AI4Lawyers – Artificial intelligence for lawyers: Guide on the use of AI and other novel IT technologies by European lawyers and law firms

Call: JUST-JACC/EJU/AG/2019, Grant number: 881527

Work Package 2 – Deliverable 2.2 – Public

OPPORTUNITIES AND BARRIERS IN THE USE OF NATURAL LANGUAGE PROCESSING TOOLS IN SME LAW PRACTICES

26 November 2021

European Lawyers Foundation
ANBI – Dutch Public Benefit Organisation
Fluwelen Burgwal 58, 2511 CJ Den Haag, The Netherlands – Tel. +31 612 99 08 18
E-mail: info@elf-fae.eu – Website: www.elf-fae.eu

The project is co-funded by
the Justice Programme
of the European Union

Publisher

Council of Bars and Law Societies of Europe

Rue Joseph II, 40

B-1000 Brussels (Belgium)

www.ccbe.eu

European Lawyers Foundation

Fluwelen Burgwal 58

2511 CJ, The Hague (Netherlands)

www.elf-fae.eu

This report was commissioned by the CCBE from Mr. Pál Vadász within the EU funded project "AI4Lawyers". He is the author of this report together with György Görög, Csaba Csáki, and István Úveges. However, this report has also benefited from other contributors who are listed below. Therefore, this report may not express the full views of the named author(s).

Contributors

Péter Homoki

Britta Kynast

Thierry Wickers

Christian Lemke

Jiri Novak

Carla Secchieri

Iain Mitchell QC

European Lawyers Foundation (ELF) - Alonso Hernández-Pinzón García, Jonathan Goldsmith

Council of Bars and Law Societies of Europe (CCBE) - Simone Cuomo, Martin Sacleux

Table of contents

Background and aims of the AI4Lawyers project

On 19th October 2017, as part of addressing emerging trends, the European Council invited the European Commission to put forward a European approach to artificial intelligence ("AI") by early 2018. In a subsequent communication of the European Commission[1], the Commission set out a European initiative on AI aiming to prepare for socio-economic changes brought about by AI through, among other things, encouraging the modernisation of education, anticipating changes in the labour market and supporting labour market transitions.

In the 2019-2023 Strategy on e-Justice, the Council also stressed that legaltech[2] areas such as AI should be closely monitored, in order to identify and seize opportunities with a potentially positive impact on e-Justice.[3]

The Council of Bars and Law Societies of Europe (CCBE), which represents more than 1 million European lawyers through its bar and law society members, has been following for more than a decade the effects new technologies have on the day-to-day operations of lawyers.[4] AI in general, and the possible changes that may be brought about by the tools which customarily use AI, has been a direct subject of numerous studies by its committees and working groups since at least 2016.[5]

The outcome has been summarised in the CCBE Considerations on the Legal Aspects of Artificial Intelligence, adopted in 2020 ('CCBE Considerations').[6] The CCBE Considerations devoted a separate chapter to the issue ('The impact of AI on legal practice')[7]. That chapter highlights the specific areas that are worth exploring in more detail, such as subfields of research within AI which are more relevant to lawyers' everyday life, the general difficulties in applying AI tools to lawyers' work, and the opportunities obtained from various tasks and process steps within the work of an "average" lawyer.

The CCBE Considerations identified that the most important aspect of AI to be studied in relation to lawyers is not simply how to approach certain technical problems, but how the technical changes to be expected will affect the rule of law through the changed operations of a lawyer, and how the core principles of the European legal profession can be preserved in the interest of clients and the rule of law. All the issues set out in the CCBE Considerations warrant further in-depth studies.

In 2019, the EU launched the adopted a 2019-2023 Action Plan on European e-Justice[8], which sets out a list of projects and initiatives ('actions') to be implemented as part of the 2019-2023 European e-Justice Strategy. The Action Plan also indicates the goals of individual actions and the envisaged

[1] European Commission. "COM(2018) 237 final Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Artificial Intelligence for Europe." Brussels, 25/04/2018. <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51625>, accessed 29 August 2021.

[2] In this overview, legaltech simply means legal technology for lawyers or the market of such software (law firm specific software market).

[3] See Official Journal of the European Union, OJ C 96, 13.3.2019, p. 6

[4] See e.g. 'Council of Bars and Law Societies of Europe (2005). 'Guidelines on electronic communication and the internet. Brussels, 12 2005. https://www.ccbe.eu/fileadmin/speciality_distribution/public/documents/IT_LAW/ITL_Position_papers/EN_ITL_20051230_Electronic_communication_and_the_internet.pdf (accessed 5 September 2021).

[5] See e.g. Council of Bars and Law Societies of Europe. 'Innovation and Future of the Legal Profession in Europe'. Bruxelles: Bruylant, 2017.

[6] Council of Bars and Law Societies of Europe. 'CCBE considerations on the legal aspects of AI 2020.' 2020. https://www.ccbe.eu/fileadmin/speciality_distribution/public/documents/IT_LAW/ITL_Guides_recommendations/EN_ITL_20200220_CCBE-considerations-on-the-Legal-Aspects-of-AI.pdf (accessed 04/11/2020).

[7] See The impact of AI on legal practice and also see separately section 4.7 on the use of AI by lawyers and defense counsels in the criminal justice systems (Council of Bars and Law Societies of Europe 2020).

[8] See project number 11 in "2019-2023 Action Plan European e-Justice" in OJ C 96, 13.3.2019, p. 17.

activities, the participants, and the expected contributions of the stakeholders involved (citizens, companies, legal practitioners and judicial authorities). The drafting of a guide on the use of AI by lawyers in the EU was mentioned in the Action Plan under the possible actions to be implemented under 'Artificial Intelligence for Justice'.

This inclusion in the Action Plan, and the subsequent publication by the European Commission of the call for proposals for action grants to support national or transnational e-Justice projects (JUST-JACC-EJU-AG-2019), encouraged the CCBE and the European Lawyers Foundation to submit a project proposal on Artificial Intelligence for Lawyers (AI4Lawyers)[9]. The project was awarded an EU Grant and officially started on 1st April 2020. It will run until 30th March 2022.

The project targets the need for European lawyers and law firms to have a clear understanding of the use of AI and other novel IT technologies in their daily practice. The project's main aim is threefold:

1. Overview of on average state of the art of IT capabilities and comparison with best practices in the United Kingdom, USA and Canada
2. Report on opportunities and barriers in the use of natural language processing tools in SME law practices
3. Guide on the use of AI by lawyers and law firms in the EU

[9] This is the acronym of the project, not to be confused with e.g. the project of the University of Oxford, called AI4LAW (used at least since around the same time). The projects are not related.

Introduction to the report on opportunities and barriers in the use of natural language processing tools in SME law practices

This deliverable is the result of the project's phase 2 , which aimed to produce a report on opportunities and barriers in the use of natural language processing tools in SME law practices.

The work on the report was carried out between February and October 2021. Taking into account that neither ELF nor the CCBE had the required in-house expertise to execute the research needed for the report, the work was subcontracted to an IT expert (Mr. Pál Vadasz). The IT expert was supported in his work by the other project subcontractor (Péter Homoki) and CCBE experts.

The goals of this report are twofold:

- to give background material to enhance and support the absorption of legaltech tools and methods by small law firms (henceforth SLF's) and law firms working within languages not widely spoken across the EU.
- to give background to the drafting of a guide on the use of AI by lawyers and law firms in the EU (project phase 3).

Questions to be answered by this study.

- What are the major technological developments in the field of NLP that may influence the future of SLFs?
- Which technological and organisational[10] barriers impede the absorption of these technologies?
- What are the opportunities in this new environment?

Methods of research employed

Most of this work was carried out by desktop research and correspondence with academic, business, and legal experts. Business and scientific publications were used as reference to acquire knowledge. To explore more deeply in specific cases data bases were referred to. Unfortunately, due to the Covid 19 pandemic there has been no opportunity for onsite consultations. Conclusions were drawn from the data of other professional sectors utilising NLP technologies that are perhaps subject to other regulation in adoption, such as management consultancy, information and communications technology (henceforth ICT), marketing, healthcare etc.

Boundaries of this study

The technology is changing so quickly, that the question is not whether any statement here will be obsolete, but simply when it will be (probably very soon). New technologies, methods and products grow like mushrooms. Hence, this paper needs to be read in the time context when it was produced (November 2021).

The scope of the project had been developing during the Q&A sessions and correspondence throughout the whole of phase 2. The fluid sharpening of the information need resulted in the understanding that in certain well-defined areas much **deeper exploratory investigations will be**

[10] Under organisational we understand sociological, political, economic, financial, legal etc. circumstances that influence efforts and processes.

desirable. Beside the barriers and opportunities, a strong focus could be experienced on the legal technologies themselves.

Though the original task was to focus on SLFs, there is another segment that needs much attention, the **small languages in the EU**. As will be seen below, legal firms and legal departments working with languages with small populations, and therefore small markets in the EU suffer under heavy disadvantages compared to those using the languages of large, dominant populations. Furthermore, attention is given to polycentric languages[11]. Since the small language problem is one of the major issues discussed in this paper, it is examined in both sections relating to barriers, in its relevant aspects.

This paper is not to be identified with the view of CCBE or that of its members, this report is the result of an independent research study undertaken by non-legal professionals for the purposes communicated by EFL and CCBE.

Since the lingua franca of academic publications is English, most sources found and referenced are in English. The majority of legaltech publications found in academic depositories such as Google Scholar come from English-speaking sources, which **may cause a bias to the disadvantage of EU scientific sources or those written in languages other than English**.

Though the main subject of this study is the SLFs, there is unfortunately **hardly any academic material available on this topic**. As of 21st August 2021, there are 26 hits on the subject "*small legal firms*" and just 398 on "*small law firms*" in Google Scholar since 2017, of which approximately 170 papers related to various EU sources, including the UK.

It must be stated that despite thorough research over several months we **could not establish significant technical barriers and opportunities that can relate solely to SLFs**. All computational linguistic efforts and results are valid for the whole legal field regardless of the size of the legal entity. This statement does not, obviously apply to the small language problem.

We would have liked to structure this study as a SWOT[12] analysis emphasising the threats and weaknesses but following the original contract these have been merged into **the barriers and opportunities sections that follow the overview of NLP based legal technologies**.

We have studied several market surveys and reports such as those from HSBC, Deloitte, KPMG, Altman Weil, Consero, Thomson Reuters, Aderant, The Law Society of England and Wales, and ABA (Clio supported), and found that though the specific methods of research, sampling, and evaluating are different, the **conclusions for the granularity required for the present study are similar enough not require separate examination**. We are grateful to Wolters Kluwer Germany for the charts from their 2020 and 2021 surveys and have used them to illustrate the most vital findings.

[11] A polycentric language is a language with several interacting codified standard forms, often corresponding to different countries. Source: Wikipedia.

[12] SWOT analysis (or SWOT matrix) is a strategic planning technique used to help a person or organization identify strengths, weaknesses, opportunities, and threats related to business competition or project planning. Source: Wikipedia.

An overview of NLP based legal technologies

"A word is characterized by the company it keeps"[13]

NLP technologies applied by legaltech

NLP based legal technologies are discussed below. These are grouped into two main categories, NLP technologies that are used in legaltech, and legaltech applications that apply NLP. The lists are not exhaustive of all technologies, since there is neither room nor need for a comprehensive overview of the whole legaltech landscape, but we try to focus on the one hand on those likely to be used by SLFs, and on the other hand on those that are most relevant to the barriers and opportunities elaborated in the coming sections. We also focused on those subjects discussed during the preparatory phase. One section follows legaltech in terms of the changes in business models used in the legal industry.

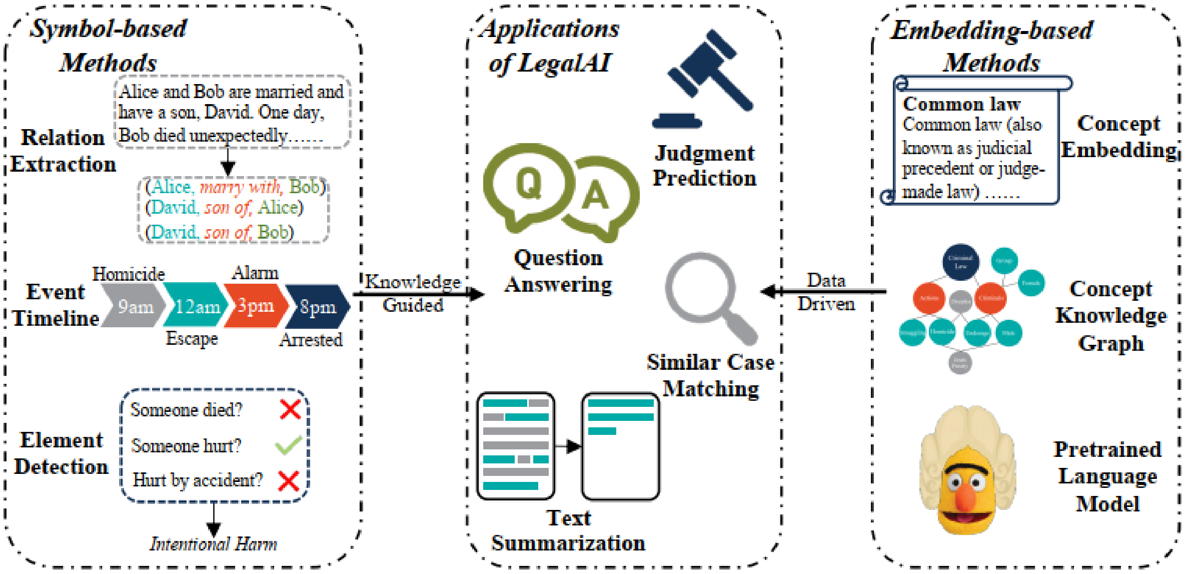


Figure 1. Workflow of legalAI applications. Source: Zhong et al.: (2020) How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 5218-5230.

A traditional lexical-based search is exhausted in the search for the exact sequence of characters that match a given combination of keywords. It produces significant "*noise*" in the search results. Even this simple approach can benefit from stemming and synonyms.

In **semantic search**, the main goal is to be able to provide an interpretation beyond the static, dictionary meaning of the words in the search expression. The key is to find the intentions(s) of the searcher, the context given by the search expression and the relationships between the individual words. By adding this kind of information, we can obtain more relevant results that better match the

[13] Firth, John Rupert (1957). 'A Synopsis of Linguistic Theory' https://czlwang.github.io/zettel/20201201162131-firth_1957_a_synopsis_of_linguistic_theory.html accessed 29 August 2021.

searcher's intention(s). To achieve this, it was necessary to integrate new approaches such as deep learning.

