Organisation des données dans un observatoire multilingue en langues écrites et signées.[1]

Bruno Mascret[1], Cédric Moreau
[1]Laboratoire d'InfoRmatique en Image et Système d'information (LIRIS)
UMR 5205
Équipe Silex (Supporting Interactions and Learning by Experience)
Université Claude Bernard Lyon 1
bruno.mascret@liris.cnrs.fr

[2]Institut National Supérieur de Formation et de Recherche pour

l'Education des Jeunes Handicapés et les Enseignements Adaptés (INS HEA)

58, 60 av des Landes 92150 Suresnes.

[2]UMR 7023, Université Paris 8

2 rue de la Liberté

93200 Saint-Denis France

cedric.moreau@inshea.fr

Résumé : En France, l'Académie française a pour mission de défendre la langue française et d'en faire un patrimoine commun à tous les Français et à tous ceux qui pratiquent cette langue. Jusqu'à ce jour la langue des signes française n'a pas bénéficié d'un tel appui.

Aujourd'hui pour faire face à cette situation, un outil de référence a vu le jour avec le soutien du Ministère de l'Éducation nationale et de la Délégation générale à la langue française et aux langues de France. Le projet « OCELLES » est un site internet collaboratif bilingue français / LSF qui a pour vocation de proposer, au minimum, pour un concept donné : sa définition et ses signifiants (dans les deux langues), dans tous les champs de la connaissance. Tous les contenus déposés sur ce site (texte, image, vidéo, présentation, …) par les utilisateurs sont ensuite validés par des experts tant sur le fond que sur la forme ou refusés avec justification.

Compte tenu des disparités géographiques ou sociologiques, pour un même concept, plusieurs signes peuvent être proposés par les utilisateurs et retenus par les experts. Le projet n'a pas vocation à choisir le « signe idéal » mais à soumettre l'ensemble des propositions aux utilisateurs identifiés qui indiquent s'ils les ont ou non rencontrés et dans quel(s) contexte(s). Une série d'informations est ainsi recueillie pour chaque signe et peut être croisée avec les profils des utilisateurs. Plateforme d'échanges, ce site sert ainsi d'observatoire linguistique.

L'une des principales problématiques de l'organisation des données a été d'arriver à concilier les différents points de vue et usages des utilisateurs, L'objectif de cette plateforme n'étant pas d'offrir un simple dictionnaire mais un réseau d'ontologies. L'autre problématique réside dans l'impossibilité de rester sur un modèle figé d'organisation, le site étant amené à évoluer et à s'enrichir à la fois par l'ajout de nouveaux concepts, mais aussi par la prise en compte de nouveaux descripteurs ou de nouvelles fonctionnalités (illustration, homonymies, antonymies, etc.)

Après une brève description des objectifs de la plateforme, nous présentons dans cet article cette organisation particulière de données, qui permet par exemple l'utilisation de plusieurs classifications. Nous

1 Avec le soutien du Ministère de l'Éducation nationale et de la Délégation générale à la langue française et aux langues de France

illustrons l'intérêt de cette approche en critiquant l'utilisation de la classification de Dewey, que nous avons dans un premier temps implémentée malgré ses limites (accès difficile à un concept précis, organisation peu intuitive, concepts récents ou spécifiques à la LSF non référençables, etc.). Nous proposons ensuite de la remplacer par des classifications réalisées directement par les utilisateurs, motivées par leurs attentes et leurs besoins. L'arborescence se construit alors au fur à mesure sous le contrôle des experts de chaque domaine.

Chaque contenu est ainsi accompagné de descripteurs et de classificateurs qui lui permettent de jouer différents rôles en fonction du contexte dans lequel il se trouve. Ainsi un contenu peut à la fois être un concept, un thème ou un sous-thème de classification, ou encore une illustration : c'est le contexte d'utilisation qui mobilise les contenus dont il a besoin en fonction de leurs descripteurs et de leurs classificateurs.

