LOISEAU MATHIEU

Mémoire :

Vers la création d'une base de données de ressources textuelles indexée pédagogiquement pour l'enseignement des langues.

SOUTENU LE 26 JUIN 2003

DEA SCIENCES DU LANGAGE
UNIVERSITÉ STENDHAL

2002-2003

TRAVAIL ENCADRÉ PAR :
GEORGES ANTONIADIS
CLAUDE PONTON

Remerciements

IMMENSE REMERCIEMENT N° 1 :

Avant toute chose, je tiens à remercier tous les enseignants qui ont bien voulu m'accorder leur temps, d'autant plus précieux que le contexte social tendu de ce printemps 2003 ne se prêtait pas forcément idéalement aux activités "*extra-scolaires*". Leurs témoignages constituent la matière première de ce mémoire, un énorme merci à :

- Myriam Béatrix
- Sophie Bourgade
- Adriana Celińska
- Maria-Elena Galoppo
- Alice Henderson
- Iwona Puchalska
- Michel Sainty
- Sonia Tendero

IMMENSE REMERCIEMENT N° 2 :

Pour mes parents, qui non contents de soutenir mes études depuis sept ans, ont relu le présent mémoire durant un week end ensoleillé, fort propice à la promenade…

IMMENSE REMERCIEMENT N° 3 :

A Mick Souvy, sans qui le nombre d'entretiens aurait été quasiment amputé de moitié. Merci pour tout.

MAIS CELA NE VEUT PAS DIRE QUE LES AUTRES N'ONT PAS JOUÉ UN RÔLE IMPORTANT, UN GRAND MERCI :

A Georges Antoniadis et Claude Ponton, pour leur disponibilité et leurs conseils avisés.

Au collège de Bissy (73), qui m'a accueilli au pied levé dans ses locaux et de m'a permis d'effectuer mes enregistrements dans de bonnes conditions.

A Noëlle Abrial, qui m'a hébergé gracieusement toute une partie de l'année.

Au Rouquin pour les bouquins.

A Audrey, Cédric, Cindy, Charlotte, Chouch', Christoph, Claude, Cyril, Elisa, Elo, Fabri, Fred, Founet, François-Karim, Jaideep, Jildaz, Laura, Lucia, Marco, Margotte, Marie, Nata, Nicole, Ola, Roubi, Sam, Sophie, Steph et Vincent pour le "*soutien psychologique*".

Table des matières

1. Présentation du sujet 1
  1.1. Le projet MIRTO 1
    1.1.1. Le but du projet 1
    1.1.2. Structure de la plate-forme 2
      1.1.2.1. Les utilisateurs 2
      1.1.2.2. Les outils TAL 2
      1.1.2.3. Scripts, activités et scenarii 3
      1.1.2.4. La base de textes 4
      1.1.2.5. Historique (le parcours des étudiants) 4
  1.2. Le mémoire 4
    1.2.1. Les trois activités liées à la base de textes 5
      1.2.1.1. Principes généraux de bases de données 5
      1.2.1.2. Ajout de textes dans la base 7
      1.2.1.3. Recherche de textes dans la base 7
      1.2.1.4. Utilisation des textes par les outils 8
    1.2.2. Le sujet 8
      1.2.2.1. Domaine 8
      1.2.2.2. « Problématique » 9
      1.2.2.2.1. Complète 10
      1.2.2.2.2. Cohérente 10
      1.2.2.2.3. Difficulté 10
  1.3. Terminologie pour le reste du mémoire 11
    1.3.1. Texte : 11
    1.3.2. Opposition annotation / description : 11
    1.3.3. Texte authentique / texte pédagogique 11
2. La demarche 13
  2.1. Base de textes 13
    2.1.1. FranText 13
    2.1.2. FRIDA 14
    2.1.3. TEI 14
  2.2. Contexte de recherche 15
  2.3. Un produit pour quels utilisateurs 16
  2.4. Les entretiens 16
    2.4.1. Stratégie 16
    2.4.2. Les « *sujets* » 17
    2.4.3. Déroulement des entretiens 19
      2.4.3.1. Enregistrement 19
      2.4.3.2. Prise de notes 19
    2.4.4. Contenu 19
3. Préparation des entretiens 21
  3.1. Le niveau d'enseignement 21
    3.1.1. Exemples d'Evaluation du niveau 21
      3.1.1.1. TOEIC 21
      3.1.1.1.1. Une notation selon un axe déterminé 21
      3.1.1.1.2. Granularité 22
      3.1.1.2. « La classe de langue » 22

3.1.2. Multilinguisme de la base 22
3.1.3. Position du problème à des enseignants chevronnés (dans diverses langues) 23
3.1.3.1. Préparation des entretiens 23
3.1.3.1.1. Fiche 24
3.2. Les emplois actuels des ressources textuelles en classe 24
3.2.1. Hypothèse de travail 24
3.2.2. Etat de l'art 25
3.2.2.1. Avant propos 25
3.2.2.2. Activités de compréhension d'un texte écrit 25
3.2.2.2.1. Inférence du sens d'un mot d'après le contexte 26
3.2.2.2.2. Lecture rapide 27
3.2.2.2.3. Questions de compréhension globale 27
3.2.2.2.4. Textes littéraires 28
3.2.2.2.5. Extraction d'informations détaillées et/ou sélectives à partir d'un texte 29
3.2.2.3. Compréhension de la structure du texte 30
3.2.2.3.1. Registre et style 31
3.2.2.3.2. La ponctuation 31
3.2.2.3.3. A partir d'une chanson (le texte sans écoute) 31
3.2.2.3.4. Structure logique et type dominant d'un texte 32
3.2.2.3.5. Production 33
3.2.2.4. Travail avec objectif linguistique 34
3.2.2.4.1. Travail sur les structures grammaticales 34
3.2.2.4.2. Travail sur le vocabulaire 36
3.2.2.4.3. Travail sur la phonétique 37
3.2.2.5. Autres pratiques 38
3.2.2.5.1. Vocabulaire 38
3.2.2.5.2. Mise en page 38
3.2.3. Comment les enseignants définissent-ils leurs usages des textes ? 38
3.2.3.1. Plan de cette phase de l'entretien 39
3.2.3.1.1. Fiche 39
3.3. La recherche de documents 40
3.3.1. Avant les entretiens 40
3.3.2. Fiche pour l'entretien 41
4. Compte-rendu des entretiens 42
4.1. Le niveau 42
4.1.1. Remarque générale 42
4.1.2. Le niveau dans le contexte de la classe 42
4.1.3. Le niveau en auto-formation 43
4.1.4. Les compétences 44
4.1.4.1. Les familles de compétences 44
4.1.4.2. Granularité 45
4.2. Activités et processus de recherche 45
4.2.1. Structure 45
4.2.2. Activités qui n'avaient pas été vues dans l'état de l'art 45
4.2.2.1. Conception grammaticale 46
4.2.2.2. Signes diacritiques 46
4.2.3. Les différentes classes d'activité 46
4.2.3.1. Bases personnelles 46
4.2.3.1.1. La recherche passive 46
4.2.3.1.2. Organisation 47

4.2.3.1.2.1. Niveau 47
4.2.3.1.2.2. Thème 48
4.2.3.1.2.3. Point grammatical 48
4.2.3.1.2.4. Civilisation 48
4.2.3.1.3. Conclusions 49
4.2.3.2. Exercices 49
4.2.3.3. Terminologie 49
4.2.4. Choix des documents 50
4.2.4.1. Prise en compte du niveau 50
4.2.4.1.1. Longueur du texte 50
4.2.4.1.2. Vocabulaire et structures grammaticales 50
4.2.4.1.2.1. Influence de la langue 51
4.2.4.1.2.2. Influence du niveau 51
4.2.4.1.3. Conclusion 51
4.2.4.2. Prise en compte des classes d'activités 51
4.2.4.2.1. Activités linguistiques 52
4.2.4.2.1.1. Alternative 52
4.2.4.2.1.2. Champs lexicaux / champs sémantiques 52
4.2.4.2.2. Activités de compréhension 52
4.2.4.2.2.1. Autres critères 53
4.2.4.2.2.2. Remarque 53
4.2.4.2.3. Activité d'approfondissement 54
4.2.4.2.4. Tableau récapitulatif 54
4.2.4.3. Autres critères 55
4.2.4.3.1. Thème 55
4.2.4.3.2. Type de texte 56
4.2.4.3.3. Style 57
4.2.4.3.4. Texte authentique, édité ou inventé 57
4.2.5. Processus de recherche 58
4.2.5.1. Remarque 58
4.2.5.2. Source 58
4.3. Récapitulatif 59
5. Modélisation informatique 62
5.1. Données candidates 62
5.2. Champs potentiels de la base 62
5.2.1. Choix d'un formalisme 62
5.2.2. Diagramme de classes UML 64
5.2.2.1. Classes et attributs 64
5.2.2.2. Clé 64
5.2.2.3. Notations 65
5.2.2.3.1. Classe 65
5.2.2.3.2. Relation / Cardinalité 65
5.2.2.3.3. Classes associatives 66
5.2.2.3.3.1. Définition 66
5.2.2.3.3.2. Notation 67
5.2.2.3.4. Héritage 67
5.2.2.3.4.1. Définition 68
5.2.2.3.4.2. Notation 68
5.2.3. La base de textes 68
5.2.3.1. La classe Texte 69

5.2.3.1.1. La clé ... 69
5.2.3.2. Autres attributs ... 70
5.2.3.3. La classe Auteur ... 71
5.2.3.3.1. La clé ... 71
5.2.3.3.2. Autre attribut ... 71
5.2.3.4. La relation entre les classes Auteur et Texte ... 71
5.2.3.5. La classe source ... 72
5.2.3.5.1. Attributs ... 72
5.2.3.5.2. La clé ... 72
5.2.3.6. La relation entre les classes Source et Texte ... 72
5.2.3.7. La classe associative Modification ... 73
5.2.3.7.1. Multiplicité de la relation ... 73
5.2.3.7.2. Clé ... 74
5.2.3.7.3. Autres attributs ... 74
5.2.4. Implantation ... 74
5.2.5. Ajout d'un texte dans la base ... 75
5.2.5.1. Champs laissés à la responsabilité de l'utilisateur ... 75
5.2.5.1.1. Auteur ... 75
5.2.5.1.1.1. Nom et prénom ... 76
5.2.5.1.1.2. Dates ... 76
5.2.5.1.1.3. Nationalité(s) ... 76
5.2.5.1.2. Source ... 76
5.2.5.1.2.1. Type ... 76
5.2.5.1.2.2. Titre ... 77
5.2.5.1.2.3. Date ... 77
5.2.5.1.2.4. PaysDePublication ... 78
5.2.5.1.3. Texte ... 78
5.2.5.1.3.1. Type ... 78
5.2.5.1.3.2. Titre ... 80
5.2.5.1.3.3. Date ... 80
5.2.5.1.3.4. Intégrité ... 80
5.2.5.1.3.5. Authenticité ... 80
5.2.5.1.4. Modification ... 81
5.2.5.1.4.1. Nature ... 81
5.2.5.1.4.2. Responsable ... 81
5.2.5.2. Champs entrés automatiquement ... 81
5.2.5.2.1. Champs concrets ... 81
5.2.5.2.1.1. Identificateurs ... 81
5.2.5.2.1.2. Mesures ... 82
5.2.5.2.1.3. Intégrité / Authenticité ... 82
5.2.5.2.2. Champs dont l'utilisation nécessite une couche logicielle intermédiaire ... 82
5.2.5.2.2.1. Principes généraux de l'indexation ... 82
5.2.5.2.2.2. Grammaire ... 83
5.2.5.2.2.3. Thème, Vocabulaire, ChampsLexicaux ... 83
5.2.5.2.2.4. Langue ... 84
5.2.5.2.2.5. Style ... 86
5.2.5.3. Structure de l'en-tête TEI ... 86
5.2.5.4. Conclusions sur l'ajout de texte dans la base ... 88
5.2.6. Recherche d'un document dans la base ... 88
5.2.6.1. Champs concrets ... 88

5.2.6.2. Thème 90
5.2.6.3. Vocabulaire 91
5.2.6.3.1. Recensement du vocabulaire 91
5.2.6.3.2. Tolérance 91
5.2.6.3.3. Idée de mise en œuvre 92
5.2.6.3.4. Représentation des résultats 92
5.2.6.4. Champs Lexicaux et Style 92
5.2.6.5. Grammaire 93
5.2.6.5.1. Expression des structures grammaticales 93
5.2.6.5.1.1. Nommage des phénomènes grammaticaux 94
5.2.6.5.1.2. Développement d'un autre formalisme 94
5.2.6.5.1.3. Par l'exemple 95
5.2.6.6. Concordances 96
5.2.6.7. Conclusions sur la recherche de documents 97
6. Conclusion 98
6.1. Remarques générales 98
6.2. Le fonctionnement de la base de texte indexée en fonction de critères pedagogiques 98
6.2.1. Ajout d'un document dans la base 99
6.2.2. Recherche d'un document dans la base 100
6.3. Poursuite du travail 101
6.3.1. Jusqu'à la réalisation d'un prototype 101
6.3.2. Vers la réalisation de la base… 101
6.3.2.1. …de textes 101
6.3.2.2. …de supports pédagogiques 101
7. Bibliographie 103
7.1. Sites 103
7.2. Livres 104

1. PRÉSENTATION DU SUJET

Le présent mémoire trouve son origine dans un projet coordonné par le DIP[1] : le projet MIRTO (Multi-apprentissages Interactifs par des Recherches sur des Textes et l'Oral). Avant d'entrer dans le détail de ce travail, nous allons brièvement expliciter le cadre dans lequel il s'inscrit.

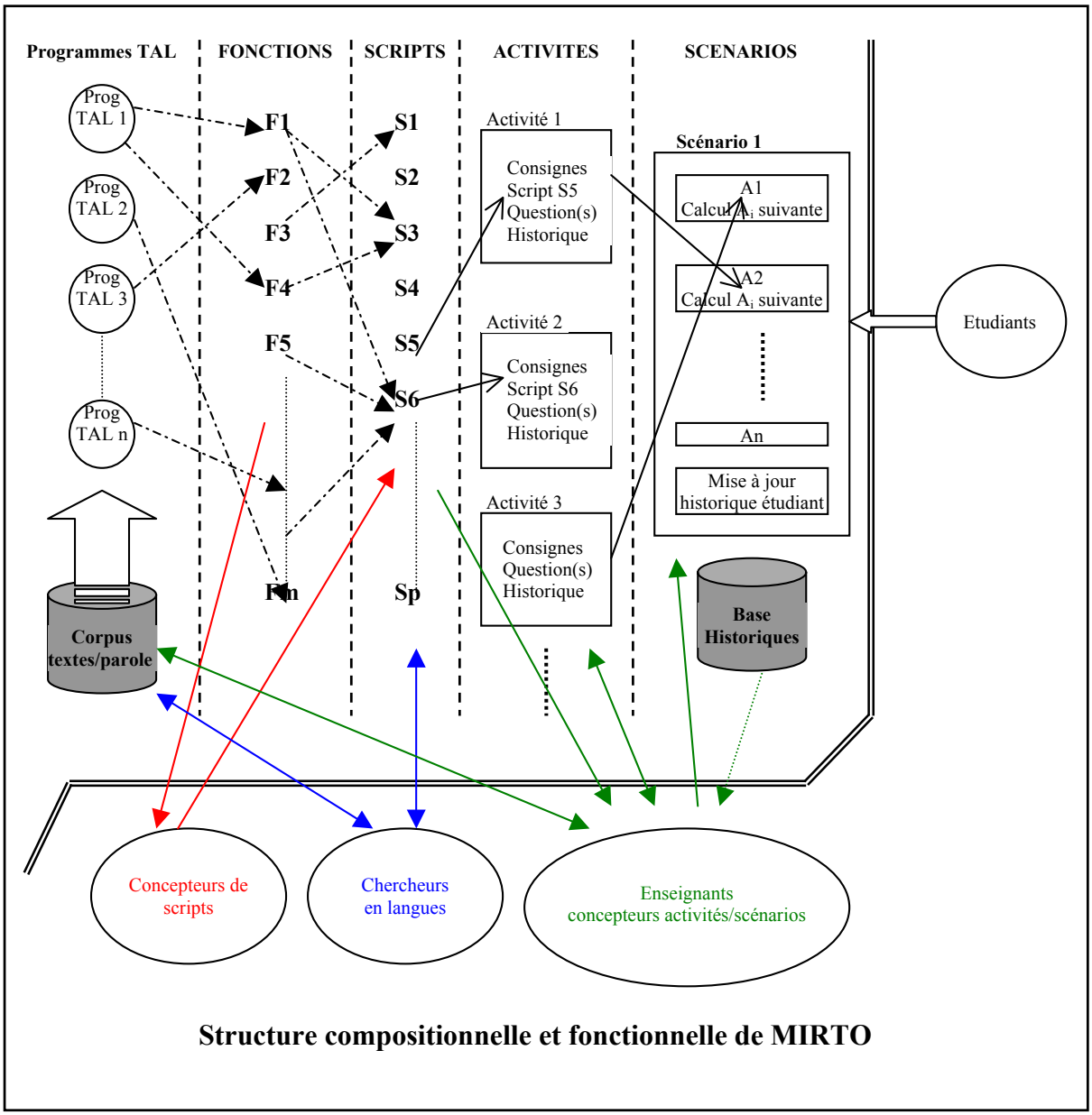
1.1. LE PROJET MIRTO

1.1.1. Le but du projet

Le projet MIRTO a pour but la création d'une plate-forme dédiée à l'apprentissage des langues. Il s'est fixé de prendre en compte les diverses évolutions du domaine des outils TAL et de les intégrer à une plate-forme pour l'enseignement des langues (plate-forme multilingue), permettant ainsi d'éviter les travers des systèmes d'apprentissage de langues existant. Ces derniers se contentent, en général, de juxtaposer l'outil informatique à l'apprentissage des langues sans vraiment l'intégrer [ANT 95]. En effet la plupart des systèmes d'apprentissage des langues se bornent à n'utiliser l'informatique que comme un outil. En tant que tel, cet outil ne permet que d'effectuer des comparaisons de deux chaînes de caractères (telles que 'un projet' ≠ 'un projet'[2], dont on peut imaginer les conséquences dans un logiciel d'apprentissage des langues) ou d'effectuer d'autres opérations extrêmement basiques. La nouveauté dans le projet MIRTO réside dans la tentative d'intégrer des outils dédiés au traitement automatique des langues naturelles à la plate-forme [A-P 02]. Ces outils tiennent compte des spécificités de la langue (par exemple en ce qui concerne la comparaison ci-dessus). Ils seront intégrés dans une plate-forme qui prendra donc en compte les spécificités du traitement automatique des langues : la plate-forme se démarquera des systèmes dédiés à l'apprentissage des langues existants, en exploitant les avancées dans le domaine des outils TAL sans oublier les limites de tout outil informatique : possibilité de fonctionner de manière générique dans certains cas définis précisément, manque de fiabilité quand on sort de ce cadre.

[1] Département d'Informatique Pédagogique (université Stendhal)
[2] 'un_projet' ≠ 'un__projet' Dans le premier cas il n'y a qu'un espace alors que dans le second il y en a deux.

1.1.2. Structure de la plate-forme



Structure compositionnelle et fonctionnelle de MIRTO

Figure 1 Structure de Mirto

1.1.2.1. Les utilisateurs

MIRTO est, d'une part, destiné aux enseignants et aux chercheurs, afin de préparer leurs cours pour les uns et de faciliter les travaux des seconds ; d'autre part, les étudiants pourront bénéficier de la plate-forme pour prolonger et approfondir leurs enseignements.

1.1.2.2. Les outils TAL

Ce terme désigne aussi bien les logiciels que les fonctions TAL. Les fonctions sont en général des sous-parties des traitements effectués par les logiciels. Les logiciels pourront être

utilisés tels quels par les apprenants (ex : encyclopédie numérique). Les fonctions quant à elles (ex : étiqueteur) seront ré-utilisées dans l'écriture d'un script (voir paragraphe concerné). Tous ces produits peuvent aussi bien venir du commerce, que d'équipes de recherche (de l'université Stendhal par exemple) ou encore d'industriels (comme XRCE) [MIR #1]. Le diagramme ci-dessus suggère que la plus grande partie des logiciels TAL fournis, seront livrés avec leurs API[3] : une API est un ensemble de fonctions mis à la disposition des programmeurs afin de permettre la réutilisation du code en ne se préoccupant que des paramètres d'appel des fonctions, sans avoir à s'inquiéter de leur implémentation. Par exemple, dans le cas d'une encyclopédie en ligne, on peut imaginer qu'une API, s'il y en a une, permettra la recherche d'un mot, par le biais d'une fonction. On pourra donc afficher l'article concernant un mot sans avoir à passer par le logiciel encyclopédie (et donc son interface). Ainsi on pourrait proposer à l'apprenant de visualiser les articles concernant les mots qu'il ne comprend pas, au moyen d'un simple clic dans l'interface de MIRTO. Cet exemple de l'encyclopédie ne se base pas nécessairement sur une fonctionnalité de la plate-forme MIRTO, il a juste pour vocation d'expliciter la notion d'API.

1.1.2.3. Scripts, activités et scenarii

Les fonctions seront ensuite utilisées pour créer des scripts, qui à leur tour seront associés à une activité. Un script correspond à un outil permettant d'automatiser une partie d'une activité, par exemple la génération d'exercices. Les scripts seront implémentés à la demande d'enseignants qui pourront par la suite créer des scenarii d'apprentissage (séquences d'activités) destinés aux étudiants (apprenants) [MIR #2].

Par exemple, l'enseignant veut faire travailler à certains étudiants la distinction entre présent simple et présent progressif en anglais, on créera un générateur de textes lacunaires, qui utilisera :

- Un étiqueteur : qui permettra de repérer dans le texte tous les verbes au présent simple ou au présent progressif.
- Un lexique des formes de l'anglais : que l'on utilisera pour retrouver la forme lemmatique (infinitif) des verbes concernés.

On obtiendra ainsi un "générateur" d'exercices lacunaires, qui à la donnée d'un texte source, rendra un exercice dans lequel tous les verbes aux présent auront été mis à l'infinitif.

[3] API : Applications Program(ming) Interface

Dans cet exemple, l'activité est un exercice généré automatiquement à la donnée d'un texte (alors que le script est le processus qui permet de générer automatiquement l'exercice).

Cette activité pourra ensuite être incluse dans un scénario, auquel seront confrontés les étudiants éprouvant des difficultés par rapport à l'utilisation contrastive du présent simple et du présent progressif en anglais.

1.1.2.4. La base de textes

Elle est appelée "*corpus*" dans la figure 1. Ce corpus va contenir les ressources textuelles qui seront utilisées pour l'élaboration des différents scénarios. Il devra être annoté et indexé, de manière à permettre aux enseignants une sélection simple des mieux adaptées à telle ou telle activité dans le scénario qu'il est en train de créer. Les données du corpus doivent aussi bien permettre la recherche de concordances (pour des exercices types, de grammaire, ou de vocabulaire) que de textes entiers, pour des questions de compréhension ou autre.

C'est cette facette du projet qui nous intéresse dans ce mémoire. Bien-sûr, il faudra garder à l'esprit les autres parties du projet MIRTO, mais c'est à cet aspect précis du problème que nous allons nous intéresser.

1.1.2.5. Historique (le parcours des étudiants)

Il ne figure pas sur le diagramme ci-dessus, mais la plate-forme devrait contenir un historique, pour permettre un suivi personnalisé des étudiants. Elle sera mise à jour automatiquement et contiendra le parcours de chaque étudiant (quelles activités ont été faites, avec quel score…)

1.2. LE MÉMOIRE

MIRTO est donc à l'origine de ce mémoire même si sa réalisation reste indépendante de l'évolution du projet et inversement. A terme, il est possible que les deux se rejoignent mais pour l'année universitaire 2002-2003, les deux sont restés des entités autonomes. Cela ne nous a, bien-sûr, pas empêcher de garder à l'esprit les caractéristiques du projet MIRTO dans la réalisation de ce travail.

Le mémoire s'inscrit dans la perspective de la base de textes ; tout ce qui touche à sa réalisation, à son utilisation et bien entendu à la prise en compte des aspects pédagogiques qui doivent gouverner les deux activités.

1.2.1. Les trois activités liées à la base de textes

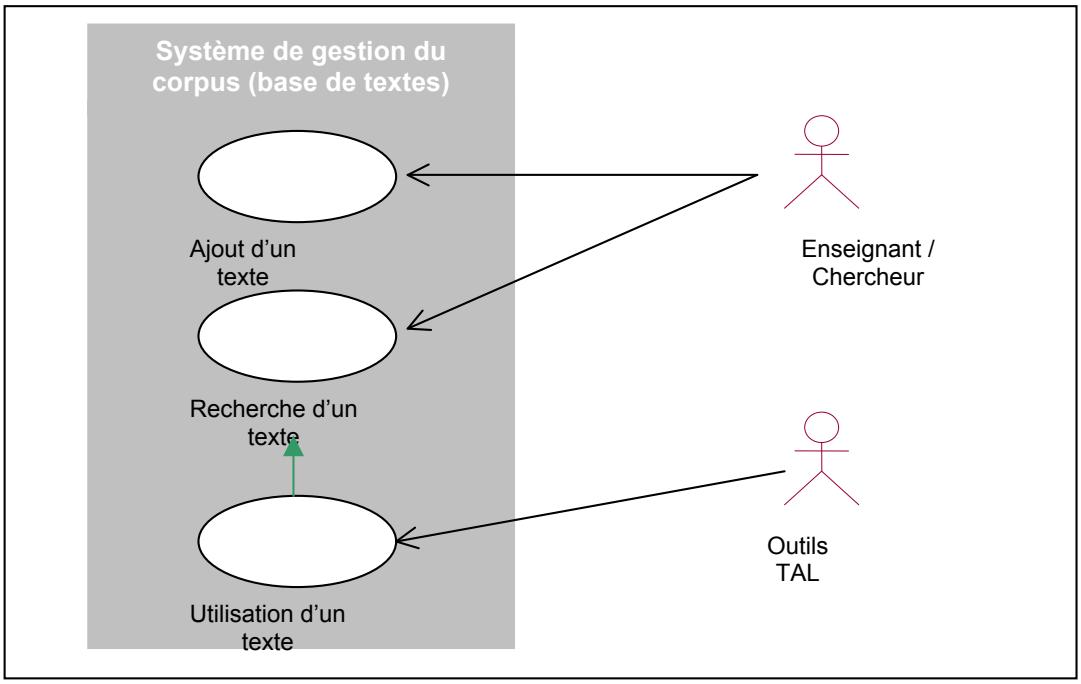


Figure 1 Diagramme UML (cas d'utilisation) [OMG 03] pour la base de textes

Le diagramme UML[4] ci-dessus représente les différents cas d'utilisation de la base de textes, ou plutôt du système de gestion de la base de texte, qui n'est pas intrinsèquement utilisable et nécessite une couche logicielle pour être exploitée. En d'autres termes, ce sont les différentes fonctionnalités qu'on attend pour le système de gestion de la base de textes.

Selon les standards UML, chaque ovale représente un cas d'utilisation (une fonctionnalité). Les "*bonshommes*" représentent des acteurs extérieurs au système de ces cas d'utilisation. Les outils TAL tels que nous les avons définis dans le paragraphe précédent satisfont ces conditions.

1.2.1.1. Principes généraux de bases de données

La base de textes (à vocation pédagogique) a des caractéristiques communes avec toute autre base de données.

Dans toute la suite du mémoire, nous allons parler de "*requête*" et de "*champs*". Afin de faire comprendre de quoi il s'agit, nous allons expliquer dans les grandes lignes les principes d'utilisation des bases de données. Nous n'entrerons pas dans les détails et allons simplifier le plus possible les notions, de manière à ce que le reste du mémoire (à l'exception peut être de la partie technique) soit le plus abordable possible.

[4] Unified Modeling Language

Une base de données, comme son nom l'indique, contient des données, correspondant à des éléments du monde réel. Nous n'allons pas entrer ici dans le détail des tables et des relations. En revanche, il faut savoir que pour identifier les données, on utilise des traits caractéristiques, que l'on appelle des champs. Par exemple dans le cas d'une base de données représentant la discothèque de quelqu'un, les traits caractéristiques des données stockées seront, à priori, si l'on s'intéresse aux albums : l'interprète, le titre, l'année de sortie, le nombre de chansons, les titres des chansons, le style musical, si c'est un album ou une compilation etc.…

Si l'on s'intéresse aux chansons une par une, les champs seront à priori : le titre, l'auteur, le compositeur, l'interprète, l'année, la durée, l'album dont elles sont tirées, la position dans l'album.

Une requête dans une base de données, consiste à spécifier des valeurs pour certains champs et à demander au système d'aller chercher tous les éléments les satisfaisant. Dans le cas précédent, si l'on s'intéresse aux albums, on pourra par exemple rechercher tous les albums des Pixies, ou tous les titres d'albums et interprètes de musique folk parus en 1966 et le système renverra toutes les données que l'on a demandées, qui soient liées aux objets trouvés. Par exemple, pour la première requête on aura :

Interprète	Titre	Année	Nb Chansons	Style	**Album**	…
The Pixies	Come on Pilgrim	1987	8	Rock	**oui**	
The Pixies	Surfer Rosa	1988	13	Rock	**oui**	
The Pixies	Doolittle	1989	15	Rock	**oui**	
The Pixies	Bossanova	1990	14	Rock	**oui**	
The Pixies	Trompe Le Monde	1991	15	Rock	**oui**	

Tableau 1 Résultat de la requête « Albums des Pixies »

On a mis en gras les données de la requête.

Alors que pour la seconde, étant donné que l'on ne demande que les titres et interprètes des albums, on aura :

Année	Interprète	Titre	**Style**
1966	Davey Graham	Midnight Man	**Folk**
1966	Bert Jansch	It don't bother me	**Folk**
1966	Bob Dylan	Blonde on blonde	**Folk**

Tableau 2 Résultat de la requête « titre d'album et interprète pour les disques folks parus en 1966 »

Enfin, pour l'ajout de données dans la base, l'utilisateur devra pour chaque objet entrer la valeur de chaque champ, afin que ceux-ci soient ensuite utilisables pour la recherche.

1.2.1.2. Ajout de textes dans la base

L'un des buts de la base de textes est qu'elle ne soit pas figée. L'intérêt d'une telle base, réside dans le fait qu'elle n'est pas définitive, mais en évolution constante. En effet, si la requête d'un enseignant n'offre pas de résultats satisfaisants et qu'il est obligé d'utiliser des moyens plus "*traditionnels*"[5] pour parvenir à ses fins, il serait dommage que le fruit de ses recherches ne puisse pas profiter à d'autres. Ce type de pratique a été mis en place pour d'autres types de bases de données avec d'excellents résultats : on peut prendre l'exemple de CDDB [CDDB]. Une base de données contenant les informations relatives à des CDs. Lors de l'insertion d'un CD dans le lecteur de l'ordinateur, un certain nombre de logiciels formulent automatiquement une requête à partir du nombre de chansons, de leurs durées et éventuellement d'autres informations (ce que l'on appellera la "*signature*" du CD), et récupèrent ainsi le titre de l'album, l'artiste, les titres des chansons et autres caractéristiques… Si le CD n'est pas trouvé, l'utilisateur est invité à fournir lui-même les informations, afin que cela puisse servir à d'autres utilisateurs de la base de données, qui contient maintenant quasiment deux millions d'albums. Un monde sépare le cas de CDDB et le nôtre (type de données traitées, nombre d'utilisateurs, diffusion, CDDB vide à sa mise en service, complètement remplie par les utilisateurs). Il y a, malgré tout, dans cette démarche quelque chose à apprendre : en rendant les entrées dans la base extrêmement faciles (toute la partie concernant la "*signature*" du CD est traitée automatiquement, l'utilisateur doit juste se contenter de retranscrire les caractéristiques), ils ont réussi à rendre la base de données extrêmement complète. Par facile, on entend ici non-ambiguë : les informations entrées par l'utilisateur ne concernent que les données concrètes qui figurent sur les pochettes des CDs. Les deux millions de CDs qui figurent dans CDDB ont été entrés par les utilisateurs eux-mêmes. Dans notre cas, le processus sera nécessairement plus complexe, mais si l'on peut en automatiser une partie, et rendre le reste aussi peu ambigu que possible, on a de bonnes chances de créer une base évolutive, qui sera d'autant plus complète qu'elle aura été utilisée.

1.2.1.3. Recherche de textes dans la base

Comme dans toute autre base de données, le but principal est de pouvoir chercher et retrouver des données (ici des ressources textuelles destinées à l'enseignement des langues).

[5] Ces moyens plus "*traditionnels*"de rechercher un document à vocation pédagogique dans le cadre d'un cours de langue seront explicités plus tard.

