EXMARaLDA : un système pour la constitution et l'exploitation de corpus oraux

Thomas SCHMIDT [1]
SFB 538, Université de Hambourg

Introduction

La constitution de corpus oraux pour la sociolinguistique est une tâche qui exige beaucoup de temps et de travail. L'enregistrement d'interactions authentiques et leurs transcriptions nécessitent non seulement des connaissances méthodologiques avancées, mais aussi un équipement technique spécialisé. Ces corpus sont donc des ressources précieuses, et il semble souhaitable, d'un point de vue économique aussi bien que méthodologique, de partager et de réutiliser les corpus oraux autant que possible. En pratique pourtant, différents obstacles technologiques, comme les incompatibilités entre formats de fichiers, entre logiciels ou encore entre systèmes d'exploitation, rendent cette entreprise difficile.

Nous présentons dans cet article EXMARaLDA, un système qui vise à surmonter quelques-uns de ces obstacles et ainsi à faciliter la construction, l'exploitation et la réutilisation de corpus oraux pour la sociolinguistique. Nos objectifs principaux dans le développement de ce système consistent à :

- améliorer la compatibilité des données et à faciliter ainsi leur échange entre les chercheurs et les environnements techniques (différents logiciels, différents systèmes d'exploitation, etc.) ;
- exploiter d'une façon conséquente les capacités hypertextuelles et multimédias des ordinateurs modernes pour la constitution et la présentation de données linguistiques ;
- créer les conditions nécessaires à l'archivage à long terme des données

Le système comprend trois outils principaux – un éditeur de transcription, un outil pour l'administration des corpus, et un outil de fouille, qui seront décrits en détail § 3. Bien qu'EXMARaLDA puisse être (et soit) utilisé dans divers domaines de recherche linguistique (p. ex. la phonologie, la dialectologie, la didactique des langues), il est optimisé pour les approches s'intéressant aux rapports entre les faits linguistiques et les paramètres ethnographiques. Ainsi, les caractéristiques suivantes montrent qu'EXMARaLDA se prête bien à des analyses sociolinguistiques :

- Les outils permettent au chercheur de relier d'une façon simple et systématique la description du comportement verbal – c'est-à-dire la transcription – à différentes variables de contexte – par exemple une caractérisation des locuteurs ou de la situation d'interaction. Toutes ces métadonnées sont disponibles lors de la recherche automatique dans les transcriptions. Le chercheur peut en profiter, soit a priori, en présélectionnant dans un corpus des transcriptions d'un certain type (e.g. toutes les données des locuteurs d'une certaine région et d'un certain âge),

1. Je remercie Cyrille Granget (Université de Nantes) pour de nombreuses corrections et suggestions d'amélioration.

soit a posteriori, en groupant et quantifiant selon de tels paramètres les résultats d'une fouille.

- Inspiré par les méthodologies de la pragmatique fonctionnelle et du contextualisme à la Sinclair, EXMARaLDA tâche de rendre possible des analyses qui sont *corpus driven* (qui émanent des corpus) plutôt qu'uniquement *corpus based* (qui exploitent un corpus, v. la distinction de Teubert 2005). En d'autres termes, le système essaie d'encourager une approche exploratoire en assistant le chercheur dans le développement graduel d'hypothèses à partir du matériel empirique. En pratique, cela nécessite la possibilité d'appliquer des catégories *ad hoc* à des résultats de recherche et de les affiner, de les améliorer, voire de les réviser dans une démarche heuristique.
- Pour les mêmes motifs, EXMARaLDA attache beaucoup d'importance aux méthodes de visualisation, car la démarche heuristique peut tirer un grand profit d'une présentation adéquate des relations structurelles (comme les relations temporelles et les relations hiérarchiques entre annotations) dans les données. Par exemple, les discours à plusieurs locuteurs sont caractérisés par des événements simultanés (chevauchement, simultanéité des actions verbales et non verbales) et par des relations séquentielles (*turn taking*, séquences question / réponse). Différents principes de visualisation qui accentuent l'un ou l'autre aspect (e.g. notation sous forme de partition musicale *vs* notation sous forme de script dialogique) peuvent être privilégiés selon les intérêts du chercheur. La technologie des langues de balisage (XML et *stylesheets*) dont EXMARaLDA se sert permet de créer pour un même ensemble de données plusieurs visualisations optimisées pour différents objets.

