Rapport de Synthèse

ECOLE CNRS

« Acquisition du langage et corpus en linguistique et psychologie :

pour une mutualisation des ressources (outils et données) »

http://acquisition.risc.cnrs.fr/

Ce document a été rédigé par K. Duvignau et A. Lacheret à partir des informations recueillies avant l'école (section 1), durant l'école - voir notamment table ronde du vendredi matin (section 2) ainsi que post école sur la base des éléments recueillis via les questionnaires d'évaluation (section 3). Nous terminons par les perspectives envisagées par l'ensemble des acteurs à l'école (section 4).

1. Objectifs et participants

1.1. Rappel des objectifs de l'école

L école organisée à Moissac en mai 2006 se situait à l'intersection de deux thématiques centrales au sein des recherches actuelles sur le langage : la thématique « corpus » (mutualisation des savoirs et des ressources : outils et données) et celle de l'acquisition. Pour mémoire, l'école s'articulait autour de trois objectifs principaux :

1- état des lieux le plus exhaustif possible dans le domaine de la francophonie sur les corpus (oral et écrit) des enfants (0-11 ans), sur les outils de recueil et de traitement disponibles,

2- formations sous forme d'ateliers pour l'utilisation d'outils de recueil et de traitement des données,

3- mutualisation des savoirs, données et outils sur l'acquisition du langage avec prise en compte des principaux niveaux d'analyse de la langue : Phonétique- phonologie - morphologie - syntaxe – lexique & sémantique – pragmatique, d'où la mise en place de groupes de travail pour ces niveaux.

1.2 Les acteurs de l'école

Cette école a réunit 15 intervenants et 40 participants soit 65 personnes.

Concernant les disciplines représentées au sein de l'école, si l'équilibre était présent au niveau des Intervenants qui se répartissaient de façon homogène en linguistique et en psychologie :

Linguistique : 45 %	**Psychologie : 55%**

on a pu noter, en ce qui concerne les participants, une homogénéité moindre entre la communauté des linguistes et celle des psychologues:

Linguistique : 60 %	**Psychologie : 40 %**

Répartition en zone géographique

Les participants à l'école couvrent une grande zone géographique puisqu'ils sont issus de 22 sites-structures différents. Les retombées de l'école concerne donc des organismes de recherche divers et variés.

Sites CNRS : 7	**Sites Université : 12**	**INSERM : 1**	**Etranger : 2**
Paris 8-UMR 7023	Nice – EA 1189	Paris U676	UK
Lyon 2 –UMR 5596	Angers		Belgique
Paris 10 –UMR 7114	Toulouse – JE 2395		

Paris 3 – UMR 7018	Nantes		
Grenoble- UMR 5105	Savoie		
Rouen - FRE 2787	Grenoble 2		
Poitiers – UMR 6215	Clermont-Ferrand		
	Lyon1 – EA 2281		
	Paris 7 – EA 333		
	Lille 3- EA 1764		
	Orléans		
	Besançon		

2. Table ronde

(slides des 4 groupes de travail accessibles sur le site de l'école & discussion du vendredi matin)

Après avoir rappelé la problématique générale visée par l'école, l'objet de cette partie est d'en préciser les pourtours, les résultats effectifs par rapport aux résultats attendus, ce qu'il ressort au bout du compte de la notion centrale de mutualisation (données, outils de traitement mais aussi méthodologies d'étude) et de rappeler les pistes qui ont été suggérées pour la mise en pratique effective de cette mutualisation.

2.1. De quoi parle-t-on ?

La thématique de l'école ainsi spécifiée et circonscrite à un domaine précis n'en reste pas moins très large tant au niveau des types de recherche impliqués que des disciplines concernées. La problématique de l'acquisition couvre tous les modules de traitement du langage, des niveaux substantiels aux niveaux symboliques et représentationnels, elle concerne les questions de production, de perception et de stockage des formes; ces questions peuvent se poser tant dans une perspective plurilingue que monolingue ; enfin elle s'inscrit au cœur de la pluridisciplinarité, faisant dialoguer non seulement linguistes et psychologues, spécialistes du développement ou de l'apprentissage, mais également informaticiens et modélisateurs,. La variété des angles d'attaque est donc bien réelle et les problèmes conceptuels et terminologiques associés ne doivent pas être sous-estimés. Il était donc dès le départ nécessaire de faire des choix étant bien entendu impossible de couvrir ce champ foisonnant en une semaine de formation. En pratique, qu'est-ce qui a guidé notre stratégie pour la structuration de l'école ? Comme toujours dans ces cas là : critère de représentativité mais aussi de créativité scientifique.

