'التخليّة' into legal English. The selection of these terms has been made on the basis of the fact they have no one-to-one correspondence in legal English. The following is the methodology steps of the integrative model Alwazna (2022) followed to translate the three terms:

**The term** {بيع التولية}. According to Alwazna (2022), it is crucial to analyse the source text term to render it into legal English following the steps contained in the integrative model. This points to the process of uncovering and identifying the conceptual aspects of the term in question, drawing on the legal sources peculiar to the legal system concerned. {بيع التولية} is derived from the Islamic legal system. It refers to a particular type of sale which is made by a buyer of a

particular commodity at its cost price (Albahūtī, 1636a, 1639b; Alwazna, 2013). Ibn Qudāmā (1997), analysing the same concept, points out that the term {بيع التولية} refers to a sale of an item at its cost price, with neither an increase nor a decrease in price. Based on the foregoing, this type of sale is different from what is known in Islamic Law as {بيع المراجعة}, which is a type of sale made by a buyer of a commodity with some increase in the price (Albahūtī, 1636a, 1639b). Moreover, {بيع التولية} differs from what is termed in Islamic Law as {بيع المواصفة}, which is a sale made by a buyer of an item with some reduction in the price (Albahūtī, 1636a, 1639b).

Having considered that the term: {بيع التولية} has no one-to-one correspondence in legal English, Alwazna (2022) considers that it is crucial to follow the second step of the integrative model, which resides in comparative legal and linguistic analysis because it is primarily grounded in the characteristics or intensions of concepts which are defined through their legal content and which represent the basis on which the comparative legal and linguistic analysis is primarily founded. The connotation of the term {بيع التولية} needs to be borne in mind when providing a descriptive paraphrase for the Arabic Islamic legal term in question.

Alwazna (2022) thinks that an English descriptive paraphrase that can be suggested to stand for the Arabic Islamic legal term: {بيع التولية} may be read as (sale of an item at its cost price') or the like. Carrying out the comparative legal and linguistic analysis between the aforementioned Arabic Islamic legal term and the English descriptive paraphrase given, it is worth asserting that the intentions of the concepts designated by the said term are to some extent conveyed by the English descriptive paraphrase. Such intentions of the concepts may rest upon the fact that the Arabic Islamic legal term: {بيع التولية} points to the sale made by a buyer of a specific commodity at its cost price. Linguistically, although the English descriptive paraphrase has managed to relay the same intended linguistic meaning in addition to the same intended legal meaning of the said Arabic Islamic legal term, the concept of economy is evidently compromised. This refers to the number of words used to make up each construction. For instance, {بيع التولية} is composed of two terms. On the other hand, the English descriptive paraphrase consists of eight words. It is worth noting that in comparative linguistic analysis, only the semantic level has been considered in the current context. Otherwise, there are evident incongruencies between the aforementioned Arabic Islamic legal term and the English descriptive paraphrase. In other words, a specific legal system may allow the owner of a particular commodity to sell it at cost price, without making a profit or applying any reductions. However, it is evident that legal English has not assigned a specific term to relay such legal

meaning. Hence, the English descriptive paraphrase ‘sale of an item at its cost price’ can be used as a functional bridge to acquaint the target reader with the appropriate knowledge concerning the concept used in the source legal system. This runs in line with Sandrini (1996), who declares that functional bridges are not meant to offer exchangeable correspondents, rather they are employed to provide the target audience with the pertinent knowledge peculiar to the legal system in question.

The third step contained in the integrative model lies in the notion of acceptability analysis of reformulation options. This type of analysis examines the pragmatic considerations related to the use of a specialised language in light of the expectations and needs of the target reader. In this regard, Sandrini (1996) also claims that the notion of acceptability can only be assessed through textual parameters and pragmatic criteria. This means that the translator’s decision in choosing specific lexical items that convey the same intended legal meaning as that conveyed by the source legal term is predominantly driven by the micro-textual considerations, bearing in mind the macro-contextual variables. The translator should be able to anticipate such textual considerations on the basis of the situation concerned. To get further with this notions, Alwazna (2022) asked the following questions: what is the linguistic situation in which such micro-textual considerations should be predicted? What type of target readers are addressed? Are they specialist of Islamic Law, non-specialist, familiar to some extent with Islamic legal concepts? He found out that the acceptability analysis prioritises the expectations and needs of the target reader and bases the selection of the English equivalent on the linguistic situation of the target audience. Hence, the English descriptive paraphrase ‘sale of an item at its cost price’ may be valid if the target reader is non-specialist of Islamic Law, aware of certain legal concepts of Islamic Law or has no knowledge thereof. It is also recommended that the Arabic Islamic legal term concerned is transliterated in English before providing the English descriptive paraphrase as this is the way terms are treated. However, if the target reader is a specialist of Islamic Law, the translators do not need to provide such English descriptive paraphrase, particularly within the text, but They may be sufficed with the transliteration of the Arabic Islamic legal term, with a footnote or endnote explaining the intended legal meaning of the term concerned in a neutral language. Alternatively, the translators may tail their translation with a linguistic analysis/commentary, discussing and analysing the type of problems encountered during the translation process and the set of techniques adopted to overcome them.

To recapitulate, Alwazna (2022) adapted the integrative model (Prieto Ramos, 2011) to translate three Arabic legal terms into English ‘بيع التولية’, ‘بيع الوفاء و الأمانة’, ‘التخلية’. The

translation results show that the terms should be transliterated and descriptively paraphrased within the text if the target reader is non-specialist of Islamic Law, however, they may only be transliterated within the text, with glossed description if the target reader is specialist of Islamic Law. On the other hand, others should be transliterated followed by lexical expansion if the target reader is a non-specialist of Islamic Law.

In the Arabic and French legal scenario, Halimi (2017) declares that Legal translation in the context of French-Arabic criminal justice terminology requires careful consideration of contextual variables and the interaction between national laws and international multilingual legal communication. She, hence, conducts a study on the contextualization translation decisions in legal system-bound and international multilingual contexts. Her case of study is on the French-Arabic criminal justice terminology. The paper uses the concept of frames as a model of concept representation to map the investigation process in three criminal justice realities: the Algerian, Egyptian, and French criminal procedure. Halimi (2017) aims to examine conceptual variations and problems of conceptualization in translation when national laws interact with one another. The study shows that when considering legal translation from an international perspective, all instances of reciprocal connection between laws and international multilingual legal communication should be examined in order to present relevant contextual variables in translation. This approach brings to light specificities at the procedural level that the legal translators and legal experts should be aware of in determining the translation, in order to avoid conceptual inaccuracies, terminological discrepancies and awkward calques (Lewis, 1980; Faiq, 2004; Halimi, 2019).

### 1.3. Resources for Translators in the Legal domain for Arabic: Terminological, Linguistic, and others

To translate in this global and fast era, legal translators need access to extensive and reliable terminology and linguistic resources like termbases, glossaries, and electronic dictionaries that serves as an ashgate guide in maintaining accuracy throughout the translation process, preventing the occurrence of ambiguities or misunderstandings that could potentially lead to legal consequences. In this regard, The European Union (EU) for example, stands as a testament to linguistic diversity in 24 official languages. In this multilingual environment, the management, standardisation, and utilisation of terminology resources in different domains become a formidable task. The EU has taken significant strides in addressing this challenge by developing a comprehensive framework for terminology management such as Interactive

Terminology for Europe (IATE)<sup>2</sup>, EU Vocabularies (EuroVoc)<sup>3</sup>, the United Nations Terminology Database (UNTERM)<sup>4</sup>, WordNet<sup>5</sup>, and Jur-WordNet<sup>6</sup> in addition to diverse parallel corpora.

**IATE.** It is a dynamic terminology database designed to support the multilingual drafting of EU texts and legal texts in particular. It aims to provide relevant, reliable, easily accessible data which represent a distinct added value compared with other sources of lexical information (e.g., translation memories, the internet, electronic archives). The project was launched in 1999 with the objective of providing a web-based infrastructure for all EU terminology resources, enhancing the availability and standardisation of the information. IATE incorporated existing terminology databases of the EU's translation services into one interinstitutional database containing approximately 1.4 million multilingual entries.

IATE has a multilingual interface in the 24 official languages of the EU (Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish and Swedish), and one non-European language (Georgian). It is intended to contain a single entry per concept but contains multiple entries for many concepts. It employs a structured method for the description of terms (term ID, domain, cross reference, definition, definition reference, term reference, term type, reliability, term in context, context reference). IATE is available to the public, this privilege gives language professionals, academics, national experts, public administrations, policy advisers, private sector companies public access to validated terminology. This resource has been an opportunity to rethink the technologies, architecture and data structure of the system in order to prepare it for future challenges, including interoperability, modularity, scalability and data exchange, among other things (Zorrilla-Agut et al., 2019).

**EuroVoc.** The EuroVoc provides access to vocabularies managed by the EU institutions and bodies. This includes controlled vocabularies, schemas, ontologies, data models, etc. As part of the EU open data initiative. It offers free public access to all of its content. This thesaurus exists in 24 official languages of the EU.

---

<sup>2</sup> <https://iate.europa.eu/home>

<sup>3</sup> <https://op.europa.eu/en/web/eu-vocabularies>

<sup>4</sup> <https://unterm.un.org/unterm2/>

<sup>5</sup> <https://wordnet.princeton.edu/>

<sup>6</sup> <http://www.ittig.cnr.it/Ricerca/materiali/JurWordNet/JurWordNetEng.htm>

**WordNet** (Miller et al., 1990) is an ontology-based lexical database containing 155,327 words organised into 175,979 synsets, resulting in 207,016 word-sense pairs that capture semantic relations between words. These relations include synonyms, hyponyms, and meronyms. Synonyms are grouped into synsets, each with a brief definition and usage examples.

**Jur-WordNet (Sagri et al., 2004).** It is an extension for the legal domain of the Italian ItalWordNet (IWN)<sup>7</sup> (Roventini et al., 1998) database, aimed at providing a knowledge base for the multilingual access to sources of legal information.

**United Nations Terminology Database (UNTERM).** It is a multilingual database for official terminology relevant to the work of the United Nations. It is a public website used by UN delegates, staff, translators, editors, and interpreters, as well as other users who are interested in the work and activities of the United Nations around the world. Official terminology is provided in the six official languages of the United Nations (Arabic, Chinese, English, French, Russian and Spanish).

Other recent research in multilingual legal terminology management has been oriented towards using institutional big data. The LETRINT project<sup>8</sup> (2015–2022) draws upon three institutional corpora from the UN, EU, and the World Trade Organization (WTO) to develop translation quality markers in international institutional settings, focusing on context-based specificities of terminology (Ramos & Cerutti, 2021). Using document mapping technology, the LYNX project<sup>9</sup> (2017–2021) relies on EU legal big data and AI to provide effective ways of accessing immense quantities of digital regulatory documents (national legislation, case laws, industry norms, etc.) based on a Legal Knowledge Graphs.

The progress made by EU bodies in their mission toward a "multilingual Europe" provides a valuable resource for legal translators, researchers, and experts, serving as a reliable bridge for translating between EU languages, but what is the status of existence and/or development of legal terminology resources for Arabic?

Studies conducted over the past decades paint a crucial picture of the state of lexicographical and terminographical production in Arabic in general and of bilingual multilingual specialised resources in particular (Al Qasimi, 1991; Ben Mourad, 1997; Galfan, 1998). According to the literature, specialised dictionaries lacks theoretical foundations and

---

<sup>7</sup> <https://www.ilc.cnr.it/progetti/italwordnet-2/>

<sup>8</sup> <https://transius.unige.ch/letrint>

<sup>9</sup> <https://www.lynxproject.org/>

guidance on their use. In other words, they have been produced without asking the fundamental questions about the process: for whom is the dictionary intended; what is the basis for the choice of data; what are the relations between the concepts to build the entries? Halimi (2023) conducts a review of the tools that are available for a bilingual search in both the French-Arabic and English-Arabic language pairs in the legal field. She starts with some of the most-known bilingual legal dictionaries. These dictionaries have not been recently revised or consolidated in a new edition, and most of them are still available in hard copy examples include Al Wahab's Law Dictionary (Al Wahab, 1063), Faruqi's Law Dictionary (Faruqi, 1969/2008), *Hakki's Dictionnaire des termes juridiques et commerciaux* (Hakki, 1982), *Moussa's Lexique des termes juridiques, français-arabe* (Moussa, 1993) and *Belefkih's Lexique pluridisciplinaire* (Belefkih, 1998), *Dictionnaire pratique de droit-commerce-finance*, *Abi Fadel's Dictionnaire des termes juridiques* (Abi Fadel, 2004), *Lexique des termes juridiques* (Chamel et al., 2010), and *Vocabulaire juridique* (Al-Qadhi, 1998). she states that these resources have four features: they represent a list of terms with no description or terms related information, they present general subject matter, they contain various term proposals without context, and they represent translations of existing dictionaries.

Halimi (2023) reviews online resources for Arabic legal translation namely dictionaries and glossaries, terminological databases such as UNTERM, International Labour Organisation Terminology Database (ILOTERM), International Monetary Funds Multilingual Terminology Database (IMF TermBase), and parallel corpora including the ones available in the online Official Document System (ODS). She states that these resources are not exclusively law-oriented per se because they are UN-system oriented, thereby restricting their applicability for translation situations, since they do not provide all the information required to meet translators needs. Secondly, they are less aware of national-system variations, thus creating uncertainties with regard to term applicability, as existing variants are not covered. The ODS contains full-text digital UN documents from the Security Council and General Assembly for administrative purposes. Thus, using these resources as a legal translation source proves to be time-consuming.

Many authors have contributed to the analysis of Arabic legal language and what makes the incongruity of its terminology translation into English and French a challenge. They try in many works mentioned above to emphasise the efficacy of the use of translation techniques, theoretical models, etc., to solve the untranslatability of legal terms in Arabic legal documents. In fact, the challenges of translating legal terms into Arabic are multifaceted and demand a

deep understanding of both the legal context and the linguistic intricacies of the Arabic language. Attempting to address this challenge through the adaptation of translation techniques often falls short and is not a recommended practice.

There is a significant lack in the literature of Arabic legal terminology in terms of efforts to develop and label legal terminology resources and other resources that can be adapted to the needs of the legal translators. Halimi (2023) in the review she conducted on online resources for Arabic legal translation compared to EU resource concludes that there is a real terminological lack and heterogeneity in Arabic electronic legal dictionaries, glossaries, terminological bases and institutional websites, which evidently compromises the reliability of the information that is presented. This absence of comprehensive terminology resources has a direct effect on many NLP application namely the effective training of MT systems to produce accurate and reliable results and without access to a robust and extensive terminology resources MT systems struggle to grasp the intricate nuances and context-specific usage not only of legal terms but in all specialised domains (Halimi, 2023).

To address this research gap, this work aims to foster collaborative initiatives among legal experts to develop a robust infrastructure of legal terminology resources. These resources will serve as a foundational knowledge base for training MT systems, enabling them to capture the specific nuances of Arabic legal terminology through specialised tools and technologies optimised for accurate and reliable legal translations.

## **2. Arabic Machine Translation in the Legal Domain**

MT is a subfield of AI and computational linguistics that focuses on the development of systems capable of automatically translating text or speech from one language to another (Poibeau, 2017). In other words, MT translates texts or speech from one natural language to another. It receives a source sentence  $S = \langle s_1, s_2, \dots, s_i \rangle$  and generates a target sentence  $T = \langle t_1, t_2, \dots, t_j \rangle$  by translating the source sentence and give the meaning of it in the target language (Alqudsi et al., 2014).

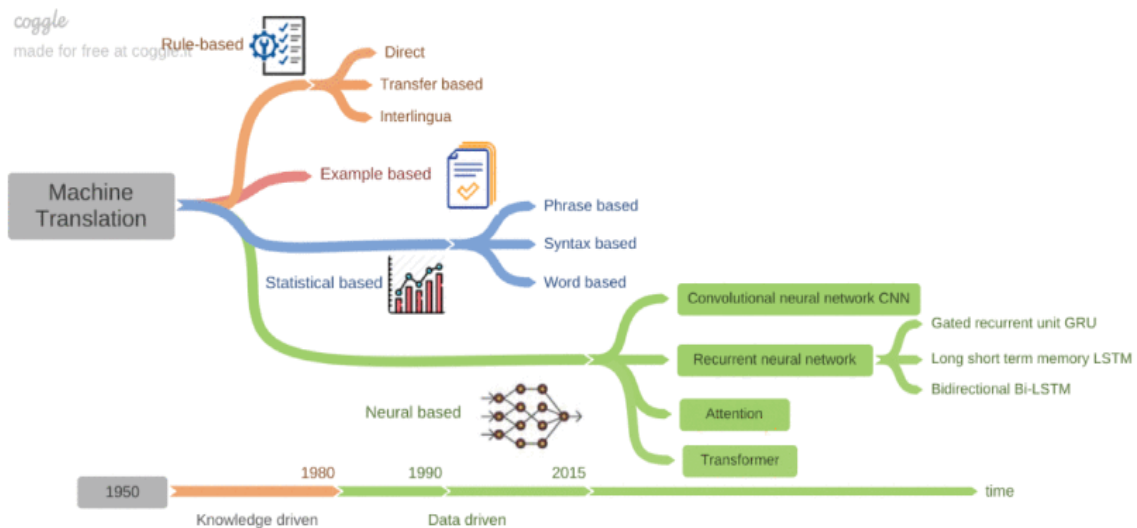


Figure 2. Evolution and Classification of Machine Translation Approaches.

MT has three main architectures: Rule-Based (also known as knowledge-driven) approaches (RBMT), data-driven approaches, which include Example-Based (EBMT), Statistical (SMT), NMT, and hybrid MT. Figure 2 visualizes the evolution and general classification of these MT approaches.

NMT has rapidly become the leading approach in MT, leveraging neural networks, primarily RNNs and transformers, to model the entire translation process. Initially, neural techniques supplemented traditional SMT by enhancing components like scoring (Schwenk, 2012; Lu et al., 2014), reordering (Yamagishi et al., 2016; Poibeau, 2022), and pre-ordering models (Jehl et al., 2015). Key developments, such as the attention mechanism and advancements like byte-pair encoding (Chorowski & Bahdanau, 2015; Jean et al., 2015b), helped NMT systems deliver high-quality translations, even for longer sentences, eventually surpassing SMT. This shift was swift by 2017, nearly all submissions in translation competitions were NMT-based. Recently, the emergence of LLMs, has further driven progress in translation, enabling models to understand and generate natural language with unprecedented fluency and accuracy, and expanding NMT's potential to handle complex multilingual contexts at scale (Koehn, 2020; Brown et al., 2020).

The emergence of these models shows significant improvements in translation quality, outperforming traditional approaches and achieving more fluent and contextually accurate translations especially in high resourced languages. Nevertheless, their performance in low-resource languages, such as Arabic, is a complex and evolving area of research. In the context of legal terminology, the challenge is heightened due to the specialised nature of legal language

that is governed by many factors the strongest of which is culture-bound connotations, and the need for specialised training data.

### 2.1. Machine Translation of Arabic Legal Terminology

This section will delve into providing a comprehensive review of the literature where it discusses studies on the quality of machine-translated legal terminology in different language pairs, in addition to the methods and approaches adapted earlier by distinguished scholars to address this problem. The outcomes achieved from these efforts will also be discussed.

Kamir et al. (2002); Shirko et al. (2010); Darwish et al. (2020) agree that in the dynamic intersection of language, law, and technology, the MT of Arabic legal terminology stands as a compelling subject demanding scholarly attention. They add that the foremost challenge for NLP in Arabic, is overcoming ambiguity and that it is quite common for a single word or term to have various possible translations, each with significantly distinct meanings (Attia, 2008), Consequently, the translation of legal terms becomes an even more formidable task, intensifying the difficulties in establishing trust between users and MT systems (Deng et al., 2022).

With regards to investigating MT pros and cons, Koehn and Knowles (2017) examine several challenges to NMT and give empirical results on how well the technology holds up compared to PB-SMT. To do this, they train both NMT and PB-SMT for German to English on domains using a series of specialised corpora (i.e., law (Acquis), Medical (EMEA), IT, Koran (Tanzil), subtitles) obtained from OPUS (Tiedemann, 2012). They note that the output of the NMT system is often quite fluent but completely unrelated to the input, while the PB-SMT output betrays its difficulties with coping with the out-of-domain input by leaving some words untranslated. They conclude that despite the recent successes, NMT must still overcome various challenges, most notably performance in out-of-domain and under-resourced conditions. Zakraoui et al. (2021) conduct a survey related to Arabic MT challenges which they split into two categories, namely linguistic (i.e., morphology richness, syntactic word reordering, Word Sense Disambiguation, vocalization, dialectal variation, gender bias, etc.) and technical (i.e., low-resource language, domain mismatch, Out-Of-Vocabulary, word alignment, sentence length, among others). They mention the limited availability of parallel data for researchers and suggests the need for multi-domain NMT approaches (Zakraoui et al., 2021). Alsohybe et al. (2017); Hadla et al. (2014); Han (2016) prove the ineffectiveness of NMT systems, mainly Google Translate (GT) when producing Arabic-English translations. In a

quantitative comparative evaluation, Ziemski & Junczys-Dowmunt (2016) report the performance of PB-SMT and NMT across fifteen language pairs and thirty translation directions on the UN Parallel Corpus v1.0 and show that for all translation directions, NMT is either on par with or surpasses PB-SMT. Bentivogli et al. (2016) analyse the output of MT systems in an English-to-German translation task from the IWSLT 2015 evaluation campaign (Cettolo et al. 2015) by considering different linguistic categories. Toral and Sánchez-Cartagena (2017) conduct an evaluation to compare NMT versus PB-SMT outputs across broader aspects (e.g. fluency, reordering) for 9 language directions. Castilho et al. (2017) conduct a more extensive qualitative and quantitative comparative evaluation of PB-SMT and NMT using automatic metrics and professional translators. Popović (2017) carries out an extensive comparison between NMT and PB-SMT language-related issues for the German to English language pair. In a nutshell, these works (Bentivogli et al. 2016; Castilho et al. 2017; Popović 2017; Toral and Sánchez-Cartagena 2017) show that NMT provides better translation quality (e.g. more fluent, and fewer lexical, reordering and morphological errors) than the previous state-of-the-art PB-SMT. This trend continues in other studies and use-cases such as in the translation of literary text (Toral & Way 2018), MT post-editing set-up (Forcada et al. 2017), industrial set-up (Shterionov et al. 2018), translation of patent documents (Long et al. 2017; Kinoshita et al. 2017), less-explored language pairs (Klubička et al. 2017; 2018), highly investigated “easy” translation pairs (Isabelle et al. 2017), and translation technical products catalogues (Beyer et al. 2017).

As far as legal terminology in NMT is concerned, Macketanz et al. (2017) conducts a linguistically driven fine-grained evaluation to compare rule-based, phrase-based and neural MT engines on English–German pair, based on a test-suite for MT quality. Their findings agree with the previous ones. From German-to-English, PB-SMT, despite having the lowest average score, performs better on named entities and terminology. However, from English-to-German, a commercial NMT engine excels in term translation. Conversely, Avramidis et al. (2020) report that their PB-SMT system outperforms NMT on terminology translation on both in-domain (IT domain) and general domain test suites in an English-to-German translation task. Specia et al. (2017) carried out an error evaluation using the Multidimensional Quality Metrics (MQM) framework (Burchardt, 2013) in an MT post-editing environment. The list of errors is divided into three main categories: accuracy, fluency and terminology. According to the annotation results, more terminology-related errors are found in NMT translations than in PB-SMT translations in an English-to-German task (139 vs 82 error), and the other way round in an

English-to-Latvian task (31 vs 34 error). They report that PB-SMT outperforms NMT on term translation, which they speculate could be because their technical term bank was part of the training data used for building their PB-SMT system. Vintar (2018) conducts an automatic and small-scale human evaluation on the terminology translation quality of GT compared to PB-SMT for Slovene to English in the domain of karstology<sup>10</sup>. Their findings confirm that NMT is slightly better than PB-SMT in terminology translation, while it shows the reversed picture with PB-SMT outperforming NMT from English to Slovene.

To investigate domain term translation in PB-SMT and NMT, Haque et al. (2020) carried out experiments with two morphologically divergent languages, English and Hindi. They adapt a technique that semi-automatically create a gold standard test set from an English to Hindi judicial domain parallel corpus. The sentences of the gold standard test set were translated with their PB-SMT and NMT systems, and the patterns of the terminology translation errors on a sample set of translations were inspected. They also propose a terminology translation typology, categorizing translations of domain-specific terms from the gold-standard test set with input from legal domain experts. Given the classification results, they carried out a comparative evaluation of PB-SMT and NMT on terminology translation. They find that NMT is less error-prone than PB-SMT as far as terminology translation is concerned (8.3% versus 9.9% and 11.5% versus 12.9% error rates in English to Hindi and Hindi to English translation tasks, respectively; differences in error rates are statistically significant). Their findings contradict some of the findings of several previous studies (Beyer et al. 2017; Burchardt et al. 2017; Specia et al. 2017; Vintar, 2018; Avramidis et al. 2020) who reports that PB-SMT outperforms NMT in term translation. As far as the terminology translation and fine-grained error categories are concerned, they find that the NMT systems commit fewer lexical, reordering and morphological errors than the PB-SMT systems. The differences in error rates with regards to the lexical selection and reordering errors are statistically significant in both MT tasks, and the difference of the morphological error rates is statistically significant in the Hindi to English task. The morphological errors are seen relatively more often in PB-SMT than in NMT when translation is performed from a morphologically rich language (Hindi) to a morphologically poor language (English).

For the Arabic-English pair and vice versa, Al Shehab (2013) conducts a study on the translatability of English legal sentences into Arabic using GT. He adapts Šarčević (2000)

---

<sup>10</sup> A field within geomorphology, specializing in the study of karst formations: <https://fr.wikipedia.org/wiki/Karstologie>.

functional equivalence that can be categorised into three levels: near-equivalence, partial equivalence, and non-equivalence. To do this, the researcher uses a test which consists of six English legal articles. To achieve the validity of the test, the researcher adapts what he calls "committee translation". Two legal professionals are chosen from the Department of Translation of two different institutions "Yarmouk and Jordanian University" in Jordan to get the reference translation. Al Shehab (2013) finds that GT results are partially acceptable, although the meaning is understood, but the system struggles with producing the right terminology choice. There is substantial agreement between Al Shehab (2013) and Alkatheery (2023) who also conducts a study examining the accuracy factor of GT in the translation of Arabic legislation into English. He finds out that terminological errors were the highest leading to faulty, inaccurate translations, and that GT cannot be trusted for the MT of Arabic legislation. In an attempt to address this issues, Berrichi and Mazroui (2021) study the limited vocabulary and Long Sentences Constraints in English to Arabic NMT systems. To do this, they perform various approaches to deal with questions raised by the limited vocabulary size in NMT systems and the poor performance of these systems for long sentences. These approaches are examined under low resources (140K sentence pairs) and high resources (750K sentence pairs) conditions. They experiment with various Factored NMT models that integrate linguistic knowledge into the input or output of the translation system using multiple factors as translation units. Experiments carried out on these models highlighted their benefits in terms of both vocabulary coverage and the translation quality compared to both the word-level Baseline system that are Byte Pair Encoding segmentation based. Regarding the long sentence issue in NMT systems, they propose two segmentation techniques to overcome this challenge. The first method consists of segmenting long sentences into sub-sentences using a lexical marker list that they collect. For the second one, they exploit the Moses system to extract aligned bilingual phrases from parallel sentences. They find that these techniques have proved a significant improvement, especially in low resource conditions like Arabic. Indeed, the SplitRules (Train-Test60)) model that they develop for English to Arabic direction improved according to BLEU score by 2.81 compared to the Baseline Model (Baseline WithoutSplit).

## 2.2. Quality Evaluation of Machine-Translated Terminology in the Legal Domain for Arabic

As the demand for efficient and accurate translations continues to surge, the evaluation process becomes pivotal in refining and advancing translation systems. Two primary methodologies have emerged to assess the efficacy of machine-generated translations: automatic evaluation

metrics and human evaluation. Automatic evaluation metrics are computational tools designed to quantify the quality of machine-generated translations without human intervention (Papineni et al., 2002). These metrics employ algorithms to compare MT against reference translations, assigning numerical scores based on various linguistic features. Human evaluation involves the subjective assessment of machine-generated translations by human judges (EIFqih & Monti., 2024). These judges, often bilingual or multilingual experts, evaluate translations based on criteria such as fluency, adequacy, and faithfulness to the source text (EIFqih & Monti., 2024). Human evaluation provides a more holistic and context-sensitive perspective, capturing nuances that automated metrics may overlook (Comparin & Mendes, 2020). As long as the automatic MT evaluation is concerned, researchers use different metrics such as TER (Translation Edit Rate) (Snover et al., 2006), Word Error Rate (WER) (Juang & Rabiner, 2022), Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Lavie & Denkowski, 2009; Banerjee & Lavie, 2005), Arabic Language- BiLingual Evaluation Understudy metric (AL-BLEU) (Bouamor et al., 2014) which extends BLEU (Papineni et al., 2002) to deal with Arabic rich morphology.