Nous expérimentons ce système depuis sept ans, surtout dans le cadre de recherches sur différents aspects du multilinguisme (en tant que phénomène social et en tant que phénomène individuel). Plusieurs grands corpus oraux ont été ou sont en train d'être construits avec l'aide d'EXMARaLDA, parmi lesquels un corpus sur la communication interscandinave, un corpus sur l'utilisation des langues régionales dans le nord de l'Allemagne, un corpus sur le bilinguisme turco-allemand et un corpus sur l'interprétation dans les hôpitaux [2].

Dans le paragraphe suivant, nous proposons un aperçu des caractéristiques principales du système. Le § 3 fournira quelques informations détaillées sur les différents logiciels du système. Nous nous concentrons ici sur les informations que nous jugeons essentielles pour un utilisateur ou une utilisatrice potentiel(le) d'EXMARaLDA (v. Schmidt 2005 et Schmidt *et al.* 2009 pour une description plus détaillée des aspects techniques du système, et Schmidt & Wörner 2009 et Schmidt & Wörner 2010 pour l'emploi d'EXMARaLDA respectivement en pragmatique et en phonologie).

2. Modèles de données

EXMARaLDA fournit un modèle de données pour représenter des données linguistiques, c'est-à-dire les descriptions du comportement communicatif des différents locuteurs d'une interaction verbale, ainsi qu'un modèle de données pour représenter des données extra-linguistiques, c'est-â-dire les caractéristiques sociographiques des locuteurs (âge, sexe, etc.), des informations sur les circonstances de la communication (lieu, type d'interaction, etc.), et des informations sur les circonstances de sa documentation (outil d'enregistrement, transcripteur, etc.).

2. Les corpus sont disponibles à partir de http://corpora.exmaralda.org et http://sin.sign-lang.uni-hamburg.de/drupal/, respectivement.

En ce qui concerne les données linguistiques, le modèle doit être capable de représenter des descriptions à différents niveaux linguistiques et leurs relations temporelles. Une transcription typique, comme en figure 1, comprend une transcription (orthographique ou phonétique) du comportement verbal des locuteurs (lignes marquées « [v] »), y compris, éventuellement, les descriptions des pauses et des passages incompréhensibles pour le transcripteur. D'autre part, la transcription du comportement verbal est souvent supplémentée par des descriptions du comportement non verbal (la gestualité, la mimique, lignes marquées « [nv] »), par des annotations prosodiques (ligne marquée « [int] ») et, le cas échéant, par d'autres annotations qui servent à rendre la transcription plus accessible au lecteur (comme les traductions anglaises dans les lignes marquées « [en] »).

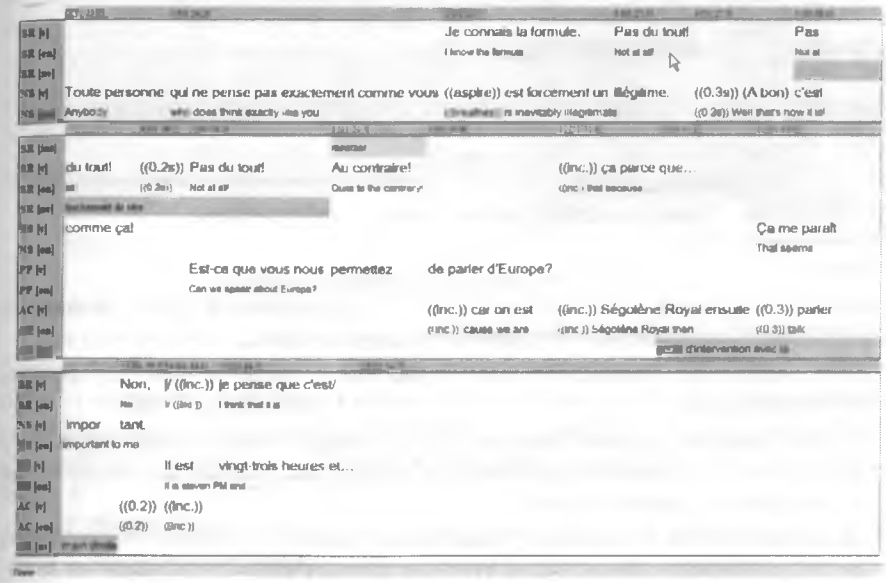

Figure 1.- Transcription sous forme de partition musicale du débat télévisé entre Nicolas Sarkozy et Ségolène Royale durant la campagne présidentielle de 2007.