Symbol-based methods (Relation Extraction, Information Extraction) organise and interpret legal knowledge and relationships between individual symbols (e.g., Named Entities such as persons, organisations etc.). Several application possibilities have been explored. For instance, relation-extraction is a special type of information extraction (IE), whereby the main goal is to build a structured database of sentences with entity pairs and relation types extracted from the text. Another example is to detect the sequence of events in a criminal case.

Document analysis is one of the symbol-based solutions. It organises the unstructured information present in the text into a form that can be interpreted by humans and software alike. This can be achieved by putting the entities of the text into relationships with each other, or by building ontologies or wordnets.

The important task here is to represent knowledge in the human mind to make it available to machine learning algorithms, i.e., to be able to use together the knowledge available from the data and the machine learning solutions. A good example of this is the construction of knowledge graphs, which also try to capture the relationship (semantics) of concepts.

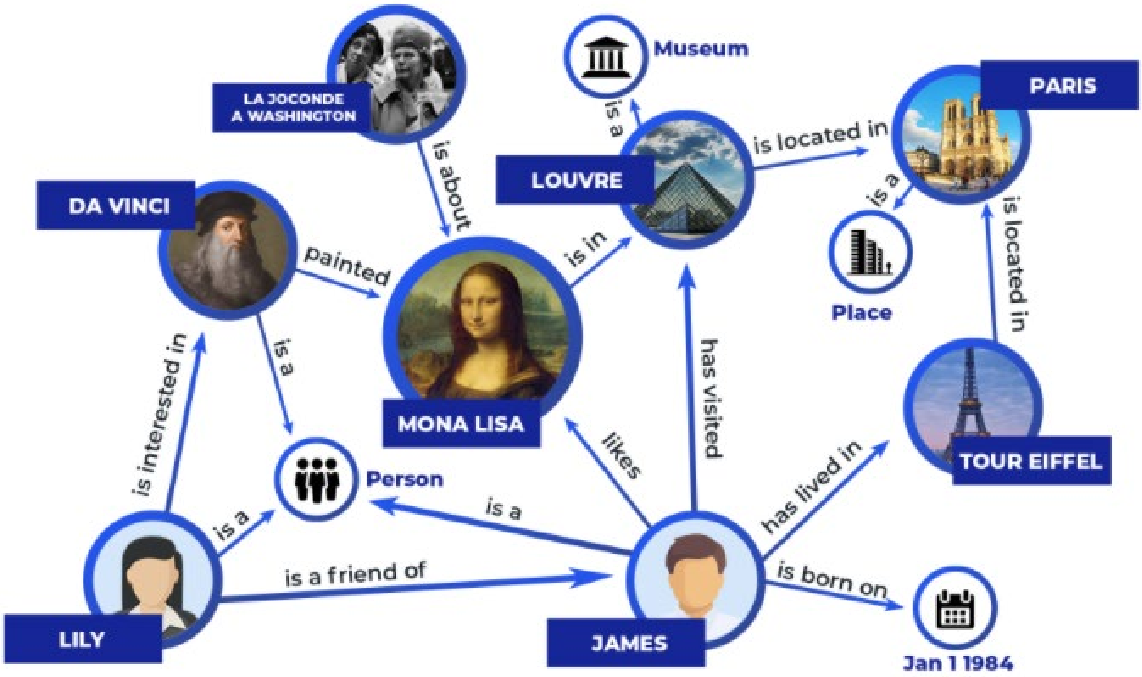


Figure 2. Syntactic information, which can be extracted using NLP tools, can help to map relations. For instance, in the sentence "da Vinci, painter of the Mona Lisa", the relation is "painter of the Mona Lisa", which connects “da Vinci”. Source: https://yashuseth.blog/2019/10/08/introduction-question-answering-knowledge-graphs-kgqa/

Word embedding-based methods were introduced in 2013. Their novelty compared to lexical-based search was that the models included contextual information in their representation of words. A good example of embedding-based solutions is the creation of **pre-learned language models** or **knowledge graphs**. The latter are essentially specialised ontologies that focus on practical applicability rather than philosophical depth. They do not encode ordinary knowledge, but rather focus on many entities.

Embeddings may have difficulty in telling apart synonyms and antonyms (occurring in similar text surroundings).

Deep Learning designates neuronal network-based solutions where the network structure is increasingly complex, computing capacity permitting. The field has been experiencing a renaissance since the release of **BERT** (Bidirectional Encoder Representations from Transformers) in 2018. BERT is a specific deep learning pre-training language model published by Google. The novelty of BERT lies in its ability to distinguish the meaning of words with the same form from each other based on context. For example, in "*They're planning to go on a second date*." and "*They're trying to find a good date for the conference.*" the word "*date*" has a very different meaning. Traditional word embeddings (such as Word2Vec or Node2Vec) assign the same mathematical representation (in particular, a vector space representation) to the word "*date*" in both sentences, whereas BERT generates its own vector for the word "*date*" in both sentences, from which the difference in meaning can be recovered.

Reconstructing such semantic differences in the processing of search terms is crucial for retrieving relevant results, especially in the case of law, where common words can also occur as specific legal terms. An additional advantage over traditional word embeddings is that while the latter cannot handle the presence of words that were not encountered during model learning (the model learns at the "*word level*"), BERT learns at the "*subword*" level and is able to overcome this problem, making it much more robust in practical applications.

Both approaches can have similarities in practice, they can, for example, be applied to solve **similar-case matching** problems, where the aim is to find previous precedents that are substantially like the case at hand. In such a case, the search for similarity implies a kind of semantic similarity hashing, which may appear at fact level, event level or element level. On the other hand, the notion of similarity can also be interpreted in several ways, e.g., similarity of facts or the solution given by the judge. Considering that similarity may be understood as similarity of facts, similarity of the legal grounds or similarity of the result (the solution given by the judge), it is also clear that language technologies will only be part of the technology, together with intelligent element classification algorithms that decide if an element is a fact, legislation reference or a decision element.

In any case, the knowledge obtained by symbol-based solutions is interpretable by humans, but the results of deep machine-learning solutions are not.

While traditional lexical-based solutions were often based on the use of rule sets created manually by experts, to exploit the potential of AI, their toolset had to be integrated. For example, training machine learning models requires large amounts of data (corpus). Some methods (e.g., BERT) are very computationally intensive during training and require significant architectural and financial investment to build. If the general language model is to be applied to a specific task, it also needs to be fine-tuned to the specific target texts. Moreover, unlike traditional machine learning models, BERT-based solutions (e.g., RoBERTa, DistilBERT) and other transformers (like XLNet) are computationally expensive not only to train, but also to use for prediction (e.g., to find similar documents).

Transfer learning could be a technology for avoiding obstacles that smaller languages with limited corpora in specific legal fields (e.g., large numbers of contracts) might face in the future. Transfer learning is a means of extracting knowledge from a source setting and applying it to a different target setting. An example is (pre)training an AI model on a general text corpus of a given language and later fine tuning the model on a special domain corpus (e.g., legal texts) of the same or another language. In transfer learning, the base network is trained on a base dataset and task at first to be able to capture very general features of the data (in our case the texts). During the fine-tuning phase the original model (entirely or partially) is used as a starting point for the new model to be developed for the specific task, therefore the original “knowledge” extracted by the first model is transferred to the latter model. However, how far this technology can be helpful when not only languages but also jurisdictions differ would need to be explored .Lately, with the advent of Google’s BERT tool and derivatives, this procedure has gained dominance in the field.

In **cross-language transfer learning** (henceforth CLTL) machine knowledge is transferred from a source (resource-rich) language to another, target (resource-poor) language. Resources in this regard are annotated text corpora and examples. The same-language and cross-language approaches often come mixed, even the training texts may be of different languages. For using certain NLP models, CLTL could become an essential technology in the future for smaller languages with limited corpora in general (e.g., GPT-3, Google Switch etc.).

There are inherent differences between the goals, means and attitudes of academic researchers and AI engineers[14]. Some innovative methods, even if published, remain obscure and are difficult to unearth, while scrutinising, understanding and reproducing every and all published papers and videos is simply impossible. The seminal paper describing the technology under BERT (transformers) by Vaswani et al. has more than 26,000 citations[15].

Among academics, the BERT model is predominant. This raised the hope that a single tool will solve any and all tasks, at least until the advent of GPT-3[16]. There are other models and a range of papers comparing them; for instance Ghavidel et al.[17] and Wenjie [18].

[14] Masatoshi Nishimura, ‘The Best Document Similarity Algorithm in 2020: A Beginner’s Guide’ (Medium, 7 May 2021) https://towardsdatascience.com/the-best-document-similarity-algorithm-in-2020-a-beginners-guide-a01b9ef8cf05 accessed 5 September 2021.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), 31st Conference on Advances in Neural Information Processing Systems (NIPS 2017) (Vol. 30, pp. 5999–6009). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[16] Sejuti Das, ‘GPT-3 Vs BERT For NLP Tasks’ (Analytics India Magazine, 11 September 2020) https://analyticsindiamag.com/gpt-3-vs-bert-for-nlp-tasks/ accessed 5 September 2021.

[17] Hadi Ghavidel, Amal Zouaq and Michel Desmarais, ‘Using BERT and XLNET for the Automatic Short Answer Grading Task’:, Proceedings of the 12th International Conference on Computer Supported Education (SCITEPRESS - Science and Technology Publications 2020) http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009422400580067 accessed 5 September 2021.

[18] Wenjie, Z. (2020). The comparison between the tools for Named Entity Recognition. Master’s Thesis. [Auckland University of Technology]. https://openrepository.aut.ac.nz/bitstream/handle/10292/13364/ZhangWenjie.pdf?sequence=1&isAllowed=y

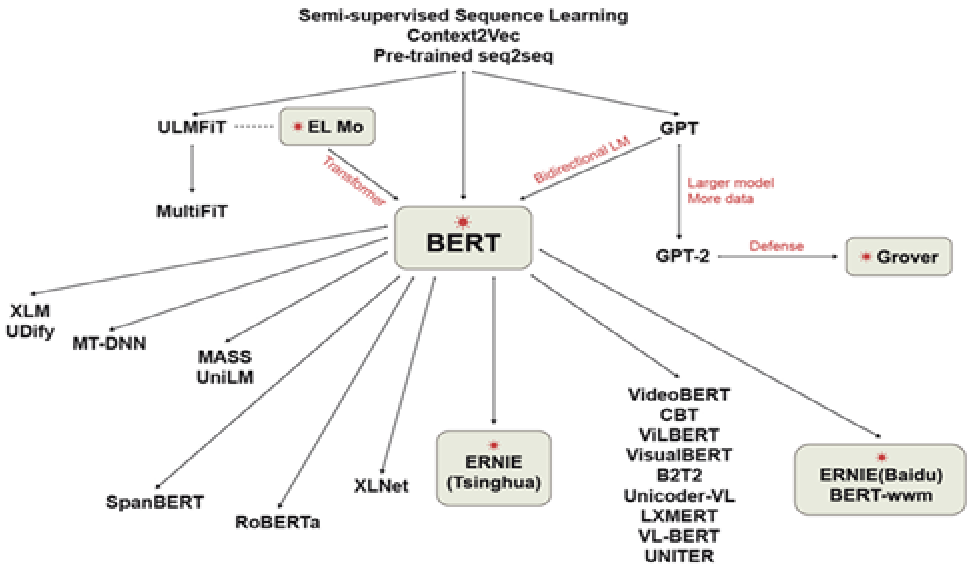


Figure 3. The academic evolution of word embedding methods.
Source: https://www.huaweicloud.com/news/2020/20201231115833.html

In industrial applications, it is common to state the use of NLP and Deep Learning without specifying the technology (e.g., "*provides AI-powered data extraction*", "*uses machine learning software*"). BERT is already used, but not widely. WeSearch, a "*solution for an infinite number of legal needs*" launched by CaseText this March is based on uploaded documents analysed with the help of BERT. The application returns relevant sentences and paragraphs from its document base, pointing to the source (within the user's uploaded documents or a public one). Their slogan "*you can finally escape the prison of the keyword*" may be accurate, and characterises recent developments in legal tech correctly.

The integration of classical NLP tools and deep learning has enabled the emergence of several functions that could not be effectively implemented earlier by using traditional methods. Examples of this can be seen in some of the main industrial development directions:

- in the search process, the aim is to allow questions to be formulated as far as possible in the user's everyday language rather than in some artificial query language;
- the aim is to return answers and not hit lists as results;
- insight is given into the context behind the data, and automatic analysis if the question allows.

A good example of this is the use of search engines that can immediately visualise the relationships between cases by reference to each other. This makes it easy, for example, to identify important cases that are often used in precedent-setting decisions.

The combined use of named entity recognition (henceforth NER) and relation mining in judicial decisions can help find documents where a person was present in a particular role (e.g., as a defendant) without the need to manually build a database by pre-processing documents beforehand.

Automatic **text generation** of sufficient quality in the context of the legal domain is currently an unsolved problem. This is probably mainly due to the linguistic characteristics of legal texts; language models taught for general purposes cannot adequately represent the specific legal meaning of colloquial words or the semantics of technical legal terms.

The most recent and promising solutions are largely based on a combination of existing methods. Such a solution could for example be one where the text generation is done by the language model, key information is represented by form slots at this stage, and then these slots are filled by a Q&A method during the final text generation. Such slots are typically assigned to those elements of the text that need to be significantly constrained by logic, such as the law or the paragraph being referred to.

The accuracy of the language model used for generation has a significant impact on the overall quality of the text, so progress made there will indirectly affect the quality of the legal text generation. However, current trends in development seem to indicate that for a drastic (commercially acceptable) improvement in quality, the development of existing knowledge representation methods is also key.

Semantic natural language understanding and text generation are related tasks, but some models perform better in the former than in the latter. GPT-3 is famous for its capability to write long documents in English.

Since their relevance to the every day legal practice can be considered further, the following technologies are not discussed, but simply mentioned for those interested in further research: optical character recognition (OCR), speech recognition, speech to text, speaker recognition, text-to-speech, word sense disambiguation, text summarisation (both extractive and abstractive[19]), grammatical error correction, machine translation, question answering, sentiment analysis, document classification, natural language understanding.