In the Arabic-English language pair and vice versa and as far as human evaluation is concerned, Jabak (2019) assessed the quality and accuracy of GT performance in the Arabic-English pair. The error analysis showed that GT output contained lexical and syntactic errors. Lexical errors were the highest leading to inaccurate translations. The study concludes that GT system cannot be used separately without human post-editing interference. Alkathery (2023) agrees with Jabak (2019) in his study where he assesses the GT accuracy in the translation of legal text from Arabic into English relying on human evaluation. The findings indicate that GT is unsuccessful in decoding the majority of the Arabic legal terms presented in their data. Their results oppose Hailat et al. (2014) study on GT and Babylon accuracy. Their work includes over 1,000 sentences categorised into the four sentence types: declarative, interrogative, imperative, and exclamatory. The automatic evaluation of these translations was conducted using the BLEU metric. Results indicated that GT produced more accurate translations overall. In a similar study, Adly and Al Ansary (2010) examined the differences among three MT systems: GT, Babylon, and Tarjim. They used an automatic evaluation approach and found that all three systems struggled with capturing the semantic and terminological cohesion as well as the stylistic elements of Arabic. Additional research efforts include statistical error analysis of MT for Arabic, allowing native speakers to review linguistic aspects of MT errors (El Marouani

et al., 2020). Other approaches employ neural networks to detect errors (Madi & Al-Khalifa, 2020) in Arabic texts or to correct them (Tufano et al., 2018).

In the ever-evolving era of AI where LLMs appears to be a game-changer in the field of NLP, particularly in the realm of MT. Models like Generative Pre-trained Transformers (GPT)<sup>11</sup> (Brown et al. 2020), Gemini<sup>12</sup> (Rohan et al. 2023) and BERT (Devlin et al., 2018) of Google show the ability to understand context, syntax, and semantics and generate translations that surpass any other existing MT system (Zhang et al., 2023). Authors like Hendy et al. (2023) have started conducting comprehensive evaluation on the MT performance of GPT models namely GPT-3.5. They experiment with eighteen different translation directions involving high and low resourced languages across four different domains (e.g., conversational domain, News domain, e-Commerce domain, and social domain). For the evaluation process, they employ both human evaluation and the latest neural network-based automatic evaluation metrics together with the conventional MT evaluation metrics (e.g., COMET-22 (Rey et al., 2020), COMETkiwi (Rey et al., 2022), ChrF (Popović, 2015), and BLEU (Papineni et al., 2002)). They find out that GPT models excel at translating well-represented languages. They also produce highly fluent and competitive translation outputs even in the zero-shot setting especially for the high-resourced language translations but fell short in under-resourced languages. They also assess translation and reasoning tasks and detect discrepancies in the level of support of the tasks for the same languages.

While the use of LLMs in MT is rapidly progressing, Hendy et al. (2023) recall research directions to improve quality and understanding of the LLMs MT tasks, focusing on under-represented languages, in-context learning, model fusion, and developing reliable metrics that can measure the contextual correctness of LLMs-generated translations. Castilho et al. (2023) conduct a study to answer a question on whether online GPT models care for context. They address the challenges of evaluating document-level MT in light of recent advancements in context-aware NMT. Their work investigates how well online MT systems deal with six context-related issues, namely lexical ambiguity, grammatical gender, grammatical number, reference, ellipsis, and terminology, when a larger context span containing the solution for those issues is given as input. To do this, they extract sentences from the DELA corpus (Castilho et al., 2021) to compile their test suite on 4 languages: German, Irish, Brazilian Portuguese, and Simplified Chinese. Their results show that, while the change of punctuation

---

<sup>11</sup> <https://chat.openai.com/>

<sup>12</sup> <https://gemini.google.com/>

in the input yields great variability in the output translations, the context position does not seem to have a great impact. Moreover, the GPT model seems to out-perform the NMT systems but fell short for Irish. They highlight that while the GPT models ability to generate multiple translation versions may lead to improved solutions, it can also result in inconsistencies, particularly in terminology. Khoshafah (2023) hypothesises that cross-cultural communication has become more accessible with the advent of ChatGPT as a translator for multiple language pairings, including Arabic and English. Hence, she believes that the accuracy of ChatGPT's translation needs to be evaluated due to limitations in its understanding of domain-specific terminology and cultural context. Therefore, Khoshafah (2023) targets evaluating the translation accuracy of ChatGPT-3.5 by comparing its output with reference translations. To do this, she uses different types of texts including general texts from the media, literary, scientific, religious and legal texts extracted from sources such as books and websites. The results indicate that ChatGPT-3.5 can be used as a translator for simple content, though it struggles with complexities that require human intervention. Despite ChatGPT-3.5 generally providing accurate translations, its limitations make it unsuitable for some texts, such as legal documents, medical reports, scientific studies, and literary works. Therefore, users should be cautious while using ChatGPT for such texts, as human intervention is required to ensure accuracy.

Han (2016) and Lee et al. (2023) evaluate several automatic metrics, including those mentioned previously. They demonstrate that no definitive conclusions can be drawn regarding the superior performance of any specific metric over others. They add that while automatic metrics capture the average performance on how well an MT model translates sentences, they do not provide insights into which linguistic aspects the MT models struggle with in producing fluent output.

### **3. Linguistic and Legal Terminology Resources for Arabic Machine Translation Applications**

The development of linguistic and domain-knowledge terminology resources for Arabic MT applications plays a pivotal role in enhancing their accuracy and fluency. Linguistic resources contribute to the understanding of the structural and grammatical aspects of languages. These resources include extensive lexicons, syntactic rules, and morphological information specific to each language. For instance, in NMT systems, the incorporation of linguistic resources enables the model to better capture the properties of sentence structures, word order, and grammatical nuances, resulting in more contextually accurate translations. Furthermore, linguistic resources aid in mitigating challenges arising from the morphological complexity of

languages, especially in for Arabic, where root-based morphological structures pose unique difficulties that require specialised linguistic knowledge. On the other hand, terminology resources, particularly in domains such as legal, medical, and others, contribute to the precision and domain-specific relevance of MT systems. Developing comprehensive glossaries, databases, and specialised terminology ensures the model accurately translates content within specific domains.

Linguistic Data Consortium (LDC)<sup>13</sup>, the backbone behind several refinements in favour of MT systems, is an open consortium of universities, libraries, corporations and government research laboratories who have joined forces to ensure that language resources reach a broad spectrum of users including students, scholars, researchers, and developers, etc. The LDC Catalogue contains over 500 holdings and grows at a rate of 30–36 publications annually. This includes resources developed at LDC as well as datasets created by researchers around the world. LDC’s Arabic language resources represent all the data types found in LDC’s Catalog: speech, text, video and lexicons. Those collections were particularly created for MT research and technology development. LDC collects large amounts of Arabic text from various sources, including newswires and web text (blogs, newsgroups, and emails). Two groups of corpora developed from that material are exceptionally dedicated to MT applications: parallel text corpora and the series of corpora used in NIST’s (US National Institute of Standards and Technology) OpenMT technology evaluations. Most translations are performed by humans and undergo a series of rigorous quality checks. Other parallel text datasets are automatically generated (see *Table 1*).

The LDC community also provides a morphological analyser called Standard Arabic Morphological analyser (SAMA). It is based on, and updates, the Buckwalter Arabic Morphological Analyzer (BAMA), developed by Buckwalter (2004). This software considers each Arabic word token in all possible ‘prefix-stem-suffix’ segmentations, and lists all known/possible annotation solutions, with assignment of all diacritic marks, morpheme boundaries (separating clitics and inflectional morphemes from stems), and all Part of Speech (POS) labels and glosses for each morpheme segment (Bies et al., 2012).

---

<sup>13</sup> <https://www ldc upenn edu/>

Table 1: The linguistic Data Consortium Resources.

<b>The Linguistic Data Consortium (LDC)</b>	
<b>Parallel Texts</b>	<b>US National Institute of Standards and Technology (NIST) Resources</b>
<ul style="list-style-type: none"> <li>• GALE Phase 1 Arabic Broadcast News Parallel Text – Part 1 LDC2007T24</li> <li>• GALE Phase 1 Arabic Blog Parallel Text LDC2008T02</li> <li>• GALE Phase 1 Arabic Newsgroup Parallel Text – Parts 1 and 2 LDC2009T03, LDC2009T09</li> <li>• ISI Arabic-English Automatically Extracted Parallel Text LDC2007T08</li> <li>• Multiple-Translation Arabic (MTA) Parts 1 and 2 LDC2003T18, LDC2005T05</li> </ul>	<ul style="list-style-type: none"> <li>• NIST 2002 Open Machine Translation (OpenMT) Evaluation LDC2010T10</li> <li>• NIST 2003 Open Machine Translation (OpenMT) Evaluation LDC2010T11</li> <li>• NIST 2004 Open Machine Translation (OpenMT) Evaluation LDC2010T12</li> <li>• NIST 2005 Open Machine Translation (OpenMT) Evaluation LDC2010T14</li> <li>• NIST 2006 Open Machine Translation (OpenMT) Evaluation LDC2010T17</li> <li>• NIST 2008 Open Machine Translation (OpenMT) Evaluation LDC2010T21</li> <li>• NIST 2009 Open Machine Translation (OpenMT) Evaluation LDC2010T23</li> </ul>

Similarly, Habash et al. (2017) develop a parallel corpus called ‘Arab-Acquis’ for evaluating MT between Arabic and 22 European Languages. It consists of over 12,000 sentences from the JRC-Acquis (Acquis Communautaire) corpus translated twice by professional translators, from English and to French, with a total of over 600,000 words. The corpus can be employed for tuning, development, and testing.

Numerous studies have focused on addressing the challenges of low-resource languages like Arabic. For example, Ren et al. (2018) introduced a novel triangular training architecture (TA-NMT) to enhance translation performance for low-resource language pairs. This architecture uses a resource-rich language as an intermediate latent variable, and it optimizes the translation models using a unified bidirectional Expectation-Maximization algorithm. Their approach significantly improves the translation quality for rare languages such as Arabic.

Liu et al. (2020) present mBART a denoising auto-encoder extended by pre-training BART (Lewis et al., 2019) on several monolingual language corpora. Their model is fine-tuned to

translation tasks without language-specific modifications or initialization schemes. mBART initialization leads to significant gains (up to 12 BLEU points) across low and medium-resource pairs (<10M bi-text pairs), without sacrificing performance in high-resource settings. For Arabic, their mBART25 (pretrained on 25 languages) has led to an increase of 10.1 BLEU points from Arabic to English pair. Similarly, Lin et al. (2020) propose multilingual Random Aligned Substitution Pre-training (mRASP), an approach to pre-train a universal multilingual NMT model for many languages, which can be used as a common initial model to fine-tune on arbitrary language pairs. It brings words and phrases with similar meanings across multiple languages closer in the representation space. They pre-train their model on 32 language pairs jointly with only public datasets. The model is then fine-tuned on downstream language pairs to obtain specialised MT models. Accordingly, the results on English-Arabic pair are improved by 1.8 BLEU points.

To improve MT models performance, Abid (2022) proposes an NMT model, by bootstrapping existing parallel sentences and complementing this with multilingual training to achieve strong baselines. The author creates a 4-way benchmark dataset between Egyptian, Levantine, Modern Standard Arabic, and English freely available. The results of the conducted experiments suggest that a multilingual model of dialects and Modern Standard Arabic, along with bootstrapping, achieves the best results by 2.56 (9%) BLEU score.

NMT training generates a target vocabulary consisting of individual, unique target tokens from the training corpora, sorted by frequency. Ideally, this vocabulary includes every token in the training data; however, it is typically restricted in practice to a fixed number of the most frequently occurring tokens. Jean et al. (2015) depict a limitation in NMT vocabulary size: as the number of target words increases, so does the complexity of both training and decoding. Recent techniques have been developed to address these limitations. For instance, Dougal & Lonsdale (2020) propose methods for improving NMT quality using terminology injection, which involve substituting or replacing an approved term where it does not emerge during translation. The method involves the use of the long short-term memory attention mechanism, which helps in correctly identifying semantic entities and aligning the tokens that represent them in both the source and target languages. During decoding, appropriate terminology is injected into matching alignments to ensure consistent translation of individual terms with an approved multilingual terminology collection. For training and evaluating the NMT models, they use three relevant language resources: EuroParl, Microsoft C# technical documentation,

and Microsoft terminology. Additionally, they apply terminology injection with an externally specified termbase to assess the system's functionality.

For NLP software application environments, Najar et al (2016) develop a large terminological dictionary of Arabic compound words using NooJ, which is a linguistic engine based on large coverage dictionaries and grammars. It uses Finite State Transducers to parse text corpora made up of hundreds of text files in real time and associate each recognised entry with its related information, such as morpho-syntactic information (POS, Gender, Number, etc.), syntactic and semantic information (e.g. transitive, Human, etc.). To do this, they collected about 63500 multiword expressions (MWEs) associated with 20 fields: economy, media, religion, financial, political, etc. educational, computer, sportive and psychological. The terms used cover fixed expressions using morphological grammar and their inflectional variants. Their results demonstrate high precision after testing the coverage of their resources on the test corpus.

The emergence of LLMs in recent AI advancements in both industry and academia has marked a transformative leap in many NLP application including but not limited to text generation, summarization, question answering, and MT (Zhang et al., 2023; Chang et al., 2023; Briva-Iglesias et al., 2024). These models like Generative Pre-trained Transformers (GPT) (Brown et al., 2020), mBART (Lui et al., 2020) and mBART-50<sup>14</sup> (Tang et al., 2020), Llama<sup>15</sup> (Touvron, 2023) by Meta AI, Gemini (Rohan et al., 2023), M2M100<sup>16</sup> (Fan et al., 2021), and No-Language-Left-Behind (NLLB) (Costa-jussà et al., 2022), etc., are trained on a vast amount of data, sometimes encompassing billions or even trillions of tokens. Given these models' ability to perform MT tasks, they struggle with idiosyncratic or context-dependent translations and are prone to issues like hallucination and the propagation of biases present in their training data, especially in specialised domains or languages with complex grammatical structures, such as Arabic (Khoshafah, 2023; Chang et al., 2023; Briva-Iglesias et al., 2024). Among the efforts to address these models' pitfalls, Al-qurishi et al. (2022) introduce AraLegal-BERT, a bidirectional encoder transformer-based model that has been thoroughly tested and carefully optimised with the goal of amplifying the impact of NLP-driven solutions on jurisprudence, legal practices, and legal documents including 336K legal cases including appealing, 3K of Laws and regulations in the Kingdom of Saudi Arabia and Gulf countries, 5K

---

<sup>14</sup> [https://huggingface.co/docs/transformers/model\\_doc/mbart](https://huggingface.co/docs/transformers/model_doc/mbart)

<sup>15</sup> <https://llama.meta.com/>

<sup>16</sup> [https://huggingface.co/docs/transformers/model\\_doc/m2m\\_100](https://huggingface.co/docs/transformers/model_doc/m2m_100)

of reports and studies, academic courses, forms, reports, contracts, and 6K of master and PhD theses, research papers, magazines, dictionaries, and Fiqh books which are foundational texts in Islamic jurisprudence. The authors use the previously mentioned dataset to fine-tune ArabLegal-BERT and evaluate it on three Natural Language Understanding (NLU) tasks. They hypothesize that by focusing on the legal domain for Arabic, their study can reveal methods for adapting an NLP model to fit any thematic domain. To test this, they examine whether pretraining BERT from scratch with Arabic legal examples provides a stronger foundation for handling documents with Arabic legal terminology than using the standard, general-purpose version of the algorithm. Their findings indicate that the base version of AraLegal-BERT achieves higher accuracy on legal texts compared to the original BERT model, rivalling the performance of computationally intensive models while delivering results more quickly and with fewer resources. They conclude that the base version alone can attain accuracy comparable to larger versions of the original BERT model when trained with domain-specific examples like those used in this study. Unfortunately, the authors inform that the model is proprietary and cannot be made publicly available.

Moslem et al. (2023) present the outcomes of fine-tuning Mistral 7B<sup>17</sup> (Jiang et al., 2023) for adaptive MT. The fine-tuning involves combining zero-shot and one-shot translation prompts within the medical domain. They experiment using a relatively small dataset of 20,000 segments incorporating a mix of zero-shot and one-shot prompts from Spanish into English. Initially, they mixed Spanish to English medical datasets from OPUS (Tiedemann, 2012), namely ELRC (Berzins et al., 2019), The European Medicines Agency (EMA)<sup>18</sup>, SciELO<sup>19</sup>, and TICO-19<sup>20</sup> (Anastasopoulos et al., 2020). Then, they filter them to exclude duplicates and too-long segments. The results show that fine-tuning a general purpose LLM such as Mistral 7B can improve its in-context learning ability. They also find that incorporating a mix of zero-shot and one-shot prompts in the training data helps improve both regular zero-shot translation, and one-shot translation that incorporates a fuzzy match.

Linguistic and legal terminology resources are crucial for improving MT systems performance. They help them make the right terminology choice, enhancing both accuracy and reliability. In

---

<sup>17</sup> <https://mistral.ai/news/announcing-mistral-7b/>

<sup>18</sup> <https://www.ema.europa.eu/en/homepage>

<sup>19</sup> <https://scielo.org/en/>

<sup>20</sup> <https://paperswithcode.com/dataset/tico-19>

what follows, we present our contributions, starting with an evaluation that guided the selection of methods for deconstructing our proposed solutions to our research problematic.

### Chapter Three: Research Design, Methodology, and Experimental Set-up

Research into MT of Arabic legal terminology is driven by the unique challenges of Arabic legal discourse, which combines classical Arabic, modern legal terms, and regional dialects, all layered with religious and cultural nuances. This complexity spans syntax, morphology, and lexicon, where precise terminology and intricate grammatical structures are essential for clarity and authority. Influenced by religious texts (e.g., the Quran), scholarly traditions, and cross-cultural exchanges, Arabic legal language is a dynamic linguistic domain that demands specialised expertise for accurate translation. Given the scarcity of domain-specific data in Arabic for the legal field and its impact on MT performance, this study seeks to bridge this gap by developing a comprehensive legal corpus and a standardised multilingual terminology resource. These resources support both preliminary and comparative evaluations of MT systems, establishing methods to optimize MT performance in Arabic legal contexts.

#### 2. Methodology

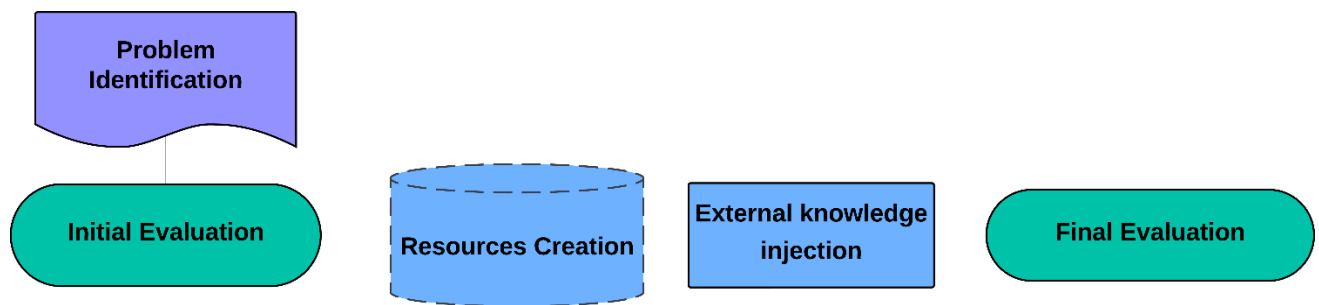


Figure 3: The General Methodology Pipeline

Figure 3 outlines the general methodology used in this work; it stages:

- **Problem Identification**
  - This is the initial stage where the specific issues related to the untranslatability of legal terms within Arabic legislation are identified.
- **Initial Evaluation**
  - After identifying the problematic areas, an initial evaluation is performed to define the research problematic. This involves a quantitative analysis of the MT output.
- **Resources Creation**

- Based on the initial findings, we start to create resources that can address the identified problems, which we believe are key to improve the MT systems performance.
- **External Knowledge Injection**
  - This stage concerns the integration of external knowledge into the MT systems to improve their performance. This involves using the resources developed, incorporating legal knowledge bases, or adapting the translation models to better handle the nuances of the legal language.
- **Final Evaluation**
  - We conduct a final evaluation to determine the effectiveness of our contributions.
- **End**
  - The final step indicates the conclusion of the methodology, where we summarize the outcomes of the research. This involves documenting the improvements made to the MT systems and discussing potential future work in the legal domain for Arabic.

### **3. Data Collection and Resources creation**

In this section, we discuss the data collected and used in this thesis and the meticulous management strategies applied to ensure its integrity and relevance. Consider the following pipeline:

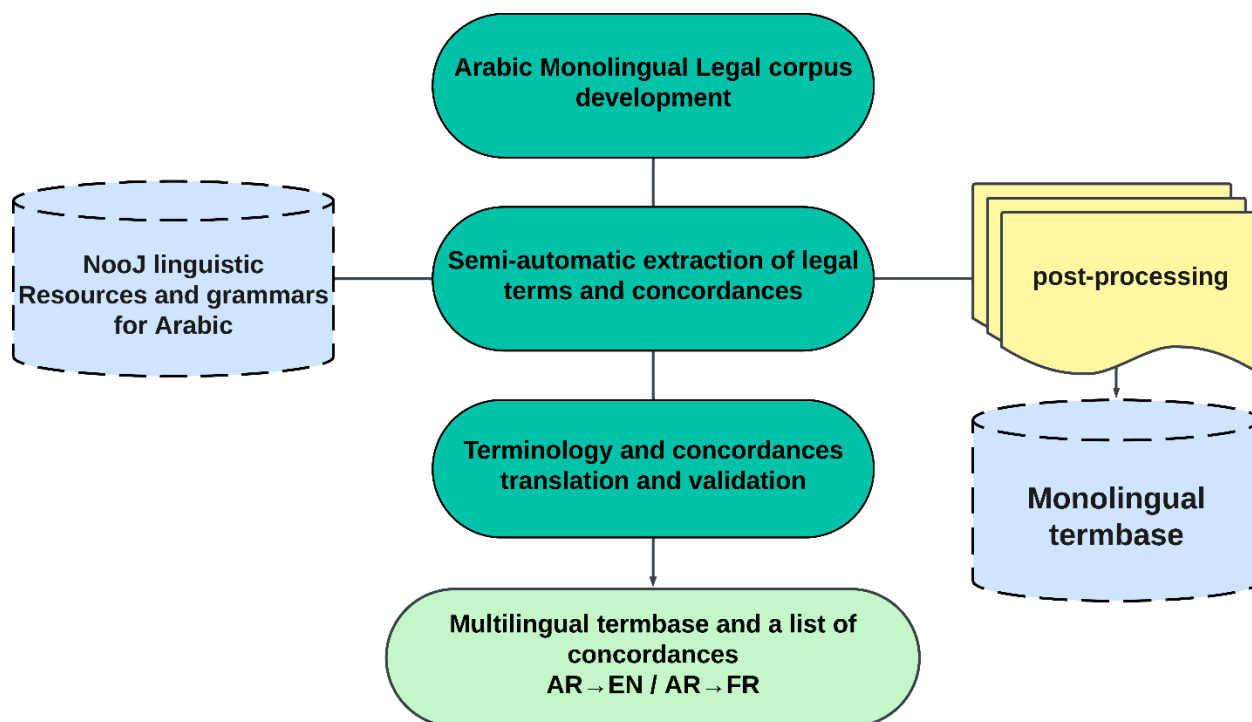


Figure 4. Dataset.

This pipeline diagram foresees the following steps:

- **Arabic Monolingual legal corpus development:** we meticulously collect a series of Arabic legislation of several types, namely marriage contracts, divorce provisions, decrees, constitutions, codes, etc., of different Arab countries including Morocco, Algeria, Tunisia, Egypt, Qatar, United Arab Emirates, Saudi Arabia, etc., (see *Table 2* for information about the size of the corpus). After the collection of the Arabic legislation, we then need to preprocess the texts to clean and standardise the data. This involves removing non-textual elements, normalising text formats, and handling common issues like typos or inconsistent formatting. Afterwards, the corpus is developed using NooJ<sup>21</sup> (Silberztein, 2016), an NLP software application, which serves as the foundational dataset for extracting legal terms.
- **Semi-automatic legal term candidate extraction and validation;** a semi-automatic process is employed to extract potential legal terms, thanks to NooJ linguistic resources developed by Mesfar (2008) and especially syntactic grammars (see *Figure 5*) which help identify patterns that are typical of legal terminology, such as specific phrase structures or recurrent legal expressions.

<sup>21</sup> <https://nooj.univ-fcomte.fr/>



For validation, the results are submitted for filtration and examination, where we delete irrelevant or inaccurate terms (i.e., terms that do not express a legal practice) and add new ones that were missed by the automatic process and choose to process 1,949 entries (70% are single terms and 30% are MWEs). Then, all the listed entries were voweled using Character-based Arabic Tashkeel Transformer<sup>22</sup> (Alasmary et al., 2024) and then manually verified. Diacritization is fundamental here because in some cases, we can find entries that have different ways of vocalization and different meanings, for example, the term ‘الْقَطْع’ with ( . ) means *pieces* but ‘الْقَطْع’ with ( َ ) means *certainty*. Therefore, manual vocalization is an extremely important step since it allows us to vowel entries depending on their semantic connotation. This helps reduce semantic ambiguities in Arabic legal texts to differentiate between non-legal and legal terms.

- **Concordances extraction:** We use NooJ syntactic grammar (see *Figure 6*), through which we create a node containing all the terms previously extracted to obtain the sentences in which they occur.

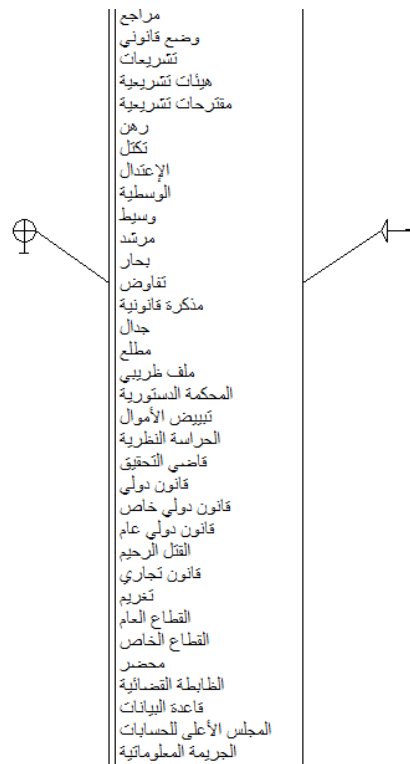


Figure 6: NooJ Syntactic Grammar for Concordances Extraction.

<sup>22</sup> <https://huggingface.co/spaces/MohamedRashad/arabic-auto-tashkeel>

- **Monolingual Termbase and concordances** concerns the semi-automatic extraction outcome, which is subject for validation by the legal expert.
- **Terminology Translation and Validation;** our approach to translating and/or checking the equivalences of legal terminology in AR→EN / AR→FR pairs involves 4 critical steps. Firstly, we develop a NooJ grammar containing all the terms in a node and apply it to the corpus (see *Figure 6 & 7*). This operation assists in concordancing the entries' occurrences to capture the context where each term occurs and/or the legal practice it conveys to assist us in the translation. Secondly, for the translation, we use MT systems along with some thesauri like WordReference<sup>23</sup>, Lexicolog<sup>24</sup>, and Almaany<sup>25</sup>. Thirdly, to pre-ensure the equivalence in the target languages, we double-check the results through online gateways to EU laws including EUR-Lex, IATE, and

After	Seq.	Before
بالنسب في المجلس الشرعي المعقود	إقرار	من زوجتي الشرعية سلمى على
الزوجية الأولاد زياد و سارة	فراش	الأولاد زياد و سارة و
تسجيل هذا الإقرار. وعليه وحيث	طلب	
الدستور	تنفيذ	
ظهير شريف رقم 1.11.91 صادر في 27 من	ظهير	تنفيذ الدستور
التنظيمي رقم 29.93 المتعلق بالمجلس الد	القانون	مشروع الدستور على الاستفتاء؛ وعلى
لاختيارها الذي لا رجعة فيه	وفاء	الدستور تصدير إن المملكة المغربية
ديمقراطية يسودها الحق والقانون، تواصل	دولة	لا رجعة فيه، في بناء
يسودها الحق والقانون، تواصل بعزم	دولة ديمقراطية	لا رجعة فيه، في بناء
يسودها الحق والقانون، تواصل بعزم	ديمقراطية	رجعة فيه، في بناء دولة
مجتمع متضامن، يتمتع فيه الجميع	دعائم	شراكة والتعددية والحكمة الجيدة، وإرساء
وواجبات المواطنة. المملكة المغربية دولة	حقوق	الكرام، في نطاق التلازم بين
كاملة، متشعبة بوحدتها الوطنية والترابية	سيادة	المملكة المغربية دولة إسلامية ذات
هويتها الوطنية، الموحدة بانصهار كل	مقومات	الوطنية والترابية، وبصيانة تلاحم وتنوع
المغربية تتميز بتبوأ الدين الإسلامي	الهوية	والأندلسية والعبرية والمتوسطية. كما أن
الإسلامي مكانة الصدارة فيها، وذلك	الدين	أن الهوية المغربية تتميز بتبوأ
والمبادئ الثابتة، وعلى إرادتها القوية	القيم	في العالم. وتأسيسا على هذه
والقانون الدولي الإنساني والنهوض بهما	حقوق الإنسان	التعاون جنوب - جنوب؛ حماية منظومتني
ومكافحة كل أشكال التمييز، بسبب	حظر	لتنك الحقوق، وعدم قابليتها للتحزبي؛
بسبب الجنس أو اللون أو	التمييز	للتحزبي؛ حظر ومكافحة كل أشكال
التمييز		

Figure 7: Concordance Sequences for Capturing the Context.

