Groupe de travail sur les technologies de la langue française

Juillet 2023

L'ÉQUIPE

Ce groupe de travail a été co-animé par :
- Thibault Grouas
- Thierry Bultingaire

Avec le soutien de Fadila Taleb Larabi pour la consultation publique

POINT DE DÉPART

On assiste aujourd'hui à une expansion des domaines liés à l'intelligence artificielle dont les technologies de la langue (*la langue sous toutes ses formes : écrite, orale, signée*). L'Europe d'une manière générale et la France en particulier, ont la volonté et l'ambition de s'imposer dans ce secteur, de faire face aux géants américains et chinois, et de développer la souveraineté numérique. Dans cette optique, l'idée d'un centre de référence français et européen sur les technologies de la langue a vu le jour. Ce projet serait implanté au sein de la cité internationale de la langue française, à Villers-Cotterêts.

PROBLÉMATIQUE

Recueillir les avis et les besoins du grand public et des professionnels sur le futur centre de référence des technologies de la langue. Il s'agit globalement de répondre à cette question : *quelles sont les attentes du public et des professionnels pour un tel type de centre : quels sont les services, les évènements, les activités attendues ?*

MÉTHODE ET CALENDRIER

❑ **Méthode de travail :** des consultations/entretiens à l'échelle nationale et européenne avec différents acteurs institutionnels (ministères), industriels et académiques spécialisés dans le domaine des technologies de la langue. Il nous paraissait pertinent d'appuyer ces démarches par la stratégie numérique culturelle portée par le SNUM, qui s'inscrit donc en complément des consultations menées par la DGLFLF.

❑ **Acteurs sollicités :** à l'échelle nationale, des entreprises comme *Orange, Airbus, EDF, ATOS, Doctrine, Zaoin, LightOn, Deezer* … ; des associations et autres organismes de recherche : *CNRS, ATALA, APIL, LISN, Voice Lab*, …Sur le plan européen, plusieurs rencontres et entretiens ont eu lieu en début 2023, avec des représentants des gouvernements dans différents états membres (EM) : *Croatie, Espagne, Grèce, Hongrie , Malte, Pologne, République Tchèque, Suède, Slovaquie, Slovénie*... En outre, un échange fréquent a lieu avec la commission européenne dans le cadre de la mise en place du projet ALT-EDIC.

❑ **Calendrier suivi :** la concertation **au niveau national** a débuté mi-2022 et s'est accélérée début 2023. Une réunion avec tous les représentants identifiés à ce stade a eu lieu le 26 mai. Au final, plus d'une vingtaine d'entretiens et une quinzaine de réunions collectives ont été menés. La concertation **au niveau européen** a débuté en décembre 2022 et est toujours en cours. Une vingtaine d'États membres ont été rencontrés dans ce cadre. Un groupe de travail formel a été mis en place par la Commission européenne en avril 2023. Il s'est réuni 6 fois et est coordonné par le ministère français de la Culture. Il compte une cinquantaine de participants. La **consultation publique** a été menée en parallèle de ces activités de février à avril 2023.