Legaltech applications using NLP technologies

Law practice management software integrates and automates the front and back-office activities of legal practices including calendaring, appointment scheduling, case management, conflict checking, messaging, project management, time tracking, billing etc. (see the results of phase 1 of this project)[20]. Legal document management software allows legal professionals to organize and quickly access electronically stored legal documents. Legal case management software coordinates documents, scheduling, conflicts, contacts, and reporting associated with legal cases etc. These tools cover a wide range of possible functions; nevertheless, NLP tools in this area are often used in helping users with filing and categorising input documents and capturing and filling out metadata and content data on specific attributes (such as the relevant cases or matters the file pertains to, a workflow step or a calendar event, identifying accounting information such as relevant journals or subledgers etc). NLP tools are also widely used in the retrieval of documents relevant for the law practice.

Legal Research software and services aggregate court judgements, legislation, regulations, and case law from a variety of sources as well as providing tools for automating legislation discovery and analysis. Legal research helps in monitoring legislation changes, generating, and backing legal opinions and precedent identification. Contracts can be generated in-house by filling in a questionnaire or form, from which data a simple program will transfer data into templates. A contract draft received from outside, however, is just a bag of words, not data. It is difficult to extract data such as parties' names or terms, or even numeric financial values into a database, but NLP tools are available or can be trained for such uses. Such tools are also used for identifying the existence or relation of certain clauses within the documents analysed.

E-discovery software is a tool used by legal professionals to gather and maintain documents and communications related to a lawsuit. Such documents and communications include emails, chats, instant messages, PDFs, audio/video files, and social media messages. The tool creates a centralized and searchable directory for such information and ensures appropriate governance for data storage.

[19] Rohan Jagtap. 'Abstractive Text Summarization Using Transformers' (Medium) https://medium.com/swlh/abstractive-text-summarization-using-transformers-3e774cc42453

[20] For an overview of the „average state of the art" IT capabilities of law firms …", see https://elf-fae.eu/wp-content/uploads/2021/03/Overview-of-the-average-state-of-the-art-IT-capabilities-in-the-EU.pdf

E-discovery can be applied among others to related documents identification, technology assisted review, fraud detection, compliance or cartel detection. The phases of the e-discovery workflow are as follows: identification of the relevant sources, data capture, data processing, review, and dissemination. ML is frequently used in the processing phase to identify relevant documents by using pretrained datasets[21] (technology-assisted review, TAR). (E-discovery as a technical field is not identical to the process of discovery as an issue of different procedural laws, i.e., such tools are usually marketed under this name even if there is no discovery process in the legal sense in the given country).

Litigation prediction or litigation risk analysis is an application of predictive analytics based on large corpora of judicial decisions. It helps decide whether one should litigate a matter or settle it and in other strategic decisions.. The gist of litigation prediction is comparing a specific case which the legal expert is working on, to cases processed in the past, and finding similar ones. Once these are listed according to relevance, their judicial decisions are evaluated and the probability of the outcome of the given specific case is established by the application.

Chatbots are NLP based human-machine question and answering (Q&A) applications that simulate a human-to-human conversation[22]. Chatbots understand questions put in natural language both in writing and through voice, the latter using speech recognition (speech-to-text S2T) technology, which is itself a branch of AI.

Legaltech Business Models Enabled by New Technologies

While technology which has appeared in recent years (or even decades), especially various ICT solutions are claimed to be the most important part of changing the value chain in most industries, it is important to consider through what mechanisms technology can create new value propositions. This is not different for the legal profession either.Recent leading technologies include big data, cloud solutions, internet-based applications, data science, analytics, or AI have, however, managed to penetrate the legal world, but how these enable the delivery of more efficient services to customers deserves a second look for legal professionals. Indeed, two aspects of the legaltech ecosystem[23] deserve attention here: (business) processes and business models[24].

Processes mean the (elementary) tasks and the roles fulfilling those tasks in the legal workflow and our concern is how technology impacts such roles and tasks. The business model, on the other hand, is defined as the way or logic through which a company creates, captures, and delivers value to its customers and other stakeholders[25]. Obviously, restructuring any part of legal workflows may lead to new business models and such restructuring may be supported or even initiated by info-communication technologies.

[21] Pál Vadász and others, 'Identifying Illegal Cartel Activities from Open Sources' in Babak Akhgar, P Saskia Bayerl and Fraser Sampson (eds), Open Source Intelligence Investigation (Springer International Publishing 2016) http://link.springer.com/10.1007/978-3-319-47671-1_16 accessed 26 August 2021.

[22] 'What Is a Chatbot and How Does It Work?' (SearchCustomerExperience) https://searchcustomerexperience.techtarget.com/definition/chatbot accessed 18 August 2021.

[23] Here, by ecosystem we mean the relationships formed between members of a co-dependent community who take different roles in the resource-flow within a certain local area, contextualized by environmental constraints. Resources in the legaltech ecosystem mean primarily data/information not just money, while the crucial element of the environment is the regulatory context, and roles include legal service providers, data providers, data cleaners, technology service providers, clients, and so on.

[24] Qian Hongdao and others, 'Legal Technologies in Action: The Future of the Legal Market in Light of Disruptive Innovations' (2019) 11 Sustainability 1015 https://www.mdpi.com/2071-1050/11/4/1015 accessed 6 September 2021.

[25] Osterwalder, A., and Y. Pigneur (2010). Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers. Vol. 1. John Wiley & Sons.

From the point of view of processes technology (especially AI) promotes the breakdown of the legal profession and the division of legal workflow into elementary tasks[26]. Then ICT enables substitution of specific tasks (with automated versions), augmentation of tasks (supporting humans to increase efficiency), and the creation of new tasks (used in forming new services) [27].

Business models describe the logic of utilizing technology and the corresponding strategic choices of a firm. Regarding the delivery of services, it is necessary to look at both sides: the service provider (in this context the law firm) and the service recipient (i.e., legal departments of larger organizations or small firms and individuals). Furthermore, the core of a business model is a customer value proposition which includes the way the business best meets the perceived needs of its customers[28].

The ability of law firms to change their (or their clients') processes and deliver value to customers while generating profit depends on what the literature calls *'assets'* (or in some contexts (dynamic) capabilities). Such assets include not only technology itself, but data (considering size and quality) and - increasingly more importantly - human capital. Some legal firms may consider hiring technology-savvy employees - which implies not only technical people (with preferably some affinity to and understanding of legal matters and proceedings), but also lawyers with knowledge and understanding of technological matters. This is what Armour and Sako call, the '*hybrid professional'*[29]. (Although it is not part of this report, it is worthwhile to note here that recent literature has been discussing possible changes in legal education considering technology-based advances in the legal practice (see for example Davis[30]) in order to train 'hybrid' legal professionals.) However, there are also other possible ways to answer technological challenges such as outsourcing to a provider who very well understands the needs of a given law firm or support from bars and law societies to their members through providing various forms of technical assistance.

When considering the changing mix and use of assets along with the changing task and role structure, different business model archetypes have been proposed with some varying similarities and overlaps[31]. We review two leading approaches here: the one proposed by Hongdao et al.[32] (refining the ideas put forward by Susskind, 2008[33] and 2013[34] considering market changes) and another proposed by Armour and Sako [35]. This does not imply, however, that such changes deemed inevitable, neither this technical report embraces any opinion about the future.

According to Hongdao et al.[36] the legal market might be split into the following five key areas (what they call 'segments') each representing a particular business model (BM) strategy:

[26] Salmerón-Manzano, E. (2021). Legaltech and Lawtech: Global Perspectives, Challenges, and Opportunities. Laws 10(2): 24.

[27] Kerikmäe, T., T. Hoffmann, A. Chochia (2018). Legal technology for law firms: Determining roadmaps for innovation. Croat. Int. Relat. Rev., 24: 91–112.

[28] Zott, C., and R. Amit (2013). The business model: A theoretically anchored robust construct for strategic analysis. Strategic Organization 11(4): 403–411.

[29] John Armour and Mari Sako, 'AI-Enabled Business Models in Legal Services: From Traditional Law Firms to next-Generation Law Companies?' (2020) 7 Journal of Professions and Organization 27 https://academic.oup.com/jpo/article/7/1/27/5734679 accessed 6 September 2021.

[30] Anthony E Davis, 'The Future of Law Firms (and Lawyers) in the Age of Artificial Intelligence' (2020) 16 Revista Direito GV e1945 http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1808-24322020000100404&tlng=en accessed 6 September 2021.

[31] Webley, L., J. Flood, J. Webb, F. Bartlett, K. Galloway, and K. Tranter (2019). The Profession(s)'engagements with Lawtech: Narratives and Archetypes of Future Law. Law, Tech. & Hum. 1(6).

[32] Hongdao and others (n 24).

[33] Richard E Susskind, The End of Lawyers? Rethinking the Nature of Legal Services (Oxford University Press 2008).

[34] Richard Susskind, Tomorrow's Lawyers: An Introduction to Your Future (1st edition, Oxford University Press 2013).

[35] Armour and Sako (n 29).

[36] Hongdao and others (n 24).

- self-service BM: law firms offering commoditized law solutions by satisfying clients' cost-effectiveness needs. In this category, most legaltech start-ups target individuals, small businesses, and enterprises, for example, for online legal-document services (including incorporation, estate plans, legal health diagnostics, legal-document automation, practice management, document storage, billing, accounting, and electronic discovery);

- manifold BM: this includes electronic legal marketplaces, networks, and multisided platforms for L2C, L2B, and L2L services. Other legal tech activities in this group include legal advice and content portals, online reverse-auction platforms, recruiting platforms, legal-database insourcing platforms, and legal-process outsourcing;

- big business model: this contains a large variety of high-tech tools for specific legal workflows, processes, and tasks. This category includes document review, e-discovery, intellectual-property asset management, automated document assembly, legal-contract management, legal-research analysis, and legal-practice management;

- online dispute-resolution (ODR): these are platforms targeting small consumer claims to resolve various disputes without court intervention. ;

- legal artificial-intelligence systems: artificially intelligent lawyer solutions – mainly appearing in the US market – are based on cognitive computer technology and can answer simple legal questions, conduct research for relevant legal source materials using advanced pattern-recognition software and other tools.

Armour and Sako[37] consider the following key business model propositions emerging in the legal profession due to legaltech advances in contrast to the traditional legal advisory business models of established law firms:

- Scaling BM: using (typically) online service offering to deliver legal services to small companies or individuals;

- legal operations BM: legaltech in improving internal process of law firms allowing reductions in costs, although there is an initial investment involved in investigating and then reengineering certain processes or tasks;

- consulting BM: legaltech in reorganizing the work of legal departments of (typically large) clients, i.e., improving through technology the legal operations of clients through technology;

- legal technology BM: developing and selling legal software or data, including various platforms (as a service).

Notice, that the first two options mean, that "*clients, under pressure to reduce internal as well as external costs, will turn to such developers and vendors of AI solutions to achieve outcomes more efficiently, faster, and more cheaply than (traditional) law firms can deliver*" [38].

Regarding the impact of the spread of the above business models (irrespective of which categorization one prefers), Webley et al. conclude - based on narratives from related literature -, that there are three potential outcomes (or legal future '*archetypes*' as they put it): the status or '*True Legal*' professional (meaning augmentation), the '*Technological Disruptor*' or new law innovator (assuming disruption as envisioned by Christensen[39]), or '*Death*' (or the end of lawyers as we know it).[40] Clearly, these are

[37] Armour and Sako (n 29).

[38] Davis (n 30) 10.

[39] Christensen, C. (1997). The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Boston: Harvard Business Review Press.

[40] Webley and others (n 34) 2.

archetypes based on advances in legaltech in the States and there is little evidence in the academic literature regarding similar analysis in continental contexts.

To understand the underpinning of these business models there are changes to be considered in three areas: pricing, scaling, and asset mix. Pricing could move from an input-based (i.e., time-based) model to an output-based model, where the price is based on the outcome of the services provided. AI predictions are often used to predict expected work hours and costs of a task or project based on experience. These are based on the scaling of services (enabled by technology) and the proper mix of assets (also including data) fine-tuned to deliver those services efficiently Such business models may need fuller consideration.The weaknesses of AI solutions include the requirement of large amounts of data being available and the need for trained staff and special expertise[41]. In addition, AI may only be used as back-office support in the case of client-facing work or client-specific services.

One need to notice, that there could be a real difference between large firms and SLFs regarding their opportunities and abilities to take advantage of new technologies and related process restructuring and business models. Large firms are likely to have more access to large amounts of data and the resources to know how to collect and interpret the data they produce. Small and medium-sized firms often do not have adequate tools and could suffer from the lack of sufficient data (either they do not have them, may not have abilities to manage their case data efficiently, and may not have resources to purchase data, if available). While ICT (especially AI) impacts the whole legal field and the results (of NLP tools, for example) are applicable by most firms, this does not mean that the needs or priorities or uptake is the same across the profession (by legal approach, by culture, by country, or by size). SLFs do need easy to use tools which have a major impact on the applied technology and methodologies that should be chosen for SLFs specifically. Furthermore, legal workflows – and the amount of time spent on certain tasks – is different for firms of various sizes. SLFs can have many small cases on their desk, whereas larger firms more likely have large cases with fewer clients per lawyer. Consequently, the number of types of workflows and tasks in the former may be higher for certain firms. Most core legal workflows contain tasks (steps) that either deal with legal documents or work on written legal statements (such as from letters or emails) or explore the legal domain to solve a problem. NLP may help in all such tasks to find the rightcontent and to discover connections to other content. The question is which of these steps may be automated, augmented or left fully for humans.

To summarise, using innovative technologies such as NLP also assumes relevant knowledge, skills, and access to data. This could be a divider of success or failure facing the future.

As a next step the issue of barriers needs to be investigated as will be discussed in the next section.

[41] Davis (n 23).

Barriers

„More for less”[42]

There are basically two kinds of barriers that hinder the absorption of NLP by SLFs. Some can be classified as technological, the others have more to do with the human nature or organisational characteristics.

Technical barriers

Complexity of law and indeterminism

According to McKamey[43] citing Simon Chester, law is messy, and it is difficult to construct algorithms that capture the law in a useful way. Unlike in the medical field, answers to legal questions can vary greatly depending on the relevant jurisdiction. Few legal problems, he says, have clear yes or no answers. Others have noted the complexity of legal reasoning as a potential barrier. One argument is that legal reasoning is an inherently "*parallel process*" in which "*the answer to one question may change which questions are subsequently asked.*" Also, legal uncertainty increases over time[44].

These concerns, long known to lawyers, reached the ICT community in the '90s when neural networks were first applied to legal problems. Where problems are complex, with few simple yes or no answers, AI can find ways through this complexity[45]. For example, reviewing documents for discovery is not a process with simple yes or no answers, and the unique context of the case often determines the degree of relevance for each document. Still, e-discovery systems are usually sensitive enough to the subtleties of a specific case and achieve better results than human-only discovery even without deep learning[46].

