Extraire automatiquement des informations de décisions des juges aux affaires familiales ?

Julien Barnier

EXTRAIRE AUTOMATIQUEMENT DES INFORMATIONS DE DÉCISIONS DES JUGES AUX AFFAIRES FAMILIALES ?

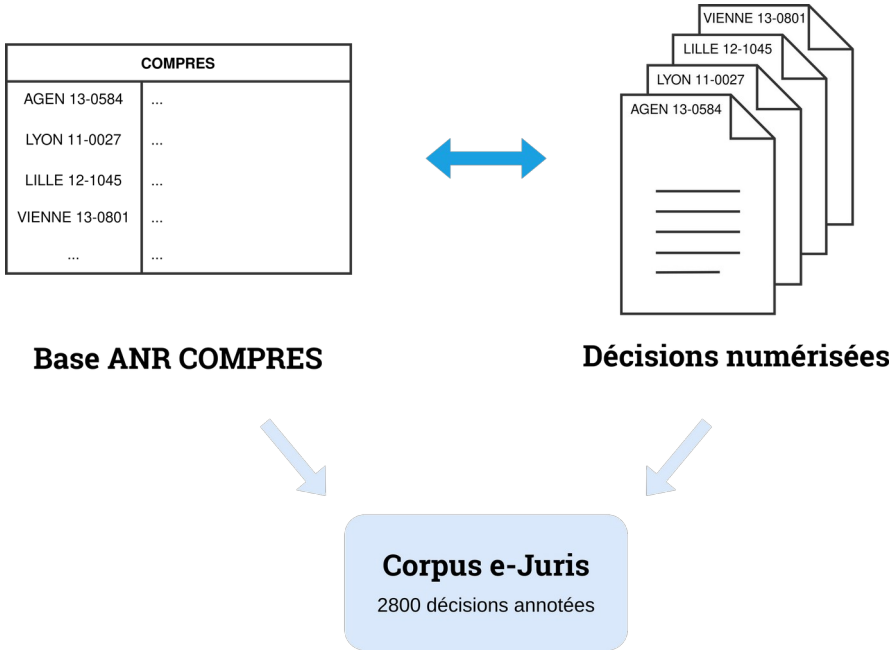
Julien Barnier, Ingénieur CNRS, Centre Max Weber

Démarche et corpus

Le projet e-Juris s'appuie sur une base de données créée dans le cadre d'un projet de recherche antérieur, l'ANR COMPRES. Cette base de données a été élaborée à partir d'un travail manuel minutieux d'analyse de décisions de divorce de première instance. Elle contient des informations extrêmement détaillées notamment sur les caractéristiques des parties, leurs offres et demandes, les décisions du juge, etc. Au final cette base de données comprend plusieurs centaines de variables, résultant d'un travail de lecture attentive et de codage rigoureux et détaillé, pour près de 5500 décisions issues de plus de 150 tribunaux différents, soit la totalité des juridictions compétentes en matière de divorce

La constitution de cette base de données a nécessité des moyens humains et techniques considérables, tant en termes de temps que de coûts. C'est pourquoi une réflexion a été menée sur la possibilité d'automatiser tout ou partie du processus d'extraction d'informations. Ainsi, il a été décidé de numériser l'intégralité des décisions utilisées dans la base COMPRES, qui étaient jusqu'alors uniquement disponibles sous format papier.

Après un travail important de numérisation et de reconnaissance optique de caractères (OCR), nous avons obtenu ce que l'on appelle dans le domaine de l'apprentissage automatique (*machine learning*) un *corpus textuel annoté* : un ensemble de textes numérisés accompagnés de données les concernant. Après avoir exclu les divorces par consentement mutuel, le corpus est composé d'un peu plus de 2800 documents textuels et de leurs données associées.



L'un des axes de travail de l'atelier e-Juris a consisté à utiliser ces données pour tester différentes méthodes d'extraction automatisée d'informations. Cette démarche avait pour objectif de renforcer les compétences des membres de l'atelier sur ces questions et de mieux comprendre ce qui était

envisageable ou réalisable en termes d'automatisation au regard des résultats obtenus par l'analyse manuelle

Préparation des textes

Les textes numérisés présentaient des formats très variés, chaque tribunal ayant sa propre mise en page, et de nombreuses annotations telles que des coups de tampon ou des surlignages. La transformation de ces images en textes via l'OCR a donc été compliquée et a donné un résultat comportant de nombreuses erreurs, en particulier au début des décisions qui présentent généralement des mises en page plus complexes.

Il a donc été nécessaire de procéder à un travail, forcément imparfait, de nettoyage des textes numérisés. En plus de la correction manuelle de certaines erreurs et de la suppression de certains éléments récurrents tels que les numéros de page, nous avons effectué une vérification systématique au niveau des mots : nous avons appliqué un correcteur orthographique à l'ensemble des termes apparaissant dans le corpus, puis avons passé en revue manuellement tous les mots jugés mal orthographiés (en excluant les termes entièrement en majuscules). Chaque terme ainsi repéré a été soit conservé, soit corrigé manuellement, soit supprimé. Nous avons également corrigé autant que possible plus spécifiquement les éléments qui nous apparaissaient importants pour la suite du travail. Nous avons notamment harmonisé toutes les mentions de « Monsieur » et « Madame », celles de sommes en euros, ou encore corrigé certaines erreurs présentes dans des dates. Malgré ce travail, il faut garder en tête que le caractère imparfait de cette correction et la présence d'erreurs de numérisation dans le corpus a forcément des conséquences négatives sur les résultats obtenus par la suite.

Méthodes d'extraction utilisées

Dans le cadre de cette recherche nous avons principalement testé deux grandes familles de méthodes d'extraction d'informations : les expressions régulières et l'apprentissage automatique.