Entretiens *Ateliers* *Consultation publique*

OBJECTIFS INITIAUX DU PROJET

Principaux objectifs	Démarches et pistes suivies
Réflexions préliminaires sur le projet du centre de Villers-Cotterêts et préparation de sa candidature dans le cadre de France 2030.	Constitution d'un pre-consortium autour du porteur du projet, composé d'experts scientifiques (associations et organismes de recherche).
Besoin de lettres de soutien pour la candidature France 2030.	Sollicitations de plusieurs entreprises françaises et organismes de recherche.
Faire adhérer des groupes industriels français au projet du centre de référence français et européen sur technologies de la langue.	Des prises de contact menées par la DGLFLF (*délégation générale à la langue française et aux langues de France*) auprès d'entreprises concernées de près ou de loin par le sujet des technologies de la langue et de l'IA; Ces prises de contact ont été suivies par des rencontres bilatérales ayant permis une réflexion et des échanges pertinents autour du projet.
Recueillir l'avis du grand public et d'un large panel des professionnels des technologies de la langue	Mener une consultation publique dans le cadre de la feuille de route du numérique au sein du ministère de la Culture.
Proposition d'une infrastructure européenne numérique à vocation industrielle (EDIC) centrée sur les technologies de la langue et la nécessité de faire adhérer des états membres de l'union européenne.	Rencontres bilatérales avec plusieurs états membres de l'union européenne et coordination d'un groupe de travail européen avec des réunions bimensuelles entre mai et juillet 2023.
Rédaction des statuts pour le projet européen d'Alliance pour les technologies des langues ALT-EDIC.	Réunions bimensuelles avec le GT européen; Appui par un prestataire extérieur spécialisé (Welcomeurope); Échanges réguliers avec la commission européenne.
Communication sur le projet national et européen (*centre de référence sur les technologies de la langue et ALT-EDIC*)	Réunions avec des acteurs industriels et académiques spécialistes du domaine des technologies de la langue; Participation à des conférences / évènements scientifiques (Big Data Value Association, Metaforum, Data Week, NLP MeetUp, journée Demotal…)

OBJECTIFS INITIAUX DU PROJET (vue détaillée)

❑ Sur le plan national

Dans le cadre de la préparation de la candidature du centre de Villers-Cotterêts pour France 2030, un pre-consortium, composé de plusieurs experts du domaine des technologies de la langue, ainsi que divers acteurs culturels et industriels, a été constitué (*CCRV, APIL, ATALA, CMN, ELDA, LISN, Voice Lab,* …) . L'objectif de ce groupe de travail était de se pencher sur le projet de ce centre de référence des technologies de la langue afin de tracer ses grandes lignes : *déterminer les principales actions, les types d'acteurs impliqués, les questions financières, les orientations scientifiques et éthiques du centre*, etc. En parallèle à ce groupe de travail, plusieurs entreprises françaises (*Orange, EDF, AirBus, Renault, Plaine Images, Aday, LightOn,* …etc.) ont répondu favorablement aux sollicitations de la DGLFLF pour des lettres de soutien.

Après la soumission de la candidature du projet pour France 2030, des prises de contact avec d'autres acteurs industriels ont été effectuées par la DGLFLF pour faire connaitre le projet du centre de référence des technologies de la langue et faire adhérer des nouveaux acteurs. Ces prises de contact ont donné lieu à plusieurs rencontres bilatérales. Parmi ces nouveaux acteurs rencontrés et qui ont manifesté un grand intérêt au projet et contribué à alimenter les réflexions sur le sujet des technologies de la langue, nous pouvons citer les entreprises suivantes : *Doctrine, Zaion, Atos, Compilatio, Thales, Lettria, Systran, Teklia, …*

Ces différents acteurs nationaux ont été réunis en mai dernier pour un point d'information général sur l'avancement du projet du centre de référence de Villers-Cotterêts, à la fois sur la perspective nationale mais aussi sur son orientation européenne (ALT-EDIC). Cette réunion a permis d'une part, de confirmer les besoins et les attentes exprimés par les professionnels à travers la consultation publique (*mise à disposition de données, de formations, lieu d'échanges, …)* ; et d'autre part, de faire émerger d'autres besoins comme la mise à disposition d'un service de communication permettant aux entreprises de gagner en visibilité et faire connaitre leurs activités ou encore la nécessité d'une aide juridique, à travers une documentation clé en main par exemple, pour accompagner les entreprises sur les questions légales et éthiques touchant aux domaines des technologies de la langue et de l'IA.

❑ Sur le plan européen

En début d'année 2023, la DGLFLF est allée à la rencontre de plusieurs états membres de l'Union Européenne (*Croatie, Espagne, Grèce, Hongrie , Malte, …*). L'objectif de ces entretiens consistait à partager l'idée d'un EDIC centré sur les technologies de la langue (ALT-EDIC) et de recueillir les avis et les soutiens de ces EM. Ces rencontres bilatérales ont été fructueuses et ont permis de faire adhérer 23 EM au projet ALT-EDIC dont 11 se sont déclarés comme membres de l'ALT-EDIC et 7 comme observateurs, au moment de la pré-soumission du projet, fin mai dernier.