Legal reasoning and argumentation are a specific set of skills[47], and although they are pragmatic, largely ignore the theoretical framework of formal logic[48]. The latter may hamper the application of general automated reasoning algorithms[49]. Besides strict logic, argumentation by analogy and

[42] Susskind (n 34)

[43] Mark McKamey, 'Legal Technology: Artificial Intelligence and the Future of Law Practice' (2017) 22 Appeal: Review of Current Law and Law Reform 45, 45 https://heinonline.org/HOL/Page?handle=hein.journals/appeal22&id=53&div=&collection=

[44] Anthony D'Amato, 'Legal Uncertainty' [2010] Faculty Working Papers https://scholarlycommons.law.northwestern.edu/facultyworkingpapers/108

[45] Mirna El Ghosh, 'Automation of Legal Reasoning and Decision Based on Ontologies' 266.

[46] Gordon, &, & Cormack, V. v. (2011). Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, 17 Rich. J.L. & Tech, 17. http://scholarship.richmond.edu/jolt/vol17/iss3/5 accessed 5 September 2021

[47] Murphy, J. B. (2006). The Lawyer and the Layman: Two Perspectives on the Rule of Law. The Review of Politics, 68(1), 100–131. https://www.jstor.org/stable/20452757 accessed 5 September 2021.

[48] Vern R Walker, 'Discovering the Logic of Legal Reasoning' (2007) 35 Hofstra Law Review 23.

[49] 'Handbook of Automated Reasoning, Volume I - 1st Edition' https://www.elsevier.com/books/handbook-of-automated-reasoning/robinson/978-0-444-82949-8 accessed 5 September 2021.

examples is an important part of the legal reasoning toolset[50]. Lawyers must argue for the rules themselves and show why a particular rule (or major premise) should apply to a particular case[51].

What is more, with a pluralistic view of law, law is inherently indeterminate because valid but contradictory legal arguments potentially exist regarding the interpretation of the law[52]; and legal arguments are often arguments about what the language means or ought to mean[53].

Eliot describes the following levels of reasoning automation:

- Level 0: No Automation for AI Legal Reasoning
- Level 1: Simple Assistance Automation for AI Legal Reasoning
- Level 2: Advanced Assistance Automation for AI Legal Reasoning
- Level 3: Semi-Autonomous Automation for AI Legal Reasoning
- Level 4: Domain Autonomous for AI Legal Reasoning
- Level 5: Fully Autonomous for AI Legal Reasoning
- Level 6: Superhuman Autonomous for AI Legal Reasoning

Most of the modelling work published concerns criminal law [54] [55]; an elegant, example-oriented, illustrated, albeit not easily understandable discussion was published by de Zoete[56]. The logic of Bayesian thinking is quite straightforward and easy to understand in paternity calculations[57].

Training text availability

When the quantity of training text is insufficient for AI (the resulting model will be an underfit, as they term it), a two-step procedure termed transfer learning is usually recommended.

The majority of legal texts is not public by nature. Yet it is very desirable to make such texts available for text mining and AI training without exposing them to unauthorised eyes.

Anonymisation (deidentification) is one of these possibilities[58]. Yet as AI representations are not human readable, the user is unlikely to be able to determine if sensitive information has been removed

[50] Jürgen Hollatz, 'Analogy Making in Legal Reasoning with Neural Networks and Fuzzy Logic' (1999) 7 Artificial Intelligence and Law 289 http://link.springer.com/10.1023/A:1008344904309 accessed 5 September 2021.

[51] Danaher, J. (2021, March 8). Understanding Legal Argument (1): The Five Types of Argument. Philosophical Disquisitions. https://philosophicaldisquisitions.blogspot.com/2021/03/understanding-legal-argument-1-five.html accessed 5 September 2021

[52] Eliot, L. (2020). AI and Legal Argumentation: Aligning the Autonomous Levels of AI Legal Reasoning. https://arxiv.org/abs/2009.11180v1 accessed 5 September 2021

[53] Anne von der Lieth Gardner, An Artificial Intelligence Approach to Legal Reasoning (MIT Press 1987).

[54] Ibs, I. C. (2016). Applications of Bayesian Networks In Legal Reasoning. B.Sc. Thesis. https://www.researchgate.net/publication/311667346_Applications_of_Bayesian_Networks_In_Legal_Reasoning accessed 5 September 2021

[55] Henry Prakken, Floris Bex and Anne Ruth Mackor, 'Editors' Review and Introduction: Models of Rational Proof in Criminal Law' (2020) 12 Topics in Cognitive Science 1053 https://onlinelibrary.wiley.com/doi/10.1111/tops.12519 accessed 5 September 2021.

[56] Jacob de Zoete and others, 'Resolving the So-Called "Probabilistic Paradoxes in Legal Reasoning" with Bayesian Networks' (2019) 59 Science & Justice 367 https://linkinghub.elsevier.com/retrieve/pii/S1355030618302922 accessed 5 September 2021.

[57] Amanda B Hepler and Bruce S Weir, 'Object-Oriented Bayesian Networks for Paternity Cases with Allelic Dependencies' (2008) 2 Forensic Science International: Genetics 166 https://linkinghub.elsevier.com/retrieve/pii/S1872497307004036 accessed 5 September 2021.

[58] Gergely Márk Csányi and others, 'Challenges and Open Problems of Legal Document Anonymization' (2021) 13 Symmetry 1490 https://www.mdpi.com/2073-8994/13/8/1490 accessed 29 August 2021.

or not. Also, requirements may change over time and published documents cannot be reliably withdrawn.

Scarcity is a main enemy of anonymisation[59]. Many people sell peanuts, so that if personal information is removed, a peanut contract's parties are not easy to identify. If, however, only a single company in the given country sells electricity, it is easily identifiable from the subject of the contract itself; if the subject and other specific details are also removed, the usefulness of the document may severely be hampered.

Polycentric languages

From a theoretical linguistic point of view, polycentric languages can be defined as languages present in more than one country and as an equivalent version of the "*base language*" (which is the same as the standard of the given country's mother tongue). Rudolf Muhr[60] considers polycentric languages as a separate category between languages and dialects. As they function as administrative and state languages or as regional official languages, they have linguistic and communicative autonomy. Muhr considers German, English, French, Greek, Italian, Dutch, Portuguese, Spanish, and Swedish as polycentric European languages.

The particular importance of such languages in the domain of legal texts is because, their different usage and traditions, the same legal term is often captured by different lexical expressions in the countries using them. One of the most iconic examples of this is German, where, for instance, in Austria and Germany, several legal concepts are referred to by separate terms, e.g.: *Jus* (Austria) vs. *Recht* (Germany) is the term refers to *Law*, Ehepakt (Austria) vs. Ehevertrag (Germany) refers to *Marriage Contract* (~Postnuptial agreement) and *Arbeitsbewilligung* (Austria) vs. *Arbeitserlaubnis* (Germany) refers to *Work Permit* or visa, if the country in the context of which the term is used is subject to visa requirements (cf. Markhardt: 2005 [61]).

To stay within the EU, this code system lists DEU as German (Germany) and DEA as German (Austria). Clearly, differences exist between the legal written languages used in Germany and Austria as well[62]. Such polycentric languages[63] may require special attention when training AI. However, differences may equally be viewed as language or jurisprudence differences. In both cases, transfer learning will pre-train AI using the larger text corpus, and fine-tune using the smaller corpus. The practical process will largely be the same, and the result will be two differently trained AI modules.

Legal language specifics

The way legal language is represented may differ substantially from everyday language. Even for AI training, except for character-level solutions, sentences and paragraphs are important clues. Regulatory texts and to some extent contracts are characterised by complex sentences including

[59] ibid.

[60] Muhr, Rudolf 2003. Die plurizentrischen Sprachen Europas – Ein Überblick. In: Gugenberger, Eva –, Blumberg, Mechthild (Hrsg.): Vielsprachiges Europa. Zur Situation der regionalen Sprachen von der Iberischen Halbinsel bis zum Kaukasus. (= Bd. 2 Österreichisches Deutsch – Sprache der Gegenwart.). Peter Lang Verlag. Frankfurt am Main. 191-233.

[61] Markhardt, Heidemarie. 2005. *Das österreichische Deutsch im Rahmen der EU.* Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

[62] Muhr, R. (2009). The differences in the legal terminology of Austria and Germany and the results for the German legal terminology within the scope of the European Union. Muttersprache, 119(3), 199–216. https://www.researchgate.net/publication/289292621_The_differences_in_the_legal_terminology_of_Austria_and_Germany_and_the_results_for_the_German_legal_terminology_within_the_scope_of_the_European_Union accessed 5 September 2021.

[63] Muhr, R., Mas Castells, J.-A., & Rueter, J. M. (2019). European pluricentric languages in contact and conflict (R. Muhr, J. Angel, M. Castells, & J. Rueter, Eds.; Vol. 21). Peter Lang. https://www.peterlang.com/view/9783631803097/html/ch10.xhtml accessed 5 September 2021

numbered lists citations etc. Sanchez [64] and Glaser et al. [65] provide details concerning US and German law, respectively. Because of numbered (and lettered) lists, sentences often have inside end-of-line characters; because of serial numbers as well as abbreviations, sentences may have internal full stops in some languages. Paragraphs may cross page boundaries, a common difficulty for OCR of any text type. Roman numbers may not be uncommon in everyday legal texts even these days . Latin text may be found in legal documents.

It is customary to abbreviate or nickname entities like such as parties, regulatory acts etc. with their full name mentioned only once in the text (“*hereinafter referred to as*”, e.g., Anti-Corruption Office, Latvia (‘the KNAB’)). Anonymisation may leave ellipses (…) or fullstop sequences (……), and there is a role for {} and [] brackets and other special characters (e.g., §►◄«») not to mention the continued use of“/text/” boundaries in place of paired brackets.

On the positive side, only a small number of typos are expected in legal texts, except perhaps for raw drafts, a rare training set. The morphological diversity of legal language is certainly smaller than in everyday speech, lacking informal addressing (or addressing at all), 1st and 2nd persons, and fewer verbs etc. On the other hand, legal texts incorporate other (regulated or contracted) professions’ special terms. This points to a *de novo* pre-training rather than cross-domain training from everyday language.

Document size and sentence length are considerable factors in AI training and must be reckoned with.

The remedy is special text preparation before AI training and special configuration of the neural network to handle document and sentence size.

Also, simpler neural network models can be trained with less text and may work quite satisfactorily. Technologies to reduce the need for large data sets were summarised by Maheswari[66]; for later developments see e.g., Kim et al.[67] and Riekert et al. [68].

The black box and lawyers’ responsibility

Deep Learning results in black box (non-transparent) neural networks that are unable to explain their results. Even if those results are reliable as revealed by extensive testing, the statistical nature of the process itself always leaves some doubts in the particular case. It is difficult to imagine an SLF filing a lawsuit because the program said so. This holds true even in the light of automatic risk-assessment (of the likelyhood of a defendant becoming a recidivist) systems like COMPAS in the US[69] and especially in

[64] George Sanchez, ‘Sentence Boundary Detection in Legal Text’, Proceedings of the Natural Legal Language Processing Workshop 2019 (Association for Computational Linguistics 2019) http://aclweb.org/anthology/W19-2204 accessed 5 September 2021.

[65] Ingo Glaser, Sebastian Moser and Florian Matthes, ‘Sentence Boundary Detection in German Legal Documents’: Proceedings of the 13th International Conference on Agents and Artificial Intelligence (SCITEPRESS - Science and Technology Publications 2021) https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010246308120821 accessed 5 September 2021.

[66] Jyoti Prakash Maheswari, ‘Breaking the Curse of Small Data Sets in Machine Learning: Part 2’ (Medium, 27 April 2019) https://towardsdatascience.com/breaking-the-curse-of-small-data-sets-in-machine-learning-part-2-894aa45277f4 accessed 13 September 2021.

[67] Kang-Min Kim and others, ‘From Small-Scale to Large-Scale Text Classification’, The World Wide Web Conference (ACM 2019) https://dl.acm.org/doi/10.1145/3308558.3313563 accessed 5 September 2021.

[68] Martin Riekert, Matthias Riekert and Achim Klein, ‘Simple Baseline Machine Learning Text Classifiers for Small Datasets’ (2021) 2 SN Computer Science 178 https://link.springer.com/10.1007/s42979-021-00480-4 accessed 5 September 2021.

[69] ‘Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing’ (UCLA Law Review, 19 February 2019) https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/ accessed 5 September 2021.

the light of the shutting down of the Dutch alternative dispute resolution (ADR) initiative e-Court[70] in 2018. Actually, the latest EU whitebook on AI (European Commission, 2020) says "*Combining symbolic reasoning with deep neural networks may help us improve explainability of AI outcomes*"[71]. An additional requirement may be that the explanation should come in natural language (possibly supported by graphs and other visualisation devices, if we may add.)

In accordance with these goals, the Europen Research Council initiated a project with the title of „Science and technology for the explanation of AI decision making" within a time span of 1 October 2019 – 30 September 2024. The project was awarded to Italian unversities and research bodies. Its counterpart, Interactive Natural Language Technology for Explainable Artificial Intelligence, has a larger budget and a more international participant list of 10 institutions.

Regarding legaltech, there are excellent and detailed surveys of the problem which offer potential remedies, although some of them seem to be outside the scope of an SLF ("*Utilize multiple AI systems*")[72]. Sutherland raises some additional points[73].

Technological differences between civil and common law countries

Since a large portion of scientific and professional publications originate from countries with Common law systems, the legitimate question was raised whether the statements in the present study have any implications on practitioners in countries with a civil law system. From the pure computational linguistical point of view there are no major implications. All NLP tools and methods can be applied in both domains. As to applying NER, finding similarities in text corpora, such as among various cases and norms, comparing the dates of the origin of texts, the methods are similar. However, the differences between civil and common law systems have a significant impact on legal research and particularly legal prediction. Although it is also necessary to search judgements in civil law systems, the codified statutes upon which these judgements and opinions are based might have changed significantly since they were considered in the earlier cases. Hence, AI systems used for legal research, and particularly legal prediction, must be capable of identifying whether and the extent to which prior judgments remain relevant under the present (codified) law. Also, as there is no formal system of precedent in continental law, codified laws and arguments based on them have a leading influence. Therefore, diverging judgements are more likely.