Expressions régulières

Les expressions régulières existent depuis les années 1950. Il s'agit d'un mini-langage informatique permettant de décrire et rechercher des motifs complexes dans un texte. Par exemple, le motif suivant permet de capturer une date au format jour/mois/année si elle est précédée du texte « mariés le » et optionnellement de la mention d'un lieu (« mariés à ... le »). Ce motif permet aussi d'extraire la valeur de la date en question :

```
mariés(?: à .*)? le ([0-3]\d\/[0-1]\d\/(?:19|20)\d\d)
```

En construisant et combinant différents motifs, on peut isoler de plus en plus précisément certains passages et soit extraire une information spécifique (un montant, une date), soit en déduire une « catégorie », par exemple décider que la présence et/ou absence combinées de ces motifs signifie que la décision du juge est qu'il s'agit d'un divorce « pour faute ».

Apprentissage automatique

L'apprentissage automatique, ou *machine learning*, est une autre méthode d'extraction d'informations qui repose sur la création d'un modèle via une procédure d'entraînement sur un corpus annoté. Cette méthode consiste à spécifier un type de modèle (arbre de décision, réseau neuronal, etc.) et à lui présenter des données accompagnées des « réponses » attendues. Cette opération est répétée un grand nombre de fois avec un grand nombre de données, et petit à petit, les paramètres du modèle sont optimisés pour renvoyer autant que possible la « bonne » réponse.

Lorsque le modèle comporte un très grand nombre de paramètres, comme dans le cas des réseaux de neurones, on parle alors d'apprentissage profond (ou *deep learning*). Dans ce cas la quantité de données d'entraînement nécessaire est d'autant plus importante et les temps de calcul sont d'autant plus longs. Des techniques d'apprentissage non-profond utilisant des arbres de décision appliquées à la

matrice termes-documents ont été testées dans le cadre de ce travail, mais elles n'ont pas donné de résultats probants et ne seront donc pas présentées ici. En effet, seuls les modèles de *deep learning* sont en mesure de prendre en compte l'ordre des termes dans le texte, ce qui est un facteur essentiel pour extraire une information précise et pertinente.

L'apprentissage profond est une méthode puissante, mais elle nécessite des volumes de données d'entraînement très importants, bien plus importants que les 2800 textes annotés du corpus e-Juris. Pour certaines tâches, une façon de pallier cette limite consiste à ne pas entraîner un modèle « de zéro » mais à utiliser un modèle préexistant et à le « spécialiser » sur une tâche spécifique via une phase « d'ajustement » (*fine-tuning),* cette phase d'apprentissage supplémentaire spécifique nécessitant moins de données. C'est la manière dont nous avons procédé ici : nous avons utilisé notre corpus non pas pour l'entraînement complet d'un modèle, mais pour un ajustement de modèles préexistant comme FlauBERT ou CamemBERT.

Identification des parties de la décision via des expressions régulières

Le juge aux affaires familiales bénéficie d'une grande liberté dans la manière dont il rédige une décision de divorce. Pour autant, la plupart des décisions suivent une structure en quatre grandes parties :

- l'*introduction*, qui contient des informations administratives diverses,
- la *procédure*, qui rappelle également les faits,
- la *discussion*, qui énonce notamment les demandes des parties,
- le *dispositif*, qui contient les conclusions et décisions rendues par le juge.

La capacité à identifier les parties d'une décision permet de cibler plus précisément une partie du texte. Par exemple, si l'on s'intéresse à une décision prise par le juge, on pourra effectuer la recherche uniquement dans la partie « dispositif ». C'est donc particulièrement intéressant en vue d'une extraction automatisée d'information.

Pour identifier ces parties de décision nous avons utilisé les expressions régulières, avec comme objectif d'identifier des termes ou des tournures caractéristiques du passage d'une partie à l'autre.

Concernant la partie « dispositif », elle débute dans la grande majorité des cas par l'expression *« par ces motifs »*. En recherchant cette expression tout en tenant compte de ses variantes en termes de majuscules/minuscules et des erreurs d'OCR, on arrive à identifier une grande part des dispositifs. En ajoutant certaines formulations plus rares comme *« décision »*, *« décisions »* ou *« décisions en conséquence »* sur une seule ligne, on parvient au résultat indiqué dans le tableau 1 : la partie dispositif est identifiable pour la quasi totalité (99,9%) des décisions du corpus.

Tableau 1. Résultats d'identification des parties de la décision

	Décisions (n)	Décisions (%)
Identification partie « dispositif »	2850	99,9
Identification partie « discussion »	2737	96,0
Identification partie « procédure »	2237	78,4

Lecture : la partie « dispositif » a été correctement identifiée dans 2 850 décisions, soit 99,9 % du corpus.

Pour la partie « discussion » les formulations à identifier sont plus variées : *« exposé des motifs »*, *« motifs de la décision »*, *« motivations »*, *« sur ce »*, *« sur quoi »*, et de nombreuses autres variantes. Sans être parfaite, l'identification fonctionne cependant relativement bien puisque cette partie est identifiée dans 96 % des textes.

Pour la partie « procédure », enfin, les résultats sont moins bons : la transition entre ces deux parties est plus souvent effectuée sans expression ou tournure de phrase identifiable. Si des expressions comme *« exposé du litige »*, *« faits et procédures »* ou *« faits et prétentions des parties »* reviennent régulièrement, on n'a au final réussi à identifier la partie « procédure » dans 78 % seulement des décisions.

Extraction d'informations via des expressions régulières

Nous avons ensuite continué d'utiliser les expressions régulières, mais cette fois pour extraire des informations directement depuis le contenu textuel des décisions.

Dates et montants

Dans un premier temps nous avons essayé de retrouver des informations directement identifiables et présentes dans les textes de manière explicite, comme des dates ou des montants. Plus précisément, nous avons tenté d'extraire la date du mariage, la date de la décision, les montants de prestation compensatoire demandés et offerts par les parties, et le montant de prestation compensatoire en capital fixé par le juge.

