Atelier DAHLIA

DigitAl Humanities and cuLtural herItAge: data and knowledge management and analysis

Comité d'organisation et du programme :

Claudia Marinica (DUKe – LS2N, Polytech Nantes)
Fabrice Guillet (DUKe – LS2N, Polytech Nantes)
Florent Laroche (IS3P – LS2N, Ecole Centrale de Nantes)

organisé par le **groupe de travail DAHLIA** soutenu par l'Association EGC

conjointement avec la conférence
Extraction et Gestion des Connaissances (EGC2025)

le 28 janvier 2025 à Strasbourg

Editeurs :

Claudia Marinica
Laboratoire LS2N, équipe DUKe - Polytech'Nantes
page web : https://claudia-marinica.polytech.univ-nantes.fr/
email : claudia.marinica@univ-nantes.fr

Fabrice Guillet
Laboratoire LS2N, équipe DUKe - Polytech'Nantes
page web : https://www.univ-nantes.fr/fabrice-guillet
email : fabrice.guillet@univ-nantes.fr

Florent Laroche
Laboratoire LS2N, équipe IS3P - Ecole Centrale de Nantes
page web : http://www.florentlaroche.net/
email : florent.laroche@ec-nantes.fr

Accès en ligne :

Atelier DAHLIA : https://dahlia.egc.asso.fr/atelierDAHLIA-EGC2025.html
Groupe de travail DAHLIA : http://dahlia.egc.asso.fr
Mailing liste : gt-dahlia@egc.asso.fr

Préserver le patrimoine linguistique comorien : Translittération bidirectionnelle entre l'alphabet latin et le système Kamar-Eddine

Naira Abdou Mohamed* **, Abdessalam Bahafid* **, Zakarya Erraji*,
Anass Allak*, Naoufal Mohamed Soibira***, Imade Benelallam* **

*Institut National de Statistique et d'Économie Appliquée
https://insea.ac.ma/
**ToumAI Analytics
https://www.toumai-voice-analytics.com/
***Sciences Po Grenoble – UGA (Université Grenoble Alpes)
https://www.sciencespo-grenoble.fr/

Résumé. Les Comores, riches de leur diversité linguistique, abritent des dialectes issus du swahili et influencés par l'arabe. Historiquement, le système Kamar-Eddine, basé sur l'alphabet arabe, fut l'un des premiers systèmes d'écriture utilisés pour le comorien. Cependant, il a progressivement été supplanté par l'alphabet latin, bien que de nombreux textes d'archives et des locuteurs plus âgés continuent de l'utiliser, témoignant de son importance culturelle et historique. Dans ce contexte, nous présentons Shialifube, un outil de translittération bidirectionnelle entre les scripts latin et arabe, conçu selon les règles du système Kamar-Eddine. Pour évaluer sa performance, nous avons appliqué une technique de translittération aller-retour, obtenant un taux d'erreur de mots de 14,84 et un taux d'erreur de caractères de 9,56. Ces résultats démontrent la fiabilité de notre système pour des tâches complexes. En outre, Shialifube a été expérimenté sur un cas concret lié à la reconnaissance vocale, montrant son potentiel pour le Traitement Automatique des Langues. Ce projet représente une passerelle entre tradition et modernité, contribuant à la préservation du patrimoine linguistique comorien tout en ouvrant la voie à une meilleure intégration des dialectes locaux dans les technologies avancées.

1. Introduction

À la croisée des chemins entre l'Afrique, l'Europe, le Moyen-Orient et l'Asie du Sud-Est (Abeid et al., 2024; Allibert, 2015), les Comores se distinguent par un riche patrimoine culturel, une diversité qui manifeste particulièrement dans les dialectes locaux présentant des similitudes notables avec plusieurs langues étrangères. Bien que ces dialectes appartiennent à la famille bantoue du fait d'une proximité plus prononcée avec le swahili (Ahmed Chamanga, 2022) et le groupe de langues Sabaki (Serva & Pasquini, 2021), ils partagent également des similitudes avec l'arabe. C'est en partie pour cette raison que, comme le swahili (Mugane,

2017) avec le script ajami[1], l'un des systèmes d'écriture les plus anciens pour les dialectes comoriens est en script arabe (Lafon, 2007).

Connu sous le nom de système Kamar-Eddine, ce système d'écriture a été introduit dans les années 1960 par le linguiste Cheikh Ahmed Kamar-Eddine. Bien que l'alphabet latin soit aujourd'hui largement utilisé pour écrire le comorien, une minorité, principalement composée de personnes âgées, ne maîtrise que l'écriture en graphie arabe. Par ailleurs, de nombreux manuscrits anciens sont rédigés en cette graphie, ce qui souligne l'importance historique et culturelle qu'a cette écriture. Disposer d'une solution pour traiter ce système pourrait jouer trois rôles majeurs : (a) Démocratiser l'accès aux technologies de Traitement Automatique des Langues (TAL), rendant les dialectes comoriens accessibles à un large public, notamment ceux n'ayant pas accès aux outils numériques modernes ; (b) Préserver et promouvoir la richesse multiculturelle de l'archipel, en valorisant le système Kamar-Eddine comme élément fondamental de l'héritage linguistique et culturel ; (c) Rendre les archives nationales comoriennes accessibles à tous, en facilitant leur numérisation et leur conservation à long terme, tout en ouvrant la voie à de nouvelles recherches et applications en TAL.

Ce travail vise donc à amorcer le TAL pour ce système d'écriture, en espérant contribuer à la préservation du patrimoine immatériel comorien. Plus concrètement, nos principales contributions se résument en trois points :

- **Étude complémentaire** : Ce travail s'appuie sur l'article de Michel Lafon (Lafon, 2007), qui, à notre connaissance, est la seule étude réalisé sur le système Kamar-Eddine.
- **Exploration fondatrice** : Nous contribuons à l'introduction du TAL non seulement pour ce système d'écriture, mais aussi pour le traitement du comorien, une langue encore peu représentée dans ce domaine.
- **Innovation partagée** : Nous mettons à disposition des résultats de ce travail, en partageant le code et les modèles développés, afin de permettre à la communauté de bénéficier de nos avancées.

2. À propos du ShiKomori

Le Comorien, ou ShiKomori, est composé de quatre dialectes, chacun parlé dans une île spécifique : le ShiNgadidja, le ShiMwali, le ShiNdzuani et le ShiMaore. Bien qu'il serait idéal, dans le meilleur des cas, de travailler sur chaque dialecte séparément, ce travail traite le comorien dans son ensemble, sans distinction entre les variantes dialectales. Deux raisons justifient ce choix :

- **Similitudes élevées entre les dialectes** : Les dialectes sont très proches les uns des autres (Ahmed Chamanga, 2022). Ainsi, un locuteur d'une île donnée peut comprendre sans grande difficulté un dialecte parlé dans une autre île, en raison des lexiques largement partagés entre ces variantes. Cette forte similarité facilite la génération de solutions TAL généralisables sur l'ensemble des dialectes.
- **Rareté des données** : Il est difficile de trouver des corpus spécifiques à chaque dialecte, en raison du faible nombre de travaux réalisés dans ce domaine. De plus, les

[1] L'Ajami est un système d'écriture dérivé de l'alphabet arabe, utilisé principalement pour transcrire certaines langues africaines telles que le swahili, le wolof et le haoussa.

locuteurs préfèrent souvent écrire en français plutôt que d'utiliser les dialectes locaux, ce qui rend l'accès aux données encore plus limité.

Les fortes similitudes entre ces dialectes et le manque significatif de données rendent plus optimal de les traiter comme une langue unique. En effet, tenter de développer une solution pour chaque dialecte nécessiterait de travailler avec de petits corpus distincts, qui risqueraient de ne pas être suffisants pour entraîner des modèles performants. L'idée est donc de tirer parti des dialectes disposant de plus de données pour améliorer les performances sur ceux qui en ont moins. Cette approche a d'ailleurs été étudiée dans (Lin et al., 2019) où, parmi les solutions proposées pour faciliter la représentation des langues peu dotées, figure l'adoption d'une approche multilingue d'apprentissage par transfert depuis une langue bien dotée partageant des similitudes importantes. De plus, le système introduit par Kamar-Eddine considère le comorien comme une langue unifiée, sans règles spécifiques à chaque dialecte.

3. Travaux connexes

Le comorien est une langue très peu étudiée dans le domaine du TAL. Si quelques travaux antérieurs ont pu apporter des solutions qui la traitent sur différents cas d'usage (Abdourahamane et al., 2016; Naira et al., 2024), selon nos connaissances il n'existe aucun travail en linguistique computationnelle qui la traite sous son écriture en graphie arabe.

Au-delà de notre volonté à vouloir préserver ce patrimoine immatériel se cache une motivation née à la suite d'observations réalisées par des travaux antérieurs à l'instar de ceux qu'on retrouve dans (Micallef et al., 2023). Ce dernier décrit des expérimentations réalisées sur le maltais dans lesquelles une curieuse observation a été faite : sur plusieurs tâches (reconnaissance d'entités nommées, analyse des sentiments, etc.), la translittération vers des caractères arabes permet d'améliorer considérablement la performance des modèles. La raison est que le maltais, bien qu'écrit en caractères latins et qu'il ait des emprunts d'italien, il reste quand même une langue sémitique très similaire à l'arabe. La proximité du comorien avec ce dernier rend donc légitime de se poser des questions sur l'éventualité d'améliorer l'existant sur le TAL en recourant à une approche similaire.

À défaut de retrouver des travaux réalisés sur le comorien écrit en caractères arabes, nous proposons sur la Table 1 quelques descriptions de travaux notables ayant traité le sujet de la translittération d'une manière générale et plus particulièrement sur des langues africaines.

4. Le système Kamar-Eddine

La standardisation de l'écriture du comorien est devenue une priorité dès les premières années suivant l'indépendance de l'archipel (Chamanga, 1977). Si l'idée d'établir des règles spécifiques pour chaque dialecte a été rapidement abandonnée, le débat sur l'utilisation de l'alphabet latin ou arabe a suscité de vives discussions : d'une part, seule une faible minorité de la population, éduquée en français, langue coloniale, savait lire l'alphabet latin et prônait donc pour le recours à cette écriture et d'une autre part, la grande majorité, ayant reçu une éducation principalement à l'école coranique, maîtrisait la lecture de l'alphabet arabe. L'opinion publique étant en faveur de ce dernier, celui-ci fut rapidement adopté pour la traduction des documents officiels.

Il est toutefois important de noter que, malgré l'utilisation généralisée de cet alphabet, il n'existait aucune règle fixe régissant son application. Et c'est justement dans ce contexte

qu'Ahmed Kamar-Eddine eut l'idée de proposer une standardisation de ce système d'écriture. Il entreprit ce projet en publiant des chroniques dans sa revue Mwando.

4.1 Premières adaptations

L'alphabet arabe présente la particularité d'être un abjad, un alphabet dont les graphèmes représentent des consonnes, tandis que les voyelles sont indiquées uniquement par des diacritiques. Il existe trois voyelles, /a/, /i/ et /u/, représentées respectivement par les diacritiques *fatha*, *kasra* et *dhamma* (voir quelques exemples sur la Table 2). L'absence de voyelle est représentée par un *sukun* comme dans le mot بِنْت (bint) qui signifie *fille.*

Titre	Année	Description
Moroccan Arabizi-to-Arabic conversion using rule-based transliteration and weighted Levenshtein algorithm (Hajbi et al., 2024)	2024	C'est un système de translittération de l'Arabizi (Arabe dialectal marocain écrit en caractères latins) vers des caractères arabes. La méthode employée utilise la distance de Levenshtein.
Exploring the Impact of Transliteration on NLP Performance: Treating Maltese as an Arabic Dialect (Micallef et al., 2023)	2023	Amélioration de l'état de l'art TAL sur plusieurs tâches par traitement du Maltais écrit en caractères arabes.
A Unified Model for Arabizi Detection and Transliteration using Sequence-to-Sequence Models (Shazal et al., 2020)	2020	Pipeline de détection de l'Arabizi dans un texte présentant des alternances codiques (de l'arabe mélangé à d'autres langues, tous écrits en caractères latins) et translittération en caractères arabes.
Arabizi Chat Alphabet Transliteration to Algerian Dialect (Klouche & Benslimane, 2020)	2020	Translittération en caractères arabes des commentaires sur l'opérateur téléphonique algérien Ooredoo afin d'entraîner un modèle d'analyse des sentiments.

TAB 1 – *Quelques travaux antérieurs sur la translittération en graphies arabes.*

Cette particularité de l'arabe à avoir seulement trois voyelles pose un problème lors de l'adaptation de certaines langues en cette graphie. C'est justement le cas du wolof qui contient neuf voyelles (Currah, 2015), du swahili (Raia, 2021) et du comorien (Lafon, 2007). Pour ce dernier, il existe en plus des consonnes supplémentaires inexistant dans l'alphabet arabe. Pour répondre à ces particularités, des actions particulières ont été introduites dans les premières tentatives d'adaptation. Parmi elles il y avait :

- **Une introduction de caractères supplémentaires** : Un emprunt a été fait du Persan pour la représentation des sons /v/ qui devient ڤ, /g/ qui devient غ et /p/ qui devient پ. Mais il y avait quand même des ambiguïtés étant donné que le son /pv/ est tantôt transcrit en ڤ comme /v/ ou en ف comme /f/.

- **Représentation des voyelles** : Le comorien, ayant cinq voyelles /a/, /e/, /i/, /o/ et /u/, des mesures ont été adoptées pour combler l'absence de /o/ et /e/ dans l'arabe. Elles étaient marquées par le choix entre l'utilisation des diacritiques ou le recourt aux voyelles longues و pour /o/ et ي pour /e/). Mais là aussi on se retrouvait dans certains cas sur des ambiguïtés puisque des termes comme *mezi* (mois) et *mizi* (racines) s'écrivaient de la même manière (مِز ou ميزي en utilisant les voyelles longues).

4.2. Innovation originale

Pour remédier aux ambiguïtés observées lors des précédentes tentatives d'adaptation, parmi les solutions proposées par Kamar-Eddine, il y avait l'abandon des diacritiques au profit des voyelles longues. Les voyelles /a/, /i/ et /u/ gardent leurs écritures initiales tandis que /e/ et /o/ sont représentées respectivement par ـهـ et ه. Cela règle catégoriquement certains cas de confusion comme pour le dernier exemple que nous venons de voir dans la sous-section précédente. En effet, avec cette correction, le terme *mezi* devient ainsi مـهـزي et *mizi*, ميزي.

Voyelle	Transcription	Signification
na	نَجم (najm)	étoile
ni	نِظام (nizām)	système
nu	نُور (nūr)	lumière

TAB 2 – *Les diacritiques dans l'écriture arabe.*

Jusqu'alors, il n'y avait pas de représentation claire des affriqués, qui sont pourtant fréquentes dans le comorien. Kamar-Eddine propose alors l'utilisation du *shadda* pour l'accentuation de ces consonnes (voir Table 3). Nous rassemblons enfin dans la Table 4 toutes les règles que nous avons identifiées.