Small languages as a technical barrier

Machine learning tasks always require large text corpora. While a language model can be built by collecting only the Wikipedia entries for a given language, a more specific task (e.g., legal NER) requires specific corpora. These may have to be annotated manually, which may require the involvement of linguists and IT experts. For example, there are different ways of preparing a corpus for sentiment analysis and meaning-clarification tasks, and different levels of expertise are required to do this. The availability of such expertise may be more limited for languages with few speakers.

[70] Willemien Netjes and Arno R Lodder, 'e-Court – Dutch Alternative Online Resolution of Debt Collection Claims' (2019) 6 International Journal of Online Dispute Resolution 96 https://www.elevenjournals.com/tijdschrift/ijodr/2019/1/IJODR_2352-5002_2019_006_001_005 accessed 5 September 2021.

[71] 'Commission-White-Paper-Artificial-Intelligence-Feb2020_en.Pdf' https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf accessed 5 September 2021.

[72] Ronald Yu and Gabriele Spina Alì, 'What's Inside the Black Box? AI Challenges for Lawyers and Researchers' (2019) 19 Legal Information Management 2 https://www.cambridge.org/core/product/identifier/S1472669619000021/type/journal_article accessed 5 September 2021.

[73] 'Some Thoughts on Black Box AI and Law' (Slaw, 18 August 2021) http://www.slaw.ca/2021/08/18/some-thoughts-on-black-box-ai-and-law/ accessed 5 September 2021.

For small languages, even if the material needed to build a language model is available online, there is often no corpus available for more specific tasks, or no publicly available corpus exists from which, for example, a fine-tuning of a model can be trained. This could be applicable to comparable small communities of a polycentric language.

If the individual needs of smaller companies are to be considered, the lack of commonly available resources increases costs (each company must create the resources for teaching itself) and the different quality of the material used for teaching can lead to very different quality of the result.

Human and organisational barriers

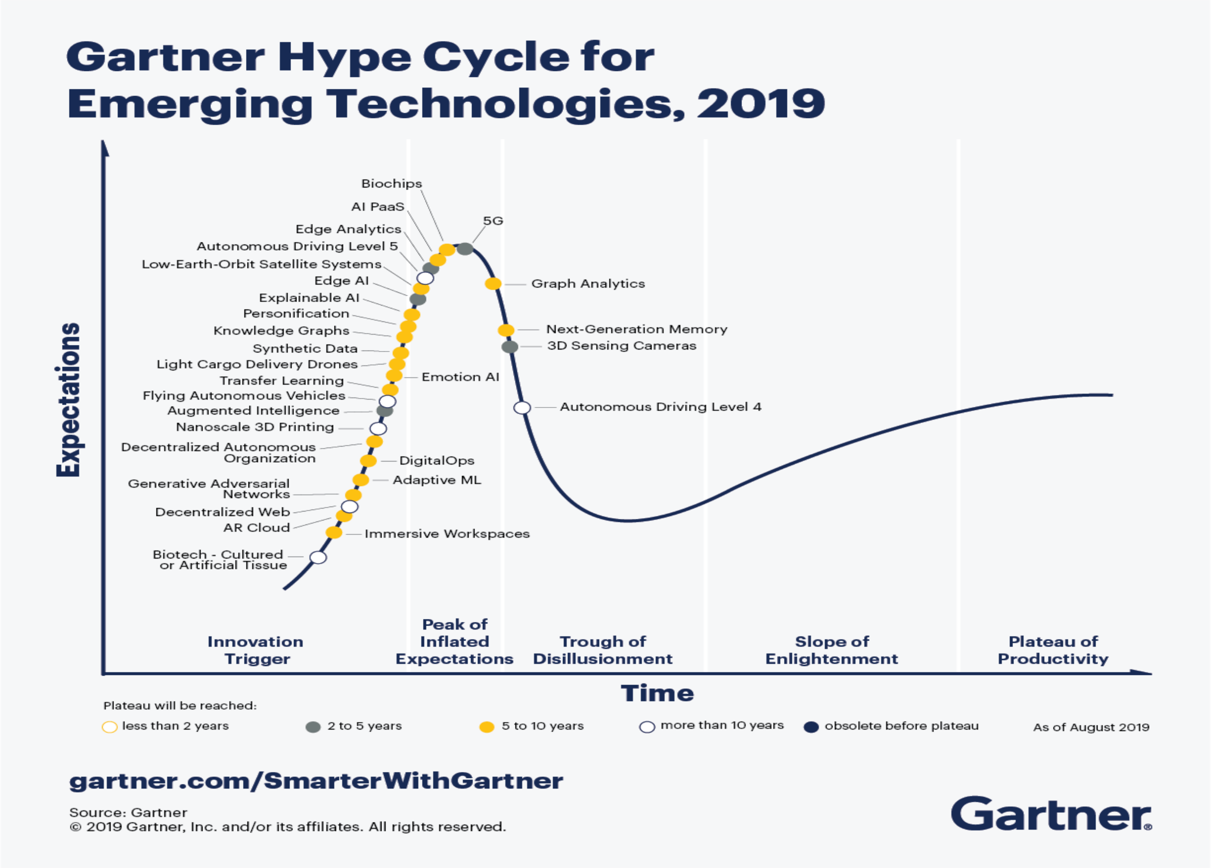
Organisational barriers

The Gartner Hype Cycle

The acceptance of technological developments, particularly in the ICT industry, shows a typical pattern, the so-called Gartner Hype Cycle. As we have seen in the case of handwriting recognition, the dotcom bubble, and the Y2000 turn, the absorption of AI shows a similar curve, as can be seen in Figure 4. The initial enthusiasm, often boasted by the press and the investment community, reaches a peak, at which point the limitations of the technology, often still immature, begin to be experienced. This is followed by a steep slump of disillusionment, from which realistic industrial growth emerges slowly much later, reaching a plateau of market saturation. What is not shown in Figure 4, however, is that the curve can have ups and downs, as happened to AI at the turn of 70's and the 80's and in the early 90's - the so-called AI winters. Since all realistic analysts agree that **NLP for legaltech is very much overhyped**, we consider it very relevant to mention this, to calm exaggerated expectations and to save potential legal clients from unpleasant surprises.

The Gartner Hype Cycle is loosely related to the Rogers Diffusion Curve which illustrates how, why and at what rate ideas or technologies spread. The subject is discussed by W.D. Henderson[74]. While the Gartner Hype demonstrates the expectations, the Rogers Diffusion Curve illustrates the market share of different kind of innovators; both show their results in the function of time. If we assume that the Rogers Diffusion Curve is not quite a symmetric bell curve, then one might say that the area under the Gartner Curve is continuously filled over time with the innovators appearing under the Rogers curve. See Fig. 5.

[74] William D Henderson, 'Innovation Diffusion in the Legal Industry' 122 DICKINSON LAW REVIEW 85, 458.



[75]

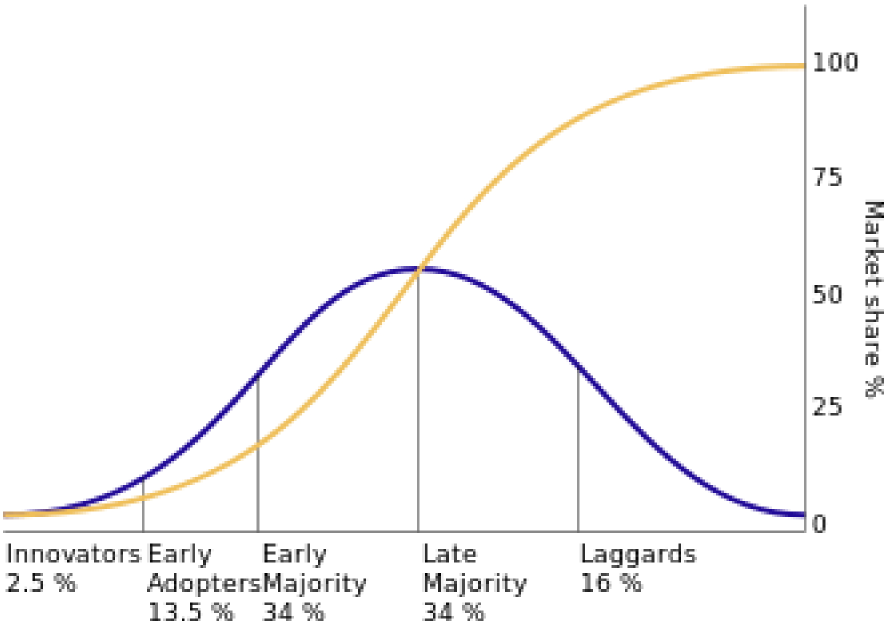


Figure 5.: the Rogers Diffusion of Innovation Curve. Source: Wikipedia

[75] Panetta, Kasey: 5 trends appear on the Gartner hype cycle for emerging technologies, 2019. Gartner. Online: http://gtnr.it/3uuYs4J

Small languages as an organisational barrier

Ethnology defines a small language as one with fewer than 10,000 speakers[76]. However, for the purpose of this study, to apply the ethnological meaning of the term would make no sense when discussing digital presence. Moreover, even if a language has millions of speakers, the population may be divided into several jurisdictions, so that the actual homogenous (in the sense discussed here) population may be even smaller. See the paragraph on polycentric languages.

A major problem is caused by the fact that small languages (including small polycentric languages) do not hit the break-even point of the economies of scale for manufacturers and service providers. The business case is far too weak for investment in applications running with small languages. In other words, the return on investment for an application or even just an upgrade of the training data for better performance for a large language/population/market is much higher than for a small one. The result is that the choice and the quality of applications for small languages is vastly poorer than that for the large ones. This is no place to make a comparison of commercial products, but a quick look at the number of articles in the English and Latvian Wikipedia will be convincing enough (2,567,509 versus 17,527)[77]. Another example is Microsoft Word's style guidance (File» Options» Proofing» When correcting spelling and grammar in Word» Settings). One only needs to see the abundance of options for English or German, and the scarcity in Hungarian, or even the complete lack of options in Latvian (as of 20th August 2021). The market leader grammar and style checker, Grammarly, is not even available in any other language but English[78] [79].

UNESCO defines four levels of language endangerment (after "*safe*"), as follows[80]:

- Vulnerable - "most children speak the language, but it may be restricted to certain domains (e.g., home)"
- Definitely endangered - "children no longer learn the language as mother tongue in the home"
- Severely endangered - "language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves"
- Critically endangered - "the youngest speakers are grandparents and older, and they speak the language partially and infrequently"
- Extinct - "there are no speakers left; included in the Atlas if presumably extinct since the 1950s".

The digital "*stamina*" of a given language depends on various factors, such as its community size, prestige, identity function, financial background, level of education, political lobby power, technological sophistication etc. A language can have a relatively small population, but, like the African

[76] 'Size and Vitality of Shan' (Ethnologue) https://www.ethnologue.com/size-and-vitality/shn accessed 24 August 2021.

[77] 'Wikipedia: Multilingual Statistics', Wikipedia (2020) https://en.wikipedia.org/w/index.php?title=Wikipedia:Multilingual_statistics&oldid=944342501 accessed 22 August 2021.

[78] Robert Dale and Jette Viethen, 'The Automated Writing Assistance Landscape in 2021' (2021) 27 Natural Language Engineering 511 https://www.cambridge.org/core/journals/natural-language-engineering/article/automated-writing-assistance-landscape-in-2021/E1B54FD65963E65D46EF440B5A13F186 accessed 21 August 2021.

[79] 'Does Grammarly Support Languages Other than English?' (Grammarly Support) https://support.grammarly.com/hc/en-us/articles/115000090971-Does-Grammarly-support-languages-other-than-English- accessed 21 August 2021.

[80] 'List of Endangered Languages in Europe', , Wikipedia (2021) https://en.wikipedia.org/w/index.php?title=List_of_endangered_languages_in_Europe&oldid=1035831048 accessed 20 August 2021.

Mandinga (1.35 million speakers), have a literacy rate below 1% and zero Wikipedia presence. The Romani language is spoken by 1.5 million people in Europe and there is no Wikipedia in Romani (although there are 690 articles in Vlax Romani, a South-European variant). Estonian, however, with 1,165 million speakers, has 221,000 articles. Of the 7139 spoken[81] (in 2016), and continuously diminishing number of languages 321 have at least 1 article in Wikipedia[82]. The Universal Declaration of Human Rights has been translated into 530 languages.[83] One could thus assume that 4.5% of languages worldwide may have some digital appearance. Considerable NLP is likely to be a portion of this figure.

A seminal work, Digital Language Death[84] arrives at the conclusion, that of the roughly 7000 languages still alive 2500 may survive for another century, and the **number of estimated digital survivors is only 250 worldwide**.

There are about 70 spoken and 24 official languages in the European Union[85]. Some of these languages are spoken by a small community. We have found no data on the digital presence of all these 70 languages, let alone their level of NLP sophistication. This has been taken into account in this study since the main barrier to employing NLP tools by SLFs is – apart from the necessary software applications – **the lack of sufficient digitally available training corpora**.

Alternative legal advisory services

Alternative Legal Advisor Services (henceforth ALSPs provide quasi legal services with a high-level NLP-based technical arsenal over the internet[86]. Having invested and specialised in these disruptive technologies they are able to function at an extremely competitive price level. ALSP robots can answer questions commonly asked by clients without them needing to speak to a lawyer either in the company's legal department or in an external legal firm.

In the US legal market, a tendency is being sensed that some corporate legal departments tend to trust ALSPs more than legal firms on legaltech recommendations[87]. This also means that ALSPs bite into the market of law firms, or in other words, they are a threat or at least competition in certain areas[88].

Chatbots as a possible alternative to human counselling

Chatbots are not only significantly cheaper than a human workforce, but they are also always available, and store crucial information once acquired until deleted, unlike humans, who take it with themselves when leave, and so newcomers must be trained again.

[81] 'How Many Languages Are There in the World?' (Ethnologue, 3 May 2016) https://www.ethnologue.com/guides/how-many-languages accessed 20 August 2021.

[82] 'List of Wikipedias - Meta' https://meta.wikimedia.org/wiki/List_of_Wikipedias accessed 20 August 2021.

[83] As of November 22, 2021, In: https://www.ohchr.org/EN/UDHR/Pages/SearchByLang.aspx

[84] András Kornai, 'Digital Language Death' (2013) 8 PLOS ONE e77056 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077056 accessed 20 August 2021.

[85] 'Languages of Europe', Wikipedia (2021) https://en.wikipedia.org/w/index.php?title=Languages_of_Europe&oldid=1036009595 accessed 20 August 2021.