Afin que le corpus soit utile, il faut que les requêtes soient faciles à formuler. Comme nous allons avoir beaucoup recours à cette notion par la suite, il convient de l'expliciter, et ce d'autant plus que "*facile*" est un terme passe partout relativement peu précis. Les requêtes doivent faire appel à des notions concrètes. La question, pour l'utilisateur (l'utilisatrice), doit être de savoir quelles caractéristiques il (ou elle) recherche dans son texte et non pas de savoir comment exprimer ces caractéristiques. Il y a nécessairement une période d'adaptation pour apprendre à utiliser une base de données quelle qu'elle soit, mais, une fois le produit pris en main, la formulation des requêtes ne doit pas souffrir de la présence de champs trop ambigus.

Pour que la base soit utile, il faudra aussi que le "*rappel*" soit aussi grand que possible. Le rappel désigne le rapport du nombre de documents pertinents fournis par rapport au nombre de documents pertinents dans la base [FLU 00]. En d'autres termes, si un texte correspond à la requête effectuée, il doit figurer parmi les résultats.

1.2.1.4. Utilisation des textes par les outils

Une fois trouvés, les textes doivent être utilisables par les outils TAL présents dans la plate-forme, ce qui suppose que toute annotation du texte soit interprétable par les outils ou à défaut n'interfère pas avec ces derniers. En outre, il faut que le texte contienne toutes les informations (annotations) nécessaires au bon fonctionnement des outils qui l'utiliseront.

Idéalement (flèche verte sur la figure 2), il pourrait y avoir un outil permettant d'étendre une activité à un texte ayant plus ou moins les mêmes caractéristiques que le texte employé, grâce à une requête faite automatiquement (ou plus vraisemblablement grâce à un ensemble de textes candidats déjà choisis par l'enseignant). Ceci permettrait en fonction des résultats d'un apprenant (historique) de prolonger son apprentissage, sans être obligé de repasser par la création d'une nouvelle activité. On ne peut pas s'intéresser au développement d'un tel outil tant que la base de textes n'est pas créée et ne fonctionne pas bien pour les autres cas d'utilisation.

1.2.2. Le sujet

1.2.2.1. Domaine

A l'origine le sujet de ce mémoire concernait la base de textes en général et donc les trois cas d'utilisation présentés ci-dessus. Après avoir travaillé sur le sujet, nous nous sommes aperçus qu'un tel sujet était bien trop vaste pour être traité en si peu de temps. En effet, le troisième cas d'utilisation (utilisation des textes par les outils) demandait l'établissement d'un

état de l'art des outils TAL qui pouvaient potentiellement être utilisés dans le projet. Cet état de l'art devait être extrêmement détaillé et présenter pour chaque outil, l'intégralité de ses API, en examinant les données d'entrée et de sortie. Une fois ce travail effectué, il aurait fallu étudier les différences et points communs, afin d'en déduire les informations qui devaient figurer dans les textes de la base (annotations). Il aurait fallu aussi s'intéresser à des outils de formatage qui auraient pris en compte les besoins des outils existants et anticipé ceux des outils qui seront amenés à exister dans le futur.

C'est à la lumière de l'anticipation du travail à effectuer que le sujet a été restreint à la recherche et à l'ajout de documents dans la base. Comme l'ajout va dépendre directement des champs définis pour la recherche, on traitera les deux aspects en même temps.

1.2.2.2. « Problématique »

Le présent travail a pour but de tenter de structurer la base de textes, c'est à dire définir les champs selon lesquels les requêtes pourront être effectuées.

Les enseignants, lorsqu'ils recherchent un texte pour leur cours, font appel à toute leur expertise et leur expérience du domaine afin d'évaluer l'intérêt pédagogique d'un texte. A force de répétition du processus de recherche, les critères qu'ils utilisent ne sont pas forcément conscients. Là se situe la grande différence avec la recherche dans une base de données. En effet, pour pouvoir effectuer une recherche dans une base de données, il faut être capable d'exprimer chacun des critères pouvant être pris en charge par la base dés le début de la recherche, les critères plus instinctifs n'intervenant que lors du choix parmi les documents candidats extraits par la base.

Il va falloir essayer de comprendre le plus possible de critères pédagogiques employés par les enseignants et de les formaliser de manière à ce qu'ils puissent être pris en charge par la base de données. On peut bien sûr anticiper un certain nombre de champs :

Ex : auteur, titre, etc.…

Mais une fois les champs définis, le travail n'est pas terminé, la création d'une base de données concerne aussi bien les valeurs qu'on pourra entrer dans ces champs que les champs eux-mêmes : nous savons, avant même le début du travail, que nous allons être confronté à des problèmes liés au fait que pour exprimer une même qualité, un même critère, plusieurs typologies existeront.

Il faudra donc trouver et expliciter les valeurs possibles pour que la base de données soit ordonnée de manière : complète et cohérente.

1.2.2.2.1. Complète

Il va falloir s'intéresser aux informations qui peuvent s'avérer utiles dans la recherche de documents. Le choix d'utiliser ou de ne pas utiliser tel ou tel trait d'un texte lors de la recherche, doit venir de l'utilisateur et non du système, qui ne devra pas être limitant dans ce domaine. En d'autres termes, tout ce qui sert au choix d'un document devra d'une manière ou d'une autre être représenté dans la base.

1.2.2.2.2. Cohérente

A un même texte devra correspondre une seule entrée (possibilité de passer des critères formels aux critères pédagogiques)

L'ajout d'un texte dans la base devra être non-ambigu : à un même texte devra correspondre un et un seul ensemble de valeurs de champs.

Si l'ajout de documents est ambigu, cela posera forcément des problèmes au niveau de la recherche : un texte peut ne pas être trouvé parce que l'on aura choisi une valeur parmi deux valeurs candidates ou si l'on ne peut pas faire la différence entre la portée de valeurs différentes pour un même champ, la formulation de la requête sera forcément rendue plus compliquée (d'autant plus que le public auquel se destine cette base, n'est pas un public de spécialistes de l'informatique).

Si l'on se réfère à l'exemple de base de données du paragraphe introduisant les principes généraux des bases de données (Base de données de CDs), le style musical est typiquement le genre de champ qui est ambigu : dés qu'un artiste a un style qui sort des stéréotypes, il sera extrêmement difficile de remplir le champ correspondant. Même un champ comme année peut s'avérer ambigu : parle-t-on de l'année de la première sortie de l'album, ou de l'année de la version présente dans la base ?

1.2.2.2.3. Difficulté

Lors de l'entrée de documents dans la base, une partie des informations pourra éventuellement être entrée automatiquement par des outils, mais il y en aura forcément une partie qui sera entrée par un agent humain. Pour pouvoir satisfaire le critère de cohérence de

la base, il faudra définir des directives suffisamment précises pour que rien (ou le moins possible) ne soit laissé uniquement à l'appréciation de l'utilisateur.

1.3. TERMINOLOGIE POUR LE RESTE DU MÉMOIRE

1.3.1. Texte :

Dans la suite de ce mémoire, nous verrons le texte comme un texte brut, ce qui signifie qu'il ne contiendra pas d'autres informations typographiques que les caractères de base et les retours à la ligne. On ne prendra en compte aucune information sur la police de caractère utilisée, la taille des interlignes, les marges ou autre. Il ne contiendra pas non plus de liens hypertextes ou d'images (à l'heure actuelle).

Contrairement à la forme, nous n'imposons aucune contrainte sur le contenu. Nous considérerons comme texte, tout ce qu'un enseignant peut avoir envie d'introduire dans la base. Le contenu de la base, incombera aux utilisateurs.

1.3.2. Opposition annotation / description :

Nous aurons aussi recours, principalement dans la partie technique de ce mémoire, aux notions d'annotation et de description, que nous opposerons. Les définitions que nous allons donner ne présupposent rien sur l'implémentation des dites annotations et descriptions. La différence concerne principalement ce qu'elles représentent.

Une annotation concernera une propriété d'une sous partie du texte, une propriété que l'on ne peut pas appliquer au texte dans son intégralité. Par exemple, si l'on décide d'étiqueter le texte en fonction des catégories grammaticales, ce sera une annotation, puisque l'intégralité du texte ne sera pas un nom ou un verbe ou autre.

La description concernera les caractéristiques générales du texte, comme son auteur, son année de parution ou dans la grande majorité des cas, sa langue.

On insiste bien sur le fait que même si les annotations semblent prédestinées à être entrées sous forme de balises dans le texte et les descriptions deviennent des champs de la base de données, on ne s'adressera au problème que dans la partie sur l'implémentation (ou la conception) de la base de textes.

1.3.3. Texte authentique / texte pédagogique

Un texte authentique est un texte écrit dans sa langue par un locuteur natif et dont le but est purement communicatif ; un tel texte pourra bien entendu être utilisé à des fins

pédagogiques mais il n'aura pas été écrit dans cette perspective. Au contraire un texte pédagogique aura été écrit par un locuteur natif ou non, avec comme objectif d'être utilisé dans une activité pédagogique. Cela ne signifie pas qu'un texte pédagogique n'aura pas de fonction communicationnelle mais que lors de son écriture, il était destiné à une utilisation dans le contexte de l'apprentissage de la langue dans laquelle il est écrit.

Comme mentionné dans le paragraphe sur la notion de texte, la question du contenu de la base, revient entièrement aux utilisateurs de la base, les deux types de textes pourront donc s'y retrouver.

2. LA DEMARCHE

La première partie du travail que nous avons effectué s'est articulée autour de la recherche d'informations sur les bases de textes existantes, après quoi nous verrons comment le contexte de recherche et les spécificités du produit nous mèneront à procéder à une série d'entretiens.

2.1. BASE DE TEXTES

Si l'on recherche sur Internet des bases de textes ou des corpus, on va pouvoir trouver de nombreuses initiatives ou travaux dans le domaine : on trouvera des corpus de diverses tailles dans différentes langues. Pourquoi devons-nous donc créer notre propre base ? Qu'est-ce qui différencie notre travail de l'existant ? Nous n'allons pas ici passer en revue tous les corpus existant, mais en citer certains et tenter d'expliquer ce qui les différencie de notre travail :

2.1.1. FranText

FranText est défini sur le site officiel de la manière suivante [FranText] :

> *« FRANTEXT peut se définir comme un vaste corpus, à dominante littéraire, constitué de textes français qui s'échelonnent du XVIe au XXe siècle. Sur l'intégralité du corpus, il est possible d'effectuer des recherches simples ou complexes (base non-catégorisée). Sur un sous-ensemble comportant des oeuvres en prose des XIXe et XXe siècles, les recherches peuvent en outre répondre à des critères syntaxiques (base catégorisée). »*

Les textes figurant dans FranText pourraient éventuellement être utilisés dans des applications d'enseignement du français (même s'ils seront probablement trop complexes pour les apprenants débutants), mais indépendamment des données qui y sont stockées, l'organisation même de la base ne conviendrait pas. Les recherches des concordances qui existent dans FranText pourraient être utiles, mais cette base de textes est destinée à des locuteurs français et n'est pas organisée pour satisfaire les besoins d'enseignants recherchant des textes pour leurs élèves. Les enseignants devraient s'adapter à la base et non le contraire (ce que nous cherchons à faire). FranText n'est pas une base à visées didactiques. Pour trouver des informations dans FranText, les enseignants seraient obligés de se satisfaire des méthodes de recherche offertes. Ces méthodes ne tiennent pas compte des aspects didactiques auxquels les enseignants ont recours pour choisir les textes qu'ils utilisent en classe.

2.1.2. FRIDA

FRIDA[6] [FRIDA] est une base de texte menée en parallèle du projet FreeText, qui comporte certaines analogies avec le projet MIRTO, mais qui se concentre sur la correction automatique des productions libres des étudiants. Pour y parvenir, le système utilisera la base FRIDA, qui contient un ensemble de textes produits par des apprenants du français (FRIDA est l 'équivalent français du ICLE[7] ; tous deux ont été créés dans le même laboratoire). Les apprenants du français dont les textes se retrouvent dans FRIDA, sont originaires de pays différents, ont des expériences d'apprentissage de langues différentes. Grâce à ce corpus, on espère pouvoir corriger automatiquement les fautes effectuées le plus fréquemment par les apprenants, en fonction de leur profil. La base de texte a ici un but pédagogique, mais a un contenu qui ne correspondrait pas à nos besoins, finalement très différent sur le problème de la base de textes (support pour les activités des apprenants pour MIRTO, et données pour la configuration d'outils de correction automatique pour FreeText).

2.1.3. TEI[8]

Les informations données dans ce paragraphe proviennent de The TEI: History, Goals and Future [I-S 95].

La Text Encoding Initiative, n'est pas à proprement parler, une base de textes, mais elle constitue une démarche dont nous pouvons nous inspirer. L'idée de base est venue de l'existence de multiples schémas de codage des textes, développés dans les années 60, 70 et 80. Chaque schéma, était conçu pour les besoins de ses créateurs, d'où une certaine incompatibilité entre les différentes normes. La TEI a donc été développée avec un souci d'unification, en se basant sur les "*principes de Poughkeepsie*" (du nom de la réunion où ils ont été définis). Le but de TEI était de créer un standard ou plutôt des "*lignes* directrices" de codage électronique pour les *« texts intended for humanities scholarship»* (textes pour les recherches en sciences humaines). Finalement, la portée du travail de la TEI ne s'est pas cantonnée aux recherches en sciences humaines, mais fut élargie au traitement du langage naturel, à la recherche d'information, à l'hypertexte et à la publication électronique. Pour chacun des domaines concernés, ils ont du jongler entre les différentes théories dissidentes, les spécificités de chaque type de texte ou de traits de textes (*alphabets, corpus, linguistique générale, dictionnaires, données terminologiques, textes parlés, hypermédia, prose littéraire,*

[6] **FR**ench **I**nterlanguage **DA**tabase
[7] International Corpus of Learner English
[8] Text Encoding Initiative

vers, tragédies, matériel historique, apparat critique) et même les différentes manières dont peuvent être perçu un texte :

- *« physical objects »* : (textes considérés sous l'angle du support : matière, époque…)

- *« typographic objects »* : (attention portée sur la mise en page, les polices de caractères utilisées…)

- *« linguistic objects »* : (le texte est vu comme une série de graphèmes ou de phonèmes, ou à un niveau d'abstraction plus élevé comme des composants lexicaux ou des phrases, tout dépendra de la théorie employée)

- *« formal objects »* : (on s'intéressera ici à la structure hiérarchique du texte)

- *« rhetorical objects »* : (séries ou hiérarchies d'actes de parole, figures rhétoriques, tropes)

- *« propositional objects »* : (se référant à des personnes, des choses, des endroits et des événements spécifiques, réels ou imaginaires)

- *« historical and cultural objects »* : (dans lesquels on ne perd pas de vue les différents témoignages de la manière dont le texte a été transmis, interprété, ré-interprété, commenté)

Le travail que nous devons effectuer, est évidemment beaucoup moins vaste et possède certaines spécificités (comme le fait de s'inscrire dans le cadre d'une base de données), mais la démarche suivie par TEI peut nous être utile. En effet, une grande partie de leur travail a concerné la manière de faire utiliser leur notation, ils ont donc du s'intéresser à comment l'expliciter, pour que les différents utilisateurs potentiels puissent la prendre en main. C'est évidemment un aspect non négligeable de notre étude.

Il y a donc certains liens (plus ou moins distants) qui existent entre notre travail et celui de la TEI, nous pouvons regretter qu'ils n'aient pas perçu les textes comme des "*educational objects*" (objets pédagogiques), ce qui nous aurait beaucoup facilité la tâche.

2.2. CONTEXTE DE RECHERCHE

Les bases de textes les plus proches de celle que nous devons réaliser sont celles que nous avons mentionnées dans le paragraphe sur les bases de textes. Ceci constitue une difficulté majeure à laquelle nous avons été confronté au cours de la réalisation de ce mémoire : il n'y a que très peu de références bibliographiques sur le domaine qui nous intéresse. Nous avons donc passé beaucoup de temps à chercher dans toutes les directions

possibles sans trouver d'information qui puisse réellement nous orienter dans nos recherches. Le travail que nous avons donc du fournir ne pouvait apporter des informations définitives sur le sujet, c'est réellement un travail de défrichage qui est effectué ici. Le but de cette étude est de pouvoir servir de base à des recherches plus approfondies dans le domaine. C'est pourquoi, il a été décidé d'orienter les recherches vers les attentes des utilisateurs potentiels.

2.3. UN PRODUIT POUR QUELS UTILISATEURS

Comme tout système informatique, le produit du projet MIRTO est destiné à une certaine classe d'utilisateurs. Il s'agit ici d'enseignants et de chercheurs dans le domaine des langues, c'est-à-dire un public qui n'est pas nécessairement expérimenté en ce qui concerne l'utilisation d'un micro-ordinateur et encore moins en matière de bases de données. Pour adapter au mieux la base de texte à ses utilisateurs potentiels, il s'agira, d'une part, d'avoir une interface claire et ergonomique (ce qui sort du cadre de ce mémoire), mais aussi de concevoir la base de données de manière à ce que les champs soient, comme nous l'avons mentionné précédemment, explicites. Pour ce faire, nous allons tenter de savoir comment les utilisateurs potentiels de cette base, procèdent afin de la modéliser en conséquences.

2.4. LES ENTRETIENS

Afin de mieux connaître le processus de recherche de textes à finalité didactique et de s'assurer que notre base de ressources textuelles soit ainsi adaptée aux futurs utilisateurs, il a été décidé de procéder à une série d'entretiens avec des enseignants. Grâce aux données recueillies au cours de ces entretiens nous espérons pouvoir formaliser une partie de leur processus de recherche de textes.

2.4.1. Stratégie

Beaucoup d'enseignants ne maîtrisent pas forcément l'utilisation de l'informatique. Les entretiens n'ont donc pas nécessairement lieu uniquement avec des enseignants ayant des notions de bases de données ou de TAL, puisque nous nous intéressons ici à leurs connaissances dans le cadre de l'enseignement des langues en général et non uniquement en ce qui concerne notre sujet.

Pour ne pas biaiser les entretiens, pour nous assurer que les enseignants raisonnent dans le cadre du cours présentiel, c'est-à-dire, de leur travail de tous les jours, nous ne mentionnerons pas le TAL dans les questions, au moins dans un premier temps. Il est très important de s'assurer ainsi que l' "*interviewé*" ne s'autocensure pas. Si le TAL est mentionné

immédiatement, nous pouvons nous attendre à ce que la personne avec laquelle nous nous entretenons, ne voyant pas de lien direct avec son travail dans le contexte de la classe, soit déstabilisée et tente d'anticiper ce qui est important pour nous et ce qui ne l'est pas. Or, il est peu probable que l'interviewé sera à même d'en juger, nous risquons donc, le cas échéant d'être privé d'informations potentiellement importantes.

Nous parlerons donc des pratiques en classes et de la manière de procéder des enseignants dans ce contexte, après quoi nous essaierons de lier ces pratiques au processus de recherche dans le but de formaliser les informations qu'ils nous fourniront. Eventuellement, si le besoin s'en fait ressentir, ou avec les enseignants ayant de réelles connaissances dans le domaine, nous pourrons poser des questions plus directes afin de préciser les informations recueillies. Bien sûr, cela ne signifie pas que nous n'expliquerons pas le but de la manœuvre aux enseignants, nous signalons juste, qu'il ne nous paraît pas judicieux d'orienter ostensiblement et immédiatement la discussion dans la direction du TAL.

De plus, il convient de rappeler que le travail que nous effectuons est une étude préliminaire qui vise à déterminer les grandes tendances avant un sondage plus précis (pour lequel un prototype serait une aide).

2.4.2. Les « *sujets* »

Afin d'avoir une couverture la plus large possible, nous avons tenté de nous adresser à tous les enseignants pouvant être concernés par un tel projet, c'est à dire tous.

Nous avons donc essayé de diversifier (en fonction des disponibilités des professeurs avec lesquels nous sommes entré en contact) : les langues enseignées, le public (à qui ils enseignent), le niveau d'enseignement (le niveau de ceux à qui ils enseignent), les relations des « *sujets* » avec les nouvelles technologies (de la technophobie, à la bonne connaissance de l'outil informatique dans un contexte d'enseignement des langues). Avec ces objectifs nous avons donc effectué huit entretiens répartis comme suit :

NOM Prénom	Langue	Niveau	Etablissement	Expérience[9]	TICE[10]
BÉATRIX Myriam	Anglais	Collège (LV1+LV2)	Collège de Bissy (73)	23	R/E
BOURGADE Sophie	Anglais	Bac+1 → Bac+4 (non spécialistes)	Université de Savoie (73)	7	C
CELIŃSKA Adriana	Polonais	Débutant (adultes)	Amicale du Dauphiné Dom Polski	1	P
GALOPPO Maria-Elena	Espagnol	Débutant, Conversation (adultes), Débutant (professionnel, apprentissage rapide) Avancé (bac+*, non spécialistes)	Université Inter-Age du Dauphiné, ESC, Université Stendhal, LOGOS (38)	5	R
HENDERSON Alice	Anglais	Bac+1 → Bac+5 (spécialistes) Bac+1 → Bac+2 (non-spécialistes)	Université de Savoie (73)	12	R/E
PUCHALSKA Iwona	Polonais	Débutants Faux-Débutants 2nde langue (niv 1) 2nde langue (niv 2) Bilingues	Université Stendhal (38)	2	R
SAINTY Michel	Anglais	Collège (LV1+LV2)	Collège de Bissy (73)	30	R
TENDERO Sonia	Espagnol	Collège (LV2)	Collège de Bissy (73)	5	P

La colonne niveau correspond à l'intitulé du cours, tel qu'il est défini par l'établissement dont dépend l'enseignant.

Remarquons que si l'outil informatique semble être largement entré dans les pratiques personnelles des enseignants, notamment pour la recherche de leurs supports de cours, il n'est que peu utilisé dans le contexte même de la classe.

[9] Nombre d'années d'enseignement

[10] Degré de connaissance des Technologies de l'Information et de la Communication pour l'Enseignement :
Ø → reticent(e) à utiliser l'outil informatique
P → utilisation de l'outil informatique à titre personnel
R → utilisation de l'outil informatique dans la recherche de matériel pédagogique
E → utilisation de l'outil informatique dans le cadre de l'enseignement
C → conception d'activités pédagogiques avec support informatique

2.4.3. Déroulement des entretiens

2.4.3.1. Enregistrement

Les entretiens sont enregistrés selon la séquence suivante :

Micro stéréo sony ECM-909 < convertisseur Analogique/Numérique Aiwa HDA-1 < Enregistreur DAT Aiwa HD-S1 + DAT sony < carte son Trident 4D wav(NX) PCI Audio < .WAV<.SHN

Les enregistrements, sont convertis au format shn, qui est un format audio compressé, sans perte de données (contrairement au format mp3), un logiciel de compression/décompression peut être téléchargé à http://etree.org/mkw.html.

Pour ce qui est de l'édition des fichiers sons (découpage, réduction du souffle), on utilise la version "*demo*" de *Cooledit pro 2.1.*, qui peut être téléchargée à http://ftp.syntrillium.com/pub/cep/cepsetup.exe.

2.4.3.2. Prise de notes

Même si les entretiens sont enregistrés, des notes seront prises au fur et à mesure, au moyen de fiches créées à cet effet. Les fiches sont disponibles en annexe 1 et 2. Nous expliciterons à quoi correspondent les différents tableaux en même temps que nous expliquerons les différentes données que nous allons tenter de recueillir.

La première fiche contient un tableau "*généralités*", qui va nous permettre de noter les éléments permettant de mettre en contexte les informations relevées par la suite et de dater l'interview. On y trouvera donc l'identité de l'enseignant, la date de l'entretien, la langue enseignée, à quel(s) niveau(x) et depuis combien de temps. Ceci va nous permettre de savoir s'il y a des spécificités liées au niveau auquel on enseigne une langue (même langue enseignée niveau différent), à la langue elle-même ou même au public auquel l'enseignement est adressé (adultes, adolescents, formation continue, cursus obligatoire…).

2.4.4. Contenu

Bien que les entretiens soient menés d'une traite, ils s'articulent autour de trois phases. Chacune de ces phases concerne un domaine précis : le premier concerne les pratiques en classe, c'est à dire la manière dont l'enseignant s'appuie sur les textes au cours de son enseignement. Nous faisons donc attention à ne négliger aucun aspect (par exemple les phrases d'énoncé d'un exercice peuvent être considérées comme des textes). Nous essayons

ensuite de savoir en quels termes et sur quels critères l'enseignant s'adresse au niveau de ses élèves. Et nous terminons en le questionnant sur les mécanismes mis en œuvre lors de la recherche de documents pour sa classe en fonction de ses objectifs pédagogiques. Si les connaissances de l'enseignant dans le domaine le permettent, nous lui demanderons aussi, comment il utilise les « nouvelles technologies » dans le cadre de son travail. Nous espérons pouvoir extraire une typologie de textes à travers ces différentes phases de l'entretien.

Pour chaque phase, nous avons pris soin de nous documenter sur le sujet.

3. PRÉPARATION DES ENTRETIENS

Comme nous l'avons dit précédemment, les entretiens s'articulent autour de trois points clés pour lesquels une phase de recherche de l'existant s'impose.

3.1. LE NIVEAU D'ENSEIGNEMENT

Dans le contexte de la base de ressources textuelles, il paraît inévitable d'avoir un ou plusieurs champs concernant le niveau de langue des apprenants auxquels sera destiné chaque texte. Il s'agit ici de bien choisir le formalisme utilisé, en effet il existe de nombreux moyens d'évaluer le niveau d'un apprenant, il va donc falloir en choisir un qui permette non seulement de décrire de manière appropriée le niveau des étudiants, mais aussi une utilisation qui soit le plus instinctive possible pour les enseignants (aussi bien au niveau de la recherche de documents que de l'entrée de ceux-ci dans la base, ce qui doit être accessible à tous les utilisateurs du système).

3.1.1. Exemples d'Evaluation du niveau

3.1.1.1. TOEIC

Le TOEIC est un exemple parmi tant d'autres de certifications du niveau en langue étrangère (l'anglais ici). Nous n'allons pas nous intéresser à sa validité ni même au processus d'évaluation mais plutôt à la manière de décrire le niveau de l'apprenant.

3.1.1.1.1. Une notation selon un axe déterminé

Comme l'indique son nom, le TOEIC (Test Of English In Communication), a pour vocation d'évaluer les compétences des sujets dans le domaine de la communication, en particulier au sein d'une entreprise. L'évaluation sera effectuée selon deux axes : l'oral et l'écrit. En fonction du score selon chaque axe, la « *can do levels table* » (disponible en annexe 3), donne les compétences du sujet évalué. Par exemple à partir du score, on peut d'après cette table savoir le type de texte que le sujet sera à même de lire et de comprendre ou le type de production qu'il sera capable de mener à bien. On ne s'intéresse pas aux compétences grammaticales, en tous cas pas directement, l'évaluation s'articule autour de compétences concrètes de communication dans le cadre de l'entreprise.

Au contraire, dans Ven (libro del alumno) [C-M-M-R 00], un manuel d'espagnol, se focalisera sur des compétences plus grammaticales (recensées dans un index pour chaque leçon).

3.1.1.1.2. Granularité

Les scores du TOEIC sont compris entre 10 et 990 et évoluent de 5 en 5, soit 197 scores globaux possibles. En outre, ces scores peuvent être décomposés en une note orale et une note écrite augmentant encore les possibilités. Bien entendu, cette granularité est trop fine pour être réellement représentative d'une différence de niveau pour des notes très proches : si deux sujets ont des notes qui diffèrent de 5 points (l'écart minimum), il sera impossible de dire juste à partir de cette donnée qui a le meilleur niveau dans les faits. D'où le découpage, dans la table des « *can do levels* » des compétences en six niveaux à l'écrit et six niveaux à l'oral. Dans cette table, on constate aussi l'existence de six niveaux globaux (somme des points de l'écrit et de l'oral). Ces six niveaux (0/0+ (10-250 points) | 1 (255-400 points) | 1+ (405-600 points) | 2 (605-780 points) | 2+ (785-900 points) | 3/3+ (905-990 points)), permettent donc d'évaluer globalement, le niveau de quelqu'un, dans le domaine de la communication en anglais au sein d'une entreprise. Cependant sur le site français du TOEIC on nous explique :

> *« Le certificat TOEIC existe en cinq couleurs, correspondant à des plages de résultats distinctes : Orange (010-215), Marron (220-465), vert (470-725), bleu (730-855) et or (860-990). »*

Or ces plages ne correspondent pas aux six niveaux énoncés ci-dessus, ce qui peut poser problème dans la mesure où les cinq niveaux de certificats, ne correspondent pas aux niveaux définis par l'organisation même en charge du TOEIC.

3.1.1.2. « La classe de langue »

Pour les activités qu'elle propose dans son ouvrage [TAG 94], Christine Tagliante spécifie le niveau des apprenants pour lesquels la pratique peut être effectuée selon la typologie suivante : débutant / moyen / avancé. C'est une typologie que l'on rencontrera aussi pour des applications multimédia d'auto-apprentissage des langues étrangères.

Le problème qui peut être posé par ce genre de typologie globale est qu'on ne sait pas réellement ce qui est évalué, en effet quelqu'un ayant un certain niveau en ce qui concerne la grammaire, n'aura pas nécessairement le même en ce qui concerne le vocabulaire ou les facultés à communiquer.

3.1.2. Multilinguisme de la base

La base de ressources textuelles dont on va s'occuper devra être multilingue en ce sens que l'on devra y faire figurer des ressources textuelles pour toutes les langues prises en charge par la plate-forme. Il sera donc très difficile de définir des critères avec la même granularité

que ce que l'on a vu dans le TOEIC (notion de compétence). En effet, s'il y a un axe d'évaluation grammatical, pour pouvoir présenter une grille complète des connaissances grammaticales d'un apprenant, il faudrait que les champs proposés soient directement dépendants de la langue. En effet, si certaines compétences se retrouveront plus ou moins dans toutes les langues (ex : « expression du passé »), d'autres dépendront directement de la langue utilisée (ex : « maîtrise du présent / présent progressif », qui pourra être une compétence en anglais ou en espagnol, mais pas en français ou « maîtrise du datif » qui ne se retrouvera que dans les langues à déclinaison comme l'allemand ou le polonais mais pas dans les langues romanes ou l'anglais).

La précision de la base s'opposera donc à son évolutivité : plus on décrira le niveau pour lequel est prévu telle ou telle ressource textuelle avec précision, plus l'expression du niveau sera dépendante de la langue concernée et plus il sera difficile d'ajouter une langue à la plate-forme.

Tous ces problèmes doivent donc être gardés à l'esprit lors des entretiens, qui devront y apporter une réponse.

3.1.3. Position du problème à des enseignants chevronnés (dans diverses langues)

3.1.3.1. Préparation des entretiens

Ce qui nous intéresse dans le cadre de ce travail n'est pas nécessairement de trouver le moyen le plus précis de décrire le niveau, mais bien d'utiliser le moyen qui sera le plus parlant pour le plus de professeurs et qui leur sera le plus utile dans la recherche de ressources textuelles. En effet, on peut s'attendre à ce que le niveau soit un champ utilisé pour les recherches, puisque lorsqu'un enseignant prépare une activité, c'est forcément pour un public défini. Pour que le texte choisi soit adapté au public, il faut s'assurer qu'il soit en adéquation avec le niveau de connaissance de la langue de ce dernier.

Pour que ce champ soit adapté, il faudra qu'il ne soit pas trop complexe à manipuler, ni pour l'ajout de ressources dans la base, ni pour leur recherche. Au cours des entretiens, nous allons nous inspirer des types de notation exposés ci-dessus, en ce sens que, nous allons d'abord nous intéresser aux axes selon lesquels les professeurs souhaitent pouvoir évaluer le niveau d'un texte (ou le niveau des apprenants qui pourront lui être confronté). Ensuite, il faudra essayer de voir avec quelle précision il est souhaitable d'évaluer les niveaux selon les axes prédéfinis : si quelques tranches de niveaux qui pourront poser un problème à cause des

limites assez floues qui peuvent exister entre deux niveaux (cf. TOEIC : opposition 6 "*tranches*" de niveaux / 5 niveaux de diplôme) seront suffisantes ou si au contraire nous essaierons d'être plus spécifiques en s'intéressant aux compétences requises pour la compréhension du texte. Ce qui peut également poser problème, puisqu'il faudra être exhaustif, aussi bien dans les choix de compétences possibles (difficulté pour la conception de la base / problème du multilinguisme), que lors de l'ajout d'un document dans la base (difficulté pour l'utilisateur).

Nous pouvons anticiper le fait que, dans le premier cas, la recherche d'un document selon le critère du niveau risque d'être affecté par du "*bruit*" (trop de résultats, pas tous pertinents) [FLU 00], alors que dans le deuxième nous risquons plutôt d'avoir un problème de "*silence*" (tous les résultats sont pertinents, mais certains résultats pertinents ont été oubliés) [FLU 00].

3.1.3.1.1. Fiche

La fiche correspondant à la manière des enseignants d'évaluer le niveau, se trouve en Annexe 1 (c'est la partie du tableau située sous les généralités). Nous tenterons tout d'abord de connaître les axes d'évaluations pour les enseignants, ensuite dans chaque axe, nous essaierons de savoir si l'évaluation est faite en termes de compétences ou non. Si c'est le cas nous reporterons toutes les compétences. Enfin nous nous pencherons sur la granularité de l'évaluation, c'est à dire l'intervalle dans lequel seront comprises les valeurs. La granularité pourra aussi bien s'adresser à chacune des compétences qu'au niveau général de l'apprenant (selon un axe d'évaluation).