Pour les dates nous avons travaillé sur l'ensemble du texte des décisions, pour le montant de prestation compensatoire fixé nous nous sommes concentrés sur la partie dispositif, et pour les montants demandés et offerts nous avons travaillé sur toute la décision sauf la partie dispositif.

Les expressions regulières utilisées peuvent rapidement devenir nombreuses et complexes. Par exemple, pour l'extraction de la date du mariage, le principe est de repérer les différentes notations possibles d'une date (*« 22 février 2003 »*, *« 22/02/03 »*, *« 22.02.2003 »*...) en combinaison avec des tournures indiquant qu'il s'agit d'une date de mariage (*« se sont mariés »*, *« ont contracté mariage »*, *« unis en mariage »*...), tout en tenant compte de nombreuses formulations intermédiaires pouvant s'intercaler : *« par devant... »*, *« sans contrat préalable »*, *« en première noce »*, etc. Il s'agit donc de développer progressivement un ensemble de règles permettant de maximiser le nombre de dates repérées, tout en évitant les faux positifs (c'est-à-dire des dates ne correspondant pas à des dates de mariage). Ce travail est relativement long, fastidieux, et jamais totalement terminé, car il est potentiellement toujours possible d'améliorer les résultats obtenus en ajoutant ou en modifiant des règles existantes.

Tableau 2. Résultats obtenus via des expressions régulières sur les dates et montants

	Texte utilisé	Valeur correcte trouvée (%)	Parmi les documents concernés (%)
Date du mariage	Décision complète	96,0	96,0
Date de la décision	Décision complète	76,7	76,7
Montant de prestation compensatoire fixé	Dispositif	95,1	86,1
Montant de prestation compensatoire demandée	Tout sauf dispositif	88,9	81,4
Montant de prestation compensatoire offerte	Tout sauf dispositif	95,8	41,5

Lecture : le montant de prestation compensatoire fixé par le juge a été correctement retrouvé dans 95,1 % de l'ensemble des décisions, et dans 86,1 % de celles comportant effectivement un tel montant.

Le tableau 2 présente les résultats obtenus. Concernant la date du mariage, ceux-ci sont plutôt bons, puisqu'on arrive à retrouver la bonne valeur dans plus de 95 % des cas. Pour la date de la décision, les résultats sont moins satisfaisants puisqu'elle n'est correctement identifiée que dans trois documents sur quatre. Cela s'explique en partie par la plus grande variabilité de la manière dont la date est formulée, mais aussi par le fait qu'elle est souvent située dans l'en-tête ou au début du document, où les erreurs d'OCR sont plus fréquentes dans notre corpus.

Concernant les montants de la prestation compensatoire, on arrive à retrouver le montant correct fixé par le juge dans 86 % des décisions concernées, et le montant demandé dans 81 % des cas. Les résultats sont par contre très insatisfaisants pour les montants de prestation compensatoire offerts puisqu'ils ne sont repérés correctement que dans 41 % des documents qui comportent effectivement une offre : les décisions concernées sont peu nombreuses, et les formulations plus variées et plus ambiguës.

Type de divorce prononcé par le juge

Dans un second temps, nous avons essayé d'appliquer la technique des expressions régulières pour extraire une information plus « qualitative » : le type de divorce prononcé par le juge. Contrairement à l'extraction d'une information explicite, cette tâche nécessite une forme « d'interprétation » à partir de certaines formulations spécifiques : il s'agit donc d'isoler le passage du texte contenant cette information, et d'identifier le type de divorce à partir de son contenu.

Dans la base COMPRES, outre les divorces par consentement mutuel exclus de notre corpus quatre modalités différentes sont présentes pour coder le type de divorce prononcé par le juge : « divorce accepté », « divorce pour faute », « divorce pour altération définitive du lien conjugal » et « conversion de séparation de corps ». Une difficulté réside dans le fait que si cette information apparaît majoritairement dans la partie « dispositif », certains juges la mentionnent dans la partie « discussion ». Nous avons donc procédé en deux temps, en recherchant d'abord uniquement dans la partie dispositif, puis en élargissant au reste du texte si aucun résultat n'a été trouvé.

Les règles mises en place sous forme d'expressions régulières comportent deux aspects. D'abord une tentative de repérage d'expressions permettant d'identifier un type de divorce :

- pour les divorces acceptés, la mention d'articles de loi tels que les articles 233, 234 ou 247, ou bien la présence de formulations du type *« sur demande acceptée »*, *« acceptation du principe »*, *« acceptation des époux »*, etc.
- pour les divorces pour faute, la mention de l'article 242 ou la présence d'expressions comme *« torts exclusifs », « torts partagés », « torts de l'époux »*, etc.
- pour les divorces pour altération définitive du lien conjugal, la mention des articles 237 ou 238, ou bien des formulations comme *« altération définitive »* ou *« définitivement altéré »*
- pour les conversions de séparation de corps, soit la mention de l'article 306 soit la présence de l'expression *« séparation de corps »*

Cependant, la simple recherche de ces expressions génère un grand nombre de faux positifs, car elles peuvent être employées dans des contextes différents. Par exemple, un article de loi peut-être cité sans pour autant être utilisé pour fonder une décision. La mention *« torts exclusifs »* peut apparaître non pas lors d'une décision du juge mais pour décrire une demande d'une des parties. Ou encore, *« torts de l'époux »* peut être utilisé dans des formulations où le juge déboute l'une des parties de sa demande. Il a donc fallu ajouter un ensemble de règles permettant de repérer autant que possible les contextes pertinents tout en ne tenant pas compte des autres.