Son	Transcription	Exemple	Traduction
/ny/	نّ	نّاما	viande
/tr/	تّ	تّونكو	herbe
/dz/	زّ	مزّو	fardeau

TAB 3 – *Utilisation du shadda pour représenter les affriqués.*

5. Méthodologie

Aujourd'hui, à moins que cela n'échappe à notre vigilance, il n'existe aucune base de données en comorien écrite en caractères arabes. Afin de mesurer l'efficacité de notre système, nous sommes donc contraints d'utiliser uniquement comme références des textes en caractères latins[2]. Composée de 17.000 entrées (phrases, mots et expressions), le jeu de données est utilisé dans un premier temps pour translittérer vers l'arabe en implémentant les règles basées sur le dictionnaire construit, puis nous effectuons une translittération inversée pour obtenir le texte

[2] https://huggingface.co/datasets/nairaxo/shikomori-texts

d'origine. Pour contrôler la qualité de notre système, nous utilisons comme métriques le taux d'erreur de mots (WER) et le taux d'erreur de caractères (CER).

Alphabet régulier						Digrammes / Affriqués		
Son	Arabe	Latin	Son	Arabe	Latin	Son	Arabe	Latin
/a/	ا	a	/m/	م	m	/ð/	ذ	dh
/ɓ/ ou /b/	ب	b ou ɓ	/n/	ن	n	/ɖ/	ڕ	dr
/t͡ʃ/	ڜ	c	/o/	ه	o	/d͡z/	ز	dz
/ɗ/ ou /d/	د	ɗ ou d	/p/	پ	p	/ʈ/	ت	tr
/e/	ہ	e	/r/	ر	r	/ɲ/	ن	ny
/f/	ف	f	/s/	س	s	/ʃ/	ش	sh
/g/	غ	g	/t/	ت	t	/β/	ۏ	pv
/h/	ح	h	/u/	و	u	/θ/	ث	th
/i/	ي	i	/v/	ڤ	v	/t͡s/	س	ts
/d͡ʒ/	ج	j	/w/	و	w			
/k/	ك	k	/y/	ي	y			

TAB 4 – *Alphabets comoriens en caractères latin et avec le système Kamar-Eddine.*

La Figure 1 résume le pipeline par lequel passe un texte en entrée lors de l'inférence de notre outil. Dans un premier temps nous utilisons des règles de calcul pour détecter le type de script utilisé, si c'est un script en arabe ou en latin. Cela permet de savoir quel dictionnaire charger (arabe_latin ou latin_arabe). Par la suite une fois le type de script bien identifié ainsi que le type de dictionnaire, nous effectuons la translittération suivie d'une translittération inverse pour tenter de générer le texte d'origine dans le but de calculer les scores de translittération aller-retour permettant de mesurer la confidence de la translittération. Ainsi, deux choses sont retournées à la sortie : la translittération et sa confidence.

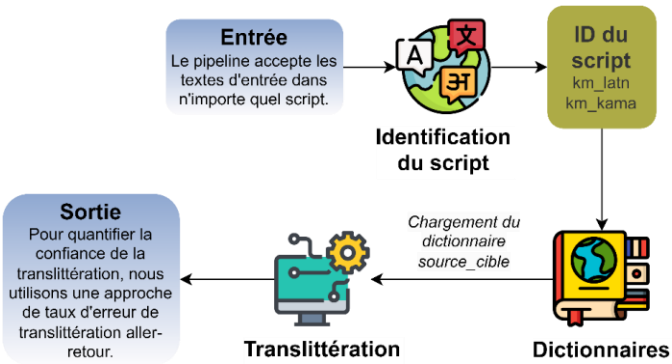


FIG. 1 – *Pipeline global de Shialifube, un système bidirectionnel de translittération du Comorien.*

5.1. Du Latin vers l'Arabe

La première étape de cette approche consiste à identifier les digrammes latins présents dans la chaîne, en les remplaçant par leurs équivalents en caractères arabes à l'aide d'un dictionnaire de correspondance préétabli. Cette étape permet de transformer efficacement les sons spécifiques représentés par deux caractères en un seul symbole arabe approprié, tel que le digramme *sh* ou *pv*. Pour comprendre pourquoi, imaginons que nous souhaitons translittérer le terme *shama* (association). Ne pas identifier au début les digrammes conduirait à une considération du *sh* comme étant deux lettres différentes (considérer le *s* comme س et le *h* comme ح), ce qui est une grave erreur. Au lieu de raisonner de cette manière, nous translittérons *sh* en ش puis on traite le reste où chaque caractère latin restant est converti en son équivalent arabe selon un second dictionnaire de correspondance des caractères isolés, garantissant ainsi la couverture des sons non représentés par des digrammes.

5.2. De l'Arabe vers le Latin

Nous réalisons la translittération d'une chaîne de caractères en écriture arabe vers une représentation en écriture latine en appliquant plusieurs transformations spécifiques. Nous effectuons ici également un remplacement des lettres arabes devant être représentées par des digrammes en Latin, par leurs équivalents. Ensuite, l'algorithme traite des caractères arabes spéciaux comme le symbole ه pour les remplacer par les caractères latins appropriés, en gérant aussi les combinaisons spécifiques comme بـ pour garantir une translittération phonétiquement correcte.

Après avoir segmenté la chaîne en caractères individuels, l'algorithme applique un ensemble de règles spécifiques pour traiter les lettres utilisées comme voyelles longues comme و et ي. En effet, si le و est utilisé comme étant non pas une voyelle longue, mais comme la lettre exprimant le son /w/, on le remplace par *w* et par *u* dans le cas contraire. Nous faisons pareil pour les translittérations du ي qui sont *y* et *i* pour les représentations respectives du son /y/ et de la voyelle longue /i/. Enfin, la chaîne de caractères est réassemblée pour produire la version finale en écriture latine, respectant les conventions phonétiques et graphiques de la langue cible.

5.3. Evaluation du système

Le WER est une mesure courante pour évaluer la précision d'un système de reconnaissance automatique de la parole et en traduction automatique. Il indique le taux d'erreurs dans la transcription produite par rapport à une transcription de référence. Le WER prend en compte plusieurs types d'erreurs, y compris les insertions, les suppressions et les substitutions de mots. Des valeurs plus basses de WER signifient une meilleure performance, indiquant que le système a moins d'erreurs par rapport à la référence. Le WER va de 0 à 100%. La formule pour le calculer est la suivante :

$$WER = \frac{S + D + I}{N}$$

Avec :

- S : le nombre de mots substitués (substitutions incorrectes).

- D : le nombre de mots supprimés (omissions).
- I : le nombre de mots insérés (ajouts incorrects).
- N : le nombre total de mots dans la transcription de référence.

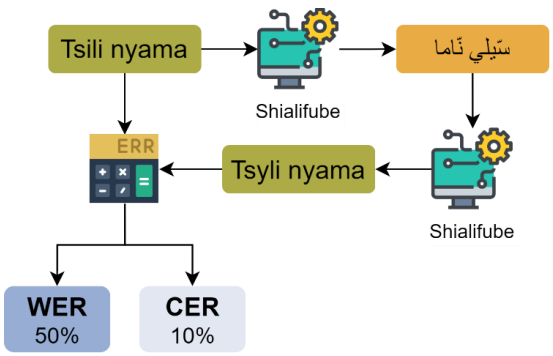
Nous utilisons la même formule pour calculer aussi le CER qui mesure quant à lui le taux de substitution des caractères plutôt que des mots. Si ces deux métriques permettent de mesurer la performance d'un système comme le nôtre, elles n'ont pas forcément les mêmes finalités. En effet, le WER tend plutôt à mesurer la divergence orthographique entre deux textes. Prenons l'exemple de la phrase سّيلي نّاما (J'ai mangé de la viande). Il se peut que lors de la translittération ce mot soit écrit سّيل نّام ce qui n'est pas difficile à comprendre malgré l'erreur d'écriture. Le WER dans ce cas est de 100%, en revanche, pour le CER est quand même bas, 22,2%.

En général, pour calculer ces métriques il faut des données labelisées, ce qui n'est pas notre cas. Pour cela, nous recourons à une technique inspirée de la traduction aller-retour (Kementchedjhieva & Søgaard, 2023) en translittérant nos textes latins en arabe utilisant notre système puis en faisant la même chose pour obtenir le texte en latin. Nous calculons alors les métriques WER et CER pour évaluer la performance de notre solution. La Figure 2 montre un exemple d'un cas de translitération aller-retour.

6. Expériences et résultats

Dans cette section, nous présentons dans un premier temps les résultats et les métriques de performance de Shialifube avec des descriptions de l'ensemble des itérations adoptées afin d'améliorer cette performance. Nous conduisons également une expérimentation sur un cas d'usage réel en reconnaissance vocale, le premier modèle d'apprentissage automatique jamais conçu dans le Comorien écrit en caractères arabes.

Convaincus que les contributions open-source sont la clé pour faire avancer la représentation des langues peu dotées dans le domaine du TAL, nous mettons publiquement à la disposition de tous la bibliothèque Shialifube[3], le code sur GitHub[4] et un *space* sur HuggingFace de reconnaissance vocale[5].



[3] https://pypi.org/project/shialifube/
[4] https://github.com/nairaxo/shialifube
[5] https://huggingface.co/spaces/nairaxo/shikomori-asr

FIG. 2 – *Exemple de translittération aller-retour et de calcul des métriques de performance.*

6.1. Métriques d'évaluation

Le processus d'application de nos règles de translittération se faisait d'une manière incrémentale et notre algorithme s'ajustait au progressivement en fonction des cas particulier que l'on rencontrait. Le but était de trouver l'approche la plus optimale qui réduit au plus les métriques. À chaque fois que l'on ajustait notre algorithme, nous calculions ces métriques. La Table 5 décrit l'ensemble des scénarios utilisés. Nous avons réalisé au total quatre itérations. Sur la dernière nous avons abouti à des métriques assez intéressantes indiquant une certaine fiabilité de notre système, bien que l'on se propose d'expérimenter dans de futurs travaux de nouvelles pistes d'amélioration.

En effet, ce qu'il faut savoir ici c'est que, bien que nous ayons fait en sorte de gérer tous les cas particuliers, on n'est pas à l'abri de constater des limitations lors de l'utilisation de notre système. Pour réduire au maximum ces limitations, nous nous proposons de continuer le perfectionnement et de mettre à jour la bibliothèque, la version actuelle étant en réalité une pré-version.

6.2. Cas d'usage : reconnaissance vocale

Nous proposons ici d'entraîner un modèle de reconnaissance vocale en comorien avec le système Kamar-Eddine. Le but de cet exercice est de tout simplement tester l'efficacité de notre système de translittération. Pour cela, nous considérons le corpus initial en alphabet latin avec son équivalent translitteré en arabe. Le premier est utilisé pour entraîner un modèle de référence que nous utilisons pour positionner la performance de notre modèle de transcription en alphabet arabe.

Expérience	Description	WER	CER
1	Première itération, sans traitement des digrammes.	68,56	34,41
2	Traitement des digrammes et gestion des voyelles longues.	43,09	21,30
3	Uniformisation et correction des séquences du corpus utilisé.	33,89	16,75
4	Traitement d'autres cas particuliers et considération de toutes les observations des précédentes approches.	14,84	9,56

TAB 5 – *Métriques d'évaluation de l'approche de translittération aller-retour.*

Quant au choix de l'architecture du modèle à utiliser, il s'est arrêté sur Whisper (Radford et al., 2022), un des modèles de reconnaissance vocale les plus performants dans l'état de l'art. En effet Whisper est pré-entraîné sur une large base de données multilingues comprenant des audios en Swahili et en Arabe. Cette phase de pré-entraînement consiste à apprendre le modèle à mieux assimiler chaque langue à partir de la compréhension de paramètres latents dans les audios. Nous ajustons le modèle en mettant à jour ses paramètres pour effectuer les tâches de

reconnaissance vocale en précisant au niveau des paramètres que l'on souhaite ajuster sur du Swahili pour le modèle en caractères latins et sur de l'Arabe pour celui en caractères arabes.

Les résultats sur la Table 6 indiquent une bonne performance du premier modèle comparé au second. En effet, deux principales raisons justifient cela :

- **Données non transformées** : en effet, transformer les données altère leur qualité, bien que dans notre cas cela soit le seul moyen d'obtenir de la data, à défaut d'effectuer de l'annotation manuelle. Cette dernière est justement une piste à sérieusement considérer dans des travaux futurs non seulement pour améliorer les performances en reconnaissance vocale, mais sur d'autres cas d'usage TAL comme l'analyse des sentiments, la détection des entités nommées, etc.
- **Vocabulaire inconnu** : l'utilisation d'un modèle pré-entraîné dépend du vocabulaire de celui-ci. En effet, bien que le Comorien soit proche de l'Arabe, il ne l'est pas plus que le Swahili. Ce qui fait que, en effectuant le découpage du texte en tokens lors de l'entraînement du modèle en alphabet arabe on se retrouve avec plus de tokens inconnus par le modèle pré-entraîné que lors de l'ajustement du modèle en alphabet latin.

Caractères	WER	CER
Latins	35.48	17.76
Arabes	37.44	21.42

TAB 6 – *Métriques d'évaluation de l'approche de translittération aller-retour.*

7. Conclusion

Le présent travail visait à poser les bases du Traitement Automatique du Langage Naturel appliqué au comorien, en mettant l'accent sur la transcription de cette langue en caractères arabes selon le système Kamar-Eddine. Dans un premier temps, nous avons compilé l'ensemble des règles d'écriture de ce système, qui ont servi de fondement à Shialifube, un système de translittération bidirectionnel du comorien.

En l'absence de données parallèles pour évaluer directement la performance de notre solution, nous avons adopté une approche de translittération aller-retour. Celle-ci a consisté à transcrire un corpus en caractères latins vers des caractères arabes, puis à retranscrire ces derniers en caractères latins. Cette méthode a permis d'obtenir des métriques prometteuses, après plusieurs itérations : un taux d'erreur de mots de 14,84 % et un taux d'erreur de caractères de 9,56 %.

Dans l'optique de tester l'utilité de cet outil pour des cas d'usage concrets, nous avons également mené des expériences en reconnaissance vocale. Nous avons observé une performance encourageante avec un taux d'erreur de mots de 37,44 % pour la version en caractères arabes, bien qu'elle reste légèrement inférieure à celle du modèle en caractères latins, qui atteint un score de 35,48 %. Enfin, il convient de souligner que ce travail est une étape préliminaire. Nous envisageons de poursuivre son perfectionnement dans le cadre de futures contributions, dans l'espoir qu'il contribue à la préservation et à la valorisation du patrimoine immatériel comorien. Afin d'encourager d'autres chercheurs à approfondir cette initiative, nous mettons à disposition l'intégralité du code source, la bibliothèque Shialifube ainsi que les modèles entraînés.