Concernant le premier critère, il s'agissait de se situer dans une perspective épistémologique et d'assurer des interventions de pionniers (ières) dans le domaine de l'utilisation de corpus pour l'acquisition du langage. Cette perspective historique justifie en particulier l'ouverture de l'école et les zooms sur le champ de l'interaction ; d'une manière générale elle explique l'orientation de la formation sur les aspects symboliques de l'acquisition, syntaxe et sémantique, au détriment du reste. Et de fait, elle bat en brèche la question posée par quelques participants sur l'absence dans cette école des représentants de la linguistique formelle qui ne s'est attelée que très récemment et encore souvent avec beaucoup de réserves aux données in vivo et à l'usage[1]. Bref, en aucun cas, il ne s'agissait « d'éviter les conflits », d'autant que la controverse scientifique est sans aucun doute garante d'une recherche seine et sans cesse renouvelée. Il s'agit là simplement de la conséquence logique des orientations scientifiques de deux courants à l'origine bien distincts dans la conception qu'ils ont de la langue elle-même et par voie de conséquence dans leur manière d'aborder les processus langagiers. Pour mémoire, le langage est défini dans l'approche générative comme un système de règles formelles fixes, universelles et innées ; le linguiste travaille en production indépendamment du contexte et de l'usage ; si au fil des années, il est apparu indispensable d'ancrer la question de la plausibilité cognitive des modèles développés au cœur de l'usage, cette préoccupation reste néanmoins récente. Dans l'approche fonctionnaliste en revanche, la question du contexte, et donc des corpus, est d'amblée cruciale puisque les structures du langage sont déterminées par les fonctions symboliques qu'il sert : les fonctions conceptuelle et communicationnelle par laquelle les sujets interagissent.

[1] Pendant longtemps linguistique formelle et usage, donc corpus, resteront antithétiques.

Pour ce qui est du critère de compétence/créativité, il s'agissait, en restant dans la problématique générale de l'école, de faire connaître et manipuler à la communauté des outils soit bien stabilisés maintenant et de fait incontournables (ex. Childes), soit nouveaux, originaux, pour certains encore en développement mais très prometteurs pour les recherches futures et leurs orientations nouvelles (Phon, Prox).

Même en restreignant le champ de la sorte, la communauté constituée à Moissac restait hétérogène. Si cela a été un point fort semble-t-il (le dépouillement des questionnaires montre que la diversité des participants a été globalement très appréciée), il reste néanmoins nécessaire de s'assurer que l'on parle tous bien de la même chose et, quand ce n'est pas le cas, de saisir précisément ce qui fait les spécificités des uns et des autres (transferts de connaissance) en envisageant des passerelles terminologiques le cas échéant (ex. définition linguistique vs. psychologique du concept de *grammaticalisation*). Sur ce point, l'école aura permis de faire émerger le consensus suivant concernant la notion même de « corpus » qui, aussi surprenant que cela puisse paraître, ne va pas nécessairement de soi. On retiendra deux points centraux (voir en particulier les précisions apportées par Garcia-Debanc lors de son intervention du jeudi 18 et par M. Guidetti à la table ronde du vendredi 19 : *a) un corpus correspond à une collection de données langagières sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage » (Sinclair, 1996*), *b) Le terme « corpus » peut désigner deux stades documentaires sensiblement différents : d'une part, un ensemble de sources, écrites, orales ou audiovisuelles brutes et, d'autre part, l'information élaborée à partir de ces sources*. Elle aura aussi mis en lumière qu'au-delà des ressources elles-mêmes, une mutualisation des outils méthodologiques pour les constituer et/ou les traiter est également souhaitable[2].