[86] 'What Is an Alternative Legal Service Provider (ALSP)? | Paralegals in ALSPs' (16 April 2018) https://www.paralegaledu.org/alternative-legal-service-providers/ accessed 17 August 2021.

[87] 'Legal Departments Likely Trust ALSPs More Than Firms on Tech Recommendations' (Legaltech News) https://www.law.com/legaltechnews/2021/08/04/legal-departments-likely-trust-alsps-more-than-firms-on-tech-recommendations/?kw=Legal%20Departments%20Likely%20Trust%20ALSPs%20More%20Than%20Firms%20on%20Tech%20Recommendations accessed 13 August 2021.

[88] ABA Journal, 'Who's Eating Law Firms' Lunch?' (ABA Journal) https://www.abajournal.com/magazine/article/whos_eating_law_firms_lunch accessed 19 August 2021.

The quality of chatbots is constantly improving. The more complex the problem, the more probable that the request must be passed on to a human. Up to a certain point chatbots can stand the Turing test.

Gartner forecast in 2019 that legal chatbots, so called *lawbots* (virtual legal assistants run by AI), will handle one third of enquiries in 2023[89]. Even if this prophecy may well be too optimistic/pessimistic, chatbots are biting more and more into the customer service cake, and there is no reason to assume that it will be different in the legal field[90].

Small vs. larger firms

Another question is whether SLFs could lose their market share to larger legal firms. Since advanced technology is investment intensive, smaller organisations cannot afford top tier solutions and fall behind because of a lack of efficiency. It is too early to draw consequences even from Anglo-Saxon legal markets; one may observe analogies from other professions.

A study prepared for the *Deutscher Anwalt Verein* predicted that by 2030 the survival of legal offices may be a matter of the size of the office.[91] The study does not elaborate on the subject, nor does it communicate estimates on any survival rate, but it does generally emphasise the importance of adaption to the trends of digital transformation. The Bellwether Report by LexisNexis from 2019[92] cites various challenges facing SLFs, such as growth potential; nevertheless, it is largely optimistic and does not even mention technology as a crucial factor, let alone NLP.

Migration to better supported linguistic environments

Law firms may opt to work in a major language, possibly in the best supported, lingua franca, English, or even more American English, if a proper NLP arsenal is not at hand in their native small language environment. This is, of course, only possible in cases where the court of arbitration, jurisdiction and language can be chosen and such choice would be in the best interest of the client.

Poor supplier support

Overenthusiastic start-ups may not be able to give sufficient support, or even disappear. On the other hand, mature, solid companies' pricing may not be palatable or even bearable to SLFs.

The EU is disadvantaged compared to the USA or China

EU countries are disadvantaged in several ways when compared to the USA and China. The EU market is divided by different jurisdictions, the regulatory burden is perceived as high. For example, despite all its advantages, GDPR might have a negative influence on VC investments in AI based start-ups[93]. The major problem is, however, that of the different languages and jurisdictions. This makes the market fragmented because products are not easily transferable from one country to another, and

[89] 'Gartner Predicts That by 2023 "Lawbots" Will Handle a Quarter of Internal Legal Requests' (Gartner) https://www.gartner.com/en/newsroom/press-releases/2020-01-30-gartner-predicts-that-by-2023-lawbots-will-handle-a-quarter-of-internal-legal-request accessed 8 August 2021.

[90] 'Applications for Legal Chatbots' (Law Technology Today, 13 July 2020) https://www.lawtechnologytoday.org/2020/07/applications-for-legal-chatbots/ accessed 18 August 2021.

[91] Prognos AG. 'Executive Summary der Rechtsdienstleistungsmarkt 2030' https://anwaltverein.de/de/anwaltspraxis/dav-zukunftsstudie?file=files/anwaltverein.de/downloads/service/DAV-Zukunftsstudie/DAV-Zukunftsstudie-Executive-Summary.pdf p. 19. accessed 11. August 2021.

[92] 'The Bellwether Report 2019: Is the Future Small?' https://www.lexisnexis.co.uk/research-and-reports/is-the-future-small-bd.html accessed 14 August 2021.

[93] Jian Jia, Ginger Zhe Jin and Liad Wagman, 'The Short-Run Effects of GDPR on Technology Venture Investment' (National Bureau of Economic Research 2018) w25248 http://www.nber.org/papers/w25248.pdf accessed 13 September 2021.

only non-homogeneous training data is available. As result, major suppliers are not interested in localising their products into small markets.

The costs and efficiency

The major challenge the legal sector - and particularly SLFs – might be facing is that in some areas functions hitherto performed by humans can be performed by NLP based applications at some point. The technologies and application fields are discussed in detail separately. SLFs that do not take advantage of these disruptive technologies can potentially lose their clients to those that do and can work more efficiently.

Running costs

Apart from the skill shortages the costs of implementation, training and support of still pricy software applications are high. Furthermore, the associated costs are easier to spread in larger legal firms with larger transactions that in SLFs, where the business case cannot be established. Moreover, the number of smaller legal tasks that can otherwise be easily automated do not reach the level of economies of scale in SLFs.

Investment

Developing, maintaining, and supporting effective NLP applications needs a high level of capital concentration which SLFs are less likely to afford. Furthermore, the more advanced the market, the higher the concentration of capital. To make any legaltech investment feasible, a locally different, but very much existent market size is inevitable. SLFs will have to make considerable efforts to differentiate themselves.

Legal liability

Legal liability for NLP applications (or rather of the humans and organisations running them) is a topic, which is not closed for discussion, though such discussion lies outside the scope of this paper. . However, mentioning legal liability for NLP applications as a possible obstacle is something that cannot be avoided.

Human factor

The future of professions in the light of new disruptive technologies

Different views have been published on the future of professions, among them the legal profession.

On the one hand, McKinsey suggests that by 2030, as the worst-case-scenario, up to 800 million jobs could disappear, and 375 million people will have to look for another occupation[94]. MIT Technology Review predicts that one tenth of white-collar workplaces will be taken by robots in the not too far future[95]. Gartner predicted in 2019, that by 2023 a quarter of internal requests will be handled by lawbots or virtual legal assistants, VLA's, substituting the paralegals who have been doing the job so far[96].

On the other hand, more optimistic gurus forecast no significant changes in the number of jobs, but rather, a significant shift in the structure of the job market: routine work will be performed more and

[94] McKinsey & Company . ' Jobs lost, jobs gained: workforce transitions in a time of automation.' [2017] https://www.mckinsey.com/~/media/BAB489A30B724BECB5DEDC41E9BB9FAC.ashx
[95] Will Knight. 'MIT Is Technology About to Decimate White-Collar Work?' (MIT Technology Review) https://www.technologyreview.com/2017/11/06/147955/is-technology-about-to-decimate-white-collar-work/
[96] Gartner Predicts That by 2023 "Lawbots" Will Handle a Quarter of Internal Legal Requests. https://www.gartner.com/en/newsroom/press-releases/2020-01-30-gartner-predicts-that-by-2023-lawbots-will-handle-a-quarter-of-internal-legal-request accessed 5 September 2021

more by robots, whereas time will be freed for humans to do more sophisticated creative work. The view of the present authors leans towards the latter, i.e., not disagreeing with McKinsey, but just taking a somewhat less dystopian view. Table 1. shows that the most industrialised countries with highly automated manufacturing still have a relatively low unemployment rate. The limited availability of a qualified workforce *per se* can be a strong motivation for automation. No reliable data could be found on the automatization level of the least developed countries; however, one can confidently assume that it is nowhere close to that of those in the upper group.

An important question is whether the legal profession would lose workplaces in general due to robotisation. Since the penetration of AI into the legal world is a long way behind that of, for example, radiology, an early adopter of AI in the medical field (not NLP but image processing Recurrent Neural Network, RNN, and Decision Tree classifier), it is worth looking at the analogy. AI based tumour detection performance is well above that of humans. Still, a Chinese study concludes that radiologists **are not losing their jobs but are just able to focus on the tasks that need higher qualifications**[97]**. The table underneath reflects employment trends in manufacturing and does not intend to suggest that there will be no painful structural changes in several areas and professions, such as in the legal field. We found no data on any job losses to date due to NLP.**

	Automatization	Unemployment
Republic of Korea	631	4,2
Singapore	488	3,6
Germany	309	4,4
Japan	303	2,9
Sweden	223	8,9
Denmark	211	5,7
USA	189	6,2
Italy	185	10,7
Belgium	184	5,3
Taiwan	177	3,8
Burkina Faso		77
Syria		50
Senegal		48
Haiti		40,6
Kenya		40
Djibouti		40
Marshall Islands		36
Namibia		34
Kiribati		30,6

Table 1: The first 10 countries in terms of the number of installed robots per 10,000 employees in the manufacturing industry 2016, and the 10 countries with the highest level of unemployment. Source: The Robotreport[98], Wikipedia[99], assembled by the author.

[97] Niklas Muennighoff, 'Diagnosing the Impact of AI on Radiology in China' [2021] arXiv:2106.07921 [cs] http://arxiv.org/abs/2106.07921 accessed 10 August 2021.

[98] Stewe Crowe. '10 Most Automated Countries in the World' (therobotreport.com) https://www.therobotreport.com/10-automated-countries-in-the-world/ accessed 5 September 2021.

[99] 'Wikipedia. 'List of countries by unemployment rate' https://en.wikipedia.org/wiki/List_of_countries_by_unemployment_rate accessed 22 August 2021.

Skilled HR

The level of ICT skills and the readiness to catch up of small law firms is insufficient. There is a shortage of skilled ICT personnel. Legal companies and legal departments of large organisations compete for scarcestaff. It may be that some SLFs will be unable to compete with the large firms in obtaining the services of these specialists.

Wolters Kluwer found that 59% of corporate lawyers expect AI to have a significant impact, yet **just 22% understand these technologies**[100].

Furthermore, legaltech courses are a rarity even in Anglo-Saxon universities, let alone in European ones. The legaltech syllabus often does not stretch to data base search techniques. Often the informatics of legal matters is interchanged with courses on the legal aspects of informatics.

Short term versus long term priority

Though all known survey studies emphasize the recognition of the importance of legaltech by legal firms and legal departments, most of them stress the inertia with which a high enough priority is given to the very implementation of legaltech applications. The short-term tactical priority of doing chargeable work often kills the long-term strategic priority of investment[101].

Lack of knowledge and experience

The lack of sufficient knowledge about legaltech, and particularly AI and NLP based applications, among senior staff can be identified as a hindrance to the proliferation of technologies.

Lack of trust, resistance to change

Fear of the unknown, a lack of proper training and knowledge and bad experiences with hastily introduced, misfitting, undeveloped software products with complex and hard-to-navigate user interfaces, too many false positive and - in legal praxis - extremely painful false negative, results, can lead to apprehension, distrust and thus, no adoption of new technologies. Due to a lack of understanding of how AI works, false expectations result in dissatisfaction. AI does not function as a general ledger module, and some applications need updates of training data, fine-tuning, and feedback. A lack of maintenance can cause deteriorating results and thus dissatisfaction[102]. Wolters Kluwer's 2020 survey found that change management difficulties and leadership resistance to change is the major barrier to implementing change for law firms (53%) and corporate departments (65%)[103]. Although we have not found any data relating specifically to SLFs, WK's data on legal departments still gives a clear indication of resistance factors (Figure 6.). An interesting comparison highlights the growing awareness of the lack of understanding and skills (Figure 7.)

[100] '2020 Wolters Kluwer Future Ready Lawyer: Performance Drivers and Change in the Legal Sector' 5 https://www.wolterskluwer.com/en/news/2020-wolters-kluwer-future-ready-lawyer-performance-drivers-and-change-in-the-legal-sector accessed 23 August 2021.

[101] 'Barriers to Legal Technology Adoption.Pdf' 5 https://events.legatics.com/hubfs/Barriers%20to%20Legal%20Technology%20Adoption.pdf accessed 22 August 2021.

[102] '2016 ABA Future of Legal Services -Report-Web.Pdf' 17 https://www.srln.org/system/files/attachments/2016%20ABA%20Future%20of%20Legal%20Services%20-Report-Web.pdf accessed 14 August 2021.

[103] '2020 Wolters Kluwer Future Ready Lawyer: Performance Drivers and Change in the Legal Sector' (n 100) 4.

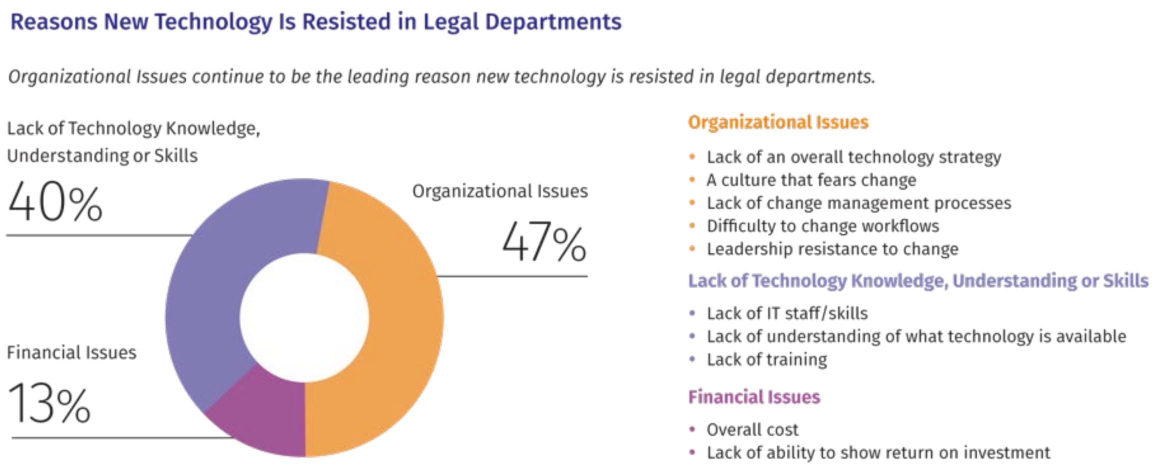


Figure 6. Reasons new technology is resisted in legal departments. Source: The 2020 Wolters Kluwer Future Ready Lawyer [104].