Références

Abdourahamane, M., Boitet, C., Bellynck, V., Wang, L., & Blanchon, H. (2016, juillet). Construction d'un corpus parallèle français-comorien en utilisant de la TA français-swahili. *TALAf (Traitement Automatique des Langues africaines)*. https://hal.science/hal-01992871

Abeid, S. N., Farhane, H., Motrane, M., Anaibar, F. E., & Harich, N. (2024). Inference on the biological history of the Comoros archipelago using the CD4 Alu/STR compound system. *Gene Reports*, *34*, 101865. https://doi.org/10.1016/j.genrep.2023.101865

Ahmed Chamanga, M. (2022). ShiKomori, the Bantu Language of the Comoros : Status and Perspectives. In *Handbook of Language Policy and Education in Countries of the Southern African Development Community (SADC)* (p. 79?98). BRILL. https://doi.org/10.1163/9789004516724_006

Allibert, C. (2015). L'archipel des Comores et son histoire ancienne. Essai de mise en perspective des chroniques, de la tradition orale et des typologies de céramiques locales et d'importation. *Afriques*, *6*. https://doi.org/10.4000/afriques.1721

Chamanga, M. A. (1977). *Recherches sur l'instrumentalisation du comorien : Problèmes d'adaptation lexicale (d'après la version comorienne de la loi du 23 novembre 1974)*. https://www.persee.fr/doc/cea_0008-0055_1977_num_17_66_2451

Currah, G. (2015, août). *Orthographe Wolofal*. http://currah.download/pages/ajamisenegal/orthographe_wolofal_harmattan_26-aout-2015_a4.pdf

Hajbi, S., Amezian, O., Moukhi, N. E., Korchiyne, R., & Chihab, Y. (2024). Moroccan Arabizi-to-Arabic conversion using rule-based transliteration and weighted Levenshtein algorithm. *Scientific African*, *23*, e02073. https://doi.org/10.1016/j.sciaf.2024.e02073

Kementchedjhieva, Y., & Søgaard, A. (2023). Grammatical Error Correction through Round-Trip Machine Translation. *Findings of the Association for Computational Linguistics: EACL 2023*, 2208-2215. https://doi.org/10.18653/v1/2023.findings-eacl.165

Klouche, B., & Benslimane, S. M. (2020). Arabizi Chat Alphabet Transliteration to Algerian Dialect. In *Artificial Intelligence and Renewables Towards an Energy Transition* (p. 790?797). Springer International Publishing. https://doi.org/10.1007/978-3-030-63846-7_76

Lafon, M. (2007). Le système Kamar-Eddine : Une tentative originale d'écriture du comorien en graphie arabe. *Ya Mkobe*, *14-15*, 29-48.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P., & Neubig, G. (2019). Choosing Transfer Languages for Cross-Lingual Learning. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3125-3135. https://doi.org/10.18653/v1/P19-1301

Micallef, K., Eryani, F., Habash, N., Bouamor, H., & Borg, C. (2023). Exploring the Impact of Transliteration on NLP Performance : Treating Maltese as an Arabic Dialect. In K. Gorman, R. Sproat, & B. Roark (Éds.), *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)* (p. 22-32). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.cawl-1.4

Mugane, J. (2017). The Odyssey of 'Ajam ? And the Swahili People. *Islamic Africa*, *8*(1-2), 193-216.

Naira, A. M., Imade, B., Abdessalam, B., & Zakarya, E. (2024). Datasets Creation and Empirical Evaluations of Cross-Lingual Learning on Extremely Low-Resource Languages : A Focus on Comorian Dialects. In S. Henning & M. Stede (Éds.), *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)* (p. 140-149). Association for Computational Linguistics. https://aclanthology.org/2024.law-1.14

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision* (arXiv:2212.04356). arXiv. http://arxiv.org/abs/2212.04356

Raia, A. (2021). *A network of copies | Scholarly Publications*. https://scholarlypublications.universiteitleiden.nl/handle/1887/3275029

Serva, M., & Pasquini, M. (2021). *The Sabaki languages of Comoros*. https://doi.org/10.31235/osf.io/qsfx7

Shazal, A., Usman, A., & Habash, N. (2020). A Unified Model for Arabizi Detection and Transliteration using Sequence-to-Sequence Models. In I. Zitouni, M. Abdul-Mageed, H. Bouamor, F. Bougares, M. El-Haj, N. Tomeh, & W. Zaghouani (Éds.), *Proceedings of the Fifth Arabic Natural Language Processing Workshop* (p. 167-177). Association for Computational Linguistics. https://aclanthology.org/2020.wanlp-1.15

Summary

The Comoros, rich in linguistic diversity, are home to dialects derived from Swahili and influenced by Arabic. Historically, the Kamar-Eddine system, based on the Arabic script, was one of the first writing systems used for Comorian. However, it has gradually been supplanted by the Latin alphabet, even though many archival texts and older speakers continue to use it, reflecting its cultural and historical significance. In this context, we present Shialifube, a bidirectional transliteration tool between Latin and Arabic scripts, designed according to the rules of the Kamar-Eddine system. To evaluate its performance, we applied a round-trip transliteration technique, achieving a word error rate of 14.84 and a character error rate of 9.56. These results demonstrate the reliability of our system for complex tasks. Additionally, Shialifube was tested in a practical case related to speech recognition, highlighting its potential for Natural Language Processing. This project serves as a bridge between tradition and modernity, contributing to the preservation of Comorian linguistic heritage while paving the way for better integration of local dialects into advanced technologies.

Liage d'entités nommées issues de pharmacopées anciennes – une première approche

Karim El Haff*,** Agnès Braud*
Florence Le Ber* Véronique Pitchon**

*Université de Strasbourg, ENGEES, CNRS, UMR 7357 ICube, F 67000 Strasbourg
kelhaff@unistra.fr, agnes.braud@unistra.fr, florence.le-ber@unistra.fr
**Université de Strasbourg, CNRS, UMR 7044 Archimède, F 67000 Strasbourg
pitchon@unistra.fr

Résumé. Les travaux sur la désambiguïsation et le liage d'entités nommées dans des textes historiques, en particulier des pharmacopées, sont encore peu développés. L'analyse de ces manuscrits se fait souvent par le biais d'experts, sans recours systématique aux technologies linguistiques qui pourraient automatiser ces processus. Cet article propose une approche intégrant une chaîne de désambiguïsation et de liage d'entités nommées appliquée aux traductions anglaises de pharmacopées arabes médiévales à travers l'utilisation de ressources comme BabelNet, Wikidata et GBIF pour lier les mentions de plantes à leurs identifiants scientifiques modernes.

1 Introduction

Le travail présenté dans cet article s'inscrit dans le cadre de l'analyse des anciennes pharmacopées, pour trouver et expérimenter des ingrédients anciens – en particulier des ingrédients issus de végétaux – afin de traiter certaines maladies. Après une première étape d'extraction des entités nommées (NER) (El Haff et al., 2023), il vise à développer et à évaluer, en utilisant des outils existants, une chaîne de désambiguïsation et de liage (NED et NEL) pour les entités nommées désignant des végétaux. Étant donné l'ambiguïté des textes en langage naturel, le développement d'une telle chaîne de traitement est nécessaire, dans l'objectif de créer un système unifié capable de traiter les corpus anciens depuis l'entrée brute jusqu'aux données liées et structurées. Dans le cas présent il s'agit d'identifier des espèces végétales.

Les premières méthodes de NED/NEL développées reposaient sur des règles avant d'incorporer des modèles syntaxiques et des dictionnaires, mais sont restées limitées par la rigidité des règles codées manuellement (Cucerzan, 2007). L'usage d'ontologies a été proposé par Volz et al. (2007), pour améliorer la désambiguïsation en structurant les entités selon des relations prédéfinies, ce qui est particulièrement efficace dans des contextes tels que la reconnaissance d'identificateurs géographiques. Les approches statistiques développées par la suite reposent sur des modèles probabilistes, comme les modèles de Markov cachés (HMM), pour traiter l'ambiguïté en considérant la séquence des mots et leurs distributions de probabilité sous-jacentes. Dans les travaux de Priya (2015), les HMM sont utilisés pour l'extraction d'entités en

exploitant des caractéristiques n-grammes et le regroupement de parties du discours. Les CRF (Conditional Random Fields) étendent les capacités des HMM en modélisant les dépendances entre les entités d'une séquence (Li et al., 2008). Les modèles d'apprentissage profond ont marqué un changement important. Borchert et Schapranow (2022) ont exploré ces approches pour la génération de candidats en combinant un vecteur TF-IDF avec un modèle SapBERT multilingue. Kolitsas et al. (2018) ont utilisé une architecture LSTM bidirectionnelle et un mécanisme d'attention dans un modèle qui découvre et relie simultanément les entités, ce qui permet d'utiliser la dépendance mutuelle entre les tâches de détection et de désambiguïsation des entités. Une autre avancée importante dans le domaine des NED/NEL a été le développement de BabelNet et Babelfy (Moro et al., 2014; Navigli et Ponzetto, 2012), qui ont permis une approche unifiée intégrant la désambiguïsation du sens des mots et la mise en relation des entités. L'intégration par BabelNet de plusieurs bases de connaissances, dont Wikipedia et WordNet, renforce ses capacités de désambiguïsation multilingue et interdomaine.

Dans le cadre de nos travaux, nous utilisons ces ressources en ligne et en particulier GBIF [1], qui permet d'exploiter une information géographique sur la localisation des végétaux. Dans la suite de l'article nous présentons (section 2) les données et la chaîne de traitement, puis nous détaillons (section 3) notre proposition d'utiliser l'information géographique pour lier les entités nommées désignant des végétaux à des taxons précis. Nous présentons une analyse de performance avant de conclure (section 4).

2 Données et Méthodes

Données et prétraitement des données. Le corpus de données est issu d'une première étape de reconnaissance d'entités nommées effectuée sur la traduction anglaise par Oliver Kahl de l'ouvrage « Dispensatory in the Recension of the Aḍudī Hospital » écrit par Sābūr ibn Sahl au IXe siècle (Kahl, 2009). Ce manuscrit décrit 292 remèdes ou préparations. L'ensemble du corpus est constitué d'environ 36 000 tokens. Ici, nous nous intéressons aux entités nommées qui correspondent à des ingrédients des remèdes, et qui ont 3 093 occurrences dans le corpus.

Ces données sont fortement bruitées, les entités étant mentionnées avec une forte variabilité, une phase de pré-traitement est donc nécessaire pour les nettoyer et les transformer en un format adapté aux étapes de désambiguïsation et de liage. Par exemple, un ingrédient peut être désigné par des mots descriptifs supplémentaires, tels que « pépins de pomme grillés », où seul le mot « pomme » est nécessaire pour une liaison précise en tant que taxon dans l'infrastructure de données moderne. En effet, les anciennes pharmacopées décrivent souvent les plantes en termes de parties (par exemple, graines, feuilles) ou de transformations (par exemple, grillées, séchées), il est donc nécessaire de supprimer ces qualificatifs pour se concentrer sur l'entité primaire. Pour ce faire, deux dictionnaires spécialisés, fabriqués manuellement, ont été créés. L'un s'est concentré sur l'identification et la suppression des parties de plantes et l'autre sur la suppression des mots-clés liés aux transformations.

Ressources utilisées. BabelNet est un dictionnaire multilingue et un réseau sémantique qui intègre des données de sources variées, telles que WordNet et Wikipedia. Babelfy, un outil basé sur BabelNet, fournit une API pour identifier et relier les entités textuelles aux concepts dans

1. Global Biodiversity Information Facility, https ://www.gbif.org/

BabelNet. Cet outil aide à obtenir des identifiants BabelNet pour chaque entité reconnue, ce qui facilite le processus de liage.

Wikidata (Vrandečić et al., 2023) est une infrastructure de données ouverte qui sert de stockage central pour les données structurées de divers projets Wikimedia. Dans le contexte du pipeline, Wikidata sert d'intermédiaire entre les entités vernaculaires identifiées par BabelNet et la ressource GBIF.

GBIF fournit un accès à des données sur la biodiversité, avec près de 2 milliards d'entrées, assorti d'une structure taxonomique qui standardise les noms d'espèces issues de différentes sources. GBIF recense également les coordonnées des endroits où les espèces ont été enregistrées à l'échelle mondiale. Ces données sont utiles pour la désambiguïsation car elles permettent de croiser le contexte historique de la mention d'une plante avec sa distribution géographique connue. GBIF apparaît ainsi comme un outil approprié pour atteindre le lien final vers le nom scientifique d'une mention de plante.

Chaîne de traitement (NED et NEL). Cette chaîne de traitement est composée de trois étapes, qui exploitent successivement les API des ressources décrites ci-dessus.

1. Désambiguïsation avec Babelfy et extraction de l'identifiant BabelNet. L'outil Babelfy est utilisé sur le texte brut pour la désambiguïsation initiale. Babelfy utilise le réseau sémantique de BabelNet sur le texte environnant pour désambiguïser avec précision les entités. Cela est utile dans les cas où une entité peut avoir plusieurs significations, comme « orange » qui peut faire référence à la fois à un fruit et à une entreprise. La capacité de Babelfy à prendre en compte le contexte textuel lui permet de sélectionner la signification la plus pertinente, en reliant l'entité au concept BabelNet approprié.

   Exemple : Pour l'ingrédient « Alhagi » dans la phrase « The prescription of the tabasheer pastille with alhagi » → BabelNet ID : bn :03344048n

2. Extraction de l'identifiant Wikidata. Chaque terme retrouvé dans BabelNet est ensuite passé par un processus de liage pour associer les mentions de plantes à des identifiants uniques dans Wikidata qui permettrait l'obtention de l'identifiant unique qui se trouve dans la base GBIF.

   Pour l'ingrédient « Alhagi », Wikidata ID : Q310762 → GBIF ID : 2945087

3. Analyse de la distribution géographique avec GBIF. Cette analyse vise a sélectionner l'espèce utilisée la plus probable, à partir des différentes espèces recensées et géo-référencées. Pour exemple, la figure 1 montre la distribution mondiale du taxon 2945087 *Alhagi sp.*, qui recouvre 6 espèces dont 3 avec des géo-références. Cette étape est détaillée dans la section suivante.

   Pour l'ingrédient « Alhagi », 3 espèces candidates ont des géo-références : Alhagi mauromum, Alhagi graecorum et Alhagi pseudalhagi

3 Le système de notation géographique

Dans cette première approche, l'hypothèse est que les plantes citées dans une pharmacopée proviennent plus probablement de la région où elle a été rédigée. Nous proposons donc de construire un score de proximité pour chaque espèce candidate, en exploitant les coordonnées

FIG. 1 – *Géo-références dans GBIF pour le taxon Alhagi possédant l'identifiant 2945087*

géographiques de leurs occurrences dans GBIF. Nous définissons pour cela trois zones de proximité à partir d'un point de référence (la ville de Bagdad), dans lesquelles la densité de chaque espèce est calculée ; sont considérées les occurences localisées dans un rayon de 50 km (espèces locales), celles localisées dans un rayon de 500 km (espèces régionales) et celles localisées dans un rayon de 1500 km (espèces éloignées). Pour calculer la distance entre les localisations des occurrences des espèces et le point de référence (Bagdad dans notre exemple), nous utilisons la formule de Haversine qui tient compte de la rotondité de la Terre. Le calcul du score pour un taxon repose sur les deux mesures suivantes.

— **Densité d'une espèce et score par zone.** Nous définissons la densité $\rho_{t,r}$ et le score $S_{t,r}$ d'un taxon t dans le cercle de rayon r de la façon suivante :

$$\rho_{t,r} = \frac{n_{t,r}}{\pi r^2} \qquad S_{t,r} = \frac{\rho_{t,r}}{\rho_{\max,r}} \cdot C_r$$

où $n_{t,r}$ est le nombre d'occurrences dans le cercle de rayon r, $\rho_{\max,r}$ est la densité maximale observée dans le cercle de rayon r, et C_r est un coefficient favorisant la proximité et de valeur fixée à 0.9 (pour $r = 50$), 0.6 ($r = 500$) ou 0.3 ($r = 1500$).