Nous terminons l'article en mettant en évidence la grande capacité d'évolution de la plateforme, en présentant nos travaux actuels pour l'intégration de recherches directes en LSF via les descripteurs de configuration spatiale. Nous montrons également que cette organisation des données permet un portage facile vers les langues des signes d'autres pays.

Mots clés : Langue des signes française, LSF, langues écrites, langues signées, expressions lexicales, classification dynamique, sourd, site internet collaboratif, concept, ontologie, expert.

I. PREAMBULE

En France, selon le rapport Gillot [1], 80% des sourds profonds sont illettrés et seuls 5% d'entre eux accèdent à l'enseignement supérieur. Dalle 2003 [2] mentionne également que « l'illettrisme ou la maîtrise insuffisante du français écrit, l'absence de diplôme ou de qualification et enfin les difficultés de communication ont de fortes répercussions sur l'insertion sociale et professionnelle des adultes sourds ».

La loi française du 11 février 2005 impose à chaque école d'accueillir tous les enfants relevant de son secteur de recrutement, et décide de rendre accessible tous les sites Internet publics. Dans ce contexte, de nouveaux besoins apparaissent et la création d'un diplôme d'état, prévue pour 2010 pour les enseignants de LSF, accélère la demande de documents pédagogiques bilingues adaptés.

Comme le Français, la Langue des Signes Française (LSF) compte des disparités régionales de vocabulaire et s'enrichit spontanément de termes inédits, à l'initiative des locuteurs qui créent des signes pour nommer des concepts nouveaux et/ou spécifiques à un champ de connaissances. Ce phénomène s'accroit au fur à mesure que les sourds accèdent aux milieux étudiants ou professionnels qui leur étaient autrefois fermés. Comme le souligne Belfais-Duquesne [3], chaque concept, une fois nommé, peut servir à définir des propriétés caractéristiques d'un autre concept, plus abstrait, et participer à la construction d'un réseau de connaissances. La nominalisation des propriétés caractéristiques du concept, permet à ce dernier de changer de statut pour d'implicite devenir explicite et participer à la construction d'un langage rigoureux.

II. LE PROJET « OCELLES »

Il n'existe à ce jour aucun outil en France ou à l'étranger comparable à « OCELLES ».

Une grande majorité des sites internet existants ne proposent qu'un lexique de signes sans leur définition et aucun n'offre une ouverture sur les concepts associés et leur réseau d'ontologies.

Peu de sites proposent un fonctionnement collaboratif mais aucun ne se positionne en tant qu'observatoire de référence des signes existants ou émergents.

C'est dans ce contexte que l'Institut National Supérieur de la formation et de la recherche pour l'éducation des jeunes Handicapés et les Enseignements Adaptés (INS HEA), sous l'égide et avec l'aide de la Délégation générale à la langue française et aux langues de France (DGLFLF) et du ministère de l'Éducation nationale, propose le projet « OCELLES », dont l'acronyme se décline en « Observatoire des Concepts et Expressions Lexicales en Langues Ecrites et Signées ».

Ce projet repose essentiellement sur un site internet collaboratif bilingue français/LSF dont la vocation est de proposer au minimum, pour un concept donné sa définition et ses signifiants (dans les deux langues), dans tous les champs de la connaissance. Tous les contenus (texte, image, vidéo, présentation, …) déposés sur ce site par les utilisateurs sont ensuite validés par des experts tant sur le fond que sur la forme ou refusés avec justification.

A. Gestion des utilisateurs et des droits

Le site sous licence GPL [4] est ouvert à tous les d'internautes avec pour chacun la possibilité d'accéder à un statut différent :

- **les visiteurs** accèdent au site et à son contenu public.
- **les utilisateurs**, issus du rang de visiteur accèdent à ce statut en s'enregistrant puis en répondant au questionnaire qui permet d'établir leur profil. Les

indications apportées sur leur parcours scolaire et professionnel, la ou les langues utilisées par chacun d'eux au sein de leur cellule familiale, sociale ou associative serviront ultérieurement de facteurs pondérateurs à leurs réponses aux questions concernant les signes. Les utilisateurs accèdent automatiquement au rang de rédacteurs sur simple demande et pourront également, à terme, choisir parmi les interfaces disponibles en fonction de leurs besoins spécifiques.