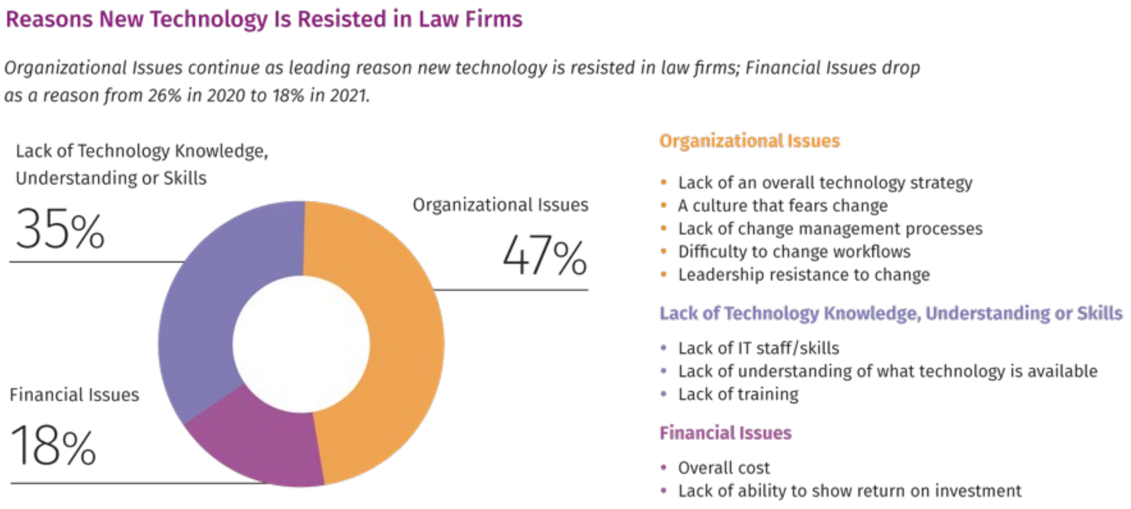


Figure 7. Reasons new technology is resisted in legal departments. Source: The 2021 Wolters Kluwer Future Ready Lawyer [105].

Ethical considerations, algorithmic biases

AI and as such, NLP, cannot be perfectly free of biases. The main reason is the imperfect model resulting from imbalanced historic training data. One type of AI bias is discrimination based on ethnicity, gender, sexual orientation, gender identity, race, political orientation, financial situation etc.

[104] '2020 Wolters Kluwer Future Ready Lawyer: Performance Drivers and Change in the Legal Sector' (n 100) 4.

[105] '2021 Wolters Kluwer Future Ready Lawyer' https://www.wolterskluwer.com/en/know/future-ready-lawyer-2021 accessed 23 August 2021.

There are several ways of reducing biases; however, complete eradication is theoretically impossible. Karen Hao analyse the COMPAS case and present a stunning paradox on this issue[106].

[106] 'Can You Make AI Fairer than a Judge? Play Our Courtroom Algorithm Game' (MIT Technology Review) https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/ accessed 22 August 2021.

Opportunities

"Now this is not the end.
It is not even the beginning of the end.
But it is, perhaps, the end of the beginning."[107]

In this section of the paper, consideration is given to the technical,the organisational, and the wider opportunities for SLFs through the use of NLP tools.

Technical opportunities

Dedicated NLP tools for SLFs

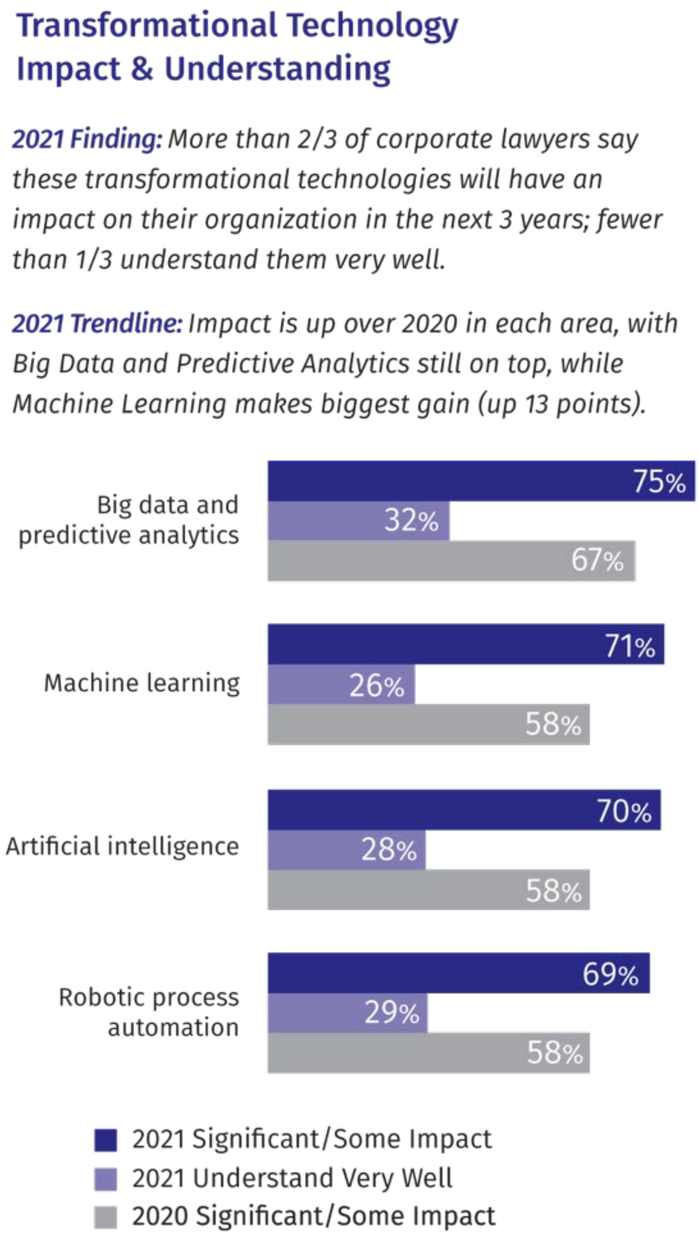
There are several ways to leverage key technologies to pave the way for SLFs and small languages towards up-to-date NLP utilization.

Thorough analysis and testing of popular applications can be carried out at a later stage. Although it is not the subject of the present study, it is highly recommended, even if for nothing else than widening the horizon of decisionmakers.

When managing the digital transition, it is worth considering building specific NLP tools for SLFs. These should be user-friendly, easy to learn how to operate and inexpensive. These tools, most likely functioning as software as service (SaaS), should cover all application fields SLFs need in their everyday business. Crucial modules depending on jurisdiction and local language, however, will have to be trained individually.

Though not SLF specific, the rapidly growing interest is reflected in the findings shown in Figure 8.

[107] Sir Winston Churchill



Transformational Technology Impact & Understanding

2021 Finding: *More than 2/3 of corporate lawyers say these transformational technologies will have an impact on their organization in the next 3 years; fewer than 1/3 understand them very well.*

2021 Trendline: *Impact is up over 2020 in each area, with Big Data and Predictive Analytics still on top, while Machine Learning makes biggest gain (up 13 points).*

Big data and predictive analytics
75%
32%
67%
Machine learning
71%
26%
58%
Artificial intelligence
70%
28%
58%
Robotic process automation
69%
29%
58%
■ 2021 Significant/Some Impact
■ 2021 Understand Very Well
■ 2020 Significant/Some Impact

Figure 8.: Transformational technology impact and understanding. Source: WK 2021[108]

[108] '2021 Wolters Kluwer Future Ready Lawyer' (n 99).

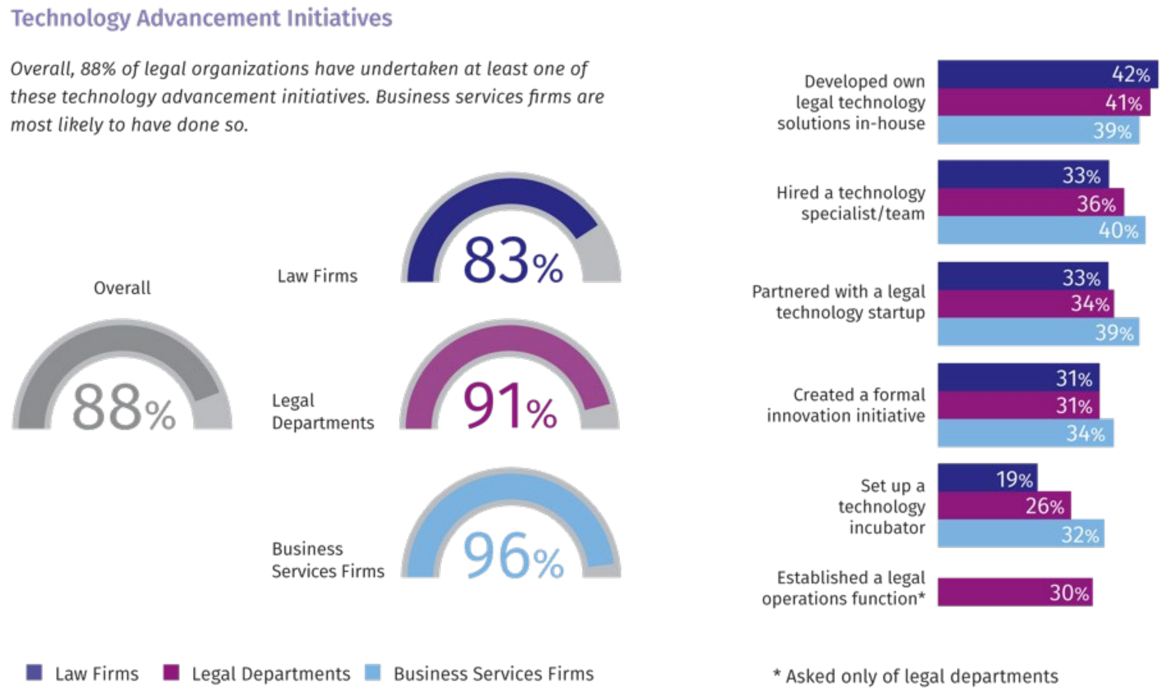


Figure 9.: Technology investment activities. Source: WK 2021[109]

Figure 9. shows the growing awareness and the actions followed. Unfortunately, there is no data on SLFs, but there is a looming discomfort about much less similar activity.

Cross lingual transfer learning (CLTL)

CLTL is one of the technologies that can potentially alleviate small language impediments.

Solution complexity

In the case of an SLF, it seems natural that it wants to keep pace with technological development, but within a single application if possible. There are still monolithic applications that do everything, but their upfront costs may be too high, and the learning curve may be too slow for an SLF. Cloud services still may prove too complex and expensive for SLFs.

A possible solution is a set of microservices on behalf of the solution provider. Each microservice provides a well-defined solution to a well-defined problem, such as anonymisation, entity recognition, finding similar documents etc.

The user is then free to proceed with consuming services as they wish. Of course, subscription/usage-based packages are needed, but if they are formulated with care, the user never pays for something they do not need. Services may also be made available through national/EU funds for free for SLFs.

This approach is very advantageous for SLFs specialising in restricted/single areas of law, e.g., damages, divorce, penalty etc. If similarly specialised microservices are available, these can be fine-tuned to the extreme and provide reliable answers.

With microservices, the user is not forced to work with the provider's user interface. Licence permitting, they themselves, or a third party, can develop user interfaces that suit their needs best.

[109] Ibid.

Building databases from legacy contracts

Legal teams often need to buy contract review software to tell them what is in their own documents. Contracts in house are also a bag of words. However, at least their templates should be available. Using these templates, data can be extracted and organised into a database. The question of writing, for example, a company's name in different ways over time remains.

In other words, the result is a set of key details tables in a database.

At a further stage of development, one must fill in the details table and the text is generated by software.

As legaltech develops, hundreds of applications flourish. Marketplaces have been organized to facilitate comparison and localisation. Software offerings are organised into categories and are searchable.

Security

Security is of the utmost importance since client data is inevitably stored by legal firms[110]. Eventual leakage, for example, caused by hackers or governmental seizure can cause a case to collapse. Larger legal firms can afford to run a fully-fledged internal network, an intranet with appropriate security measures and tools. However, even large firms or corporate legal departments probably cannot completely rely on internally run applications and escape cloud-based tools. For SLFs SaaS cloud-based legaltech applications are a must, and with these sensitive client data can be encrypted by methods US or other government agencies cannot decrypt within reasonable time (dozens or hundreds of years). There is no room to discuss encryption methods and products in this paper. One should, however, note that quantum computing will be able to easily break codes currently considered secure. This disruptive technology will change the way ICT currently operates in many ways.

Security is a substantial barrier to the widespread use of AI- and NLP-based legaltech because legal firms do not, respectively cannot, necessarily trust cloud-based services such as communication channels (TEAMS, Zoom, Skype, GTalk etc.), or file storage platforms (Sharepoint, GDrive, ProtonDrive etc.). Stored data can be covertly accessed by intelligence agencies in the USA or elswhere. Security is the obverse side of comfort or usability. A more secure application or environment can be more troublesome to operate it. Comfort can be the price to be paid for safety. Microsoft introduced Microsoft 365 for Legal to provide a comprehensive platform for legal firms to perform several legaltech functions. However, since the US government has the right to blind subpoena Microsoft, legal firms are not confident about their clients' data being held secure unconditionally under all circumstances[111].

Homomorphic encryption (HE), originally described in a Stanford PhD thesis[112] is an evolving technology. Its use for text mining was described in an academic paper[113], and it has been developed by Microsoft as an open source software component (Microsoft SEAL (Simple Encrypted Arithmetic

[110] 'Small Legal Firms Are Struggling with the Adoption of Digital Platforms' https://www.americanbar.org/groups/business_law/publications/blt/2021/05/digital-platforms/ accessed 22 August 2021.

[111] Caroline Hill. 'Microsoft 365 for Legal –The four phases to transformation' (Legal IT Insider, (legaltechnology.com) https://www.legaltechnology.com/2021/02/11/microsoft-365-for-legal-the-four-phases-to-transformation/ accessed 5 September 2021

[112] Craig Gentry, 'Computing Arbitrary Functions of Encrypted Data' (2010) 53 Communications of the ACM 97 https://dl.acm.org/doi/10.1145/1666420.1666444 accessed 5 September 2021.

[113] Gianpiero Costantino and others, 'Privacy-Preserving Text Mining as a Service', 2017 IEEE Symposium on Computers and Communications (ISCC) (IEEE 2017) http://ieeexplore.ieee.org/document/8024639/ accessed 5 September 2021.

Library)). Microsoft scientists demonstrated that deep learning on homomorphically encrypted data is feasible[114]. There are many other implementations.

Unlike other encryption models used these days, homomorphic encryption is safe from being broken by quantum computers. Fully homomorphic encryption (FHE) is, on the other hand, considered rather slow in practice and not suitable for mass processing.

Secure multiparty computation (SMPC) is only suitable for sharing encrypted information among trusted parties and will not be discussed here.