2.2. Une telle école : pourquoi, pour qui, dans quel but ?

Cette école s'inscrit dans le prolongement de celle qui s'est tenue à Caen en juin 2004 consacrée aux corpus, à l'archivage et au traitement des données langagières écrites et orales. L'école caennaise avait en effet été conçue comme la première étape d'un cycle de formations où il apparaissait fondamentalement nécessaire à un moment où un autre du cycle de prendre en compte la thématique de l'acquisition, tant le domaine est riche et convoque à lui seul un ensemble d'hypothèses théoriques à valider sur corpus, et de questions pratiques associées. Pensés en termes de **mutualisation** et d'échange, les deux objectifs majeurs de l'école étaient : 1° l'information, 2° la formation. Concernant le point 1, le chercheur doit savoir quels sont les corpus et/ou les outils, et où il pourra les trouver, potentiellement utilisables pour sa recherche et celle de ses collègues ; il a de ce point de vue un rôle de renseignement et de veille, crucial pour la dynamique de son laboratoire et de l'équipe qu'il encadre le cas échéant. Ceci afin de ne pas réinventer l'eau chaude à chaque fois, par exemple la collecte de nouvelles données pour une recherche dédiée alors que d'autres, du même type et a priori utilisables pour la dite recherche, traînent probablement dans plus d'un laboratoire. Sur ce sujet, un objectif majeur des groupes de travail était d'initier un premier référencement des données potentiellement disponibles pour différents types de recherche; ce n'est qu'un premier pas et il apparaît nécessaire de le systématiser à l'avenir (cf. *infra*, § 3). Cet objectif vaut aussi pour les outils de traitement (annotation, exploitation, quantification) et les méthodologies valides reconnues comme telles par les pairs pour l'exploration d'une thématique donnée (voir les « valeurs sûres » présentées par Heather Hilton sur les versants des outils de recueil (slides 5 de sa présentation du vendredi). Concernant le volet formation, un telle école devait mettre le pied à l'étrier aux chercheurs afin qu'ils soient rapidement opérationnels et autonomes dans la manipulation d'outils spécifiques, et aujourd'hui incontournables. Là encore, il reste du chemin à parcourir et des pistes concrètes de formations complémentaires ont été proposées (cf. *infra*, § 3.1). Information et formation sont ainsi les préalables pour une recherche cumulative qui va de l'avant. Connaître pour mieux transmettre, connaître pour mieux avancer et ouvrir de nouveaux chantiers sur des bases solides et contrôlées. Elles ouvrent également la voie pour une recherche plurielle dans un domaine qui la convoque particulièrement par

[2] protocoles de collectes de données; outils d'évaluation complémentaires aux données elles-mêmes (échelle d'acquisition, rapports parentaux); protocoles standards d'annotation et de transcription pour les données orales (ex. annotation de la prosodie, pauses et ratures orales, identification des fillers) et gestuelles, nouvelles données en neuro-imagerie qui pourraient aider la recherche sur le développement, etc.

sa complexité et sa richesse (nécessité de travaux complémentaires élaborés en partenariat, organisations en réseaux). En aval, la **valorisation** du travail de recherche s'imposera d'elle-même : un corpus qui continuera à vivre et à circuler en dehors du laboratoire où il aura été conçu, qui sera cité dans plus d'une publication scientifique, obtiendra sans problème ses lettres de noblesse et sa propre légitimité scientifique. De même, un outil de traitement particulier, d'abord utilisé par une poignée de chercheurs, deviendra vite incontournable, si l'information circule bien et si sa maintenance est correctement assurée, quitte à ce que d'autres équipes rejoignent le projet, voire prennent le relais pour assurer la pérennité de l'outil (mise à jour, tutorial en ligne, liste de discussion, etc, voir pour exemple le phénomène « PRAAT »). De ce point de vue, un bon logiciel n'est pas nécessairement le plus performant technologiquement, il faut d'abord qu'il soit bien documenté et le plus ouvert au partage ; nous touchons là le cœur de la mutualisation. Mais l'objectif de valorisation ne s'arrête pas là : à l'instar des sciences de l'ingénieur, à partir du moment où l'on adopte une approche expérimentale, fondamentalement basée sur l'observation et l'induction, il devient nécessaire pour être crédible de donner tous les moyens à la communauté de falsifier la théorie proposée et les hypothèses défendues, ce qui veut dire très concrètement la transparence méthodologique, l'accès aux ressources utilisées, la comparabilité des données et des résultats. De ce point de vue, l'école avait comme objectif d'impulser une dynamique de mutualisation dans la recherche et d'illustrer pourquoi une conception individualiste du travail du chercheur n'est plus une position tenable aujourd'hui, en quelques sortes un combat d'arrière garde, bref de convaincre les plus frileux au partage des ressources et à une pratique collective de la recherche.