- **les rédacteurs** issus du rang d'utilisateur, après avoir accepté les conditions de publication du portail ont la possibilité de proposer aux experts des contenus et de nouveaux concepts assortis ou non d'une classification. Ils peuvent également ajouter une vidéo d'interprétation à une zone de texte dont ils ne sont pas les auteurs et réciproquement.

- **les experts**, issus du rang de rédacteur décident de la validité des contenus déposés par les rédacteurs. Dans le cas où la proposition ne convient pas, une explication est systématiquement apportée au rédacteur. Le site utilise les compétences d'experts de chaque champ conceptuel proposé, mais également d'experts de la langue des contenus déposés.

- **l'administrateur**, est expert sur l'ensemble des domaines du site, mais son rôle essentiel est de gérer le portail sans participer aux expertises ou aux publications.

Il est important de préciser que les statuts de rédacteurs et d'experts ne sont pas attribués pour l'ensemble des contenus, mais limités au domaine de compétence de l'utilisateur : on peut être expert en mathématique et en physique, rédacteur en philosophie et simple utilisateur dans les autres domaines.

Cette gestion fine des droits d'utilisation devrait permettre un usage plus démocratique et communautaire du site, la hiérarchie dépendant uniquement du thème abordé.

B. Les concepts

Une page dynamique propose pour un concept donné, sa définition, une traduction en langue écrite (le français par exemple) accompagnée d'un ou plusieurs signes dans la langue signée (en LSF par exemple).

Compte-tenu des disparités géographiques ou sociologiques, plusieurs signes peuvent être proposés pour un même concept. Le projet n'a pas vocation à choisir le « signe idéal » mais à soumettre l'ensemble des propositions aux utilisateurs.

La rédaction des définitions répond à certaines règles d'écriture.

Une définition en langue écrite :

- doit donner le sens du concept et ses propriétés essentielles.
- doit être précise.
- doit commencer par un mot ou groupe de mots, de sens plus général que le mot à définir.
- doit pouvoir remplacer le mot inconnu, dans une phrase.
- ne doit pas intégrer d'autres définitions.
- ne doit pas contenir un mot de la famille du terme défini.
- ne doit pas être circulaire.

Les illustrations, exemples, remarques, commentaires (pédagogiques, linguistiques, …) nuances, usages régionaux, connotations, … ne font pas partie de la définition mais viennent en complément.

Une définition signée :

- ne doit contenir aucun code ou appellation régionale. Les signes locaux ont toutefois leur place pour désigner un concept.
- doit éviter d'utiliser des néologismes.
- doit être ponctuée afin d'en permettre un séquençage.

Le débit de signation est adapté à un public d'apprenants, entendants ou sourds.

L'objectif de cette plateforme n'est pas d'offrir un simple dictionnaire mais un réseau d'ontologies : des liens associatifs entre concepts existent (exemple le concept « thèse », renvoie à « argumentation »…) sous les formes « voir aussi », concepts proches, opposés, contraires. Les capacités associatives d'« OCELLES » ne sont pas figées et d'autres types d'associations peuvent facilement être pris en compte.

Fig.1 : Copie d'écran d'une « page concept ».

C. Les signes

Chaque proposition de signe dispose de sa propre page, qui permet d'ajouter des remarques épistémologiques, linguistiques, des exemples en contexte et d'autres descripteurs.