En parallèle à ces rencontres avec les EM, des échanges réguliers ont lieu avec la Commission européenne pour la préparation de la candidature de l'ALT-EDIC et dont la France s'est déclarée comme coordinateur du groupe de travail européen.

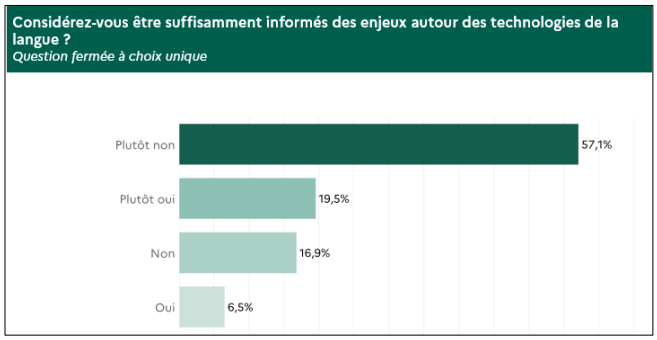
Dans cette optique, entre mai et juillet 2023, la France a animé des réunions bimensuelles avec un groupe de travail constitué de près de 50 personnes (des représentants des EM, des membres de la CE, …). Ces réunions ont permis de définir le programme de l'ALT-EDIC, son schéma de gouvernance et son budget.

CONSULTATION PUBLIQUE

Objectif du questionnaire

- ❑ L'objectif est de récolter l'avis du grand public et des professionnels sur l'ouverture prochaine d'un centre de référence des technologies de la langue;
- ❑ L'enquête a été menée sous la forme de deux questionnaires distincts (composés de questions fermées et ouvertes) : l'un à destination des potentiels visiteurs du futur centre, l'autre à destination des professionnels,
- ❑ Au total, il y a eu **467** participants dont **344** présentés comme potentiels visiteurs du centre et **123** comme professionnels des technologies de la langue,

Résultats clés : questionnaire 1 « potentiels visiteurs du centre »



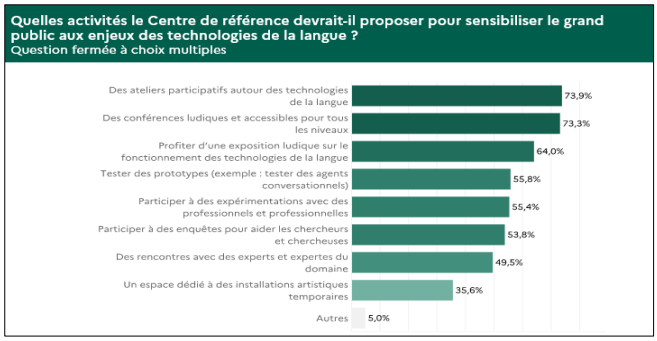
Commentaire :

- Plus de la moitié (74%) des participants (grand public) ne sont pas informés sur les enjeux autour des TL,
- Il y a donc un manque à combler sur ce terrain. Cela met en valeur l'apport du centre auprès du grand public : un centre d'information sur les enjeux des technologies de la langue (à travers des *compagnes de sensibilisation, ateliers et formations sur les divers sujets touchant les TL, etc.*)

Commentaire :
Deux groupes d'activités ressortent du graphique :

- Ateliers participatifs et ludiques
- Recherche implicative et participative

La participation du public, l'animation d'activités ludiques (médiation/vulgarisation) devraient être une partie intégrante du centre.