3.2. LES EMPLOIS ACTUELS DES RESSOURCES TEXTUELLES EN CLASSE

3.2.1. Hypothèse de travail

Re-précisons que le but de ce travail est d'utiliser dans la base de textes des champs aussi naturels que possible pour que les enseignants puissent l'utiliser intuitivement. Nous voulons que l'outil soit adapté à leurs pratiques. Dans ce contexte, il nous a été nécessaire de formuler une hypothèse : la recherche d'un texte par un enseignant sera conditionnée par ce qu'il veut faire du texte. En effet, on peut s'attendre à établir une sorte de typologie de texte en fonction de la manière de l'utiliser.

Pour tenter de confirmer ou infirmer l'hypothèse précédente, nous allons devoir recenser les différentes activités qui peuvent découler de l'utilisation d'un texte dans le contexte de la

classe de langue étrangère ou seconde. Cela nous permettra d'avoir une idée de typologie provisoire (dans le sens de l'hypothèse) sur laquelle nous pourrons baser nos entretiens.

3.2.2. Etat de l'art

3.2.2.1. Avant propos

Dans cet état de l'art, nous allons essayer de couvrir aussi largement que possible les pratiques s'appuyant sur un support textuel dans le contexte de la classe de langue. Cependant, cette étude ne pourra bien évidemment pas être exhaustive. Son véritable propos étant la préparation des entretiens avec les enseignants, elle nous permettra en effet d'anticiper certaines pratiques des enseignants, diminuant ainsi les explications quant aux pratiques elles-mêmes, nous permettant de nous concentrer sur les autres informations (processus de recherche des documents). Il faut bien noter ici, qu'en aucun cas nous n'allons émettre des critiques sur la valeur pédagogique des pratiques exposées, mais que nous allons seulement nous renseigner sur l'existant dans le domaine (ou ce que nous avons pu en voir).

Dans le contexte du projet MIRTO, nous n'exploiterons bien entendu pas toutes les pratiques ci-dessous ; du fait des limitations liées à l'utilisation de l'informatique, nous parlerons très peu, de toute pratique qui nécessite une réelle interaction entre les apprenants ou apprenants - enseignant.

Il convient en outre de signaler les difficultés auxquelles nous avons été confronté au cours de cette recherche. Il nous a été impossible de trouver des documents traitant uniquement de l'utilisation des textes en classe de langue, celle-ci étant plus généralement évoquée au détour d'un paragraphe. C'est finalement dans les ouvrages les plus généralistes sur l'enseignement des langues étrangères que nous avons trouvé les meilleures informations.

3.2.2.2. Activités de compréhension d'un texte écrit

> *« Todos los textos se pueden trabajar con un objetivo lingüístico, es decir, utilizarse para presentar o practicar lenguaje : bien sea pronunciación, ortografía, vocabulario o estructuas grammaticales. Pero también los textos pueden ser trabajados sin ninguno de estos objetivos lingüísticos, sino con un objetivo de comprensión y, de forma indirecta, de adquisición del lenguaje. »* [ALO 94]
>
> *(A partir de n'importe quel texte, il est possible d'effectuer un travail à objectif linguistique, c'est à dire utiliser un texte pour présenter ou pratiquer la langue : que ce soit la prononciation, l'orthographe, le vocabulaire ou les structures grammaticales. Mais les textes peuvent aussi être utilisés sans aucun de ces objectifs linguistiques, mais plutôt avec un*

objectif de compréhension, et de manière indirecte d'acquisition du langage.)

C'est sur cette dernière utilisation des textes (avec objectif de compréhension) que nous allons nous concentrer dans ce paragraphe. Même si la limite avec les utilisations plus purement « linguistiques » est parfois ténue, nous nous efforcerons de rester dans ce contexte, sans rentrer forcément dans le détail des cinq stratégies de lecture que, d'après Christine Tagliante [TAG 94], l'on cherche à faire adopter aux élèves à travers les différents exercices :

- le repérage : recherche d'informations précises et ponctuelles, le repérage est utilisé pour la compréhension globale d'un texte écrit.
- l'écrémage : pour aller à l'essentiel, recherche des mots-clés significatif de ce qui est important, intéressant et / ou nouveau. L'écrémage est d'avantage utilisé pour la compréhension détaillée d'un texte écrit.
- le survol : compréhension globale d'un texte long, en en dégageant l'idée directrice et l'enchaînement des idées, et en en sélectionnant des passages intéressants.
- l'approfondissement : réflexion, analyse détaillée, mémorisation ; pour la compréhension de l'implicite dans un texte écrit.
- la lecture de loisir.

On retrouve aussi des aspects de cette classification chez Alonso [ALO 94], ainsi que chez Cuq et Gruca [C-G 02], la terminologie n'est pas nécessairement la même mais certains aspects reviennent. Le repérage sera, par exemple, pour Cuq et Gruca [C-G 02], le balayage. On peut également noter chez ces derniers l'existence d'une lecture intensive ou studieuse dont le but est de retenir le plus d'informations possible, qui, bien que faisant intervenir une forme de mémorisation, ne semble pas coïncider exactement avec la notion d'approfondissement de Tagliante [TAG 94].

Les activités suivantes de compréhension, ne sont pas mentionnées dans un ordre particulier.

3.2.2.2.1. Inférence du sens d'un mot d'après le contexte

Le but de cette activité est de montrer qu'il n'est pas nécessaire de connaître le sens de tous les mots d'un texte pour en comprendre le sens. L'activité vient sous forme d'exercice

lacunaire (dont nous verrons par la suite qu'il peut avoir beaucoup d'autres utilisations) : on donnera un texte inconnu duquel on a retiré toutes les occurrences d'un unique mot clé.

Par exemple : *Il était une fois une paire de _____ qui étaient mariées ensemble. La _____ droite, qui était le monsieur, s'appelait Nicolas, et la _____ gauche, qui était une dame, s'appelait Tina. Elles habitaient dans une belle boîte en carton, où elles étaient roulées dans du papier de soie. Mais voilà qu'un beau matin une vendeuse les sortit de leur boîte afin de les passer au pied d'une cliente…*

Cette activité s'adressera plus particulièrement à des débutants. [D-G 90]

3.2.2.2.2. Lecture rapide

Ces activités s'adressent directement aux stratégies de lecture évoquées ci-dessus, en particulier la lecture « repérage ». Pour ce faire García Hernandez suggère, d'une part, une mise en page favorisant la segmentation et mettant en relief les structures grammaticales pour faciliter la visualisation :

Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres
•
•
•
Il les perdait toutes de la même façon […]
•
•

[GAR 90]

D'autre part, de poser des questions concernant des mots ou propositions clés auxquelles il faudra répondre en un temps minimum [GAR 90]. Cette activité est aussi proposée par Christine Tagliante [TAG 94] à ceci près qu'elle spécifie le type de données à faire chercher (chiffres, dates, noms propres, etc. …).

3.2.2.2.3. Questions de compréhension globale

Par ce terme peu rigoureux, on entend les questions qui concernent le thème du texte, l'idée centrale. C'est le type de question qui peut être posée après la lecture du texte. En effet, les questions ne porteront pas sur un aspect précis, on ne demande pas à l'apprenant d'avoir compris le texte dans les moindres détails mais seulement d'avoir une idée de ce dont on parle.

Dans Formas de trabajo en el aula[LÓP 90], on suggère d'utiliser un exercice lacunaire, créé à partir d'un résumé du texte. Ce procédé est évoqué aussi chez Marcus [MAR 99], à ceci près que dans le premier cas le résumé est extrêmement simple mais les mots "*retirés*" ne

figurent nulle part, alors que dans le deuxième cas, le résumé est légèrement plus complexe, mais l'apprenant est aidé par la liste des mots effacés présentés dans le désordre.

López Hernandez suggère également de poser des questions portant sur le thème du texte et dont on modifiera la forme en fonction du niveau de langue de l'apprenant. Nous résumons cette évolution par le graphique ci-dessous :

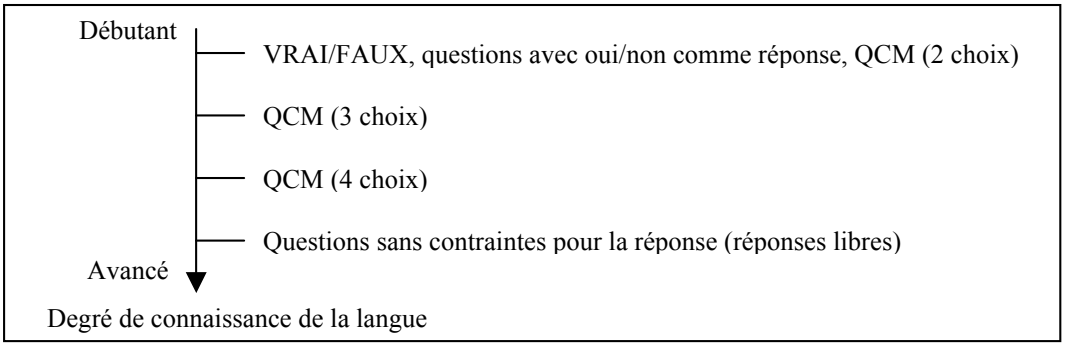


Figure 2 Degré de liberté dans les réponses par rapport au degré de connaissance de la langue

Mieux l'apprenant connaît la langue, plus on lui laisse de latitude dans le choix des réponses.

Quand l'apprenant est un grand débutant, la compréhension même des questions est un exercice en soi.

Encina Alonso suggère d'autres activités dans ce domaine comme donner un titre au texte ou en faire un résumé en un nombre de mots prédéterminé. Elle suggère aussi un autre exercice qui consiste à remettre dans l'ordre des textes qui ont été découpés en plusieurs parties ensuite mélangées [ALO 94]. On retrouve cette activité chez López Hernandez [LÓP 90].

3.2.2.2.4. Textes littéraires

López Hernandez [LÓP 90] suggère pour les textes littéraires les pratiques suivantes :

- Division en parties avec sous-titres.
- Justification du titre à partir de mots ou phrases pris dans le texte.
- Mise en relation avec d'autres textes lus préalablement.

Ces activités sont aussi évoquées par Tagliante [TAG 94] dans le cadre de la lecture écrémage. Nous pouvons remarquer que, comme le suggère cette dernière, ce type d'activité pourrait aussi bien être mis en œuvre pour des articles de presse que pour des recueils de textes courts.

Dans Français langue seconde, 36 lectures pour les collèges [MAR 99], Catherine Marcus propose, pour la construction du sens, l'élaboration de "*cartes d'identité*" de personnages, qui contiennent un certain nombre de champs.

Ex : Nom / Prénom / Age approximatif / Lieu de naissance / Lieu d'habitation / Profession / amis …

Ce genre d'activité concernera plutôt des œuvres entières ou de longs textes.

3.2.2.2.5. Extraction d'informations détaillées et/ou sélectives à partir d'un texte

Alonso suggère, dans son ouvrage [ALO 94], beaucoup d'activités que nous ne détailleront pas. Bien qu'applicable en classe, ces activités ne fonctionneront pas dans le cadre de notre travail : elles font appel à des dessins (à mettre dans l'ordre, à apparier ou comparer avec un texte), à des pratiques gestuelles (mimes) à du travail de groupe, préalable à la lecture ou à l'écriture de textes.

En revanche, les exercices de questions ou lacunaires paraissent possibles à mettre en place. Les exercices lacunaires concerneront des textes dont on aura ôté plusieurs mots-clés. La différence avec l'exercice proposé dans le cadre de l'inférence est que plusieurs mots-clés différents ont été ôtés mais que toutes les occurrences d'un même mot (ou mot de la même racine) ne seront pas forcément effacées. Nous n'entrerons pas plus dans le détail de cet exercice puisque, d'après l'exemple donné, il ne fait pas appel uniquement à la compréhension du texte mais aussi à un travail linguistique :

Titre du texte : Sentarse con salud
La sociedad actual se ha vuelto sedentaria. Las personas pasan demasiado tiempo _____ y muchas veces parecen dolores del cuello, hombros y espalda.
→ sentadas (compréhension, puisque employé auparavant, mais fait appel à des notions de grammaire puisqu'il aura fallu prendre le participe passé et le mettre en relation de part sa morphologie avec le SN auquel il se raporte : "las personas")
[…] Ya no _____ que mirar solo la estética o el precio […]
→ hay (hay que : formule pour indiquer l'obligation). [ALO 94]

Nous reviendrons donc sur ce type d'exercice par la suite, pour en exprimer les autres buts possibles.

En ce qui concerne les questions, elles seront très pointues (ce qui les différencie de celles pour la compréhension globale) et à ce titre devraient être données avant la lecture afin de mettre en évidence le travail demandé (écrémage).

Dans une optique d'extraction des idées principales d'un texte, on peut aussi demander aux apprenants d'indiquer les parties du texte qui correspondent à différentes thématiques, en les reportant dans un tableau :

« MY FAVOURITE CLASS
Name : Anne Jackson
I like school. Maths is my favourite class. I love numbers because they are fun. History is interesting too: English history, the Vikings, pirates, captain Drake, Queen Elizabeth.
I do not like sports very much. Judo is difficult and swimming is boring. I **never** play football!
My friends like sports, but they do not like school. Mr. Chips, the math teacher, is always angry because they never do any homework. »

Dans Formas de trabajo en el aula[LÓP 90], les thématiques pour le texte ci-dessus sont : "*personal, physical, routines, hobbies, abilities".*

3.2.2.3. Compréhension de la structure du texte

La compréhension d'un texte ne sera pas totale si l'on s'en tient uniquement aux éléments du paragraphe précédent. En effet, nous nous sommes appliqué à ne voir que les éléments de compréhension pour ainsi dire pure, c'est à dire qui ne portent que de manière indirecte sur l'acquisition du langage. Il vient un moment où il est utile d'expliciter les phénomènes mis en œuvre dans les textes, que ce soit pour arriver à une compréhension plus profonde ou pour formaliser un peu plus certaines connaissances acquises.

> *« Podríamos afirmar, sin duda, que a escribir mejor se aprende leyendo más y no escribiendo más únicamente. […] Durante todo el proceso de aprendizaje de la escritura es de gran ayuda utilizar una serie de técnicas que lleven al alumno a observar y experimentar primero para luego escribir. » [D-G 90]*
>
> *(On pourrait affirmer, sans aucun doute, que l'on apprend à écrire mieux en lisant plus et pas seulement en écrivant plus.[…] Durant tout le processus d'apprentissage de l'écriture, l'utilisation d'une série de techniques qui amènent l'apprenant à observer et expérimenter tout d'abord, pour plus tard écrire, constitue une grande aide.)*

Ce qui signifie en substance que la lecture a un rôle des plus importants dans l'apprentissage de l'écriture et de ses techniques, que ne peuvent assumer à elles seules les activités de production écrite. Les activités décrites dans ce paragraphe viseront donc à

permettre à l'apprenant une meilleure compréhension du texte et de sa structure en vue de passer à la phase de rédaction par la suite.

3.2.2.3.1. Registre et style

Del Mar Martín Viaño et Gómez Casañ proposent aux élèves d'analyser et de contraster différents types de textes, afin de mettre en évidence le vocabulaire, les expressions et les conventions qui apparaissent dans les textes. On pourra par exemple les amener à comparer deux lettres l'une formelle, l'autre informelle au moyen du tableau suivant [D-G 90] :

Aspects analysés	Lettre formelle	Lettre informelle
En-tête		
Salutation initiale		
Salutation finale		
Relation rédacteur/lecteur		
Phrases indiquant la relation		

Dans La comprensión lectora, on s'intéresse à la fonction remplie par les pronoms en terme de cohésion du texte. Pour ce faire, on propose un texte dans lequel les substantifs sont constamment répétés sans qu'on ne leur substitue aucun pronom. L'apprenant qui observe cette répétition réalise qu'elle n'est pas nécessaire et corrige le texte, lui donnant une forme plus appropriée. Cette activité s'adresse à une notion grammaticale, mais peut donner lieu aussi à une approche stylistique (lourdeur des répétitions).

3.2.2.3.2. La ponctuation

La ponctuation influe sur la compréhension [GAR 90]. Un élève capable de comprendre les nuances apportées à un texte par la ponctuation sera probablement plus à même de l'utiliser à bon escient dans ses propres productions. Les exercices proposés dans la expresión escrita [GAR 90] vont permettre à l'élève de mettre en évidence sa compréhension d'un texte (pour les deux exercices) puis de mettre en œuvre ses compétences en terme d'utilisation de la ponctuation dans le contexte de la production (cf. exercice 1) :

- Exercice 1 : Un texte non ponctué est donné ; l'apprenant doit retrouver la ponctuation correcte.
- Exercice 2 : Deux textes identiques, à la ponctuation près, sont donnés ; l'apprenant doit les différencier.

3.2.2.3.3. A partir d'une chanson (le texte sans écoute)

Tagliante [TAG 94] décrit les activités qui peuvent être mises en place avant l'écoute d'une chanson. Etant donné le fait qu'elles ne font pas intervenir le son, elles pourront pour

certaines être implémentées directement dans un projet tel MIRTO. Notons que le traitement de l'oral est prévu à court terme pour ce projet. La description de ces activités n'est donc pas dénuée de tout rapport avec ce projet.

La première activité consiste en un exercice de "remise dans l'ordre" des phrases de la chanson. Il est assez proche de l'exercice du même type qui traitait de la compréhension globale d'un texte, à ceci près que celui de la chanson fait intervenir d'autres éléments que la compréhension. La structure de la chanson a été présentée sous forme de cadres ce qui permet à l'apprenant d'isoler le refrain par exemple, les rimes l'aideront aussi à produire un texte aussi proche que possible de la chanson.

Dans le même ordre d'idée, la pratique dite de "*lexique éparpillé*" perdra par son passage à l'informatique un peu de son intérêt qui réside dans la confrontation / explication des différentes solutions des élèves entre eux. En effet l'activité s'articule autour de la structure suivante, elle n'est pas déterministe au niveau de la solution, et ce quel que soit l'état de connaissance de la langue de l'apprenant :

with	your				air
and		head		the	
try		trick	and		
	head		collapse		
and	there		nothing		it
and		will		yourself	
where				?	

On accompagne la grille[11] d'une liste de vocabulaire

Adverbes et prépositions	Pronoms possessifs et personnels	Substantifs	Articles	Verbes
in	your	feet	the	spin
on	it	ground	this	will
	my	mind		is
				ask

Ce genre d'exercice pourrait tout à fait être appliqué à des poèmes (ce qui ne nécessiterait pas l'usage du son), qui pourraient faire intervenir les rimes et les caractéristiques des vers (nombre de pieds etc.…)

3.2.2.3.4. Structure logique et type dominant d'un texte

[11] The Pixies (paroles et musique : Black Francis) : "Where is my mind ?", 1er couplet, tiré de l'album "SURFER ROSA", 4AD, ©1988.

Nous avons déjà évoqué les parties dans le paragraphe sur les textes littéraires mais ce n'est pas le seul exercice qui puisse être mis en œuvre afin d'analyser la structure d'un texte. Ainsi Tagliante [TAG 94] suggère de faire surligner les mots de liaison et les mots de structuration. Ceci permettra non-seulement à l'apprenant de comprendre plus en profondeur la structure du texte qu'il est en train de lire mais également de les réutiliser dans ses propres productions.

A partir d'activités de ce type (question guidant l'élève dans un processus de mise en évidence de certains traits caractéristiques de différents types de textes), l'apprenant va pouvoir extraire le type dominant d'un texte ainsi que ses éventuelles sous-dominantes. Par dominante, Tagliante [TAG 94] entend intention de l'auteur qui domine les autres de part l'expression d'un lexique, d'une "grammaire" et d'une articulation particulière des phrases. Pour arriver à tirer ses conclusions, l'apprenant pourra s'appuyer sur des règles d'organisation des différents types de textes.

Type dominant	Lexique	Syntaxe	Articulation
Narratif (construit sur un axe temporel)	De caractérisation des personnages, des lieux, des moments : qui, quoi, où, quand, avec quel résultat	Présent, futur, imparfait et passé composé, passé simple et temps composés	Indicateurs temporels : il y a un an, de nuit, le jour suivant, vers dix heures…

Tableau 3 exemple de règle d'organisation pour les textes de type "*narratif*"

Cuq [C-G 02], relaye en quelque sorte ce type d'analyse structurelle ayant pour but une meilleure compréhension du texte, puisqu'il prône, pour « *initier une construction globale du sens* » :

> *Une « lecture orientée vers certains éléments pertinents du texte qui vont initier la compréhension : […]*
>
> - *Perception du texte à travers son organisation pour en déceler son architecture : articulateurs logiques, rhétoriques, déictiques spatio-temporels, éléments anaphoriques, etc. ;*
> - *Attention portée sur les entrées et fins de paragraphes ;[…] »*

Soit autant d'aspects concernés par les pratiques ci-dessus.

3.2.2.3.5. Production

Pour les apprenants débutants, les exercices de production seront relativement systématiques, comme l'exercice suivant :

Harry is a funny boy. He is ten years old. He has brown hair and eyes and a round face. He is always dirty. He likes football.	Describe his brother Hud. Use these words : serious too clean school

Notons que, dans l'exercice original, des images étaient associées aux deux protagonistes (Harry & Hud) permettant ainsi un rappel du vocabulaire utilisé. Bien qu'utiles ces images n'ont pas paru être une condition *sine qua non* à la réalisation de l'exercice.

En général, les exemples d'activités de production que nous avons trouvés font intervenir un travail d'équipe, une interaction avec la classe (et accessoirement de l'oral). Ces composantes (à l'exception du travail à l'oral) seront difficilement intégrables dans un projet comme Mirto et sont toujours exclusivement réservées au contexte de la classe de langue.

3.2.2.4. Travail avec objectif linguistique

Comme le suggérait Alonso [ALO 94], on peut se baser sur des textes pour effectuer un travail beaucoup plus spécifiquement linguistique, qui portera moins sur fond du texte et plus sur sa forme, permettant ainsi d'adresser des questions de vocabulaire, d'orthographe, de phonétique ou de structures grammaticales.

3.2.2.4.1. Travail sur les structures grammaticales

Les exercices lacunaires ont été déjà explicités dans le contexte de la compréhension, cependant, ils sont tout aussi utiles pour l'utilisation des connaissances en termes de grammaire [GAR 90]. La différence résidera, comme on a commencé à le suggérer (voir paragraphe sur l'extraction d'informations détaillées et/ou sélectives à partir d'un texte), dans le choix des mots retirés. Si un tel exercice fait toujours appel à la compréhension, les connaissances grammaticales ne sont pas forcément mises en jeu. Dans le cas où les mots à retrouver sont fournis dans le désordre aux élèves, le rôle de la morphologie de ces derniers deviendra encore plus prépondérant. Par exemple :

« No todo _____$_1$ que brilla es oro, no todo.
No todo lo que vale _____$_2$, no todo.
No _____$_3$ los que tienen algo que decir, suenan en tu stereo, no todos.
No todo lo que brilla es oro, no todo.
No _____$_4$ lo que suena vale, no todo.

No todos _____$_5$ que tienen arte, _____$_6$ su camino »[12]

Mots proposés : $_1$lo / $_5$los / $_4$todo / $_3$todos / $_6$consiguen / $_2$suena.

Ici pour pouvoir remplir les blancs, il n'est pas nécessaire de connaître les mots proposés, ni même de comprendre réellement le texte : les mots proposés fonctionnent par paires : un mot au singulier et un mot au pluriel. Le choix pourra s'effectuer comme suit : à chaque emplacement correspond une nature de mot restreignant le choix à deux possibilités, le choix entre ces deux possibilités, pourra ensuite être fait en fonction des informations morphologiques des mots auxquels ils se rapportent.

Tagliante propose, quant à elle, un certain nombre d'activités de grammaire qui, bien que non explicitement liées aux textes, peuvent utiliser ces derniers comme énoncé [TAG 94]. Comme nous n'avons pas donné de limite de taille aux ressources textuelles dans la base, nous pouvons considérer une phrase comme un texte ou comme une sous-partie d'un texte. Tagliante sépare les exercices en différents types.

Elle évoque tout d'abord les exercices structuraux qui visent à « *faire acquérir la maîtrise d'une structure par la mise en place d'automatismes créés par la répétition de transformations structurales, à partir d'un modèle unique proposé au début de l'exercice* » :

ex : Que dites-vous ? → je vous demande ce que vous dites.

Les exercices de transformation, qui sont une *sous-classe* des exercices structuraux, peuvent par exemple concerner la position d'une question par rapport à un mot souligné dans une phrase (Le groupe qu'il écoute le plus est *the smashing pumpkins*. → Quel est le groupe qu'il écoute le plus ?), une transposition au style indirect (voir ci-dessus, la différence est qu'ici on ne donne pas forcément d'exemple dans l'énoncé) ou encore un changement de registre (vous savez ce que je pense ? → Savez-vous ce que je pense ?).

Elle évoque aussi les exercices de classement et d'appariement, dans lesquels on donne une liste de questions et les apprenants doivent retrouver les réponses correspondantes qui sont présentées dans le désordre.

Les QCM, qui « *se prêtent à toutes sortes de connaissances* », pourront avoir deux, trois ou quatre choix : « **Ella** *a) es ; b)esta* **muy simpatica.** »

[12] Macaco (paroles : Macaco el Mono Loco, musique : Macaco el Mono Loco / Martin Fuks) : "Oro en el stereo", tiré de l'album "Rumbo Submarino", Edel Music S.A., © 2001

Tagliante [TAG 94] propose ici un type d'exercice lacunaire non exposé précédemment : on donne des indications sur les mots qui devront remplir les blancs. Par exemple : mettez le verbe entre parenthèses au temps convenable / complétez par le relatif, la préposition, l'article, le possessif… convenable. Le choix des mots effacés sera bien évidemment fait en fonction de la notion que l'on veut faire travailler.

Enfin, on parlera d'exercices de reconnaissance quand, à la donnée d'un énoncé/texte, l'apprenant devra essayer de reconnaître certains traits de cet énoncé, comme le registre (soutenu / commun / populaire : T'as pas l'heure ? / Auriez-vous l'heure), le rôle et le statut de locuteurs d'un dialogue (Vous désirez quelque chose ? / Deux cafés et l'addition, s'il vous plait.) ou encore de « *reconnaître, dans les écrits, les formes morphosyntaxiques qui permettent de déterminer qui sont l'énonciateur et le(s)destinataire(s) du message* » (Dans le message suivant, quels sont les sexes respectifs de l'énonciateur et du destinataire : « *Chère Dominique, je t'ai appelée hier soir mais tu n'étais pas là. Je serais désolé si tu ne pouvais pas te joindre à nous pour mon anniversaire. Claude.* »).

Un texte peut également servir de base à ce que Tagliante [TAG 94] appelle la conceptualisation grammaticale, le but de cette approche est de faire que les apprenants à travers « l'*examen d'un corpus d'éléments constitutifs d'un fait linguistique ou d'une structure dont on souhaite faire découvrir le fonctionnement* » arrivent par eux-mêmes à une explication. En classe, cette pratique nécessite un échange entre les apprenants afin d'arriver à un consensus. Cependant on pourrait s'inspirer de ce genre de pratique dans l'autoformation, même si ce faisant elle perdra tout un aspect de son intérêt ; l'apprenant pourra être guidé dans sa réflexion à travers diverses étapes, pour arriver à une règle.

Dans ¿ Cómo ser profesor/a y querer seguir siéndolo ? [ALO 94], un autre type d'exercice est évoqué, il consiste à prendre les mots d'une phrase et à les donner aux apprenants mélangés, à eux ensuite de reconstituer les phrases (qui contiennent des tournures sur lesquelles les apprenants ont déjà travaillé).

3.2.2.4.2. Travail sur le vocabulaire

Dans le contexte de la classe, un grand nombre d'activités traitant de vocabulaire est tributaire des interactions entre les différents apprenants et l'enseignant (*brainstorming* selon un thème, situations où ce sont les élèves qui tentent de s'expliquer entre eux, d'après leurs connaissances préalables, le vocabulaire nouveau dans un texte…). Ces interactions n'étant

pas réellement modélisables en auto-formation, nous nous intéresserons plus aux activités qui seront transposables sur support informatique.

Dans La expresión escrita, García Hernandez suggère des explications de vocabulaire, mots ou phrases, concernant des mots qui ne sont pas forcément connus des apprenants.

Une autre activité lexicale possible, très proche des exercices lacunaires, proposée par Tagliante [TAG 94], consiste à changer les mots, au lieu de les supprimer, de manière à rendre le texte amusant, à la limite du surréalisme et de demander aux apprenants de retrouver le mot juste. Ceci requiert bien sûr un travail de compréhension, mais il permet aussi de travailler le vocabulaire : pourquoi tel mot ne « va » pas à tel endroit et par quel mot connu (lexique) peut-il être remplacé ?

On pourra aussi mettre en œuvre la pratique dite de « regroupements lexicaux » [TAG 94], faire rechercher tout le lexique se rapportant à un même thème.

Les QCM et exercices lacunaires peuvent également être utilisés pour effectuer un travail sur le vocabulaire [ALO 94]. Pour les QCM on recherchera particulièrement des phrases qui définissent le mot à chercher, par exemple :

> « La persona que trabaja en un restaurante es un : *a) dependiente b)cliente* ***c)camarero***. »
> ou « Pour conduire une voiture en France on a besoin d' : ***a)un permis*** *b)une licence c)une attestation*. »

Nous ne reviendrons pas sur les exercices lacunaires ci-dessus puisque nous en avons déjà beaucoup parlé, Alonso propose cependant une activité similaire mais dans laquelle ce sont des lettres et non des mots entiers qui sont effacées. On pourra par exemple retirer les voyelles d'une phrase.

3.2.2.4.3. Travail sur la phonétique

Pour les étudiants avancés, il existe des exercices de transcription en API (Alphabet Phonétique International) ou de positionnement des accents toniques, ou inversement partir de l'API pour arriver à un texte. Par exemple, dans le cas des « *compounds* » en Anglais, demander à un apprenant de passer de l'API à l'écriture normale en alphabet latin, les phrases suivantes :

> /juːˊgəʊˊpast ðəˊ griːnˊhaʊs ənd ðenˊtɜːnˊleft/ → You go past the green house and then turn left. (Tu tourneras à gauche après la maison verte).

/juːˊgəʊˊpast ðəˊgriːnhaʊs ənd ðenˊtɜːnˊleft/ → You go past the greenhouse and then turn left. (Tu tourneras à gauche après la serre).

3.2.2.5. Autres pratiques

Nous traiterons ici de pratiques, intéressantes de part leur originalité ou au contraire la fréquence avec laquelle elles sont employées. Cependant elles ne seront probablement pas prise en considération (puisque hors-sujet ici, ou trop marginales).

3.2.2.5.1. Vocabulaire

Un exercice très fréquent en termes de vocabulaire consiste à faire des paires mots / définitions. On retrouve cette technique dans l'ouvrage de Encina Alonso [ALO 94], mais ce type d'exercice ne se base sur aucun texte, à moins que l'on considère un article de dictionnaire comme un texte, ce qui n'est à priori pas le cas. Nous ne tiendrons donc pas compte de ce type d'exercice ici.

3.2.2.5.2. Mise en page

Catherine Marcus [MAR 99] invite à étudier « *espèce d'espaces* » de Georges Perec sous l'angle de la mise en page, afin de rechercher un lien entre le sens et l'aire scripturale. L'étude de la mise en page ne sera pas toujours intéressante en terme de compréhension, mais dans certains textes comme celui cité par Catherine Marcus ou « *Oh captain ! my captain !* » de Walt Whitman, elle jouera un rôle non-négligeable.

3.2.3. Comment les enseignants définissent-ils leurs usages des textes ?

A travers cet état de l'art nous avons pu nous familiariser avec certaines des principales pratiques qui risquent d'être évoquées au cours des entretiens. Cela nous aura permis également de relever certains paramètres qui entrent en compte dans la caractérisation d'une activité : le public auquel elle est destinée (âge, niveau), le but pédagogique, le type d'exercices (certains comme les exercices lacunaires revenant régulièrement).

Tout cela nous permet donc de préparer la série d'entretiens à travers :

- La reformulation de notre hypothèse de départ : le choix du texte dépendra principalement du but pédagogique. Nous allons donc tenter de voir confirmer ou infirmer cette hypothèse. Si d'une certaine manière elle se confirme, il nous faudra essayer de formaliser les paramètres influençant ce choix, pour un but pédagogique donné.

- La connaissance du domaine : grâce à elle, nous pourrons nous focaliser, au cours des entretiens, sur l'objet de cette étude (la recherche des textes). Si nous n'avions pas connu toutes ces pratiques, une grande partie de notre attention aurait été mobilisée par la description des activités mises en place par les enseignants. Nous avons pu également préciser le déroulement et le plan des entretiens et en dériver la fiche dédiée à la prise de notes.