Entre ces différents éléments de description, il existe trois relations structurelles différentes : l'enchaînement (comme entre les énoncés d'un même locuteur), la simultanéité (comme dans les chevauchements, et entre les actions verbales et non verbales d'un locuteur) et l'équivalence (comme entre un énoncé français et sa traduction anglaise). Dans le format de partition musicale illustré ci-dessus, les éléments qui se suivent temporellement sont alignés de gauche à droite, tandis que les éléments simultanés ou équivalents sont superposés. Le modèle d'EXMARaLDA pour les données linguistiques est construit sur la base de partitions musicales. Il suit la proposition de Bird & Liberman (2001) de représenter les annotations linguistiques sous forme de « graphe d'annotation ». Ainsi, le modèle de données d'EXMARaLDA et ses formats de fichiers deviennent compatibles et interopérables avec un grand nombre d'autres formats de fichiers utilisés pour les corpus oraux (comme Praat, ELAN, Transcriber, TASX, etc., v. Schmidt *et al.* 2009).

En ce qui concerne les données extra-linguistiques, le modèle de données d'EXMARaLDA s'appuie sur deux éléments fondamentaux : la communication (en tant qu'événement communicatif) et le locuteur. Ces deux éléments peuvent être décrits par un certain nombre de métadonnées (c'est-à-dire des paires attribut-valeur

comme « sexe : féminin » ou « lieu : Paris »), et ils peuvent être mis en relation n : m, représentant le fait que d'une part plusieurs locuteurs peuvent participer à une communication, et, d'autre part, un locuteur spécifique peut participer à plusieurs communications. Le lien entre données extra-linguistiques et linguistiques est établi en attribuant une ou plusieurs transcriptions et un ou plusieurs enregistrements à une communication.

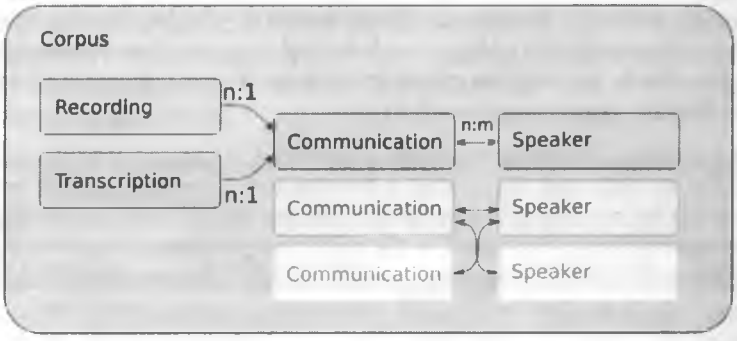

Figure 2.- Modèle pour les données extra-linguistiques

3. Outils

EXMARaLDA fournit trois outils principaux qui répondent aux trois démarches principales de la constitution et de l'analyse de corpus oraux :

- La transcription et l'annotation des enregistrements audio ou vidéo se font dans le *Partitur-Editor*.
- L'assemblage de plusieurs transcriptions et enregistrements en un corpus et la description des événements communicatifs et des locuteurs par des métadonnées se font dans le *Corpus Manager*.
- La recherche dans un corpus de phénomènes transcrits ou annotés, la corrélation de ces phénomènes linguistiques avec des données extra-linguistiques et le filtrage, la catégorisation et la quantification de résultats de recherche se font dans *EXAKT*.

Les sections suivantes se proposent d'illustrer ces trois outils plus en détail.

3.1 Partitur-Editor

Le Partitur-Editor présente une transcription sous forme de partition musicale par un ensemble de lignes qui correspondent à un locuteur et à un niveau descriptif spécifiques (verbal, non verbal, etc.). Chaque ligne est composée de plusieurs événements qui contiennent les transcriptions ou descriptions des éléments de l'interaction, et qui sont arrangés d'une façon qui correspond à leur occurrence dans l'enregistrement. L'enregistrement même apparaît sous forme d'oscillogramme (s'il s'agit d'un fichier audio .wav) ainsi que dans un écran (s'il s'agit d'un fichier vidéo). La navigation dans la transcription est synchronisée avec la navigation dans l'enregistrement de sorte que la partie sélectionnée dans la partition musicale corresponde toujours à la partie sélectionnée dans l'oscillogramme ou dans l'outil vidéo. Cela permet au transcripteur de jouer et rejouer des extraits de l'enregistrement tout en se concentrant sur leur description adéquate dans la transcription. Les intervalles temporels et les descriptions textuelles peuvent être modifiés à tout moment dans le processus de transcription. Il est aussi possible d'ajouter (ou de supprimer) des lignes ou des locuteurs après avoir commencé la transcription.