— **Score et nombre relatif d'occurrences d'un taxon.** Chaque taxon est caractérisé par deux valeurs, son score S_t, qui est le maximum des scores calculés dans chaque zone et son nombre relatif d'occurrences, O_t :

$$S_t = \max_r(S_{t,r}) \qquad O_t = 0.1 \cdot \frac{n_t}{n_{\max}}$$

où $n_t = \sum_r n_{t,r}$ est le nombre total d'occurrences du taxon t, et $n_{\max}$ est le nombre maximum d'occurrences observé parmi tous les taxons candidats pour l'entité nommée dans GBIF.

Le score final d'un taxon est composé comme de la somme S_t et O_t. Ce mode de calcul permet de prioriser les taxons ayant une densité élevée dans les premiers cercles (petits rayons), tout en prenant en compte le nombre total relatif d'occurrences. À l'issue de ce calcul, nous pouvons lier la plante mentionnée dans le texte au meilleur candidat situé dans la base GBIF, puis extraire les détails importants, notamment le nom scientifique et la famille de la plante.

Le tableau 1 présente les résultats des différents critères pour les trois candidats associés à l'exemple « Alhagi » [2]. L'espèce *Alhagi maurorum* (GBIF ID : 2945092) a obtenu le score final

2. Chaque valeur est ajustée à 4 décimales, à l'exception de la densité ρ_{tr} qui est une petite valeur

Taxon ID	Radius (km)	ρ_{tr}	S_{tr}	S_t	O_t	S_{ot}
2945092 *Alhagi maurorum*	50	0.0	0.0000	0.6000	0.1000	**0.7000**
	500	5.092958178940651e-06	0.6000			
	1500	4.031925224994682e-05	0.3000			
2945091 *Alhagi graecorum*	50	0.0	0.0000	0.6000	0.0958	**0.6958**
	500	5.092958178940651e-06	0.6000			
	1500	3.8621599523633275e-05	0.2874			
11374752 *Alhagi pseudalhagi*	50	0.0	0.0000	0.2063	0.0678	**0.2741**
	500	0.0	0.0000			
	1500	2.7728327863121322e-05	0.2063			

TAB. 1 – *Comparaison des trois taxons candidats basée sur les rayons et le calcul des scores.*

le plus élevé, tandis que le candidat suivant, *Alhagi graecorum* (GBIF ID : 2945091), obtient un score très proche. Le dernier candidat, *Alhagi pseudalhagi*, obtient un score très inférieur et n'est donc pas retenu. Le système permet donc un premier tri, mais une expertise en botanique et en histoire reste nécessaire pour départager les candidats proches.

4 Discussion et conclusion

Une analyse de performance a été menée, en appliquant la chaîne de traitement à un échantillon de 1 000 entités issues de la pharmacopée de Sabur Ibn Sahl, dont 820 (82 %) désignaient des ingrédients à base de plantes. Après le traitement du texte via Babelfy, 767 entités parmi les 820 (93,54 %) ont été analysées ; 53 (6,46 %) n'ont pas été détectées, surtout des translittérations arabes (zanbaq, kaukab) ou des appellations rares (gum-arabic). Cette couverture de 93,54 % montre l'efficacité de BabelNet pour détecter des termes variés. Cependant, certains homonymes, comme "rose" interprété à tort comme le verbe "to rise", ont entraîné des erreurs dans 7,32 % des cas (60 entités). Ensuite, 572 des 767 entités ont été correctement liées à Wikidata, soit une perte d'environ 25,42 %, due à l'absence d'identifiants Wikidata pour certaines entités identifiées par BabelNet. Cette étape constitue un goulet d'étranglement, car les entités désambiguïsées par BabelNet ne correspondent pas toujours à des identifiants disponibles dans Wikidata. Enfin, 266 des 572 entités ont été liées avec succès à GBIF, soit une réduction supplémentaire de 53,50 %. Ce taux reflète les défis liés à la correspondance entre les noms vernaculaires et les taxonomies contenues dans GBIF, entravée par l'absence d'ID GBIF dans Wikidata. Finalement, 266 entités sur les 820 initiales (32,44 %) ont été correctement désambiguïsées et liées à GBIF à la fin de la chaîne de traitement. Malgré cette perte progressive, l'algorithme de filtrage géographique apparaît intéressant pour prioriser les correspondances en fonction de l'origine géographique du texte analysé, facilitant le travail des experts.

Finalement, bien que la chaîne de traitement ait réussi à désambiguïser et à lier plusieurs entités végétales, notre travail révèle plusieurs défis quant aux ressources utilisées. Ces ressources, Wikidata ou GBIF, ont été conçues d'abord pour traiter des données contemporaines, et ne couvrent pas suffisamment les niches historiques et les langages vernaculaires intrinsèques à notre recherche. Cela met en évidence la nécessité de développer des ressources plus spécialisées, en collaboration avec des experts en histoire et botanique, pour couvrir de manière

exhaustive les entités vernaculaires. Par ailleurs, la robustesse de BabelNet, en tant que réseau multilingue, ne doit pas être négligée, car elle permettrait d'appliquer ces méthodes à des textes en différentes langues. Enfin, le filtrage géographique montre un potentiel qui doit être affiné par la prise en compte de routes commerciales historiques et de zonages écologiques.

Références

Borchert, F. et M.-P. Schapranow (2022). HPI-DHC @ BioASQ DisTEMIST : Spanish Biomedical Entity Linking with Pre-trained Transformers and Cross-lingual Candidate Retrieval. *CLEF (Working Notes)*.

Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *EMNLP-CoNLL, 2007 Joint Conf*, pp. 708–716.

El Haff, K., W. Antoun, F. Le Ber, et V. Pitchon (2023). Reconnaissance des entités nommées pour l'analyse des pharmacopées médiévales. In *EGC 2023 - Extraction et Gestion des Connaissances*, pp. 329–336.

Kahl, O. (2009). *Sābûr Ibn Sahl's Dispensatory in the Recension of the Aḍudī Hospital*. BRILL.

Kolitsas, N., O.-E. Ganea, et T. Hofmann (2018). End-to-End Neural Entity Linking. In *Proc. of the 22nd Conf. on Computational Natural Language Learning*, pp. 519–529.

Li, H., T. Liu, W.-Y. Ma, T. Sakai, K.-F. Wong, et G. Zhou (Eds.) (2008). *Information Retrieval Technology : AIRS 2008, Revised Selected Papers*. LNCS 4993.

Moro, A., A. Raganato, et R. Navigli (2014). Entity Linking meets Word Sense Disambiguation : a Unified Approach. *Trans. of the Assoc. for Computational Linguistics 2*, 231–244.

Navigli, R. et S. P. Ponzetto (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence 193*, 217–250.

Priya, N. (2015). A Name Entity Detection and Relation Extraction from Unstructured Data by N-gram Features. *IOSR Journal of Computer Engineering*, 25–28.

Volz, R., J. Kleb, et W. Mueller (2007). Towards ontology-based disambiguation of geographical identifiers. *I3*.

Vrandečić, D., L. Pintscher, et M. Krötzsch (2023). Wikidata : The Making Of. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 615–624.

Summary

Today, work on the disambiguation and linking of named entities in historical texts, particularly pharmacopoeias, is still underdeveloped. The analysis of these manuscripts is often carried out by experts, without systematic recourse to linguistic technologies that could automate these processes. This paper proposes an approach integrating a chain of disambiguation and named entity linking applied to English translations of medieval Arabic pharmacopoeias through the use of resources such as BabelNet, Wikidata and GBIF to link plant mentions to their modern scientific identifiers.

Le « Management de la Connaissance » : la clé stratégique de la réflexion sur l'apport de la « Mémoire d'Entreprise »

Alain Berger*

*Ardans SAS,
6 rue Jean Pierre Timbaud, «Le Campus» Bâtiment B1,
78180 Montigny-le-Bretonneux, France
aberger@ardans.fr,
https://www.ardans.fr

Résumé. En 2024, l'Observatoire B2V des Mémoires® s'est emparé de la question de la « *mémoire de l'entreprise* » et a lancé des actions concrètes pour pointer ce sujet dans les sphères managériales. Membre de son Conseil Scientifique, le Pr Jean-Gabriel Ganascia nous a invité à éclairer cette réflexion par notre vision industrielle des démarches en cours et de l'apport de l'intelligence artificielle dans le KM, la question étant ainsi formulée :

« *Comment positionner le « Management de la Connaissance » dans cette réflexion à propos de la « Mémoire d'Entreprise »?* »

L'article présente dans une première partie la démarche considérée comme pionnière, singulière et exemplaire du Commissariat à l'Énergie Atomique et aux Énergies Alternatives.

Dans un deuxième temps, un regard est posé sur l'arrivée de la norme ISO30401 et les exigences attendues sur les SKM ou « *Systèmes de Management de la Connaissance* ».

La contribution de l'Intelligence Artificielle déjà significative et la venue sur les grands modèles de langage sont abordés avec les limites industrielles de prudence qui s'imposent.

La conviction de l'auteur est claire : le KM est une clé stratégique pour adresser le sujet de la « *Mémoire d'Entreprise* ».

Mots-clés : Management de la Connaissance, Knowledge Management, KM, Mémoire d'Entreprise, Capitalisation et Exploitation des Connaissances, PARNASSE , KB_ Scope®, Observatoire B2V des Mémoires®, Intelligence Artificielle, Connaissance clé, Connaissance cruciale, Ingénierie de la Connaissance, Transfert de Connaissance d'Expert, Système de Knowledge Management, ISO30401, Grands Modèles de Langage.

Keywords : Knowledge Management, KM, Corporate Memory, Knowledge Capitalisation and Exploitation, PARNASSE , KB_ Scope®, Observatoire B2V des Mémoires®, Artificial Intelligence, Key knowledge, Crucial knowledge, ISO30401, Knowledge Engineering, Expert Knowledge Transfer, Knowledge Management System, Large Language Models.

1 Avant-Propos

Quand début 2024, l'Observatoire B2V des Mémoires® [1] affirme que « *Sans mémoire, pas d'avenir* », la question de la *« Mémoire d'Entreprise »* est au cœur du sujet. Son Conseil Scientifique valide l'engagement pris par la Direction Générale d'instruire ce sujet stratégique pour les entreprises et leurs salariés. Le Pr Jean-Gabriel Ganascia nous a alors invité à éclairer cette réflexion par notre vision industrielle, justifiée par vingt-cinq années de réalisation de telles opérations et, complétée par notre connaissance de l'intelligence artificielle : ceci afin d'illustrer son apport dans le *« Management de la Connaissance »* (ou Knowledge Management *i.e.* (KM). La question était donc initialement ainsi formulée :

« *Comment positionner le « Management de la Connaissance » dans cette réflexion à propos de la « Mémoire d'Entreprise » ?* »

Il convient de rappeler que l'action de l'Observatoire B2V des Mémoires® s'est déjà concrétisée par un premier sondage d'opinion [2], par une conférence à Lille le 7 juin 2024 [3] par un enseignement sous la forme d'un Certificat de formation Continue « *Mémoire de l'Entreprise* » [4] avec l'Université Paris Dauphine-PSL et la Fondation Maison de Salins, et par une première action mémorielle auprès de retraités chez un industriel.

Notre intervention lors du Certificat a mis en lumière toute la problématique de cette réflexion sur l'apport du KM dans la *« Mémoire d'Entreprise »*, d'où cette proposition de clarification de notre perception.

2 Introduction

Comme tout organisme vivant, l'entreprise se dote naturellement d'une mémoire. Elle construit des documents, des procédures, des archives tant pour son existence administrative que pour exercer ses activités métiers. La mise en place de systèmes d'information pour les différentes fonctions de soutien ou de production fait que si l'enjeu de l'efficience collective est basée sur ce support technologique informatique, la pertinence de la justification « *métier* » reste dans la tête des humains qui font montre de discernement et surtout d'expertise.

Comment pérenniser ces savoirs, comment les expliciter, comment les transmettre, comment les exploiter ? Dans les systèmes de management, celui de la qualité (ISO9001:2015) est depuis 2018 consolidé par celui dédié au système de management de la connaissance (ISO30401:2018).

A ceux qui souhaitent se doter d'une solution « base de connaissance » augmentée par une intelligence artificielle, il convient de leur recommander de débuter par se doter d'un « *Système de Management de la Connaissance* » (ou SKM pour « *Système de Knowledge Management* ») autour d'une équipe dédiée, ceci afin de pallier un tel déficit organisationnel.

1. https://www.observatoireb2vdesmemoires.fr
2. Sondage réalisé par l'institut IFOP sur le sujet « *mémoire de l'entreprise* » en mars 2024 auprès de 1000 cadres français : https://www.observatoireb2vdesmemoires.fr/sondage-dopinion
3. Conférence avec le Medef des Hauts de France et l'Institut Choiseul présentant la démarche en cours et les premiers résultats obtenus https://www.observatoireb2vdesmemoires.fr/lobservatoire/memoire-de-lentreprise/conference-lentreprise-en-memoires
4. https://executive-education.dauphine.psl.eu/formations/certificat/memoire-entreprise

S'il faut donner du temps au temps pour une telle mise en place, les différentes étapes qui seront réalisées consolideront tant la maîtrise des processus métier que celles des compétences nécessaires pour identifier les savoirs clés et pérenniser les connaissances cruciales. Un tel SKM renforce l'identité culturelle de l'organisme, améliore la qualité des échanges par un langage commun partagé, accélère l'intégration de nouvelles ressources humaines, consolide la qualité des produits ou services rendus, et appuie la R&D pour anticiper les innovations futures. C'est certes un long chemin, cependant il procure en général un résultat particulièrement fructueux pour ceux qui l'ont emprunté.

Nous observerons que le *« Management de la Connaissance »* est bien devenu la clé stratégique de la réflexion sur l'apport de la *« Mémoire d'Entreprise »*.

L'article introduit cette démarche considérée comme pionnière, singulière et exemplaire conduite au sein du Commissariat à l'Énergie Atomique et aux Énergies Alternatives.

Dans un deuxième temps, un regard est posé sur l'apport de la norme ISO30401, et les exigences attendues sur les SKM.

A la contribution préliminaire de l'Intelligence Artificielle déjà significative, la venue des grands modèles de langage est abordée avec les limites industrielles de prudence qui s'imposent : halte aux ultracrépidarianistes [5] [Ganascia (2023)].

3 La remarquable exemplarité du CEA

Il paraît important de préciser que le choix de cet organisme est aussi une preuve que les structures dites étatiques ont mis en place des mécanismes remarquables qui durent dans le temps indépendamment des évolutions politiques à la tête de l'état ; les intérêts fondamentaux de la Nation [6] et une « *Mémoire de la Nation* » est alors instaurée.

Voici comment la mission initiale de cet EPIC [7] de développer de nouvelles connaissances scientifiques et transfère des innovations technologiques auprès du monde industriel, a pris un nouvel essor ces trente dernières années.

Dans le prolongement de la signature du traité d'interdiction des essais nucléaires décidé en 1996, le Président de la République, Jacques Chirac, en visite au Commissariat à l'Énergie atomique-Direction des applications militaires (CEA DAM), au Centre de Bruyères-le-Châtel, le 7 septembre 2006, déclarait [8] qu'il « *assignait au CEA, et notamment à la DAM, la mission afin qu'il continue à assurer la crédibilité de notre dissuasion et à participer au développement et au rayonnement mondial de la science française* » ; c'est un message clair où l'institution doit rester à l'état de l'art.