<sup>23</sup> <https://www.wordreference.com/aren/%D9%82%D8%A7%D9%85%D9%88%D8%B3>

<sup>24</sup> [https://www.lexilogos.com/english/arabic\\_dictionary.htm](https://www.lexilogos.com/english/arabic_dictionary.htm)

<sup>25</sup> <https://www.almaany.com/>

Table 3: Extract from Out-of-Context Reference Dataset.

Out-of-Context Reference Translations			
Term ID	Arabic	English	French
1	صداق	Dowry	Dote
2	متعة	Compensation	Compensation
3	مأذون	Superintendent Registrar	Maire
4	القطع	Certainty	Certitude
5	الطعن	Rebuttal	Refutation
1015	الفراش	Marriage Consummation	Consummation du Mariage

Juremy, the intelligent online concordance search tool in EU law data available in all the 24 EU languages. Following the completion of the data preprocessing, the fourth and last crucial stage is the validation process. In this phase, the translations of the entries and the concordances are subject to validation by our legal expert, who also possesses translation skills. This hybrid approach involves ensuring the accuracy and consistency of the data from AR→EN / AR→FR. The outcome of this process is a comprehensive and validated reference dataset ready for experiment-use see an extract of our data. An extract of our data can be seen in Table 3 & 4.

After	Seq.	Before
بالنسب في المجلس الشرعي المعقود	إقرار	من زوجتي الشرعية سلمى على
الزوجية الأولاد زياد و سارة	فراش	الأولاد زياد و سارة و
تسجيل هذا الإقرار. و عليه وحيث	طلب	
الدستور	تنفيذ	
ظهير شريف رقم ا	ظهير	تنفيذ الدستور
شريف رقم 1.11.91 صادر في 27 من	القانون	مشروع الدستور على الاستفتاء؛ و على
التنظيمي رقم 29.93 المتعلق بالمجلس الد	وفاء	الدستور تصدير إن المملكة المغربية
لاختيارها الذي لا رجعة فيه	دولة	لا رجعة فيه، في بناء
ديمقراطية يسودها الحق والقانون، تواصل	دولة ديمقراطية	لا رجعة فيه، في بناء
يسودها الحق والقانون، تواصل بعزم	ديمقراطية	رجعة فيه، في بناء دولة
يسودها الحق والقانون، تواصل بعزم	دعائم	شاركة والتعددية والحكامة الجيدة، وإرساء
مجتمع متضامن، يتمتع فيه الجميع	حقوق	الكريم، في نطاق التلازم بين
وواجبات المواطنة. المملكة المغربية دولة	سيادة	المملكة المغربية دولة إسلامية ذات
كاملة، متشعبة بوحدتها الوطنية والترايبية	مقومات	الوطنية والترايبية، وبصيانة تلاحم وتنوع
هويتها الوطنية، الموحدة بانصهار كل	الهوية	والأندلسية والعبيرية والمتوسطية. كما أن
المغربية تتميز بتبني الدين الإسلامي	الدين	أن الهوية المغربية تتميز بتبني
الإسلامي مكانة الصدارة فيها، وذلك	القيم	في العالم. وتأسيسا على هذه
والمبادئ الثابتة، وعلى إرادتها القوية	حقوق الإنسان	التعاون جنوب - جنوب؛ حماية منظومتي
والقانون الدولي الإنساني والنهوض بهما	حظر	لتنكس الحقوق، وعدم قابليتها للتجزئ؛
ومكافحة كل أشكال التمييز، بسبب	التمييز	للتجزئ؛ حظر ومكافحة كل أشكال
بسبب الجنس أو اللون أو		

Figure 8: Concordance Sequences for Capturing the Context.

Table 4: Extract from In-Context Reference Dataset.

In-Context Reference Translations			
Sentence ID	Arabic	English	French
1	تَسْتَحِقُّ الْمُطَلَّقةَ قَبْلَ الْبِنَاءِ نِصْفَ الصَّدَاقِ الْمُسَمَّى لَهَا بِمُقْتَضَى عَقْدِ الزَّوْاجِ	The divorced woman deserves half of her dowry, according to the marriage contract terms and conditions, only before the marriage consummation.	La femme divorcée n'a droit à la moitié de sa dot, conformément aux termes du contrat de mariage, qu'avant la consommation du mariage.
2	تشمل مستحقات الزوجة: الصداق المؤخر إن وجد، ونفقة العدة، والمتعة التي يراعى في تقديرها فترة الزواج والوضعية المالية للزوج.	The wife's entitlements include: the rest of the dowry, if any, the waiting period expenses, and the compensation that is estimated according to the marriage duration and the financial situation of the husband.	Les droits de la femme comprennent : le rest de la dot, s'il y en a, les dépenses de la période d'attente, et l'indemnité qui est estimée en fonction de la durée du mariage et de la situation financière du mari.
3	يعتمد المأنون في معرفة بلوغ أحد الزوجين السن القانونية على شهادة الميلاد أو أي مستند رسمي آخر	To confirm that a person has reached the legal age of getting married, the superintendent registrar relies on the birth certificate or on any other official proof.	Pour savoir si quelqu'un atteint l'âge légal pour se marier, le maire se base sur l'acte de naissance ou sur toute autre preuve officielle.
4	يعتبر الفراش بشروطه حجة قاطعة على ثبوت النسب، لا يمكن الطعن فيه إلا من الزوج عن طريق اللعان، أو بواسطة خبرة تفيد القطع.	Marriage consummation is considered a strong proof of paternity, it can be rebutted only by the husband through accusation or through certain evidence.	La consommation du mariage est considérée comme une preuve solide signifiant la paternité, il ne peut être réfutée que par le mari soit à travers l'accusation ou bien une certaine preuve.
1015	تسكن الزوجة خلال العدة في بيت الزوجية، أو للضرورة في مسكن ملائم لها وللوضعية المادية للزوج، وإذا تعذر ذلك حددت المحكمة تكاليف السكن في مبلغ يودع كذلك ضمن المستحقات بكتابة ضبط المحكمة.	During the waiting period, the wife will live in the marital home, or if necessary, in a residence suitable for her and the husband's financial situation. If this is not possible, the court will determine the accommodation costs which will also be deposited within the dues in the Court Clerk's Office.	La femme doit résider dans le domicile conjugal pendant la période de l'attente, ou, en cas de nécessité, dans un logement approprié à sa situation et à celle financière du mari. Si cela n'est pas possible, le tribunal fixera les frais de logement, qui seront déposés au greffe du tribunal en tant que partie des droits.

## 4. Transformer Models as Legal Translators of Arabic Legislation

### 4.1. Introduction

GPT models, Gemini and other NMT systems are all based on the transformer architecture (Vaswani et al., 2017), but they differ in several aspects. For example, NMT models are specifically designed for translation tasks, using an encoder-decoder architecture with attention mechanisms to process and translate text by converting the source language into a continuous representation and generating the target language text from this. In contrast, GPT is a large, versatile language model pre-trained on diverse datasets and fine-tuned for various tasks, including translation, by predicting the next word in a sequence while considering the context of the entire input. Gemini, on the other hand, represents an advanced AI model by Google DeepMind, which combines the strengths of both NMT and LLMs, using deep learning techniques and extensive pre-training to generate translations. Indeed, this progress has prioritised model scalability, efficiency, and human-like understanding and translation capabilities. However, their potential in domain-specific MT remains under-explored (Zakraoui et al., 2021; Moslem et al., 2024). Consider example 1 from the Moroccan family code, taking the terms { مستحقات, الصداق المؤخر, المتعة, العدة } into consideration:

AR→EN

1. **Arabic:** تشمل مستحقات الزوجة : الصداق المؤخر إن وجد، ونفقة العدة، والمتعة التي يراعى في تقديرها فترة الزواج والوضعية المالية للزوج.

**English (ChatGPT-4):** the wife's **dues** include: the **deferred dowry**, if any, the maintenance of the **period of waiting**, and the **pleasure** that is taken into account in its estimation of the period of marriage and the financial status of the husband.

**English (Gemini 1.5 Pro):** The wife's **entitlements** include: **delayed dowry** if applicable, post-marriage support, and **temporary marriage**, which takes into account the duration of the marriage, the husband's financial situation.

**English (GT):** The wife's **dues** include: **deferred dowry**, if any, **maintenance** for the waiting period, and **entertainment**, the estimation of which takes into account the period of marriage and the husband's financial situation.

**Human Translation (EN):** The wife's **entitlements** include: the **rest of the dowry**, if any, the **waiting period** expenses, and the **compensation** that is estimated according to the marriage duration and the financial situation of the husband.

## AR→FR

**French (ChatGPT-4):** Les **droits** de l'épouse comprennent : **la dot différée** si elle existe, **la pension alimentaire** pendant la **période d'attente**, et la **jouissance** évaluée en fonction de la durée du mariage et de la situation financière du mari.

**French (Gemini 1.5 Pro):** Les **droits** de l'épouse comprennent: **la dot différée** si elle existe, la **pension alimentaire** pendant la période d'attente, et une **indemnité** dont le montant est évalué en fonction de la durée du mariage et de la situation financière du mari.

**French (GT):** Les **droits** de l'épouse comprennent : **une dot différée**, le cas échéant, une pension alimentaire pendant la période d'attente et des **divertissements** dont l'estimation tient compte de la durée du mariage et de la situation financière du mari.

**Human Translation (FR):** Les droit de la femme comprennent: **le rest de la dot**, s'il y en a, les **dépenses** de la période d'attente, et **l'indemnité** qui est estimée en fonction de la durée du mariage et de la situation financière du mari.

The bold terms in example 1 are specific to the domain and dependent on the context, so their translation requires considering their context, i.e., the cultural, lexical, morphological, and semantic properties. Additionally, their equivalents across languages and legal systems (i.e., English and French) must be taken into account, as the Human Translation (HT) does. All models, in fact, produce results with high levels of multidimensional critical errors. This example highlights the fact that ChatGPT-4 and Gemini 1.5 Pro can indeed capture general language patterns, generate coherent responses, and perform excellent MT tasks, while GT tries to leverage neural networks to understand and produce fluent translations. Nonetheless, they all struggle with translating the given terminology due to their limited domain-knowledge and limited training data. As mentioned in section 2 in chapter two, many authors, namely Hendy et al. (2023) indicate that GPT models perform competitively for high-resource languages but show limited capabilities for low-resource languages. They also show that developing metrics that can measure the contextual correctness of LLM-generated translations is essential, and that further research can explore new ways to evaluate the quality of MT more accurately, especially when using LLMs. Therefore, in this chapter, we present a preliminary and comparative evaluation between GT, ChatGPT-4, and Gemini 1.5 Pro as legal translators of Arabic legislation to test their accuracy and the extent to which they care for context and terminology across two language pairs AR→EN / AR→FR. The study targets the evaluation of Zero-Shot

prompting for in-context and out-of-context scenarios relying on a gold standard dataset, verified by professional translators who are also experts in the field. We evaluate the results applying the Multidimensional Quality Metrics (MQM) (Burchardt, 2013) to classify translation errors, compare the models' abilities, and suggest directions for future work.

#### 4.2. Experimental Set-up and Methodology

In this fundamental section, that is a gate to the upcoming gap fillings in this work, we present a comprehensive evaluation of GT, ChatGPT-4, and Gemini 1.5 Pro performance in the translation of Arabic legislation. We aim at investigating whether these models care for context and terminology in two language pairs AR→EN / AR→FR based on our gold standard dataset, which to the best of our knowledge is the first formalised legal terminological resource for Arabic developed specifically for assessing the quality and accuracy of machine-translated Arabic legal terms (see *Figure 4*).

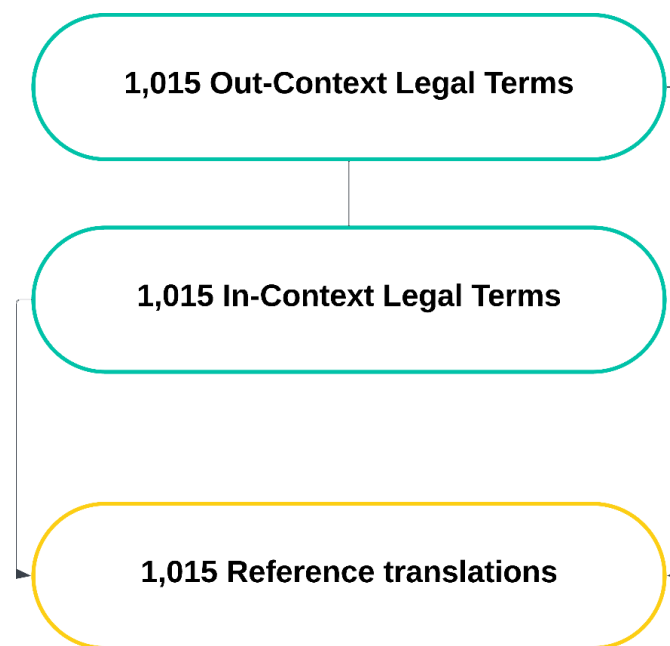


Figure 9: Our Gold Standard Dataset in the Legal Domain for Arabic.

**Prompting.** To start with the ChatGPT-4 and Gemini 1.5 Pro assessment, we employ a Python script that iterates through our data to generate prompts. It implements a series of prompts for Out-of-Context and In-Context translation evaluation in two language pairs AR→EN / AR→FR.

- **Prompt 1** = {Write the corresponding English/French term for "{Arabic legal term}". Write only the English/French

translation without providing additional context or comments.' 'Write the translations following the order provided.}

- **Prompt2** = {Write the corresponding English/French sentence for "{Arabic legal sentence}". Write only the English/French translation without providing comments.' 'Write the translations following the order provided.}

Prompts 1 and 2 generate from the LLMs the translation for Out-of-Context and In-Context legal terms and sentences into English and then into French, each separately respecting the order provided. Prompt 1 is instructed to provide translations without providing additional context or comments and Prompt 2 is directed to provide no additional comment. For GT, we use the webApp download feature, where we upload our files containing the source language data for both scenarios each separately. The generated output, subject for evaluation, is refined and classified as follow:

- Out-of-Context and In-Context translation results of ChatGPT-4 from AR→EN / AR→FR,
- Out-of-Context and In-Context translation results of Gemini 1.5 Pro from AR→EN / AR→FR.
- Out-of-Context and In-Context translation results of GT from AR→EN / AR→FR.

**Methodology.** Our evaluation consists of three fundamental phases. The first is to assess the obtained results of all the models, ChatGPT-4, Gemini 1.5 Pro and GT based on the accuracy criterion against our gold dataset. The second is the evaluation of the results applying the MQM (Burchardt, 2013), which is a framework for analytic TQE. The third consists of evaluating the general LLMs outputs to verify their correctness, consistency, and completeness.

*Table 5* presents machine-translated legal terms results of the 3 models from AR→EN / AR→FR based on accuracy criterion. In the out-of-context evaluation, where the models have no prior knowledge or context about the legal terms being translated, GT yields lower false percentage than ChatGPT-4 and Gemini 1.5 Pro from AR→EN. Whereas, in AR→FR pair all the three models approximately score equal false counts. Notably, all the models demonstrate a consistent trend of higher false counts than true counts across both language pairs, highlighting the challenge of accurately translating legal terminology without contextual cues. While GT generally exhibits slightly fewer false results than ChatGPT-4 and Gemini 1.5 Pro

from AR→EN, ChatGPT-4 and Gemini 1.5 Pro showcase marginally better performance in Arabic into French, indicating nuanced differences in its capabilities across different language pairs.

Table 5: Comparison of ChatGPT-4, Gemini 1.5 Pro, and GT MT of Legal Terms from AR→EN / AR→FR in Out-of-Context and In-Context Scenarios, Based on Accuracy Criterion.

	<b>Out-of-Context AR→EN</b>	<b>In-Context AR→EN</b>	<b>Out-of-Context AR→FR</b>	<b>In-Context AR→FR</b>
	<b>Google Translate</b>			
<b>False</b>	43%	45%	58%	55%
	<b>ChatGPT-4</b>			
<b>False</b>	54%	41%	60%	59%
	<b>Gemini 1.5 Pro</b>			
<b>False</b>	57%	55%	61%	56%

However, in the In-context scenario, all the three models demonstrate significant challenges, with GT 45% in AR→EN and 55% AR→FR false count, ChatGPT-4 exhibits 41% false results in AR→EN, and 59% false results in AR→FR pair. Gemini 1.5 Pro performs shows 55% false results in AR→EN, and 56% false results in AR→FR. While the accuracy measure is important, these first findings do not, however, capture the nuances of the translation quality beyond word-for-word accuracy i.e., the Arabic linguistic structures and cultural nuances Arabic legal terminology denotes. They also do not provide an accountable decision on the overall quality of the models.

Therefore, we apply the MQM<sup>26</sup> (Burchardt, 2013) framework. It can be used to evaluate MT, HT, and AI-generated translation (Monti & Nolano, 2019). It leverages a set of error types and a scoring model. Error types are organised in a hierarchical system<sup>27</sup> under seven high-level core error dimensions, together with subordinate error types and their associated severity levels (neutral, minor, major, and critical). The scoring model<sup>28</sup> features a system of weights and parameters assigned to the error types and severity levels, as well as a scoring formula used

<sup>26</sup> Full details here: <https://themqm.org/>, consider also the paper in reference section number 4 & 5.

<sup>27</sup> Available here: <https://themqm.org/error-types-2/typology/>

<sup>28</sup> Detailed process available here: <https://themqm.org/error-types-2/detailed-process/>

to calculate a numerical score that represents the quality of the evaluated translation i.e., Absolute Penalty Total according to agreed-upon specifications and the Error Counts. To achieve the MQM scoring stage, it is fundamental to first annotate the errors applying the MQM error typology framework. *Table 6, 7 & 8* detail per each model i) six high level dimensions namely terminology, style, linguistic convention, accuracy and audience appropriateness, ii) the errors subtypes iii) and the number per error type for the two language pairs AR→EN / AR→FR in Out-of-Context and In-Context scenarios.

Table 6: Errors Count and Classification of Machine-translated Legal Terms by Google Translate for AR→EN / AR→FR in Out-of-Context and In-Context Scenarios.

Error Types	Error sub-types	Google Translate			
		Out-of-context AR→EN	In-context AR→EN	Out-of-context AR→FR	In-of-Context AR→FR
Terminology	Inconsistent with terminology resource	980	1011	707	999
	Inconsistent use of terminology				
	Multiple terms for concept in source (multiple terms for concept)				
	Wrong term				
Style	Language register	1008	1013	1000	992
	Awkward style				
Linguistic convention	Duplication	596	655	612	521
	Grammar				
	Word form				
	Agreement				
	Word order				
	Function words				
	Grammatical register				
	Transliteration				
Coherence					
Accuracy	Undertranslation	1002	1008	999	1003
	Overtranslation				
	Addition				
	Omission				
	Untranslated				

	<b>Incomplete procedure</b>				
	<b>Mistranslation</b>				
	<b>Ambiguous target content</b>				
	<b>False friend</b>				
	<b>Completeness</b>				
	<b>Incomplete List (incomplete list)</b>				
	<b>Incomplete procedure (incomplete procedure)</b>				
<b>Audience appropriateness</b>	<b>Culture-specific reference</b>	1000	1015	1000	1015
	<b>Locale-specific content</b>				
	<b>Legal requirements</b>				
	<b>End-user suitability</b>				

Table 7: Errors Count and Classification of Machine-translated Legal Terms by ChatGPT-4 for AR→EN / AR→FR in Out-of-Context and In-Context Scenarios.

Error Types	Error sub-types	ChatGPT-4			
		Out-of-Context	In-of-Context	Out-of-Context	In-of-Context
		AR→EN	AR→EN	AR→FR	AR→FR
<b>Terminology</b>	<b>Inconsistent with terminology resource</b>	1000	1006	998	1001
	<b>Inconsistent use of terminology</b>				
	<b>Multiple terms for concept in source (multiple terms for concept)</b>				
	<b>Wrong term</b>				
<b>Style</b>	<b>Language register</b>	1008	1005	768	1000
	<b>Awkward style</b>				
<b>Linguistic convention</b>	<b>Duplication</b>	645	760	765	864
	<b>Grammar</b>				
	<b>Word form</b>				
	<b>Agreement</b>				
	<b>Word order</b>				
	<b>Function words</b>				
	<b>Grammatical register</b>				
	<b>Transliteration</b>				
<b>Coherence</b>					

<b>Accuracy</b>	<b>Undertranslation</b>	986	999	921	1001
	<b>Overtranslation</b>				
	<b>Addition</b>				
	<b>Omission</b>				
	<b>Untranslated</b>				
	<b>Incomplete procedure</b>				
	<b>Mistranslation</b>				
	<b>Ambiguous target content</b>				
	<b>False friend</b>				
	<b>Completeness</b>				
	<b>Incomplete List (incomplete-list)</b>				
	<b>Incomplete procedure (incomplete-procedure)</b>				
<b>Audience appropriateness</b>	<b>Culture-specific reference</b>	1002	1010	1002	1002
	<b>Locale-specific content</b>				
	<b>Legal requirements</b>				
	<b>End-user suitability</b>				

Table 8: Errors Count and Classification of Machine-translated Legal Terms by Gemini 1.5 Pro for AR→EN / AR→FR in Out-of-Context and In-Context Scenarios.

<b>Error Types</b>	<b>Error sub-types</b>	<b>Gemini 1.5 Pro</b>			
		<b>Out-of-context AR→EN</b>	<b>In-context AR→EN</b>	<b>Out-of-context AR→FR</b>	<b>In-of-Context AR→FR</b>
<b>Terminology</b>	<b>Inconsistent with terminology resource</b>	998	1009	862	1000
	<b>Inconsistent use of terminology</b>				
	<b>Multiple terms for concept in source (multiple terms for concept)</b>				
	<b>Wrong term</b>				
<b>Style</b>	<b>Language register</b>	980	1003	950	989
	<b>Awkward style</b>				
<b>Linguistic convention</b>	<b>Duplication</b>	500	800	750	987
	<b>Grammar</b>				

	<b>Word form</b>				
	<b>Agreement</b>				
	<b>Word order</b>				
	<b>Function words</b>				
	<b>Grammatical register</b>				
	<b>Transliteration</b>				
	<b>Coherence</b>				
<b>Accuracy</b>	<b>Undertranslation</b>	909	1005	901	999
	<b>Overtranslation</b>				
	<b>Addition</b>				
	<b>Omission</b>				
	<b>Untranslated</b>				
	<b>Incomplete procedure</b>				
	<b>Mistranslation</b>				
	<b>Ambiguous target content</b>				
	<b>False friend</b>				
	<b>Completeness</b>				
	<b>Incomplete List (incomplete-list)</b>				
	<b>Incomplete procedure (incomplete-procedure)</b>				
<b>Audience appropriateness</b>	<b>Culture-specific reference</b>	1006	1011	963	1001
	<b>Locale-specific content</b>				
	<b>Legal requirements</b>				
	<b>End-user suitability</b>				

Subsequently, to be more precise towards the evaluation and the final judgement of the MT results produced by the models, we calculate the final quality using the MQM scoring model. The MQM scoring model<sup>29</sup> allows to convert error type and severity level counts into numeric scores that can be used to determine whether the content meets specifications by assigning scores to the following:

- **The Error Severity Level.** It is defines depending on the extent to which the error poses a risk to the quality of the translation, it features four Severity Levels: Neutral, Minor, Major, and Critical;

<sup>29</sup> <https://themqm.org/error-types-2/the-mqm-scoring-models/>

- **Neutral Severity Level.** The Severity Level of an error that differs from a quality evaluator's preferential translation or that is flagged for the translator's attention but is an acceptable translation.
- **Minor Severity Level.** The Severity Level of an error that does not seriously impede the usability, understandability, or reliability of the content for its intended purpose, but has a limited impact on, for example, accuracy, stylistic quality, consistency, fluency, clarity, or general appeal of the content.
- **Major Severity Level.** The Severity Level of an error that seriously affects the understandability, reliability, or usability of the content for its intended purpose or hinders the proper use of the product or service due to a significant loss or change in meaning or because the error appears in a highly visible or important part of the content.
- **Critical Severity Level.** The Severity Level of an error that renders the entire content unfit for purpose or poses the risk for serious physical, financial, or reputational harm. In many systems, a single Critical Error would automatically trigger a Fail Rating.
- **The Severity Penalty Multiplier (SPM).** It is the score which reflects the increased risk and impact between the Error Severity Levels (for example, in our case study, we give *1* to neutral, *1* to minor, *5* to major, and *25* to critical),
- **Error Type Weight (ETW).** It reflects the importance of certain error types that should be given more prominence than others. For example, we give *5* ETW score for terminology, accuracy, and audience appropriateness because they represent the highest error rates and only *3* to style and linguistic convention.
- **Error Type Penalty Total (ETPT).** It is the sum of penalty points calculated for the individual error types annotated. The error count for a specific error type and severity level is multiplied by the respective SPM and ETW to obtain the ETPT. For example, the ETPT for terminology in *Table 13* is determined as follows:  $0 \times 1 \times 5 = 0$ ;  $6 \times 1 \times 5 = 30$ ;  $10 \times 5 \times 5 = 250$ ;  $979 \times 25 \times 5 = 122,375$  then,  $(0 + 30 + 250 + 122,375) \times 5 = 613,275$  therefore, *613,275* is the ETPT for terminology errors in the Out-of-Context from AR→EN scored by ChatGPT-4.
- **Errors Count (EC).** It is the number of errors set for each dimension and its types which we classify according to the level of severity across the error severity level.

- **Absolute Penalty Total (APT).** It is considered the most important value used for quality score calculation, and the one that we consider deciding over the translation quality and compare the models' quality and performance on MT of legal terms in and out of context from AR→EN / AR→FR. The formula to calculate the APT is as follows:

$$\sum_{ij} \text{Error Count}_{ij} \times \text{Severity Multiplier}_j \times \text{Error Type Weight}_i$$

**Where:  $i$  = index for Error Types,  $j$  = index for Severity Level.**

The evaluation comparing MT application of GT, ChatGPT-4 and Gemini 1.5 Pro in the AR→EN / AR→FR language pairs using the MQM scoring model provides valuable insights into their performance, both in and out of context (see from *Table 9* to

*Table 20*), and what follows is a detailed discussion about the evaluation and the results comparing the three systems.

#### 4.3. Results and Evaluation

The evaluation comparing MT results of GT, ChatGPT-4 and Gemini 1.5 Pro in the AR→EN / AR→FR language pairs using the MQM scoring model provides valuable insights into their performance, both in and out of context. Tables 9, 10, & 11 present the scorecards of GT performance in both scenarios and language pairs. For terminology, in Out-of-Context, GT surpasses ChatGPT-4 and Gemini 1.5 Pro in both language pairs with the evidence that GT scores less critical and major results than its counterparts. However, in In-Context scenario both models surpass GT in both language pairs as the critical errors went higher but the difference is slight. For style GT yields the highest error rate ranging from major to critical but very few of them are neutral and minor in In-Context and Out-of-Context scenarios in both language pairs. For the linguistic convention dimension, most of the critical errors has to do with transliteration and coherence in both language pairs and scenarios. GT seems to be inaccurate as it yields a significant number of errors with regards to accuracy. It shows a lot of omissions, mistranslation, and many untranslated statements. It also produces additions and false friends in both language pairs in the In-Context scenario. Audience appropriateness is the dimension that almost led the above errors go rogue especially its sub-types including Culture-specific reference, Local-specific content, Legal requirements, and End-user suitability and this is explained by the error counts mentioned during annotation in Table 6, 7, 8.

Table 9: MQM Evaluation Scorecard of Google Translate Performance in Out-of-Context  
AR→EN.