Pensez-vous à d'autres activités qui pourraient vous inciter à y passer du temps ?
Question ouverte

Exemples de verbatims

Activités créatives, participatives et ludiques

« Ateliers de créations pour adulte »

« des ateliers courts (une ou deux heures) m'initiant aux rudiments d'une langue régionale grâce a un robot conversationnel »

« ateliers permettant une appropriation active de certaines technologies »

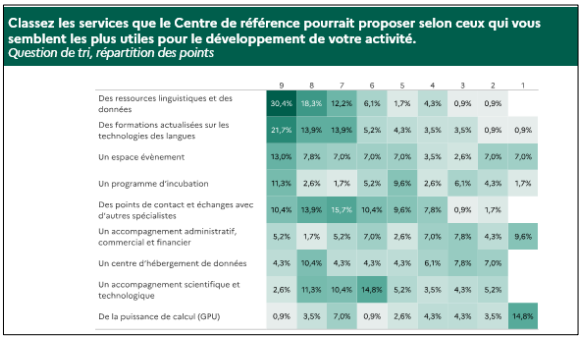
« Apprendre à utiliser des outils numériques qui font appel aux technologies du langage (sessions de formation sur inscription ou libres). »

« Des jeux : activités interactives et ludiques (exemple jeu d'évasion) »

« Parcours ludiques thématiques (par exemple : enquête pour enfant comme adulte) »

CONSULTATION PUBLIQUE

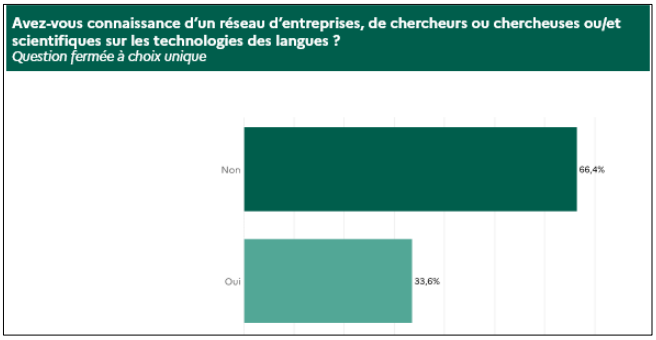
Résultats clés : questionnaire 2 « professionnels des TL »



Commentaire :
Le classement établi dans ce graphique donne une idée sur les services prioritaires à mettre à disposition des professionnels :

- Mise à disposition de ressources linguistiques
- Programme de formations pour viser l'excellence dans le domaine
- Disposer d'un lieu fédérateur pour la communauté des TL

Commentaire :
Ce graphique montre tout l'intérêt et les besoins autour d'un centre de référence des technologies de la langue dont l'un des objectifs serait un lieu de rencontre et d'échange pour les différents acteurs (chercheurs, entreprises, grand public), spécialistes des technologies de la langue.



Exemples de verbatims

Un centre de ressources linguistiques

« Une plateforme d'hébergement des archives et du patrimoine linguistique pour les acteurs qui effectuent la collecte »
« Des leviers permettant d'accéder à certaines ressources linguistiques, inaccessibles pour des petites entreprises ou des indépendants, par exemple certains corpus de la BnF »
« Des données statistiques sur les besoins »

Un espace d'échanges et de collaborations

« Un espace d'échanges entre scientifiques et acteurs de terrain »
« De point d'échange entre structures au niveau départemental ou régional/par taille de structure »
« Un lieu de rencontre pour organiser des conférences scientifiques sur le thème de la langue française »
« Opportunités de sensibilisation du grand public aux enjeux du TAL, p.ex. en participant à des conférences »

Conclusion

Les résultats de la consultation publique confirment la nécessité d'un centre de référence des technologies de la langue. Certaines attentes du grand public et des professionnels ayant participé vont dans le sens des objectifs et des actions que se fixe le centre Villers-Cotterêts. D'autres besoins exprimés vont servir à prioriser les activités et le plan d'action du futur centre.