5. « *Sutor, ne supra crepidam* », littéralement, le cordonnier (*sutor*), pas plus haut que la sandale (*crepidam*). Rapportée par Pline l'ancien dans son Histoire naturelle, cette sentence latine signifie que, « *de ce qui va au-delà de son métier, et que l'on ignore, on ne devrait parler* ».

6. Article 410-1 du Code pénal https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006418343.

7. EPIC pour Établissement Public Industriel et Commercial : c'est une personne morale de droit public gérant un service public spécialisé, distincte de l'État, et des collectivités territoriales, mais rattachée à eux.

8. https://www.vie-publique.fr/discours/163298-declaration-de-m-jacques-chirac-president-de-la-republique-sur-la-dis

3.1 La traduction de la mission

Après avoir réalisé un programme de simulation numérique qui garantisse la fiabilité et la sûreté des armes, la question de la pérennité du savoir des scientifiques impliqués se posait alors. Il convenait aussi de se doter de l'outil scientifique qui allait valider les phénomènes de

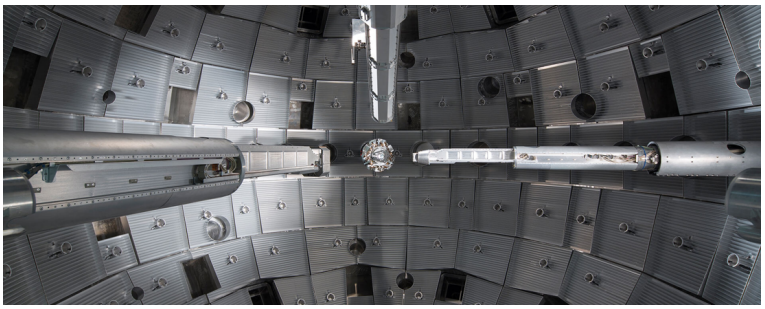


FIG. 1 – *L'intérieur de la chambre d'expériences du Laser MégaJoule - Photo © CEA*

la physique fondamentale prévus par des expérimentations en laboratoire : c'est la finalité du Laser MégaJoule (cf. figure 1).

3.2 De l'esprit humain à l'explicitation et la modélisation à des fins de pérennisation pour l'exploitation et la transmission

Le programme CEC pour « Capitalisation et Exploitation des Connaissances » allait se voir confié dès avril 1996 la mission de, pérenniser et transmettre les connaissances initialement acquises lors de ces essais, puis, de préserver les savoirs critiques ou cruciaux avant le départ d'experts, et enfin, lorsque les expérimentations d'un projet se heurtent à un mur technologique, de conserver tous les acquis et retours d'expérience à la suspension de ces travaux de recherche. Bien évidemment, cette dimension stratégique de gestion de la connaissance avait été anticipée et confirmée bien avant par la Direction du CEA, dès 1994, comme une directive de son manuel qualité [van Craeynest et al. (1997)].

3.3 Une mémoire d'entreprise pour une double efficience

Cet exemple est particulièrement riche car il s'agit pour un tel organisme de se doter de l'outil de mémoire qui lui confère une double efficience technique et économique ; en disposant d'un moyen d'éviter de refaire deux fois la même étude et la même expérimentation, et donc, d'éviter de perdre du temps en étant sûr de consacrer l'argent de son budget à aller plus avant dans la recherche.

3.4 Une mémoire d'entreprise multi-facettes

Cet outil de mémoire est multi-facettes dans la mesure où elle se traduit par différents types d'actions :

- « *Agréger* » : dans des codes de calculs les fruits des expériences analysées et de leur modélisation associées,
- « *Numériser* » : des documents (avant qu'il ne s'effacent), les référencer et les archiver selon les règles,
- « *Filmer* » : les gestes métiers appropriés dans les opérations manuelles,
- « *Recueillir et expliciter* » : les retours d'expériences, les savoirs, les expertises des sachants avant qu'ils ne quittent leurs fonctions.

3.5 Une cartographie partagée pour analyser le contenu de cette mémoire

Ce qui est intéressant, c'est que ces travaux ont permis d'approcher une nouvelle vision de cette mémoire du savoir : une vision « méta » traduite sur une carte (cf. figure 4). Le processus de l'activité métier qui consomme k ressources cognitives qui sont détenues par x acteurs répartis sur y sites (cf. figure 2) [Vexler et al. (2013)]. Cette prise de conscience révèle les sujets qui sont prioritaires parmi ceux qui sont à prendre en compte. On parle de **connaissances clés** - celles pour la réalisation des activités qui constituent le savoir-faire métier et sont l'élément différenciant vis-à-vis de la concurrence - qui peuvent devenir des **connaissances cruciales** [Belloni et al. (2017)] - celles sans lesquelles les problèmes essentiels de l'entreprise n'ont pas de solution - en particulier lorsque le maintien de ce savoir n'est pas garanti à un horizon défini. L'exemple de cartographie [9]

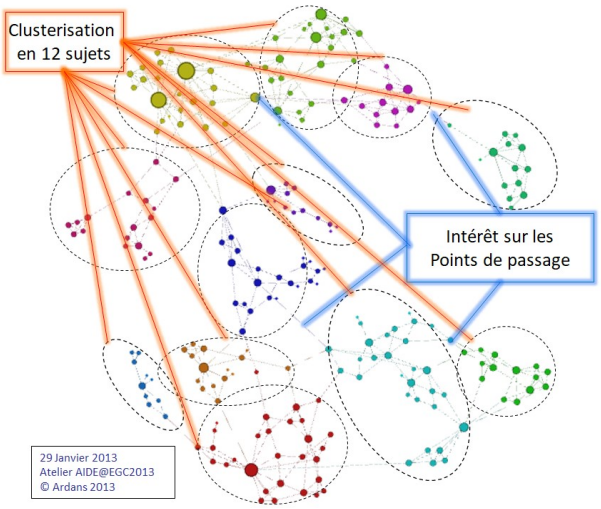


FIG. 2 – *Les « clusters » du graphe de l'expertise modélisée avec AKM [Vexler et al. (2013)]*

9. AKM ou Ardans Knowledge Maker® v2025 est l'implantation de la méthode Ardans [Mariot et al. (2007)].

3.6 L'émergence d'une stratégie d'une organisation pour manager cette mémoire

Les actions détectées sont alors ordonnées selon des critères bien établis : il est clair qu'une stratégie de management de la pérennisation de la connaissance émerge et se déroule. Ceci a nécessité une réflexion sur l'organisation à mettre en place pour réaliser ces cartographies, les consolider, les agréger, les analyser pour proposer la stratégie pertinente fondée sur ce processus d'élicitation d'une solution optimale pour toute l'entité.

Si cette démarche peut sembler être une tautologie, il faut considérer deux notions fondamentales liée à l'IGI1300 [Premier-Ministre (2021)] qui définit les exigences de sécurité des systèmes d'information amenés à traiter des informations ou supports classifiés. Celle du « *droit à en connaître* » et celle de la « *protection des personnes physiques* » dont la conséquence est une distribution sur le territoire des visions partielles du sujet à résoudre collectivement. A ce point, les actions peuvent être planifiées par une équipe dédiée maîtrisant les outils de capitalisation [10] de cette mémoire, et ce, afin de garantir la mission assignée au CEA.

4 La question du SKM traduite en norme

Ainsi, le savoir individuel a été échangé, montré, expliqué (ou Socialisé) avant d'être traduit, formalisé, explicité pour être transmis (ou Externalisés) à une communauté choisie. Elle va alors le structurer et l'intégrer (ou le Combiner) pour générer de nouveaux éléments qui appropriés (ou Internalisés) vont constituer de nouveaux savoirs individuels qui à leur tour... On retrouve ce qui a été identifié et qualifié de SECI par Nonaka [Nonaka et Takeuchi (1995)].

Depuis 2018, la question s'est ainsi naturellement replacée vers la question du SKM ou Système de Management de la Connaissance. Les organismes confrontés au §7.1.6 de la norme ISO9001:2015, ont pu apprécier la retranscription de la question du management de la connaissance organisationnelle dans la norme ISO30401:2018 [Secretary (2018)]. Il ne s'agit pas d'une révolution mais d'une consolidation où coexistent les travaux de Nonaka (cf. supra), de Grundstein [Grundstein (2003)] et de Deming [11] !

L'apport majeur de cette norme est de poser les bases d'un langage commun sur les exigences attendues pour qu'un organisme puisse se prévaloir de l'implantation d'un Système de Management de la Connaissance.

4.1 Une vision processus pour s'approprier et se mesurer à la norme

La question du management de la connaissance est *in fine* une question de progrès pour l'espèce humaine. Le passage de l'oralité à l'écrit ne s'est pas passé en une génération. De la même façon que l'échange oral est l'objet d'une langue avec ses propres codifications, celle de la représentation écrite a été l'objet d'avancée technique de la pierre, au papyrus, au papier, du pictogramme, idéogramme, hiéroglyphe, à une écriture syllabique puis alphabétique avec des caractères... la formalisation étant une codification et la lecture un décryptage. A l'heure

10. Capitaliser sur les connaissances de l'entreprise, c'est considérer les connaissances utilisées et produites par l'entreprise comme un ensemble de richesses constituant un capital, et en tirer des intérêts contribuant à augmenter la valeur de ce capital [Grundstein (1995)]

11. La roue de Deming est une représentation graphique de la méthode d'amélioration continue des processus et de gestion de la qualité dite PDCA (Plan-Do-Check-Act)

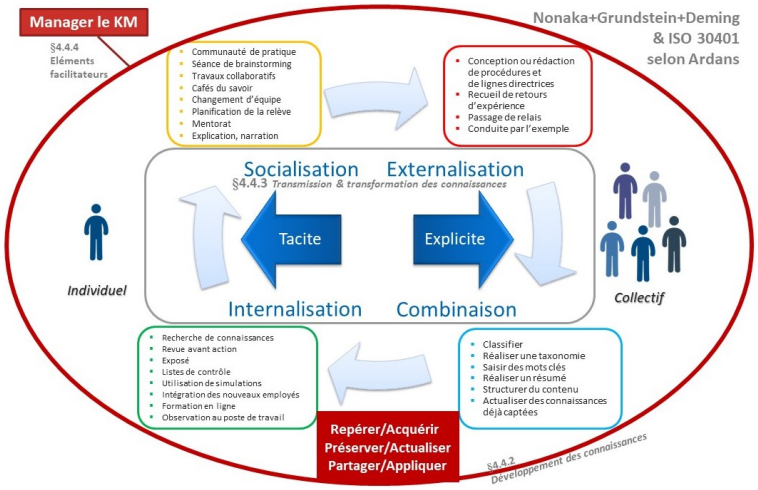


FIG. 3 – *L'ISO30401 sur les exigences des SKM revisité par Ardans [Berger (2023)]*

du numérique, les sujets se sont décalés mais sont de même nature : comment représenter une image ? En point, en vecteur ? en discret ou en continu ? et la couleur ? Et le format. . .lequel sera celui de référence dans 10 ans ou dans un siècle ? Idem pour l'animation de l'image, la vidéo, le son, le texte. . . quel est le bon format qui va s'avérer pérenne ? La question est bien de savoir comment un organisme s'approprie la bonne organisation pour faire grandir le savoir contenu dans sa « mémoire collective d'entreprise », partagée par ses collaborateurs et sécurisée par rapport à des éventuelles agressions hostiles.

Si la réponse est « ce n'est pas avec un outil informatique », il est clair aussi que l'outil informatique accompagne une démarche de management de la connaissance afin que l'utilisateur habilité puisse consulter, contribuer, questionner ses pairs, actualiser les contenus de son domaine de compétence. La prise en compte de la culture et du métier est essentielle pour que le dispositif puisse se fondre dans le quotidien. Il se révèle enfin indispensable de mettre en place la gouvernance qui s'impose et qui dispose des moyens adéquats, les comités qui modèrent, le langage commun qui est partagé dans le métier, et les processus qui concourent à la bonne hygiène de la vie du Système de Management de la Connaissance.

La mémoire de l'entreprise est définitivement de la responsabilité des humains qui y collaborent à commencer par les responsables qui la dirigent.

4.2 Le Portail pour manager en connaissance le KM

Il convient de citer à ce point l'excellente initiative de l'association « Club Gestion des connaissances » qui, après avoir participé depuis son origine à l'établissement de la norme ISO30401 pour la France (via l'Afnor), l'a traduit avec une vision processus [Coustillière (2022)] au sein d'une méthode et d'un outil : PARNASSE [12] . L'idée est de rendre audible la

12. PARNASSE acronyme de Portail Associant la Référence Normative avec un Référentiel Structuré d'Entreprise

Le KM, clé stratégique sur l'apport de la « Mémoire d'Entreprise »

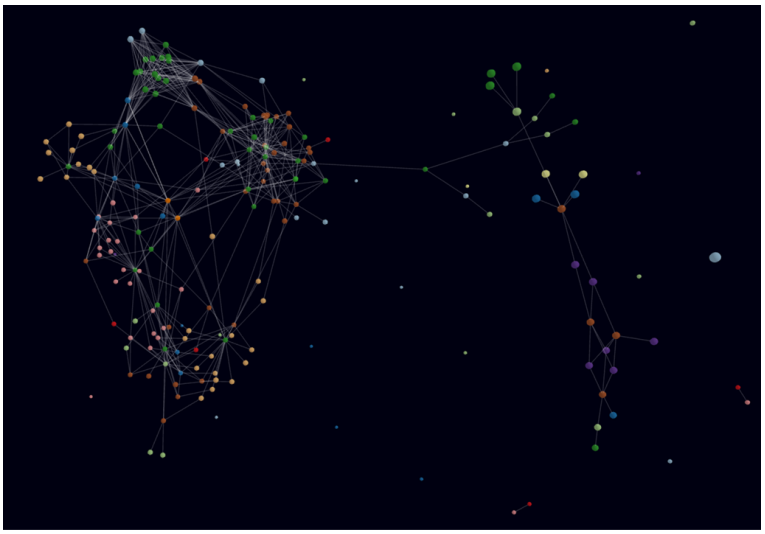


FIG. 4 – *Le graphe "3D" d'une expertise selon l'œil du KB_ Scope® d'AKM*

norme, d'aider celui ou celle qui aura le rôle de Knowledge Manager par la mise à disposition d'un outil qui clarifie à quoi ressemble le KM dans son organisation.

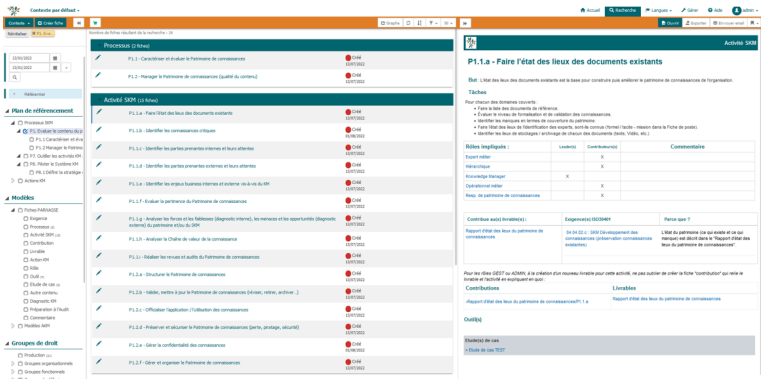


FIG. 5 – PARNASSE *guide le knowledge manager dans la visualisation du processus SKM*

Da la même façon qu'une expertise peut s'illustrer sous forme du graphe (2D/ 3D) de la base de connaissance (KB) (cf. figure 4), le système KM peut se décrire sous forme de 8 processus et 20 activités (cf. tableau 1) et alors se manipuler avec de simples liens dans PARNASSE . Avec une telle modélisation, le Knowledge Manager dispose de l'outil pour maîtriser et manager en parfaite connaissance le SKM (cf. figure 5) et ainsi le processus de mémoire de son entreprise.