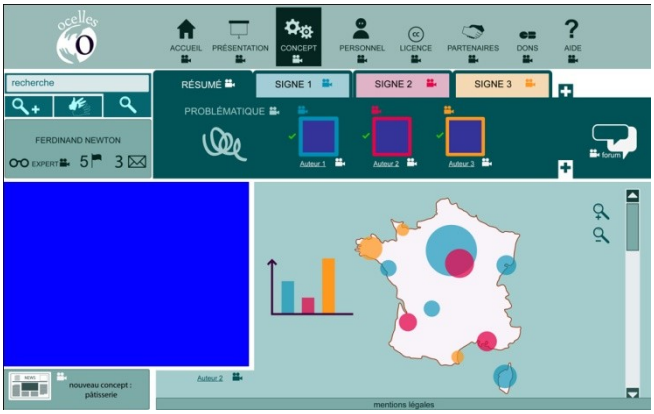

Fig.2 : Copie d'écran d'une « page résumé ».

Les utilisateurs sont invités à répondre à des questions fermées sur :

- le ou les contexte(s) d'utilisation (en classe, au travail, …) de chaque signe entré ;
- les caractéristiques du signe (réalisé à une ou deux mains, formé à partir d'un transfert [5] de forme, situationnel, personnel, ses configurations, ses emplacements, son mouvement, la mimique faciale, ...).

D. Résumé des propositions

Pour un concept donné les résultats recueillis pour chaque signe sont résumés sur une page. Ces résultats peuvent être croisés avec les profils des utilisateurs, ce qui permet notamment de connaître chaque localisation géographique et le nombre de réponses associées.

Plateforme d'échanges, ce site sert ainsi d'observatoire linguistique dans une approche synchronique.

III. MODELE D'ORGANISATION

A Limite d'une classification figée

La première classification utilisée dans notre projet fut celle décimale de Dewey (CDD) utilisée dans les bibliothèques. En effet, ce système développé par Melvil Dewey en 1876 [6] a l'avantage d'exister et de permettre de classer l'ensemble du savoir humain.

Les tests utilisateurs composés d'un public mixte sourds/entendants ont rapidement mis en avant les difficultés pressenties. La CDD répartit les ouvrages en dix classes générales. Chacune d'elles est divisée en dix divisions, elles mêmes divisées en dix subdivisions et ainsi de suite. Une recherche de concept à travers cette arborescence impose donc :

- de connaître les thèmes et les sous-thèmes contenant le concept souhaité ;
- de ne pas commettre d'erreur à chaque nœud de l'arborescence sur le choix de chaque sous-thème : avec cette approche peu intuitive, un concept de « bas niveau » se situe loin dans l'arborescence et nécessite la connaissance de concepts de « haut niveau ».

La LSF, quant à elle, est un système linguistique qui allie un aspect catégoriel inhérent à la structure de son vocabulaire. Selon les travaux de C. Courtin [7] l'utilisation de la LSF chez les enfants sourds nés de parents sourds et signeurs, amène à de meilleures performances de catégorisations que chez des enfants entendants ou sourds oralisants. Ce phénomène s'applique pour les catégorisations qui respectent la complexité du monde réel en utilisant des prototypes ou des schémas (type écologique [8]). Ainsi, pour énoncer un concept de niveau de catégorisation supérieur comme un « être vivant » ou un « mammifère », il est fréquent d'utiliser une suite de signes d'exemples prototypiques de cette catégorie définissant ainsi la classe en extension. Une classification figée et arbitraire pourrait donc perturber les utilisateurs sourds.

D'autre part comment classer des concepts nouveaux directement issus de la langue des signes, tels que « signaire » (ensemble des signes des langues des signes, comme il existe l'ensemble des mots d'une langue donnée) ou « LS Vidéo » (enregistrements vidéos de LSF formalisée en vue d'une communication différée) dans une classification figée construite à partir d'une langue écrite ?

B Classification dynamique

L'une des principales problématiques de l'organisation des données a été d'arriver à concilier les différents points de vue et usages des utilisateurs. Le portail devait pouvoir évoluer et s'enrichir facilement, par l'ajout de nouveaux concepts, mais aussi par la prise en compte de nouveaux descripteurs ou de nouvelles fonctionnalités (illustrations, homonymies, antonymies, etc.). Il était impossible de rester sur un modèle figé d'organisation.