2.3. Problématique de la mutualisation :

L'objectif de l'école étant posé, la mutualisation a ses limites et il serait naïf de les ignorer. La question est finalement la suivante : quelle est la juste mesure à adopter entre deux attitudes opposées : a) « tout est partageable, tout est distribuable ici et maintenant, tout est recyclable » vs. b) « étant donné les centres d'intérêts, les cadres théoriques et les objectifs fixés extrêmement variés des uns et des autres, la quête d'une mutualisation semble illusoire, notamment en termes de données, les situations d'étude étant très hétérogènes (interactions libres, narrations, situations naturelles, expérimentales ou aménagées) ; de fait, il n'existe pas et il ne peut pas exister de corpus omnibus, la notion même de **corpus de référence** devient caduque ? Le caractère optimiste de la position a) est louable et séduisant mais posé comme tel, sans aucun garde fou, risque de conduire à une mutualisation non contrôlée autrement dit à plus de mutualisation du tout. Quelle stratégie adopter donc pour qu'une réserve de corpus et d'outils ne se réduise pas au bout du compte à un vaste four-tout inexploitable, mais que les ressources ainsi référencées et/ou mises à disposition soient effectivement réutilisables par d'autres ?

2.3.1. Mutualisation des données

Concernant les données langagières, on eut envisager plusieurs niveaux d'accessibilité : a) une donation immédiate sans attendre d'être sollicité, b) une simple vitrine, *i.e.*L'intérêt de la vitrine est son côté rassurant pour le chercheur qui peut garder le contrôle sur les données en bénéficiant du feed back des utilisateurs. Mais même dans le cas a), un contrôle minimum s'impose concernant en particulier la dimension éthique, juridique et sociolinguistique. Là encore la documentation doit être exhaustive, d'où l'importance des méta-données facilement lisibles, thématique peu abordée à l'école. La donation sans réserve est, elle aussi, à géométrie variable : s'agit-il de mettre en ligne uniquement les données brutes, uniquement la transcription ou les deux ? Sur le plan **éthique**, Peut-on tout mettre en ligne (problème d'anonymisation des visages pour les études kinésiques, des noms propres, etc, coupure des données sensibles en particulier pour les contextes pathologiques, auquel cas comment couper sans perdre trop d'informations ?) ? Quel que soit le choix retenu, la mise en place d'une charte d'utilisation des données s'avère nécessaire (voir intervention Ch. Parisse vendredi 19). Sur le plan **technique**, quelles sont les contraintes imposées au chercheur pour rentrer ses données dans la base ? La question de **format** est ici essentielle : des contraintes trop fortes risquent d'en dégoûter plus d'un, aucune contrainte risque de rendre les données inexploitables, la souplesse des transcriptions devant donc être compensée par une description exhaustive des formats utilisés.

2.3.2. Mutualisation des outils de traitement

Bien que n'ayant pas fait le tour de la question, loin de là, l'école a montré la richesse et la diversité des outils sur les versants de l'annotation, de l'exploration et de l'analyse, enfin de la modélisation et de la simulation (simuler pour mieux comprendre, simuler pour falsifier). Si, là encore, un outil est toujours conçu au départ pour une recherche bien spécifique, parfois même étrangère à la thématique de l'acquisition, il est pourtant nécessaire de ne pas ignorer ce paradigme. De ce point de vue là, l'expérience de Lexique 3, au départ développé complètement en dehors du champ, est révélatrice. Il en va de même pour le logiciel Nooj dérivé d'Intex et son utilisation pour l'étude de l'acquisition du français au Canada, le logiciel PRAAT, ainsi que des systèmes plus récents (logiciel PROX présenté à l'école), voire encore à l'état d'expérimentation (logiciel de traitement de la prosodie ANALOR discuté au sein du GT « phonologie »).