Tableau 3. Résultats obtenus via des expressions régulières sur le type de divorce

	Précision (%)	Rappel (%)
Divorce accepté (n=1551)	99,8	96,4
Faute (n=513)	98,1	98,8
Altération définitive du lien conjugal (n=766)	94,9	97,9
Conversion de séparation de corps (n=15)	32,1	90,0

Lecture : 98,1 % des divorces identifiés comme « pour faute » le sont effectivement. Par ailleurs, 98,8 % des divorces « pour faute » du corpus ont été correctement identifiés.

Les résultats obtenus sont présentés tableau 3. Ils sont bons voire très bons pour les divorces acceptés, les divorces pour faute et pour les divorces pour altération définitive du lien conjugal, avec des valeurs de précision et de rappel de l'ordre de 95 à 99 %[1]. Ils sont en revanche beaucoup moins bons pour les

[1] Dans le cadre d'une classification de documents, la « précision » est le pourcentage de documents classés dans la bonne catégorie (par exemple, le pourcentage de documents classés dans « divorces acceptés » qui sont bien des « divorces acceptés ». Le « rappel » est le pourcentage de documents d'une catégorie qui ont été correctement identifiés (par exemple, le pourcentage des documents de type « divorces acceptés » qui ont bien été classés comme « divorces acceptés »). Une « bonne » classification doit avoir des valeurs proches de 100 % pour ces deux indicateurs.

conversions de séparation de corps, mais celles-ci étaient très peu nombreuses dans notre corpus (seulement 15 décisions).

Au final, on voit que l'extraction ou la recherche d'informations via la méthode des expressions régulières peut donner des résultats intéressants. Elle présente cependant deux inconvénients majeurs :

- elle demande un travail de développement assez long et fastidieux : il faut en effet rechercher les motifs de texte pertinents, prendre en compte les différentes formulations possibles, éviter les « faux positifs », retourner sans arrêt au corpus, etc.
- les résultats obtenus sont assez peu robustes, car totalement dépendants des formulations présentes dans le corpus utilisé pour développer les règles. Il suffit que la rédaction change légèrement pour que le codage des expressions régulières produit précédemment ne fonctionne plus.

Apprentissage automatique et reconnaissance d'entités nommées

Pour aller plus loin que les simples expressions régulières, nous avons décidé de tester des méthodes plus récentes basées sur de l'apprentissage automatique, et plus précisément sur de la reconnaissance d'entités nommées (NER). La NER a pour objectif de retrouver dans un texte certains termes ou groupes de termes catégorisables dans des classes particulières. Par exemple, on pourra l'utiliser pour essayer d'identifier des noms de personnes, des noms de lieux, des dates, des quantités, etc.

Nous avons donc décidé d'utiliser cette méthode pour extraire une partie des variables déjà travaillées précédemment avec les expressions régulières :

- la date du mariage
- les montants de prestation compensatoire demandés
- les montants de prestation compensatoire offerts
- le montant de prestation compensatoire fixé par le juge

Annotation du corpus

Pour utiliser une méthode d'apprentissage automatique, il est nécessaire de fournir à un modèle d'apprentissage un corpus « annoté ». Cela signifie que le corpus doit comporter les informations recherchées, afin que le modèle puisse apprendre progressivement à les retrouver. En d'autres termes, le modèle doit être entraîné sur un ensemble de données qui ont été préalablement étiquetées avec les informations à extraire. Cela permet au modèle d'apprendre à reconnaître les motifs et les caractéristiques qui sont associés à ces informations, et de les utiliser pour extraire de nouvelles informations à partir de données non annotées.

Pour la reconnaissance d'entités nommées, l'annotation du corpus consiste à « surligner » et étiqueter les différentes informations recherchées dans les textes. Pour chaque document, il est donc nécessaire d'indiquer la position des mentions de la date du mariage, celle du montant de prestation compensatoire éventuellement fixé par le juge, etc. Un exemple de résultat d'annotation d'une partie de décision effectué à l'aide du logiciel libre *Label studio* est présenté encadré 1.

Encadré 1. Exemple d'annotations avec Label Studio

EXPOSE DU LITIGE
Madame et Monsieur se sont mariés le février 1993 DATE_MARIAGE devant l'officier de l'état-civil de la commune de sans avoir établi de contrat de mariage.
A la suite de la requête en divorce déposée le mars 2011 par Madame le juge aux affaires familiales a, par ordonnance de non-conciliation du octobre 2011, autorisé les époux à introduire l'instance en divorce et a statué sur les mesures provisoires.
Par acte d'huissier de justice du décembre 2011, Madame a fait assigner son conjoint en divorce sur le fondement de l'article 233 du Code civil.
Dans ses dernières conclusions récapitulatives reçues au greffe le mars 2013, auxquelles il convient de se référer pour un plus ample exposé des prétentions et moyens, Madame maintient sa demande en divorce.
Elle sollicite principalement qu'il soit :
- prononcé le divorce sur le fondement de l'article 242 du Code civil ;
- ordonné la mention du jugement en marge des actes d'état civil des époux,
- dit n'y avoir lieu à liquidation du régime matrimonial,
- donné acte à ce qu'elle reprendra l'usage de son nom de naissance,
- condamné M à lui verser la somme de 15000,00 PC_DEMANDEE euros au titre de prestation compensatoire,
- condamné M aux dépens.
Par conclusions reçues au greffe le février 2013, auxquelles il convient de se référer pour un plus ample exposé des prétentions et moyens, M sollicite principalement qu'il soit :
- dit que Madame est irrecevable en sa demande de divorce,
- débouté Madame de sa demande de divorce pour faute,
- déclaré subsidiairement mal fondée en sa demande de prestation compensatoire et l'en débouté,
- dit que subsidiairement cette prestation compensatoire sera limitée à 6 000,00 PC_OFFERTE euros,
- fixé les modalités de paiement du capital dans la limite de huit années,
- dit que M pourra s'en libérer par mensualités de 100,00 euros sur cinq années,
- condamné Madame outre des dépens de l'instance, à 2 000,00 euros au titre de l'article 700 du Code procédure civile.