3.2.3.1. Plan de cette phase de l'entretien

Au cours de cette phase, il n'y a pas vraiment de plan prédéfini, à priori. Nous tenterons d'aller dans le même ordre que dans l'état de l'art ci-dessus, mais nous ne couperons pas l'enseignant dans son explication s'il ne suit pas le plan. Par contre nous veillerons à suggérer les activités ci-dessus afin de savoir s'il les utilise. Bien sûr, il est possible que certaines activités utilisées n'aient pas été anticipées dans l'état de l'art.

3.2.3.1.1. Fiche

La partie de la fiche concernée par cette partie est jointe en annexe 2. Au cours des entretiens nous tenterons de passer en revue les différentes activités impliquant des textes auxquelles a recours l'enseignant interrogé, pour chaque activité, nous remplirons une "*fiche activité*".

Nous tenterons, dans un premier temps, d'en savoir plus sur l'activité, de la même manière que nous avons exposé les activités précédentes en tentant d'en présenter le but et le fonctionnement. Nous demanderons à l'enseignant de les préciser pour nous, ce qui permettra de voir si nous nous trouvons dans le cadre d'une activité connue ou d'une nouvelle.

Le public est une donnée intéressante puisqu'un type d'activité ne conviendra pas forcément à tout le monde.

La partie « commentaires », m'est destinée pour dire si l'activité est transposable sur une plate-forme électronique ou pour faire toute autre remarque inspirée par l'entretien.

Enfin la dernière case correspondant à cette partie, est la case : « caractéristiques du texte », qui regroupera les critères de choix d'un texte pour l'activité concernée. Tous les critères, aussi subjectifs soient-ils seront notés. C'est lors de l'exploitation des résultats, et seulement à ce moment là, que nous déciderons de ce qui est exploitable et de ce qui ne l'est pas.

3.3. LA RECHERCHE DE DOCUMENTS

3.3.1. Avant les entretiens

Nous avons décidé de nous intéresser au processus de recherche de documents des enseignants pour essayer de formaliser les critères qu'ils utilisent plus ou moins consciemment lors de cette démarche, rendant ainsi la recherche dans la base de ressources textuelles, la plus proche possible de ce que les enseignants connaissent.

Nous nous attendions à ce que ce processus soit très personnel et très peu formalisé, nous n'avons donc pas été surpris de trouver aussi peu de documents se rapportant à cet aspect du travail d'enseignant. En effet, le peu d'informations que nous avons pu trouver sur cette étape du travail d'un enseignant, ne nous en a pas appris vraiment plus sur le sujet [GAR 90] :

> *« Lecturas de* artículos de prensa *extranjera […] bien seleccionados en cuanto a vocabulario y estructuras, son muy útiles como "documentos reales" que aportan una información cultural del país. »*
>
> *(La lecture d'articles de presse étrangère bien sélectionnés en terme de vocabulaire et de structures, son très utiles comme "documents réels" qui apportent une information culturelle du pays.)*

Ce qui signifie, en substance, que les documents authentiques apportent une information sur la culture du pays et qu'il faut bien les choisir en ce qui concerne le vocabulaire et les structures. Mais rien n'est dit d'autre sur les éléments de la recherche ou du choix du document. Tagliante [TAG 94], quant à elle, donne plus d'informations, elle écrit même un paragraphe sur les critères de choix des documents écrits (reproduit quasiment intégralement ci-dessous) :

> *« - Que le texte dose convenablement les éléments linguistiques (morphologiques et lexicaux) connus et inconnus.*
>
> *- Que le contenu socioculturel permette une comparaison avec la réalité locale.*
>
> *- Que les différents textes proposés soient représentatifs des différents types de textes([…])*
>
> *- Que les documents soient écrits pour un public de même âge et de mêmes motivations (surtout au début de l'apprentissage)*
>
> *- Que le texte soit toujours une source de curiosité et d'information. »*

Bien que non-explicités, ces critères de choix nous donnent une indication, de ce qui peut être regardé par un enseignant au cours du processus de recherche d'un document,

cependant. Cela ne diminue en rien l'intérêt (espéré) des entretiens. Au contraire, les entretiens vont nous permettre, de valider ou non les champs qui peuvent découler de ces critères, de les compléter et d'essayer d'associer des ensembles de valeurs sur les champs qui seront retenus.

Pour cette partie des entretiens on tentera donc d'entrer le plus profondément possible dans les détails de la recherche de documents.

3.3.2. Fiche pour l'entretien

Pour chaque activité définie (voir paragraphe sur les activités), on s'intéressera au processus de recherche des documents (on en connaîtra déjà les critères de choix). On essaiera donc de savoir, vers quel type de source les enseignants se tournent (journaux, revues, livres…), où ils vont les chercher (bibliothèque, Internet) et de quels outils ils disposent pour le faire (base de données, moteur de recherche, aucun). Nous savons à l'avance que dans la plupart des cas, les pratiques sortiront du cadre défini ci-dessus (réponses anticipées entre parenthèses) et qu'il nous faudra alors nous adapter.

4. COMPTE-RENDU DES ENTRETIENS

Avant toute chose, il convient de remercier une nouvelle fois les enseignants, d'avoir bien voulu se prêter à cet exercice et de m'avoir accordé leur temps. Sans leur aide, la réalisation de ce travail n'aurait pas été possible.

4.1. LE NIVEAU

4.1.1. Remarque générale

Il a été très difficile, au cours des entretiens, d'orienter la discussion vers la manière d'évaluer le niveau des élèves dans la perspective du choix des textes. Ceci était probablement du à une mauvaise formulation de ma part. En général durant cette phase la conversation commençait à dévier sur les caractéristiques des textes, j'obtenais ainsi des informations sur la suite du travail, sans pour autant avoir de réponse à ce sujet. Malgré tout, la question du niveau revenait spontanément lorsque nous parlions des différentes activités effectuées en classe ou des critères qui influençaient le choix d'un texte. Nous avons donc pu extraire les deux tendances principales suivantes.

4.1.2. Le niveau dans le contexte de la classe

Lorsque les enseignants interrogés préparent un cours, ils le préparent dans l'optique de le présenter à une classe ou à un groupe précis. La terminologie qu'ils utilisent pour désigner un groupe de niveau est donc une conséquence directe des intitulés donnés par l'établissement pour lequel ils travaillent. Nous avons pu relever dans les entretiens des expressions du type :

- *« Moi, ils sont classés par niveau : 1ère année, 2ème année, 3ème année, maîtrise, pour les licences, ils sont classés par licence. »*[13]
- *« C'est variable selon les niveaux, on ne va pas procéder de la même façon en 6e et en 3e. »* [14]
- *« Pour les avancés j'aime beaucoup travailler avec les textes.* »[15]
- *« It's thematic, basically, with the non-specialists. »* [16]

(Au fond, avec les non-spécialistes [le choix des documents] est thématique.)

[13] Entretien avec Sophie Bourgade
[14] Entretien avec Myriam Béatrix
[15] Entretien avec Maria-Elena Galoppo
[16] Entretien avec Alice Henderson

Il est inutile de relever toutes les occurrences dans les différents entretiens, la tendance générale étant de décrire le niveau grâce aux intitulés des formations qui avaient été donnés en début d'entretien, lorsque nous nous intéressions à leur public. Ce phénomène s'est vérifié au cours de chacun des entretiens à l'exception de celui d'Adriana Celińska, cette dernière n'enseignant qu'à un seul niveau. Nous pouvons dire que si nous nous basons sur le groupe d'enseignants interviewés, tous se réfèrent au niveau de leurs élèves selon les termes de l'établissement qui les emploie.

Notons également que cette manière de décrire le niveau prend également en compte le public. Quand ils se réfèrent à un groupe en ces termes, les enseignants s'adressent aussi bien au niveau qu'à l'âge des apprenants ou à leurs attentes. Ainsi quand Sophie Bourgade dit qu'elle classe « *par licence* » ses documents, ce n'est plus le niveau qui est concerné, mais le public : elle n'utilisera pas toujours les mêmes textes avec les licences de physique, les licences d'informatique ou les licences environnement. Même si certains textes peuvent être utilisés avec tous ces groupes, le choix se fait en général de manière à ce que le thème des textes ait un certain rapport avec la formation considérée. Alice Henderson insiste elle aussi sur le fait qu'elle ne peut pas utiliser n'importe quel texte avec n'importe quel groupe. Lorsqu'elle a affaire aux étudiants en sciences humaines (principalement première et deuxième année), certains sujets sont à éviter, à cause du rapport conflictuel qu'ont les étudiants avec la langue anglaise, qu'ils associent à l'impérialisme américain. C'est un type de problème que, bien sûr, elle ne rencontre pas avec les spécialistes ou auquel Sophie Bourgade n'est pas confrontée avec son public de scientifiques.

4.1.3. Le niveau en auto-formation

Certains enseignants, parmi ceux que nous avons interviewés, ont recours en parallèle à leurs cours présentiels à des séances en auto-apprentissage. Le recours à des séances en auto-apprentissage a pour vocation de permettre de travailler à son rythme sur des points adaptés à ses besoins. Et c'est dans ce contexte qu'il sera utilisé à l'université de Savoie (73), pour les non-spécialistes (que ce soit en Sciences Humaines à Jacob-Bellecombette ou en sciences au Bourget-du-lac). Ces séances auront pour but de fixer des points grammaticaux non-assimilés au cours des années précédentes du processus d'apprentissage ou apprendre du nouveau vocabulaire. Le programme du travail effectué pendant ces heures est fixé à l'avance en parallèle du cours. En ce qui concerne le cas du Bourget-du-lac, le programme était cette année calqué sur le cours, mais sera l'an prochain adapté au niveau de chaque étudiant, évalué

en début d'année grâce au Oxford Placement Test (OPT). Une séquence d'apprentissage sera assortie à chaque niveau de l'OPT en fonction des compétences associées au niveau.

4.1.4. Les compétences

4.1.4.1. Les familles de compétences

Lors des conversations sur le niveau, un certain nombre d'axes et de compétences ont été évoqués, parmi lesquels, on retrouvera :

- Les compétences grammaticales qui s'articulent en général autour des structures qui sont maîtrisées ou non.
- Les compétences lexicales, c'est à dire l'étendue du vocabulaire, que ce soit en termes de mots ou d'expressions idiomatiques.
- Les compétences communicationnelles : prise de parole, *fluency* (fluence verbale), capacité à participer à un débat ou même accent
- Les compétences de compréhension orale (ici, la capacité à comprendre un locuteur natif, que ce soit pour retranscrire ce qui aura été dit ou au moins pour comprendre dans les grandes lignes)

Cependant, à la lumière des informations obtenues et du but de la base de textes, nous ne serons amené à traiter que les deux premiers ensembles de compétences (ou axes d'évaluation). En effet, seuls ces deux axes joueront un rôle dans le choix d'un texte.

Comme nous l'a fait remarquer Michel Sainty, les enseignants (de collège, en ce qui le concerne, mais ceci est également applicable dans les autres situations) sont « *tenus par un programme officiel* », que ce soit celui de l'éducation nationale, celui voté par une UFR (à l'Université) ou celui qui sera mis en place par un organisme. C'est donc ce programme qui dans la majeure partie des cas fera le lien entre l'intitulé du niveau et les compétences. Pour chaque niveau, un pré-requis est supposé et une progression définie. On s'intéresse ici principalement aux aspects lexicaux et grammaticaux de ces programmes. Bien sûr, on peut regretter que cela ne soit pas personnalisé à chaque apprenant, mais c'est rarement le cas d'un cours présentiel, qui fait en général intervenir trop d'étudiants pour que cela ne soit possible. Dans le cas des initiatives d'auto-apprentissage, on se ramènera à une situation analogue puisque les compétences acquises et à acquérir seront définies en fonction du test de placement (qui a aussi cours dans des pratiques de cours présentiels comme à la POL[17] de

[17] Politique ouverte des langues : http://www.u-grenoble3.fr/stendhal/stendhal/pol/polmlc_me.html

l'université Stendhal). Une fois le test de placement effectué, les apprenants appartiennent à un groupe et suivent le programme correspondant.

4.1.4.2. Granularité

La précision quant aux compétences dépendra du niveau global. On a remarqué, principalement dans le cas du polonais, mais aussi à un degré moindre dans les autres langues représentées dans les entretiens, que plus le niveau des apprenants était faible (dans le sens où ils n'apprennent la langue que depuis peu de temps), plus leurs connaissances sont évaluées avec précision. Le fait que cela soit plus flagrant pour le polonais tient probablement au fait que la grammaire polonaise est plus complexe et plus difficile d'accès pour les francophones que celles de l'espagnol ou de l'anglais, d'où une attention toute particulière accordée aux structures utilisées dans les supports du cours (voir plus loin : prise en compte du niveau dans le choix des documents).

4.2. ACTIVITÉS ET PROCESSUS DE RECHERCHE

Afin de satisfaire la condition de "*non-autocensure*", nous avons passé en revue autant de pratiques qu'il nous a paru nécessaire au moment de l'entretien. Nous avons donc parlé de pratiques orales autant que d'exercices et nous avons trouvé autant de choses intéressantes dans le contexte de cette base de textes écrits dans le premier cas que dans le second.

En outre, comme le suggèrent les fiches activités (Annexe 2), nous avons parlé du processus de recherche en le liant avec les activités, nous allons donc les traiter de concert dans cette partie.

4.2.1. Structure

Dans la suite du travail nous allons beaucoup parler de structures grammaticales et utiliser ce terme pour désigner des exemples des différentes règles de morphologie, de syntaxe et de grammaire sémantique ou encore des structures rhétoriques, que les enseignants seront amenés à transmettre à leurs élèves.

4.2.2. Activités qui n'avaient pas été vues dans l'état de l'art

Nous n'allons pas énumérer ici toutes les activités qui n'étaient pas dans l'état de l'art, puisque l'état de l'art se limitait aux activités qui peuvent être modélisées dans la plate-forme. Les entretiens ayant déjà été effectué, l'intérêt de ce paragraphe est principalement de compléter l'état de l'art ci-dessus et de donner éventuellement des idées d'activités. Les

activités faisant intervenir le groupe ou l'oral, aussi intéressantes soient-elles, ne seront pas détaillées.

4.2.2.1. Conception grammaticale

Dans le processus d'introduction d'un nouveau temps, Sonia Tendero fait passer ses élèves par une phase de repérage. Elle leur demandera de lire le texte et de relever les verbes présents dans le texte, d'essayer d'en trouver l'infinitif et de dire s'ils appartiennent au premier, au deuxième ou au troisième groupe ou encore s'ils sont irréguliers. A condition que les outils soient capables d'offrir des informations sur le groupe[18] auquel appartient un verbe, cette activité sera très facile à modéliser.

4.2.2.2. Signes diacritiques

Iwona Puchalska, utilise un autre type d'exercice, que l'on pourrait rapporter aux exercices lacunaires : elle donne à ses élèves un texte duquel elle aura retiré tous les signes diacritiques (qui sont relativement nombreux en polonais par rapport à d'autres langues). Ce type d'exercice, tient aussi bien du vocabulaire et de l'orthographe que de la phonétique (dans la mesure où dans beaucoup de langues les signes diacritiques affectent la prononciation, que ce soit en termes de phonèmes ou d'accentuation des mots).

4.2.3. Les différentes classes d'activité

Après ces entretiens, il apparaît que l'hypothèse que nous avions formulée (voir paragraphe sur les activités) se vérifie en partie, dans la mesure où en effet, il existe des différences de caractéristiques d'une activité à l'autre. Cependant, il y aura un certain nombre d'activités (dans le sens où on l'a défini dans l'état de l'art), qui partageront les mêmes textes. La différence s'effectue en fait entre différentes classes d'activités.

4.2.3.1. Bases personnelles

Lorsque nous nous sommes intéressé au processus de recherche de documents des enseignants, il y a une certaine pratique qui est revenue régulièrement, que nous allons nommer la "*recherche passive*" dans la suite du travail.

4.2.3.1.1. La recherche passive

La recherche passive a été constatée principalement au cours des interviews de Sophie Bourgade, Michel Sainty et Myriam Béatrix (dans l'ordre des entretiens). Cette pratique

[18] En espagnol comme en français, on sépare les verbes réguliers en trois groupes de conjugaison : le premier groupe concerne la conjugaison des verbes se terminant en –ar à l'infinitif, le deuxième, les verbes en –er et le troisième les verbes en –ir.

n'exclue pas d'autre types de recherches. Tous les trois m'ont indiqué, que lorsqu'ils lisent des documents en anglais, la langue qu'ils enseignent, que ce soit pour leur plaisir ou dans le cadre d'une recherche précise, s'ils rencontrent par hasard un texte dont ils ont l'impression qu'ils pourront plus tard l'exploiter d'une manière ou d'une autre, ils le remisent dans un recueil de textes. Nous nous sommes donc intéressé à la manière dont ils les classaient.

4.2.3.1.2. Organisation

Pour représenter graphiquement l'organisation des "*bases*" de chaque enseignant, nous utiliserons tout simplement une arborescence. Il n'est pas nécessaire d'avoir recours à un formalisme typique de bases de données (entité-association ou même UML) puisque ces bases ne se servent pas d'un SGBD[19], mais sont, soit des classeurs / trieurs, soit, s'ils sont informatisés, des fichiers triés d'une certaine manière dans une arborescence windows.

Les diagrammes ont été effectués en se basant sur les informations fournies au cours des entretiens :

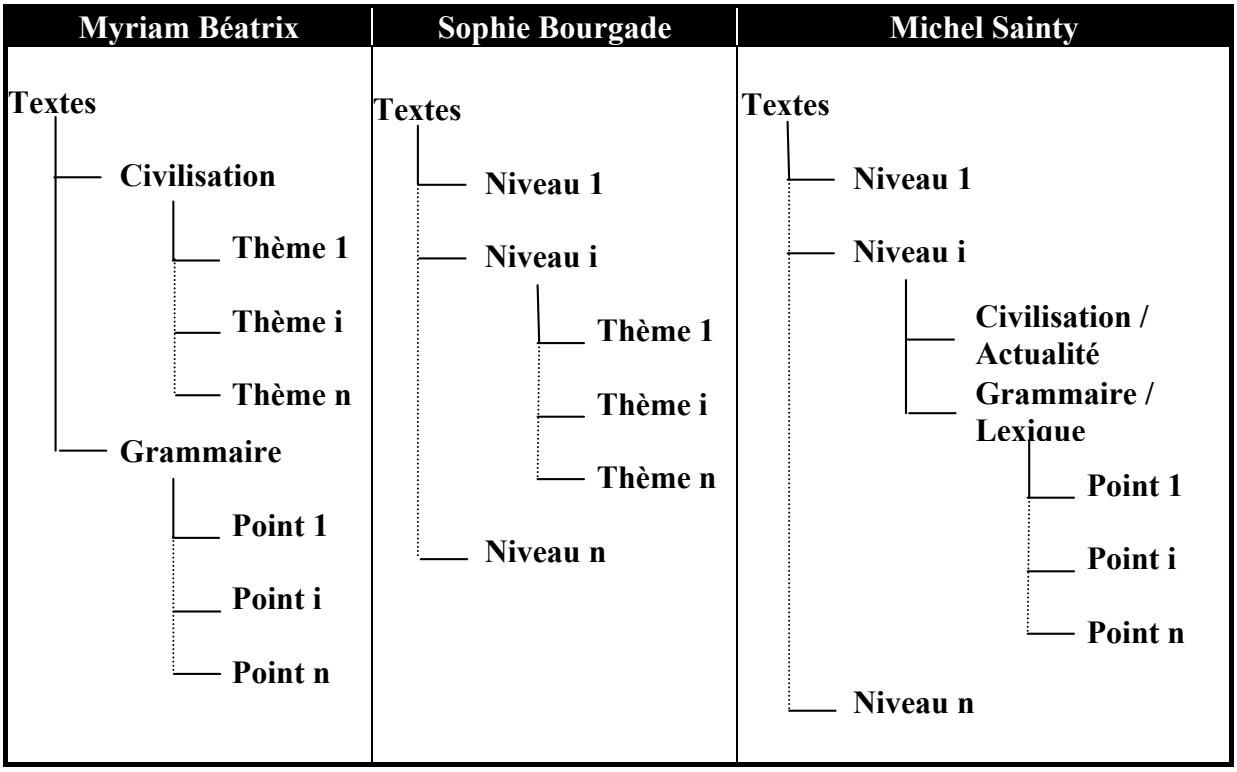


Tableau 4 Organisation des bases de textes personnelles de trois enseignants

4.2.3.1.2.1. Niveau

Par niveau, nous entendons les différentes classes ou groupes (voir paragraphe concerné). Michel Sainty aura donc un dossier pour les 6e, un pour les 5e, un pour les 4e, un

[19] Système de Gestion de Bases de Données

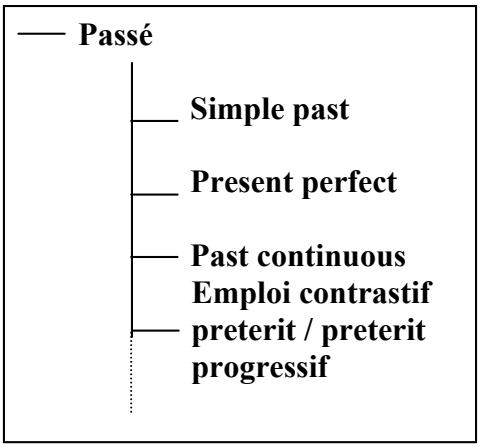
pour les 3e, un pour les 4e deuxième langue et un pour les 3e deuxième langue. Sophie Bourgade aura, quant à elle un dossier pour les 1ère année, un pour les 2ème année, un pour chaque licence et enfin un pour les maîtrise. Ces différentes bases personnelles constituaient donc une confirmation de ce que nous avancions dans le paragraphe concernant les niveaux : à partir de la conversation, cette manière de procéder transparaissait, mais ce sont ces séquences sur les bases personnelles qui l'ont ratifiée.

4.2.3.1.2.2. Thème

Par thème, nous entendons le contenu du texte, le sujet qu'il traite. Sophie Bourgade fait beaucoup travailler ses élèves à l'oral, elle recherche donc des textes qui amèneront ses élèves à parler, le thème est donc une caractéristique très importante dans son travail.

4.2.3.1.2.3. Point grammatical

Ce sera toute notion qui pourra être l'objet d'un cours. Myriam Béatrix articule son travail autour des temps principalement (tout comme Sonia Tendero qui articule ainsi la progression de son cours d'espagnol). Elle pourra donc avoir des arborescences du type (exemple basé sur ce qu'elle a exprimé dans l'entretien et non pas sur l'étude de sa base de texte) :



Par Past continuous on entend les constructions de la forme : (be+ed) + (V+ing) comme "he was listening to the music"

Figure 3 Exemple d'arborescence possible pour la base de Myriam Béatrix

Michel Sainty, quant à lui met aussi des textes dont le but est d'introduire du vocabulaire ou des tournures idiomatiques dans son dossier grammaire (grammaire / lexique).

4.2.3.1.2.4. Civilisation

Myriam Béatrix nous a donné quelques exemples de thèmes qui peuvent se retrouver dans son classeur "civilisation" :

- Culture
- Histoire de l'Angleterre
- Halloween

Michel Sainty inclut, quant à lui, des textes d'actualité dans cette catégorie.

4.2.3.1.3. Conclusions

A travers l'étude de l'organisation de ces recueils de textes, nous avons pu extraire deux classes d'activités que nous allons utiliser par la suite. La première et la plus évidente sera celle des activités "*linguistiques*", qui concerne principalement l'introduction de nouvelles structures, mais aussi de nouveau vocabulaire. Les catégories civilisation, civilisation/Actualité, correspondraient à la classe des activités de compréhension.

4.2.3.2. Exercices

Bien que non-évoqués dans le paragraphe précédent, il convient de ne pas oublier une troisième classe d'activité, celle des exercices destinés à fixer les tournures ou le vocabulaire appris dans les activités linguistiques. Cela ne signifie pas qu'il n'y ait pas d'exercices dans la classe des activités linguistiques, mais que les exercices dont nous parlons ici satisfont d'autres caractéristiques. Ils sont destinés à revoir et approfondir les notions introduites des activités linguistiques. Cette distinction est particulièrement flagrante dans le cas d'un cours présentiel, puisque les activités que l'on a appelées linguistiques sont principalement effectuées à l'oral et guidées par l'enseignant alors que dans le cas des exercices d'approfondissement les élèves sont beaucoup plus autonomes. Des différences au niveau du fonctionnement des activités découlent des différences dans le choix des documents.

4.2.3.3. Terminologie

Dans la suite du mémoire, on se référera aux trois classes d'activité avec les termes suivants :

- Activités linguistiques : concernant l'introduction de nouveau vocabulaire et de nouvelles structures.
- Activité de compréhension : le thème du texte sera variable, mais le but de l'activité la compréhension d'un texte écrit.
- Activité d'approfondissement : tous les exercices visant à approfondir des notions issues d'activités linguistiques.

Même si toutes ces activités existeront dans la plate-forme MIRTO, les plus représentées seront à priori les activités d'approfondissement.

En outre une activité peut être transversale à la classe compréhension et à la classe linguistique.

4.2.4. Choix des documents

Connaissant maintenant les différentes classes d'activités, nous allons tenter ici de les mettre en relation avec le processus de recherche de documents, mais avant cela, nous allons nous intéresser à la manière dont le niveau est pris en compte.

4.2.4.1. Prise en compte du niveau

Le niveau des apprenants, sera pris en compte principalement à travers trois aspects.

4.2.4.1.1. Longueur du texte

Le plus facilement mesurable concerne la longueur du texte, il a été évoqué dans la majorité des entretiens, particulièrement par Alice Henderson et Adriana Celińska. L'unité standard de mesure de la longueur d'un texte est le mot (nombre de mots), même si certains comptent les phrases.

Cette mesure pose un problème supplémentaire : celui de l'intégralité du texte. En effet, on nous a fait remarquer à plusieurs reprises l'importance d'avoir un texte avec un début et une fin (au moins dans les activités de compréhension et celles que nous avons appelées linguistiques [20]). Et sans cette contrainte de l'intégralité d'un texte, ce critère de longueur n'aurait pas de sens puisqu'il suffirait de couper la fin du texte pour en obtenir un de la bonne taille. Le formatage d'un texte, n'est pas interdit, mais dans le cadre des activités sus-nommées il doit être fait de manière raisonnée et prendra beaucoup de temps, d'où l'intérêt de pouvoir mesurer la longueur de textes entiers.

4.2.4.1.2. Vocabulaire et structures grammaticales

Cet aspect de la prise en compte du niveau des étudiants concerne évidemment l'étendue de leur vocabulaire et l'adéquation de celui employé dans le texte par rapport à leurs connaissances, et de la même manière l'étendue de leurs connaissances en termes de structures grammaticales par rapport à celles qui sont présentes dans le texte. En fonction de la classe d'activité, de la langue et du niveau des apprenants, on tolérera plus ou moins de

[20] entretiens avec Alice Henderson et Sonia Tendero

vocabulaire / structures inconnus dans un texte. Nous détaillerons l'influence de la classe d'activité dans le paragraphe qui lui est consacré.

4.2.4.1.2.1. Influence de la langue

D'après notre expérience d'apprenant et les données des entretiens, nous extrayons les facteurs suivants. Plus la langue enseignée est proche de la langue des apprenants, plus on tolèrera un grand nombre de mots / structures inconnus dans un texte. La transparence du sens de certains mots permettra de comprendre le texte même si le mot n'était pas connu au départ, par contre dans une langue comme le chinois ou à un degré moindre le polonais, tout mot nouveau doit être explicité ou pour le moins contextualisé à partir de mots connus. De même, la ressemblance de certaines structures grammaticales avec celles de la langue source permettra de rendre la confrontation avec ces dernières moins problématiques. Les enseignantes de polonais interviewées font donc particulièrement attention au vocabulaire et aux structures grammaticales (cas, temps, genre, nombre…) employées. Cette manière de procéder doit être assez généralisée dans le domaine de l'enseignement du polonais puisque certains recueils d'exercices recensent, pour chaque exercice, le vocabulaire utilisé. En d'autres termes, on pourra confronter un apprenant français à plus de nouveau vocabulaire d'un seul coup dans l'apprentissage de l'espagnol, de l'italien ou même de l'anglais que dans celui du polonais ou encore plus dans celui du chinois.

4.2.4.1.2.2. Influence du niveau

Le niveau des apprenants est pris en compte selon deux points de vues : le premier est le point de vue des connaissances, c'est à dire savoir ce que connaît l'apprenant pour être capable de dire quelle proportion d'un texte emploie du vocabulaire / des structures connus. Le second point de vue est celui de la progression : plus le niveau de l'apprenant est faible plus les nouvelles notions seront introduites progressivement.

4.2.4.1.3. Conclusion

Le niveau des apprenants est principalement pris en compte à travers la longueur du texte, le vocabulaire et les structures, ce qui nécessite, particulièrement pour les premiers niveaux d'avoir une connaissance précise des compétences dans ces deux domaines (compétences lexicales et grammaticales).

4.2.4.2. Prise en compte des classes d'activités

Comme on l'a vu, chaque classe d'activité a un but précis et à ce titre, les documents qui s'y prêteront auront des caractéristiques précises. Toutes les remarques suivantes sont des

remarques générales et seront à combiner avec les manières de prendre en compte les niveaux et avec les autres informations sur la recherche de textes que nous verrons par la suite.

4.2.4.2.1. Activités linguistiques

Ce seront les activités qui commenceront à priori toute séquence d'enseignement. Ces activités ayant pour but l'introduction d'un point de grammaire ou de vocabulaire, on recherchera donc un texte qui contient les formules à introduire. Le texte devra donc contenir principalement des structures connues et des instances des structures / mots de vocabulaire à introduire. Par contre, on tentera de recourir à un texte qui ne présente pas d'occurrences de tournures inconnues qui ne soient pas celles que l'on veut introduire[21].

4.2.4.2.1.1. Alternative [22]

Avant de confronter les élèves à des textes entiers mettant les tournures en situation, il est possible de commencer par les confronter à une série de phrases proposant un contexte suffisant pour comprendre le fonctionnement de la tournure à introduire. Alors qu'il sera possible, si le contexte le permet (voir paragraphe sur la prise en compte du niveau : langue, niveau des apprenants), de tolérer une occurrence isolée d'une tournure inconnue dans un texte plus long, lorsque l'on passe par des phrases on les refusera.

4.2.4.2.1.2. Champs lexicaux / champs sémantiques

L'une des activités linguistiques, en ce qui concerne le lexique est l'étude des champs lexicaux / sémantiques. Les phases de l'enseignement du polonais aux débutants par Adriana Celińska s'articulent autour de champs sémantiques : chaque phase concerne un champ sémantique dont le vocabulaire sera introduit petit à petit, cours par cours.

Alice Henderson a recours à certains champs lexicaux typiques de l'anglais (par exemple des manières de regarder : to stare at, to look at, to peep, to glare, to glance…) qui n'auront pas leurs équivalents en français (à moins de remplacer les verbes anglais par des expressions complètes, dans l'exemple précédent).

Ce type de pratique orientera bien sûr leurs recherches de documents.

4.2.4.2.2. Activités de compréhension

Dans le cadre d'un travail de compréhension, il est moins important que les éléments grammaticaux et lexicaux employés soient déjà connus. On tentera d'éviter un texte rempli de

[21] entretien avec Sonia Tendero
[22] entretien avec Myriam Béatrix

vocabulaire et de structures grammaticales inconnus, mais elles ne constituent pas un "interdit" de ce type de textes, même pour le polonais ; même si, au moins pour les débutants, quelques réserves sont émises pour les textes écrits (ce qui nous intéresse ici). Cela semble poser moins de problèmes dans le cas d'une vidéo, puisque l'image va aider pour le contexte [23]. Dans ce type d'activité, le vocabulaire sera plus limitant que la grammaire, on tolérera donc plus de structures grammaticales inconnues que de vocabulaire inconnu, qui nuira plus à la compréhension du texte [24].

En outre, si l'on se réfère au paragraphe sur les pratiques en classe de langue pour la compréhension, on peut remarquer que certains exercices visent à évaluer la compréhension globale d'un document. Et n'est-ce pas là, la vocation de ce type d'exercice que de faire prendre conscience aux apprenants que l'on a pas besoin de comprendre chacun des mots d'un texte pour en saisir la substance ?

4.2.4.2.2.1. Autres critères

Pour les activités de compréhension, un critère est revenu deux fois [25], c'est la possibilité de résumer le texte. Pour Alice Henderson, cela requiert un texte complet[26] (*« qui a une idée, la développe jusqu'à la fin et se termine »*), alors que pour Sonia Tendero, qui travaille avec un public plus jeune, cela dépend surtout de la présence d'actions dans le texte.