Google Translate						
Out-of-Context AR→EN						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	15	25	132	808	5	522,500
Style	0	12	222	774	3	184,248
Linguistic Convention	25	38	225	308	3	79,989
Accuracy	0	5	20	977	5	613,250
Audience Appropriateness	0	0	55	945	5	597,500
<b>Absolute Penalty Total</b>						1,997,487

Table 10: MQM Evaluation Scorecard of Google Translate Performance in In-Context  
AR→EN.

Google Translate						
In-Context AR→EN						
Error Severity Multiplier	Neutra 1	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	0	15	19	977	5	613,375
Style	0	20	32	961	3	217,845
Linguistic Convention	33	59	88	475	3	111,663
Accuracy	0	0	68	940	5	596,000
Audience Appropriateness	0	0	15	1,000	5	626,875
<b>Absolute Penalty Total</b>						2,165,758

Table 11: MQM Evaluation Scorecard of Google Translate Performance in In-Context AR→FR.

Google Translate						
In-Context AR→FR						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	15	52	216	716	5	476,175
Style	0	12	102	878	3	202,248
Linguistic Convention	45	65	98	313	3	75,825
Accuracy	0	0	39	964	5	607,375
Audience Appropriateness	0	0	23	992	5	252,250
<b>Absolute Penalty Total</b>						1,613,873

Table 12: MQM Evaluation Scorecard of Google Translate Performance in Out-of-Context AR→FR.

Google Translate						
Out-of-Context AR→FR						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	31	25	31	620	5	392,775
Style	0	9	54	937	3	213,336
Linguistic Convention	35	68	112	403	3	96,642
Accuracy	0	19	21	959	5	602,475
Audience Appropriateness	0	0	22	978	5	614,000
<b>Absolute Penalty Total</b>						1,919,228

In the Out-of-Context scenario for AR→EN translation, ChatGPT-4 exhibited a higher absolute penalty total compared to Gemini 1.5 Pro (see Table 9, 13 & 17). This suggests that ChatGPT-4 struggled more with maintaining accuracy and coherence when translating legal terms without the surrounding context unlike GT’s acceptable performance.

We assign 5 ETW for terminology (ranging from 0 neutral, 6 minor, 10 Major, and 979 critical terms), accuracy (ranging from 0 neutral, 23 minor, 35 Major, and 916 critical terms), and audience appropriateness (ranging from 0 neutral, 23 minor, 11 Major, and 959 critical terms), as ChatGPT-4 shows an inconsistent use of terminology, the use of wrong terms and false friends, untranslated entries, lack of legal requirements, and culture-specific references etc., which all lead to present wrong concepts in the target language with a total number of 1000 errors out of 1015 term. For example, the English translation of the term {متعة} ‘**compensation**’ is left untranslated by ChatGPT-4 and results instead in the following comment ‘**nan (there doesn't seem to be an English translation for this term)**’, ‘البناء’ ‘**marriage consummation**’ is translated as ‘**construction**’, {القطع} ‘**certainty**’ is translated as ‘**severance**’, {مسطرة} ‘**procedure**’ translated as ‘**ruler**’. The high critical severity rate, considering the number, dimensions, and types of errors, underscores the importance of addressing these translation issues, especially in legal contexts where precision is paramount.

Table 13: MQM Evaluation Scorecard of ChatGPT-4 Performance in Out-of-Context AR→EN.

ChatGPT-4						
Out-of-Context AR→EN						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	0	6	10	979	5	613,275
Style	0	15	40	950	3	213,750
Linguistic Convention	0	90	55	430	3	96,750
Accuracy	0	23	35	916	5	572,500
Audience Appropriateness	0	23	11	959	5	601,325
<b>Absolute Penalty Total</b>						2,097.600

Table 14:MQM Evaluation Scorecard of ChatGPT-4 Performance in In -Context AR→EN.

ChatGPT-4						
In-Context AR→EN						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	0	25	60	915	5	580,000
Style	0	16	58	913	3	208,179
Linguistic Convention	0	54	56	650	3	149,256
Accuracy	0	17	52	930	5	587,875
Audience Appropriateness	0	13	55	934	5	590,950
<b>Absolute Penalty Total</b>						2,116.26

Table 15: MQM Evaluation Scorecard of ChatGPT-4 Performance in Out-of-Context AR→FR.

ChatGPT-4						
Out-of-Context AR→FR						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	0	20	65	913	5	579,250
Style	0	55	65	648	3	149,220
Linguistic Convention	0	54	67	644	3	148,806
Accuracy	0	11	23	887	5	557,525
Audience Appropriateness	0	9	22	971	5	609,850
<b>Absolute Penalty Total</b>						2,044.651

Table 16: MQM Evaluation Scorecard of Gemini Performance in Out-of-Context AR→FR.

ChatGPT-4						
In-Context AR→FR						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	0	8	35	958	5	603,325
Style	0	12	33	955	3	216,468
Linguistic Convention	0	11	66	787	3	180,144
Accuracy	0	8	23	970	5	609,325
Audience Appropriateness	0	5	27	970	5	609,750
<b>Absolute Penalty Total</b>						2,219.012

Conversely, in the In-Context scenario (*Table 14 & 18*) for AR→EN both models, despite the slight variations in favour of ChatGPT-4, face a notable failure in performance when contextual information was provided but their APT, however, surpasses GT (*Table 10*). Consider example A from the Moroccan Family Code<sup>30</sup> and B from The Universal Declaration of Human Rights:

#### AR→EN

A **Arabic**: تشمل مستحقات الزوجة: الصداق المؤخر إن وجد، ونفقة العدة، والمتعة التي يراعى في تقديرها فترة الزواج والوضعية المالية للزوج.

**English (GT)**: The wife's **needs** include: the **dowry**, **maintenance** for the waiting period, and **enjoyment**, the estimation of which takes into account the period of marriage and the financial situation of the husband.

**English (ChatGPT-4)**: the wife's **dues** include: the **deferred dower**, if any, the **maintenance** of the **period of waiting**, and the **pleasure** that is taken into account in its estimation of the period of marriage and the financial status of the husband.

<sup>30</sup> <https://shorturl.at/EG567>

**English (Gemini):** The wife's **entitlements** include: **delayed dowry** if applicable, post-marriage support, and **temporary marriage**, which takes into account the duration of the marriage, the husband's financial situation.

**Human Translation:** The wife's **entitlements** include: the **rest of the dowry**, if any, the **waiting period** expenses, and the **compensation** that is estimated according to the marriage duration and the financial situation of the husband.

#### AR→FR

**French (GT):** Les **droits** de l'épouse comprennent : une **dot différée**, le cas échéant, une **pension** alimentaire pendant la **période d'attente** et des divertissements dont l'estimation tient compte de la durée du mariage et de la situation financière du mari.

**French (ChatGPT-4):** Les **droits** de l'épouse comprennent : **la dot différée** s'il y en a, **la pension de l'iddah**, et le **plaisir** qui est estimée en tenant compte de la durée du mariage et de la situation financière du mari.

**French (Gemini 1.5 Pro):** Les **droits** de la femme comprennent, le cas échéant, la **dot différée**, la **pension pendant la 'idda**, et un **plaisir** financière dont le montant est évalué en tenant compte de la durée du mariage et de la situation financière du mari.

**French (HT):** Les **droits** de la femme comprennent : **le rest de la dot**, s'il y en a, **les dépenses de la période d'attente**, et **l'indemnité** qui est estimée en fonction de la durée du mariage et de la situation financière du mari.

#### AR→EN

B **Arabic:** كل شخص متهم بجريمة يعتبر بريئاً إلى أن تثبت إدانته قانوناً بمحاكمة علنية تؤمن له فيها الضمانات الضرورية للدفاع عنه.

**English (GT):** Everyone charged with a penal offence has the right to be presumed innocent until proved guilty according to law in a public trial at which he has had all the guarantees necessary for his defence.

**English (ChatGPT-4):** Every person accused of a crime is considered innocent until proven guilty by law in a public trial that ensures the necessary guarantees for their defense.

**English (Gemin 1.5 Pro):** Everyone charged with a penal offense shall be presumed innocent until proved guilty according to law in a public trial at which he has had all the guarantees necessary for his defence.

**Human Translation (EN):** Everyone being accused of a crime shall be presumed innocent until proven guilty by law in a public trial that ensures the necessary guarantees for their defence.

## AR→FR

**French (GT):** Toute personne accusée d'un crime est considérée comme innocente jusqu'à ce qu'elle soit légalement reconnue coupable lors d'un procès public au cours duquel elle bénéficie des garanties nécessaires à sa défense.

**French (ChatGPT-4):** Toute personne accusée d'un délit est présumée innocente jusqu'à ce que sa culpabilité soit légalement établie au cours d'un procès public où toutes les garanties nécessaires à sa défense lui auront été assurées.

**French (Gemini 1.5 Pro):** Toute personne accusée d'une infraction est présumée innocente jusqu'à ce que sa culpabilité ait été légalement établie par un jugement public au cours duquel lui auront été garanties toutes les garanties nécessaires à sa défense.

**Human Translation (French):** Toute personne accusée d'un acte délictueux est présumée innocente jusqu'à ce que sa culpabilité ait été légalement établie au cours d'un procès public où toutes les garanties nécessaires à sa défense lui auront été assurées.

In example A, we find out that GT completely fails to render all the highlighted terms in their context as from the results provided, we find some are false friends, and part of a multiword term is omitted. The result also shows the use of awkward style and does not respect the legalese i.e., the formal and specialised style of language used in law. Whereas ChatGPT-4 and Gemini 1.5 Pro only translate part of the highlighted terms. However, the models' translations also demonstrate several pitfalls including the inconsistent use of terminology, the use of wrong terms, word order that is not compliant with the target language norms. They also fail render how '**compensation**' is estimated which is a critical and important information, part of acknowledging the husband duties towards the wife after divorce. The models show lack of legal requirements and undertranslate or mistranslate legal terms in a legal system where religion, culture, and law meet and mutually influence each other. The critical severity rate and the EC being high in GT, while ranging from neutral to major in ChatGPT-4 and Gemini 1.5 Pro indicate persistent translation issues that need attention. If we consider the English results in example B compared to the ones in example A, we find that they are quite close from being perfect because the terms and concepts found therein are simple and common thus easily translatable. In this regard, Alkhawaja et al. 2020; Moslem et al. 2023; Moselm et al. 2024; Alghamdi et al. 2023 state that LLMs MT applications can translate terms or sentences governed by law, religion, and culture, but with varying degrees of accuracy and nuance. While these technologies are fine-tuned to translate texts based on extensive data, they often struggle with context-specific subtleties found in legal, religious, and cultural texts. These areas are deeply

rooted in tradition, interpretation, and specific societal norms, which can lead to translations that may be technically correct but miss the intended meaning, or worse, convey a message that is culturally insensitive or legally inaccurate. Therefore, while advanced MT systems can provide translations, human expertise is often required to ensure that these translations are accurate, appropriate, and culturally sensitive.

In the In-Context AR→FR pair, ChatGPT-4 and Gemini 1.5 Pro demonstrate a high critical number of errors than the major and minor ones compared to the Out-of-Context with 0 neutral errors, and low rates of minor ones, but high rates from major to critical (see *Table 15, 19*). Whereas GT (*Table 10*) performance is also high in terms of major and critical severity multipliers but has got more neutral and minor severity multipliers than ChatGPT-4 and Gemini 1.5 Pro especially in terminology and linguistic convention error dimensions. In Out-of-Context, ChatGPT-4 and Gemini 1.5 Pro (see *Table 14 & 20*) score a higher number of minor and major errors than the critical ones compared to GT (see *Table 11*). In other words, ChatGPT-4 approximately generates around 28% and Gemini 44% of correct terms in Out-of-Context, and only 7% (ChatGPT-4) and 15% (Gemini) of correct phrases in In-Context Scenario with, however, higher ETW score in both scenarios for both models. This implies that the two models struggle with leveraging and disambiguating legal phrases that solely rely on contextual and cultural knowledge, leading to inaccuracy, transliterations, non-compliant results with legal requirements, and violation of culture-specific references, etc.

Table 17: MQM Evaluation Scorecard of Gemini Performance in Out-of-Context AR→EN.

Gemini 1.5 Pro						
Out-of-Context AR→EN						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	0	2	23	973	5	611,050
Style	0	11	33	936	3	212,184
Linguistic Convention	0	10	44	446	3	102,420
Accuracy	0	5	25	879	5	552,625
Audience Appropriateness	0	14	21	971	5	609,850
<b>Absolute Penalty Total</b>						2,088.129

Table 18:MQM Evaluation Scorecard of Gemini performance in In-Context AR→EN.

Gemini 1.5 Pro						
In-Context AR→EN						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	0	2	55	941	5	595,050
Style	0	5	13	962	3	217,080
Linguistic Convention	0	2	11	487	3	110,088
Accuracy	0	8	54	847	5	536,325
Audience Appropriateness	0	6	55	945	5	597,650
<b>Absolute Penalty Total</b>						2,056.193

Table 19:MQM Evaluation Scorecard of Gemini Performance in Out-of-Context AR→FR.

Gemini 1.5 Pro						
Out-of-Context AR→FR						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	0	2	55	805	5	510,050
Style	0	12	21	917	3	207,348
Linguistic Convention	0	15	32	703	3	159,750
Accuracy	0	11	39	851	5	537,025
Audience Appropriateness	0	12	42	909	5	573,675
<b>Absolute Penalty Total</b>						1,987.848

Table 20: MQM Evaluation Scorecard of Gemini Performance in Out-of-Context AR→FR.

Gemini 1.5 Pro						
In-Context AR→FR						
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	Error Type Weight	Error Type Penalty Totals
Error Types Dimension	Error Count					
Terminology	0	3	30	967	5	608,200
Style	0	9	22	958	3	216,621
Linguistic Convention	0	5	32	950	3	215,235
Accuracy	0	2	15	982	5	615,675
Audience Appropriateness	0	3	44	954	5	120,425
<b>Absolute Penalty Total</b>						1,776.156

Our terminological dataset consists of judicial documents (i.e., contracts, provisions, codes, decrees, etc.) of different Arab countries namely Morocco, Algeria, Tunisia, United Arab Emirates, Saudi Arabia, Egypt. Most of them use distinct legal terminology to convey similar legal practices. This significantly affects the precision required in legal language, as GT, ChatGPT-4, and Gemini may struggle to accurately capture the intended meaning of Arabic legal terms in both scenarios. As a result, this can lead to mistranslations, misinterpretations, and errors that may have serious legal consequences. For example, the term {مأذون} is used mostly in Qatar, Egypt, and Saudi Arabia. It is used to refer to the person certified by the judge to perform certain legal formalities, especially to draw up or certify marriage contracts, deeds, and other documents for use in other jurisdictions<sup>31</sup>. GT translates it as ‘**licensed**’ into English which could mean someone who has a driving licence or retailer, and in French as ‘**autorisé**’. On the other hand, ChatGPT-4 translates it as ‘authorized’ into English and ‘**autorisé**’ into French. Whereas Gemini 1.5 Pro, as well, translates it as ‘authorized’ into English and ‘**autorisé**’ into French. Therefore, we notice that all models not only transform the grammatical category of the term from a noun,

<sup>31</sup> <https://www.almaany.com/ar/dict/ar-ar/>

which represents a person into an adjective, but they also misinterpret the intended legal practice in the target legal systems, and went far beyond to giving general term thus creates ambiguity and mistranslation. Hence, in France, the equivalence of {مأذون} is ‘**maire**’ (i.e., the person who chairs the municipal council; according to Local and Regional Governments in Europe: Types of Institutions, bodies and agencies<sup>32</sup>). It is the one who oversees approving and drawing up marriage contracts. In England the person in charge of approving and celebrating the marriage requests is called the ‘**superintendent registrar**<sup>33</sup>’ of the district. This unveils that these models are not trained on a diverse and comprehensive datasets that covers a wide range of legal terminologies from different countries. In other words, these models need to be equipped with region-specific legal dictionaries and context-aware algorithms that consider the nuances of each country's legal language. Additionally, leveraging parallel legal texts in different terms can help train the models to better handle these variations.

#### 4.4. Conclusion

The results show that GT surpasses ChatGPT-4 and Gemini 1.5 Pro in achieving less critical errors in Out-Of-Context scenario in both language pairs, conversely, in In-Context scenario, ChatGPT-4 and Gemini 1.5 Pro demonstrate less critical results, but more minor and major errors compared to GT with only slight differences regarding the critical error severity count in both language pairs. This evaluation reveals significant shortcomings, particularly with GT, which fails to accurately render key terms within their legal context, often omitting crucial parts of multiword terms and producing awkward, non-legal style translations. ChatGPT-4 and Gemini 1.5 Pro perform better in translating some highlighted terms, but they also exhibit issues such as inconsistent terminology, incorrect word order, and misinterpretation of critical concepts like ‘**compensation**’ as in example A which is essential in the legal context of divorce. These models struggle to meet legal requirements and often produce translations that are inaccurate or culturally insensitive, especially in contexts where law, religion, and culture intersect. The critical severity and error classification ratings further underscore these persistent translation issues. Conversely, in example B where the terms and concepts are simpler and more universal, the models perform much better, demonstrating that while LLMs and MT tools can handle general translations, they often fall short in specialized, context-sensitive areas. This

---

<sup>32</sup>EESC/COR-FR, d'après le Conseil des communes et régions d'Europe (CCRE), «[Gouvernements locaux et régionaux en Europe — Structures et compétences](#)» (2016) (3.5.2022), page 26

<sup>33</sup> Term reference: <https://www.citizensadvice.org.uk/family/living-together/marriage-and-civil-partnership/getting-married/>

underscores the need for human expertise to ensure translations are accurate, culturally appropriate, and legally sound.

Furthermore, these models are pre-trained on vast amount of text from diverse sources and on a large amount of unlabelled corpora, and they might not have access to sufficient amounts of domain-specific texts for every possible domain. In this regard, we believe that efforts should be invested on enhancing context sensitivity of LLMs' MT tasks by i) developing models capable of handling longer text sequences, such as Transformer-XL (Dai et al., 2019), which can capture broader dependencies, ii) incorporating hierarchical modelling to process text at multiple levels (sentence, paragraph, document) to ensure better context continuity, iii) training with datasets that preserve discourse structure and using dynamic context integration techniques, like Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) which combines generative models with retrieval mechanisms to dynamically incorporate relevant context from external sources during the generation process, vi) leverage prompt engineering strategies to optimise MT output stimuli of these models. Indeed, MQM offers an insightful and comprehensive approach to evaluating machine-translated legal terms for Arabic by GT, ChatGPT-4, and Gemini 1.5 Pro. This framework enables a granular assessment of the accuracy state of the models better than any automatic metric would do, by considering various dimensions such as terminology, style, linguistic convention, accuracy, and audience appropriateness, and their respective sub-dimensions. This detailed analysis provides us with actionable feedback for improvement as it presents detailed insights that go beyond automatic metrics.

## **5. Towards a Linguistic Annotation of Arabic Legal Texts: a Multilingual Electronic Dictionary for Arabic 'ULTRA'.**

### 5.1. Introduction

The challenges identified in the previous chapter regarding the translation of Arabic legislation by MT systems namely GT, ChatGPT-4, and Gemini 1.5 Pro, reveal numerous limitations. The general-purpose nature of the dataset used to train these models makes them insufficient for responding to the accuracy criteria required in the translation of Arabic legislation. This underscores the need for datasets that capture the unique nuances of Arabic legalese and optimise these systems performance accordingly.

This section, therefore, focuses on the development of a multilingual AR→EN / AR→FR legal dictionary able to solve context-dependent issues, automatizing the process of annotating Arabic legal texts, and obtaining the automatic translation of technical legal terms from AR→EN / AR→FR. To do this, we use NooJ that gives the possibility of processing different natural languages and their related information including morpho-syntactic, and semantic information. The proposed resource will serve as a tool for bridging the gap between MT capabilities and the complex demands of the legal language, ensuring that key terms and concepts are correctly interpreted within their appropriate contexts. As pipeline, we proceed with the development of our dictionary using NooJ morpho-syntactic information (*POS, gender, number, etc.*), syntactic information (*transitive, intransitive, Naqis, etc.*), and the proposition of our semantic tags that describe our domain-knowledge terms including legal, Juri-religion, etc., and *geoUsage* to indicate where a given term is adapted to express a legal practice. Finally, we incorporate the validated translations (see the process in Section 2 in Chapter Three: Research Design, Methodology, and Experimental Set-up).

## 5.2. Experimental Set-up and Methodology

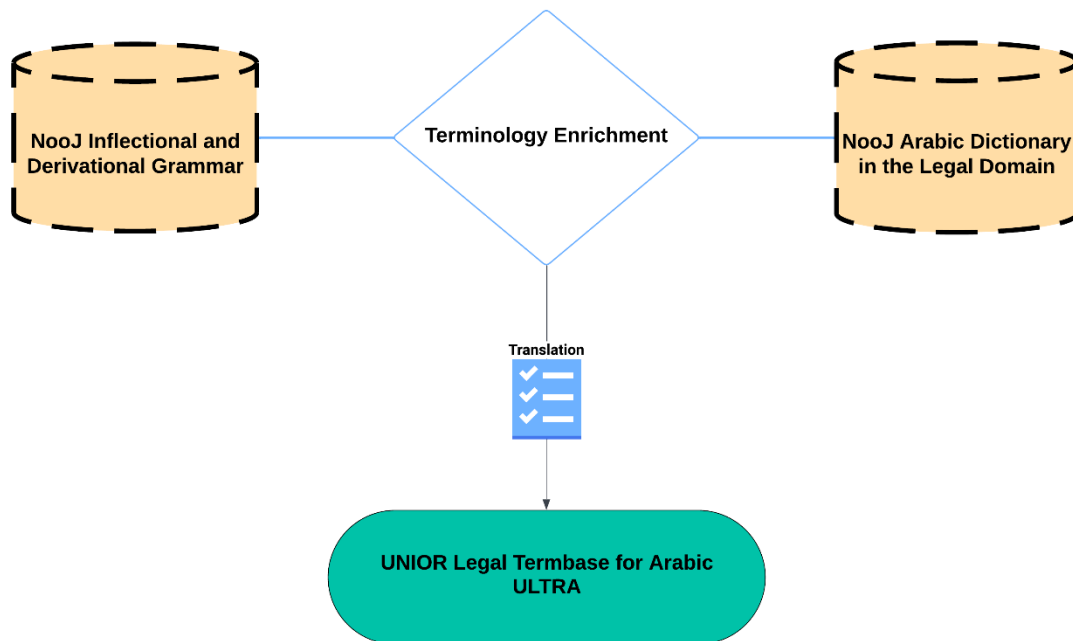


Figure 10: Resource Creation Workflow

The pipeline designed for developing the AR→EN / AR→FR legal dictionary includes the following steps:

- **Terminology Enrichment** is the point where the development of our dictionary starts after the validation of the list of terms, where we:

- Associate and/or enrich each entry with its grammatical and semantic properties,
- Associate the translations,
- Develop a dictionary that incorporates Arabic inflectional and derivational rules,
- **Translation** step pertains to the association of the AR→EN / AR→FR translation output submitted and validated to the legal expert see section 3 in Chapter Three: Research Design, Methodology, and Experimental Set-up
- The final product is our **UNIOR Legal TeRmbase for Arabic (ULTRA)**.

**Terminology Enrichment.** Following the validation process undertaken for the Arabic multilingual legal termbase, we proceed to initiate the creation of our dictionary, incorporating the validated and refined terminology as a foundational component. In this phase, we associate each entry with a set of properties as follow:

- **PoS tags:**
  - The single terms,
  - For MWEs, we provide the syntactic phrase structure composition of our Arabic legal MWEs, giving each entry its component elements (noun + noun, noun + adj, noun + preposition+noun... see Table 21).
- **Semantic properties:**
  - Domain field (legal, Juri-religion)
  - GeoUsage, following the ISO 20 771:2020 standard, for legal translation requirements, to distinguish where a given term is adapted to express a particular legal practice (Most Arab Countries (MAC), MAR (Morocco), Tunisia (TN), United Arab Emirates (UAE), Mon-archies, QA (Qatar), etc.,
- **TermNote** which can include several types of information related to Usage Note, Jurisdictional Information, Legal Reference, Clarification or distinction, etc.
- **Gender and number:**
  - Feminine or masculine, singular, or plural.
- **Translations AR-EN / AR-FR.**

The properties are encoded and defined in a NooJ dictionary (see Annex A, A1), and they are then added to our entries. These codes are instantly available and can be used in any NooJ query or grammar (Silberztein, 2003). *Table 21* provides details about the linguistic annotation and the terminology specifications for developing our resource.

Table 21: Terminology Annotation Table.

Part of Speech		Semantic Information	Domain Field	GeoUsage	Gender	Translations
Single Terms	MWE	Hum=Human LOC=Location	Juri_religion Legal	The use of ISO 3166 standard to provides codes for the names of countries.	S=feminine M=Masculine	English French
N ADJ V...	PoS	MWE PoS annotation				
	N+ADJ+V	N_ADJ1_ADJ 2_ADJ3 N_ADJ N1_N2_ADJ N_PROP_AD J...				

**Inflectional and Derivational Dictionary Rules.** The development of an inflectional and derivational dictionary is a fundamental undertaking in the field of linguistics, aimed at capturing and organizing the intricate system of word forms and morphological processes within a language (Mesfar, 2010). This academic endeavour involves meticulous analysis and documentation of inflectional patterns and derivational relationships, providing invaluable resources for researchers, language learners, and computational linguists alike. By systematically cataloguing the inflected forms and derived lexemes, such a dictionary enables a comprehensive exploration of the language's morphological structure, shedding light on its grammatical intricacies, lexical productivity, and semantic nuances (Najar, 2016).

In this phase, we create a NooJ dictionary of inflectional and derivational rules of Arabic<sup>34</sup>. This step consists of the inflection and derivation process of the entries using grammars and the transformation of grammars into transducers (Silberztein, 2016). For encoding the base form of our entries, this dictionary includes information about gender, number, and case needed to inflect our entries as well as any irregular inflection pattern that may apply such as vowels or the so called 'الشكل' in Arabic.

**Inflection.** Consider the following example in *Table 22*:

<sup>34</sup> Available here: <https://github.com/Kaitelfqih/NooJ-Inflectional-derivational-dictionary-for-Arabic.git>

Table 22: Annotation of the Entry {خَبِيرٌ} ‘Expert’.

Entry	Annotation							
	PoS	Number	FLX	Sem	DOM	GeoUsage	EN	FR
خَبِيرٌ	N	S	FLX1	Hum	Legal	MAC	Expert	Expert

The term {خَبِيرٌ} ‘expert’ refers to a person recognised as a source of expertise in a branch of knowledge, particularly in the legal field, commerce, etc<sup>35</sup>. The rule of inflection -FLX1- applied to this entry inflects its masculine noun, feminine noun, masculine singular forms, feminine singular forms, masculine plural forms, and feminine plural forms. This grammar also encodes the dual forms of both masculine and feminine genders, consider these results:

- **Feminine dual form:** خَبِيرَتَيْنِ, خَبِيرٌ+f+d+a
- **Feminine plural form:** خَبِيرَاتٌ, خَبِيرٌ+f+p+u
- **Masculine dual form:** خَبِيرَيْنِ, خَبِيرٌ+m+d+a
- **Masculine plural form:** خُبَرَاءُ, خَبِيرٌ+m+p+u

-FLX1- is diacritics-based rule, it plays an important role in conveying grammatical information about nouns. In NooJ grammar rules, diacritics are referred to as {a, u, i}: {a} for *fatha*, {u} for *damma*, {i} for *kasra*. They can be added to Arabic nouns to indicate their grammatical functions and form different case endings and gender markers. For example, the addition of the vowel {a} or (*fatha*) and the {<T>} (ة: *Tah marbouta*) to the end of a noun can indicate the nominative case for feminine nouns as in ‘خَبِيرَةٌ, خَبِيرٌ’, while the addition of the vowel {u} or (*dama*) indicates the genitive case for masculine noun as in ‘خَبِيرٌ, خَبِيرٌ’.

Arabic diacritics form an essential component of Arabic language processing in computational linguistics, i.e., they must be taken into careful consideration as they play an important role in forming cases and gender markers for our terms. Furthermore, the automatic diacritization techniques are recommended to accurately process our corpus, as they provide important cues for disambiguating the morphological and syntactic structures and distinguish between different concepts, which is what we will discuss more with the derivational rules' implementation for our entries.

**Derivation.** Derivation in Arabic linguistics is the process of forming new words by adding affixes to a root. A root is a set of consonants that carries a core meaning and can be modified

<sup>35</sup> <https://www.almaany.com/ar/dict/ar-ar/خَبِير/>

by adding various vowels and affixes, suffixes, or infixes to create different words. However, in addition to deriving words from roots, it is also possible to derive words from other words, even if they are not based on a root. Consider the following term {مَسْئُولٌ} ‘responsible’ in Table 23:

Table 23: Annotation of the Entry {مَسْئُولٌ} ‘Responsible’.