PARNASSE **: Les processus du SKM de référence du Club Gestion des Connaissances**
Processus 1. Evaluer le contenu du patrimoine et le gérer
▷ P1.1 - Caractériser et évaluer le Patrimoine de connaissances
▷ P1.2 - Manager le Patrimoine de connaissances (qualité du contenu)
Processus 2. Faire vivre le patrimoine de connaissances et garantir son application
▷ P2.1 - Formaliser et mettre à disposition les connaissances
▷ P2.2 - Garantir l'application des connaissances
▷ P2.3 - Recenser les connaissances utiles à l'Organisation
▷ P2.4 - Gérer les Communautés de savoir et gérer l'expertise
Processus 3. Gérer et piloter les dispositifs d'acquisition de connaissances
▷ P3.1 - Processus RH - Recenser le besoin en formations nécessaires à l'activité (actuelle et future)
▷ P3.2 - Processus RH - Gérer et piloter l'apprentissage individuel (MOOC, e-learning, Coaching, ...)
▷ P3.3 - Gérer et piloter l'apprentissage en interaction collective (groupes d'expertises, séminaires, communautés d'apprentissage...)
▷ P3.4 - Définir les besoins en recrutement en lien avec les connaissances critiques de l'Organisation
▷ P3.5 - Processus RH - Gérer et piloter la construction des formations et solutions d'apprentissage
Processus 4. Soutenir les dispositifs de créativité et d'innovation
▷ P4.1 - Soutenir les activités de créativité
▷ P4.2 - Soutenir l'activité d'innovation
▷ P4.3 - Faire le bilan des connaissances acquises au cours des activités d'innovation / créativité
Processus 5. Soutenir les processus opérationnels
Processus 6. Transformer l'information externe en connaissance utile pour l'organisation
Processus 7. Outiller les activités KM
▷ P7.1 - Interagir avec les outils d'IA
Processus 8. Piloter le Système KM
▷ P8.1 - Définir la stratégie et les objectifs KM
▷ P8.2 - Construire le plan KM accepté par la direction de l'Organisation
▷ P8.3 - Évaluer le Système KM : les audits
▷ P8.4 - Superviser le Système KM : processus de décision, revues de pilotage, tableaux de bord des indicateurs, ressources humaines et matérielles, niveau de compétence...
▷ P8.5 - Organiser et conduire les actions de mise en place et d'amélioration du Système KM : sensibiliser, communiquer, mobiliser les acteurs, conduire les actions, ...

TAB. 1 – PARNASSE *: Processus du SKM de référence du Club Gestion des Connaissances.*

4.3 Une vision utilisateur pour exprimer ses attentes par rapport au SKM

Pour être plébiscité dans l'industrie, la base de connaissance ou le Système de Management de la Connaissance doit satisfaire aux attentes des acteurs (cf. figure 6) dont notamment :

- « *exhaustivité* » : il convient que la connaissance soit exhaustive sur le périmètre sur laquelle elle est porte afin d'obtenir la confiance de l'utilisateur à commencer par lui transmettre la réponse pertinente ;

FIG. 6 – *Parmi les attentes d'un utilisateur de SKM*

- « *clarté* » : les contenus sont clairs, dénués de toute ambiguïté, cela pour faciliter l'adhésion, l'appropriation et le bon usage par l'utilisateur,
- « *consistance* » : les résultats de « *navigation* » pour obtenir les contenus sont consistants ; cette stabilité rassure l'utilisateur ;
- « *contextualisation* » : il est fondamental de bien décrire le contexte dans lequel cette connaissance est valide pour être exploitée en toute sérénité ;
- « *position* » : l'élément de connaissance consulté est au cœur d'un réseau (implicitement sémantique) d'éléments de connaissance au sein desquels il doit être positionné dans une représentation cartographique multidimensionnelle : un réseau précieux pour évaluer la qualité de la base comme son homogénéité, ses relations, ses trous, ses densités ;
- « *diffusion* » : la connaissance est un actif précieux et est restreinte à ceux habilités à en connaître, celui qui en bénéficie doit savoir le mesurer ;
- « *convivialité* » : plus que jamais l'ergonomie d'un système à base de connaissance moderne doit être d'une ergonomie intuitive et fluide et démontrer qu'elle offre un retour sur investissement à l'usager sans pareil ;
- « *validité* » : la connaissance est vivante, comme elle s'affine dans le temps, elle est intrinsèquement « non monotone » et doit être datée ;
- « *argumentation* » : les contenus sont argumentés et disposent des niveaux de preuve nécessaires pour une bonne appropriation par le lecteur ;

5 L'apport de l'intelligence artificielle

Le *« Management de la Connaissance »* pratiqué par les ingénieurs de la connaissance [Berger (2015) est directement issu de la branche "connaissance" de l'intelligence artificielle telle que définie par McCarthy [McCarthy et al. (1955)] lors du . Les systèmes experts et les systèmes à base de connaissance s'ils ne fonctionnent pas de la même façon que les bases de connaissance actuelles, les deux dispositifs nécessitent tous de colliger de la connaissance, c'est à dire de l'expliciter et de suivre un processus pour faire qu'elle soit validée avant d'être mise à disposition pour être dans le futur actualisée. La validation des règles des systèmes experts était déjà très complexe (maîtrise du déclenchement de la règle ou validation du système

général) ; celle des bases de connaissance actuelle est plus abordable même si elle requiert la même attention et la même précision.

5.1 La contribution dans l'abstraction et la représentation

L'outil reconnu maintenant comme le référent dans les démarches de capitalisation d'expertise intègre des notions industrialisées issues des techniques de l'intelligence artificielle. Quand on parle de « *solution hybride* » cela se confirme car il exploite ces dispositifs :

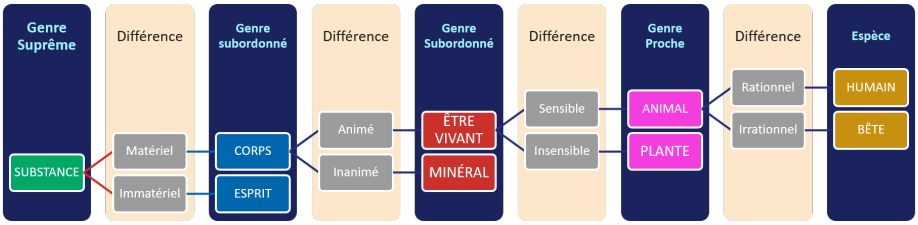


FIG. 7 – *L'arbre de Porphyre : à chaque étape un genre se différencie du précédent*

- « *ontologie* » : La constitution d'arborescences de concepts classifiés pour décrire le langage du métier est très précieux : cela aide le novice à comprendre cette hiérarchie de termes, à les positionner les uns par rapports aux autres, cela aide à décrire les environnements qu'ils soient physiques (comme dans l'ingénierie système) ou fonctionnels, que cela soit des contextes de travail ou des notions de priorité de droit ; L'arbre de Porphyre (234-305) (cf. figure 7) est l'ancêtre de cette représentation.
- « *objet* » : La représentation des connaissances est très friande de l'usage de cette forme de langage ; Que l'on parle de Classe, Attribut, et Instance, ou de Modèle, Rubrique, et Fiche, il s'agit de peu ou prou de la même chose !
- « *héritage* » : L'héritage est une notion essentielle de la programmation orientée objet qui permet de définir une nouvelle classe à partir de classes existante ; idem pour les modèles qui peuvent hériter de modèles.
- « *graphe* » : Le liage entre les éléments qui existent dans les bases de connaissance est extrêmement précieux. Il s'agit d'un véritable « *réseau sémantique* » qui est élaboré au fil de l'eau et par construction. Il est autant utilisé pour l'élaboration que pour la consultation de la base de connaissance [Vexler et al. (2013)].
- « *apprentissage* » : L'indexation n'est pas uniquement syntaxique, elle est aussi sémantique. elle se réalise par un apprentissage sur les contenus validés de la base de connaissance pour appuyer l'utilisateur en consultation comme en contribution [Vexler et al. (2020); Berger et al. (2020)].
- « *hypertexte, url, web* » : Le premier succès de l'IA est cette capacité de lien [Ganascia (2024)] si implicite à nos systèmes actuels qu'il serait indécent d'omettre de le citer !

5.2 La question de la confiance dans la réponse

La réalisation d'une action de *« Mémoire d'Entreprise »* par une opération de *« Management de la Connaissance »* correspond à ce que l'on qualifie de « *EKT* » pour Transfert de Connaissance d'Expert (*i.e. Expert Knowledge Transfer*). Nous avons évoqué supra le processus long qui concerne l'explicitation de la connaissance avec en particulier le travail méticuleux et précis de la validation d'une base de connaissance : la photographie (cf. figure 8) illustre ici concrètement à quoi ressemble un entretien de validation d'une base de connaissance. L'ingénieur de la connaissance échange avec l'expert face à lui sur l'explicitation qui a été traduite dans la base de connaissance. Ce travail est présenté sur l'écran de l'ordinateur qui est retransmis simultanément via la vidéoconférence à l'expert receveur qui dans ce cas se trouve à 1350 km.

On observe ainsi que la retranscription de cet « *élément de connaissance* » est soumis à l'expert pour valider la fidèle élicitation de son point de vue. Ce contenu est alors proposé au regard du receveur, qui étant déjà un spécialiste du domaine va naturellement « *stresser* » ce contenu. De cette réaction, un ajustement est réalisé si nécessaire afin de consolider l'appropriation par le receveur. Dans certains cas, un complément nécessitant un nouvel élément de connaissance est produit et ajouté : la formulation de Davenport et Prusak (1998) a été justement transcendée en « *Transfert de connaissances = Transmission + Absorption & Utilisation + Enrichissement* » [Berger et al. (2024)] [13]. Ce mode de construction garantit une excellente

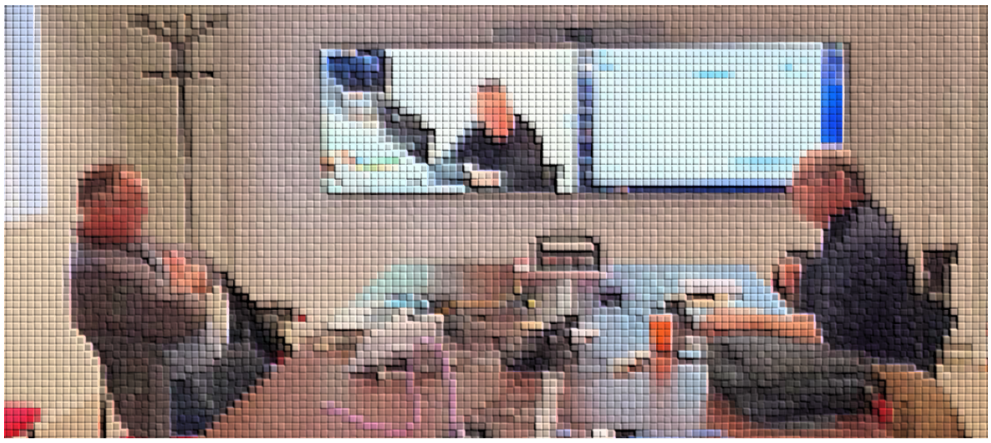


FIG. 8 – *Le travail méticuleux et précis de la validation d'une base de connaissance*

confiance dans le système qui gère cette mémoire : à charge de l'organisation de maintenir la bonne actualisation du travail consciencieux initialement produit.

13. A la version **Knowledge Transfer = Transmission + Absorption & Use** est ajouté **+ Enrichment**

5.3 L'interrogation sur l'exploitabilité en confiance des LLM pour cette *« Mémoire d'Entreprise »*

L'arrivée des grands modèles de langage (en anglais « *Large Language Models* » ou LLM) est en train de chahuter la question de l'exploitation des outils informatiques pour interroger de grands volumes de données et de textes en particulier. Le sujet sur lequel il n'y a pas de question est que les bases de connaissance qui contiennent une mémoire stratégique de l'entreprise, ne sont pas connectées à l'extérieur pour des questions triviales de sécurité. La problématique concerne donc du volume de "contenu" qui est nécessaire pour disposer de la capacité à disposer d'un apprentissage pertinent afin d'avoir des réponses « *pertinentes* » et de celui à partir duquel les « *hallucinations* » vont insidieusement apparaître. Aujourd'hui, nous observons sur nos bases qu'une progression dans le séquencement des étapes suivantes délivre des résultats prometteurs selon Moris (2024) :

1. Disposer d'une ontologie et d'une modélisation pertinente.
2. Avoir un corpus de connaissance validé significatif.
3. Avoir réalisé une indexation sémantique sur ce corpus.
4. Exploiter un RAG au LLM.
5. Générer le bon « *Prompt* » pour converger vers une réponse pertinente.

Le dernier point est essentiel pour conserver la confiance dans le dispositif. On note les impératifs suivants :

— Il faut que l'utilisateur puisse disposer de la « *justification* » de cette réponse avec la ou les éléments de connaissance qui ont permis de répondre à la requête.
— Le système doit être en mesure de s'abstenir de générer des extrapolations si la base de connaissance est vide sur le sujet du questionnement.
— Le système doit être en mesure de garantir le fait de rester dans le périmètre maîtrisé par la base et donc de ne pas générer des hallucinations en mélangeant allègrement des notions présentes mais décontextualisées.

Le danger étant que l'utilisateur soit fasciné par la réponse et ne soit pas en mesure de discerner une assertion fantasque magnifiquement sublimée par une écriture délicieusement spécieuse.

6 La « mémoire de l'entreprise » utile valorisée par le management de la connaissance

Si Friedrich Nietzsche affirme que « *le futur appartient à celui qui a la plus longue mémoire* », il faut admettre que l'entreprise a un véritable enjeu patrimonial à savoir valoriser sa mémoire collective. Les processus métiers, qui évoluent pour suivre les changements réglementaires ou législatifs, pour anticiper les mouvements des marchés qui les composent, pour enrichir l'offre en continu pour satisfaire la clientèle, doivent être parfaitement appropriés par les personnels concernées. Ils doivent aussi disposer de l'entière adhésion de tous les collaborateurs qui contribuent aux produits ou aux services considérés. Seulement la transmission d'un savoir et son assimilation pour qu'il devienne une compétence réelle n'est pas immédiate.

De plus, s'il ne s'agit pas de noyer le collaborateur par des informations superfétatoires (car prochainement obsolètes), il convient de réfléchir à deux fois si l'obstacle qui bloque

l'innovation aujourd'hui ne sera pas contourné demain ; ainsi mémoriser une telle situation serait potentiellement profitable à l'entreprise demain.

Entre infobésité et pertinence le choix implique du discernement qui doit être fourni si demandé à ce même collaborateur. C'est d'une part une question de confiance, et d'autre part la fondation de cette justification sera éclairante pour rassurer l'acteur, comme pour le questionner, et donc, pour suggérer de sa part une réaction qui consolidera l'édifice ou l'enrichira par un nouvel actif.