Ce travail d'annotation est inévitablement long et fastidieux, surtout quand on doit l'appliquer à un corpus de plusieurs milliers de documents. Dans notre cas, nous avons automatisé une partie de ce travail en réutilisant les résultats obtenus avec les expressions régulières à l'étape précédente. En effet, pour toutes les informations correctement repérées à cette étape, les règles utilisées permettent non seulement d'extraire les informations, mais aussi de récupérer leur position dans le texte : cela nous a permis de gagner du temps et d'annoter automatiquement une grande partie des documents. Au final, ce sont environ 400 décisions pour lesquelles toutes les informations n'ont pas pu être annotées automatiquement et qui ont donc dû faire l'objet d'un « surlignage » manuel dans le logiciel *Label studio.*

Corpus d'entraînement et modèles

Une fois le corpus entièrement annoté, l'étape suivante a consisté à mettre en œuvre des processus d'apprentissage automatique. Pour cela, nous avons entraîné différents modèles afin d'apprécier leurs capacités respectives.

Ces apprentissages ont été effectués sur deux corpus différents (tableau 4) :

1. uniquement sur la partie dispositif pour la date du mariage et le montant de prestation compensatoire fixé par le juge.
2. sur toute la décision sauf la partie dispositif pour la date du mariage et les montants de prestation compensatoire demandés et offerts par les parties.

Tableau 4. Corpus et données utilisés pour la NER

Corpus	Variable	Nombre d'annotations
Dispositif	Date du mariage	2 756
	PC fixée	838
Hors dispositif	Date du mariage	2 837
	PC demandée	1 294
	PC offerte	153

Plusieurs modèles de langue adaptés au français étant disponibles au moment de la réalisation de cette étude, nous avons décidé d'entraîner chacun d'entre eux afin de comparer les résultats obtenus. Les modèles de langue testés sont présentés tableau 5.

Tableau 5. Modèles de langue testés

Modèle	Type	Année	Origine
CamemBERT	BERT	2020	INRIA / Facebook
FlauBERT	BERT	2020	CNRS / UGA / Université Paris Diderot
Barthez	BART	2021	École Polytechnique
JuriBERT	BERT	2021	HEC / École Polytechnique

Ces modèles utilisant une architecture de type *transformer* sont issus d'un entraînement sur de vastes corpus textuels francophones. À partir d'une tâche d'entraînement relativement simple (par exemple, pour les modèles de type BERT, arriver à prédire correctement un mot masqué au milieu d'une phrase), ils permettent une transformation contextuelle du contenu des décisions en séquences de vecteurs numériques. Celles-ci peuvent alors être utilisées en entrée d'un modèle d'apprentissage profond spécialisé pour la reconnaissance d'entités nommées.

Parmi les modèles testés, CamemBERT[2], FlauBERT[3] et Barthez[4] sont des modèles généralistes entraînés sur des corpus francophones de très grande taille. JuriBERT[5] est un modèle plus spécialisé, entraîné sur un corpus plus réduit de textes juridiques issus de Légifrance et de la cour de cassation. Par ailleurs, un modèle est souvent fourni dans différentes « tailles » (*small, base, large*) selon le nombre de paramètres qu'il comporte : dans ce cas on a entraîné et comparé chacune des différentes tailles disponibles.

Concrètement, le corpus annoté a été séparé en deux : un corpus d'entraînement comportant 85 % des décisions, et un corpus de test comportant les 15 % restant. Après une phase d'ajustement des hyperparamètres[6], chaque modèle a été entraîné à partir du corpus d'entraînement, puis évalué sur le corpus de test. Les calculs ont été effectués via la bibliothèque *spaCy* sur le supercalculateur Jean Zay du CNRS. Au total, ces opérations correspondent à une centaine d'heures de calcul GPU.

[2] L. MARTIN *et al*, « CamemBERT: a Tasty French Language Model », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 2020. https://doi.org/10.18653/v1/2020.acl-main.645

[3] H. LE *et al.*, « FlauBERT: Unsupervised Language Model Pre-training for French », *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020. https://aclanthology.org/2020.lrec-1.302

[4] M. K. EDDINE *et al.*, « BARThez: a Skilled Pretrained French Sequence-to-Sequence Model », *arXiv preprint*, 2020. https://doi.org/10.48550/arXiv.2010.12321

[5] S. DOUKA *et al.*, « JuriBERT: A Masked-Language Model Adaptation for French Legal Text », *Proceedings of the Natural Legal Language Processing Workshop 2021*, 2021. https://dx.doi.org/10.18653/v1/2021.nllp-1.9

[6] Les hyperparamètres d'un modèle incluent les paramètres liés à son architecture et à son processus d'entraînement. Dans le cas présent, les valeurs de *batch size* ont été fixées à 128 pour les modèles de base et 512 pour les modèles « *large* », tandis que les valeurs de *window* et de *stride* ont été fixées respectivement à 64 et 40.

Résultats sur le corpus « dispositif »

Le premier corpus ne comportait que la partie « dispositif » des décisions, et les annotations concernant la date du mariage et le montant de prestation compensatoire fixé par le juge. Les résultats obtenus pour chaque modèle sur ce corpus sont présentés tableau 6[7].