Differential privacy (DP) enables statistics and machine learning on a dataset while ensuring that information about individual records in the dataset cannot be extracted or inferred. It has been applied to text analytics, most notably BERT pre-training, very recently[115]. It remains to be evaluated for practical purposes.

Human and organisational opportunities

Organisational opportunities

The range of applications to be used in EU SLFs in a few years is limited by the SLFs' cost-bearing capacity, availability of applications/services and expertise, as well as market pressures, including clients' preferences.

These solutions should ideally be implemented, fully functional and seamlessly integrated. Presently, SLFs are usually expected to be proactive, going after new AI solutions based on some internal motivation, or at least reactive to marketing (from legal tech vendors) and market information (e.g., seeing the next-door office prosper and clients diverted). The rate of progress may also depend on administrative and legal labour costs vs. AI costs, which differ widely among EU countries.

Wolters Kluwer's survey in 2020 confirmed that the increasing importance of legaltech is the top trend, with impacts on 76% of respondents across Europe and the US across law firms, corporate legal departments, and business services firms. Only one third of lawyers felt ready for the coming changes[116].

[114] Nathan Dowlin and others, 'CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy' 10.

[115] Xiang Yue and others, 'Differential Privacy for Text Analytics via Natural Text Sanitization' [2021] arXiv:2106.01221 [cs] http://arxiv.org/abs/2106.01221 accessed 5 September 2021.

[116] '2020 Wolters Kluwer Future Ready Lawyer: Performance Drivers and Change in the Legal Sector' (n 100) 1.

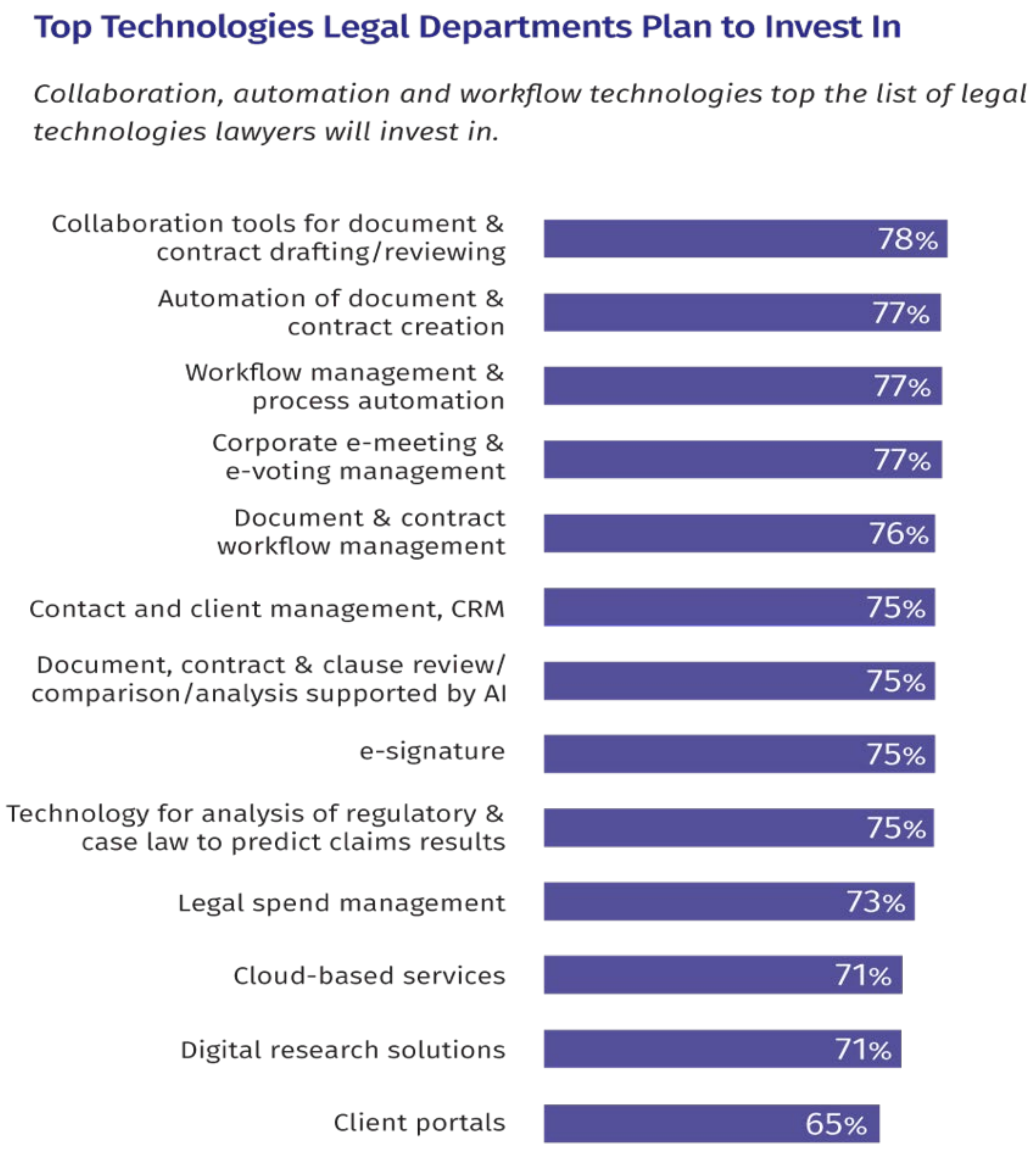
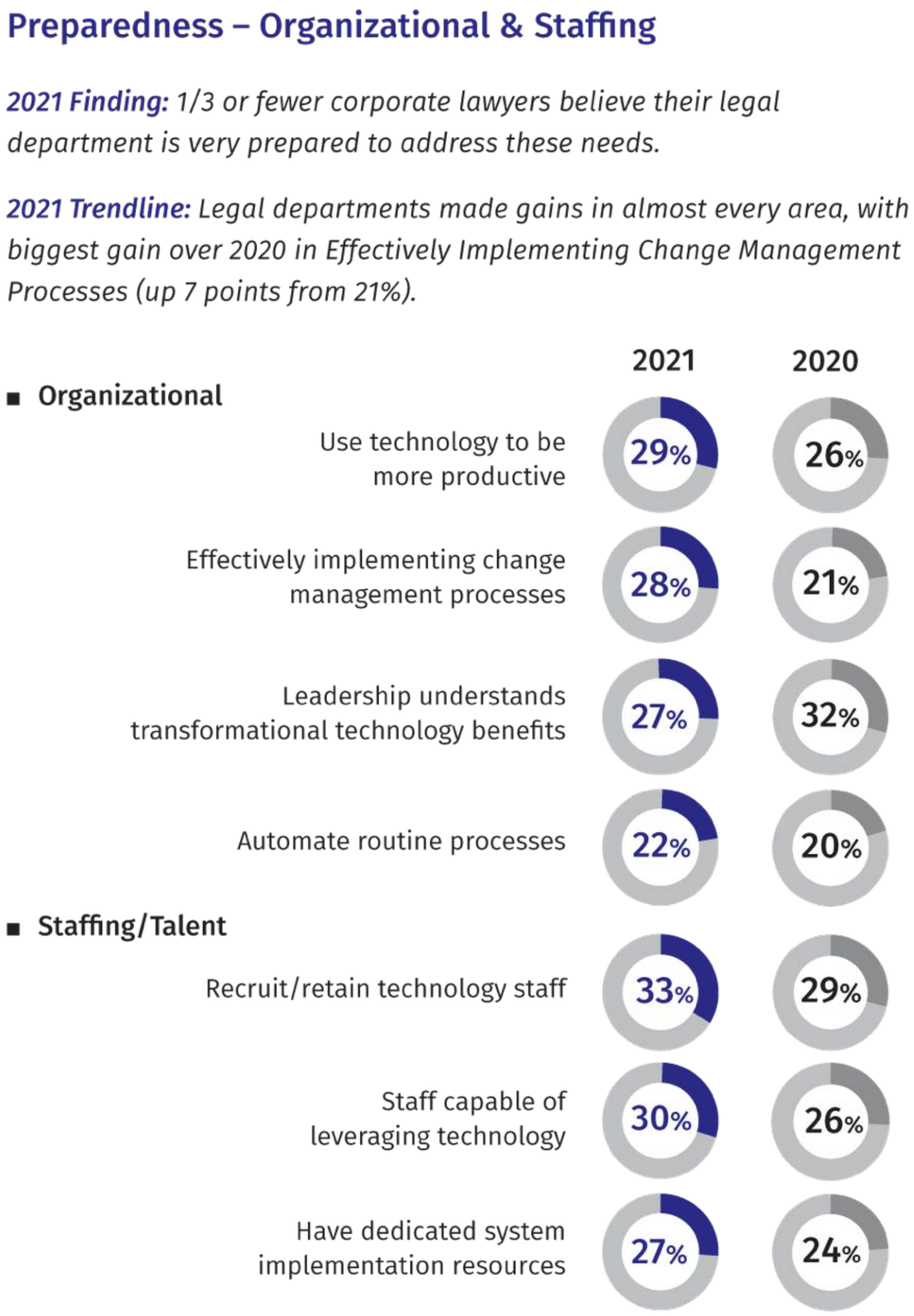


Figure 10.: Top technologies legal departments plan to invest in. Source: [117]

Figure 10 shows the investment priorities in legaltech applications by legal departments.

[117] '2020 Wolters Kluwer Future Ready Lawyer: Performance Drivers and Change in the Legal Sector' (n 100).



Preparedness – Organizational & Staffing

2021 Finding: *1/3 or fewer corporate lawyers believe their legal department is very prepared to address these needs.*

2021 Trendline: *Legal departments made gains in almost every area, with biggest gain over 2020 in Effectively Implementing Change Management Processes (up 7 points from 21%).*

	2021	2020
■ **Organizational**		
Use technology to be more productive	29%	26%
Effectively implementing change management processes	28%	21%
Leadership understands transformational technology benefits	27%	32%
Automate routine processes	22%	20%
■ **Staffing/Talent**		
Recruit/retain technology staff	33%	29%
Staff capable of leveraging technology	30%	26%
Have dedicated system implementation resources	27%	24%

Figure 11.: Preparedness – Organisational and staffing. Source: WK 2021[118]

[118] '2021 Wolters Kluwer Future Ready Lawyer' (n 105).

WK’s survey shows in Figure 11. that though self-confidence is decreasing, change management is improving markedly.

Human factors

Recommendations

As discussed above, NLP solutions need a functional language model, which in turn needs sufficient training data. Therefore, even small languages (including small polycentric languages) need enough legal corpora to feed their language model, something which may just not be available. To generate and preserve the digital presence of these languages in the legal field, these efforts must be supported. Language models, and particularly legal language models for small languages are an absolute must to prevent small populations falling by the wayside.

It is highly recommended that a detailed survey discovers the status of language models generally, and legal ones specifically, country-by-country in the EU, with special focus on the small languages (including small polycentric languages).

Programs should be established for more advanced countries to support others in the development of software and language models. Even more important is to establish **technology transfer** competence centres regionally to train, consult and support SLF staff. **Legaltech education** should be introduced, or the level thereof substantially enhanced in all EU universities. Collaboration among universities should be supported in all known ways.

Conclusion

Little attention has been paid to the problems faced by small law firms (SLFs), which is probably why there is no literature on the subject. As a result of this lack of attention, legaltech literature does not tend to refer to barriers and opportunities which are specific to SLFs. Most papers focus on large law firms and legal departments, or do not distinguish according to the size of the law firm.

This is unhelpful as, in most countries[119] , SLFs account for the majority of lawyers. Moreover, SLFs play an essential role in assisting individuals and small and medium-sized enterprises. Their location, which is much less concentrated than that of the large firms, ensures that both individuals and SMEs have access to local legal services, which access is of considerable importance to them.

This study has identified the obstacles that most SLFs will face in implementing AI tools based on NLP. Although larger firms are likely to face similar barriers, these will be more prevalent in SLFs. In addition, unlike larger firms, SLFs will more frequently face an accumulation of multiple barriers.

The major barriers are not just technological but human, financialand organisational.

In many cases, it can be seen that SLFs lack the human resources to deal with new technologies and working practices. Similarly, these firms may have limited financial resources.

Although various authors mention the appearance of new business models as a consequence of the use of artificial intelligence, it is doubtful that SLFs will be able simply to adopt, these new models easily by themselves.

The lack of human resources also means that SLFs will have to rely on external providers for the natural language tools that have been identified and that look promising.

The price of these tools may also become a barrier to their use by SLFs, while, at the same time, the tools available may not be as well adapted to the requirements of SLFs as they are to the requirements of large firms.

There is a danger that client interest may shift from law firms to alternative legal service providers (ALSPs) due to perceived cost savings. ALSPs often attempt implement cost savings by replacing qualified personnel by automation, even where such replacement might not be justified. As a result, these new providers, which try to focus on automating simple legal services,, are likely to compete more with small firms than with large firms.

Natural language proficiency tools reveal a new risk of inequality, depending on the language used.

Small languages are endangered. Without dedicated efforts, digital presence, even in the legal industry, could diminish.

SLFs are not usually internationally based. They operate only within the national framework or borders. They will be, again, more affected by the difficulties associated with the existence of small languages.

Consequently, specific measures will have to be put in place for SLFs which play a crucial role in providing local legal services to citizens and small and medium sized enterprises. SLFs will need help to benefit from the advances brought about by NLP tools, and then to pass these on to ordinary customers.The emergence of NLP could otherwise be a source of new inequalities for citizens who need legal services.

[119] Including the USA

Further, education for SLFs in the use of NLP tools is crucial. Training and ICT support centres are needed and NLP education could also be addressed in university legal courses.

In short, although NLP tools might create considerable opportunities for growth and development for SLFs, they are also attended with challenges to the continued provision by SLFs of local legal services to SMEs and individuals, particularly: the development of tools which are suited to large, international law firms and which are of little use and relevance to small, local firms; and the development of NLP (especially in the legal sphere) simply passing by those smaller languages where developers do not see a sufficiently large market.

Clearly, targetted and specific measures will be needed to address these and the other issues highlighted in this Report, if the full potential of NLP tools is to be unlocked.

Acknowledgements

We would like to thank Mr. Christian Dirschl of Wolters Kluwer Deutschland GmbH and Gabor Berend of Szeged University for reviewing the paper and providing invaluable suggestions.

We would also like to acknowledge the help of Wolters Kluwer Deutschland GmbH for permission to incorporate the charts from their 2020 and 2021 Future Ready Lawyer Studies.