Figure 3.- Capture d'écran du Partitur-Editor

Les transcriptions complétées peuvent être imprimées ou exportées en différents formats de présentation pour l'affichage dans un traitement de texte (p.ex. Word) ou dans un navigateur Web (p.ex. Firefox). La présentation n'est pas limitée au format de partition musicale (comme dans la figure 1 ci-dessus) – il est également possible de visualiser la transcription sous forme d'une simple liste d'énoncés comme dans la figure 4.

1 NS : Toute personne qui ne pense pas exactement comme vous ((respire)) est forcément illégitime.

en : Anybody | who does not think exactly like you | ((breathes)) is inevitably illegitimate.

2 SR : Je connais la formule.

en : I know the formula.

3 SR : Pas du tout !

en : Not at all!

4 NS : ((0.3s)) (A bon) c'est comme ça !

en : ((0.3s)) Well that's how it is!

5 SR : Pas du tout !

en : Not at all!

6 SR : ((0.2s)) Pas du tout !

en : ((0.2s)) | Not at all!

7 PP : Est-ce que vous nous permettez de parler d'Europe ?

en : Can we speak about Europe?

8 SR : Au contraire !

int : montant

en : Quite to the contrary!

9 AC : ((inc.)) car on est ((inc.)) Ségolène Royal ensuite ((0.3)) parler ((0.2)) ((inc.))

en : ((inc.)) cause we are | ((inc.)) Ségolène Royal then | ((0.3)) talk | ((0.2)) | ((inc.))

10 SR : ((inc.)) ça parce que…

en : ((inc.) that because…

11 NS : Ça me paraît important.

en : That seems important to me.

Figure 4.- Visualisation d'une transcription sous forme de liste d'énoncés

Outre cette fonctionnalité de base – créer et visualiser une transcription – le Partitur-Editor offre d'autres fonctions pour favoriser un travail efficace avec des transcriptions :

- Des fonctions pour échanger des données avec d'autres logiciels de transcription. À présent, des fonctions permettent d'importer et d'exporter des données de Praat, d'ELAN, de TASX et de FOLKER, et d'exporter une transcription en format CHAT ou TEI (Schmidt 2005).
- Des fonctions pour vérifier la consistance des données, c'est-à-dire pour vérifier si la transcription est conforme à une certaine convention de transcription [3], et s'il ne contient que des relations structurelles permises [4].
- Une fonction pour afficher une liste alphabétique de mots, avec une quantification des types et des occurrences (*tokens*), dans la transcription.
- Un outil (« clavier virtuel ») pour transcrire en alphabet phonétique international.
- Un outil (*annotation panel*) pour appliquer systématiquement une liste de catégories (*tag set*) au texte transcrit.
- Un outil (*Praat panel*) pour afficher des visualisations du signal sonore (spectrogramme, contour d'intonation etc.).

3.2. Corpus Manager

Le Corpus Manager sert à constituer et administrer un corpus avec ses métadonnées. L'outil affiche l'ensemble des données en deux listes – une liste avec des communications auxquelles sont attribuées les transcriptions et les enregistrements, et une liste avec des locuteurs. Tous ces éléments peuvent être décrits séparément par une liste de métadonnées. L'utilisateur peut lui-même définir les attributs de cette liste et leurs valeurs, mais il a aussi la possibilité de choisir des éléments prédéfinis pour des descriptions fréquentes (comme les lieux, les dates, les langues etc.). Entre les communications et les locuteurs, il existe une relation n : m, c'est-à-dire, une communication peut être liée à plusieurs locuteurs, et un locuteur peut être lié à plusieurs communications. En reliant communications et locuteurs d'une telle manière, non-hiérarchique, on évite de répéter des informations identiques en plusieurs lieux dans la description du corpus.

3. À présent, les systèmes de transcription suivant sont soutenus, entre autres : HIAT (Rehbein *et al.* 2004), GAT (Selting *et al.* 1998) et CHAT (MacWhinney 2000).

4. Par exemple, un chevauchement d'un locuteur avec lui-même ou une annotation pour laquelle n'existe pas de texte annoté sont des relations structurelles interdites.