4.2.4.2.2.2. Remarque

Dans certains cas l'enseignant n'a pas toute latitude pour choisir les documents. A l'université de Savoie (Jacob Bellecombette), par exemple, pour les licences de sociologie, il existe un module appelé lecture sociologique dans lequel les étudiants doivent lire un article de recherche de 30 pages, choisi par l'administration. Les niveaux de langue dans les groupes de ce type là étant en général très disparates, ce type d'exercice est très difficile pour les étudiants. Cependant d'après Alice Henderson :

> *« Basically, you can take any text and make it accessible to people, […by] changing what you do with the text. […] So you have to do something with that text, to make it accessible, you can't actually change what is written there, but you add supports to it, you break it down into manageable bits. You look at the overall structure, you get them to work sometimes on listening exercises that are created from the text. […] The question for me*

[23] Entretien avec Adriana Celińska
[24] Entretien avec Myriam Béatrix
[25] Entretiens avec Alice Henderson et Sonia Tendero
[26] cf. § longueur du texte dans la prise en compte du niveau

is, how am I going to make it accessible for them ? And then the levels come in, in terms of what kind of exercise I am going to use. »

(Au fond, il est possible de prendre n'importe quel texte et de le rendre accessible, en changeant ce que l'on fait avec le texte. Alors il faut faire quelque chose avec ce texte pour le rendre accessible, on ne peut pas changer ce qui est écrit, mais on peut ajouter des supports, le découper en parties abordables. On peut aussi se baser sur la structure globale, leur faire faire des exercices d'écoute qui sont créés à partir du texte. La question pour moi est : comment vais-je leur rendre ce texte accessible ? Ce n'est qu'à ce moment là que le niveau entre en compte en termes de type d'exercice.)

Cela signifie que par les supports que l'on ajoute à un texte, on peut le rendre accessible à un public auquel il est pourtant complètement opaque au premier abord. Un constat partagé par Michel Sainty :

« Tout document peut être bon, à condition que l'on sache sélectionner dedans ce qui peut être exploitable. »

4.2.4.2.3. Activité d'approfondissement

Ces activités ont pour but de revenir sur le travail déjà effectué pour approfondir la compréhension des phénomènes mis en œuvre, de créer des automatismes et par l'essence même de ces exercices, ils ne doivent pas introduire de nouvelles notions. On sera dans ce cas là, et ce quelle que soit la langue (pour le polonais bien sûr, mais même pour l'espagnol ou l'anglais), à la recherche de textes ne présentant pas d'autres difficultés que celles de l'exercice que l'on générera [27].

4.2.4.2.4. Tableau récapitulatif

	Compréhension	**Linguistique**	**Approfondissement**
Vocabulaire inconnu	Toléré	A introduire : Impératif Autre : Interdit	Interdit
Structures inconnues	Toléré+ [28]	A introduire : Impératif Autre : Interdit	Interdit

Figure 4 Caractéristiques types d'un texte en termes de vocabulaire et de structures en fonction de la classe d'activité

Il convient de noter que ce tableau n'est pas une vérité absolue mais bien une ligne directrice. On ne considère ici que quelques paramètres, il faudra les combiner avec ceux que l'on a déjà exprimés sur le niveau, ceux qui viendront et la remarque d'Alice Henderson sur l'ajout de supports.

[27] Entretiens avec Sonia Tendero, Iwona Puchalska et Adriana Celińska
[28] on a vu que l'on était plus tolérant avec les structures grammaticales qu'avec le vocabulaire qui doit quand même rester abordable.

Un texte pourra donc convenir à différents publics en fonction de ce que l'on veut en faire, rien n'interdit d'utiliser un texte qui sert dans une activité de compréhension pour un niveau donné comme support pour une activité linguistique à un niveau supérieur.

4.2.4.3. Autres critères

Lors des entretiens, nous avons évoqué d'autres critères influençant la recherche et/ou le choix des documents, les critères suivants sont indépendants de la classe d'activité.

4.2.4.3.1. Thème

Le thème du texte était, en règle générale, le premier critère évoqué au cours des différents entretiens. Que ce soit pour les cours orientés vers la conversation ou pour les autres, le thème du texte sera directement en rapport avec le public :

> *« [En parlant des caractéristiques recherchées dans un texte] d'un côté c'est l'actualité et de l'autre côté c'est l'intérêt professionnel de mes élèves, c'est à dire s'il y a des élèves qui font de l'histoire ou de la géographie [je prendrai des textes qui parleront de l'histoire de l'Amérique latine][…], s'il y a des scientifiques j'essaye toujours de leur ouvrir un peu la tête à la culture. »* [29]
>
> *« What I've tried to do is make it so that we have got an organized set of texts that appeal to the students and that are in harmony with what they expect from language teaching to some extent. […] For psychology, the first semester, they really want to work on psychology stuff. So the themes […] they are interested in deal with psychology, so you take like a popularised text […] and they love that. They eat that sort of stuff up. »*
>
> *« Second term, they are so fed up with psychology, they want to read about Stephen King or "ants on the moon", they want to read about something that has nothing to do with psychology. So you are choosing your text based on what you think they are going to want to work on, what you imagine their motivations to be. »* [30]
>
> *(Ce que j'ai essayé de faire c'est de m'arranger pour avoir un ensemble organisé de textes qui intéressent les étudiants et qui, d'une certaine manière, sont en harmonie avec ce qu'ils attendent des cours de langue. En ce qui concerne les étudiants en psychologie, le premier semestre, ils veulent vraiment travailler sur la psychologie. Alors les thèmes qui les intéressent sont liés à la psychologie, on peut prendre un texte vulgarisé, ils adorent cela. Ils dévorent ce genre de texte.)*
>
> *(Quand arrive le deuxième semestre, "ils n'en peuvent plus de la psychologie", ils veulent lire des textes sur Stephen King ou sur "*des fourmis sur la lune", *ils veulent lire quelque chose qui n'ait rien à voir*

[29] Entretien avec Maria-Elena Galoppo
[30] Entretien avec Alice Henderson

avec la psychologie. Alors je choisis les textes en me basant sur ce sur quoi je pense qu'ils vont avoir envie de travailler, ce que j'imagine être leurs motivations.)

Alice Henderson explique donc que l'on essaie de faire que les textes employés soient non seulement intéressants pour les étudiants, mais aussi, qu'une certaine proportion d'entre eux corresponde à l'idée que se font les étudiants des textes qui doivent être utilisés dans l'apprentissage des langues. Elle relate ensuite l'exemple des DEUG de psychologie de l'université de Savoie. Ils insistent pour travailler sur des textes en rapport avec leur domaine au premier semestre, mais en général préfèrent parler de tout sauf de psychologie dans le cadre de leur cours de langue au deuxième semestre.

« Je ne vais pas parler [à mes élèves] que de ce qu'ils connaissent, ce n'est pas très intéressant, je vais aussi leur parler de ce qu'ils ne connaissent pas, parce que moi, ça m'intéresse plus. Je [ne] vais pas leur parler que de téléphones portables ou d'ordinateur […] parce qu'il y a des textes, à part l'aspect linguistique, [dont] il n'y a rien à tirer. » [31]

En règle générale, ce qui est recherché à travers le choix du thème d'un texte est soit de trouver un thème qui corresponde à ce qui intéresse les apprenants, quelque chose qu'ils connaissent, soit au contraire de les ouvrir à d'autres horizons, ou encore d'en profiter pour ajouter des aspects culturels à l'enseignement des langues.

4.2.4.3.2. Type de texte

La question du type de texte, au sens traditionnel du terme est revenue régulièrement, que ce soit pour faire travailler les élèves sur les spécificités de chaque type de texte ou pour diversifier les écrits auxquels ils sont confrontés. Un certain nombre de types de textes ont été évoqués au cours des entretiens :

Texte Littéraire	Autre
Extrait de roman	Article d'actualité (journaux)
Nouvelle	Article de recherche
Conte	Article scientifique vulgarisé
Poème	Texte documentaire (de société)
Paroles de chanson	Texte documentaire (historique)
Fiction	Dialogue
Scenario	

[31] Entretien avec Sonia Tendero

4.2.4.3.3. Style

Dans le prolongement du type de texte, certaines composantes que l'on regroupera sous le terme style, sont utilisées pour évaluer la lisibilité du texte :

> *« […Il y a certains textes] qui sont tellement littéraires que tu ne peux pas les donner aux scientifiques. »* [32]

> *« Un article de journal, je le laisserai plus volontiers tel qu'il est, parce que le vocabulaire [est plus simple] et les tournures sont concises, il n'y a pas d'effet de style, c'est peut–être plus facile à comprendre. »* [33]

> *« There are certain news magazines, or newspapers I will never choose a text from, never ever. I will never use a text from Time and Newsweek. Forget it. It is packed with idiomatic expressions and what turn out to be often non-standard constructions. The guardian is full of typing mistakes, but […] it has a clear style. »* [34]

> *(Il y a certains magazines d'information ou certains journaux desquels je ne choisirai jamais un texte, "jamais de la vie". Je n'utiliserai jamais un texte de Time ou de Newsweek. C'est hors de question. Ils sont remplis d'expressions idiomatiques et de constructions qui s'avèrent souvent ne pas être standard. The Guardian est plein de fautes de frappes, mais le style est clair.)*

Les composantes de ce que l'on a appelé le style, sont donc la présence de figures de style, d'expressions idiomatiques et nous allons rajouter aussi la présence de vocabulaire spécialisé, que l'on retrouvera par exemple dans les articles de recherche ou même dans certains articles de vulgarisation scientifique.

4.2.4.3.4. Texte authentique, édité ou inventé

> *« Si je trouve une page dans un roman qui m'intéresse dans un roman, je la photocopie et je l'utilise, il m'arrive de la réadapter, […] je ne la prends pas obligatoirement brut de décoffrage, il m'arrive de la réécrire pour qu'elle soit plus facile à comprendre, pour [ne] pas que ça ait un aspect rebutant. »*[35]

Ce n'est pas l'objet de ce travail de dire quel texte trouvera sa place dans la base de textes ou non. C'est là le travail des enseignants. Or pour certains niveaux (les plus débutants) il sera difficile d'avoir recours à des textes authentiques, principalement dans les langues les

[32] Entretien avec Maria-Elena Galoppo
[33] Entretien avec Michel Sainty
[34] Entretien avec Alice Henderson : « *Il y a certains magazines ou journaux desquels je ne prendrai jamais un texte, jamais de la vie. Je n'utiliserai jamais un texte de Time ou de Newsweek. Pas question. C'est rempli d'expressions idiomatiques, qui souvent se révèlent être des constructions non-standard. The guardian, est plein de fautes de frappes […] mais le style est clair.* »
[35] Entretien avec Michel Sainty (est suivi par la citation que l'on a prise dans le paragraphe sur le style)

plus éloignées du français[36], et même dans le cas de l'espagnol, il faut parfois avoir recours à l'édition des textes[37].

4.2.5. Processus de recherche

4.2.5.1. Remarque

> *« Je crois qu'avec l'expérience on trouve toujours. Tu sais ce que tu vas trouver avec vocable, ce que tu vas trouver dans le bouquin de grammaire, ce que tu vas trouver par Internet. J'ai beaucoup d'intuition toujours pour trouver les textes. […] »*[38]

Comme nous nous y attendions, le processus de recherche est très instinctif et peu d'information ont pu être recueillies sur le processus même. Nous nous sommes donc concentré sur les critères énoncés plus haut, et avons malgré tout réussi à glaner quelques informations.

4.2.5.2. Source

De ce que nous avons pu entrevoir du processus de recherche, l'expérience permet surtout de savoir où chercher certains types de textes ; par type de texte nous entendons aussi bien la typologie en fonction de ce que nous allons en faire (classe d'activité) que dans le sens plus traditionnel (paragraphe sur le type). Dans le choix de la source, les critères énoncés ci-dessus interviennent. Nous pourrions reprendre ici la citation d'Alice Henderson qui exclut Time et Newsweek de ses recherches (paragraphe sur le style). Nous pourrons tout aussi bien citer Maria-Elena Galoppo qui, dans sa recherche de courts textes littéraires, s'orientera vers certains périodiques :

> *« Même dans el Pais, […] ou dans la presse qui vient d'Amérique latine, le dimanche, il y a des histoires[…] qui sont [très] littéraires. »*[39]

Lorsque l'on parle de source, cela concerne aussi bien des journaux ou sites Internet, qui sont largement utilisés dans la recherche de documents, que des auteurs :

> *« Roald Dahl, fantastic ! All his little short stories are packed with these verbs […] for emotion and gestures and things, that in French, you need like sort of a whole phrase for."* [40]

---

[36] voir entretiens avec Myriam Béatrix (utilisation de dialogues de la méthode pour les 6e), Iwona Puchalska et Adriana Celińska (cas du Polonais, parfois nécessité d'inventer les textes soi même).
[37] Entretien avec Sonia Tendero (utilisation de textes authentiques en espagnol, mais parfois raccourcis ou édités, surtout pour les débutants)
[38] Entretien avec Maria-Elena Galoppo
[39] Entretien avec Maria-Elena Galoppo
[40] Entretien avec Alice Henderson

(Roald Dahl, fantastique ! Toutes ses petites nouvelles sont remplies de ces verbes […] pour les émotions, les gestes et tout, pour lesquels, en français, vous utilisez presque une phrase entière.)

En fonction du type de champ lexical concerné (verbes d'émotions / gestes), Alice Henderson choisira un auteur, qui en utilise beaucoup. Pour avoir des textes entiers et pas juste des extraits, elle s'orientera vers des nouvelles.

4.3. RÉCAPITULATIF

Nous détaillons ici la tendance que nous avons dégagée au cours de nos entretiens. Et c'est en tant que tendance que les informations relatées ici sont présentées. Le nombre de personnes interviewées et la durée des entretiens ne permettent pas d'aller plus loin.

Au cours de ces entretiens, nous avons pu nous apercevoir que les enseignants, que nous avons interrogés, évaluaient le niveau principalement en fonction de la classe ou du groupe auquel ils enseignaient. Chaque groupe a ses propres caractéristiques qui influenceront les critères de recherche et de choix des textes qui seront utilisés :

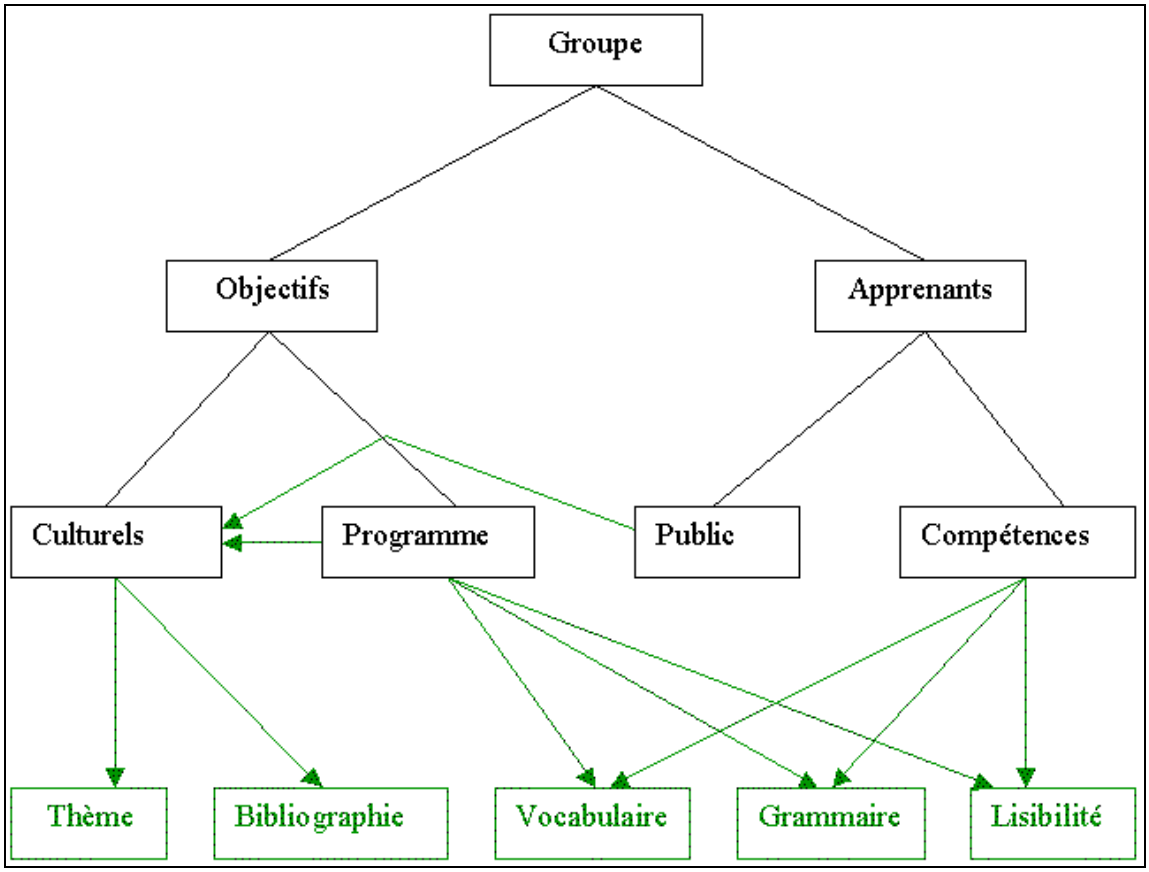


Figure 5 Influence des caractéristiques des groupes d'apprenants sur les recherches de documents

Sur la figure précédente, les flèches vertes expriment une influence.

Nous avons séparé les caractéristiques d'un groupe en deux classes : les objectifs pédagogiques et les caractéristiques des apprenants eux-mêmes.

En ce qui concerne les objectifs pédagogiques, nous les séparons entre, d'un côté le programme imposé par l'organisme pour lequel l'enseignant travaille (éducation nationale, organisme privé…) et de l'autre les objectifs culturels (quels aspects culturels aborder à travers le cours de langue : civilisation, arts…). Nous désignons principalement par programme la partie linguistique du programme, c'est à dire les compétences grammaticales et lexicales. Mais les programmes ont aussi une composante culturelle exprimée par la flèche représentant leur influence sur les objectifs culturels de l'enseignant. Ce dernier définit lui aussi des objectifs culturels et prend en considération le public auquel il enseigne.

Le terme public, tel que nous l'avons utilisé plus tôt désigne, les caractéristiques du groupe d'apprenants : âge, motivation de la participation au cours, milieu social… Les compétences désignent le niveau général de la classe en terme de grammaire et de lexique. Chacune de ces composantes influence la recherche.

Le contenu culturel du cours influence la recherche en ce qui concerne le thème du texte mais aussi ce que nous avons appelé la bibliographie : la source (type de source, titre, date, lieu de publication), l'auteur (nom, prénom, dates de naissance et de décès, nationalité) et le texte lui-même (type, titre, date, intégrité[41] et authenticité[42]).

Le contenu culturel dépendant directement du public, tous les aspects sus-nommés sont donc influencés par le public (via le contenu).

De la même manière, le programme affecte ces traits (toujours par l'intermédiaire du contenu culturel). Cependant, dans la majorité des cas, le programme concerne plus particulièrement les compétences à faire acquérir aux apprenants. Nous avons spécifié que, dans le contexte de cette étude, nous allions nous focaliser sur les compétences grammaticales et lexicales. Pour les transmettre aux apprenants, les enseignants s'appuieront sur des textes sélectionnés en partie par rapport à leurs contenus lexicaux et grammaticaux. Le programme impose ainsi des caractéristiques au texte. Malgré tout, le programme seul ne suffit pas pour décider de l'adéquation du texte (lexicalement et grammaticalement), les compétences des

[41] Le texte a-t-il été édité ?
[42] Le texte est-il authentique au sens où on l'a défini dans ce mémoire ? (chapitre 1)
Un texte peut être authentique mais avoir été édité (coupé par exemple)

apprenants ont également leur rôle à jouer. Si le programme impose certains traits, les compétences autorisent ou excluent des textes en fonction de ces critères (voir tableau récapitulatif sur les classes d'activité).

Enfin les compétences des apprenants seront aussi prises en compte (de manière plus informelle) pour choisir un texte en fonction de sa lisibilité, de son accessibilité. Cette dernière dépend de la longueur du texte (parfois imposée par le programme) en nombre de phrases ou bien de mots, de la fréquence du vocabulaire spécialisé, des expressions idiomatiques et des figures de style.

Dans le cadre d'une recherche classique (sans base de textes), certains traits des textes ne sont pas disponibles immédiatement (i.e. avant la lecture). Les remarques du paragraphe sur la recherche de documents en utilisant les sources en sont de bons exemples.

En outre, nous avons pu voir l'influence de la classe d'activité sur les critères de recherche : la classe d'activité définit la tolérance lors des recherches en fonction des aspects grammaticaux et lexicaux des textes.

Dans la suite du mémoire, nous allons tenter de tirer parti des données de ce chapitre, pour préparer la réalisation informatique de la base de textes indexée pédagogiquement. La première partie de cet ultime chapitre (exploitation informatique des résultats) est légèrement moins technique que la suite et concerne l'organisation logique de la base de textes, ce qui donne l'occasion d'expliciter les champs.

5. MODÉLISATION INFORMATIQUE

Une fois les entretiens réalisés et les informations réunies, nous devons les utiliser pour pouvoir créer une base de données adaptée à ses utilisateurs. Les données de la présente partie ne sont pas définitives, mais elles peuvent constituer le point de départ d'un travail plus fini.

5.1. DONNÉES CANDIDATES

Au cours de nos entretiens nous avons retenu un certain nombre de traits, décrivant les textes. Avant de pouvoir s'intéresser à comment les formaliser ou comment les intégrer à la base de données nous allons les rappeler ici :

- Langue du texte
- Longueur du texte
- Source (Support, Auteur)
- Texte édité ou non
- Thème
- Niveau de vocabulaire
- Niveau grammatical
- Présence de champs lexicaux
- Style (figures de style, vocabulaire spécialisé, formules idiomatiques)

Nous détaillerons la manière de les utiliser dans une partie plus technique sur la réalisation de la base.

5.2. CHAMPS POTENTIELS DE LA BASE

Avant d'entrer dans le détail des champs de la base, nous allons choisir un formalisme qui nous permettra de la décrire.

5.2.1. Choix d'un formalisme

Nous allons ici modéliser les données que nous allons stocker dans la base de textes. Pour ce faire nous allons devoir choisir un formalisme. Pour pouvoir choisir le formalisme, il convient de s'interroger sur le but de la manœuvre : nous voulons ici, choisir un formalisme qui soit assez clair pour permettre au non informaticien de comprendre de quoi il retourne, et qui puisse servir de base à une réflexion future. En effet, comme nous l'avons déjà dit plusieurs fois, le but de ce travail n'est pas d'arriver à un produit fini ; s'il est possible de créer effectivement une base de données à partir du travail réalisé, elle ne constituerait de

toutes façons qu'un prototype, voué à être repensé dans certaines lignes. Ce que nous voulons représenter ici, ce sont les conséquences sur la conception d'une base de données, que représentent les informations glanées au cours des entretiens. On s'intéresse donc ici au niveau conceptuel du processus de création d'une base de données.

Les formalismes candidats sont donc :

- UML : diagramme de classes
- Merise : Modèle Conceptuel de Données (modèle entités – associations)
- Modèle relationnel

Afin de permettre une certaine lisibilité du schéma, on peut d'ores et déjà écarter le modèle relationnel, qui est des trois formalismes ci-dessus le plus difficile d'accès et de plus, va bien plus loin que le niveau conceptuel, ce qui n'est pas nécessaire ici. Le choix devra donc se faire entre UML et le modèle entités – associations.

Les diagrammes de classes d'UML et les modèles entités – associations sont très proches :

> *« The UML object model is essentially just an extended Entity-Relationship (ER) model. The use of ER models for designing databases is well accepted and we use UML object models in a similar, but more powerful, manner. The primary advantage of OO models is that the same model addresses both programming and databases. »*[B-P 99]
>
> *(Le modèle objet de UML est, dans son essence, qu'un modèle entité – association étendu. L'utilisation de modèles entité – association pour la conception de base de données est bien acceptée et il est possible d'utiliser les modèles objet UML d'une manière similaire, mais plus puissante. L'avantage premier des modèles orientés objets est que le même modèle s'adresse aussi bien à la programmation qu'aux bases de données.)*

Bien que les auteurs puissent être influencés dans l'écriture de leur article [43], nous suivrons ce conseil dans la mesure où les modèles sont très proches et que si l'opportunité de tester le travail effectué par le biais d'un prototype se présente, une partie du travail sera déjà effectuée.

[43] L'article cité est sujet à un *copyright* de l'entreprise Rational, un des leaders mondiaux dans le développement d'outils pour UML.

Afin de ne pas surcharger les diagrammes et de ne pas entrer dans des détails de la réalisation, nous ne préciserons pas toujours les types des données employés.

5.2.2. Diagramme de classes UML

Afin que la modélisation conceptuelle de la base de données soit aussi compréhensible que possible, nous allons rappeler les rudiments des diagrammes de classes UML et les mettre en rapport avec les bases de données. Bien évidemment, nous n'exposerons ici que de brefs rappels des notations qui seront utilisées dans le diagramme, cela ne représente qu'une infime partie des possibilités d'UML.

5.2.2.1. Classes et attributs

> *« A class represents a concept within the system being modeled. Classes have data structure and behavior and relationships to other elements. »*
> [OMG 03]

Les classes représentent un concept dans le système modélisé, les classes sont caractérisées par des structures de données, un comportement et des relations avec d'autres éléments. Ici, nous allons surtout s'intéresser à la structure de données (voir attributs) et aux relations avec les autres éléments. Nous pouvons, d'ores et déjà, voir le lien avec les diagrammes entités – associations : nous allons nous intéresser aux entités et à leurs structures de données ainsi qu'à leurs relations avec les autres éléments (les associations). Lorsque nous parlons de comportement nous nous référons aux méthodes de la programmation objet, c'est à dire aux fonctions qui seront applicables à un objet donné. Ici, nous sommes dans le cas d'une base de données et le but est de pouvoir retrouver une instance d'une classe à partir de ses attributs ; dans le modèle conceptuel les méthodes n'ont pas lieu d'être (elles alourdiraient le diagramme inutilement).

Les attributs correspondent à des traits caractéristiques des éléments de la classe.

5.2.2.2. Clé

Nous ne détaillerons pas ici le choix des clés primaires des entités parmi les clés candidates, nous nous contenterons d'appeler clé une clé primaire.

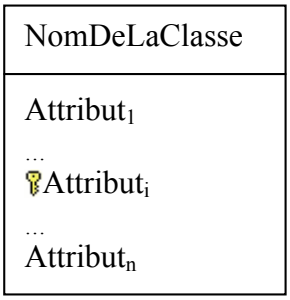
Une clé est un attribut (ou un ensemble d'attribut) qui servira d'identifiant pour un élément d'une classe.

« Ces identifiants ont ceci de remarquable que l'ajout d'une propriété (attribut) *ne leur fait pas perdre leur qualité. Le retrait d'une seule propriété, par contre, si. »* [A-V 01]

En d'autres termes, si on décide de prendre en compte une propriété supplémentaire dans le cadre d'une clé, celle-ci ne perd pas son statut d'identifiant, par contre le simple fait de retirer d'une clé une de ses composantes (attribut), fait qu'elle n'est plus identifiant.

5.2.2.3. Notations

5.2.2.3.1. Classe



Convention : pour indiquer la clé d'une classe, on la préfixera par un pictogramme clé : . En général ce pictogramme est destiné à représenter d'autres données (la visibilité d'un attribut ou d'une méthode), qui ne nous intéressent pas ici. On peut donc l'utiliser sans créer d'ambiguïtés.

Dans le cadre ci-dessus, la clé primaire est l'attribut $Attribut_i$.

5.2.2.3.2. Relation / Cardinalité

Alors que dans un diagramme entité association on parle de cardinalité d'une relation, on parlera dans UML de la multiplicité d'une association, cependant on décrit exactement le même phénomène. Nous allons reprendre ici l'exemple donné dans son cours par Didier Donsez [DON 02]:

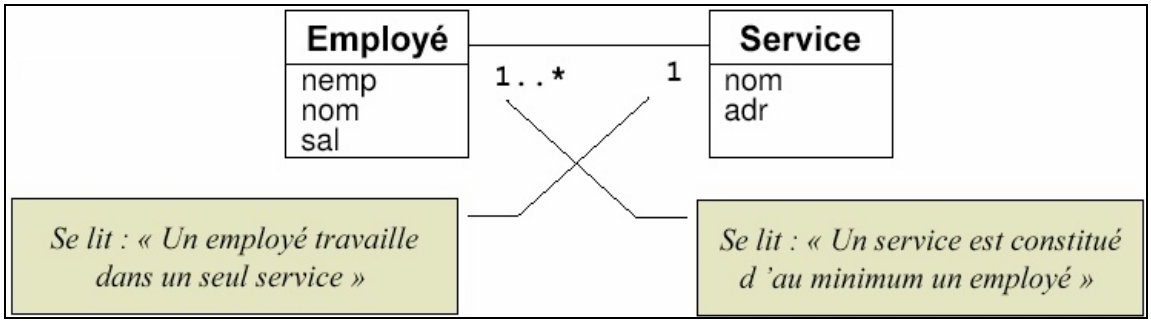


Figure 6 Lecture de la multiplicité d'une association dans un diagramme de classe

Les deux cas ci-dessus ne constituent pas l'intégralité des cas auxquels on peut être confronté :

- Si aucune information n'est donnée, alors on est dans le cas d'une multiplicité 1. Le diagramme ci-dessus pourrait s'écrire :

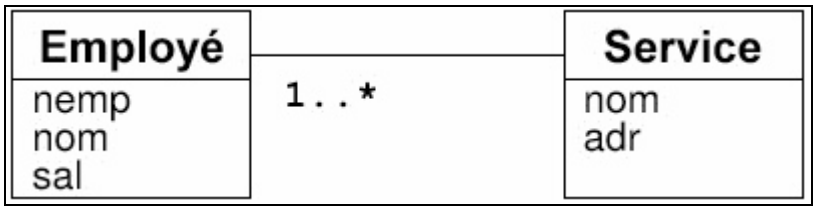


Figure 7 Un employé travaille dans un seul service

- La multiplicité * est équivalente à la multiplicité 0..* :

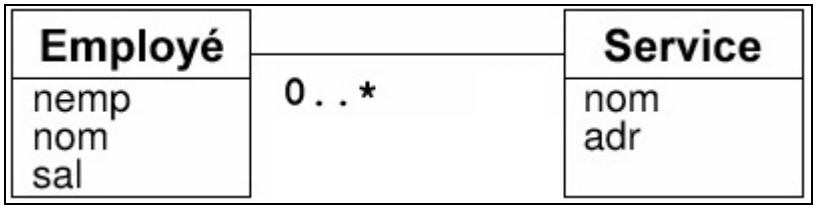


Figure 8 Un service peut être désaffecté et sera constitué d'un nombre indéterminé d'employés (0 ou plus)

- Cas général, la multiplicité $N_0..N_1$, la multiplicité sera comprise entre N_0 et N_1 :

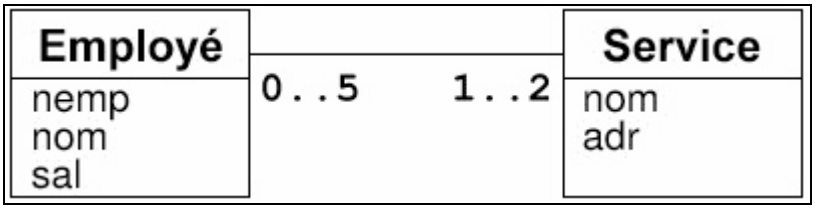


Figure 9 Un service peut être désaffecté et sera constitué au maximum de 5 employés. Un employé peut travailler dans un ou deux services.

5.2.2.3.3. Classes associatives

5.2.2.3.3.1. Définition

Nous allons nous appuyer, pour la définition, sur le cours de Jean-Marie FAVRE [FAV] :

Elles servent à associer des attributs et / ou des méthodes à des associations, d'où le nom de classes associatives. Le nom de la classe correspondra au nom de l'association.

5.2.2.3.3.2. Notation

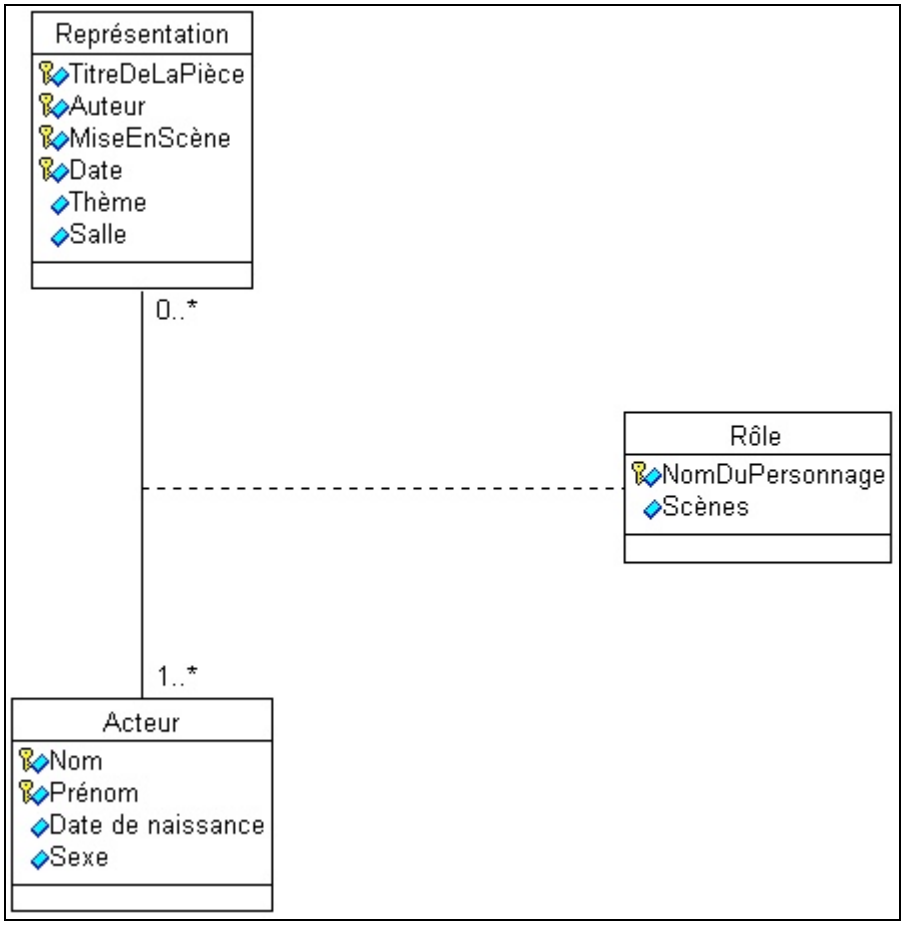


Figure 10 Exemple d'utilisation d'une classe associative.