Nous avons donc opté pour une organisation des données considérant tout élément du site comme contenu sans a priori. En parallèle, une structure associative et dynamique permettant de lier les contenus entre eux en fonction de rôles et de descripteurs pour ce rôle [9] a été mise en place.

Ainsi, un même contenu peut être utilisé plusieurs fois et dans des contextes différents, suivant les associations (rôles et descripteurs) dont il fait partie [10].

Par exemple : dans notre site, un nœud de classification joue un rôle de « thème ». Un thème est également un concept, par exemple « Langage ». Un thème répertorie d'autres thèmes et des concepts. Le thème « Langage » contient le thème « Langue », « pragmatique », …. De plus, le contenu « Langage » joue également un rôle d'illustration pour le concept « Langue ». Nous avons

donc pour un même contenu (« Langage ») plusieurs rôles (thème, concept), et chacun de ces rôles dispose de ses propres descripteurs (illustration du concept, concept associé au thème, etc.).

Pour finir, l'arborescence doit permettre de classer un concept dans plusieurs thèmes sans duplication. Ce problème est rencontré dans les bibliothèques lorsqu'il s'agit de répertorier un ouvrage concernant plusieurs thématiques : un livre sur les sciences au XIXème siècle en France va également concerner les Sciences et l'Histoire...

C Discussion sur une classification dynamique à partir des paramètres linguistiques de la LSF.

Selon Cuxac [5], deux stratégies discursives d'énonciations coexistent en LSF. Le signeur via le canal visuel-gestuel, choisit de dire sans montrer ou bien de dire en montrant. Il peut ainsi « donner à voir » l'expérience en s'appuyant sur la ressemblance maximale – séquence de signes - expérience, ou utiliser le signe standard sans ressemblance avec le référent. S'appuyant sur cette théorie de l'iconicité, notre recherche a pour hypothèse une hiérarchisation des paramètres linguistiques utilisés dans les signes comme éléments porteurs de sens.

Si ces structures de grande iconicité font intervenir des éléments linguistiques infra-lexicaux, n'appartenant pas au lexique, elles apparaissent très majoritairement dans certaines activités de récit sans pour autant figurer dans aucun dictionnaire de langue des signes. Néanmoins, si l'on considère ces structures comme faisant partie intégrante de cette langue, comment les intégrer à notre site ?

Pour répondre à cette question, deux axes de travail sont envisagés.

Un premier axe consiste à s'interroger sur les structures minimales de réalisation en langue des signes. Les paramètres linguistiques que sont la configuration, le mouvement, l'emplacement [11], l'orientation [12] [13], ...[14], [15] ne peuvent pas être considérés comme tels. En effet, si l'esprit peut les isoler de manière pertinente comme étant des éléments porteurs de sens, pour pouvoir les mettre en activité dans une communication, il faut obligatoirement les réaliser simultanément. Contrairement à ce qui se passe dans les langues vocales, la réalisation d'une forme significative en langue des signes ne peut avoir lieu avec des éléments réalisables isolément non porteurs de sens.

Un échelonnement des structures minimales de réalisation en langue des signes peut être établi sur une échelle de complexité croissante allant du transfert de forme (niveau infra-conceptuel) jusqu'au double transfert (niveau pouvant faire intervenir, plusieurs actants, locatifs ou énoncés), alors que pour le même laps de temps de réalisation, ces différentes structures monopolisent les mêmes paramètres linguistiques.

Un second axe de travail repose sur la dialectique entre syntagme et paradigme. Dans des récits utilisant des structures fortement iconiques, la valeur que prend un élément à un instant donné, subit une sorte de pression syntagmatique, mais pas forcément de la part d'unités qui ont précédé ou qui vont succéder, mais d'unités co-occurrentes qui vont également participer à cette forme minimale de réalisation. Néanmoins la simultanéité ne permet pas d'établir qu'il s'agit du paradigme, puisque ces pressions donnent à voir. En langue des signes, les pressions du contexte ne s'exercent pas sur la seule dimension temporelle. La dimension spatiale introduit également des contraintes non plus successives mais simultanées.