ENSEIGNEMENTS DE LA CONSULTATION SUR LES BESOINS DES FUTURS UTILISATEURS DU CENTRE DE VILLERS-COTTERETS

Principaux besoins exprimés	Solutions et opportunités identifiées
Disposer de grandes capacités de calculs pour adapter finement et développer des grands modèles de langue	Offrir un accès privilégié au supercalculateur Jean-Zay et d'autres partenaires européens dans le cadre du centre français et européen ALT-EDIC
Rendre accessible des grands volumes de données et favoriser leur partage entre les différents acteurs impliqués dans le développement des technologies de la langue	Concevoir une plateforme de stockage sécurisée et pérenne et accompagner et inciter les partenaires industriels à partager leurs données via le centre ALT-EDIC
Besoin de support sur les questions juridiques, éthiques liées aux technologies des langues et aux traitement de données.	Offrir un accompagnement aux entreprises incubées ou non (recrutement d'un juriste spécialisé, mise à disposition de documentations..). Mise en place dans le projet ALT-EDIC d'un nouveau comité consultatif sur les questions juridiques et éthiques.
Soutenir et favoriser la création de nouvelles entreprises dans le domaine des technologies de la langue et de l'IA : mise à disposition de locaux, de bureaux, d'espaces de réunion et de présentation, accès à des services support (secrétariat, administratif, financier…)	Mise en place d'un incubateur de jeunes pousses (*start-ups*) implanté au sein du centre de Villers-Cotterêts, avec des surfaces dédiées à leur hébergement (bureaux, salles de réunion, auditorium). Recrutement au sein de l'équipe de chargés d'affaires et d'ingénieurs financiers. Mise en place de services support pour les jeunes pousses.
Faire connaitre les activités de certaines TPE et PME auprès du grand public	Mise en place d'un service de communication/mercatique pour accompagner les TPE et les PME dans la diffusion de leurs activités
Mettre en relation le tissu industriel avec les acteurs du monde académique pour favoriser la recherche et développement autour des technologies de la langue	Le centre de Villers-Cotterêts a vocation à être le lieu de rencontre et d'échange entre les différents acteurs académiques et industriels concernés de près ou de loin par les technologies de la langue dans l'objectif de favoriser l'innovation
Rapprocher le grand public du monde industriel et académique	Mise en place d'un service de médiation et vulgarisation dédié à l'élaboration d'activités autour des technologies de la langue et de l'IA accessible aux néophytes

PASSAGE À L'ACTION

ACTION 1. **Projet national** : conforter l'assise budgétaire du projet au niveau ministériel et avec le MEFSIN. Mettre en place des lignes budgétaires pérennes en complément des dispositifs d'amorçage identifiés (Appel France 2030 Pôles territoriaux d'innovation, dispositif Cluster IA…)

ACTION DÉJÀ INITIÉE

ACTION 2. **Projet européen** : déterminer la contribution financière directe du ministère de la Culture (complétée au niveau français par le MEFSIN). Le minimum requis est de 275 000 euros par an. Déterminer la contribution en nature du ministère : jeux de données, personnel affecté au projet, …

ACTION DÉJÀ INITIÉE

ACTION 3. **Mobiliser les acteurs publics culturels** autour du sujet traitement des langues et les associer à ALT-EDIC (notamment ses comités scientifiques, juridiques et éthiques de l'IA).

ACTION À INITIER

ACTION 4. Structurer avec les opérateurs culturels **un jeu de données directement exploitable pour des besoins de traitement des langues** et notamment de conception de modèles de fondation et de grands modèles de langue, avec pour objectif de verser ce jeu de données « culture» à ALT-EDIC et de le valoriser.

ACTION À INITIER

ACTION 5. **Faire évoluer le dispositif de dépôt légal français** opéré par la BNF (contenus sur internet) et l'INA (contenus audiovisuels) pour permettre l'utilisation et l'exploitation des données collectées par ALT-EDIC dans le cadre de la recherche et de la conception de modèles de fondation ou grands modèles de langue aux niveaux national et européen exclusivement.

ACTION À FAIRE ARBITRER

SUITES DE LA DÉMARCHE

- *Faire valider un budget ministériel dédié pour les actions 1 et 2 d'ici la fin de l'été 2023 et organiser rapidement une prise de décision au niveau interministériel pour bénéficier d'autres soutiens ;*

- *Mettre en place l'établissement ALT-EDIC dès la fin 2023 et demander la reconnaissance de cet établissement en tant qu'organisme international (décret en conseil d'État) ;*

- *Préparer le site cible (Villers-Cotterêts), en lien avec le CMN, pour l'accueil des premières équipes d'ici fin 2024.*