Annotation								
Entry	PoS	Number	FLX	Sem	DOM	GeoUsage	EN	FR
مَسْئُولٌ	ADJ	s	FLXDRV	Hum	legal	MAC	responsible	responsable

The term {مَسْئُولٌ} ‘responsible’ is an adjective that describes a person or entity that is accountable for their actions or decisions, and who accepts the consequences of their behaviour<sup>36</sup>. This term can be transformed into a noun by adding the suffix {مَسْئُولِيَّةٌ} (يَّة) ‘responsibility’ using the NooJ grammar rule -FlxDRV- to indicate the state or quality of being accountable for one's actions, duties, decisions, and obligations<sup>37</sup>.

Besides, this term can at the same time be inflected to indicate different cases and gender markers by the application of the same NooJ grammar rule -FlxDRV- that uses suffixes, prefixes, infixes, and diacritization, and the results are the followings:

- **Masculine form:** مَسْئُولٌ, مَسْئُولٌ +m+s+un
- **Feminine form:** مَسْئُولَةٌ, مَسْئُولَةٌ +f+s+un
- **Feminine dual form:** مَسْئُولَتَيْنِ, مَسْئُولَتَيْنِ +f+d+a
- **Feminine plural form:** مَسْئُولَاتٌ, مَسْئُولَاتٌ +f+p+un
- **Masculine dual form:** مَسْئُولَيْنِ, مَسْئُولَيْنِ +m+d+a
- **Masculine plural form:** مَسْئُولُونَ, مَسْئُولُونَ +m+p+u

**Inflectional and derivational Grammar for Singular-plural nouns.** In Arabic grammar, مفرد جمع (mufrad-jam’) refers to the distinction between singular and plural nouns. Arabic nouns can be either singular (مفرد, mufrad) or plural (جمع, jam’), and the formation of plurals can take different forms. Broken plurals (جمع تكسير, jam’ taksīr) are formed by making internal changes to the structure of the noun, while sound plurals (جمع سالم, jam’ sālim) are created by adding regular suffixes.

<sup>36</sup> <https://iate.europa.eu/search/result/1689279604635/1>

<sup>37</sup> <https://iate.europa.eu/search/result/1689279604635/1>

The term {جُمْهُورِيَّةٌ}, which means ‘republic’ serves as an example of a singular noun. For instance, {الجُمْهُورِيَّةُ الْعَرَبِيَّةُ} refers to ‘The Arab Republic’. When referring to multiple republics, the plural form {جُمْهُورِيَّات} is used, as in {إِتْحَادُ الْجُمْهُورِيَّاتِ الْعَرَبِيَّةِ} ‘Federation of Arab Republics’. In Arabic, nouns also have dual forms (المثنى, *al-muthannā*) to indicate exactly two of something. For example, to refer to two republics, the dual forms are:

- جُمْهُورِيَّتَانِ in the nominative case,
- جُمْهُورِيَّتَيْنِ in the accusative or genitive case.

To model this linguistic feature in NooJ, we use the inflectional rule  $FLX=FI$ , which employs the language-specific command  $\langle T \rangle$  to replace the *Teh Marbuta* (ة) with the *Teh Mabsutah* (ت). This allows the noun to combine with the suffixes 'نِ (-ayn) and 'انِ (-ān) for the dual forms. The following results were generated:

- جُمْهُورِيَّةٌ, جُمْهُورِيَّتَيْنِ +d+i
- جُمْهُورِيَّةٌ, جُمْهُورِيَّتَانِ +d+u
- جُمْهُورِيَّةٌ, جُمْهُورِيَّتَيْنِ +d+a

Therefore, the grammar rule helps in automatically replacing the *Teh Marbuta* (ة) with the *Teh Mabsutah* (ت) and adding the appropriate suffixes to form the dual of the term.

**Transducers:** As of now, we explained how to generate all possible noun cases and gender markers for our entries marked by the suffixes and diacritics ‘الشكل’ for each entry in our dictionary. Now we shall explain how to generate forms for entries that need suffixes, prefixes, and infixes to be inflected and derived using NooJ transducers.

NooJ formalizes the inflectional and derivational paradigms by means of finite state transducer (Silberztein, 2016). They are based on a string of letters as morphological operators performing transformations within the input lemmas. These transformations are based on the use of certain predefined generic commands:

- $\langle L \rangle$ : move left (Left arrow),
- $\langle R \rangle$ : move right (Right arrow),
- $\langle B \rangle$ : delete the last character (Backspace),
- $\langle S \rangle$ : delete the current character (Delete),
- $\langle N \rangle$ : move to the next word (Next word form),
- $\langle P \rangle$ : move to the previous word (Previous word form),
- $\langle D \rangle$ : duplicate the current character (Duplicate)
- $\langle E \rangle$ : empty string (Empty string)

and can be associated with two types of argument

- A number: for example:
  - <B2>: delete the last two characters
  - <L3>: move left three times
  - <R4>: move right four times
  - <S5>: delete the next 5 characters
- a "W": for example:
  - <BW>: delete from the current letter to the first character of the word
  - <LW>: go to the beginning of the word
  - <RW>: go to the end of the word
  - <SW>: delete from the current letter to the last character of the word

All these morphological operators work on a stack, they require a transformation time in  $O(n)$  (Mesfar, 2010). Thus, they guarantee a correspondence between the lemma and its inflected form in linear time (Mesfar, 2010). To illustrate the use of these transformational operators, consider entries in *Table 24*:

Table 24: Annotation of the Terms {عَقْدٌ} 'Contract' and {وَتِيقَةٌ} 'Document'.

Annotation							
Entry	PoS	Number	FLX	DOM	GeoUsage	EN	FR
عَقْدٌ	N	s	Oqoudon	Legal	MAC	Contract	Contrat
وَتِيقَةٌ	N	s	Wataaiq	Legal	MAC	Document	Document

To generate the plural form, we simply develop the above transducers that contains the command <LW> which means go to the beginning of the word then <R> which says skip to the other letter, while doing so we have to insert the diacritic of each letter, and after skipping two letter '<R>ُ<R>' I add the infix 'ُ' that helps generate the plural noun of 'عَقْدٌ' 'contract' plus another <R> and finally the command <SW> at the end of the annotation, and the result after inflecting the dictionary from NooJ lab is the following:

- Oqoudon=<LW><R>ُ<R>ُ<R>ُ<R>ُ<SW>/N+p;
  - عَقْدٌ, عَقُودٌ, N+Nombre=S+FLX=Oqoudon+DOM=legal+GeoUsage=MAC+EN=Contract+FR=Contrat+N+p

For the second example ‘وثيقة’ ‘document’ to become plural, the following transducer with the infix ‘اي’ and the genitive case *kasra-i* has been modelled:

- Wataaiq= <LW><R><R>اي<S><R><S><SW>/N+p;  
 - وثيقة,وثائق,N+Nombre=s+FLX=Wataaiq+DOM=legal+GeoUsage=MAC+EN=Document+FR=Document+N+p

**Multiword Expressions.** ULTRA contains MWEs, hence classifying and annotating them would have a major impact on the disambiguation of applications at the linguistic and multilingual levels working with Arabic texts (Constant et al., 2017). The diverse parsers, based on morphological aspects, are not able to recognize MWEs (Mesfar, 2008), in addition to that, morphological parsers usually separate MWEs into single terms. Therefore, recognizing MWEs as single lexical units is essential to preserve the semantics of texts. Consider the entry in Table 25:

Table 25: Annotation of the Term {مِسْطَرَّةٌ جَنَائِيَّةٌ} ‘Criminal Procedure Law’.

Entry	Annotation						
	PoS	Pattern Composition	FLX	DOM	GeoUsage	EN	FR
مِسْطَرَّةٌ جَنَائِيَّةٌ	N	N_ADJ	F1<P>F1	Legal	MAC	Criminal Procedure Law	Droit de la procédure pénale

This entry consists of two terms {مِسْطَرَّةٌ} ‘procedure’ and {جَنَائِيَّةٌ} ‘criminal’. To annotate it, we provide the POS as a single lexical unit <N> and its syntactic structure composition <N\_ADJ> along with other properties including domain, geoUsage, translations, etc. Note that the terms {مِسْطَرَّةٌ} ‘procedure’ and {جَنَائِيَّةٌ} ‘criminal’ already exists in my dictionary, so to obtain the inflected forms of the MWE, NooJ is able to reuse the inflectional paradigms -FLX=F1<P>F1- of the components without the need for separating or duplicating them. As a result, NooJ recognizes the following forms:

- مِسْطَرَّتَيْنِ جَنَائِيَّتَيْنِ, مِسْطَرَّةٌ  
 جَنَائِيَّةٌ,N+N\_ADJ+FLX=F1<P>F1+Genre=s+DOM=legal+GeoUsage=MAR+EN=criminal procedure law+FR=droit de la procédure pénale+d+i

- مسطرتان جنائيتان, مسطرة  
جناية, N+N\_ADJ+FLX=F1<P>F1+Genre=s+DOM=legal+GeoUsage=MAR+EN=criminal procedure law+FR=droit de la procédure pénale+d+u
- مسطرتين جنائيتين, مسطرة  
جناية, N+N\_ADJ+FLX=F1<P>F1+Genre=s+DOM=legal+GeoUsage=MAR+EN=criminal procedure law+FR=droit de la procédure pénale+d+a

**Fixed Terms.** Fixed terminology refers to a set of standardised and defined terms that are used in a specific field or domain to ensure consistent and accurate communication (Cabré, 2010). Given that we are working with high technical data from the legal domain in Arabic, most of our entries are fixed semantically and, in some cases, morphologically as well. Hence, no inflectional or derivational grammar rules are needed. Consider the following single fixed term and MWE {فَراش}, {طَلاقٌ رَجْعِيٌّ}, {مَأْدُونٌ}, {مُنْعَةٌ}, {طَلاقٌ بَائِنٌ بَيْنُونَةٌ صُغْرَى}, {طَلاقٌ بَائِنٌ بَيْنُونَةٌ كُبْرَى}, {فَراش}, {الطَّعْنُ}, {الْقَطْعُ} in Table 26.

Table 26: Annotation of Some fixed Terms.

Annotation								
Entries	PoS	PoS Pattern	Number	SEM	DOM	GeoUsage	EN	FR
طَلاقَ بَائنَ بَيِّنُونَةَ كُيَرَى	N	N_ADJ1_ADJ2_ ADJ3	S	-	Juri- religion	MAC	Major irrevocable divorce	divorce irrevocable majeur
طَلاقَ بَائنَ بَيِّنُونَةَ صَغْرَى	N	N_ADJ1_ADJ2_ ADJ3	S	-	Juri- religion	MAC	minor irrevocable divorce	divorce irrévocable mineur
مُتَعَةَ	N	-	S	-	Juri- religion	MAC	Compensation	Compensation
مَادُون	N	-	S	Hum	Legal	QA_EG_ SA_UAE	Superintendent registrar	Maire
طَلاقَ رَجْعِيَّ	N	N_ADJ	S	-	Juri- religion	MAC	Revocable Divorce	Divorce révocable
فِرَاش	N	-	S	-	Juri- religion	MAC	Marriage consummation	Consummation du mariage
الطَّعْن	N	-	S	-	Juri- religion	MAC	Rebuttal	Contre-preuve
الْقَطْع	N	-	S	-	Juri- religion	MAC	Certainty	Certitude

After the creation of our dictionary (ULTRA), see an extract of it in *Annexes A: A3*, the following actions are taken:

1. Save the derivational/inflectional grammar and the property definitions in the same directory as the dictionary file.
2. Implement commands to integrate the inflectional/derivational rules and property definitions into the dictionary's functionality.
3. Apply inflection to the dictionary and output the generated list of inflected and derived word forms, then verify if the dictionary is successfully compiled.

### 5.3. Evaluation and Results

In this experiment, we develop a legal electronic dictionary for Arabic that contains 1,949 entries shown in *Annex A: A3*. We believe this rate is high especially considering all the variations that result in 2,910 inflectional/derivational forms. Our electronic legal dictionary for Arabic represents 70% of single terms and 30% of MWEs.

To test the lexical coverage of ULTRA, we firstly undertake the linguistic analysis of our corpora. Afterward, we develop a NooJ grammar of recognition (*Figure 11*) and calculate the precision (**0.91**) and recall (**0.85**) according to the parsing results coverage from NooJ grammar of Recognition in *Figure 12*. These results indicate that we have achieved high quality results of recognition. However, these are possible interpretations for the low rate of the recall:

- False vocalization of terms: misplaced vowels,
- Common typographical errors: substitution of errosand ( , ) at the end of the word, and the substitution of (أ) هَمْزَةُ الْقَطْعِ and هَمْزَةُ الْوَصْلِ (ل) at the beginning, middle or end of the word,
- Lack of entries in our dictionary,
- Typographical errors, etc.

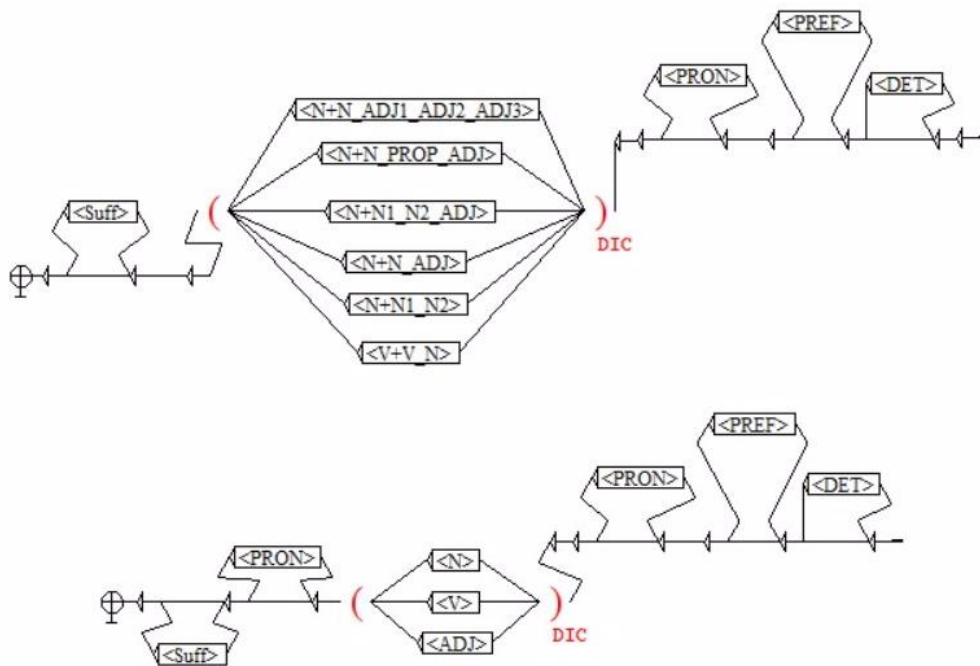


Figure 11: Grammar of Recognition for MWEs and Single Legal Entries.



keywords, providing context, and shaping the input in a way that guides the model to produce the desired response. Through prompt engineering, you can control a model's tone, style, and domain expertise without more involved customization and expensive measures like fine-tuning (AWS, 2024).

LLMs have demonstrated impressive emergent abilities as they scale, allowing effective prompting for various downstream applications (Brown et al., 2020; Kaplan et al., 2020; Wei et al., 2022b; Sahoo et al., 2024). Different from task-specific finetuning, prompting constructs task-specific prompts by rephrasing test examples with descriptive task instructions and executes the task by feeding prompts to LLMs directly. It can be further enhanced through in-context learning by providing a few labelled examples (or prompt examples) as a demonstration (Brown et al., 2020). As a new paradigm, prompting LLMs has achieved state-of-the-art performance over a range of NLP tasks (Hu et al., 2024; Trad et al., 2024). In this section, we focus on designing prompts to shape the behaviours and output of Gemini 1.5 Pro and ChatGPT-4 for MT of Arabic legislation into English and French, on the Zero-shot level.

The capability of prompting heavily depends on its surface representation, where small modifications to the prompt could cause high variance in its performance. This inspires researchers to develop advanced prompting strategies to get the most from LLMs. Gao et al. (2021) propose to generate prompt templates automatically using mT5 (Xue et al., 2020) rather than adapting manual templates. Liu et al. (2021) report selecting prompt examples close to the test input via a kNN-based retriever, Sorensen et al. (2022) resort to an information-theoretic approach based on mutual information, while Zhang et al. (2022) formulate example selection as a sequential decision problem and solved it by reinforcement learning along with Chen et al. (2023); Tonmoy et al. (2024). For reasoning tasks, Wei et al. (2022) develop chain-of-thought (CoT) prompting letting the model output the intermediate reasoning steps, which inspires researchers to further explore CoT selection (Fu et al., 2022) and decomposition (Zhou et al., 2022). The studies mentioned focus solely on other NLP tasks, while a few studies have reported translation results (Zhang et al., 2023). This, therefore, calls the need for a systematic study on how prompting must be engineered to work effectively especially in domain-specific MT. In this section, we explore prompting strategies exclusively for the translation of Arabic legal terminology in Out & In context scenarios.

## 6.2. Experimental set-up and Methodology

*Table 7 & 8* details the error counts and classification of machine-translated legal terms by ChatGPT-4 and Gemini 1.5 Pro, in and out of context scenarios from AR→EN / AR→FR, applying the MQM framework (Burchart, 2013) that uses an error typology and a scoring model. The reported results show that GT surpasses ChatGPT-4 and Gemini 1.5 Pro in achieving less critical errors in Out-Of-Context scenario in both language pairs, conversely, in In-Context scenario, ChatGPT-4 and Gemini 1.5 Pro demonstrate less critical results, but more minor and major errors compared to GT with only slight differences regarding the critical error severity count in both language pairs. This evaluation reveals significant shortcomings, particularly with GT, which fails to accurately render key terms within their legal context, often omitting crucial parts of multiword terms and producing awkward translations. The prompts provided to the models were straightforward, without any additional explanations or background information about the specific terms being translated. The only clue or context given was the general nature of the domain, but no further details were provided about the definitions or usage of the terms themselves. The focus was solely on translating the terms or sentences from AR→EN / AR→FR. We conclude that by leveraging prompt engineering strategies, we can optimise the LLMs output stimuli to generate domain-specific MT, as this approach proves to enhance the LLMs ability in handling specific translation tasks (Chu et al., 2020). In this regard, we aim at filling this gap by thoroughly examining different Zero-shot prompting set-ups on ChatGPT-4 and Gemini 1.5 Pro and compare their performance. Our methodology addresses the following key elements: the prompting strategy, the feasibility of transfer learning, and the injection of metadata for our terms both in and out of context (this includes the definitions in the source language (AR), the country where the term is used to express a legal practice, the reference to the legal document from which the term is extracted, and the publication and update dates for both source and target terms (EN & FR).

Zhang et al. (2023) mention that while performing MT, prompting need to cast the translation problem into a language modelling problem via the prompt because the format of the prompt, including its wording, differently affects how the LLM understands the task and its behaviours. In this regard, we are interested in investigating the following two questions:

1. Which template should we use for MT prompting to get the most accurate translation of Arabic legal terminology in & out of context into English and French? And what language (s) should we employ for the templates?

- What are the essential characteristics of a prompt that could shape the behaviour and output of Gemini 1.5 Pro and ChatGPT-4 for MT of Arabic legislation into English and French?

**Prompts.** Table 27 & 28 show the out-of-context and In-Context terminology translation prompts designed for ChatGPT-4 and Gemini 1.5 Pro, along with the types of errors targeted by each prompt and the language of the prompts.

Table 27: Zero-Shot Prompts for Out-of-Context MT of Arabic Legal Terminology.

		<b>Out-of-Context Prompts</b>	<b>Language of the Prompt</b>
<b>Region Specific Legal Terms Translation</b>	<b>Prompt 1</b>	You are an expert translator in the legal domain. You are specialised in the translation of Arabic legislation. Translate these terms "{source_language_terms}" into their corresponding {target_language} terms. Give only the {target_language} terms without providing extra context.	English
	<b>Prompt 2</b>	Translate the following legal terms "{source_language_terms}" into {target_language}. Give only the {target_language} terms without providing extra context. But please note that: <ul style="list-style-type: none"> <li>- You are an expert legal translator with extensive experience in translating region-specific legal terms from Arabic into {target_language}. You possess a deep understanding of legal terminology and cultural nuances, allowing you to provide accurate translations that maintain the original meaning and context.</li> </ul>	English
	<b>Prompt 3</b>	Translate the following legal terms "{source_language_terms}" into English/French. Give only the {target_language} terms without providing extra context. But please note that: <ul style="list-style-type: none"> <li>- You are an expert legal translator with extensive experience in translating region-specific legal terms from Arabic into English. You possess a deep understanding of legal terminology and cultural nuances, allowing you to provide accurate translations that maintain the original meaning and context. To assist you very well with the task, consider the metadata of the following {source_language} legal terms which I need you to translate: {source_language_term(s)}: is a term used in {GeoUsage}.</li> <li>- Also note that both terms convey the same legal practice in the target language, this is for you to avoid the word-for-word translation and generate the terms used in the target languages.</li> <li>- {Source_language_term} is extracted from {document name, date of publication, last update, the reference of the document}.</li> </ul>	English
<b>Non-Equivalent Entries and Unrecognised Legal Practices in Target Legal Systems</b>	<b>Prompt 1</b>	You are an expert translator in the legal domain. You are specialised in the translation of Arabic legislation. Translate these terms "{Source_language_terms}" into their corresponding {target_language} terms. Give only the {target_language} terms without providing extra context.	English
	<b>Prompt 2</b>	Translate the following legal terms "{source_language}" into {target_language}. Give only the {target_language} terms without providing extra context. But please note that:	English

		<ul style="list-style-type: none"> <li>- You are an expert legal translator with extensive experience in translating region-specific legal terms from {source_language} into {target_language}. You possess a deep understanding of legal terminology and cultural nuances, allowing you to provide accurate translations that maintain the original meaning and context.</li> </ul>	
	<b>Prompt 3</b>	<p>Translate the following Arabic legal terms "{source_language}" into {target_language}. Give only the {target_language} terms without providing extra context. But please note that:</p> <ul style="list-style-type: none"> <li>- You are a highly skilled legal translator with extensive experience in translating complex legal terms and concepts across multiple jurisdictions. You excel at identifying terms without direct translations and understand how to interpret the source context to generate accurate, legally-correct translations.</li> <li>- Keep in mind that the translation must respect the intricacies of legal language and the specific terminology used in Shariah Law. It is crucial to maintain fidelity to the original text while also ensuring that the translated terms are understood accurately in the target language. Be prepared to provide clarity on any term that may have multiple meanings or interpretations.</li> <li>- Please examine the legal system of the source language and the target legal system. If the provided term has no direct translation and/or is not recognised in the practices of the target legal systems, ensure to refer to the source context and generate the most accurate interpretation based on that context. To assist you very well with the task, consider the metadata of the key legal terms which I need you to take into consideration while translating: <ul style="list-style-type: none"> <li>o {source_language_terms}: {definition}+ {Document Name, Article Number, Date of Publication, Last Update, the Reference of the Document }</li> </ul> </li> </ul>	English & Arabic
<b>MT Inconsistent Generation of terminology</b>	<b>Prompt 1</b>	<p>You are an expert translator in the legal domain. You are specialised in the translation of Arabic legislation. Translate these terms "{source_language}" into their corresponding {target_language} terms. Give only the {target_language} terms without providing extra context.</p>	English
	<b>Prompt 2</b>	<p>Translate the following Shariah law terms "{source_language}" into their corresponding {target_language} terms. Provide the {target_language} equivalents with no extra context. However, ensure the following:</p> <ul style="list-style-type: none"> <li>- The translations are precise and maintain the legal nuances of Shariah law.</li> <li>- If the term lacks a direct equivalent in the target legal system, carefully examine both the source and target legal systems.</li> <li>- Ensure that the translated terms either match their usage in Islamic legal contexts or reflect the closest possible interpretation, preserving the integrity of the original legal meaning.</li> <li>- Avoid over-simplification or generalization, especially for legal terminology unique to Shariah law.</li> </ul>	English

Table 28: Zero-Shot Prompts for In-Context MT of Arabic Legal Terminology.

		In-Context Prompts	Language of the Prompt
<b>Region Specific Legal Terms Translation</b>	<b>Prompt 1</b>	<p>You are an expert translator in the legal domain. You are specialised in the translation of Arabic legislation. Translate the following "{source_language}" legal sentences into their corresponding English/French terms. Give only their corresponding {target_language} sentences without providing extra context.</p> <ul style="list-style-type: none"> <li>- "{source_language_sentences}"</li> </ul>	English
	<b>Prompt 2</b>	<p>translate the following Shariah law sentences from "{source_language}" into {target_language} without providing any additional comments:</p> <ul style="list-style-type: none"> <li>- "{source_language_sentences}"</li> </ul> <p>Please note that:</p> <ul style="list-style-type: none"> <li>- In the above legal sentences, the terms "{source_language_terms}" must be given prominent attention to their translation;</li> <li>- You are an expert legal translator with extensive experience in translating region-specific legal terms from "{source_language}" into {target_language}. You possess a deep understanding of legal terminology and cultural nuances, allowing you to provide accurate translations that maintain the original meaning and context. To assist you very well with the task, consider the metadata of the key legal terms which I need you to take into consideration while translating; {source_language_term(s)}: is a term used in {GeoUsage}.</li> <li>- Also note that both terms convey the same legal practice in the target language. This is to avoid the word-for-word translation and generate the terms as used in the target languages and in their context;</li> <li>- Please refer to the {target_language} regulations that sets out the penalties for {the_type_of_the_legislation}.</li> </ul>	English
<b>Non-Equivalent Entries and Unrecognised Legal Practices in Target Legal Systems</b>	<b>Prompt 1</b>	<p>You are an expert translator in the legal domain. Translate the following "{source_language}" legal sentences into {target_language} without providing any additional comment:</p> <ul style="list-style-type: none"> <li>- {source_language_sentences}</li> </ul>	English
	<b>Prompt 2</b>	<p>You are an expert translator in the legal domain. Translate the following Shariah law sentences from "{source_language}" into {target_language} without providing any additional comments:</p> <ul style="list-style-type: none"> <li>- {source_language_sentences}</li> </ul> <p>Please note that:</p> <ul style="list-style-type: none"> <li>- In the above legal sentences, the terms "{source_language_terms}" must be given prominent attention to their translation;</li> <li>- You are a highly skilled legal translator with extensive experience in translating complex legal terms and concepts across multiple jurisdictions. You excel at identifying terms without direct translations and understand how to interpret the source context to generate accurate, legally-correct translations.</li> <li>- Keep in mind that the translation must respect the intricacies of legal language and the specific terminology used in Shariah Law. It is crucial to maintain fidelity to the original text while also ensuring that the translated terms are understood accurately in the target language. Be prepared to provide clarity on any terms that may have multiple meanings or interpretations.</li> </ul>	English & Arabic

		<p>- Please examine the legal system of the source language and the target legal system. If the provided term has no direct translation and/or is not recognised in the practices of the target legal systems, ensure to refer to the source context and generate the most accurate interpretation based on that context. To assist you very well with the task, consider the metadata of the key legal terms which I need you to take into consideration while translating:</p> <ul style="list-style-type: none"> <li>• {source_language_terms}: {definition} + {Document Name, Article Number, Date of Publication, Last Update, the Reference of the Document }</li> </ul>	
<b>MT Inconsistent Generation of Shari'ah Law terminology</b>	<b>Prompt 1</b>	<p>You are an expert translator in the legal domain. Translate the following "{source_language}" legal sentences into {target_language} without providing any additional comment:</p> <ul style="list-style-type: none"> <li>• {source_language_sentences}</li> </ul>	English
	<b>Prompt 2</b>	<p>You are an expert translator in the legal domain. You are specialised in the translation of Arabic legislation. Translate the following Shariah law sentences from "{source_language}" into {target_language} without providing any additional comments.</p> <ul style="list-style-type: none"> <li>• {source_language_sentences}</li> </ul> <p>Please note that:</p> <ul style="list-style-type: none"> <li>- In the above legal sentences, the terms "{source_language_terms}" must be given prominent attention to their translation.</li> <li>- The translations are precise and maintain the legal nuances of Shariah law.</li> <li>- If the terms lack a direct equivalent in the target legal system, carefully examine both the source and target legal systems.</li> <li>- Ensure that the translated terms either match their usage in Islamic legal contexts or reflect the closest possible interpretation, preserving the integrity of the original legal meaning.</li> <li>- Avoid over-simplification or generalization, especially for legal terminology unique to Shariah law.</li> </ul>	English

The following is a breakdown of the prompts designed to address each error dimension, starting with the Out-of-Context scenario:

- **Region-Specific Legal Terms Translation.** This set of prompts is aimed at instructing the models to recognize and translate commonly used legal terms specific to various Arab legal systems. The terms are provided with the instruction to output only the English or French equivalents. By removing any additional context, the models are instructed to focus solely on the accurate translation of legal terminology, ensuring that they build the capacity to recognize and convert domain-specific legal terms without relying on extra explanatory information. This approach helps refine the model's ability to produce concise and correct translations in different legal contexts.
- **Non-Equivalent Entries and Unrecognised Legal Practices in Target Legal Systems.** This scenario is designed to instruct the models on more nuanced legal concepts that may not have direct equivalents in the target legal systems, especially the

terms that are deeply embedded in the cultural and legal practices of Shari'ah law. The prompts progressively introduce additional layers of guidance to help the models develop a more advanced understanding of how to handle such terms.