Aborder cette question de la mutualisation des outils a, du même coup, permis de mettre en exergue les manques associés à des niveaux de traitement linguistique spécifiques. Si par exemple la question de l'annotation morphosyntaxique des corpus semble résolue aujourd'hui et s'il y a là effectivement des outils à mutualiser, il n'en va pas de même partout. Or technologiquement, les problèmes de traitement automatique, gages de gain de temps pour le chercheur, semblent surmontables aujourd'hui. En conséquence, si les limitations sont bien réelles dans certains domaines, ce n'est sans doute pas au niveau instrumental qu'il faut se poser des questions mail bel et bien sur le plan théorique : pour ces domaines, en effet, un véritable travail reste à faire en vue de stabiliser les unités de description et aboutir à un minimum épistémologique commun, sur lequel peuvent se fonder et se développer les traitements automatiques (voir en particulier les efforts à faire en pragmatique).

2.3.3. Mise en ligne et consultation

En amont des différentes questions présentées en *supra*, se pose la question capitale de la centralisation, du stockage et de la maintenance des ressources, outils et données pour qu'il y aie mutualisation effective (cf. intervention Ch. Parisse vendredi 19). Ce type d'entreprise a évidemment un coût en termes de besoins matériels et humains. Outre, la maintenance des ressources existantes, il faut également être prêt à répondre à une demande nouvelle de la communauté en termes de développement d'outils. Nous avons vu que, pour éviter les risques de dérive en particulier, cette tâche ne pouvait en aucun cas être confiée à un laboratoire ou une équipe seule. Br. Gaume a rappelé la mise en place récente de deux **centres de compétences**, l'un sur les corpus oraux (dir. Ph. Blache, M. Jakobson), l'autre sur les corpus écrits (dir. J.M. Pierrel). Pour l'heure, ces centres de compétences sont encore à un état embryonnaire, nous n'avons aucune visibilité sur la façon dont ils vont fonctionner et sur les services qu'ils pourront rendre concrètement à la communauté, il est donc prématuré d'envisager une action dans ce cadre. K. Duvignau a proposé d'adopter une solution d'attente : continuer à faire vivre et à enrichir le site de l'école qui sera hébergé par le **RISC** pendant au moins deux ans. En parallèle, il est nécessaire **qu'un groupe de réflexion** (pour l'heure bénévole) soit constitué pour réfléchir sur des propositions très concrètes à formuler à l'issue des deux ans. Dans cette optique un module « mutualisation des ressources en acquisition du langage » (Responsables : K. Duvignau, A. Lacheret, A. Morgenstern, Y. Rose) a été mis en place au sein d'un GDR « acquisition du langage » proposé à la création auprès du CNRS en juin 2006 (responsable du GDR : M. Hickmann). Même si, nous n'avons pas trop de visibilité sur l'évolution de l'ILF dans les années à venir, A. Lacheret se propose également d'étudier les possibilités d'hébergement de ce côté là. Le groupe de réflexion aura en parallèle pour rôle de construire la maquette d'une plate-forme web pérenne hébergeant les différents types d'outils. De multiples questions se posent concernant la distribution de l'information, la spécification du cadre éthique, la mise en place de formats communs pour les données et les méta-données : standardisation ou du moins harmonisation des codages des différentes sources, etc. Sur ce point, le terrain n'est pas vierge, il sera donc fort utile de s'inspirer d'expériences passées ou en cours ; voir entre autres les recommandations de codage de la *Text Encoding Initiative* qui permettent en principe de représenter de manière compositionnelle des corpus de données orales hétérogènes dans un format unifié, les recommandations formulées dans le cadre du guide des bonnes pratiques de la DGLFLF (éditions du CNRS 2006), les problématiques récurrentes posées dans les différentes réponses à l'appel d'offre « corpus » de l'ANR en mai 2006, la vitrine sur les corpus oraux BDCOIFA hébergée sur le site du CRISCO.