Tableau 6. Résultats obtenus sur le corpus « dispositif »

Corpus dispositif

	DATE_MARIAGE		PC_FIXEE	
	Précision	Rappel	Précision	Rappel
FlauBERT large cased	98.6	99.5	90.0	95.1
Barthez	98.2	99.8	88.3	91.9
FlauBERT small cased	97.7	100.0	90.5	92.7
FlauBERT base cased	97.3	100.0	87.8	93.5
FlauBERT base uncased	97.9	98.8	88.3	91.9
CamemBERT large	96.6	99.8	87.0	92.7
JuriBERT base	97.9	98.1	91.1	82.9
JuriBERT small	96.6	99.1	85.5	95.9
JuriBERT mini	96.8	97.9	86.0	90.2
JuriBERT tiny	95.5	98.6	77.2	93.5
CamemBERT base	96.0	89.2	74.8	77.2

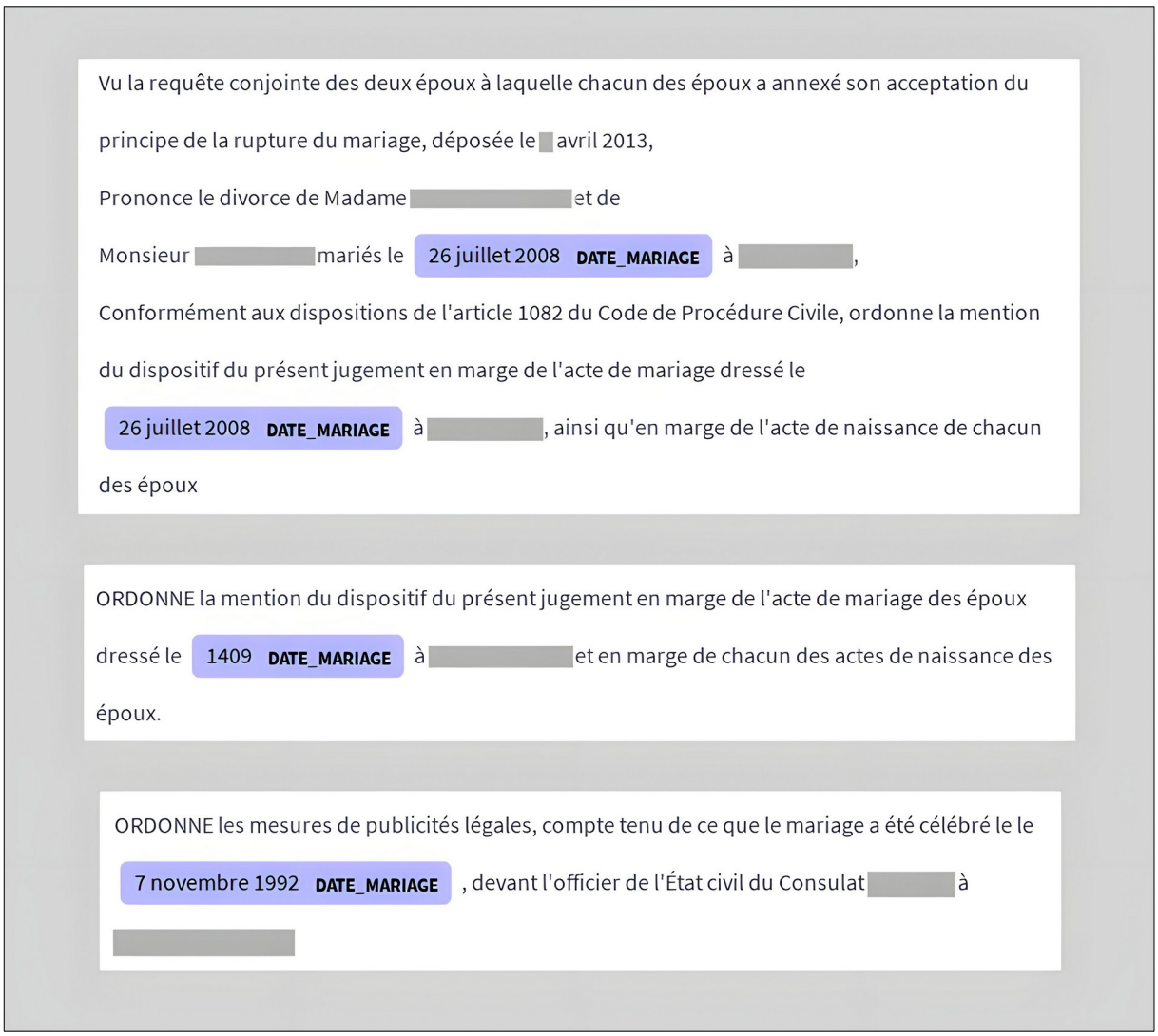
Lecture : 98,2 % des dates de mariage repérées par le modèle Barthez sont bien des dates de mariage. Par ailleurs, le modèle Barthez parvient à identifier 99,8 % de l'ensemble des dates de mariage du corpus « dispositif ».

On peut noter que certains modèles parviennent à repérer la quasi-totalité des dates de mariage, avec des valeurs de précision et de rappel proches de 100 %. Ce résultat est d'autant plus remarquable que le corpus peut présenter des erreurs d'OCR qui compliquent encore cette tâche.

L'encadré 2 montre quelques résultats obtenus lorsqu'on essaie d'identifier les dates du mariage avec le modèle résultant de l'entraînement du modèle Barthez. On voit dans le premier exemple que le modèle est capable de prendre en compte les contextes de manière assez fine car il parvient à différencier une date de mariage d'un autre type de date, malgré la présence du mot « mariage » à proximité de cette dernière. Les deux autres exemples illustrent des cas où le modèle parvient à identifier correctement la date alors que les expressions régulières n'y seraient pas parvenues : dans le premier cas, à cause d'un problème d'OCR sur la date, et dans le second cas, à cause d'une répétition du mot « le ».

[7] À noter que les résultats obtenus par l'entraînement d'un réseau de neurones sont « non déterministes » : ils comportent une part d'aléatoire et peuvent donc changer légèrement d'un entraînement à l'autre.