In most advanced prompts, the models are instructed to consider legal metadata (e.g., *document name, date of publication, last update, the reference of the document*, etc) when translating, thereby training them to pay attention to jurisdictional details and legal contexts. The goal here is to teach the models to identify terms that might not have a direct translation and provide the best possible equivalent based on the context of the target legal system. This method helps the models balance legal accuracy with cultural sensitivity.

- **MT Inconsistent Generation of Shari'ah Law terminology.** These prompts focus on training the models to deal with inconsistencies that may arise when translating terms that are deeply rooted in Shari'ah law and other Arabic legal contexts. The models are instructed to avoid oversimplification and ensure they maintain the integrity of the original meaning into English and French translations. Additionally, the prompts encourage the models to recognize when a term lacks an exact counterpart in the target legal system, instructing them to provide translations that respect the original term's complexity and significance. This helps the models learn to bridge gaps between different legal systems while maintaining accuracy across translations.

**For the In-Context scenario:**

- **Region-Specific Legal Terms Translation.** In this set of prompts for this error dimension, the models are trained to translate legal sentences that include terms which have region-specific legal meanings that vary across Arab countries. The objective is to instruct the models to:
  1. Translate legal sentences holistically, recognizing that legal terms within these sentences must be adapted based on the specific legal systems in different regions.
  2. Contextualise legal terminology to avoid literal translation errors to help the models grasp how these terms function in the sentence and generate translations that are legally accurate within both the source and target legal frameworks.
  3. Leverage legal metadata embedded in the prompt, guiding the models toward the correct legal interpretation based on jurisdiction.

- **Non-Equivalent Entries and Unrecognised Legal Practices in Target Legal Systems.** These prompts focus on translating legal sentences that include complex legal concepts, such as different types of divorces under Sharī'ah law (e.g., { طلاق بائن بينونة كبرى } and { طلاق رجعي }, etc). Since many of these terms do not have direct equivalents in the English and the French legal systems, the prompts are designed to guide the models towards:
  1. Interpreting nuanced legal concepts in context, recognizing that terms like { طلاق بائن بينونة كبرى } may not exist in Western legal systems. By embedding these terms within legal sentences, the models learn to avoid literal translation and instead focus on conveying the legal meaning and the specific legal practice
  2. Focusing on legal accuracy while translating, understanding that each sentence is tied to culturally specific legal codes. This requires the models to process the embedded metadata and generate context-aware translations that remain faithful to the original meaning.
  3. Dealing with legal discrepancies and adapt when the legal practice is not recognised in the target system, providing a translation that reflects the closest possible interpretation of the source text without losing the legal intent.
- **MT Inconsistent Generation of Sharī'ah Law terminology.** The prompts designed for this error dimension try to optimise the models to handle legal inconsistencies that are inherent in translating legal terms related to Islamic law and that represent practices in Sharī'ah law but have no direct equivalent in non-Islamic legal systems. These prompts are used to instruct the models to:
  1. Accurately interpret legal terms within the broader legal sentence, ensuring the model does not misrepresent the term's meaning by oversimplifying it.
  2. Focus on maintaining the legal nuances of Islamic law while translating into English or French, ensuring that these culturally bound practices are correctly represented. The model needs to learn how to adapt these legal terms to the target legal systems limitations while retaining the legal implications of the original Arabic sentence.
  3. Consider the surrounding legal text, which helps the model better understand how terms are being used and generate translations that preserve the legal meaning in its context. This avoids literal translations that might distort legal processes due to the lack of direct equivalence in the target language.

The prompts are presented in a bilingual framework, incorporating Arabic for the presentation of the metadata and any crucial instructions related to the entries and sentences, and English for the entire prompts design. With the experimental set-up and methodology outlined, we now turn our attention to present the evaluation and the results.

### 6.3. Evaluation and Results

AWS documentation on models customization recommend dedicating time to prompt engineering before considering fine-tuning a model on additional data (AWS, 2024). They add that the goal is to provide sufficient context and guidance to the model so that it can generalize and perform well on unseen or limited data scenarios (AWS, 2024). Therefore, in our experiment, we focus on shaping the behaviour and output of ChatGPT-4 and Gemini 1.5 Pro for MT of Arabic legislation into English and French. We employ a mixed-methods approach, including zero-shot prompting, contextual prompting, and transfer learning prompting. Each method plays a key role in optimising the models' MT performance for Arabic legal terms, both in and out of context. The architecture of the prompts, as shown in *Table 27 & 28*, follows a structured process: (i) defining the task, (ii) understanding the models' capabilities, including their strengths and weaknesses, (iii) selecting the appropriate format for each prompt, (iv) providing advanced contextual details, (v) demonstrating the desired style and tone, and (vi) testing and iterating to reinforce the context when needed.

Our evaluation proceeds in two main phases: first, we assess the effectiveness of the prompts designed for each specific objective and scenario; second, we compare the behaviour and output of ChatGPT-4 and Gemini 1.5 Pro across these prompts, benchmarking their results against our gold standard dataset.

Table 29. Prompting Results in Out-of-Context scenario.

		Out-of-Context				
		ChatGPT-4			Gemini 1.5 Pro	
		Source Term	AR→EN	AR→FR	AR→EN	AR→FR
Region Specific Legal Terms Translation	Prompt 1	مدونة الأسرة	Family Code	Code de la famille	Family Code	Code de la famille
		قانون الأسرة	Family Code	Code de la famille	Family Law	Code de la famille
		مجلة الأحوال الشخصية	Personal Status Code	Code des personnes	Personal Status Law	Code du statut personnel

	<b>Prompt 2</b>	نظام الأحوال الشخصية	Personal Status Law	Système des personnes	Personal Status System	Code de la famille	
		مدونة الأسرة	Family Code	Code de la famille	Family Code	Code de la famille	
		قانون الأسرة	Family Code	Code de l'état civil	Family Law	Code de la famille	
		مجلة الأحوال الشخصية	Code of Personal Status	Système de l'état civil	Personal Status Law	Code de la famille	
		نظام الأحوال الشخصية	Personal Status System	Système de l'état civil	Personal Status System	Code de la famille	
	<b>Prompt 3</b>	مدونة الأسرة	Family Law	Code de la famille	Family Law	Code de la famille	
		قانون الأسرة	Family Law	Code de la famille	Family Law	Code de la famille	
		مجلة الأحوال الشخصية	Family Law	Code de la famille	Family Law	Code de la famille	
		نظام الأحوال الشخصية	Family Law	Code de la famille	Family Law	Code de la famille	
	<b>Non-Equivalent Entries and Unrecognised Legal Practices in Target Legal Systems</b>	<b>Prompt 1</b>	طلاق إتفاقي	Amicable Divorce	consentement mutuel	Mutual consent divorce	Divorce par consentement mutuel
			طلاق بائن بينونة كبرى	Major Irrevocable Divorce	Divorce irrémédiable majeur	Irrevocable divorce	Divorce revocable
			طلاق بائن بينونة صغرى	Minor Irrevocable Divorce	Divorce irrémédiable mineur	Revocable divorce	Divorce irrevocable
<b>Prompt 2</b>		طلاق إتفاقي	Consensual Divorce	Divorce par consentement mutuel	Mutual consent divorce	Divorce par consentement mutuel	
		طلاق بائن بينونة كبرى	Major Irrevocable Divorce	Divorce irrémédiable majeur	Irrevocable divorce	Divorce irrévocable majeur	
		طلاق بائن بينونة صغرى	Minor Irrevocable Divorce	Divorce irrémédiable mineur	Revocable divorce	Divorce irrévocable mineur	
<b>Prompt 3</b>		طلاق إتفاقي	Agreed divorce	Divorce par consentement mutuel	Mutual consent divorce	Divorce par consentement mutuel	
		طلاق بائن بينونة كبرى	Major Irrevocable Divorce	Divorce irrévocable majeur	Major irrevocable divorce	Divorce irrévocable mineur	
		طلاق بائن بينونة صغرى	Minor Irrevocable Divorce	Divorce irrémédiable mineur	Minor irrevocable divorce	Divorce irrévocable majeur	
<b>MT Inconsistent Generation of terminology</b>	<b>Prompt 1</b>	اللعان	Li'an	Déni par serment	Oath of repudiation	La lacération	
		فراش	Consummation of Marriage	Présomption de paternité	Bed	Le lit conjugal	
		البناء	Presumption of Paternity	Cohabitation légale	Cohabitation Corporal	La consommation du mariage	
		الطعن	Appeal	Recours	punishment (or amputation,	Le recours	

					depending on context)	
<b>Prompt 2</b>	اللعان	Oath of Condemnation	Imprécation légale	Repudiation	Accusation	
	فراش	Consummation of Marriage	Consommation légale du mariage	Bed	Le lit conjugal	
	البناء	Presumption of Paternity	Présomption légale de paternité	Marriage consummation	La consommation du mariage	
	الطعن	Appeal	Recours	Appeal	Le recours	

Table 30. Prompting Results in In-Context Scenario.

		<b>In-Context</b>				
		<b>ChatGPT-4</b>			<b>Gemini 1.5 Pro</b>	
		<b>Source Sentence</b>	<b>AR→EN</b>	<b>AR→FR</b>	<b>AR→EN</b>	<b>AR→FR</b>
<b>Region Specific Legal Terms Translation</b>	<b>Prompt 1</b>	يطبق قانون العقوبات على كافة الجرائم التي ترتكب في أراضي الجمهورية	The <b>Penal Code</b> applies to all crimes committed within the territory of the Republic.	<b>Le Code pénal</b> s'applique à tous les crimes commis sur le territoire de la République.	The <b>Penal Code</b> shall apply to all crimes committed within the territory of the Republic.	Le <b>code pénal</b> s'applique à tous les crimes commis sur le territoire de la République.
		تتميمًا لأحكام المادة 49 أعلاه، بمقتضى المادة الرابعة من القانون رقم 86.14 القاضي بتغيير وتتميم بعض أحكام مجموعة القانون الجنائي وقانون المسطرة الجنائية المتعلقة بمكافحة الإرهاب	In addition to the provisions of Article 49 above, pursuant to Article 4 of Law No. 86.14 amending and supplementing certain provisions of the <b>Penal Code</b> and the Criminal Procedure Code related to combating terrorism.	En complément des dispositions de l'article 49 ci-dessus, conformément à l'article 4 de la loi n° 86.14 modifiant et complétant certaines dispositions du <b>Code pénal</b> et du Code de procédure pénale relatives à la lutte contre le terrorisme.	In addition to the provisions of Article 49 above, pursuant to Article 4 of Law No. 86.14 amending and supplementing certain provisions of the <b>Penal Code</b> and the Code of Criminal Procedure relating to the fight against terrorism.	En complément des dispositions de l'article 49 ci-dessus, et conformément à l'article 4 de la loi n° 86.14 modifiant et complétant certaines dispositions du <b>code pénal</b> et du code de procédure pénale relatives à la lutte contre le terrorisme.
	<b>Prompt 2</b>	يطبق قانون العقوبات على كافة الجرائم التي ترتكب في أراضي الجمهورية	The <b>Criminal Law</b> applies to all crimes committed within the territory of	Le <b>Code pénal</b> s'applique à tous les crimes commis sur le territoire de la République.	The <b>Criminal Code</b> shall apply to all crimes committed within the territory of the Republic.	Le <b>code pénal</b> s'applique à tous les crimes commis sur le territoire de la République.

			the Republic.			
		تتميماً لأحكام المادة 49 أعلاه، بمقتضى المادة الرابعة من القانون رقم 86.14 القاضي بتغيير أحكام مجموعة القانون الجنائي وقانون المسطرة الجنائية المتعلقة بمكافحة الإرهاب	In addition to the provisions of Article 49 above, pursuant to Article 4 of Law No. 86.14 amending and supplementing certain provisions of the <b>criminal code</b> and the Criminal Procedure Code related to combating terrorism.	En complément des dispositions de l'article 49 ci-dessus, conformément à l'article 4 de la loi n° 86.14 modifiant et complétant certaines dispositions du <b>code pénal</b> et du Code de procédure pénale relatives à la lutte contre le terrorisme.	In addition to the provisions of Article 49 above, pursuant to Article 4 of Law No. 86.14 amending and supplementing certain provisions of the <b>criminal law</b> and the Code of Criminal Procedure relating to the fight against terrorism.	En complément des dispositions de l'article 49 ci-dessus, et conformément à l'article 4 de la loi n° 86.14 modifiant et complétant certaines dispositions du <b>droit pénal</b> et du code de procédure pénale relatives à la lutte contre le terrorisme.
Non-Equivalent Entries and Unrecognised Legal Practices in Target Legal Systems	Prompt 1	كما تم إقرار حق الطلاق الاتفاقي تحت مراقبة القاضي	The right to <b>consensual divorce</b> has been approved under the supervision of the judge.	Le droit de <b>divorce par consentement</b> a été reconnu sous la surveillance du juge.	The right to an <b>agreed-upon divorce</b> was also acknowledged under the supervision of the judge.	Il a également été reconnu le droit au <b>divorce par consentement mutuel</b> sous le contrôle du juge.
		وحيث تحقق إلينا سبق الزوجية بينهما وعدم الدخول والخلوة الصحيحة بإقراره وإفادة المعرفين المذكورين فقد أفهمته أنه وقع منه طلاق بانن بينونة صغرى على زوجته المذكورة	Since it has been established that there was a prior marriage between them and that there was no consummation or valid privacy, as acknowledged by him and confirmed by the mentioned witnesses, I informed him that he had issued a <b>definitive divorce (minor separation)</b>	Étant donné que le mariage entre eux a été établi et qu'il n'y a pas eu de relations sexuelles ni de cohabitation valide, selon sa déclaration et les témoignages des témoins mentionnés, il a été informé qu'il avait prononcé un <b>divorce irrévocable de type mineur</b> avec sa femme mentionnée.	And since it has been established to us that they were previously married, and there was no consummation of the marriage or proper seclusion, as per his acknowledgment and the testimony of the aforementioned witnesses, we have informed him that he has issued a <b>revocable divorce</b> to his aforementioned wife.	Et comme il nous a été établi qu'il y avait eu mariage entre eux sans consommation ni cohabitation régulière, sur la base de son aveu et des déclarations des témoins susmentionnés, il lui a été expliqué qu'il avait prononcé un <b>divorce révocable</b> sur sa femme susmentionnée.

			from his aforementio ned wife.			
		في حالة لطلاق البيان بينونة كبرى، لا يجوز للرجل أن يستعيد زوجته إلا إذا تزوجت من غير مطلقها	In the case of a <b>definitive divorce (major separation)</b> , a man cannot reclaim his wife unless she has married someone other than her divorcer.	Dans le cas d'un <b>divorce irrévocable de type majeur</b> , l'homme ne peut récupérer sa femme que si elle s'est mariée avec un autre que son ancien mari.	In the case of an <b>irrevocable divorce</b> , a man cannot take back his wife unless she marries someone other than her former husband.	En cas de <b>divorce irrévocable définitif</b> , l'homme ne peut récupérer sa femme que si elle se marie avec un autre homme que lui.
	Prompt 2	كما تم إقرار حق الطلاق الاتفاقي تحت مراقبة القاضي	The right to <b>consensual divorce</b> has been approved under the supervision of the judge.	Le droit de <b>divorce par consentement</b> a été reconnu sous la surveillance du juge.	Moreover, the right to a <b>mutually consent divorce</b> was recognized under the supervision of the judge.	Il a également été reconnu le droit au <b>divorce par consentement mutuel</b> , sous le contrôle du juge.
		وحيث تحقق إلينا سبق الزوجية بينهما وعدم الدخول والخلوة الصحيحة بإقراره و إفادة المعرفين المذكورين فقد أفهمته أنه وقع منه طلاق بيان بينونة صغرى على زوجته المذكورة	Since it has been established that there was a prior marriage between them and that there was no consummati on or valid privacy, as acknowledg ed by him and confirmed by the mentioned witnesses, I informed him that he had issued a a <b>minor irrevocable divorce</b> from his aforementio ned wife.	Étant donné que le mariage entre eux a été établi et qu'il n'y a pas eu de relations sexuelles ni de cohabitation valide, selon sa déclaration et les témoignages des témoins mentionnés, il a été informé qu'il avait prononcé <b>un divorce irrévocable de type mineur</b> avec sa femme mentionnée.	Given that we have ascertained the previous marriage between them, and that there was no consummation of the marriage or appropriate seclusion as per his admission and the testimony of the mentioned witnesses, we have informed him that he has issued a <b>minor irrevocable divorce</b> to his aforementioned wife.	Et comme il a été établi qu'il y a eu mariage entre eux sans consommation ni cohabitation régulière, sur la base de son aveu et des déclarations des témoins susmentionnés, il lui a été expliqué qu'il avait prononcé <b>un divorce mineur irrévocable</b> sur sa femme susmentionnée.
		في حالة لطلاق البيان بينونة كبرى، لا يجوز للرجل	In the case of a a <b>major irrevocable divorce</b> , a	Dans le cas d'un <b>divorce irrévocable de type majeur</b> ,	In the case of a <b>major irrevocable divorce</b> , a man	En cas de <b>divorce irrévocable majeur</b>

		أن يستعيد زوجته إلا إذا تزوجت من غير مطلقها	man cannot reclaim his wife unless she has married someone other than her divorcer.	l'homme ne peut récupérer sa femme que si elle s'est mariée avec un autre que son ancien mari.	may not take back his wife unless she marries another man and subsequently divorces him.	définitif, l'homme ne peut récupérer sa femme que si elle se marie avec un autre homme que lui.
MT Inconsistent Generation of terminology	Prompt 1	يعتبر الفراش بشروطه حجة قاطعة على ثبوت النسب، لا يمكن الطعن فيه إلا من الزوج عن طريق اللعان، أو بواسطة خبرة تفيد القطع.	The <b>bed</b> with its conditions is considered conclusive proof of the establishment of paternity, and it can only be <b>contested</b> by the husband through <b>li'an</b> or by an expert report providing <b>certainty</b> .	Le <b>lit conjugal</b> , sous ses conditions, est une preuve <b>irréfutable</b> de la filiation, qui ne peut être <b>contestée</b> que par le mari à travers la procédure de la <b>li'an</b> ou par une expertise apportant une preuve <b>certaine</b> .	The <b>marital bed</b> , under its conditions, is conclusive proof of paternity and cannot be <b>challenged</b> except by the husband through the <b>oath of denial (li'an)</b> or through expert testimony that provides <b>certainty</b> .	Le <b>lit conjugal</b> , dans ses conditions, est une preuve concluante de paternité et ne peut être <b>contesté</b> que par le mari par le <b>serment de dénégation (li'an)</b> ou par un témoignage d'expert apportant la <b>certitude</b> .
		يعتبر الفراش بشروطه حجة قاطعة على ثبوت النسب، لا يمكن الطعن فيه إلا من الزوج عن طريق اللعان، أو بواسطة خبرة تفيد القطع.				
		إذا لم يتراض الزوجان بعد البناء على قدر الصداق في زواج التفويض، فإن المحكمة تحده مراعية الوسط للزوجين.				

Prompt 2	يعتبر الفراش بشروطه حجة قاطعة على ثبوت النسب، لا يمكن الطعن فيه إلا من الزوج عن طريق اللعان، أو بواسطة خبرة تفيد القطع.	The marital bed, under its conditions, constitutes conclusive proof of the establishment of lineage, which can only be contested by the husband through the process of li'an, or by an expert opinion providing definitive proof.	Le lit conjugal avec ses conditions constitue une preuve irréfutable de l'établissement de la filiation, et ne peut être contesté que par le mari par le biais du li'an, ou par une expertise apportant une certitude.	The marital bed, under its conditions, is conclusive proof of paternity and cannot be rebutted except by the husband through the oath of denial or by expert testimony providing certainty.	Le lit conjugal, sous réserve de certaines conditions, constitue une preuve irréfutable de la paternité, qui ne peut être contestée que par le mari par le biais de l'accusation ou par une expertise concluante.
	يعتبر الفراش بشروطه حجة قاطعة على ثبوت النسب، لا يمكن الطعن فيه إلا من الزوج عن طريق اللعان، أو بواسطة خبرة تفيد القطع.				
	إذا لم يتراض الزوجان بعد البناء على قدر الصداق في زواج التفويض، فإن المحكمة تحدد مراعية الوسط الاجتماعي للزوجين.				

Table 29 & 30 present the different outputs generated by both models from AR→EN / AR→FR in out and in context, using the various prompts designed (see Table 27 & 28) for each error dimension being addressed.

### A. Out-of-Context AR→EN/ AR→EN

**Region Specific Legal Terms Translation.** For the Out-of-Context scenario in *Table 29*, the first prompt designed for shaping the LLMs behaviour when it comes to Region Specific Legal Terms Translation, is very basic. ChatGPT-4 and Gemini 1.5 are simply asked to translate the legal terms into their English and French equivalents without additional instructions or context. For example, in the AR→EN pair, ChatGPT-4 produces different translations for {مجلة الأحوال الشخصية} ‘Personal Status Code’ and {نظام الأحوال الشخصية} ‘Personal Status Law while {مدونة الأسرة} and {قانون الأسرة} are correctly rendered. These terms are often used to refer to similar legal frameworks in different regions, but ChatGPT-4 distinguishes between them. In AR→FR pair, the translation shows some inconsistency, especially in translating {نظام الأحوال الشخصية} which appears as ‘Système des personnes’ an awkward and less accurate result. Similar to ChatGPT-4, in AR→EN pair, Gemini 1.5 Pro distinguishes between the terms but introduces variations like ‘Personal Status System’ instead of ‘Code’ or ‘Law’. The use of ‘System’ here feels less appropriate for legal terms, suggesting that the model struggles with contextualizing itself. However, in AR→FR pair, Gemini 1.5 Pro is consistent, translating the same term {نظام الأحوال الشخصية} as ‘Code de la famille’ which indeed captures the precise legal concept and demonstrates precision compared to ChatGPT-4. Prompt 2 shows some improvements as ChatGPT-4 demonstrates better consistency in English, generating ‘Family Code’ and ‘Code of Personal Status’, but these still fall short of matching the reference term ‘Family Law’. In French, it continues to use incorrect terms like ‘Système de l’état civil’ instead of ‘Code de la Famille’. Gemini 1.5 Pro also fails to align with the reference translations, still producing terms such as ‘Personal Status Law’ and ‘Personal Status System’ in English. In French, the model continues to vary its output, generating ‘Code du statut personnel’. Although both models show slight improvement with more detailed instructions, they still struggle without the inclusion of regional metadata.

Prompt 3 highlights a significant improvement in both models' performances. ChatGPT-4, when provided with detailed regional metadata about the source entries, generates accurate output for all terms in English, fully matching the reference translation ‘Family Law’. Similarly, in French, it consistently applies the correct term ‘Code de la Famille’ across all instances. Gemini 1.5 Pro also shows substantial improvement under these conditions, correctly translating all terms as ‘Family Law’ in English and ‘Code de la Famille’ in French. This analysis underscores the importance of including regional context and detailed

instructions in legal translation tasks, as both models were able to consistently match the reference translations when the metadata were included.

**Non-Equivalent Entries and Unrecognised Legal Practices in Target Legal Systems.** As long as this error type is concerned in Out-of-Context scenario, ChatGPT-4's performance improves as more context and/ or metadata is provided. However, it consistently misses the precise term 'Divorce by mutual consent' for {طلاق إنتفاقي}, opting for 'Amicable' or 'Agreed Divorce' especially in Prompt 2 & 3. This shows that while it captures the broader meaning, it could benefit from stricter legal terminology. Gemini 1.5 Pro starts off less precise, especially in distinguishing '(Major) and (Minor) Irrevocable Divorce', but it shows significant improvement with more detailed prompts, particularly in prompt 3. It also nails the 'Divorce by mutual consent' term earlier than ChatGPT-4 in both language pairs. Overall, both models benefit from detailed metadata and more specific instructions. The more comprehensive the legal background, the better the translations. This demonstrates how both models rely heavily on well-designed prompts to make legally accurate translations.

**MT Inconsistent Generation of terminology.** Two prompts have been considered for this type of error; it includes prompts designed to solve the issue posed by the terms that might be literally translated by the LLMs, such as {اللعان, فراش, البناء, الطعن}. ChatGPT-4 struggles with {اللعان} 'Sworn allegation' and {الطعن} 'Rebuttal'. The added legal nuance in Prompt 2 improved some translations, but ChatGPT-4 still had issues with consistency and accuracy, especially when dealing with terms like {البناء} and {فراش}, which should both be 'Marriage consummation'. In some cases, such as with {اللعان} 'Oath of Condemnation' and {الطعن} 'Appeal', it adds unnecessary legal qualifiers, complicating translations.

Gemini 1.5 Pro often mistranslates key terms such as {اللعان} 'Repudiation' or 'Accusation' and {فراش} 'Bed', showing less accuracy than ChatGPT-4, even with additional context. Similar to ChatGPT-4, it benefits from more detailed instructions, but it still struggles to capture the nuances of Sharī'ah legal terminology. Both models exhibit inconsistent generation of legal terms when translating from Arabic, especially in those that are specific to Sharī'ah law. Some terms are handled well, like {البناء} 'Marriage consummation', but other terms are either over-simplified or over-complicated, leading to significant deviations from the reference translations.

In some cases, even with advanced prompt adjustments, the models request the context where specific terms occur, generating the following: **'please provide the context so I can**

**offer the most accurate and appropriate translation.** Zhang et al. (2023) believe that this happens because language models rely heavily on context to understand the nuances of words and phrases. Isolated terms can have multiple meanings depending on the surrounding text, cultural factors, or specific fields of use. Without sufficient context, the model may misinterpret the intended meaning, leading to inaccurate translations. By asking for more information, the model ensures it can account for these ambiguities and provide accurate translation (Zhang et al., 2023), but how this hypothesis can be verified through the analysis of the translation of these legal terms in their context?

### **B. In-Context AR→EN/ AR→EN**

**Region Specific Legal Terms Translation** In reaction to Prompt 1 for both language pairs, ChatGPT-4 and Gemini 1.5 Pro translate {قانون العقوبات} and {القانون الجنائي} as ‘Penal Code’ into English and as ‘Le code pénal’ into French, so in terms of alignment with the reference translation which is ‘Criminal law’ in English and ‘Le Droit Pénal’ in French, it does not match precisely (see *Table 30*). Hence prompt 1 couldn’t instruct well the models nor do they learn from the context. Surprisingly, Prompt 2 seems to shape the behaviour of Gemini 1.5 Pro in the AR→EN/ AR→FR pairs, through injecting fundamental metadata as well as emphasising the terms to whom prominent attention must be given to. Gemini 1.5 Pro excels ChatGPT-4 in terms of alignment with the reference translations and also through generating the formal tone required in the legal language. Nevertheless, ChatGPT-4 indeed captures the legal meaning in both language pairs but lacks the use of the exact term in the target languages unlike Gemini 1.5 Pro does.

### **Non-Equivalent Entries and Unrecognised Legal Practices in Target Legal Systems.**

Prompt 1 serves as the designated legal translator for Arabic legislation, tasked with translating legal sentences strictly without additional comments. When translating the terms {طلاق بائن} and {بينونة كبرى} and {بينونة صغرى}, ChatGPT-4 performs almost more fluently and accurately in French than in English. In contrast, Gemini 1.5 Pro fails to translate the critical components {صغرى, كبرى}, resulting in a distortion of the legal meaning. However, Gemini 1.5 Pro accurately translates the term {طلاق انفائي} into French, closely matching the reference translation.

With regards to Prompt 2, both models respond well due to the inclusion of metadata such as key terms, their definitions, document names, article IDs, publication dates, last updates of the law, and references. The structured prompts, particularly Prompt 2, notably influence the

models' behaviour, enabling them to accurately translate complex legal terms without generating on word-for-word translations, despite the terms representing unfamiliar legal practices with no direct equivalents in the target language.

**MT Inconsistent Generation of terminology.** Prompt 1 is instructed to translate the legal sentences where the terms {اللعان, الفراش, البناء, الطعن} occur. Both models down perform and show no accuracy. For examples both ChatGPT-4 and Gemini 1.5 Pro transliterate the term {اللعان} as 'li'an' and literally translate the rest of the terms into the target languages.

Conversely, Prompt 2 contributes to effectively shape the models performance in both language pairs AR→EN/ AR→EN, as it plays a crucial role in maintaining the legal nuances of Sharī'ah law, consider the terms with no equivalence through the examination of both target languages, ensure that the translated terms match their usage in Islamic legal contexts and reflect the closest possible interpretation, preserve the integrity of the original legal meaning, and Avoid over-simplification or generalization.