3. L'école a-t-elle globalement répondu aux attentes des participants ?

Les remarques qui suivent ont été établies à partir du dépouillement des questionnaires d'évaluation (notations et commentaires) de l'ensemble des acteurs de l'école (Intervenants et Participants). Quelques mots sont formulés concernant les aspects matériels : infrastructure générale, hébergement, restauration, puis sont abordées la structuration pédagogique de l'école (cours, ateliers, posters et démo, groupes de travail), la couverture de la problématique et des thématiques envisagées.

3.1.Aspects matériels : [1 (très bien– 2 (assez bien – 3 (pas très bien à 4 (pas bien du tout]

Pour ce qui est du contexte « extra-scolaire », le lieu choisi, l'accueil et le matériel mis à disposition, enfin l'organisation générale de l'école festive ont été globalement très appréciés.

HEBERGEMENT

L'hébergement a été apprécié avec une évaluation globale de 1,38. Les différences de confort d'hébergement (Moulin vs Carmel) ont quelquefois interloqué, nous soulignerons simplement que cette solution a permis d'ouvrir l'école au plus grand nombre.

REPAS

En ce qui concerne les repas ils ont été appréciés avec une note de 1,29. Notons que pour certains ils auraient pu être moins copieux.

AMBIANCE ET ORGANISATION FESTIVE

L'ambiance et l'organisation festive de cette école ont été bien appréciés : respectivement 1,13 et 1,15.

3.2.Modalités pédagogiques [1 (très bien– 2 (assez bien – 3 (pas très bien à 4 (pas bien du tout]

DUREE DE L'ECOLE :

La durée de l'école a été appréciée avec une note moyenne de 1, 5

NOMBRE ET DIVERSITE DES PARTICIPANTS

Le nombre et la variété des participants ont été évalués positivement avec respectivement 1,11 et 1,38.

DIVERSITE DES INTERVENANTS

La diversité des intervenants a été positivement évaluée avec une note moyenne de 1,34.

MATERIEL MIS A DISPOSITION

L'école était équipée de portables wifi pour servir les besoins des ateliers mais aussi pour permettre un accès à Internet durant les sessions de groupes de travail et également pour proposer un accès au courrier électronique durant les temps libres. Cette mise en place a été globalement très appréciée par les acteurs de l'école avec une note moyenne de 1,02.

ANIMATION DES DEBATS

L'animation des débats a été bien évaluée avec une note moyenne de 1,4

COURS, ATELIERS, GROUPES DE TRAVAIL ET POSTERS-DEMOS :

Dans l'ensemble, les quatre modalités pédagogiques mis en place ont été bien évaluées : entre « très bien » (note = 1) et « assez bien » (note = 2) pour l'ensemble des acteurs de l'école (participants et intervenants) avec comme moyenne les notes suivantes sur une échelle de 1 (très bien) – 2 (assez bien) – 3 (pas très bien) à 4 (pas bien du tout) :

COURS : 1,4
ATELIERS : 1,3
GROUPES DE TRAVAIL : 1,8
POSTERS ET DEMOS : 1,6

Néanmoins un bémol ressort en ce qui concerne les horaires des « Groupes de Travail » et tout particulièrement ceux des sessions « Posters et Démos » qui ont été soulignés comme trop tardifs par un nombre non négligeable de participants. Pour l'avenir nous retiendrons donc pour ce type de manifestation l'idée suggérée par certains d'écourter les sessions Cours afin de redistribuer le temps

avec des journées moins denses et de pouvoir maintenir la possibilité d'une session Posters, qui d'une part permet aux étudiants de valoriser leur recherche et qui, d'autre part, pour nombre d'entre eux, rend possible le financement de leur venue par leur laboratoire.

Un autre point qui mérite réflexion pour les suites à envisager est le fait que malgré des thématiques précises associées à chaque groupe de travail, les centres d'intérêt étaient parfois trop hétérogènes et n'ont sans doute pas donné toute la mesure de leurs possibilités. De ce point de vue, il aurait peut être été préférable d'organiser les Groupes de Travail non pas par module de traitement d'autant que certains sont fondamentalement transversaux (ex prosodie) mais par centre d'intérêt réel (ex. acquisition précoce vs. tardive, acquisition L1 vs. L2, développement vs ; apprentissage, etc).