On modélise des représentations de théâtre : on a d'un côté la classe représentation et de l'autre la classe acteur. Il faut les mettre en relation pour savoir dans quelles pièces ont joué les acteurs et qui sont les acteurs pour une représentation donnée. Cependant ceci n'est pas suffisant, car on aimerait savoir quel rôle a joué l'acteur. On ne peut pas rajouter un attribut rôle dans la classe acteur car un acteur peut avoir plusieurs rôles dans sa carrière, cela ne correspondrait donc plus à une entité acteur. De même on ne peut pas rajouter un attribut rôle dans la classe représentation car la plupart des pièces ont plus d'un protagoniste. Si l'on se contente d'ajouter une relation entre les deux classes on n'aura pas de précision quant au rôle joué par l'acteur. D'où l'intérêt de la classe associative : elle permet de préciser la relation qui existe entre les deux classes en fournissant les attributs NomDuPersonnage (qui, comme il sera en relation avec une représentation donnée, ne sera pas ambigu) et scènes qui permettra d'indiquer les scènes dans lesquelles le personnage est un protagoniste.

5.2.2.3.4. Héritage

La définition se base à nouveau sur le cours de Jean-Marie FAVRE [FAV] :

5.2.2.3.4.1. Définition

> *« Une classe peut être la généralisation d'une ou plusieurs autres classes. Ces classes sont alors des spécialisations de cette classe. Les sous-classes héritent des propriétés des super classes (attributs, associations, méthodes, contraintes) »*

Ce que Jean-Marie Favre entend par-là, c'est qu'une classe peut être la sous-classe d'une classe C et qu'à ce titre elle héritera de ses propriétés.

5.2.2.3.4.2. Notation

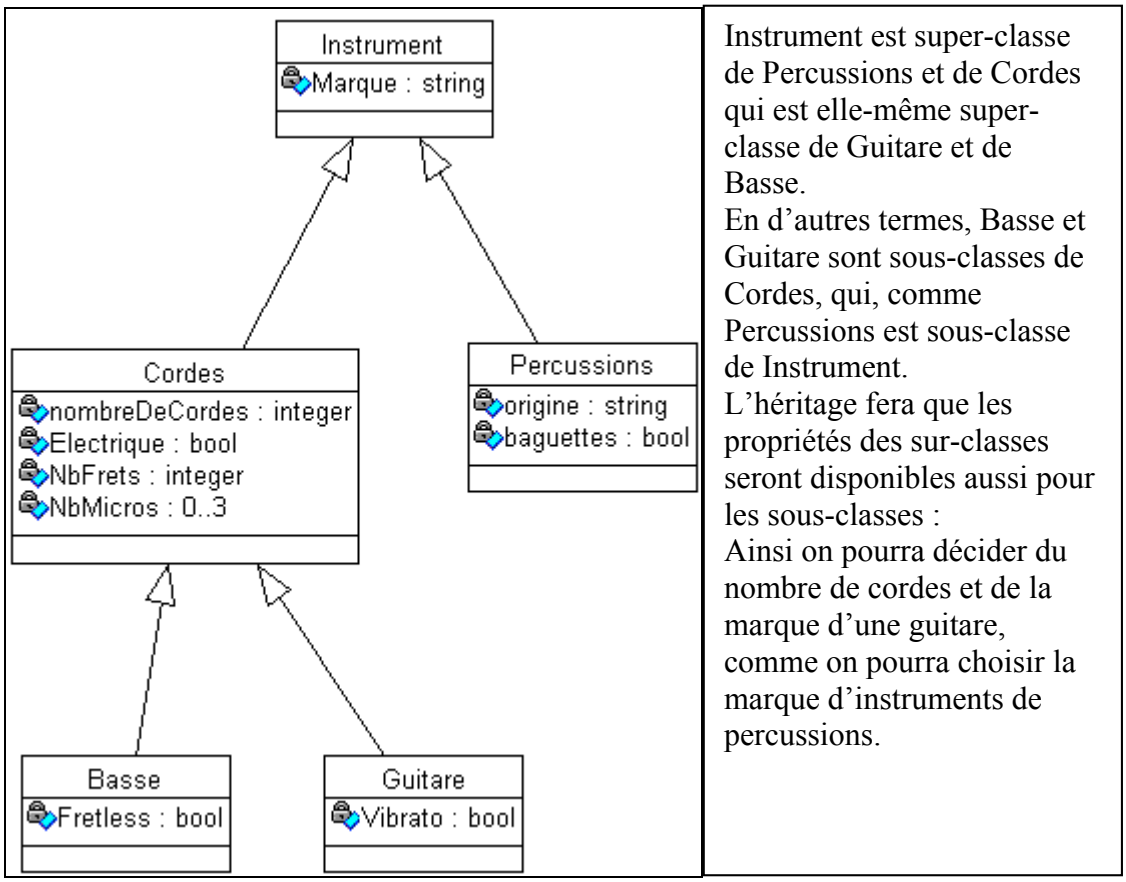


Figure 11 Exemple d'utilisation de l'héritage.

5.2.3. La base de textes

Dans le diagramme, ci-dessous, nous ne nous intéresserons qu'aux informations qui sont utiles à l'utilisateur dans ses requêtes, nous nous intéresserons plus loin à l'implémentation.

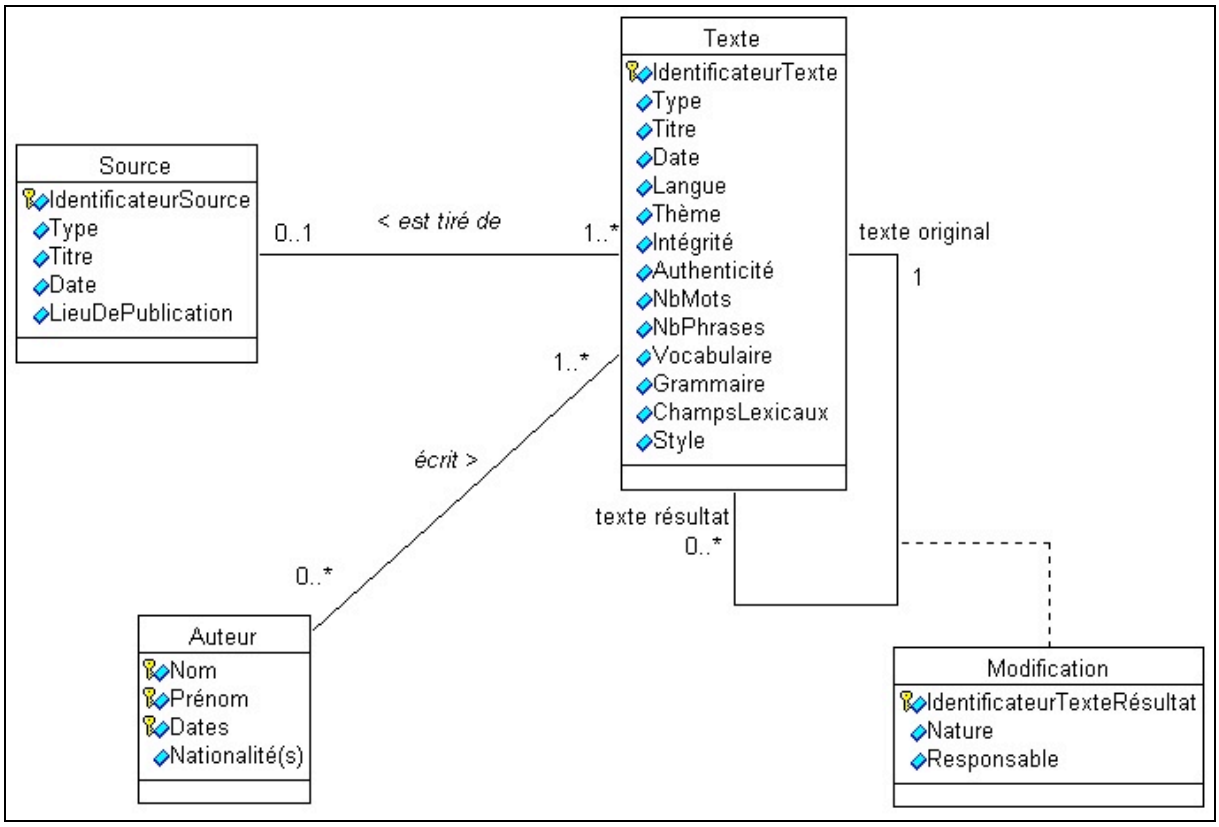


Figure 12 Diagramme de classes UML : Base de textes, conçue sous un angle pédagogique.

5.2.3.1. La classe Texte

C'est évidemment ici la classe centrale, puisque le but de la base de données est de retrouver les textes.

5.2.3.1.1. La clé

Nous avons décidé de prendre une clé qui ne s'appuie pas sur des caractéristiques du texte, tout d'abord parce que nous aurions eu besoin de prendre en compte beaucoup d'attributs (Titre, Date, Clé de Source, Clé d'Auteur et Clé de Modification), et nous aurions eu une clé relative (Clé ayant « *besoin d'un identifiant provenant d'une autre entité pour distinguer sans ambiguïté les occurrences d'une entité.* » [A-V 01]). En effet, même si cela est peu probable, deux auteurs différents peuvent avoir écrit des textes différents portant le même titre au même moment. Même si l'on se trouve dans le cadre du même texte d'origine (même auteur, même date, même titre), on peut se trouver en présence de deux textes différents dans la base si l'un provient d'une édition « *texte intégral* » et l'autre d'une édition qui aura privilégié l'aspect pédagogique. Enfin, même si le texte est issu de la même source, il y a toujours la possibilité qu'il ait été édité pour les besoins d'une activité précise (Michel Sainty nous disait dans son entretien qu'il était parfois obligé de modifier légèrement ses textes en fonction de son public ou de ce qu'il voulait en faire). La base contiendra donc deux

versions d'un texte identique et la différence entre les deux se fera par le biais des modifications. Même s'il convient de prendre tous ces aspects en compte, il est plus simple de recourir à une clé plus artificielle, qui en outre permettra d'indexer les textes dans la base et éventuellement de les référencer. L'identificateur texte sera donc une codification propre à Mirto.

5.2.3.2. Autres attributs

Les autres attributs concernent chacun une caractéristique du texte, qui pourra être prise en compte au moment de la recherche. Ils dépendent directement des résultats des entretiens.

- Le type permettra de savoir si l'on a affaire à un texte journalistique ou littéraire (voir paragraphe types de textes dans la partie ajout d'un texte à la base de données).
- Le titre sera évidemment rempli avec le titre du texte concerné.
- La date, concernera la date de l'écriture du texte (si on la connaît) : Cela n'aura pas de sens d'attribuer comme date à un texte celle de parution de la source, qui peut être un recueil paru des années après le texte lui-même. Et ce d'autant plus que l'année peut être un critère de recherche : si l'on désire étudier la littérature d'une période donnée, on ne veut pas qu'un texte correspondant aux critères ne soit pas trouvé, sous prétexte que la version disponible dans la base, provient d'un recueil paru des années plus tard.
- Langue : Langue(s) dans laquelle (lesquelles) est écrit le texte.
- Thème : de quoi parle le texte. Les bases de textes des enseignants que nous avons interviewés (pour ceux qui en avaient) sont, en général, triées (en partie) selon une approche thématique.
- Intégrité : Le texte est-il tel que l'auteur l'avait écrit, ou a-t-il été modifié ?
- Authenticité : a-t-on affaire à un texte authentique, ou à un texte uniquement à vocation pédagogique ? Entre le texte authentique et le texte pédagogique, on trouvera les textes authentiques, modifiés à des fins pédagogiques.
- NbMots / NbPhrases : longueur du texte selon les deux mesures qui ont été abordées durant les entretiens.
- Vocabulaire : Evaluation du vocabulaire présent dans le texte par rapport aux connaissances du public.

- Grammaire : Evaluation des structures grammaticales présentes dans le texte par rapport aux connaissances du public.
- ChampsLexicaux : Champs lexicaux présents dans le texte.
- Style : Utilisation du style pour l'évaluation de la lisibilité du texte, selon les trois axes définis précédemment (voir entretiens) : expressions idiomatiques, vocabulaire spécialisé et figures de style.

5.2.3.3. La classe Auteur

Sémantiquement, la classe Auteur est destinée à recenser toute personne ayant écrit un texte qui se trouvera dans notre base de textes à usage pédagogique. Cela ne concernera pas les personnes qui éditeront un texte déjà existant.

5.2.3.3.1. La clé

La clé est composée des trois attributs : Nom, Prénom et Dates. Les contenus des deux premiers sont évidents. Dates contiendra la date de naissance et le cas échéant, la date de décès. Nous estimons que ces trois traits sont tout à fait suffisants pour identifier une personne dans le cadre de ce travail. On aurait peut être même pu se limiter aux Nom et Prénom, mais il y a toujours le risque de se trouver avec deux homonymes.

5.2.3.3.2. Autre attribut

La nationalité nous a paru importante dans le cadre d'une base multilingue, de surcroît dédiée à l'apprentissage des langues. Comme nous le faisait remarquer Sonia Tendero au cours des entretiens, si elle étudie avec ses élèves un texte de Neruda, elle en profitera pour leur parler du Chili. La nationalité de l'auteur pouvant permettre d'aborder un aspect culturel (civilisation), cela peut influer dans le choix ou la recherche d'un document et ce même si le texte ne traite pas directement du pays concerné. Nous avons considéré que dans le cas d'un auteur ayant une double (ou une triple) nationalité, il faudrait que cela soit possible de toutes les lister.

5.2.3.4. La relation entre les classes Auteur et Texte

La relation porte le nom "*écrit* >" dans la mesure où l'auteur écrit un texte. Pour figurer dans la base un auteur aura écrit au minimum un texte d'où la multiplicité 1..* : un auteur a écrit au moins un texte. Au début, nous pensions ne permettre qu'un auteur et interdire l'entrée d'un texte sans son auteur. Il apparaît cependant que dans certains cas un texte pourra avoir plusieurs auteurs (cas qui se retrouve particulièrement dans le monde des publications de recherche). Ensuite nous avions pensé qu'une multiplicité 1..* serait appropriée, mais il

n'est pas toujours évident de connaître l'auteur d'un texte (textes pédagogiques pris dans des méthodes, textes très anciens…), d'où la multiplicité 0..* (bien que, dans les faits, un texte ait forcément un auteur).

5.2.3.5. La classe source

On a vu que lors du processus de recherche utilisé par les enseignants, le choix de la source jouait un rôle important.

5.2.3.5.1. Attributs

Les attributs seront tous les traits qui peuvent intéresser l'enseignant dans sa recherche :

- Le type : Il doit permettre de différencier un périodique, d'un « livre », d'une base de textes déjà numérisés ou d'un recueil de textes (les directives d'utilisation de ces termes seront explicitées plus tard)
- Titre : Ce sera le nom dans le cas d'un périodique, ou effectivement le titre dans les autres cas.
- PaysDePublication : Ce champ sera particulièrement utile dans le cas des périodiques. Il sera juste une indication bibliographique dans les autres cas.
- Date : Date de parution, particulièrement importante dans le cas des articles de journaux. Aussi bien Alice Henderson, que Maria-Elena Galoppo ou Iwona Puchalska faisaient état de la nécessité de s'appuyer sur la date dans la recherche d'articles « *d'actualité* ». Maria-Elena Galoppo recherche principalement des articles très récents pour parler d'actualité, pour Alice Henderson cela dépend du sujet traité. Iwona Puchalska nous a donné l'exemple d'une activité au cours de laquelle elle recherchait des articles sur les élections qui avaient eu lieu en Pologne deux ans auparavant pour en discuter avec ses élèves.

5.2.3.5.2. La clé

Ici encore nous avons choisi un identificateur "*artificiel*", les types de sources étant différents, il est difficile de trouver une clé qui les regroupe tous et il est plus simple d'avoir recours à un identificateur.

5.2.3.6. La relation entre les classes Source et Texte

Un texte sera tiré de 0 ou une source : Si un enseignant doit écrire lui-même un texte et le rentre dans la base, il ne sera pas associé à une source. Nous excluons ensuite le cas d'une même version d'un texte provenant de plusieurs sources, partant du principe que si le texte figure déjà dans la base, on ne l'ajoutera pas une seconde fois (pour éviter une trop grande redondance d'informations).

5.2.3.7. La classe associative Modification

Cette classe permet de préciser les aspects qui entrent en jeu pour décrire une modification.

5.2.3.7.1. Multiplicité de la relation

Un texte original pourra ne jamais être modifié, ou bien être modifié plusieurs fois. Par contre, un texte résultat (le texte après modification) ne pourra être issu que d'un seul texte original. Si un texte subit plusieurs modifications successives, les états intermédiaires seront stockés, on pourra accéder au texte original par transitivité. Par exemple soit un texte T. Pour l'utiliser avec sa classe, un enseignant décide de remplacer quelques expressions idiomatiques par des tournures plus simples, pour le rendre accessible. Le texte obtenu après modification T' est conservé dans la base et peut à son tour être modifié pour obtenir le texte T''. Chaque texte résultat n'est relié qu'à un seul texte original.

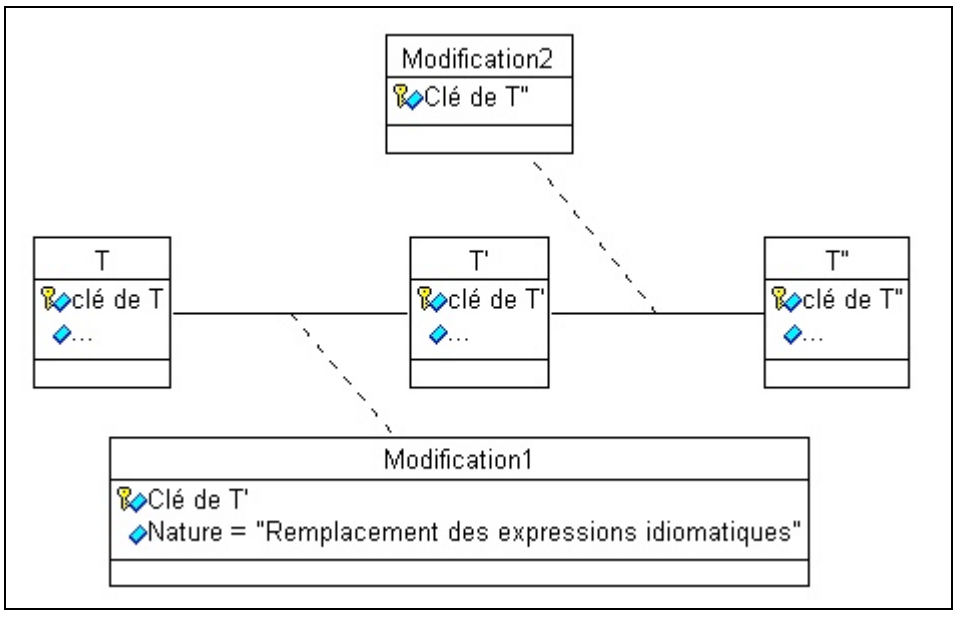


Figure 13 Diagramme d'objets (3 textes pris dans la base, dont deux dérivent du premier) exprimant la transitivité de la relation de modification

5.2.3.7.2. Clé

La clé de la modification, sera une clé relative. En effet, la multiplicité de la relation, nous indique que le fait de pouvoir identifier le texte produit sera suffisant pour identifier la modification. Un texte produit ne pouvant être la modification que d'un et un seul texte original, il n'y aura qu'une seule modification qui aura ce texte comme texte produit. Attention, cela ne signifie pas le texte produit d'une modification, ne pourra pas pourra se retrouver comme texte original d'une autre modification (voir diagramme d'objets du paragraphe précédent).

La clé de la modification sera donc l' "*IdentificateurTexte*" du texte produit.

5.2.3.7.3. Autres attributs

Dans ce cas précis, il nous est apparu que deux attributs étaient amplement suffisants.

- Nature : la nature de la modification a été ajoutée, plus à des fins de traçabilité qu'à des fins de recherches. Pour la recherche, les enseignants procèderont de la même manière que pour un texte normal. Ce champ est essentiellement présent pour leur permettre ensuite de faire leur choix.
- Responsable : toujours à des fins de suivi de l'évolution d'un texte, nous avons inclus un attribut responsable.

5.2.4. Implantation

Pour la création de la base de textes, nous suggérons l'utilisation de la TEI pour l'encodage des documents. Il y a plusieurs raisons à cela.

La dtd[44] proposée par la TEI est adaptable aux besoins de chaque projet et la documentation sur les lignes directrices (*guidelines*) [TEI 02] fournit des directives pour la modification de la dtd. Si certains aspects qui nous paraissent nécessaires sont négligés dans le cadre de la TEI, nous avons la possibilité de modifier le schéma d'encodage en conséquences.

La TEI est utilisé dans au moins 107 projets d'encodage de documents dans le monde. Ces 107 projets recouvrent une quarantaine de langues [TEI 02]. Certains projets sont même multilingues. Le nombre de projets suggère la possibilité d'accéder à une grande quantité de documents informatisés. La grande variété de ces projets peut permettre de n'adopter que

[44] Document Type Definition : spécification de la structure d'un document XML ou SGML.

ceux qui utilisent la TEI d'une manière suffisamment proche de la notre, pour diminuer la quantité de travail nécessaire à l'encodage d'un document.

En outre, si l'expérience prouve que la TEI n'est pas utilisable dans notre cas (ce qui serait contraire aux intentions des équipes à l'origine de la TEI), il y aura malgré tout certains aspects dont nous pourrons nous inspirer.

Enfin, le fait de stocker toutes les données dans la version électronique du texte va permettre une meilleure évolutivité de la base, puisque cela diminue la dépendance entre l'application et les données.

Nous ne prendrons en compte que l'en-tête TEI (TEI header), puisque l'annotation du document dépendra directement de l'utilisation que pourront en faire les outils. Cette dernière nécessitera, comme mentionné lors de la présentation du sujet, toute une étude supplémentaire.

La prise en compte de la TEI implique la création de nouveaux champs pour satisfaire les "*guidelines*" P4. On détaillera ces nouveaux champs dans un paragraphe qui leur sera dédié.

5.2.5. Ajout d'un texte dans la base

Maintenant que nous avons extrait les champs qui seront utilisés dans la base, nous pouvons nous intéresser aux valeurs qu'ils peuvent prendre et aux directives à donner afin que la base soit aussi cohérente que possible. Lors de l'ajout d'un texte dans la base, certaines tâches pourront être prises en charge automatiquement par des outils informatiques, nous allons donc séparer les champs en deux grandes familles : les champs qui seront entrés manuellement et ceux qui seront entrés automatiquement.

5.2.5.1. Champs laissés à la responsabilité de l'utilisateur

Les champs entrés manuellement, sont les champs pour lesquels seul l'utilisateur est à même de renseigner. Pour chaque champ concerné, les lignes directrices pour la saisie des données sont indiquées. Elles doivent être très précises, pour rendre la saisie le moins ambiguë possible et pour permettre que la base soit cohérente.

5.2.5.1.1. Auteur

5.2.5.1.1.1. Nom et prénom

On a séparé les nom et prénom des auteurs afin de ne pas avoir d'ambiguïté due au formatage des données : alors l'être humain arrivera à faire la relation entre "Alessandro Baricco" et "Baricco, Alessandro", cela sera problématique pour un programme informatique, d'où la nécessité d'imposer certaines contraintes.

Les initiales seront reportées dans le prénom, ainsi "Homer J. Simpson" sera répertorié : "Nom : Simpson, Prénom : Homer J."

L'interface devra permettre à l'utilisateur d'avoir accès aux auteurs qui ont déjà été entrés dans la base (on peut imaginer avoir recours à un processus analogue à celui de la saisie semi-automatique de *microsoft Internet explorer*), ce sera ensuite la responsabilité de l'utilisateur de s'assurer que l'auteur qu'il s'apprête à ajouter, n'y figure pas déjà sous une autre forme. Ceci a pour but d'éviter les doublons. Prenons l'exemple de "T.S. Elliot". S'il a été entré une fois sous la forme "Nom : Elliot, Prénom : T.S.", il ne faudra pas l'inscrire une nouvelle fois d'une autre manière : "Nom : Elliot, Prénom : Thomas Stearns".

5.2.5.1.1.2. Dates

L'interface devra faire appel à des menus déroulants, de manière à prévenir toute ambiguïté au niveau du format (20/06/03 ≠ 20 Juin 2003 ≠ 2003-06-20 ≠ 20/06/2003). Si la date de décès n'est pas renseignée, alors l'auteur est toujours en vie.

5.2.5.1.1.3. Nationalité(s)

Contiendra la liste de toutes les nationalités de l'auteur (cf. diagramme de classes et explications liées).

5.2.5.1.2. Source

5.2.5.1.2.1. Type

En renseignant le type de source, nous ne renseignons pas le type de texte contenu. Si le type de source permet d'inférer sur le contenu, c'est à travers la connaissance du monde de l'utilisateur. En effet, comme le faisait remarquer Maria-Elena Galoppo (voir paragraphe sur les sources dans le compte-rendu des entretiens), on peut trouver des textes très littéraires dans des périodiques.

Pour cette typologie, nous nous sommes inspiré de la typologie développée dans la TEI pour les sources [TEI 02]. On ne va donc autoriser dans le cadre de cette étude que quatre valeurs :

- Périodique
- Livre
- Recueil
- Source Numérique

La différence entre un livre et un recueil réside dans le fait que dans un livre les différents chapitres sont interdépendants, alors que dans un recueil, chaque sous-partie est autonome (même si un recueil n'est pas effectué au hasard et qu'il existe un lien entre les sous-parties, que ce soit par le sujet qu'elles traitent, l'auteur, le type de textes réunis).

Le type Source Numérique concerne tous les textes qui ont été trouvés sur Internet, reçus d'organismes travaillant à l'encodage électronique de textes. Tout texte prélevé sur un support numérique satisfera cette catégorie.

5.2.5.1.2.2. Titre

Le titre de la source sera dans le cas d'un périodique le nom du périodique, dans le cas d'un recueil ou d'un livre le titre, dans le cas d'un site web, le nom du site et dans celui d'une initiative d'encodage de textes, le nom du projet.

5.2.5.1.2.3. Date

La date sera la date de parution pour le périodique, le livre ou le recueil.

Dans le cas d'un site web, il est malheureusement rarement possible de connaître la date de mise en ligne. Dans le meilleur des cas, on pourra savoir la date de la dernière mise à jour du site, ce qui ne donne pas d'information sur la date de mise en ligne d'une page précise du site. Par contre, il peut être intéressant de savoir à quelle date le site fut visité ; dans le cas où le site ne serait plus disponible, la date de visite permettra de savoir si l'on a des chances de pouvoir le retrouver (si la visite était récente, le problème peut être ponctuel, alors que si la visite est plus ancienne, il y a de fortes chances que le site ait été fermé).

Dans le cas d'un document électronique non prélevé sur Internet, la date de création du document devrait être disponible, c'est donc celle-là qui sera stockée.

Pour la date, on fonctionnera de la même manière que pour les dates des auteurs, à ceci près qu'ici seule l'année sera obligatoire, puisque si la date de publication d'un ouvrage fait parfois état du mois de publication, le jour a rarement un sens.

5.2.5.1.2.4. PaysDePublication

Ce champ sera obligatoire pour tous les types de sources, sauf pour les sources électroniques. En effet, nous ne pouvons pas être sûr que l'utilisateur sera en mesure de renseigner le champ pour ces sources. Dans le cas des sites web, le contenu et le domaine peuvent aider (URL terminant en : .fr → France, .it → Italie, .pl → Pologne, .ec → Equateur, .uk → Royaume Uni, .es → Espagne, (.com, .net ou .org) → pas d'indication…). Pour le cas d'initiative d'encodage de texte, le pays de publication sera celui du laboratoire dans lequel est basé le projet. Cependant, s'il n'est pas possible d'être sûr du pays concerné, il pourra rester non-renseigné.

5.2.5.1.3. Texte

Pour décrire un texte, nous allons avoir recours à tous les champs suivants.

5.2.5.1.3.1. Type

La première question à se poser ici est de savoir quelle typologie adopter. Si l'on consulte la littérature, la typologie qui revient le plus régulièrement est celle que nous avons retrouvé chez Christine Tagliante, à savoir : narratif, informatif, explicatif, injonctif, argumentatif ou descriptif, une typologie qui rejoint en substance celle proposée par Jean-Michel Adam [ADA 85].Cette typologie est destinée à décrire la dominante d'un texte que l'on pourra compléter à l'aide de sous-dominantes [TAG 94]. Cela suggère que son utilisation, lors de l'ajout d'un texte dans la base, pourra générer des ambiguïtés. Marie-Claude Albert fourni d'ailleurs un exemple illustrant les problèmes liés à une telle typologie. Elle prend l'exemple d'un texte de Voltaire (article « Guerre », paru chez *Classiques Garnier*, en 1967). D'après son analyse, le texte peut aussi bien être rangé sous l'étiquette "*texte narratif*" que sous celle de "*texte argumentatif*". Cette typologie ne satisfait donc pas notre critère de non-ambiguïté.

En outre cette typologie n'a été évoquée au cours d'aucun entretien, nous nous orientons donc vers les types de textes mentionnés. Il faudrait donc réunir sous une terminologie commune les différents types mentionnés en tentant de ne pas trahir les intentions des enseignants interrogés et s'assurer que la typologie ainsi créée est concrète et n'est pas source

d'ambiguïtés. Il n'est pas possible, dans le temps réservé à notre étude, de réaliser un travail rigoureux dans ce domaine et de traiter les autres aspects liés à la création de la base de textes indexée pédagogiquement. Pour faire, un travail rigoureux, il faudrait partir des résultats que nous avons ici, organiser un réel sondage qui ferait intervenir un nombre d'enseignants bien supérieur à celui que nous avions ici et ne s'articulant qu'autour de cette notion de typologie de textes. Une fois la première typologie établie, il faudrait la confronter à nouveau aux enseignants déjà consultés pour s'assurer d'être en accord avec leurs pratiques. Il faudrait affiner la typologie au fur et à mesure des entretiens tout en restant fidèle à la "*philosophie*" de la base : une bonne typologie pour l'ajout de textes dans la base devra effectuer une partition des textes de la base (un texte sera forcément conforme à un type et un seul type) et surtout devra être adaptée aux attentes des enseignants. Dans cette optique nous réalisons ci-dessous un exemple de ce qui pourrait être une typologie. L'exemple a uniquement vocation d'idée et présente certains inconvénients, que nous détaillerons après l'avoir exposé.