A partir de ces axes de travail, les utilisateurs sont interrogés lors de leur première consultation d'une « page signe » sur leurs perceptions et représentations de ces unités infra-conceptuelles porteuses de sens du signe consulté.

L'objectif n'est pas de recueillir des réponses exactes mais d'obtenir un maximum de réponses identiques. Le classement repose donc sur un consensus des utilisateurs.

Toutefois les experts peuvent imposer un classement et orienter les recherches d'un signe à partir des réponses proposées sans pour autant l'utiliser comme classement final.

A partir de l'ensemble des réponses des utilisateurs, des descripteurs/classificateurs sont affectés à chaque signe en fonction de la synthèse établie à partir d'un nombre dynamique de réponses identiques significatives.

Le modèle de base de données proposé est basé sur la modélisation non pas du contenu à proprement parler mais des interactions qui lui donnent sens. Ce polymorphisme dans l'utilisation des contenus implique une organisation des informations basée non pas sur le contenu lui-même mais sur le rôle qu'on souhaite lui voir jouer, comme expliqué plus haut. Ainsi le contenu « réponse » joue à la fois le rôle de réponse à la question, mais également celui d'élément de classification et d'élément de recherche. A chaque rôle correspond un ou plusieurs descripteurs qui lui sont propres.

Cette approche de classement dynamique des concepts à partir des paramètres linguistiques de la LSF inhérents à chaque signe permet de proposer aux utilisateurs signeurs une entrée directement en LSF sans connaître le signifiant écrit du concept.

Par la suite une classification dynamique pourrait également être basée sur sign writing [16].

IV. BIBLIOGRAPHIE

[1] http://cis.gouv.fr/spip.php?article1516

[2] P. Dalle, "La place de Langue des Signes dans le milieu institutionnel de l'éducation : enjeux, blocage et évolution", *in Langue française n°137, La Langue des Signes : statuts linguistiques et institutionnels, Larousse, Paris*, 2003

[3] F. Duquesne-Belfais "Activité et langages dans la conceptualisation mathématique : des apprentissages des élèves sourds à la formation de leurs enseignants", *Thèse de doctorat en Sciences de l'Éducation, Université Lille 1*, 2007

[4] http://www.gnu.org/licenses/licenses.fr.html

[5] C. Cuxac. "La Langue des Signes Française : les voies de l'iconicité. " *Paris : OPHRY*, 2000.

[6]http://fr.wikipedia.org/wiki/Classification_d %C3%A9cimale_de_Dewey

[7] C. Courtin, "Surdité, langue des signes et développement cognitif", *Thèse de doctorat en psychologie, Université Paris V*, 1998.

[8] J. Bideaud, O. Houde, J.-L. Pedinielli, "L'homme en développement", *Paris, PUF*, 1993.

[9] .C. Moreau, B. Mascret "Lexique LSF", *3rd Workshop on the Representation and Processing of Sign Languages. 6th International Conference on Language Resources and Evaluation*, LREC 2008

[10] A. Bénel. "Chapters 5 and 7 in Consultation assistée par ordinateur de la documentation en sciences humaines : considérations épistémologiques, solutions opératoires et applications à l'archéologie", *Ph.d. thesis, Institut National des Sciences Appliquées, LIRIS, Lyon, France*, 2003

[11] W.C. Stockoe, D.C. Casterline, C.G. Cnroneberg. "A dictionary of American Sign Language", *Washingtion, D.C. : Gallaudet College Press*, 1965.

[12] L, Friedman. "On the other hand, new perspectives on american Sign Language", *New York : Academic Press*, 1977.

[13] S.K Lidell. "American Sign Language Syntax", *The Hague : Mouton Publisher*, 1980.

[14] B. Moody. "Histoire et Grammaire", tome I. Paris : Ellipses. 1983

[15] S.-C. Yau. "Création Gestuelle et début du Langage – Création de langues gestuelles chez les sourds isolés", *Hong Kong : Éditions Langages Croisés*, 1992.

[16] http://www.signwriting.org/