COUVERTURE DE LA PROBLEMATIQUE ET DES THEMATIQUES

La couverture de la problématique et les thèmes abordés ont globalement satisfait les acteurs de l'école avec respectivement une note de 1,38 et de 1,45. Etant donné l'hétérogénéité du domaine abordé, nous 'lavons vu,il était nécessaire de faire des choix évidemment criticables : voir les remarques de certains participants dont nous devons tenir compte pour l'avenir, concernant notamment la surepréssentation du point de vue structuraliste et la sous-représentation du courant génératiiviste durant cette école.

Concernant les modules de traitement : la pragmatique et la sémantique lexicale ont été abondamment évoquées, ce n'est pas le cas de la sémantique grammaticale, grande absente de l'école. Or, s'il est un domaine où le corpus est convoqué à part entière, c'est bien celui-ci ; en outre, des travaux existent en particulier dans le cadre des recherches sur la cohésion discursive (gestion de la temporalité, construction des chaînes de coréférences, etc.). Au niveau substanciel et formel, si la phonétique et la phonologie segmentale ont été également sous-représentées, ce n'est pourtant pas faute de sollicitations. Des contacts avaient été pris mais les interlocuteurs sollicités n'ont pu répondre présents. Concernant les aspects suprasegmentaux et la dimension prosodique, transversale aux différents niveaux de traitement, le silence de l'école en dit long sur la dynamique de la recherche à l'heure actuelle, aussi paradoxal que cela puisse paraître : c'est le premier module de traitement déclenché pour l'activité de langage bien avant la construction du lexique. Il permet notamment, par delà le langage articulé, l'expression des émotions et la communication d'informations, il s'inscrit dès le départ au cœur de l'intersubjectivité, et pourtant la prosodie reste en partie le mouton noir des recherches sur l'acquisition. Est-ce la complexité du domaine lui-même qui explique la frilosité des chercheurs, tant sur le plan conceptuel (primitives prosodiques et unités de description encore flottantes et sujettes à divergences chez les spécialistes du domaine), qu'instrumental (nécessité de manipulations complexes et coûteuses en temps de la matière sonore) ? Cela a été vrai jusqu'à il y a encore peu de temps mais aujourd'hui cette frilosité marque un certain retard par rapport à la situation scientifique ; autrement dit, nous disposons bien d'un minimum épistémologique commun et d'un outillage, certes pas parfait et évidemment amené à évoluer encore (segmentation syllabique automatique à venir) mais suffisamment élaboré pour entreprendre des recherches poussées concernant en particulier les contraintes métriques qui structurent le message enfantin (ex. occurrence et distribution des fillers), l'instanciation de la structure communicative dans le message enfantin, les différentes étapes de configuration, reconfiguration et finalement stabilisation du système prosodique chez l'enfant en phase d'acquisition précoce du langage (corrélatives à la mise en place du lexique d'abord, de la syntaxe ensuite), la relation entre le formatage des objets prosodiques et les formats interactionnels. La liste ainsi dressée est bien sûr loin d'être exhaustive mais elle donne déjà une idée des domaines prosodiques qui devraient se systématiser au sein des recherches en AL.

Par ailleurs, certains regrettent l'absence de la neurophysiologie cognitive dans la formation. La thématique retenue pour l'école (zoom sur le travail de terrain, recueil de données écologiques) explique ce phénomène. Pour autant, une ouverture s'impose à l'avenir de ce côté là pour une formation complémentaire, tant le domaine est foisonnant et en pleine expansion.

Pour ce qui est des ressources à proprement parler, nous avons beaucoup discuté des outils d'annotation et d'exploitation, mais finalement peu des données elles-mêmes et surtout de la

constitution de ces données[3] (méthode d'enregistrement audio-vidéo, erreurs à éviter pour avoir une qualité sonore minimale, compression et échantillonnage).

Pour la dimension sonore, A. Lacheret se propose de mettre au plus vite sur le site de l'école quelques conseils de base qui peuvent être utiles préalablement à toute prise de sons.