Notre exemple consiste à exprimer les caractéristiques des types de textes énoncés, afin que l'utilisateur se concentre sur des faits concrets. Nous remplaçons le champ type par une relation d'héritage et d'autres champs plus concrets :

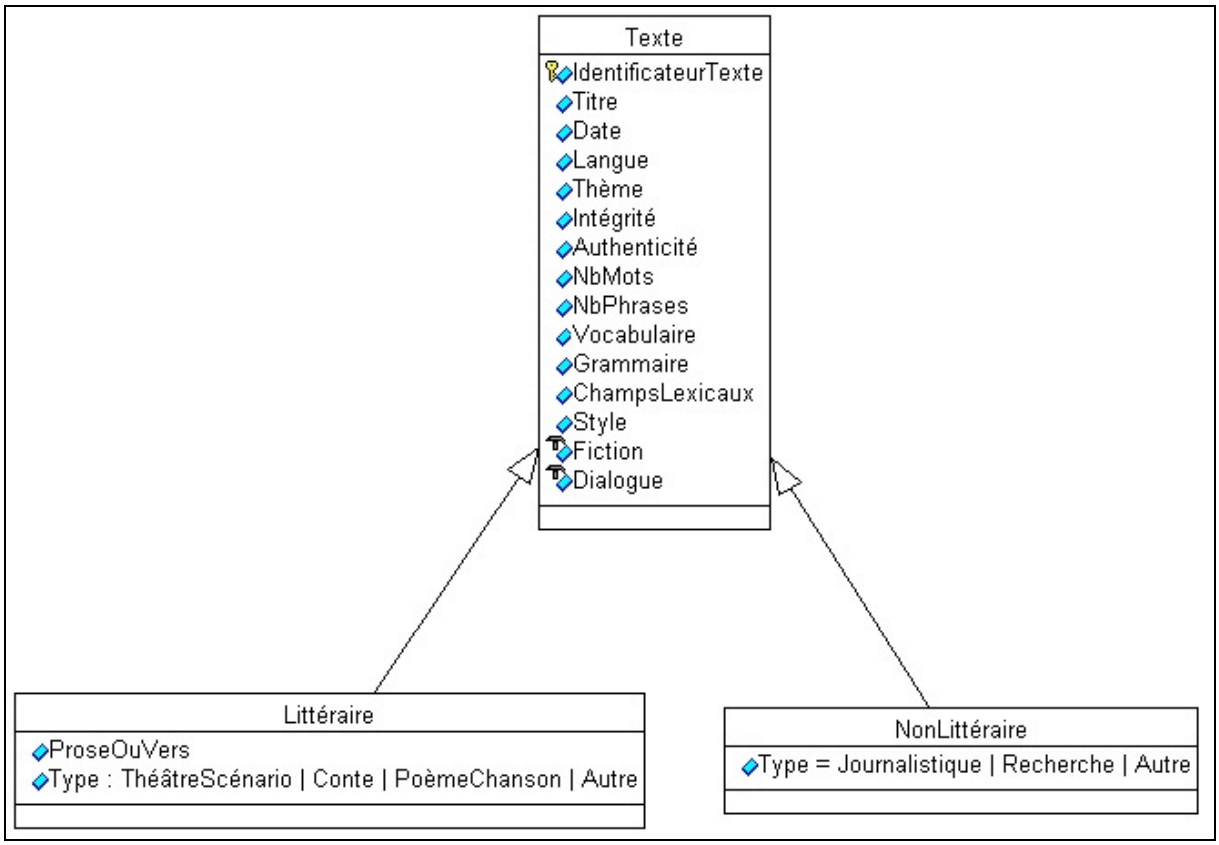


Figure 14 Spécialisation de la classe texte.

Les champs remplaçant le type dans la classe texte ont été marqués d'un . En fonction du fait qu'un texte est littéraire ou non, nous n'examinerons pas les mêmes propriétés, cela n'a, par exemple, pas de sens de se demander si un texte non littéraire est écrit en prose ou en vers. Parmi les termes employés au cours des entretiens, nous avons entendu, entre autres, "*dialogue*", qui désignait un texte d'une méthode d'anglais. Pour le retrouver, il faudra en entrer les caractéristiques : Texte NonLittéraire("dialogue : oui | Fiction : oui | Authentique : non | type : autre").

Les inconvénients que nous mentionnions sont les suivants. Tout d'abord, la notion de texte littéraire n'est pas aussi évidente qu'elle en a l'air, quels critères allons-nous mettre en place pour la caractériser ? Ensuite, que ce soit dans le cas des textes littéraires ou des textes non littéraires, nous avons eu recours à un cas "*fourre-tout*", qu'il faudra à tout prix éviter dans la réalisation de la véritable base de textes. Ce type de pratique est générateur de "*bruit*" (présence de résultats non pertinents) [FLU 00].

5.2.5.1.3.2. Titre

Il faudra s'assurer de donner au texte son titre original (et non une traduction).

5.2.5.1.3.3. Date

Nous avons déjà évoqué les directives concernant la date lorsque nous avons décrit les différents attributs du texte. La date sera donc, si elle est connue, celle de l'écriture du texte, puisqu'il peut y avoir un laps de temps conséquent entre l'écriture et la première parution d'un texte.

5.2.5.1.3.4. Intégrité

Deux valeurs sont possibles pour l'intégrité : vrai ou faux. Si le texte a été tronqué de quelque manière que ce soit avant son ajout dans la base, l'intégrité sera initialisée à faux.

5.2.5.1.3.5. Authenticité

Comme l'intégrité, l'authenticité est une valeur binaire qui traduira la différence entre un texte authentique (champ initialisé à vrai) et un texte « *pédagogique* » (champ initialisé à faux), au sens où nous avons défini ces deux notions dans le paragraphe concernant la terminologie.

5.2.5.1.4. Modification

5.2.5.1.4.1. Nature

La nature d'une modification ne sera pas un champ de recherche, mais une manière de savoir par quelles étapes est passé un texte. Elle pourra donc être écrite en langage naturel par l'auteur de la modification.

5.2.5.1.4.2. Responsable

Ce sera le nom de la personne qui a effectué la modification. Ce champ a lui aussi une fonction descriptive.

5.2.5.2. Champs entrés automatiquement

Le reste des champs sera pris en charge automatiquement par des outils. Parmi les champs qui seront entrés automatiquement par le système, nous en distinguons deux classes. Nous appellerons la première classe, celle des champs concrets, dont les valeurs seront entrées telles qu'elles seront utilisées. Il ne sera pas possible de traiter tous les champs de cette manière, à moins de perdre de vue certains des objectifs que nous nous étions fixés (comme celui de faire de la base un outil le plus conforme possible aux pratiques des enseignants).

Les autres champs sont utilisés différemment : lors de l'entrée d'un texte dans la base, des informations concrètes sont répertoriées. Cependant ces informations seraient trop fastidieuses à utiliser telles quelles. Pour que la base de texte soit utilisable, il faut donc réaliser une couche logicielle intermédiaire qui permette de faire le lien entre les aspects pédagogiques mis en œuvre au cours de la requête et les données concrètes stockées dans la base.

5.2.5.2.1. Champs concrets

5.2.5.2.1.1. Identificateurs

Tous les champs que nous avons appelés identificateurs, seront entièrement pris en charge par le système. Ceci ne posera aucun problème d'un point de vue informatique et cela évitera la présence de doublons qui pourrait découler de l'intervention de l'humain dans cette tâche.

Ces identificateurs n'auront de valeur qu'au sein de la base de textes et pourront éventuellement être utilisés pour mettre en relation les différents documents répertoriés.

5.2.5.2.1.2. Mesures

Par mesures, nous entendons la taille d'un texte en mots ou en nombre de phrases. Ces deux mesures devront être effectuées automatiquement, car elles seraient beaucoup trop fastidieuses pour un être humain. De plus, ce sont des tâches pour lesquelles les outils TAL les plus basiques sont efficaces dans le contexte de ce travail.

5.2.5.2.1.3. Intégrité / Authenticité

Ces champs ne seront pas entrés automatiquement, mais mis à jour automatiquement : le système n'a pas de moyen de savoir lorsqu'un texte est entré dans la base s'il a été modifié au préalable ou non. Par contre lors de l'édition d'un texte déjà présent de la base, il sera aisé de s'assurer que le système passe la valeur des champs à "*faux*".

La valeur initiale des champs intégrité et authenticité sera entrée par l'utilisateur, qui est le seul à même de le faire. Mais après toute édition d'un texte déjà présent dans la base, les champs intégrité et authenticité du texte résultat sont mis à jour automatiquement.

5.2.5.2.2. Champs dont l'utilisation nécessite une couche logicielle intermédiaire

Avant de nous concentrer sur les champs, il convient de rappeler certains principes.

5.2.5.2.2.1. Principes généraux de l'indexation

Nous n'allons pas rentrer en profondeur dans le phénomène d'indexation, mais juste en donner les rudiments de manière à ce que la suite soit compréhensible. Ce paragraphe tire la plupart de ses informations de Indexation et recherche d'information textuelle [FLU 00].

Comme l'index dans un ouvrage papier, l'indexation sert à retrouver des informations dans un texte. Il existe plusieurs techniques d'indexations, manuelles ou automatiques.

En règle générale, l'indexation automatique d'un texte s'effectue selon le processus suivant : le texte est parcouru, au fur et à mesure du parcours, les informations que l'on veut indexer sont relevées, il s'agit en général de formes fléchies de la langue. Dans l'indexation de textes intégraux, on ne tiendra pas compte « *des mots outils dits ''mots vides''* » [FLU 00]. Dans le modèle vectoriel, le traitement effectué est une analyse statistique concernant le nombre d'occurrences de chaque mot.

Il existe en outre un processus d'indexation manuelle avec vocabulaire contrôlé, qui ne sera pas utilisée dans le cadre de l'indexation de la base à partir de critères pédagogiques,

puisque ce processus doit être effectué par un public aguerri à cette technique, ce que l'on ne peut pas demander aux utilisateurs de la base.

Il est à noter que si l'indexation est possible en fonction de mots du langage naturel, elle l'est tout autant à partir de n'importe quelle autre donnée.

5.2.5.2.2.2. Grammaire

Pour pouvoir relever les structures présentes dans un texte, il faudra nécessairement passer par une analyse syntaxique. Une fois cette analyse effectuée, l'indexation devra se faire en fonction des structures grammaticales présentes dans le texte. Cette pratique constitue, elle aussi et à elle seule, un sujet de recherche. Elle va nécessiter la création d'un formalisme permettant de décrire les différentes structures grammaticales d'une langue donnée (en effet le formalisme dépendra vraisemblablement de la langue). Il est possible que l'utilisation du formalisme issu de l'analyse syntaxique, assorti de mécanismes analogues à ceux de recherche des concordances, soit suffisante dans notre cas. Cependant, seule l'expérience pourra le dire. Et même une fois ce problème réglé, il faudra encore se pencher sur la manière de formuler les requêtes.

A priori l'annotation dépendra de l'outil utilisé et devra représenter le résultat de l'analyse syntaxique du texte.

5.2.5.2.2.3. Thème, Vocabulaire, ChampsLexicaux

L'analyse syntaxique nécessaire au recensement des structures grammaticales dans le texte nécessite une analyse morphologique [BOU 98] qui pourra être utilisée dans tous les champs mentionnés ci-dessus.

D'après Fluhr [FLU 00], l'un des problèmes liés à l'indexation automatique de documents réside dans la présence de formes fléchies de mots : les résultats seront bien meilleurs si toutes les formes conjuguées d'un verbe sont considérées comme un seul et même mot. L'indexation pourra, comme une analyse morphologique sera effectuée de toute façon, ne considérer que les formes lemmatiques.

En ce qui concerne le thème, la recherche sera faite par l'utilisateur à partir de mots-clés, ce qui pose le problème de la synonymie et de la polysémie, problème auquel on remédie par l'utilisation de thésaurus dans le cas de l'indexation manuelle avec vocabulaire contrôlé.

Cependant il semblerait que le choix des descripteurs puisse être automatisé avec un certain succès[FLU 00]. On peut donc imaginer utiliser cette technique pour l'indexation du thème.

Cette approche ne sera pas utilisable pour la description du vocabulaire qui nécessite l'indexation de chacun des mots puisque si un apprenant connaît un mot, il n'en connaît pas pour autant tous les synonymes. Les indexations du texte selon le vocabulaire présent et le thème seront donc indépendantes, à ceci près qu'elles utiliseront, toutes deux, la forme lemmatique des mots.

La recherche de champs lexicaux ne nécessite pas l'ajout d'autres informations dans le texte, celle-ci peut être gérée par le biais des requêtes.

Pour tout ce qui touche aux champs Vocabulaire, Thème et ChampsLexicaux, les données qui devront être ajoutées dans le texte (annotations) sont les données relatives à l'analyse morphologique, dont le lemme.

5.2.5.2.2.4. Langue

La reconnaissance automatique de la langue, fonctionne dans certains cas. Un système comme *Microsoft Word* ou SILC [SILC] est capable de reconnaître automatiquement la langue d'un texte (à partir du moment où celle-ci est prise en charge). Deux problèmes peuvent malgré tout se poser.

La langue du texte n'est pas prise en charge par les outils. On peut exclure à priori ce cas, puisque l'ajout d'un texte dans une langue qui n'est pas gérée par la plate-forme, n'aura pas de sens.

Problème des variétés d'une langue : comment différencier deux variétés d'une même langue ? Il est important que la variété figure, puisqu'elle pourra être un critère pour le choix d'un texte (civilisation, activité lexicale). Or, que ce soit pour un humain ou pour un programme, il sera parfois difficile d'identifier la variété d'une langue, en particulier à l'écrit (puisqu'on ne dispose pas des informations données par la prononciation).

Nous avons pensé à un moment utiliser la nationalité de l'auteur pour décider de la variété d'une langue, mais des cas comme celui de Jorge Semprun nous l'interdisent. Jorge Semprun est un auteur espagnol, mais qui a écrit plusieurs livres en français, devons-nous en déduire que la langue employée est le français d'Espagne ? Utiliser la nationalité de l'auteur pour décrire la variété d'une langue constituerait donc la porte ouverte à certaines inepties.

La solution que nous avons décidé de mettre en place est la suivante : la reconnaissance de la langue fonctionnera automatiquement, ensuite pour la variété, il faut revenir sur le but d'une telle information. On y voit principalement deux intérêts : l'intérêt en terme de civilisation ; l'introduction de structures grammaticales et, surtout, de vocabulaire spécifique à une région.

En ce qui concerne la présentation de la civilisation, c'est essentiellement dans le choix du thème que l'on retrouvera cette composante.

Pour les structures grammaticales ou le vocabulaire, on ne cherchera pas forcément un texte entier dans la variété considérée, mais juste un texte où figurent certaines expressions ou structures. C'est, par exemple, la manière de procéder dans la méthode d'espagnol Ven [C-M-M-R 00]. Nous aborderons le problème sous une nouvelle approche : la langue ou plutôt les langues (puisqu'il ne faut pas exclure les textes qui utiliseraient plusieurs langues) seront regroupées dans un champ concret qui contiendra la langue ainsi que le pourcentage du texte concerné (grâce aux outils de comptage des mots).

Enfin, pour ce qui intéresse les variétés, chaque tournure typique d'une variété de la langue sera annotée en conséquence dans le texte. Ces annotations pourront ensuite être utilisées pour l'indexation du texte, permettant ainsi une recherche selon ce critère. Par exemple : « *Ella me dijo que es una vida buena allá. Bien rica, bien <variété lieu="Equateur">chévere<variété>.*»[45] Il n'est pas garanti que la balise garde ce nom, mais en substance l'annotation sera conforme à ce schéma. Un outil reconnaît la langue : l'espagnol. Un autre passe en revue le vocabulaire et remarque la présence du mot « *chévere* » qui n'existe pas en castillan mais qui est très couramment employé en Equateur. Pour représenter le cas où le mot est utilisé dans d'autres régions (c'est probablement le cas pour « *chévere* » mais on se contente de l'Equateur et de Puerto Rico pour ne pas écrire d'erreur), l'attribut lieu contiendra une liste de lieux séparés par "|" : *<variété lieu="Equateur|Puerto Rico">chévere<variété>*. Pour indexer ces phénomènes, chaque lieu présent dans une balise sera conservé, une requête indiquera tous les textes contenant ne serait-ce qu'une expression typique d'un lieu donné.

[45] The Pixies (paroles et musique : Black Francis) : "Vamos", tiré de l'album "SURFER ROSA", 4AD, ©1988.

5.2.5.2.2.5. Style

Nous avons extrait au cours des entretiens trois composantes du style : la présence de formes idiomatiques, celle de vocabulaire spécialisé et la présence d'effets stylistiques. Pour les deux premières, il existera (ou il sera possible de réaliser) des outils qui pourront les localiser dans un texte. Chaque fois qu'une telle forme sera rencontrée dans le texte, elle sera marquée par une balise. Nous n'indiquons pas ici le nom de la balise puisqu'elle existe éventuellement déjà dans TEI, si ce n'est pas le cas nous devrons la créer.

Un processus analogue serait souhaitable pour les effets de style, mais l'existence d'un outil les détectant est plus problématique. Demander à l'enseignant qui ajoute le texte dans la base de baliser chaque effet stylistique serait bien trop fastidieux. La conservation de ce critère dépend donc de ce qui existe dans le domaine en termes d'outils.

5.2.5.3. Structure de l'en-tête TEI

En s'appuyant sur la spécification de l'en-tête TEI (TEI-header P4) [TEI 02], nous avons établi la structure suivante pour l'en-tête de nos documents. Les balises sont celles de TEI. Les noms des champs de notre base seront précédés d'un ◆ et indiqueront qu'il existe une équivalence entre la valeur à reporter et la valeur du champ. Les balises rouges représentent celles qui demandent une modification de la dtd de l'en-tête TEI (modifications conforme aux lignes directrices définies dans [TEI 02]). Les commentaires sont entre /* et */ et écrits en vert.

```
<teiHeader>
      <fileDesc>
            <titleStmt>
                  <title>une version électronique de ◆Texte::Titre</title>
                  <author>◆Auteur::Nom, ◆Auteur::Prénom (◆Auteur::Dates),
(◆Auteur::nationalité)</author>
                  <respStmt>
                        <resp>Ajout du texte dans la base</resp>
                        <name>nom de l'utilisateur ayant rajouté le
texte</name>
                  </respStmt>
            </titleStmt>
            <extent>◆Texte::NbMots mots</extent>
            <extent>◆Texte::NbPhrases phrases</extent>
            <publicationStmt>
                  <distributor>MIRTO</distributor> /*sujet à modification
en fonction de details légaux*/
            <publicationStmt>
            <seriesStmt>
                  <title level="s">MIRTO</title>
                  <idno type="MIRTO-Txt">◆Texte::IdentificateurTexte</idno>
            </seriesStmt>
            <sourceDesc>
                  /*pas de source*/
```

```
                  <p>pas de source, créé pour la base</p>
                  /*source électronique :
                        si TEI, on recopie ici le contenu du file desc de
la source, si non TEI :*/
                  <bibl>
                        <title>Source::Titre</title>
                        <date value="Source::Date"/>
                        <pubPlace>Source::LieuDePublication</pubPlace>
                        <idno type="MIRTO-Src">
                        Source::IdentificateurSource</idno>
                  </bibl>
                  /*source non électronique*/
                  <biblStruct>
                        <analytic>
                              <title>une version électronique de
Texte::Titre</title>
                              <author>Auteur::Nom, Auteur::Prénom
(Auteur::Dates), (Auteur::nationalité)</author>
                        </analytic>
                        <monogr>
                              <title level=Source::Type /*'m' si recueil,
'j' si périodique, 'b' si livre*/>Source::Titre</title>
                              <imprint>
                                    <pubPlace>
                                    Source::LieuDePublication</pubPlace>
                                    <publisher>éditeur /*à rajouter comme
champ de la base*/</publisher>
                                    <date value="Source::Date"/>
                              </imprint>
                        </monogr>
                        <idno type="MIRTO-Src">
                        Source::IdentificateurSource</idno>
                  </biblStruct>
            </sourceDesc>
      </fileDesc>
      <encodingDesc>
            /*indications fixes sur le projet et le codage, à définir en
fonction des outils et du format des annotations*/
      </encodingDesc>
      <profileDesc>
            /*pour chaque langue présente dans le texte*/
            <langUsage id=Texte::Langue usage="/*% calculé grâce aux
outils de mesure du nombre de mots*/"/>
            /*pour chaque variété de langue présente dans le texte*/
            <langUsage id=Texte::Langue>(liste de lieux) dont on trouve
des expressions dans le texte</langUsage>
            <keywords>/*s'il ne sont pas trop nombreux pour être gérés
ainsi, nécessité d'une étude sur la meilleure implémentation*/ indexation
en rapport avec Texte::Thème</keywords>
      </profileDesc>
      <revisionDesc>
            <change>
                  <date value=Texte::Date/>
                  <respStmt>
                        <name>Modification::Responsable</name>
                        <resp>Modification::Nature</resp>
                  </respStmt>
            </change>
      </revisionDesc>
</teiHeader>
```

Nous avons montré ici une possibilité de description des textes. Elle peut servir de base à un premier prototype, mais n'est absolument pas définitive.

5.2.5.4. Conclusions sur l'ajout de texte dans la base

Le travail sur cette première version de la base de textes nous a permis d'établir certains champs et certaines stratégies, même si l'étude que nous effectuons ici n'est que provisoire et ne peut constituer que le point de départ de recherches plus avancées.

Nous avons pu remarquer que le champ type pour le texte, bien qu'ayant un sens pour les utilisateurs, ne peut pas être utilisé tel quel dans la base. Le problème qu'il pose peut être élargi à toutes les typologies qui seront nécessaires aux utilisateurs : il sera extrêmement difficile de mettre en place une typologie qui puisse satisfaire tous les utilisateurs potentiels et qui permette de faire un bon champ (partition de l'espace des valeurs). Le travail que nous avons effectué ici pour les classes d'utilisation des textes devra être effectué pour toute typologie utile aux utilisateurs : la première étape concerne le recueil des données concrètes qui sont nécessaires à l'évaluation de l'appartenance d'un texte à telle ou telle classe. Ensuite, nous nous sommes intéressé à la façon de formaliser ces données. Nous n'avons d'ailleurs pas de solution définitive dans le cas présent puisque nous sommes tributaires des outils qui seront utilisés et des technologies d'indexation de documents. Il reste à faire beaucoup de travail dans cette direction. En partant de l'hypothèse qu'il est possible d'indexer les documents selon les critères que nous avons définis ici, nous allons détailler maintenant des idées concernant l'utilisation de ces critères dans le cadre de la recherche de documents.

5.2.6. Recherche d'un document dans la base

L'une des particularités de la base de textes indexée de manière pédagogique réside dans le fait que, de part la nature même des champs, les requêtes ne vont pas pouvoir être effectuées de manière traditionnelle.

Chaque champ constitue un axe de recherche qui peut être associé aux autres champs tout en respectant sa propre logique d'interrogation. Nous allons donc, pour chaque champ, expliquer comment les recherches pourraient être menées.

5.2.6.1. Champs concrets

Nous rappelons que ce que nous appelons champs concrets sont les champs qui peuvent être utilisés tels quels dans une recherche. Dans notre cas les champs concrets sont :

- Tous les champs de la classe source
- Tous les champs de la classe auteur
- Dans la classe texte :
    - L'identificateur
    - Titre
    - Date
    - Langue (qui d'après les remarques que nous avons faites plus tôt deviendra Langue + Variété)
    - Intégrité
    - Authenticité
    - NbMots
    - NbPhrases
- D'une certaine manière, les champs de la classe modification. Cependant, ces derniers ne serviront vraisemblablement pas à faire une recherche. Ils ne seront utilisés, qu'une fois un texte trouvé, pour en suivre l'historique. Nous ne parlerons donc pas de ces champs dans le cadre de la recherche de documents.

Ces champs sont ceux qui se rapprochent le plus des champs traditionnels d'une base de données de par leur utilisation. A chaque champ pourra correspondre dans l'interface un menu déroulant qui ne permettra de sélectionner que des données figurant dans la base.

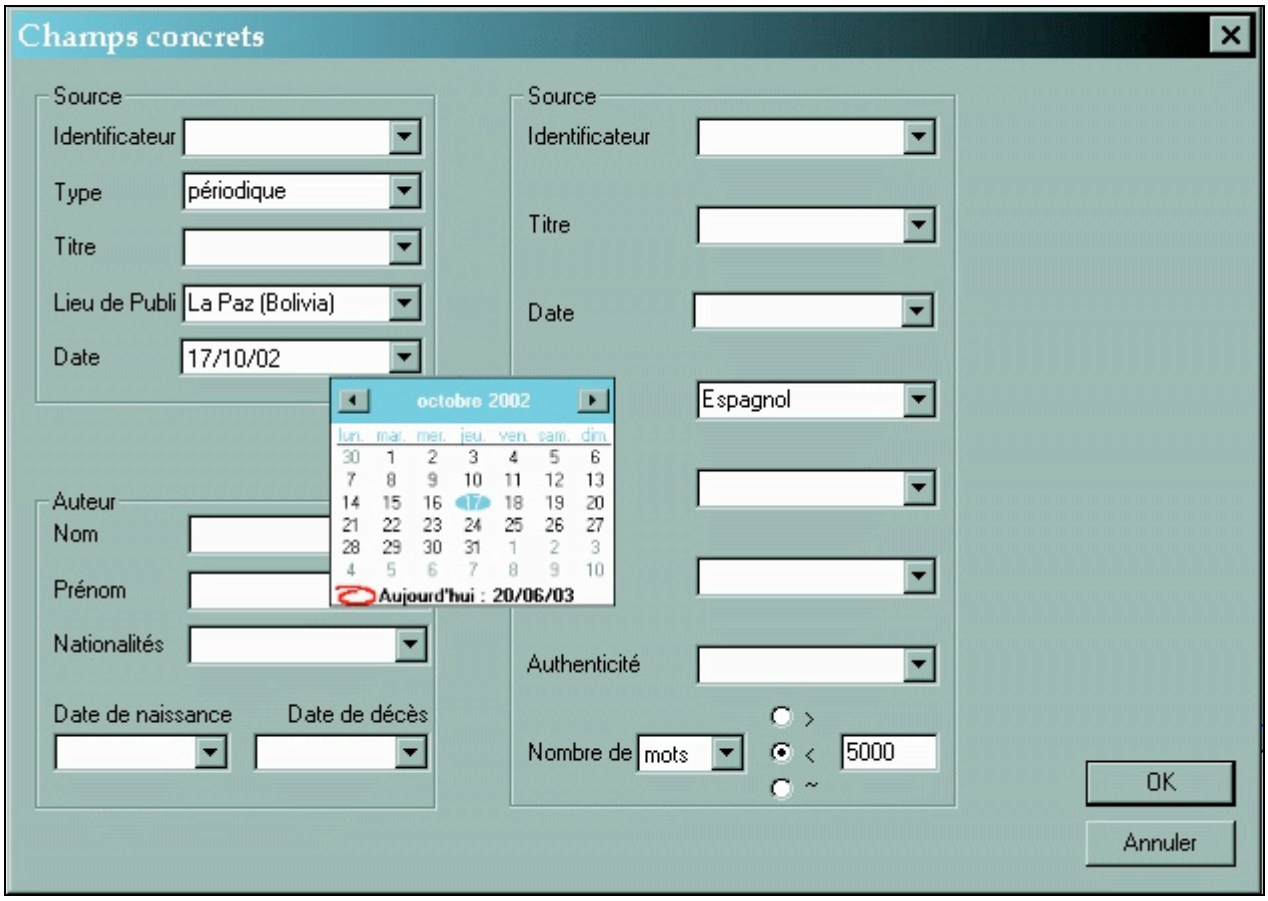


Figure 15 Maquette d'interface pour les champs concrets. Exemple de requête du type : tous les textes de moins de 5000 mots écrits en espagnol et parus dans un périodique de la Paz le 17 octobre 2002

Les recherches liées aux autres champs ne figurent pas ici. Cela tient au fait qu'elles sont plus difficiles à implémenter, ce qui ne signifie en aucun cas qu'elles ne pourront être liées (ex : textes n'utilisant que le présent, écrits en espagnol et contenant moins de 2500 mots). Sur l'interface ci-dessus, ne figurent pas les nouveaux champs correspondant au type de texte, qui n'était qu'indicatifs et nécessiteraient une nouvelle série d'entretiens.

5.2.6.2. Thème

Nous avons pu remarquer que la plupart des enseignants avaient l'habitude d'effectuer des recherches sur Internet. Le moyen le plus simple de leur faire retrouver un texte en fonction de son thème sera donc d'adopter un fonctionnement similaire (par mots-clés). Ceci nécessite l'indexation des documents en fonction de leur contenu. Sur Internet la méthode utilisée est en général la méthode « texte intégral » [FLU 00].

5.2.6.3. Vocabulaire

D'après les données recueillies durant les entretiens, les requêtes en terme de vocabulaire seront du type : doit_contenir {ensemble de mots}, peut_contenir {ensemble de mots}, ne_peut_pas_contenir{ensemble de mots}. Ces trois types de requêtes correspondent à la méthode de recherche d'un texte utilisée par les enseignants : un texte doit contenir le vocabulaire à introduire, il peut contenir ce que les élèves connaissent (les mots, que le texte doit contenir, ne figureront à priori pas dans le vocabulaire que le texte peut contenir puisque l'on introduit en général du vocabulaire inconnu). Ce que le texte ne peut pas contenir est, en règle générale, le vocabulaire que les élèves ne connaissent pas et qui ne doit pas être introduit. On ne se préoccupe ici que de vocabulaire, les mots seront donc dans leur forme lemmatique.

Ceci pose deux problèmes :

- Comment faire pour recenser tous les mots qu'un apprenant est censé connaître ?
- Est-il possible de trouver des textes qui ne contiennent que les mots que l'on recherche ?

5.2.6.3.1. Recensement du vocabulaire

Il n'est bien évidemment pas question, pour les enseignants, de rentrer pour chaque requête l'intégralité des mots que connaissent les élèves. Cependant, une manière de procéder pourrait être d'entrer, pour chaque niveau (au sens auquel cela a été défini pendant les entretiens : intitulé d'un groupe), la liste des textes auxquels ils ont été confrontés. Le système pourrait grâce à ces textes établir les mots qui peuvent figurer dans un texte. Cela sera plus complexe pour les niveaux avancés, mais on pourra considérer qu'un élève d'un niveau donné aura vu tous les textes auxquels ont été confrontés les élèves des niveaux inférieurs.

5.2.6.3.2. Tolérance

Nous n'avons pas effectué d'expériences dans ce sens (il faudrait pour cela que la base soit déjà implémentée), mais nous imaginons qu'il sera difficile de trouver des textes ne contenant que les mots d'une liste. Il faut donc un paramètre tolérance lors des recherches : nous avons vu, qu'en fonction de la classe d'activité d'enseignement, la tolérance est différente (figure 5). Nous avons aussi vu que la langue influait : dans les langues plus proches de la langue maternelle, on tolère plus de nouveau vocabulaire (ex : différence entre le polonais et l'espagnol pour les étudiants francophones). Enfin, le niveau même des

étudiants joue un rôle : plus un étudiant a un niveau élevé, plus on pourra le confronter à du vocabulaire inconnu.

5.2.6.3.3. Idée de mise en œuvre

Une manière pour mettre en place un tel système serait d'utiliser des profils utilisateurs pour les enseignants : un enseignant pourrait ainsi définir les différents niveaux auxquels il enseigne (en leur donnant un ordre croissant en termes de vocabulaire). Il pourrait ensuite définir sa progression en entrant les textes utilisés avec les élèves. L'idéal serait que tous soient entrés dans la base. Cependant, si cela s'avère trop fastidieux, on peut convenir d'envoyer les textes uniquement à l'outil qui établira les connaissances des apprenants. Ultérieurement, l'enseignant pourra lui-même définir la tolérance. Après quoi une requête sera de la forme : trouver avec la tolérance δT, un texte contenant l'ensemble de mots E_1, sachant que le vocabulaire des apprenants se limite au vocabulaire de l'ensemble de textes T. Une autre possibilité serait une requête du type : trouver des textes candidats pour un exercice de compréhension pour un élève du cours intitulé C de tel enseignant.

La base d'historique, contenant pour chaque élève les activités qu'il aura effectuées et menées à bien, pourra mettre à jour en conséquence l'état de ses connaissances présumées (à travers les textes qu'il aura étudiés).

Notons que l'on pourra imaginer un traitement différent pour les mots dits outils.

5.2.6.3.4. Représentation des résultats

Pour afficher les résultats, il faudra surligner d'une certaine manière les mots que le texte devait contenir et ceux que le texte ne devait pas contenir (tolérance). Cela permettra à l'enseignant d'évaluer plus vite le contexte de tous ces mots, afin de savoir s'ils sont dans un contexte adéquat ou s'ils vont gêner la compréhension.

5.2.6.4. Champs Lexicaux et Style

Comme nous l'avons dit au moment de l'ajout des textes, la recherche de champs lexicaux pourra se faire en incluant les mots du champ lexical dans les mots que le texte devra comporter. Pour faciliter le travail, on pourra tenter de rechercher / créer une sorte de dictionnaire de champs lexicaux qui sera une aide à la rédaction de requête. Il est à noter que la tolérance ne concerne pas seulement les mots dont l'utilisateur ne voulait pas qu'ils se trouvent dans le texte, mais aussi ceux qu'il aurait aimé trouver et qui en sont absent.

Le style, tel qu'il nous a été présenté n'est pas un critère de recherche mais de choix. Il doit être encodé lors de l'ajout dans la base c'est pour cela qu'il est resté un champ sur le diagramme, mais lors de la requête il ne sera pas pris en compte. Par contre, lors de la présentation des résultats à l'utilisateur, le système pourra soit présenter des statistiques de style (nombre d'expressions idiomatiques, d'occurrences de mots appartenant à du vocabulaire spécialisé et quel champ : médical, technique…), soit souligner les mots ou séquences de mots mises en cause de manière à ce que l'enseignant les remarque facilement. On peut même imaginer que les deux solutions soient implémentées.

5.2.6.5. Grammaire

Notons que, dans toute cette partie, nous ne traiterons que de l'aspect formulation de la requête. Il y a bien évidemment tout un travail d'indexation et / ou de reconnaissance de structure[46] qui devra être effectué, mais celui-ci dépendra des outils utilisés pour l'annotation, il ne pourra donc pas être traité ici.