La complexité comme la richesse des objets manipulés dans les organisations font que, à la connaissance verticale d'un expert, il s'ajoute une connaissance transversale d'un architecte, voire une expérience en profondeur de la vie de l'objet en exploitation par le mainteneur.

La synthèse du cas CEA DAM illustre l'exemplarité de cette démarche de mise en place d'une mémoire avec une pérennisation du savoir à plusieurs niveaux : celui de la stratégie et de la finalité de cette mémoire d'entreprise, celui de la maîtrise du processus métier dans le contexte de sécurité attendu, celui de l'organisation pour détecter, acquérir, pérenniser, exploiter, actualiser les acquis, avec celui essentiel de la mise en place d'une équipe dotée des ressources pour éliciter et faire vivre cet actif patrimonial.

Acquérir une vision holistique de la connaissance dans certaines organisations est une gageure aujourd'hui, mais il va bien falloir s'attacher à rendre cette mémoire vivante et accessible dans tous les sens du terme. On se gardera de croire encore pendant quelque temps que les moteurs fondés sur les grands modèles de langage résoudront le sujet en un clic : La mémoire d'entreprise est définitivement un sujet stratégique. Elle ne peut se construire que collectivement, en intégrant la culture et l'ADN de l'organisme afin de s'appuyer sur les acquis du passé et du présent pour mieux préparer l'avenir.

Références

Moris, E. (2024). Sémantique et LLM dans AKM. In *8ème édition d'Ardans Users'Group Meeting (AUGM2024)*. Ardans, Paris-Saclay, France.

Ganascia, J.-G. (2024). *L'I.A. expliquée aux humains*. Paris : Edition du Seuil.

Berger, A., S. Boblet, T. Cartié, J.-P. Cotton, et F. Vexler (2024). Implanter une approche hybride dans une démarche d'ingénierie de la connaissance pour manager les avis techniques relatifs au retour d'expérience d'exploitation d'un équipement sensible complexe. In *35èmes Journées francophones d'Ingénierie des Connaissances (IC 2024)*. Association Française pour l'Intelligence Artificielle and Laboratoire L3i La Rochelle Université.

Berger, A. (2023). Regard sur l'ingénierie de la connaissance face à l'ISO30401. In *34es Journées francophones d'Ingénierie des Connaissances (IC 2023)*, Volume `https://hal.science/hal-04152777/file/PFIA2023IC_IC_%26_Iso30401_Alain_Berger.pdf`, Strasbourg, France. Plate-Forme Intelligence Artificielle (PFIA 2023) and AfIA & iCube.
french

Ganascia, J.-G. (2023). Le cordonnier, l'ultracrépidarianiste et chatgpt. In *Sciences et Avenir - Mai 2023*, Volume `https://lirelactu.fr/source/sciences-et-avenir/567d4840-b6d9-4e8b-9b78-fbc036cf8a4f`.

Coustillière, P. (2022). L'ingénierie système, un outil pour le km manager ? In *Club Gestion des Connaissances*, Volume https://www.clubgc-km.fr/articles/68927-05 of *Revues des Nouvelles Technologies de l'Information*.

Premier-Ministre (2021). Instruction générale interministérielle 1300 sur la protection du secret de la défense nationale. In *https://cyber.gouv.fr/instruction-generale-interministerielle-n1300*, Paris, France, pp. 225. JORF n°0185 du 11 août 2021.

Berger, A., F. Vexler, C. Mary, et J.-P. Cotton (2020). Réflexion sur le choix d'un classifieur sémantique destiné à aider le cogniticien dans l'élaboration d'une base de connaissance et la garantie de sa consistance dans le temps. In *6ème Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle, APIA 2020*, Volume http://pfia2020.fr/wp-content/uploads/2020/08/Actes_CH_PFIA2020_V3.pdf, Angers, France, pp. 66–73.

Vexler, F., C. Mary, A. Berger, et J.-P. Cotton (2020). Management des connaissances augmenté : usage d'un classifieur sémantique pour l'aide à l'élaboration et au maintien en cohérence d'une base de connaissance. In *20èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2020*, Volume RNTI-E-36 * https://editions-rnti.fr/?inprocid=1002598, Bruxelles, Belgique, pp. 393–400.

Secretary, I. C. (2018). Knowledge management systems — requirements iso30401:2018. In *https://www.iso.org/standard/68683.html*, International Organization for Standardization. Geneva, CH.

Belloni, A., A. Berger, et J. Cotton (2017). Cibler une action de gestion des connaissances appropriée dans un cadre industriel : retour d'expérience d'ardans. In S. Bringay (Ed.), *3ème Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle, APIA 2017, Caen, France, July 3-4, 2017*, pp. 35–43.

Berger, A. (2015). Évolution dans l'industrie du métier d'ingénieur cogniticien ou d'ingénieur de la connaissance entre 1985 et 2015. In *1st Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA 2015) at the Plate-forme Intelligence Artificielle*, Rennes, France, pp. 23–33.

Vexler, F., A. Berger, J.-P. Cotton, et A. Belloni (2013). Eléments d'appréciation et d'analyse d'une base de connaissance : l'expérience industrielle d'Ardans. In *Actes Atelier AIDE à EGC'2013, 13ème Conférence Francophone sur l'Extraction et la Gestion des Connaissances*, Volume https://eric.univ-lyon2.fr/aide/actesAIDE_EGC2013ENLIGNE.pdf, Toulouse, France, pp. 59–72.

Mariot, P., C. Golbreich, J.-P. Cotton, et A. Berger (2007). Méthode, Modèle et Outil Ardans de capitalisation des connaissances. In *RNTI E12 Modélisation des Connaissances*, Volume https://editions-rnti.fr/render_pdf.php?p=1000709, pp. 187–206.

Grundstein, M. (2003). De la capitalisation des connaissances au management des connaissances dans l'entreprise, les fondamentaux du knowledge management. In *Management des connaissances en entreprise*, pp. 25–54. Economics Papers from University Paris Dauphine.

Davenport, T. et L. Prusak (1998). *Working Knowledge : How Organizations Manage what They Know*, Volume https://www.researchgate.net/publication/229099904_Working_Knowledge_How_Organizations_Manage_What_

They_Know of *EBSCO eBook Collection*. Harvard Business School Press.

van Craeynest, J.-M., J.-L. Ermine, et C. Chagnot (1997). Capitalisation des connaissances dans le cadre d'un transfert industriel. In *IC'97, Ingénierie des Connaissances*, Roscoff, France.

Grundstein, M. (1995). La capitalisation des connaissances de l'entreprise, système de production des connaissances. In *Actes du Colloque L'Entreprise Apprenante et les Sciences de la Complexité*, Aix en Provence, France. Jeanne Mallet : L'organisation Apprenante. Faire, chercher,comprendre.

Nonaka, I. et H. Takeuchi (1995). *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press.

McCarthy, J., M. Minsky, N. Rochester, et C. Shannon (1955). A proposal for the dartmouth summer research project on artificial intelligence. In *http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf*.

Summary

In 2024, the B2V Memory Observatory® took up the issue of "corporate memory" and launched concrete actions to bring this subject to the forefront of management. Prof. Jean-Gabriel Ganascia, a member of the Scientific Advisory Board, invited us to shed light on the subject, based on our industrial vision of current approaches and the contribution of artificial intelligence to KM: "How can we position Knowledge Management in this reflection on Corporate Memory? In the first part of the article, we present the pioneering, singular and exemplary approach of the Commissariat à l'Énergie Atomique et aux Énergies Alternatives. The second part looks at the arrival of the ISO30401 standard and the expected requirements for SKMs or "Knowledge Management Systems". The already significant contribution of Artificial Intelligence and the arrival of major language models are discussed, with the necessary industrial limits of caution. The author's conviction is clear: KM is a strategic key to addressing the the subject of "Corporate Memory".

A Survey of Semantic (Indoor) Trajectory Models: Towards Modelling Museum Visitors' Journey

Alaa Eddine Siouane*, Claudia Marinica*, Fabien Picarougne*, Fabrice Guillet*

*Nantes University, LS2N, Nantes 44300, France

1 Abstract

This article is part of a PhD research project, where its primary objective is to collect and analyze trajectory data within cultural spaces, specifically museums, to gain insights into visitor behavior. Museums renowned for their complex architecture and vast collections of artwork and historical artifacts, attract an increasing number of visitors worldwide. However, these spaces also present challenges, such as congestion, the presence of exhibits that may not appeal to all visitors, the complexities of organizing exhibits effectively for curators and so on. By studying visitor trajectories and behavior, this work aims to provide valuable insights to enhance the visitor experience and support curators in better managing and organizing museum spaces.

Mobility data has garnered significant interest over the years, driven by the advancements in sensor technologies and tracking systems that have made it increasingly easier to monitor the movement of objects. The movement of these objects forms trajectories, which can occur in two primary spaces: outdoor and indoor environments. While early trajectory studies focused on positional data, the growing availability of mobility datasets has highlighted the limitations of purely positional information, leading to the integration of semantic information that enriches trajectories with contextual meaning.

In this article, we delve into the field of trajectory data, with a particular emphasis on *semantic trajectories*, which combine spatial, temporal, and semantic dimensions to provide deeper insights into movement patterns. Our primary objective is to review and analyze existing semantic trajectory models developed for either outdoor or indoor environments. We aim to compare these models to uncover their strengths and limitations.

Moreover, to address the diversity and lack of standardization in terminology and definitions across the literature, we propose a benchmark framework referred to as the *pivot model*. This model encompasses the varying terminologies and conceptual definitions into an unified framework, facilitating a comprehensive and systematic comparison of semantic trajectory models. Through this work, we provide a clearer understanding of current semantic trajectory modeling and lay the foundation for advancing more robust approaches in this domain.

In the pivot model, we begin by defining a raw trajectory, which is the evolution of the position of a moving object (MO) over time. Based on this definition and considering the components (space and time), we can identify different types of trajectories. From the spatial perspective, trajectories can be geometric or symbolic, while from the temporal perspective,

they can be sequential or temporally annotated. The addition of semantic information transforms the raw trajectory into a semantic trajectory.

To enhance readability and comprehension, the semantic information can be further classified based on its usage, nature and category. To this end, we propose six classes of semantic data that can be integrated into the trajectory: (1) Temporal Semantics, (2) Spatial Semantics, (3) Moving Object Semantics, (4) External Environment Information, (5) Point of Interest and (6) Interactions. To evaluate the richness of semantic annotations in the different models, we introduce a three-level scale (low, medium, and high) enabling comparisons between different models. This framework provides a hierarchical and comprehensible representation of the semantic information added to trajectory data, offering a clearer understanding of its structure and relationships. Additionally, it serves as a benchmark and reference point for previous works in the field.

Next, we present prior research contributions in the field of semantic trajectory modelling. Various models have been proposed to represent and analyze semantic trajectory data, each employing distinct definitions, terminology, and conceptual frameworks. These different approaches reflect diverse perspectives on how semantic trajectory data can be understood and utilized. We selected these models based on their proposal of a model for mobility data that incorporates at least a minimal level of semantic aspects in their overall design.

We demonstrate the applicability of the proposed benchmark by instantiating selected models within the pivot model framework, showcasing its performance and robustness. Moreover, the Table 1 provides a comparison across models based on several criteria. First, we include the six different semantic classes identified in pivot model. Additionally, we consider the division of trajectories into stops and moves as a significant criterion to determine which models incorporate this concept. Another important criterion is the division of trajectories into subsequences, which allows for detailed analysis and processing of trajectory data. Lastly, we examine the representation of space, specifically whether the trajectory models were designed for outdoor, indoor, or both types of environments. The importance of this last criterion lies in the fact that the collection, representation and processing of trajectory data vary depending on the type of space. Data collection in indoor environments, for example, often requires dedicated sensors and techniques to track the movement of objects due to the architectural complexity, whereas GPS sensors are sufficient for outdoor environments.

Criteria / Model	Time	Space	MO	POI	Interaction	Environment	Stops and moves	Sub-sequence	Indoor or outdoor
Stops and moves Spaccapietra et al. (2008)	/	Medium	Low	Medium	/	Low	Yes	Yes	Outdoor
Alvares et al. (2007)	/	Medium	/	Medium	/	/	Yes	Yes	Outdoor
Bogorny et al. (2010)	Low	Medium	Low	Medium	/	/	Yes	Yes	Outdoor
Andrienko et al. (2011a,b)	Medium	Low	Low	Low	/	Low	Yes	Yes	Outdoor
SeMiTri Yan et al. (2011, 2010, 2013)	Low	Medium	Medium	Medium	/	/	Yes	yes	Outdoor
Spaccapietra and Parent (2011); Parent et al. (2013)	Low	Medium	Medium	Medium	/	/	Yes	Yes	Outdoor
CONSTAnT Bogorny et al. (2014)	Medium	Low	Medium	Medium	/	Medium	No	Yes	Outdoor
Ontology models Yan et al. (2008); Renso et al. (2013)	High	High	Medium	High	/	/	Yes	Yes	Outdoor
MASTER Mello et al. (2019)	Medium	Medium	High	Low	/	Medium	No	No	Outdoor
STriDE Cruz (2017)	Low	Medium	Low	Low	/	/	No	Yes	Indoor
HiOutdoor-Indoor model Model for Noureddine et al. (2020, 2021)	/	High	Low	High	/	Medium	No	yes	Indoor & Outdoor
SITM Kontarinis et al. (2021)	/	High	Medium	High	/	/	No	Yes	Indoor
MAMLSTR Cayèré et al. (2021)	Low	High	Medium	High	/	High	No	Yes	Indoor & Outdoor
TSTM-in Qin et al. (2024)	/	High	Low	High	/	/	No	Yes	Indoor

TAB. 1 – *Comparison of models based on different criteria.*

From the Table 1, one notable observation is that the *interaction criterion* is not supported by any model. However, we argue that it is essential, as it could play a significant role in understanding the trajectories of MOs, especially in the context of museums. This criterion, particularly its second aspect-focusing on the interaction between the MO and its environment-was highlighted in the work of Ceccarelli et al. (2024). Their study analyzes the trajectories of museum visitors using dedicated sensors and computer vision to identify interactions between MOs and exhibitions. The ultimate goal is to improve the understanding of visit dynamics and optimize the management of exhibition spaces based on visitor habits and preferences.

The *stops and moves* concept, introduced by Spaccapietra et al. (2008), has been a foundational benchmark for many trajectory models up to 2014. However, its application was primarily limited to outdoor environments. By its nature, the concept implies the segmentation of trajectories into meaningful sub-sequences, which is why all models based on this concept inherently support the sub-sequence criterion outlined in the Table 1. Regarding semantic classes, we observe variability in how different models address this criterion. Notably, the most recent models that incorporate the *stops and moves* concept exhibit stronger performance, with the majority of their semantic values categorized as *Medium*. This suggests an evolution in the granularity and complexity of semantic representation in newer models.

To the best of our knowledge, the first appearance of a model supporting *indoor spaces*, was in 2017 in the work of Cruz (2017). This work demonstrated strength in representing the *semantics of space*, as it was specifically dedicated to indoor environments. However, it placed less focus on time and MO semantics, with a complete absence of values for other semantic aspects. The model supports the segmentation of trajectories into subsequences, a criterion supported by all models. In contrast, Mello et al. (2019) introduced a multi-aspect model dedicated to outdoor environments, where the authors viewed trajectory segmentation as an analytical step, so their model does not address this criterion. Instead, their focus was on representing trajectories with semantic information represented as aspects, distinguishing their approach from others.