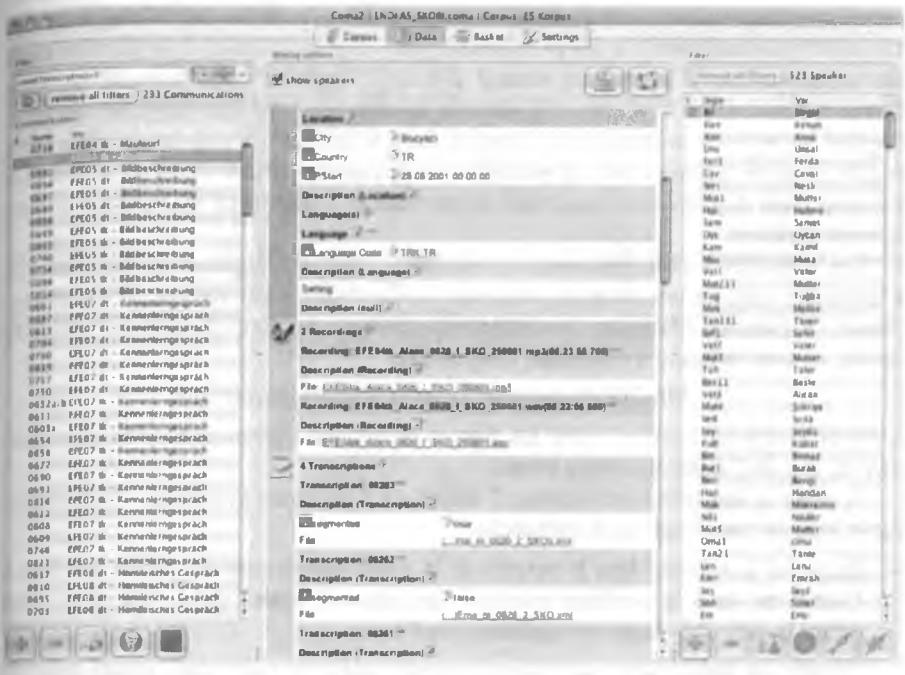

Figure 5.- Capture d'écran du Corpus Manager

Avec l'aide du Corpus Manager, le chercheur peut aussi filtrer son corpus selon les métadonnées qu'il a fournies. Par exemple, il peut créer, avec l'aide de filtres appropriés, un sous-corpus qui ne contient que les communications ayant lieu à Paris et auxquelles participent au moins deux locuteurs masculins de moins de vingt ans. De tels sous-corpus, comme le corpus entier, peuvent alors être analysés au moyen du logiciel EXAKT. Au-delà de cette fonctionnalité, le Corpus Manager comprend aussi des fonctions pour vérifier l'intégrité et l'homogénéité des métadonnées ainsi que des opérations pour la vérification de l'ensemble des transcriptions.

3.3 EXAKT

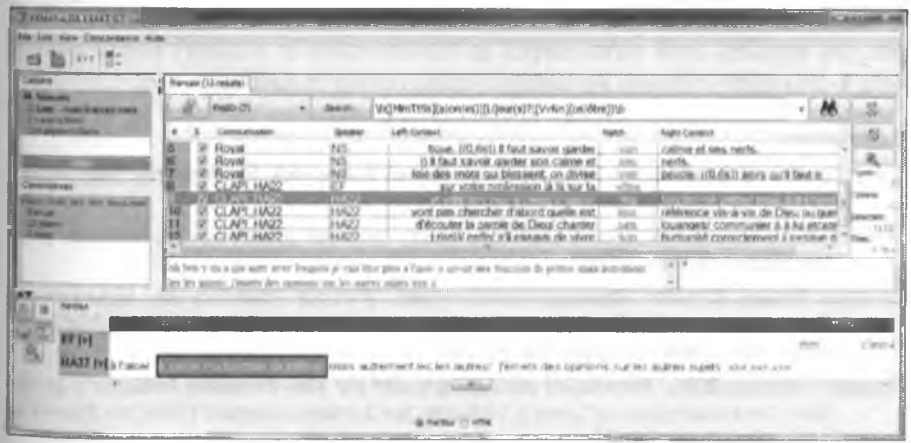

Figure 6.- Capture d'écran d'EXAKT

EXAKT (*EXMARaLDA Analyse- und Konkordanztool,* Outil d'analyse et de concordance) permet au chercheur d'ouvrir un corpus créé avec le Corpus Manager et d'y chercher systématiquement des phénomènes transcrits ou annotés et les corréler avec les métadonnées. L'instrument central d'EXAKT est un concordancier de type KWIC (*Keyword in Context*) : l'utilisateur formule une requête de recherche et le logiciel lui retourne toutes les occurrences dans le corpus qui correspondent à cette requête avec un peu de contexte gauche et droit, ainsi que l'identifiant de la communication dans laquelle a été trouvée l'occurrence et l'identifiant du locuteur auquel elle est attribuée.