Pour les structures grammaticales, nous suggérons de procéder de manière analogue. On ne pourra, toutefois, pas fonctionner avec les textes déjà étudiés, car contrairement au vocabulaire qui est en général traité lorsqu'il est rencontré, dans l'enseignement des structures grammaticales, seules les structures introduites sont traitées. Sonia Tendero nous faisait remarquer, au cours de son entretien, qu'il lui arrivait de dire à ses élèves que telle structure grammaticale exprimait telle notion mais qu'elle serait traitée plus tard. Pour pouvoir dire qu'une structure est assimilée, il ne suffit pas de savoir ce qu'elle exprime, il faut être capable de l'utiliser et de la reconnaître sous toutes ses formes. En effet savoir que « *viajaré* » est, en espagnol, la première personne du singulier du verbe "viajar" au futur, ne signifie pas que l'on maîtrise le futur, pas même pour les verbes du premier groupe.

Nous ne pourrons donc pas utiliser ici les textes auxquels les apprenants ont été confrontés pour exprimer leurs connaissances grammaticales.

5.2.6.5.1. Expression des structures grammaticales

Il faudra de toute manière trouver un formalisme pour exprimer les structures en termes interprétables par la machine. Ce formalisme sera utilisé pour indexer les textes dans la base. Il dépendra directement des outils employés, il ne sert donc à rien d'essayer de l'anticiper ici. Cependant, l'une des caractéristiques de ce type de formalismes est qu'ils sont en général

[46] Reconnaissance de patron / pattern matching

relativement opaques pour les humains et en particulier pour les non-informaticiens. Il est donc exclu de forcer les enseignants à utiliser un tel formalisme.

Trois solutions se présentent alors pour régler ce problème : l'utilisation de titres pour chacune des structures grammaticales, la création d'un formalisme plus instinctif (qui permettrait aux enseignants d'exprimer des calques de structures) et l'utilisation d'exemples.

5.2.6.5.1.1. Nommage des phénomènes grammaticaux

Pour chacune des langues présentes dans la base, il faudra alors trouver un intitulé pour décrire chaque point grammatical.

Exemple : obligation personnelle en espagnol → structures de la forme « tener que + verbe à l'infinitif »

Chaque intitulé sera alors associé dans le système à une expression la décrivant dans le formalisme de la machine. Cette expression sera utilisée par la suite pour effectuer les recherches du texte.

Cette solution présente comme avantage d'être proche de la manière dont les enseignants se représentent ces points grammaticaux. Lors des entretiens, quand un point grammatical était évoqué dans la conversation, il l'était par un intitulé : l'exemple ci-dessus provient de l'entretien avec Sonia Tendero.

Cette solution présente aussi un certain nombre de désavantages. Le premier est que pour chaque langue il faudra recenser, nommer et donner une équivalence avec le formalisme de la machine pour chaque point grammatical pouvant être abordé. Cela demandera un très long travail. En outre, le choix des intitulés risque d'être problématique, dans la mesure où il sera difficile d'adopter une terminologie consensuelle. En anglais par exemple quel terme sera le plus adéquat : prétérit progressif, past continuous, imparfait...

5.2.6.5.1.2. Développement d'un autre formalisme

Comme nous l'avons dit en préambule au compte-rendu des entretiens, quand nous parlons de structures grammaticales, nous ne nous bornons pas à la décomposition syntaxique de la phrase, mais aussi aux règles morphologiques. Le formalisme devra donc donner la possibilité de faire intervenir des mots du lexique et la morphologie aussi bien que des mots clés permettant de décrire les propriétés des structures syntaxiques elles-mêmes.

Sonia Tendero nous expliquait que lorsqu'elle introduisait un nouveau temps elle présentait en général le premier groupe pendant une première phase de son cours, le deuxième et le troisième groupes pendant une seconde phase et les verbes irréguliers au cours d'une troisième. Comment effectuer une recherche en conséquences sans avoir recours à des informations morphologiques ? Nous pouvons aussi revenir à l'exemple de l'obligation personnelle, le même problème se pose, comment décrire la structure sans avoir recours à "*tener que*".

Ce formalisme devra en fait constituer un véritable langage de description de la langue, qui comporte, en plus comme contrainte, d'être plus accessible que ce qui est utilisé pour annoter morphologiquement / syntaxiquement un texte. Nous n'allons pas rentrer dans le détail de ce que devra être le langage, car c'est un problème très complexe, pour lequel nous n'avons pas de réponses.

En revanche il nous est déjà possible de donner le principal avantage de cette manière de procéder : une fois le langage développé, les enseignants seront libres de rechercher toute tournure, toute structure, sans être limités par le cadre de patrons pré-établis.

Mais cette solution aura aussi son lot de désavantages. La prise en compte de données morphologiques fait qu'il faudra développer pour chaque langue un formalisme différent. Et ce, même si certaines parties pourront être réutilisées d'une langue à l'autre. Des données, comme le cas, n'existeront pas en anglais alors qu'elles existeront en allemand.

Un autre désavantage majeur réside dans le fait que les enseignants seront obligés d'apprendre à utiliser le formalisme et même s'il sera forcément moins abscons que celui utilisé par le système. Cette prise en main rendra l'emploi de la base de données beaucoup moins instinctive et plus fastidieuse, ce qui peut détourner certains enseignants de son utilisation.

5.2.6.5.1.3. Par l'exemple

La dernière solution, qui me paraît être la meilleure, consiste à permettre une recherche à partir d'exemples.

L'utilisateur effectue sa requête initiale en fournissant un exemple de structure qu'il aimerait trouver dans les textes résultats. Le système renvoie une liste de séquences candidates. L'utilisateur affine sa recherche en choisissant parmi les séquences candidates

celles qui conviennent et en excluant les autres. Le processus est reproduit jusqu'à ce que l'utilisateur considère les résultats suffisamment bons, après quoi la recherche de document est lancée à partir de la requête ainsi formulée.

Un tel travail peut être effectué en utilisant des réseaux neuroniques[47] : ces derniers peuvent être utilisés pour déterminer l'importance de chaque trait de l'exemple fourni par l'utilisateur. Cette première phase constituerait la phase d'apprentissage. Les réseaux de neurones pourront être utilisés pour le "*pattern matching*" ou non, selon que l'on considère que leur but est de reconnaître la requête ou la structure elle-même.

Cette approche présente un certain nombre d'avantages par rapport aux deux autres. Elle ne nécessite que peu d'adaptation de la part des enseignants qui n'auront pas de problème pour formuler un exemple. Elle ne nécessite pas d'adaptation d'une langue à l'autre : elle ne fait intervenir que des données informatiques qui ne seront pas affectées par le passage d'une langue à l'autre (seuls les outils d'annotations changent). Par contre, on ne peut pas savoir, sans tests préalables, si cette méthode va donner de bons résultats.

5.2.6.6. Concordances

Les requêtes concernant le vocabulaire et les structures devront pouvoir être utilisées avec des concordanceurs[48]. En effet, si les champs que nous avons appelés concrets peuvent être utilisés pour réduire le nombre de textes dans lesquels on effectuera les recherches et si le thème n'aura pas de sens (puisqu'il s'adresse au texte comme entité), l'utilisation des requêtes, définies précédemment pour le vocabulaire et la grammaire, sera particulièrement intéressante pour la création d'exercices [JOH] ou même pour présenter certaines notions. Les concordances sont la base d'une approche de l'enseignement des langues dirigée par et vers les données (data-driven learning). Par données, on entend ici exemples d'énoncés dans la langue enseignée [BLA 97]. On partira de ces exemples (bien formés ou mal formés, le contraste entre les deux peut être utilisé), pour permettre aux apprenants d'assimiler les notions.

[47] cours concernant les réseaux neuroniques disponible à l'adresse : http://www.cavalex.com/pdf/livre_touzet.pdf

[48] Quelsques succintes explications sur les concordanceurs sont disponibles à : http://pot-pourri.fltr.ucl.ac.be/one/corpora.htm

Cette donnée sera donc à prendre en compte, lors du choix des différents outils, pour l'annotation des textes et leur recherche dans la base. Il faudra s'assurer, soit qu'ils peuvent servir de concordanceurs, soit qu'ils sont compatibles avec des concordanceurs.

5.2.6.7. Conclusions sur la recherche de documents

La recherche de documents présente deux aspects : la manière de présenter les requêtes ou l'interface homme-machine du système et la réalisation du système lui-même qui effectuera les recherches et retrouvera les textes. Le système est beaucoup trop dépendant des outils qui seront utilisés pour être défini très clairement ici. En revanche, il est possible d'avoir une bonne idée de la façon dont les requêtes pourront être formulées et de voir comment elles pourront être liées avec le système.

6. CONCLUSION

6.1. REMARQUES GÉNÉRALES

La réalisation de ce travail s'est avérée difficile du fait du peu de références bibliographiques en rapport avec le sujet.

Nous avons, tout d'abord, été submergé d'informations tant la littérature fourmille de références à des bases de textes, à l'indexation, à la recherche de données dans une base de texte ou à l'utilisation des concordances. Mais après nous être acharné dans cette direction, il nous a fallu nous rendre à l'évidence : parmi tous ces travaux de recherche, la plupart n'avait aucun lien avec notre sujet et pour ceux qui en avaient un, il ne constituait en général pas une avancée décisive (l'une des informations les plus concrètes que nous ayons eu était un e-mail sur la mailing-list de TEI ; l'auteur du message nous expliquait qu'il allait se pencher sur un problème proche du notre d'ici quelques mois…). Nous avons donc du restreindre l'envergure du sujet et ne pas présumer de la portée du mémoire, qui est un travail de défrichage et doit être perçu comme tel.

La plus grande partie des informations que nous avons recueillies nous est donc venue des entretiens que nous avons effectués après une phase de préparation concernant les pratiques pédagogiques utilisant un support textuel.

6.2. LE FONCTIONNEMENT DE LA BASE DE TEXTE INDEXÉE EN FONCTION DE CRITÈRES PEDAGOGIQUES

De ces entretiens et de l'exploitation des résultats, il nous est apparu qu'il était possible d'établir un certain nombre de critères objectifs de choix d'un texte en fonction de l'usage que les enseignants allaient en faire. Ces critères objectifs s'opposent à des tentatives de typologies, qui auraient rendu ambigu l'ajout de documents dans le texte et auraient probablement limité les recherches : si on prend l'exemple de la typologie en fonction des classes d'activités, un texte pourra servir dans n'importe quelle classe, mais pas pour n'importe quel public.

Il est possible d'effectuer une recherche en fonctions de critères objectifs de vocabulaire et de grammaire, que rien ne nous empêche d'englober sous une couche logicielle intermédiaire qui pourrait faire le lien entre une requête utilisant une typologie et les critères objectifs codés dans la base de textes.

6.2.1. Ajout d'un document dans la base

Nous pouvons donc représenter l'entrée d'un document dans la base de la manière suivante.

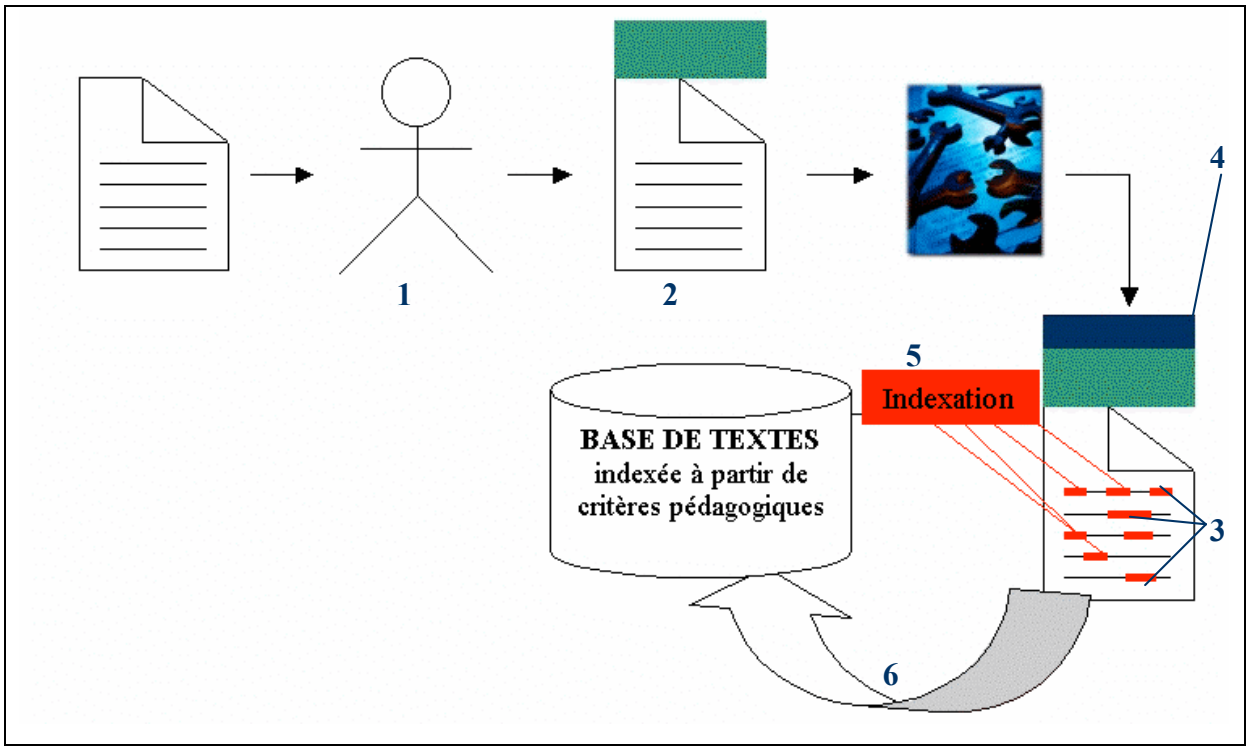


Figure 16 Ajout d'un texte dans la base.

Ce diagramme suit pas à pas les étapes pour un texte donné. L'utilisateur commence par renseigner les champs que nous avons laissés à sa charge **(1)** :

- pour la source : type, titre, date, lieu de publication, auxquels nous pouvons rajouter l'éditeur pour se conformer à la spécification TEI
- pour l'auteur : le nom, le prénom, les dates de naissance et de décès ainsi que la ou les nationalité(s)
- pour le texte : le titre, la date, l'intégrité, l'authenticité et les composantes que nous avons extraites pour le type, à savoir : fiction, dialogue, si le texte est littéraire, prose et son type, son type s'il n'est pas littéraire dans l'éventualité où elles seraient suffisantes. Comme nous l'avons dit une nouvelle série d'entretiens serait utile ici.
- pour les modifications : aucun champ puisqu'ils ne seront renseignés que lors d'une modification.

Une fois tous ces champs renseignés, ils sont ajoutés au texte sous forme d'en-tête (2). Les outils annotent ensuite le texte (pour pouvoir effectuer ensuite les recherches suivant le thème, le vocabulaire, la grammaire, les champs lexicaux et choisir les documents en fonction du style) (3). En même temps, les données provenant de l'annotation et d'autres outils servent à compléter l'en-tête avec la langue et les mesures en nombre de mots et en nombre de phrases (4). Le texte est alors indexé (5) et ainsi entré dans la base (6). Au cours de cet ajout, les champs dépendant de la position du texte dans la base, sont renseignés (identificateurs).

6.2.2. Recherche d'un document dans la base

Une fois un nombre suffisant de documents entrés dans la base, il est possible de les rechercher suivant le processus suivant :

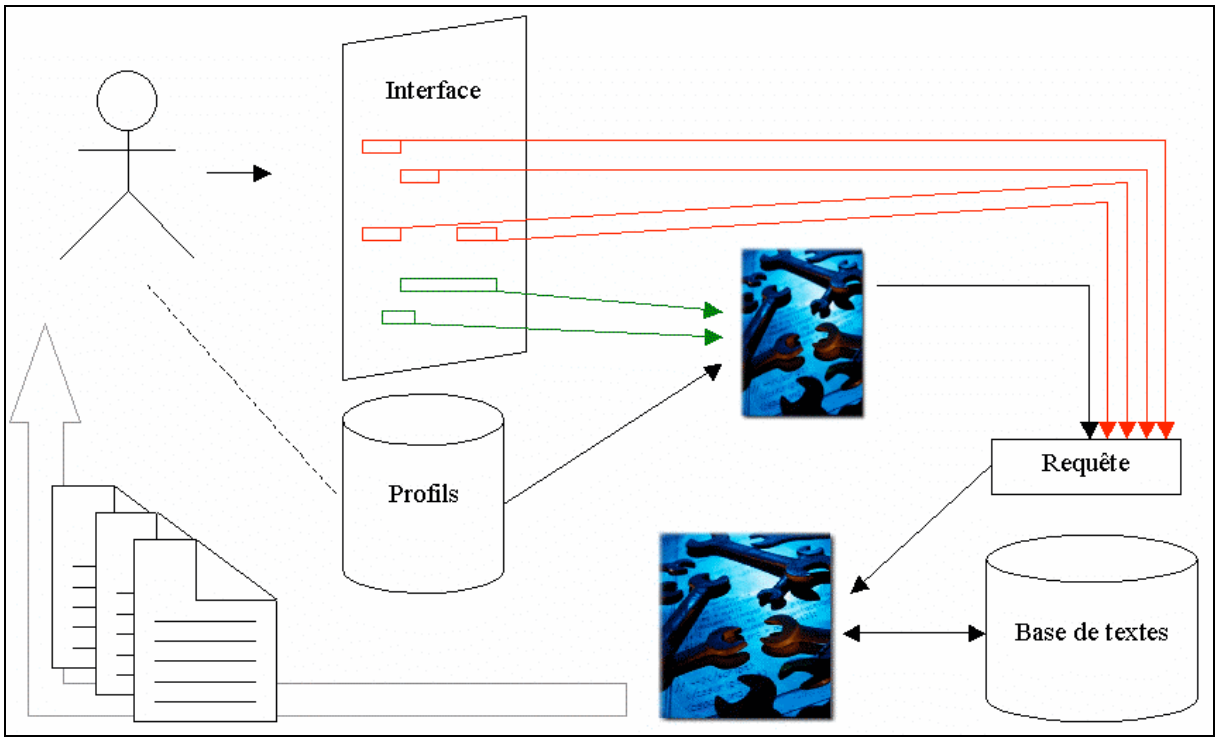


Figure 17 Recherche d'un texte dans la base

L'utilisateur renseigne les champs qui l'intéressent pour faire sa requête. Sur la figure les champs rouges sont ceux qui seront interprétables tels quels (champs concrets), les champs verts sont ceux qui nécessitent une aide à la formulation de la part d'outils. Parmi ces champs, on trouvera le thème, ce que l'on a appelé vocabulaire, grammaire et champ lexical.

En fonction des données fournies et d'autres informations comme le profil que l'utilisateur aura préalablement rempli, les outils formuleront automatiquement la dernière partie de la requête.

A partir de cette requête, d'autres outils effectueront la recherche dans la base de texte, la mettront en page (surlignage de certains éléments en fonction de la requête elle-même et du style) et renverront tous les textes candidats à l'utilisateur.

6.3. POURSUITE DU TRAVAIL

6.3.1. Jusqu'à la réalisation d'un prototype

Ce travail constitue une première étape dans le processus de réalisation d'un premier prototype. Mais avant de se lancer dans cette phase, il faudra effectuer tout le travail sur les outils, à savoir effectuer un état de l'art et entrer dans la phase de conception en fonction des données recueillies, des outils qui seront intégrés à la plate-forme et de ceux qui pourront l'être.

Le travail sur les outils nécessitera une recherche supplémentaire sur les activités pédagogiques, mais cette fois-ci, dans le cadre de l'*e-learning* et non pas uniquement dans le cadre de la classe de langue. Une fois toutes les informations recueillies on pourra commencer à implémenter le premier prototype.

6.3.2. Vers la réalisation de la base…

6.3.2.1. …de textes

Ce premier prototype pourra ensuite servir de base à une étude menée conjointement avec des utilisateurs potentiels. Cette étude aura pour but d'aller beaucoup plus en profondeur que l'on a pu le faire ici dans les aspects pédagogiques (quelles données manque-t-il aux enseignants pour pouvoir effectuer des recherches comme ils l'entendent ?) et aussi dans les aspects interface homme-machine (que faut-il changer dans les processus d'ajout et surtout de recherche de documents ?).

Une fois cette étude accomplie, il sera possible d'exploiter d'un côté le travail effectué ici et de l'autre la réponse des enseignants, pour entrer dans la réalisation de la véritable base de textes indexée pédagogiquement pour l'enseignement des langues.

6.3.2.2. …de supports pédagogiques

Dans le cadre d'un projet comme MIRTO, une base ne contenant que des ressources textuelles ne sera pas suffisante pour l'exploitation de toutes les possibilités du support informatique.

L'ajout de textes oraux est d'ores et déjà prévue.

Mais même au niveau du texte, des extensions possibles. Nous ne considérons ici que du texte brut. La définition du texte pourrait être étendue à des formats prenant en compte la mise en page, voire utilisant les liens hypertextes. L'hypertexte permettrait en outre de prendre en compte cette notion de soutien (*support* en Anglais) de texte défendue par Alice Henderson.

Enfin, même si ce n'est pas d'actualité pour le moment, l'image et la vidéo jouent un rôle de plus en plus important dans l'enseignement des langues "en présentiel", il faudra donc peut être un jour intégrer ces nouveaux médias à la plate-forme MIRTO.

7. BIBLIOGRAPHIE

7.1. SITES

Tous les liens suivants ont été vérifiés le 15 Juin 2003.

- [A-P 02] - Georges ANTONIADIS, Claude PONTON, Le TAL : une nouvelle voie pour l'apprentissage des langues, colloque untele[49], 28 au 30 Mars 2002, l'Université de Technologie de Compiègne, France, présentation PowerPoint :

  http://www.utc.fr/~untele/abst_2002/antoniad.html

- [ANT 95] - Georges ANTONIADIS, Enseignement des langues et informatique : De l'apprentissage de l'autre pour la pérennité du couple, Conférence invitée, 2e Congrès national des professeurs de français, 28 septembre au 1er octobre 1995, Salonique, Grèce :

  http://www.u-grenoble3.fr/stendhal/stendhal/dip/publis/salonique.ps

- [B-P 99] - Michael BLAHA, William PREMERLANI, Using UML to design database applications, 1999 :

  http://www.umlchina.com/Indepth/usinguml.htm

- [BLA 97] - Troy BLAPPERT, Data-driven Learning : Theory and classroom implementation :

  http://www.well.com/user/greg/KOTESOL/1997-proceedings/blappert_troy.pdf

- [CDDB] - Gracenote CDDB, description :

  http://www.gracenote.com/gn_products/onesheets/Gracenote_CDDB.pdf

  http://www.gracenote.com/corporate/press/article.html/date=1999042700

- [DON 02] - Didier DONSEZ, Le modèle entité association et les bases de données relationnelles (cours), 1998-2002 :

  http://www-adele.imag.fr/~donsez/cours/eaumlrel.pdf

- [FAV] - Jean-Marie FAVRE, UML diagramme de classes (cours) :

  http://www-adele.imag.fr/~jmfavre/ENSEIGNEMENT/TRANSPARENTS/DiagrammesDeClasses/9/DiagrammesDeClasses-9.pdf

- [FranText] - Présentation de la base Frantext :

[49] Usages des Nouvelles Technologies dans l'Enseignement des Langues Etrangères

http://www.inalf.cnrs.fr/_ns/produits/frantext.htm

- [FRIDA] - Pages officielle FRIDA / FreeText :

  http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-projects/index.htm

  http://www.aramis-research.ch/e/7018.html
- [JOH] - TIM JOHNS : Data-Driven Learning Library :

  http://web.bham.ac.uk/johnstf/ddl_lib.htm
- [MIR #1] - Projet MIRTO, logiciels existants :

  http://www.u-grenoble3.fr/stendhal/stendhal/dip/mirto/Logiciels_dispo.html
- [MIR #2] - Demande d'association de projet opérationnel au dispositif GreCO :

  http://www.u-grenoble3.fr/stendhal/stendhal/dip/mirto/MIRTO_GreCO.htm
- [OMG 03] - OMG [50], Mars 2003, UML v1.5, UML 1.5 chapter 3 - UML Notation Guide :

  http://www.omg.org/cgi-bin/apps/doc?formal/03-03-10.pdf
- [SILC] - Projet SILC : http://www-rali.iro.umontreal.ca/ProjetSILC.fr.html
- [TEI 02] - Site officiel de la TEI :

  http://www.tei-c.org/

  Version HTML de la TEI P4 : http://www.tei-c.org/Guidelines2/p4html.tar.gz

  Projets utilisant la TEI : http://www.tei-c.org/Applications/index.html
- [TOEIC] - Site officiel TOEIC :

  http://www.toeic-europe.com/pages/fr/le_test_pres.htm

  Table des niveaux :

  http://www.toeic-europe.com/pdf/TOEIC_Can_Do_Levels.pdf

7.2. LIVRES

- [ADA 85] - Jean-Michel ADAM, Quels types de textes, dans Le français dans le monde, n° 192, 1985
- [ALB 91] - Marie-Claude ALBERT, De l'utilisation des textes en français langue étrangère : apport de la réflexion sur les typologies textuelles, dans Les cahiers du CRESLEF, n° 32, 1991-2

[50] Object Management Group

- [ALO 94] - Encina ALONSO, ¿ Cómo ser profesor/a y querer seguir siéndolo ?, (collection : Investigacíon didáctica), Ediciones Eurolatinas S.A., Madrid, 1994 – ISBN : 84-7711071-9
- [A-V 01] - Pascal ANDRÉ, Alain VAILLY, Conception des systèmes d'information, Ellipses, Paris, 2001 – ISBN : 2-7298-0479-X
- [BOU 98] - Pierrette BOUILLON, Traitement automatique des langues naturelles, AUPELF-UREF – Editions Duculot (De Boeck & Larcier s.a.), Paris, 1998 – ISBN : 2-8011-1181-3
- [C-G 02] - Jean-Pierre CUQ, Isabelle GRUCA, cours de didactique du français langue étrangère et seconde, Presses Universitaires de Grenoble, 2002 – ISBN : 2-7061-1082-1
- [C-M-M-R 00] - Francisca CASTRO VIULEZ, Fernando MARIN ARRESE, Reyes MORALES GÁLVEZ, Soledad ROSA MUÑOZ, Ven 1, Libro del Alumno, Edelsa Grupo Didascalia, S.A., Madrid, 2000 – ISBN : 84-7711-045-X
- [D-G 90] – M. DEL MAR MARTIN VIAÑO, P. GÓMEZ-CASAÑ, La expresión escrita : de la frase al texto tire de Didáctica de las segundas lenguas (Estrategias y recursos básicos), Ed. Santillana S.A., Madrid, 1990 – ISBN : 84-294-3152-7
- [FLU 00] - Christian FLUHR, Indexation et recherche d'information textuelle, tiré de Ingénierie des langues, Hermes Sciences Europe, Paris, 2000 – ISBN : 2-7462-0113-5, pp 235-253
- [GAR 90] – T. GARCÍA HERNANDEZ, La comprensión lectora : la lectura como actividad didactica tiré de Didáctica de las segundas lenguas (Estrategias y recursos básicos), Ed. Santillana S.A., Madrid, 1990 – ISBN : 84-294-3152-7
- [I-S 95] - Nancy M. IDE, C.M. SPERBERG-McQUEEN, The TEI: History, Goals and Future tiré de TEI background and context, Ed. Kluwer Academic Publishers, Dordrecht, Pays-Bas, 1998 (texte de 1995) – ISBN : 0-7923-3689-5, pp 5-15
- [KEV 96] - Martine KERVRAN, L'apprentissage actif de l'anglais à l'école, Ed. Armand Colin / Masson, Paris, 1996 – ISBN : 2-200-01399-X

- [LÓP 90] – J. LÓPEZ HERNANDEZ, Formas de trabajo en el aula : individualización y socialización tiré de Didáctica de las segundas lenguas (Estrategias y recursos básicos), Ed. Santillana S.A., Madrid, 1990 – ISBN : 84-294-3152-7
- [MAR 99] - Catherine MARCUS, Français seconde langue, 36 lectures pour les collèges, CRDP de l'académie de Grenoble et Delagrave, 1999 – ISBN : 2-206-08161-X
- [N-M 01] - Eric J. NAIBURG, Robert A. MAKSIMCHUK, UML for Database Design, Addison-Wesley, 2001 – ISBN : 0-201-72163-5, Traduit en français par Alain DUCROT, Bases de données avec UML, CampusPress, Paris, 2002 – ISBN : 2-7440-1256-4
- [S-B 95] - C.M. SPERBERG-McQUEEN, Lou BURNARD, The design of the TEI Encoding Scheme tiré de TEI background and context, Ed. Kluwer Academic Publishers, Dordrecht, Pays-Bas, 1998 (texte de 1995) – ISBN : 0-7923-3689-5, pp17-39
- [TAG 94] - Christine TAGLIANTE, La classe de langue, (collection : Techniques de classe), Ed. CLE International, Paris, 1994 – ISBN : 2-09-033112-7

Annexe 1 : Entretiens, fiche 1

Généralités	Nom / Prénom						Date
	Langue						
	Niveau						
	NB Années d'enseignement						
	Etablissement						

Description du niveau	Les Axes d'évaluation				
	¿Compétences?				
	Granularité				

Annexe 2 : Entretiens, fiche activité

But de l'activité	
Public (âge / niveau)	
Fonctionnement de l'activité	
Commentaires personnels	*(Mes commentaires)* */*si certains traitements sont nécessaires par rapport au texte, et s'ils sont automatisables*/*
Caractéristiques du texte	/*critères de choix*/
Processus de recherche	

Annexe 3 : Can-do Levels Table

CAN-DO LEVELS TABLE

The descriptions below are intended to serve as guidelines to understanding the competence reflected by the corresponding scores and apply in most cases. The levels (0 to 3+) have been developed from the scale used by the *Foreign Service Institute* and the *Inter-Agency Language Roundtable*.

PART 1: LISTENING SECTION + PART 2: READING SECTION

Listening Score	Listening	Speaking	Reading Score	Reading	Writing	LEVELS / TOTAL SCORES
455-495	Can : • understand mother-tongue speakers of English in meetings • function in all of the situations described below whether professional or social, concerning concrete or abstract subjects	Can : • conduct meetings with mother-tongue speakers of English • perform all of the below with a greater degree of ease...	455-495	Can : • read adequately for most professional needs • read highly technical manuals in own area • read all of the below...	Can : • write effectively, both formally and informally; however, work for publication will still require review • produce the documents described below without undue difficulty	3/3+ General Professional Proficiency (>960 Advanced) 905 - 990
395-450	• understand most work related situations • understand most speakers of English in international meetings • function in all of the situations described below but with a greater degree of facility and accuracy	• satisfy most work requirements • conduct a job interview in own area of expertise • sustain fluency, accuracy and appropriate register in known situations	395-450	• read most types of documents with varying degrees of ease • read even highly-technical subjects with little use of dictionary • experience difficulties with sophisticated menus, novels...	• write an employment application • write a letter of complaint • write the documents below with increasing degrees of accuracy and ease	2+ Advanced Working Proficiency 785 - 900
305-390	...understand: • explanations of work problems • requests for products on phone • discussions of current events by mother-tongue speakers of English • headline news on radio	• adapt language use for different audiences in most cases • make short (30 minute) formal presentations if prepared • discuss topics of general interest using non-elaborate structures	305-390	• read with only the occasional use of a dictionary: - technical manuals - many news articles - popular novels • identify inconsistencies in points of view	• write with some effort: - letters to potential clients - 5 page formal reports - summaries of meetings - job application letters	2 Basic Working Proficiency 605 - 780
205-300	...understand:- • explanations related to routine work tasks in one to one situations • some travel announcements • limited social conversations	• describe own job responsibilities and academic background • discuss past and future projects • make travel arrangements over the phone	205-300	• understand basic technical manuals for beginners • use a dictionary to understand more highly technical documents • read agenda for a meeting	• write with some difficulty: - short memos - letters of complaint - descriptions of processes • fill out simple application forms	1+ Intermediate 405 - 600
130-200	• understand simple exchanges in everyday professional or personal life with a person used to speaking with non mother-tongue speakers • take simple phone messages	• produce simple if hesitant language adequate for elementary functions with patient listeners: introductions, directions, requesting information, ordering food...	130-200	• use a directory • understand simple instructions • read simple, standardized business correspondence	• write short notes, directions and lists with difficulty • not fill out forms, write detailed memos, letters or reports	1 Elementary 255 - 400
05-125	• understand adequately for immediate survival needs, directions, prices... • comprehend simple questions in social situations	• name objects, colors, clothes, people, days, months, dates, & give the time • only reproduce formulaic language - telegraphic style	05-125	• understand odd words e.g. shop names • read simple memos and menus, train or bus schedules, traffic signs...	• write odd words, formulaic language • not write creative sentences	0/0+ Novice 10 - 250

Don't forget to enclose a copy of this chart when communicating your score or your CV.

The Chauncey Group Europe S.A., 66, av. des Champs Elysées (Imm. E), F-75008 Paris. Email: info@toeic-europe.com

For all information on the TOEIC Test, Europe- and worldwide,
visit **www.toeic-europe.com** & **www.toeic.com**

09/2000