4. Et la suite ?

Les suites à donner envisagées s'organisent autour de 3 volets : 1) formation complémentaire le cas échéant, 2) publication d'un ouvrage, 3) forum, liste de discussion et plate-forme web pour lesquels d'une manière générale les intervenants ont moins d'attente que les participants.

4.1 Formation(s)

La plupart des participants sont séduits par l'idée de formations complémentaires, certains même sont partants pour en organiser, avec des sessions ateliers et démo peut-être plus nombreuses et diversifiées. Différentes thématiques ont été proposées dans les formulaires d'évaluation :

- Protocoles de recueil des données
- Traitements statistiques
- Même organisation mais en restreignant la thématique soit :
    - à une tranche d'âge précise (ex. 0-3 ans)
    - à un niveau de traitement particulier (ex. lexique et sémantique ou phonologie formelle et processus d'acquisition ou encore interaction adulte-enfant)
- Zoom sur les outils sous forme d'ateliers uniquement : modélisation en acquisition du langage ; programmation en Perl, langage XML ; formation sur Intex et Nooj, Praat[4]
- Recherche et pratique clinique
- Aspects socio-politiques et institutionnels reliés à la question des corpus et de leur mutualisation
- Conventions de transcription et de formats d'échange
- Outils méthodologiques et instrumentation en neurophysiologie pour l'étude de l'AL
- Ressources pour l'étude de l'acquisition des langues secondes et bilinguisme
- Développement tardif et didactique

4.2 Publication

Pour une minorité de participants, les informations déjà disponibles sur le site de l'école suffisent (cours en ligne, bibliographie) et une publication complémentaire ne s'impose pas. Pour d'autres, beaucoup de chemin reste à parcourir pour être prêt à publier un ouvrage sur la mutualisation des données et outils. Mais la plupart sont favorables à la publication d'un ouvrage collectif, si possible accompagné d'un CD. Les thématiques proposées sont les suivantes :

- Domaine pragmatique : l'enfant et la connaissance du monde à travers l'acquisition du langage, conversations et interactions
- Les différents niveaux de traitement en acquisition (pragmatique, lexique et sémantique, morphosyntaxe, phonétique et phonologie)
- Approche instrumentale : outils et applications en acquisition du langage
- Inventaire des données écrites et orales distribuables pour les recherches en AL

[3] Notons toutefois les outils de recueil selon les tranches d'âge rappelées par H. Hilton vendredi 19 (diapositive 4 et 5).

[4] Certains proposent même judicieusement des formations récurrentes aux outils un ou deux jours par an

- Définition d'un protocole standard de description précise des ressources (outils et données) sous la forme du guide des bonnes pratiques distribué par la DGLFLF.
- synthèse des présentations en un volume (publication des actes de l'école)

4.3 Forum & plate-forme web

L'idée d'un forum et d'une plate-forme web font quasiment l'unanimité en complément de ce qui existe déjà (cf. Info-childes) mais certains se demandent pourquoi limiter le champ à la francophonie. Il s'agirait concrètement d'envisager un carrefour d'informations, géré par un modérateur, en proposant :

- Des liens vers les corpus et outils existants
- Des exemples d'analyse et de recherches croisées sur les mêmes données le cas échéant
- De la bibliographie commentée et discutée
- Des manuels d'utilisation de logiciels en ligne
- Des cours et des ateliers en ligne
- Des annonces (colloques, journées d'étude et workshops, bibliographie, etc)
- Un forum d'échange de points de vues scientifiques
- Des liens vers d'autres sites par thèmes et par domaines

4.4. Constitution d'une structure CNRS pour la mutualisation des ressources en acquisition

Nous terminerons le bilan de cette école en refaisant mention de l'un de ses impacts immédiats directement liés aux vœux de tous ses acteurs : la création d'un GDR « Acquisition » au sein du CNRS coordonné par M. Hickmann et dans le lequel la quasi-totalité des intervenants mais aussi certains participants à cette école ainsi que ses responsables scientifiques ont un rôle structurateur. Ce GDR a été doté d'un groupe de travail transversal aux thèmes de recherche envisagés dont l'objectif est de mettre en place une démarche de mutualisation des ressources en acquisition du langage dans une dimension pluridisciplinaire mais aussi translinguisitique et transcognitive.