The model proposed by Kontarinis et al. (2021) does not incorporate *semantics for time*, relying instead on simple time annotations. However, it excels in indoor spaces, with a high level of semantic representation for space and point-of-interest. The MO dimension, by contrast, is represented at a low to medium level, reflecting a more limited focus in this area.

During the years 2020-2021, new models were introduced to bridge the gap between indoor and outdoor space representation. Among these, the model by Noureddine et al. (2020, 2021) stands out for introducing environmental semantics at a medium level, marking a significant improvement over previous works. This model also introduces a novel concept of representing both indoor and outdoor spaces, expanding its applicability. On the other hand, Cayèré et al. (2021) demonstrate a preference for enhancing temporal semantics, albeit at a low level, while providing medium and high levels of representation for MOs and environmental semantics, respectively. Their approach relies on the *aspect* concept first proposed by Mello et al. (2019), allowing for a more structured integration of multiple semantic dimensions. Finally, the most recent model, introduced in 2024 by Qin et al. (2024), is explicitly dedicated to indoor spaces. This model exhibits a high level of semantic representation for both spatial and MO dimensions, although its treatment of MO semantics remains at a low level, reflecting a nuanced but selective focus.

Overall, in this article, we proposed a pivot model, which is an exhaustive representation

of existing semantic trajectory models. The key advantage of this model is that it supports a wide array of models, whether dedicated to outdoor, indoor, or mixed indoor & outdoor environments. It is important to note that the pivot model does not focus on data storage or redundancy, as its primary objective is not to serve as an analysis tool but rather to establish a standardized vocabulary for translating and defining various semantic trajectory models. This common framework is crucial for facilitating cross-model comparisons and ensuring greater consistency in the field.

By reviewing and comparing the current models, we were able to position the pivot model as a unifying structure that highlights the strengths and limitations of existing approaches. This work not only provides a comprehensive summary of the state-of-the-art in semantic trajectory modeling but also sets the stage for the development of more robust models. The pivot model paves the way for integrating the advantages of each existing model, eventually leading to a more powerful and adaptable system that can be used for both analysis and real-world applications.

Ultimately, the pivot model serves as a foundation for future advancements, offering a flexible framework that can evolve with emerging technologies and new data types, since it is an extensible proposition in terms of semantic data.

References

Alvares, L. O., V. Bogorny, B. Kuijpers, J. A. F. de Macedo, B. Moelans, and A. Vaisman (2007). A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, pp. 1–8.

Andrienko, G., N. Andrienko, P. Bak, D. Keim, S. Kisilevich, and S. Wrobel (2011b). A conceptual framework and taxonomy of techniques for analyzing movement. *Journal of Visual Languages & Computing 22*(3), 213–232.

Andrienko, G., N. Andrienko, and M. Heurich (2011a). An event-based conceptual model for context-aware movement analysis. *International Journal of Geographical Information Science 25*(9), 1347–1370.

Bogorny, V., C. A. Heuser, and L. O. Alvares (2010). A conceptual data model for trajectory data mining. In *Geographic Information Science: 6th International Conference, GIScience 2010, Zurich, Switzerland, September 14-17, 2010. Proceedings 6*, pp. 1–15. Springer.

Bogorny, V., C. Renso, A. R. de Aquino, F. de Lucca Siqueira, and L. O. Alvares (2014). Constant–a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS 18*(1), 66–88.

Cayèré, C., C. Sallaberry, C. Faucher, M.-N. Bessagnet, P. Roose, M. Masson, and J. Richard (2021). Multi-level and multiple aspect semantic trajectory model: application to the tourism domain. *ISPRS International Journal of Geo-Information 10*(9), 592.

Ceccarelli, S., A. Cesta, G. Cortellessa, R. De Benedictis, F. Fracasso, L. Leopardi, L. Ligios, E. Lombardi, S. G. Malatesta, A. Oddi, et al. (2024). Evaluating visitors' experience in museum: Comparing artificial intelligence and multi-partitioned analysis. *Digital Applications in Archaeology and Cultural Heritage 33*, e00340.

Cruz, C. (2017). Semantic trajectory modeling for dynamic built environments. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 468–476. IEEE.

Kontarinis, A., K. Zeitouni, C. Marinica, D. Vodislav, and D. Kotzinos (2021). Towards a semantic indoor trajectory model: application to museum visits. *GeoInformatica 25*(2), 311–352.

Mello, R. d. S., V. Bogorny, L. O. Alvares, L. H. Z. Santana, C. A. Ferrero, A. A. Frozza, G. A. Schreiner, and C. Renso (2019). Master: A multiple aspect view on trajectories. *Transactions in GIS 23*(4), 805–822.

Noureddine, H., C. Ray, and C. Claramunt (2020). Semantic trajectory modelling in indoor and outdoor spaces. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pp. 131–136. IEEE.

Noureddine, H., C. Ray, and C. Claramunt (2021). A hierarchical indoor and outdoor model for semantic trajectories. *Transactions in GIS 26*, 214–235.

Parent, C., S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan (2013). Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR) 45*.

Qin, J., L. Wang, T. Wu, Y. Li, L. Xiang, and Y. Zhu (2024). Indoor mobility data encoding with tstm-in: A topological-semantic trajectory model. *Computers, Environment and Urban Systems 110*, 102114.

Renso, C., M. Baglioni, J. A. F. de Macedo, R. Trasarti, and M. Wachowicz (2013). How you move reveals who you are: understanding human behavior by analyzing trajectory data. *Knowledge and information systems 37*, 331–362.

Spaccapietra, S. and C. Parent (2011). Adding meaning to your steps (keynote paper). In *Conceptual Modeling–ER 2011: 30th International Conference, ER 2011, Brussels, Belgium, October 31-November 3, 2011. Proceedings 30*, pp. 13–31. Springer.

Spaccapietra, S., C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot (2008). A conceptual view on trajectories. *Data & knowledge engineering 65*(1), 126–146.

Yan, Z., D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer (2011). Semitri: a framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th international conference on extending database technology*, pp. 259–270.

Yan, Z., D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer (2013). Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology (TIST) 4*(3), 1–38.

Yan, Z., J. Macedo, C. Parent, and S. Spaccapietra (2008). Trajectory ontologies and queries. *Transactions in GIS 12*, 75–91.

Yan, Z., C. Parent, S. Spaccapietra, and D. Chakraborty (2010). A hybrid model and computing platform for spatio-semantic trajectories. In *Extended Semantic Web Conference*, pp. 60–75. Springer.

Estimation automatique de caractéristiques acoustiques pour l'étude diachronique du français oral dans les médias

Simon Devauchelle [*,**], David Doukhan[**],
Lucas Ondel-Yang[*], Benjamin Élie[***], Albert Rilliard [*]

[*] Université Paris Saclay, CNRS, LISN, France.
[**] Institut national de l'audiovisuel (INA), France.
[***] The University of Edinburgh, UK.

1 Introduction

La description des évolutions diachroniques des langues orales nécessite de relever un certain nombre de défis méthodologiques : obtenir des données suffisamment représentatives des populations en contrôlant des facteurs explicatifs tels que l'âge, le genre, les années et les contextes d'élocution. Les rares corpus collectés à des fins de recherche contiennent plusieurs limitations : périodes d'enregistrement et nombre de locuteur·rice·s limités, absence de certaines catégories de personnes (Pemberton et al., 1998; Berg et al., 2016). Les analyses de ces corpus ont conclu que la hauteur de la voix des femmes avait baissé au cours du XX^e siècle : un résultat relayé par un grand nombre d'études académiques ainsi que par la presse grand public et souvent présenté comme une évidence (Ellis et al., 2023; Robson, 2018). En l'absence de comparaisons avec des mesures réalisées à partir de voix d'hommes, il devient compliqué de comprendre si ces tendances sont exclusives à un genre en particulier ou non.

Si les archives TV et radio peuvent permettre de réduire les biais d'échantillonnage liés aux faibles quantités de données des études précédentes, leur exploitation nécessite de mettre en place des méthodologies d'analyses adaptées aux particularités de ce matériau, pour lequel le contenu lexical n'est pas contrôlé, les prises de parole et l'identité des personnes ne sont pas connues a priori, le signal de parole peut être mélangé à d'autres sources tels que les bruits d'environnement et la musique de fond, susceptibles de perturber les méthodes de description automatique de la voix. Cette communication vise à détailler les méthodes proposées pour analyser la parole dans les fonds TV et radio dans le but de décrire l'évolution des voix des années 1950 à nos jours : qu'il s'agisse de la hauteur de la voix estimée à l'aide de la fréquence fondamentale (f_0) et de la longueur du conduit vocal (VTL), ou encore des paramètres articulatoires (protrusion des lèvres, hauteur du larynx), et tenter de confirmer ou d'infirmer l'hypothèse selon laquelle la voix des femmes a pu baisser au cours du XX^e siècle.

2 Estimation automatique des caratéristiques acoustiques

Le corpus se compose de 111 heures de parole, obtenues par une sélection de 1028 locuteur·rice·s équilibrée en genre et en termes de catégories d'âge (20 - 35, 36 - 50, 51 - 65, < 65

ans), identifiés dans les archives de l'Institut national de l'audiovisuel à l'aide d'une chaîne de traitements semi-automatiques développée par Uro et al. (2022), et d'une aide documentaliste indispensable pour la présélection des documents.

Les segments de paroles ont été transcrits automatiquement en utilisant Whisper (Radford et al. (2022)) puis alignés phonétiquement en utilisant MFA (McAuliffe et al. (2017)). L'évaluation de cette approche sur nos données est associée à un score WER de 12.8% et une précision de 96.9% sur les voyelles (Elie et al. (2024)), des résultats jugés suffisants pour réaliser des études automatisées. Les analyses portent sur les trames centrales de plus d'un million de voyelles en prenant soin d'exclure les phones non voisés grâce à un processus de filtrage décrit par Rilliard et al. (2023). Les quatre premiers formants (F_1, F_2, F_3, F_4) ont été estimés avec l'algorithme de Burg implémenté dans Pratt (par Boersma et Weenink (2024)), en appliquant la méthode d'optimisation proposée par Escudero et al. (2009). À partir de ces formants et des équations proposées par Lammert et Narayanan (2015), des estimations de la longueur du conduit vocal ont également été analysées. Des régressions linéaires mixtes ont ensuite été apprises sur les différentes mesures acoustiques. Cette approche automatique a également donné lieu à une autre étude sur l'évolution des configurations articulatoires en utilisant une technique d'inversion détaillée par Elie et al. (2024), s'intéressant quant à elle à l'évolution de la hauteur du larynx et de la protrusion des lèvres.

3 Résultats et discussions

Un allongement du conduit vocal en fonction du temps est capturé par le modèle – c.-à-d. une baisse de la hauteur de la voix – mais dans le cadre de notre étude, cette observation se constate autant pour les hommes que pour les femmes. On observe également des valeurs de F_1 plus importantes pour les voyelles ouvertes dans les années 1950 comparées aux périodes plus récentes, toujours indépendamment du genre. Au niveau de la f_0, on remarque une baisse des valeurs chez les femmes en fonction de l'âge et le phénomène contraire chez les hommes. Au niveau articulatoire, on observe au cours du temps une baisse du larynx et une faible tendance vers plus de protrusion.

Les résultats de notre analyse acoustique réalisée sur de la parole issue des archives audiovisuelles ne soutiennent pas la thèse d'un changement de la hauteur de la voix au cours du temps qui serait seulement spécifique au genre féminin – comme soutenu par Pemberton et al. (1998) –, mais avancent plutôt l'hypothèse d'une baisse générale de la hauteur de la voix. Néanmoins, notre corpus est principalement constitué d'interviews, provenant de *talk shows* – avec un style de parole propre qui peut donner lieu à des variations acoustiques de la parole (voir Hollien et al. (1997)) – limitant intrinsèquement nos conclusions à ce type de matériel. Par exemple, la baisse observée des valeurs des F_1 au cours du temps peut typiquement s'expliquer par un effort vocal moins important. La cause de cette tendance pourrait en partie résider dans l'évolution des pratiques technologiques médiatiques, comme celle de la distance des microphones aux locuteur·rice·s, plus importante dans les archives plus anciennes.

Pour mieux saisir la subtilité des trajectoires phonétiques et améliorer l'interprétation des changements diachroniques, nous cherchons à nous affranchir du paradigme d'analyse fréquentiste (Candea et al. (2013)). La conclusion de la présentation portera sur les travaux en cours qui privilégient désormais une modélisation acoustique probabiliste de la parole, permettant de mieux isoler les différentes sources potentielles de variation phonétique.

Références

Berg, M., M. Fuchs, K. Wirkner, M. Loeffler, C. Engel, et T. Berger (2016). The speaking voice in the general population : Normative data and associations to sociodemographic and lifestyle factors. *Journal of Voice 31*.

Boersma, P. et D. Weenink (2024). Praat : doing phonetics by computer [computer program]. version 6.4.23. Technical report.

Candea, M., M. Adda-Decker, et L. Lamel (2013). Recent evolution of non-standard consonantal variants in french broadcast news. pp. 412–416.

Elie, B., D. Doukhan, R. Uro, L. Ondel-Yang, A. Rilliard, et S. Devauchelle (2024). Articulatory configurations across genders and periods in french radio and tv archives.

Ellis, S., S. Goetze, et H. Christensen (2023). Moving towards non-binary gender identification via analysis of system errors in binary gender classification.

Escudero, P., P. Boersma, A. S. Rauber, et R. A. H. Bion (2009). A cross-dialect acoustic description of vowels : Brazilian and european portuguese. *The Journal of the Acoustical Society of America 126*(3), 1379–1393.

Hollien, H., P. A. Hollien, et G. de Jong (1997). Effects of three parameters on speaking fundamental frequency. *Journal of the Acoustical Society of America 102*(5), 2984–2992.

Lammert, A. C. et S. S. Narayanan (2015). On short-time estimation of vocal tract length from formant frequencies. *PLOS ONE 10*, 1–23.

McAuliffe, M., M. Socolof, S. Mihuc, M. Wagner, et M. Sonderegger (2017). Montreal forced aligner : Trainable text-speech alignment using kaldi. In *Interspeech 2017*, pp. 498–502.

Pemberton, C., P. McCormack, et A. Russell (1998). Have women's voices lowered across time ? a cross sectional study of australian women's voices. *Journal of Voice 12*(2), 208–213.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, et I. Sutskever (2022). Robust speech recognition via large-scale weak supervision.

Rilliard, A., D. Doukhan, R. Uro, et S. Devauchelle (2023). Evolution of voices in french audiovisual media across genders and age in a diachronic perspective. In *20th International Congress of Phonetic Sciences (ICPhS)*.

Robson, D. (2018). The reasons why women's voices are deeper today. bbc. *BBC*.

Uro, R., D. Doukhan, A. Rilliard, L. Larcher, A.-C. Adgharouamane, M. Tahon, et A. Laurent (2022). A semi-automatic approach to create large gender- and age-balanced speaker corpora : Usefulness of speaker diarization & identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, pp. 3271–3280.

Summary

This communication details the implementation of an automatic pipeline to extract acoustic cues from audiovisual speech excerpts in order to describe the diachronic evolution of spoken French. Acoustic results and difficulties inherent to archive data will be discussed.