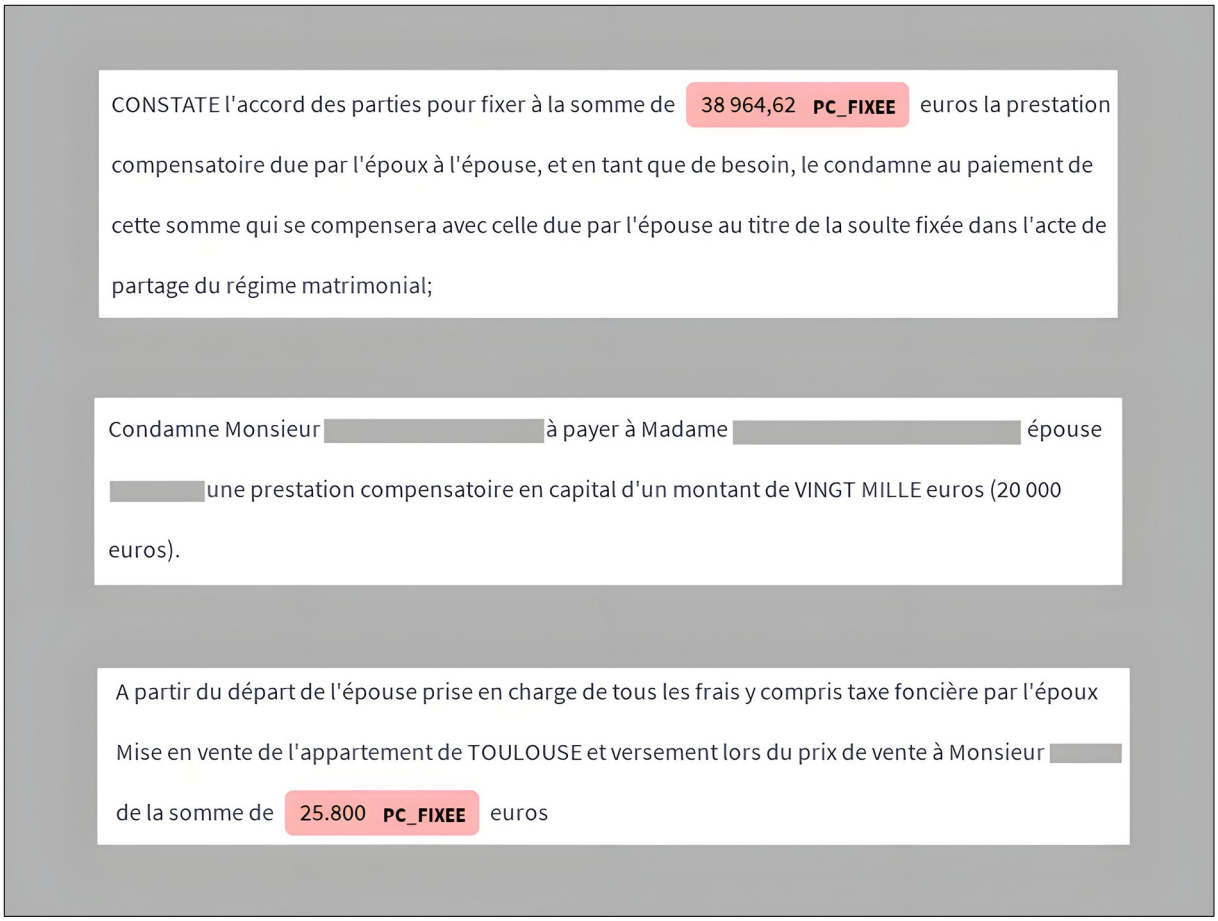
Encadré 2. Résultats de NER sur la date du mariage (modèle Barthez, corpus dispositif)



Concernant le montant de prestation compensatoire fixé par le juge, les résultats sont plus variables selon les modèles mais certains donnent tout de même des valeurs de précision et de rappel[8] élevées, de l'ordre de 90 à 95 %. L'encadré 3 montre quelques résultats obtenus, toujours avec le modèle Barthez : le premier exemple illustre un cas où le montant de prestation compensatoire a été correctement identifié, le second est un exemple où un montant n'a pas été repéré par le modèle, et dans le troisième exemple c'est un autre type de montant qui a été incorrectement identifié.

Encadré 3. Résultats de NER sur le montant de PC fixé (modèle Barthez, corpus dispositif)

[8] La précision mesure le pourcentage de valeurs identifiées comme un montant de prestation compensatoire fixé par le juge par le modèle qui le sont effectivement. Le rappel indique le pourcentage de l'ensemble des montants de prestation compensatoire fixé par le juge dans notre corpus qui ont été « retrouvés » par le modèle.



On notera que de manière générale les modèles JuriBERT, entrainés à partir de documents relevant du domaine juridique, ne donnent pas spécialement de meilleurs résultats que les modèles « généralistes », mais leur performance reste très bonne compte tenu de la taille plus réduite de leur corpus d'entraînement.