#### 6.4. Conclusion

In the domain of AI, prompt engineering has become a transformative force, unlocking the vast potential of LLMs. It has established itself as an essential technique for optimising their performance. In this section, we present a systematic study on prompting ChatGPT-4 and Gemini 1.5 Pro for MT of Arabic legal terminology in and out of context into English and French using our labelled data. We explore a mixed-method approach ranging from incorporating zero-shot prompting, contextual prompting, and transfer learning prompting. We ensure that the designed prompts follow systematic process; (i) defining the task, (ii) evaluating the models' capabilities, including their strengths and limitations, (iii) selecting the appropriate prompt format, (iv) providing detailed contextual information, (v) demonstrating the required style and tone, and (vi) testing and refining to strengthen the context as necessary. Our findings highlight the critical role of well-designed prompts and metadata in optimising the models accuracy with regards to Arabic legal translation, demonstrating that both models rely heavily on context and structured instructions to interpret complex legal terminology and respond to the accuracy criteria required in the legal domain for Arabic.

#### **Chapter Four: Discussion and Research Limitations**

This thesis tries to carefully address the challenges posed by the MT of Arabic terminology in the legal domain into English and French due to the complexities of the Arabic legal language

and its foundations. This work proposes that MT in the legal domain for Arabic is still far from being perfect. To test and solve this problem, we ask the following research questions:

1. Why is the translation of Arabic terminology in the legal domain challenging?
2. How accurate is the automatic translation of Arabic legal terms?
3. How can we improve the MT systems performance in the legal domain for Arabic?

To start answering our research question, our first fundamental stop is conducting an initial evaluation to test our research problem. Based on the initial findings, we start to create the resources that can address the identified problems, this involves i) the development of our Arabic legal corpus that comprises a series of Arabic legislation from different Arab countries, which undergoes a preprocessing phase to clean and standardise it, ii) conduct a semi-automatic legal term candidate and concordances extraction using NooJ grammars, which were then been submitted for filtration and examination, where we delete irrelevant or inaccurate terms (i.e., terms that do not express a legal practice) and add new ones that were missed by the automatic process and choose to process 1,949 entries (70% are single terms and 30% are MWEs), iii) then we translate and validate this data, this step involves 4 critical phases; the first is locating the legal terms occurrences using NooJ grammars and concordances to capture the context, second, the use of MT systems along with some thesauri for the translation, third, double-check the results through online gateways to EU laws including EUR-Lex, IATE and Juremy, the intelligent online concordance search tool in EU law data available in all the 24 EU languages, to re-ensure the equivalence in the target languages AR→EN/ AR→EN, the fourth and last crucial stage is the validation process, where the entries and the concordances are subject to validation by our legal expert.

Initially, we conduct an extensive evaluation on the performance of MT systems across GT, ChatGPT-4, and Gemini 1.5 Pro in the legal domain for Arabic into English and French against our gold standard dataset. To do this, we use the MQM framework that can be applied to both HT, MT and AI-generated translations. The findings reveal notable strengths and weaknesses, particularly in scenarios with varying contextual complexity. GT performs better in Out-of-Context scenarios, producing fewer critical errors but often failing to capture key legal terms accurately. In contrast, ChatGPT-4 and Gemini 1.5 Pro perform better in In-Context scenarios, showing a higher capacity for maintaining contextual understanding, but they introduce more minor and major errors, such as inconsistent terminology and misinterpretation of essential legal concepts. Across all models, translating complex legal terms and culturally nuanced

content remains a significant challenge, emphasizing that while these tools can handle more straightforward translations, they are less accurate for specialized, context-sensitive domains like law. The results suggest that these models, though pre-trained on large and diverse datasets, struggle with responding to the accuracy criteria required in the legal domain for Arabic.

While other studies suggest the use of automatic metrics to evaluate MT results, such as BLEU (Papineni et al., 2002), METEOR (Lavie & Denkowski, 2009; Banerjee & Lavie, 2005), or COMET (Rey et al., 2020), we find out that human evaluation is often regarded as more accurate for several reasons. Automatic metrics rely on predefined algorithms to compare translations to a reference set, typically focusing on word or n-gram overlaps. This can fail to capture the full spectrum of translation quality, especially in terms of context, fluency, and idiomatic expressions. In contrast, human evaluation allows for subjective assessments of these nuanced factors, taking into account semantic preservation, grammatical correctness, cultural appropriateness, and overall readability. Humans can detect subtle errors, like unnatural phrasing or tone mismatches, which automated systems often overlook. Additionally, translations can be judged in terms of adequacy (how well the meaning is conveyed) and fluency (how natural the language sounds), accuracy (how faithfully the source message is preserved) which are difficult for automatic metrics to capture, as confirmed by previous studies (Killman, 2014; Loupaki, 2018; Sepesy Maučec & Donaj, 2019; Haque et al., 2019; Haque et al., 2020; Zakraoui et al., 2021; Gamal et al., 2022; He et al., 2024; Zouhar & Bojar, 2024). Therefore, our evaluation method using the MQM framework enables a granular assessment of the accuracy state of the models by considering various dimensions such as terminology, style, linguistic convention accuracy and audience appropriateness and their respective sub-dimensions. It paves the way for actionable feedback in response to the second fundamental research question.

One of our interesting experiments conducted concerns the development of ULTRA the UNIOR Legal TeRmbase for Arabic using NooJ. It represents a significant advancement in addressing the challenges of translating Arabic legislation into English and French. This resource is crucial for the morphological analysis, providing the system with metadata about word forms, inflections, and derivations. This allows for accurate POS tagging, gender, number, verb transitivity (e.g., transitive, intransitive, Naqis, etc), syntactic parsing, and semantic annotation (e.g., legal, Juri-religious, geoUsage, which indicate the geographical areas where certain terms are applicable, etc). By doing so, our resource enables NooJ to automate the processing of legal texts, applying finite-state automata for recognizing patterns

and sequences in words, which is essential for accurate linguistic annotation. Needless to mention the support it provides for the disambiguation of entries with multiple meanings (polysemy), ensuring that NooJ selects the correct sense based on the context of the text. For MT, our resource provides translations that are not only faithful to the source text but also relevant to the specific legal framework of the target audience. The lexical coverage of our resource achieves a high-quality result of recognition, precision (0.91) and recall (0.85). To the best of our knowledge, we believe that this is the first labelled and formalised resource in the legal domain for Arabic that demonstrates its ability of reducing linguistic, multilingual, and terminological ambiguities and improves the precision of the results. We look forward into testing its ability in conducting many other NLP tasks.

Regarding the performance optimisation of the LLMs' MT task, we conduct a systematic study on prompting ChatGPT-4 and Gemini 1.5 Pro for translating Arabic legal terminology into English and French, both in and out of context, using our labelled data. By leveraging a mixed-method approach that incorporates zero-shot, contextual, and transfer learning prompting, we explore how different strategies impact translation quality both in and out of context. Zero-shot prompting tests the models' baseline capabilities without additional guidance, while contextual prompting provides crucial legal background to enhance translation accuracy. Transfer learning prompting further builds on the models' prior knowledge, improving performance for domain-specific tasks. Central to our approach is the systematic process of designing prompts, which includes task definition, evaluating model strengths and limitations, selecting appropriate prompt formats, and providing detailed contextual information. These elements are vital because legal language is precise and context-dependent, making accurate translation challenging. For instance, a single Arabic legal term can have different meanings depending on its use in civil or criminal law, requiring prompts to be carefully crafted to provide adequate context. Additionally, demonstrating the required legal style and tone is essential, as legal documents often demand a formal, structured language that must be maintained across translations (Hu et al., 2022).

Our findings underscore that well-designed prompts, paired with relevant metadata, play a critical role in optimizing translation accuracy. Both ChatGPT-4 and Gemini 1.5 Pro show a heavy reliance on structured instructions to accurately interpret complex legal terminology, particularly when out-of-context terms could lead to ambiguous translations. Furthermore, the iterative process of testing and refining prompts strengthens the models ability to handle nuanced terms. This suggests that prompt engineering is not only a way for unlocking the full

potential of LLMs but also a necessity for improving performance in specialised translation tasks. The results indicate that LLMs can achieve higher levels of precision and consistency when equipped with well-designed, context-rich prompts, proving the transformative impact of prompt engineering in legal and domain-specific AI applications. These findings align with Agrawal et al. (2022), Chen et al. (2023), Zhang et al. (2023), Vatsal & Dubey (2024), Chang et al. (2024), Moslem (2024).

While this work provides valuable insights, it is important to acknowledge the limitations posed by the scarcity of dataset in the legal domain for Arabic. Developing our resource was both time-consuming and expensive due to the need for specialised expertise and extensive manual work. Nevertheless, it was more of an opportunity than a limitation, as developing this resource allowed us to pioneer a solution that not only filled an essential void but also empowered researchers, translators, and experts in the field with tools that were previously unavailable. By taking on this complex task, we have opened the door to further advancements in Arabic legal technology, setting the stage for more innovation and collaboration in the field.

## **Chapter Five: Conclusion and Future Work**

This work sets out to investigate approaches to optimise the MT systems performance in the legal domain for Arabic. It focuses on developing a monolingual Arabic corpus specific to the legal domain to create a multilingual terminological dataset for AR→EN & AR→FR translation in collaboration with a group of legal experts in the field. Using this dataset, TQE frameworks are applied to assess the accuracy, consistency, and completeness of MT systems in translating Arabic legal terms. Additionally, a multilingual legal dictionary for AR→EN & AR→FR is developed to address context-dependent translation issues and facilitate the automatic annotation and translation of Arabic legislation. Finally, strategies to enhance MT models in-context learning abilities are explored, optimizing translation quality and adherence to domain-specific terminology in Arabic legal translations.

Our findings confirm that HT, rather than automated metrics alone, provide a fuller picture of translation adequacy and fluency, especially in specialised domains. Additionally, the development of ULTRA, marks a breakthrough by offering improved linguistic annotation, disambiguation, and higher lexical precision, thereby advancing NLP applications for Arabic legal language. Leveraging prompt engineering methods was also essential for achieving better MT performance, as it enables the LLMs namely ChatGPT-4 and Gemini 1.5 Pro to produce more precise legal translations, underlining the transformative role of prompt engineering.

This work is beneficial for researchers in NLP field, translators, and specialists in the legal field. For researcher in NLP, my work offers a structured resource and frameworks that bridge linguistic theory and computational applications within a challenging language and domain. Translators and legal specialists gain access to validated multilingual termbase that improve accuracy and reduce ambiguity, facilitating reliable multilingual translation work and bridging communication gaps between Arabic-speaking and non-Arabic-speaking legal communities.

While the design and execution of our experiments were guided by a commitment to making research findings highly applicable and beneficial to real-world production scenarios, we acknowledge that there is always room for further improvement. In the future, we plan to further consider the followings:

- Expand the volume of our corpus hence the reference dataset.
- Develop an automatic system that translates the inflectional and derivational forms of our resource using NooJ grammars.
- Efficient Fine-tuning of more LLMs in the legal domain from Arabic into more languages to enhance their ability to generate and interpret Arabic legislation. This experiment should employ sufficient fine-tuning approaches (Wan et al., 2023) such as QLoRA (Hu et al., 2021; Dettmers et al., 2023).
- Leverage context integration techniques in legal AI, particularly cross-Lingual RAG (Lewis et al., 2020), where we try to integrate real-time access to legal databases and up-to-date case law. This approach enables AI systems to retrieve and generate responses based on the latest legal information, allowing legal professionals to obtain accurate, multilingual insights tailored to their specific queries. The benefits include improved accuracy in legal advice, enhanced decision-making capabilities, and efficient navigation of multilingual legal environments. However, successful implementation will require attention to data quality, ethical considerations, and user training to ensure that these tools effectively meet the needs of legal practitioners.
- We plan to also leverage RAG (Lewis et al., 2020) potentials to address LLMs hallucinations or when the models are confidently generating translations that are not accurate, by grounding their outputs from relevant and reliable data sources in the legal domain for Arabic.

## Bibliography

- Abi Fadel, P. (2004). *Dictionnaire des Termes Juridiques: Français-arabe*. Beyrouth: Librairie du Liban.
- Abid, W. (2020). The SADID Evaluation Datasets for Low-Resource Spoken Language Machine Translation of Arabic Dialects. *International Conference on Computational Linguistics*.
- Abulhassan, B. (2014). *Between English and Arabic: A practical course in translation*. Cambridge Scholars Publishing.
- Adly, N., & Al Ansary, S. (2010). Evaluation of Arabic machine translation system based on the universal networking language. In *Natural Language Processing and Information Systems: 14th International Conference on Applications of Natural Language to Information Systems, NLDB 2009, Saarbrücken, Germany, June 24-26, 2009. Revised Papers 14* (pp. 243-257). Springer Berlin Heidelberg.
- Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., & Ghazvininejad, M. (2022). In-context Examples Selection for Machine Translation. *Annual Meeting of the Association for Computational Linguistics*.
- Al Aqad, M. H. (2014). Translation of legal texts between Arabic and English: the case study of marriage contracts. *Arab World English Journal*, 5 (2), 110-121.
- Al Wahab, I. (1963). *Law dictionary: English-Arabic*. Baghdad: The National Printing and Publishing Co. W.L.L.
- Alasmary, F., Zaafarani, O., & Ghannam, A. (2024). CATT: Character-based Arabic Tashkeel Transformer. arXiv preprint arXiv:2407.03236.
- Albahūtī, M. (1636a). *Kashshāf al-qinā' 'an matn al-'iqnā'* (Vol. 3). Cairo: Almaṭba'atu Al'āmiratu Alsharqiyya.
- Albahūtī, M. (1636b). *Sharḥ Muntahā al-'irādāt* (Vol. 2). Almadīna Almunawwara: Almaktabatu Alsalafiyya.
- Alcaraz Varó, E., & Hughes, B. (2014). *Legal translation explained*.

- Alghamdi, E. A., Zakraoui, J., & Abanmy, F. A. (2023). Domain Adaptation for Arabic Machine Translation: The Case of Financial Texts. arXiv preprint arXiv:2309.12863.
- Alkhatib, M., & Shaalan, K. (2018). The key challenges for Arabic machine translation. *Intelligent Natural Language Processing: Trends and Applications*, 139-156.
- Alkhawaja, L., Ibrahim, H., Ghnaim, F., & Awwad, S. (2020). Neural machine translation: Fine-grained evaluation of Google translate output for English-to-Arabic translation. *International Journal of English Linguistics*, 10(4), 43.
- Almansor, E. H., & Al-Ani, A. (2018). A hybrid neural machine translation technique for translating low resource languages. In *Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part II 14* (pp. 347-356). Springer International Publishing.
- Al-Miftah, B., & Belefkih, A. (1998). *Lexique pluridisciplinaire: Français-arabe/arabe-français*. Madrid: Printex.
- Al-Qadhi, M. (1998). *Dictionnaire du vocabulaire juridique: Français-arabe. Traduction en arabe du Vocabulaire juridique* (G. Cornu, dir.) (1996). Beyrouth/Paris: Majd, Centre universitaire pour la recherche, l'édition et la distribution / Presses Universitaires de France.
- Alqudsi, A., Omar, N., & Shaker, K. (2014). Arabic machine translation: a survey. *Artificial Intelligence Review*, 42, 549-572.
- Al-Qurishi, M., AlQaseemi, S., & Soussi, R. (2022). AraLegal-BERT: A pretrained language model for Arabic Legal text. arXiv preprint arXiv:2210.08284.
- Al-Shehab, M. (2013). The translatability of English legal sentences into Arabic by using Google translation. *International Journal of English language and linguistics research*, 1(3), 18-31.
- Alsohybe, N. T., Dahan, N. A., & Ba-Alwi, F. M. (2017). Machine-translation history and evolution: Survey for Arabic-English translations. *arXiv preprint arXiv:1709.04685*.
- Alwazna, R. Y. (2013). Testing the precision of legal translation: The case of translating Islamic legal terms into English. *International Journal for the Semiotics of Law Revue internationale de Sémiotique juridique*, 26, 897-907.

- Alwazna, R. Y. (2019). Translation and legal terminology: Techniques for coping with the untranslatability of legal terms between Arabic and English. *International Journal for the Semiotics of Law-*Revue internationale de Sémiotique juridique**, 32, 75-94.
- Alwazna, R. Y. (2022). The Efficacy of the Integrative Model Proposed by Prieto Ramos (2014) in Surmounting Terminological Problems of Arabic-English Legal Translation. *International Journal for the Semiotics of Law-*Revue internationale de Sémiotique juridique**, 35(2), 737-752.
- Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federmann, C., Genzel, D., Guzmán, F., Hu, J., Hughes, M., Koehn, P., Lazar, R., Lewis, W., Neubig, G., Niu, M., Oktem, A., Paquin, E., Tang, G., and Tur, S. (2020). TICO-19: the Translation Initiative for Covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.
- Anderman, G. M., & Rogers, M. (Eds.). (2005). *In and out of English: for better, for worse? (Vol. 1). Multilingual matters.*
- Asensio, R. M. (2007). Specialised translation: A concept in need of revision. *Babel*, 53.
- Asensio, R. M. (2014). *Translating official documents.* Routledge.
- Attia, M. A. (2008). *Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation.* The University of Manchester (United Kingdom).
- Avramidis, E., Macketanz, V., Strohriegel, U., Burchardt, A., & Möller, S. (2020). Fine-grained linguistic evaluation for state-of-the-art machine translation. *arXiv preprint arXiv:2010.06359.*
- AWS. (2024). *Prompt engineering for foundation models.* AWS Documentation. <https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart-foundation-models-customize-prompt-engineering.html>.
- Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016, March). End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4945-4949). IEEE.

- Baker, M. (1992). A coursebook on translation. *London and New York: Routledge*.
- Balboni, P. (1986). LGP versus LSP: Which Way to The Razor's Edge?. *Unesco Alsed-LSP Newsletter (1977-2000)*, 9(1).
- Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65-72).
- Bapna, A., Arivazhagan, N., & Firat, O. (2019). Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*.
- Ben Mourad, I. (1997). *Muqadima li nadhariyat al-mu'jam [Introduction aux théories lexicographiques]*. Beyrouth: Dar Al Gharb Al Islami.
- Benhima, M., & Elimadi, I. (2023, March). The use of archaic terms by legal translators: Arabic into English translation. In *Conference Language and Power* (p. 133).
- Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.
- Berrichi, S., & Mazroui, A. (2021). Addressing limited vocabulary and long sentences constraints in English–Arabic neural machine translation. *Arabian Journal for Science and Engineering*, 46(9), 8245-8259.
- Berzins, A., Choukri, K., Giagkou, M., Lösch, A., Mazo, H., Piperidis, S., ... & Zumrik, M. (2019). Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe-Why Language Data Matters: ELRC White Paper. *ELRC Consortium*, 2, 6-12.
- Beyer, A. M., Macketanz, V., Burchardt, A., & Williams, P. (2017). Can out-of-the-box NMT Beat a Domain-trained Moses on Technical Data. *Proceedings of EAMT user studies and project/product descriptions*, 41-46.
- Bies, A., DiPersio, D., & Maamouri, M. (2012). Linguistic resources for Arabic machine translation. *Challenges for Arabic Machine Translation*, 9, 15.

- Bouamor, H., Alshikhabobakr, H., Mohit, B., & Oflazer, K. (2014, October). A human judgement corpus and a metric for Arabic MT evaluation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 207-213).
- Briva-Iglesias, V., Camargo, J. L. C., & Dogru, G. (2024). Large Language Models "Ad Referendum": How Good Are They at Machine Translation in the Legal Domain?. arXiv preprint arXiv:2402.07681.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems (Article 159, pp. 1–25). Curran Associates Inc.
- Buckwalter, T. 2004. Issues in Arabic orthography and morphological analysis. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.
- Burchardt, A. (2013). Multidimensional quality metrics: a flexible system for assessing translation quality. In Proceedings of Translating and the Computer 35.
- Cabré, M. T. (1999). Terminology: Theory, methods, and applications (Vol. 1). John Benjamins Publishing.
- Cabré, T. (2023). Terminology: Cognition, language and communication. Cambridge University Press.
- Cao, D. (2007). Translating law (Vol. 33). Multilingual Matters.
- Castilho, S., Cavalheiro Camargo, J. L., Menezes, M., & Way, A. (2021, November). DELA corpus-a document-level corpus annotated with context-related issues. Association for Computational Linguistics (ACL).
- Castilho, S., Mallon, C., Meister, R., & Yue, S. (2023). Do online machine translation systems care for context? What about a GPT model?.

- Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, P., & Gialama, M. (2017). A comparative quality evaluation of PBSMT and NMT using professional translators. In *Proceedings of Machine Translation Summit XVI: Research Track* (pp. 116-131).
- Catford. (1965). *A Linguistic theory of Translation*. London: Oxford University Press.
- Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., & Federico, M. (2015). The IWSLT 2015 Evaluation Campaign. *International Workshop on Spoken Language Translation*.
- Chamel, B., Bechara, C., & Rahmé, R. (2010). *Lexique des Termes Juridiques: Français-arabe-anglais. Traduction en arabe du Lexique des termes juridiques (2009)*. Paris/Liban: Dalloz; Hachette-Antoine.
- Chang, K., Xu, S., Wang, C., Luo, Y., Xiao, T., & Zhu, J. (2024). Efficient Prompting Methods for Large Language Models: A Survey. *ArXiv*, abs/2404.01077.
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.
- Chen, B., Zhang, Z., Langren'e, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *ArXiv*, abs/2310.14735.
- Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.
- Chromá, M. (2011). Synonymy and polysemy in legal terminology and their applications to bilingual and bijural translation. *Research in language*, 9 (1), 31-50.
- Chu, C., & Wang, R. (2020). A survey of domain adaptation for machine translation. *Journal of information processing*, 28, 413-426.
- Clark, J. H., Lavie, A., & Dyer, C. (2012). One system, many domains: Open-domain statistical machine translation via feature augmentation. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*.

- Comparin, L., & Mendes, S. (2017). Using error annotation to evaluate machine translation and human post-editing in a business environment. Proceedings of EAMT 2017, Prague, May 29, 31.
- Constant, M., Eryiğit, J., Monti, J., Plas, L.V., Ramisch, C., Rosner, M., & Todirascu-Courtier, A. (2017). Survey: Multiword Expression Processing: A Survey. *Computational Linguistics*, 43, 837-892.
- Darwish, A. (2009). *Terminology and translation: A phonological-semantic approach to Arabic terminology*. Writescop Publishers.
- Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H.S., Al-Natsheh, H.T., El-Beltagy, S.R., Bouamor, H., Bouzoubaa, K., Cavalli-Sforza, V., El-Hajj, W., Jarrar, M., & Mubarak, H. (2020). A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, 64, 72 - 81.
- Deng, W. H., Mehandru, N., Robertson, S., & Salehi, N. (2022). Beyond General Purpose Machine Translation: The Need for Context-specific Empirical Research to Design for Appropriate User Trust. arXiv preprint arXiv:2205.06920.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- Dougal, D.K., & Lonsdale, D.W. (2020). Improving NMT Quality Using Terminology Injection. *International Conference on Language Resources and Evaluation*.
- Douglas, D. (2001). Language for Specific Purposes assessment criteria: where do they come from?. *Language Testing*, 18(2), 171-185.
- Dr. Ali Al Qasimi's efforts in the historical dictionary methodology: Through his membership in the two scientific committees for the two historical dictionary projects of the Arabic language in (Doha) and (Sharjah). (2024). *Al-Adab Journal*, 149, 123-144.
- El Marouani, M., Boudaa, T., & Enneya, N. (2020, December). Machine Translation Evaluation using Textual Entailment for Arabic. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1-5). IEEE.

- El-Farahaty, H. (2010). Lexical and syntax features of English and Arabic legal discourse: a comparative study. *Comparative Legilinguistics*, 4, 61-80.
- El-Farahaty, H. (2015). *Arabic-English-Arabic Legal Translation*. Routledge.
- El-Farahaty, H. (2016). Translating lexical legal terms between English and Arabic. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 29, 473-493.
- ElFqih, K. A., & Monti, J. (2023, September). On the Evaluation of Terminology Translation Errors in NMT and PB-SMT In the Legal Domain: A Study on the Translation of Arabic Legal Documents into English and French. In *Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC)* (pp. 26-35).
- Engberg, J. (2013). Comparative law for translation: The key to successful mediation between legal systems. In *Legal translation in context: Professional issues and prospects* (pp. 9-25). Peter Lang.
- Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Çelebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2020). Beyond English-Centric Multilingual Machine Translation. *J. Mach. Learn. Res.*, 22, 107:1-107:48.
- Faiq, S. (Ed.). (2004). *Cultural encounters in translation from Arabic* (Vol. 26). *Multilingual Matters*.
- Faruqi, H. S. (1969/2008). *Faruqi's law dictionary: English-Arabic*. Beyrouth: Librairie du Liban Publishers.
- Forcada, M. L., Sánchez-Martínez, F., Esplà-Gomis, M., & Specia, L. (2017). Towards Optimizing MT for Post-Editing Effort: Can BLEU Still Be Useful?.
- Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2022). Complexity-Based Prompting for Multi-Step Reasoning. *ArXiv, abs/2210.00720*.

- Galfan, M. (1998). Al mu'jam al muwahad li mustalahāt al-lisānīyāt: Ayyi mustalah li ayyī lisānīyāt? [Dictionnaire unifié des termes linguistiques; quel terme pour quelle linguistique?]. *Majalat Al Lisan Al Arabi / Revue de la langue arabe*, 46.
- Gambier, Y. & Doorslaer, L. V. (2010). *Handbook of Translation Studies*. Amsterdam: John Benjamins B.V.
- Gao, T., Fisch, A., & Chen, D. (2021). Making Pre-trained Language Models Better Few-shot Learners. *Annual Meeting of the Association for Computational Linguistics*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805,2023.
- Habash, N., Zalmout, N., Taji, D., Hoang, H., & Alzate, M. (2017, April). A parallel corpus for evaluating machine translation between Arabic and European languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 235-241).
- Hadla, L. S., Hailat, T. M., & Al-Kabi, M. N. (2014). Evaluating Arabic to English machine translation. *International Journal of Advanced Computer Science and Applications*, 5(11).
- Hakki, M. (1982). *Dictionnaire des termes juridiques et commerciaux: Français-arabe*. Beyrouth: Librairie du Liban.
- Halimi, M. I. S. (2023). Bilingual legal resources for Arabic: State of affairs and future perspectives. *International Journal for the Semiotics of Law*, 15. <https://doi.org/10.1007/s11196-023-10059-2>
- Halimi, S. A. (2017). Contextualizing translation decisions in legal system-bound and international multilingual contexts: French-Arabic criminal justice terminology. *Translation and Translanguaging in Multilingual Contexts*, 3(1), 20-46.
- Halimi, S. A. (2019). Rethinking the English–Arabic Legal Translation Course: Restructuring for Specific Competence Acquisition. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 32(1), 117-134.

- Halimi, S. A., & Alwazna, R. Y. (2023). Issues addressed in Arabic legal translation: a future perspective. In *Research Handbook on Jurilinguistics* (pp. 436-450). Edward Elgar Publishing.
- Halliday, M.A.K. (1994). *An Introduction to Functional Grammar*. 2nd ed. London: Edward Arnold.
- Han, L. (2016). Machine translation evaluation resources and methods: A survey. arXiv preprint arXiv:1605.04515.
- Hanford M (2011). The Language for Business Meetings. *English for Specific Purposes* 30 313-314.
- Haque, R., Hasanuzzaman, M., & Way, A. (2019). TermEval: An automatic metric for evaluating terminology translation in MT.
- Haque, R., Hasanuzzaman, M., & Way, A. (2019, September). Investigating terminology translation in statistical and neural machine translation: a case study on English-to-Hindi and Hindi-to-English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 437-446).
- Haque, R., Hasanuzzaman, M., & Way, A. (2020). Analysing terminology translation errors in statistical and neural machine translation. *Machine Translation*, 34, 149-195.
- Hatim, B. (2009). Translating text in context. In *The Routledge companion to translation studies* (pp. 50-67). Routledge.
- Hayati, A. M. (2008). Teaching English for special purposes in Iran: Problems and suggestions. *Arts and Humanities in Higher Education*, 7(2), 149-164.
- He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., ... & Wang, X. (2024). Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12, 229-246.
- He, Z., Wang, X., Jiao, W., Zhang, Z., Wang, R., Shi, S., & Tu, Z. (2024). Improving machine translation with human feedback: An exploration of quality estimation as a reward model. arXiv preprint arXiv:2401.12873.

- Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., ... & Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210.
- Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., ... & Xu, H. (2024). Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, ocad259.
- Hu, Z., Lee, R. K. W., Aggarwal, C. C., & Zhang, A. (2022). Text style transfer: A review and experimental evaluation. *ACM SIGKDD Explorations Newsletter*, 24 (1), 14-45.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
- Ibn Qudāmah, M. (1997). *Al-Mughnī* (Vol. 6, 3rd ed.). Riyadh: Dāru ‘Ālami Alkutubi Lilṭibā‘ati Walnashri Waltawzī‘.
- Id-Youss, H., Steurs, F., & Alswlaiman, A. A. (2018). Legal-based Ontologies Between serious needs and challenging realities. In *TOTh 2017 Proceedings* (pp. 249–264). University of Savoie; Chambéry.
- Jabak, O. (2019). Assessment of Arabic-English translation produced by Google translate. *International Journal of Linguistics, Literature and Translation (IJLLT) ISSN*, 2617-0299.
- Jean, S., Firat, O., Cho, K., Memisevic, R., & Bengio, Y. (2015, September). Montreal neural machine translation systems for WMT’15. In *Proceedings of the tenth workshop on statistical machine translation* (pp. 134-140).
- Jehl, L., Gispert, A., Hopkins, M., & Byrne, B. (2014). Source-side Preordering for Translation using Logistic Regression and Depth-first Branch-and-Bound Search. *Conference of the European Chapter of the Association for Computational Linguistics*.
- Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W.E. (2023). Mistral 7B. ArXiv, abs/2310.06825.
- Jopek-Bosiacka, A. (2018). Theoretical and logical prerequisites for legal translation. *JLL*, 7, 47.

- Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition—a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, 1(67), 1.
- Kamir, D., Soreq, N., & Neeman, Y. (2002, July). A comprehensive NLP system for modern standard Arabic and modern Hebrew. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *ArXiv, abs/2001.08361*.
- Khoshafah, F. (2023). ChatGPT for Arabic-English translation: Evaluating the accuracy. doi:10.21203/rs.3.rs-2814154/v1. PPR:PPR645536.
- Killman, J. (2014). Vocabulary accuracy of statistical machine translation in the legal context. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas* (pp. 85-98).
- Kinoshita, S., Oshio, T., & Mitsuhashi, T. (2017, November). Comparison of SMT and NMT trained with large Patent Corpora: Japio at WAT2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)* (pp. 140-145).
- Klubička, F., Toral, A., & Sánchez-Cartagena, V. M. (2017). Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1), 121.
- Klubička, F., Toral, A., & Sánchez-Cartagena, V. M. (2018). Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. *Machine Translation*, 32(3), 195-215.
- Koehn, P. (2020). *Neural machine translation*. Cambridge University Press.
- Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.
- Lavie, A., & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine translation*, 23, 105-115.

- Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A Survey on Evaluation Metrics for Machine Translation. *Mathematics*, 11(4), 1006.
- Lewis, B. (1980). Translation from Arabic. *Proceedings of the American Philosophical Society*, 124(1), 41-47.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Annual Meeting of the Association for Computational Linguistics*.
- Lin, Z., Pan, X., Wang, M., Qiu, X., Feng, J., Zhou, H., & Li, L. (2020). Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information. *Conference on Empirical Methods in Natural Language Processing*.
- Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). What Makes Good In-Context Examples for GPT-3? *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out*.
- Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8, 726-742.
- Long, Z., Utsuro, T., Mitsuhashi, T., & Yamamoto, M. (2017). Translation of patent sentences with a large vocabulary of technical terms using neural machine translation. *arXiv preprint arXiv:1704.04521*.
- Loupaki, E. (2018). EU Legal Language and Translation-Dehumanizing the Refugee Crisis. *JLL*, 7, 97.
- Lu, S., Chen, Z., & Xu, B. (2014, June). Learning new semi-supervised deep auto-encoder features for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 122-132).
- Macketanz, V., Avramidis, E., Burchardt, A., Helcl, J., & Srivastava, A. (2017). Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation. *Cybernetics and Information Technologies*, 17(2), 28-43.

- Madi, N., & Al-Khalifa, H. (2020). Error detection for Arabic text using neural sequence labeling. *Applied Sciences*, 10(15), 5279.
- Mahraj, M., & Hdouch, Y. (2020). Lexical issues in English-Arabic legal translation. *IOSR Journal of Humanities and Social Science*, 25(11), 9-16.
- Marchisio, K., Duh, K., & Koehn, P. (2020). When Does Unsupervised Machine Translation Work?. arXiv preprint arXiv:2004.05516.
- Matulewska, A. (2016). Semantic relations between legal terms. A case study of the intralingual relation of synonymy. *Studies in logic, grammar and rhetoric*, 45(1), 161-174.
- Mediouni, M. (2016). Towards a functional approach to Arabic-English legal translation: The role of comparable/parallel texts. *New insights into Arabic translation and interpreting*, 115-160.
- Mesfar, S. (2008). Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard (Doctoral dissertation, Besançon).
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4), 235-244.
- Monti, Johanna and Nolano, Gennaro (2019). Il Multidimensional Quality Metrics nell'analisi qualitativa della traduzione automatica. Project Report. Università degli Studi di Napoli "L'Orientale", Napoli.
- Moslem, Y. (2024). Language Modelling Approaches to Adaptive Machine Translation. *ArXiv, abs/2401.14559*.
- Moslem, Y., Haque, R., & Way, A. (2023). Fine-tuning Large Language Models for Adaptive Machine Translation. arXiv preprint arXiv:2312.12740.
- Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023). Adaptive machine translation with large language models. arXiv preprint arXiv:2301.13294.
- Moussa, F. (1993). *Lexique des termes juridiques: Français-arabe*. Tunis: Orbis Impression.
- Müller, M., Rios Gonzales, A., & Sennrich, R. (2019). Domain Robustness in Neural Machine Translation. ArXiv, abs/1911.03109.

- Munday, J. (2010). Translation studies. *Handbook of translation studies*, 1, 419-428.
- Najar, D., Mesfar, S., & Ghezela, H. B. (2016). A large terminological dictionary of Arabic compound words. In *Automatic Processing of Natural-Language Electronic Texts with NooJ: 9th International Conference, NooJ 2015, Minsk, Belarus, June 11-13, 2015, Revised Selected Papers 9* (pp. 16-28). Springer International Publishing.
- Newmark, P. (1988). Pragmatic translation and literalism. *TTR: traduction, terminologie, rédaction*, 1(2), 133-145.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).
- Paula, Zorrilla-Agut., Thierry, Fontenelle. (2019). IATE 2: Modernising the EU's IATE terminological database to respond to the challenges of today's translation world and beyond. *Terminology*, 25(2):146-174. doi: 10.1075/TERM.00034.ZOR
- Picht, H., & Draskau, J. Kewley. (1985). *Terminology: an introduction*. Guildford: University of Surrey, Dep. of Linguistic and International Studies.
- Poibeau, T. (2017). *Machine translation*. MIT Press.
- Poibeau, T. (2022, June). On "Human Parity" and "Super Human Performance" in Machine Translation Evaluation. In *Language Resource and Evaluation Conference*.
- Popović, M. (2015, September). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation* (pp. 392-395).
- Popović, M. (2017). Comparing language related issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics*, 108(1), 209.
- Prieto Ramos, F. (2011). Developing legal translation competence: An integrative process-oriented approach. *Comparative legilinguistics*, 5, 7-22.
- Prieto Ramos, F. (2020a). The use of resources for legal terminological decision-making: Patterns and profile variations among institutional translators. *Perspectives*, 29 (2), 278–310. <https://doi.org/10.1080/0907676X.2020.1803376>

- Prieto Ramos, F., & Cerutti, G. (2021). Terminology as a source of difficulty in translating international legal discourses: an empirical cross-genre study. *International Journal of Legal Discourse*, 6, 155 - 179.
- Prieto Ramos, F., & Guzmán, D. (2023). Translating Legal Terms at International Organisations: Do Institutional Term Banks Meet Translators' Needs?. In *New Advances in Legal Translation and Interpreting* (pp. 243-264). Singapore: Springer Nature Singapore.
- Publication Office of the European Union (2023). IATE the User's Handbook. October 1, 2023, <https://iate.europa.eu/assets/handbook.pdf>
- Rashed Alkatheery, E. (2023). Google Translate Errors in Legal Texts: Machine Translation Quality Assessment. *AWEJ for Translation & Literary Studies*, 7.
- Rei, J., van der Lee, C., Stojanovski, J., & Duh, K. (2020). COMET: A Contextualized Embedding Metric for Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.
- Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., ... & Martins, A. F. (2022). CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. arXiv preprint arXiv:2209.06243.
- Ren, S., Chen, W., Liu, S., Li, M., Zhou, M., & Ma, S. (2018). Triangular architecture for rare language translation. arXiv preprint arXiv:1805.04813.
- Rosenhouse, J. (1988) 'Occurrences of the Passive in Different Types of Texts in English, Hebrew and Arabic'. *Babel*, vol. 34 (2): pp. 90–103.
- Roventini, A., Alonge, A., Bertagna, F., Calzolari, N., Zampolli, A., Girardi, C., Magnini, B., Marinelli, R., Speranza, M., & Cancila, J. (1998). "ItalWordNet" : Building a Large Semantic Database for the Automatic Treatment of Italian.
- Sager, J. C. (1990). *Practical course in terminology processing*. John Benjamins Publishing.
- Sager, J. C., Dungworth, D., & McDonald, P. F. (1980). *English special languages: principles and practice in science and technology*. Wiesbaden: Brandstetter.

- Sagri, M. T., Tiscornia, D., & Bertagna, F. (2004). Jur-wordnet. In Proceedings of the 2nd International Global Wordnet Conference (pp. 305-310).
- Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.
- Sandrini, P. (1996). Comparative analysis of legal terms: Equivalence revisited. *Terminology and knowledge engineering*, 96.
- Sarcevic, S. (1997). New approach to legal translation. Kluwer Law International BV.
- Šarcevic, S. (2000, February). Legal translation and translation theory: A receiver-oriented approach. In International Colloquium, 'Legal translation, theory/ies, and practice', University of Geneva (pp. 17-19).
- Šarčević, S. (2016). Language and culture in EU law: Multidisciplinary perspectives. Routledge.
- Schwenk, H. (2012, December). Continuous space translation models for phrase-based statistical machine translation. In Proceedings of COLING 2012: Posters (pp. 1071-1080).
- Sepesy Maučec, M., & Donaj, G. (2019). Machine translation and the evaluation of its quality. In A. Sadollah & S. Tilendra (Eds.), *Recent Trends in Computational Intelligence*. IntechOpen. <http://dx.doi.org/10.5772/intechopen.89063>.
- Shirko, O., Omar, N., Arshad, H., & Albared, M. (2010). Machine translation of noun phrases from Arabic to English using transfer-based approach. *Journal of Computer Science*, 6(3), 350.
- Shiyab, S. (2019). The dilemma of legal terminology in the Arab world. *Handbook of Terminology*, 2, 139-154.
- Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'dowd, T., & Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32(3), 217-235.

- Shunnaq, A. T. (2006). Semantic Redundance and Translation from Arabic into English. In *CIUTI-Forum*.
- Silberztein, M. (2003). NooJ Manual. Available at: <https://atishs.univ-fcomte.fr/nooj/downloads.html>
- Silberztein, M. (2016). *Formalizing natural languages: The NooJ approach*. John Wiley & Sons.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers* (pp. 223-231).
- Sorensen, T., Robinson, J., Rytting, C., Shaw, A., Rogers, K., Delorey, A., Khalil, M., Fulda, N., & Wingate, D. (2022). An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.60>
- Specia, L., Harris, K., Blain, F., Burchardt, A., Macketanz, V., Skadin, I., ... & Turchi, M. (2017). Translation quality and productivity: A study on rich morphology languages. In *Proceedings of Machine Translation Summit XVI: Research Track* (pp. 55-71).
- Štajner, S., Querido, A., Rendeiro, N., Rodrigues, J. A., & Branco, A. (2016, May). Use of domain-specific language resources in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 592-598).
- Tang, Y., Tran, C., Li, X., Chen, P. J., Goyal, N., Chaudhary, V., ... & Zettlemoyer, L. (2020). mBART-50: Multilingual Translation with Multilingual Pre-training. arXiv preprint arXiv:2010.12893.
- team, N., Costa-jussà, M.R., Cross, J., cCelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G.M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K., Rowe, D., Spruit, S.L., Tran, C., Andrews, P.Y., Ayan, N.F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzm'an, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S.,

- Schwenk, H., & Wang, J. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. *ArXiv*, abs/2207.04672.
- Thelen, M. (2002b). Relations between terms: a cognitive approach. The interaction between Terminology, Lexicology, Translation Studies and translation practice. *Linguistica Antverpiensia, New Series—Themes in Translation Studies*, 1.
- Thelen, M. (2012). The structure of the lexicon: Incorporating a cognitive approach in the TCM lexicon, with applications to lexicography, terminology and translation. Academia Press.
- Thelen, M. (2015). The interaction between terminology and translation or where terminology and translation meet. *Trans-kom*, 8(2), 347-381.
- Thelen, M., & van der Vliet, H. (2008). *Perspectieven voor de Nederlandstalige Terminologie [Perspectives for Dutch terminology (MT)]*.
- Tiedemann, J. (2012, May). Parallel data, tools and interfaces in OPUS. In *Lrec* (Vol. 2012, pp. 2214-2218).
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *ArXiv*, abs/2305.14314.
- Tonmoy, S.M., Zaman, S.M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *ArXiv*, abs/2401.01313.
- Toral, A., & Way, A. (2018). What level of quality can neural machine translation attain on literary text?. *Translation quality assessment: From principles to practice*, 263-287.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, abs/2302.13971.
- Trad, F., & Chehab, A. (2024). Prompt engineering or fine-tuning? a case study on phishing detection with large language models. *Machine Learning and Knowledge Extraction*, 6(1), 367-384.

- Tufano, M., Watson, C., Bavota, G., Di Penta, M., White, M., & Poshyvanyk, D. (2018, September). An empirical investigation into learning bug-fixing patches in the wild via neural machine translation. In Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (pp. 832-837).
- Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Neural Information Processing Systems*.
- Vatsal, S., & Dubey, H. (2024). A survey of prompt engineering methods in large language models for different nlp tasks. *arXiv preprint arXiv:2407.12994*.
- Vintar, Š. (2018, May). Terminology Translation Accuracy in Phrase-Based versus Neural MT: An Evaluation for the English-Slovene Language Pair. In LREC 2018 Workshop (p. 34).
- Wagner, A. (2003). Translation of the language of the common law into legal French: myth or reality. *International Journal for the Semiotics of Law*, 16(2), 177-193.
- Wagner, A., Sin, K. K., & Cheng, L. (2016). Cultural transfer and conceptualization in legal discourse. In *The Ashgate Handbook of legal translation* (pp. 27-42). Routledge.
- Wan, Z., Wang, X., Liu, C., Alam, S., Zheng, Y., Liu, J., ... & Zhang, M. (2023). Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *ArXiv, abs/2206.07682*.
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2020). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *North American Chapter of the Association for Computational Linguistics*.
- Yamagishi, H., Kanouchi, S., Sato, T., & Komachi, M. (2016, December). Controlling the voice of a sentence in Japanese-to-English neural machine translation. In Proceedings of the 3rd Workshop on Asian Translation (WAT2016) (pp. 203-210).
- Zakraoui, J., Saleh, M., Al-Maadeed, S., & AlJa'am, J. M. (2020, April). Evaluation of Arabic to English machine translation systems. In 2020 11th International conference on information and communication systems (ICICS) (pp. 185-190). IEEE.

- Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2021). Arabic Machine Translation: A Survey With Challenges and Future Directions. *IEEE Access*, 9, 161445-161468.
- Zhang, B., Haddow, B., & Birch, A. (2023, July). Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning* (pp. 41092-41110). PMLR.
- Zhang, Y., Feng, S., & Tan, C. (2022). Active Example Selection for In-Context Learning. *ArXiv*, *abs/2211.04486*.
- Zhou, D., Scharli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Bousquet, O., Le, Q., & Chi, E.H. (2022). Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *ArXiv*, *abs/2205.10625*.
- Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations parallel corpus v1. 0. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 3530–3534.
- Zouhar, V., & Bojar, O. (2024). Quality and Quantity of Machine Translation References for Automated Metrics. *arXiv preprint arXiv:2401.01283*.

## Annexes A

### A1. Properties definitions dictionary « Properties.def »

```
# NooJ V1

# Dictionary properties' definition

# Language is: ar

# Special Characters: '=' '+' '#' ''

# List categories and properties associated with features

# Example: N_Number = m + f;

# Special KEYWORD: INFLECTION lists all inflectional features (used by variables $xF)

# Example: INFLECTION = m + f + Present + Futur;
```

INFLECTION = m + f + s + p +

1 + 2 + 3 + d + 5 +

P + I + S + C +

A + K +

Y + F;

ADJ\_Genre = m + f;

ADJ\_Cas = a + u + i + an + un + in;

ADJ\_Nombre = s + p;

N\_Genre = m + f;

N\_Nombre = s + d + p;

N\_Cas = a + u + i + an + un + in;

N\_Synt = PartActif + PartPassif;

N\_Sem = Ville + Hum + LOC + HumColl + AsmaKhamssa + Conc + Anim + Abst + Pol + PR;

DOM = Juri\_religion + legal + eco + Sport + Chimie + Alim;

GeoUsage = MAC + QA + EG + MAR + TN + DZ + SA + UAE + Monarchies;

PRON\_Genre = m + f;

PRON\_Nombre = s + d + p;

ADV\_Cas = a + u + i;

PREP\_Cas = a + u + i;

NSUFF\_Cas = a + u + i;

NSUFF\_Nombre = s + d + p;

DET\_Genre = m + f;

DET\_Nombre = s + p;

V\_Pers = 1 + 2 + 3 + d + 5;

V\_Voix = A + K;

V\_Temps = P + I + S + C + Y + F;

V\_Synt = Tr + TrInd + Intr + Naaqis + Kaana + Kaada;

V\_Genre = m + f;

V\_Nombre = s + p;

## **A2. Extract from the inflectional and derivational dictionary «ARLD2.nof»**

# NooJ V7

# Inflectional/Derivational description

# Language is: ar

# Generic Commands:

# <B>: keyboard Backspace

# <C>: change Case

# <D>: Duplicate current char

# <E>: Empty string

# <L>: keyboard Left arrow



يَينَ/d+i);

F2=

يَينَ/d+a |

انَ/d+u |

يَينَ/d+i ;

FlexionPL=

اَ/a |

ُ/u |

ِ/i |

اَ/an |

ُ/un |

ِ/in;

Flx1=

#Singular masculin forms

اَ/m+s+a |

ُ/m+s+u |

ِ/m+s+i |

اَ/m+s+an |

ُ/m+s+un |

ِ/m+s+in |

#Les formes au féminin singulier

(ة)

(/f+s+a |

ُ/f+s+u |

ِ/f+s+i |

َّ/f+s+an |

ُّ/f+s+un |

ِ/f+s+in) |

#Les formes duales et plurielles

َيْنَ/m+d+a |

َانَ/m+d+u |

َيْنَ/m+d+i |

َيَ/m+d+a |

َا/m+d+u |

َيَ/m+d+i |

ِينَ/m+p+a |

ُونَ/m+p+u |

ِينَ/m+p+i |

ِي/m+p+a |

ُوا/m+p+u |

ِي/m+p+i |

وُ/m+p+u |

تَيْنَ/m+d+a |

تَانِ/m+d+u |

تَيْنِ/m+d+i |

اتِ/f+p+a |

اتُ/f+p+u |

اتِ/f+p+i |

اتِ/f+p+an |

اتُ/f+p+un |

اتِ/f+p+in;

#####

# Derivation

#####

FlxDRV=

#Les formes au masculin singulier

اَ/m+s+a |

اُ/m+s+u |

اِ/m+s+i |

اَّ/m+s+an |

اُّ/m+s+un |

مِ/m+s+in |

#Les formes au féminin singulier

(ة)

(/f+s+a |

مُ/f+s+u |

مِ/f+s+i |

مُ/f+s+an |

مُ/f+s+un |

مِ/f+s+in) |

#Les formes duales et plurielles

مَيْنَ/m+d+a |

مَانَ/m+d+u |

مَيْنَ/m+d+i |

مَيْنَ/m+d+a |

مَانَ/m+d+u |

مَيْنَ/m+d+i |

مَيْنَ/m+p+a |

مُونُ/m+p+u |

مَيْنَ/m+p+i |

مَيْنَ/m+p+a |

مُونُ/m+p+u |

مَيْنَ/m+p+i |

وُ/m+p+u |

نَيْنَ/f+d+a |

نَانِ/f+d+u |

نَيْنِ/f+d+i |

اتِ/f+p+a |

اتُ/f+p+u |

اتِ/f+p+i |

اتِ/f+p+an |

اتُ/f+p+un |

اتِ/f+p+in;

#####

# Noms composés

#####

FlexionNC1=

◌<P>/a |

◌<P>/u |

◌<P>/i;

#####

# Nouns

#####

Oqoudon=

<LW><R><R>ُ<R><SW>/N+p;

MawaniI=

<LW><R>ُ<R><R><R><SW>/N+p;

Wokalaa=

<LW><R><R><S><R>ء<SW>/N+p;

Mawad=

<LW><R>ُ<R><R><S><SW>/N+p;

Qadaya=

<LW><R><R>|<R>|<S><SW>/N+p;

Masatir=

<LW><R>|<R><R><S><SW>/N+p;

Shikayat=

<LW><R><R><R><R>|<S>ت<SW>/N+p;

Mahakim=

<LW><R><R>|<R><R><S><SW>/N+p;

Ahkaam=

<LW>أ<R><R>|<R><SW>/N+p;

Qonsoliaat=

<LW><R><R><R><R>|<S>ت<SW>/N+p;

Wataaiq=

<LW><R><R>ءى<S><R><S><SW>/N+p;

Bonoudon=

<LW><R><R>و<R><SW>/N+p;

Shahadaat=

<LW><R><R><R><R>ء<S><SW>/N+p;

Jonah=

<LW><R><R><R><S><SW>/N+p;

Qodaa=

<LW><R><S>ض<S>ء<R>ء<SW>/N+p;

Morafa'at=

<LW><R><R><R><R><R>ء<S><SW>/N+p;

Jaraim=

<LW><R><R>ء<S>ء<R><S><SW>/N+p;

Dala'îl=

<LW><R><R>ء<S>ء<R><SW>/N+p;

Waqalî=

<LW><R><S><R>ء<R><S><SW>/N+p;

Tadabiir=

<LW><R><R>ء<R><R><R><SW>/N+p;

### **A3. An Extract from the UNIOR Legal TeRmbase for Arabic (ULTRA)**

# NooJ V7

# Dictionary

#

```

# Language is: ar

#

# Alphabetical order is not required.

#

# Use inflectional & derivational paradigms' description files (.nof), e.g.:

# Special Command: #use paradigms.nof

#

# Special Features: +NW (non-word) +FXC (frozen expression component) +UNAMB
(unambiguous lexical entry)

#           +FLX= (inflectional paradigm) +DRV= (derivational paradigm)

#

# Special Characters: '\ '""''','+'-'! '#'

#

#use ARLD2.nof

```

صداق,N+Genre=m+DOM=legal+GeoUsage=MAC+EN=dowry+FR=dote

عدل,N+FLX=F1+DOM=legal+GeoUsage=MAR+EN=superintendent registrar+FR=maire

مأذون,N+Genre=m+Nombre=s+DOM=legal+Sem=Hum+GeoUsage=QA\_EG+EN=superinten  
dent registrar+FR=maire

قسم,N+Genre=m+Nombre=s+DOM=legal+GeoUsage=MAC+EN=department+FR=departme  
nt

اللعان,N+DOM=Juri\_religion+GeoUsage=MAC+EN=accusation+FR=accusation

القطع,N+DOM=Juri\_religion+GeoUsage=MAC+EN=certainty+FR=certitude

الطعن,N+DOM=Juri\_religion+GeoUsage=MAC+EN=rebuttal+FR=contre-preuve

فراش, N+DOM=Juri\_religion+GeoUsage=MAC+EN=marriage

consummation+FR=consommation du mariage

خبير, N+FLX=Flx1+DOM=legal+sem=Hum+GeoUsage=MAC+EN=expert+FR=expert

قاصر, N+f+s+legal+Hum+GeoUsage=MAC+EN=Under-age+FR=Mineur

عقد, N+Genre=m+Nombre=S+FLX=Oqoudon+DOM=legal+GeoUsage=MAC+EN=Contract+FR=Contrat

زواج, N+Nombre=s+DOM=legal+GeoUsage=MAC+EN=marriage+FR=mariage

ممانع, N+s+FLX=Madariss+DOM=legal+GeoUsage=MAC+EN=Obstacle+FR=Empêchement

خطوبة, N+Genre=f+Nombre=s+DOM=legal+GeoUsage=MAC+EN=Engagment+FR=Fiançailles

ممتعة, N+Genre=f+Nombre=s+DOM=Juri\_religion+GeoUsage=MAC+EN=Compensation+FR=Compensation

طلاق, N+Genre=m+Nombre=s+DOM=legal+GeoUsage=MAC+EN=divorce+FR=divorce

إيجاب, N+Genre=m+Nombre=s+DOM=legal+GeoUsage=MAC+EN=Proposal+FR=Proposition

قبول, N+Genre=m+Nombre=s+DOM=legal+GeoUsage=MAC+EN=Consent+FR=Consentment

وكيل, N+Genre=m+Nombre=s+FLX=Wokalaa+DOM=legal+Sem=Hum+GeoUsage=MAC+EN=appointee+FR=délégué

مادة, N+Genre=f+Nombre=s+FLX=Mawad+DOM=legal+GeoUsage=MAC+EN=Article+FR=Article

قضية, N+Genre=f+Nombre=s+DOM=legal+FLX=Qadaya+GeoUsage=MAC+EN=case+FR=affaire

مداولة, N+Genre=f+Nombre=s+DOM=legal+GeoUsage=MAC+EN=Deliberation+FR=Délibération

راشد, N+Genre=m+FLX=Flx1+DOM=legal+Sem=Hum+GeoUsage=MAC+EN=adult+FR=adult

مسطرة, N+Genre=f+Nombre=s+FLX=Masatir+DOM=legal+GeoUsage=MAC+EN=Procedure+FR=Procédure

شكايية, N+Genre=f+Nombre=s+FLX=Shikayat+DOM=legal+GeoUsage=MAC+EN=Complaint+FR=Plainte

براءة, N+Genre=f+Nombre=s+DOM=legal+GeoUsage=MAC+EN=innocence+FR=Innocence

قربانية, N+Genre=m+Nombre=s+DOM=legal+GeoUsage=MAC+EN=affinity+FR=Parenté

عازب, N+FLX=Flx1+legal+Hum+GeoUsage=MAC+EN=Single+FR=Célibataire

قران, N+Genre=m+Nombre=s+DOM=legal+GeoUsage=MAC+EN=Marriage+FR=mariage

رضى, N+Genre=m+Nombre=s+DOM=legal+GeoUsage=MAC+EN=satisfaction+FR=satisfaction

البناء, N+Genre=m+Nombre=s+DOM=Juri\_religion+GeoUsage=MAC+EN=Consummation+FR=Consummation

محكمة, N+Genre=f+Nombre=s+FLX=Mahakim+DOM=Juri\_religion+GeoUsage=MAC+EN=Court+FR=Tribunal

معاينة, N+Genre=m+Nombre=s+DOM=Juri\_religion+GeoUsage=MAC+EN=Inspection+FR=Inspection

نزاع, N+Genre=f+Nombre=s+DOM=legal+GeoUsage=MAC+EN=Dispute+FR=Litige

حكم, N+Genre=m+Nombre=s+FLX=Ahkaam+DOM=legal+GeoUsage=MAC+EN=Judgment+FR=Jugement

قنصلية, N+Nombre=s+FLX=Qonsoliaat+DOM=legal+GeoUsage=MAC+EN=Consulate+FR=Consulat

وثيقة, N+Nombre=s+FLX=Wataaiq+DOM=legal+GeoUsage=MAC+EN=Document+FR=Document

بند, N+Nombre=s+FLX=Bonoudon+DOM=legal+GeoUsage=MAC+EN=term+FR=terme

وكالة, N+Nombre=s+DOM=legal+GeoUsage=MAC+EN=term+FR=terme

نفقة, N+Nombre=s+DOM=legal+GeoUsage=MAC+EN=Maintenance+FR=Aliment

دعوة, N+Nombre=s+DOM=legal+GeoUsage=MAC+EN=Counterclaim+FR=Demande  
reconventionnelle

شهادة, N+Nombre=s+FLX=Shahadaat+DOM=legal+GeoUsage=MAC+EN=Cerificate+FR=Ce  
rtificat

ميثاق, N+Nombre=s+DOM=legal+GeoUsage=MAC+EN=Charter+FR=Charte

سراح, N+Nombre=s+DOM=legal+GeoUsage=MAC+EN=Release+FR=Liberation

جناية, N+Nombre=s+FLX=Jonah+DOM=legal+GeoUsage=MAC+EN=Crime+FR=Délit

قاضي, N+Genre=m+Nombre=s+FLX=Qodaaat+Sem=Hum+DOM=legal+GeoUsage=MAC+E  
N=Judge+FR=Juge