Les recherches sont typiquement formulées comme « expressions régulières », c'est-à-dire comme une formule désignant une certaine combinaison de caractères. Ainsi, l'expression régulière « tou(s|t|te|tes) » désigne les formes *tous*, *tout*, *toute* et *toutes,* et dans la figure 6 ci-dessus, l'expression

\b([MmTtSs](a|on|es)|[Ll]eur(s)?|[VvNn](os|ôtre))\b

a été utilisée pour trouver toutes les formes de pronoms possessifs (*mon, ma, mes, ton, ta, tes,* etc.) dans un corpus donné.

À partir d'une telle concordance, l'utilisateur a plusieurs possibilités pour explorer et analyser ses résultats de recherche :

- Regarder le résultat dans le cotexte de la transcription entière, qui est affichée à l'écran ou dans la partition musicale qui apparaît également à l'écran ou bien parmi la liste dans la partie supérieure de l'écran.
- Écouter ou voir l'extrait de l'enregistrement audio ou vidéo dans lequel apparaît le phénomène recherché.
- Faire apparaître pour chaque occurrence recensée les locuteurs correspondant et certaines métadonnées liées à la communication.
- Ajouter, dans des colonnes supplémentaires, des catégorisations issues de l'analyse des résultats de la fouille automatique.

En sélectionnant ou en filtrant les résultats, EXAKT permet ainsi d'affiner progressivement une recherche dans un corpus.

4. Informations pratiques

Les outils d'EXMARaLDA, utilisables sous Windows, Macintosh, Linux et Unix, peuvent être téléchargés gratuitement à partir du site http://www.exmaralda.org.

Au même endroit se trouvent de la documentation et des tutoriels pour l'utilisation des logiciels, ainsi qu'un corpus de démonstration et plusieurs corpus publiés suite à des projets effectués au centre de recherche sur le multilinguisme.

Références

BIRD Steven and LIBERMAN Mark, 2001, "A formal framework for linguistic annotation", *Speech Communication,* 3, p. 323-360.

MACWHINNEY Brian, 2000, *The CHILDES project : tools for analyzing talk*, Mahwah [NJ], Lawrence Erlbaum (3rd ed.).

REHBEIN Jochen, SCHMIDT Thomas, MEYER Bernd, WATZKE Franziska und HERKENRATH Annette, 2004, *Handbuch für das computergestützte Transkribieren nach HIAT, Arbeiten zur Mehrsprachigkeit,* Série B 56.

SCHMIDT Thomas, 2005, *Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech, Arbeiten zur Mehrsprachigkeit (Working Papers in Multilingualism)*, Série B 62.

SCHMIDT Thomas, DUNCAN Susan, EHMER Oliver, HOYT Jeffrey, KIPP Michael, LOEHR Dan, MAGNUSSON Magnus, ROSE Travis and SLOETJES Han, 2009, "An exchange format for multimodal annotations", in M. Kipp, J.-Cl. Martin *et alii* (eds.), *Multimodal Corpora,* Dordrecht, Springer, p. 207-221.

SCHMIDT Thomas and WÖRNER Kai, 2009, "Creating, analysing and sharing spoken language corpora for pragmatic research", *Pragmatics*, 19–4, p. 565–582.

SCHMIDT Thomas and WÖRNER Kai, 2010, « EXMARaLDA », dans J. Durand, U. Gut and G. Kristoffersen (eds.), *Oxford Handbook of Corpus Phonology*, Oxford, Oxford University Press [sous presse].

SELTING Margret *et alii*, 1998, „Gesprächsanalytisches Transkriptionssystem (GAT)", *Linguistische Berichte*, 173, p. 91-122.

TEUBERT Wolfgang, 2005, "My version of corpus linguistics", *International Journal of Corpus Linguistics*, 10-1, p. 1–13.