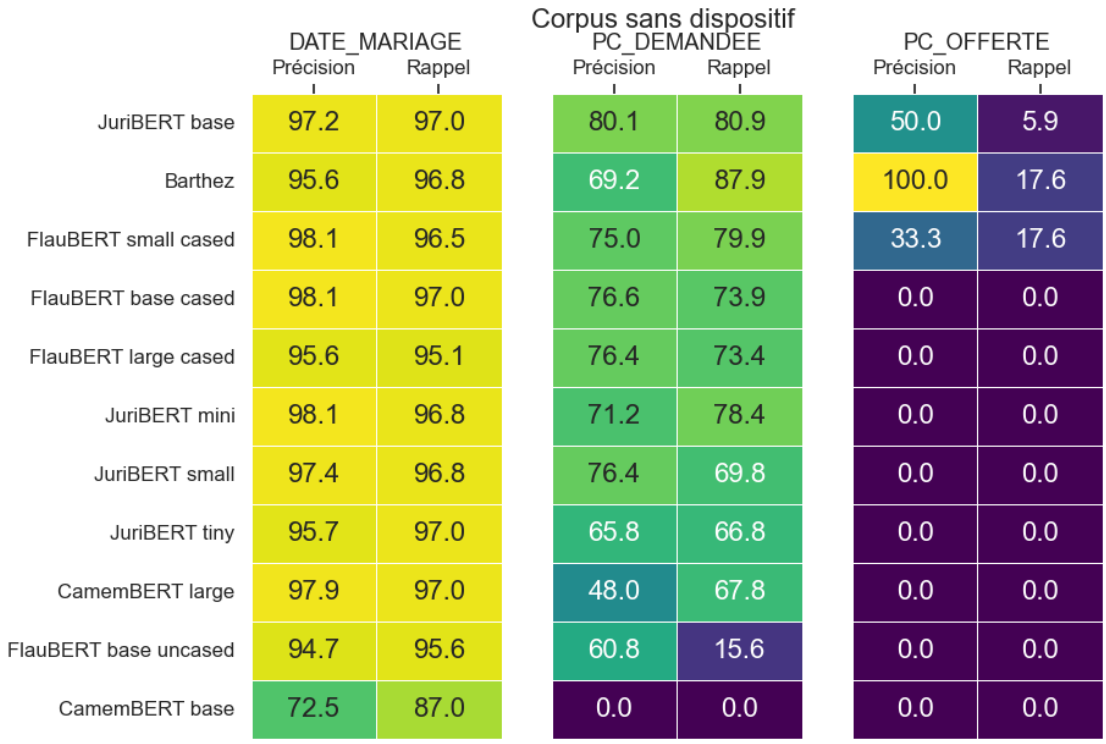
Résultats sur le corpus « hors dispositif »

Le second corpus était constitué de l'ensemble du texte des décisions à l'exception de la partie « dispositif ». Les données annotées comportaient la date du mariage (celle-ci est en effet souvent présente à la fois en début et en fin de décision), et les montants de prestation compensatoires éventuellement demandés et offerts par les parties. Les résultats obtenus pour chaque modèle sur ce corpus sont présentés tableau 7.

Concernant la date du mariage, les résultats obtenus sont un peu moins bons que ceux du corpus « dispositif », mais cette baisse était prévisible du fait que les textes sont plus longs et que les erreurs d'OCR sont plus fréquentes en début de document. Les valeurs de précision et de rappel restent cependant très bonnes, au-dessus de 95 % pour quasiment l'ensemble des modèles.

En ce qui concerne le montant de prestation compensatoire demandé par les parties, les résultats sont nettement moins satisfaisants : on parvient au mieux à une précision et un rappel de 80 %. Cette baisse de qualité s'explique par le fait que les formulations sont plus diverses et plus ambiguës, et qu'en même temps les exemples d'apprentissage moins nombreux. Comme indiqué tableau 4, on ne dispose que de 1294 annotations de montants de prestation compensatoire demandés, contre 2837 pour la date du mariage.

Tableau 7. Résultats obtenus sur le corpus hors dispositif



Enfin, pour le montant de prestation compensatoire offert, la reconnaissance d'entité nommée ne fonctionne pas du tout du fait du très petit nombre de décisions concernées et donc du manque de données d'entraînement : on ne dispose au total que de 153 annotations de montants de ce type dans notre corpus.

De manière générale, on voit donc que la reconnaissance d'entités nommées donne de bons résultats dès qu'on a un nombre suffisant d'exemples annotés dans le corpus. Le fait de « cibler » plus précisément une partie du texte des décisions permet aussi d'améliorer la précision de l'extraction d'informations . On voit enfin que si les « gros » modèles de langue généralistes obtiennent de bons résultats, des modèles plus petits mais plus spécialisés comme JuriBERT s'en sortent également plutôt bien.

Limites et perspectives

Les essais d'extraction automatisée d'informations présentés ici peuvent offrir des pistes relativement prometteuses compte tenu de certains résultats obtenus, surtout si on tient compte du fait que notre corpus de données annotées est de taille réduite et de qualité moyenne compte tenu des problèmes d'OCR.

Il convient cependant de souligner que ces pistes restent limitées, et en particulier ne concernent que des informations parmi les plus faciles à récupérer. Dans tous les cas, elles sont très insuffisantes pour envisager de remplacer, même partiellement, un travail d'analyse manuelle tel que celui réalisé dans le cadre de l'ANR COMPRES qui a permis la constitution d'une base de plusieurs centaines de variables nécessitant une interprétation parfois complexe du texte des décisions.

Une autre limite de ce travail réside dans l'évolution actuelle extrêmement rapide des méthodes de traitement des langues, en particulier dans le domaine des modèles de langue de grande taille. Ainsi, si CamemBERT et FlauBERT pouvaient être considérés comme l'état de l'art en termes de modèles francophones fin 2021, d'autres modèles sont apparus depuis tels que Bloom[9], un modèle multilingue comportant 176 milliards de paramètres, entraîné sur le supercalculateur Jean-Zay du CNRS.

[9]T. Le Scao *et al.*, « BLOOM: A 176B-Parameter Open-Access Multilingual Language Model », 2022. https://doi.org/10.48550/arXiv.2211.05100

Par ailleurs, la montée en puissance des modèles génératifs de type GPT laisse entrevoir d'autres méthodes d'extraction d'informations telles que le *prompting* : on peut envisager de fournir le texte intégral d'une décision à un modèle de ce type (éventuellement pré-entraîné sur un corpus de taille réduite) et d'extraire ensuite des informations sous forme de « questions » posées au modèle : « quelle est la date du mariage ? », « quelle est la cause de divorce fixée par le juge ? », etc.

Un travail d'évaluation de la possibilité et de la qualité des résultats de ces techniques serait donc très utile. Il se heurte cependant à plusieurs difficultés, telles que l'accessibilité des modèles (les paramètres des modèles tels que GPT-3.5 ou GPT-4 ne sont pas accessibles publiquement), la puissance de calcul nécessaire, et surtout la question du respect de la vie privée et de la confidentialité des informations contenues dans les décisions.