Cahiers de l'ILSL-45 (2016 : 45-64)

Prévisibilité dans le développement linguistique et importance des corpus électroniques[1]

Wolfgang U. DRESSLER[1]
Katharina KORECKY-KRÖLL[2]
Karlheinz MÖRTH[3]
Institut für Corpuslinguistik und Texttechnologie der Österreichischen Akademie der Wissenschaften (A) (1, 2)
Austrian Centre for Digital Humanities, Österreichische Akademie der Wissenschaften (A) (3)
Correspondance : wolfgang.dressler@univie.ac.at

1. Prévisibilité et prédictibilité

Les notions de prévisibilité et de prédictibilité, la seconde plus puissante que la première (cf. BARRETT & STANFORD 2006), sont discutées dans de nombreux champs de la linguistique, mais souvent dans une perspective différente. Par exemple, en phonétique expérimentale, les résultats des expérimentations sont prédits sur la base d'hypothèses que la méthodologie a opérationnalisées, parallèlement à ce qui se pratique dans les autres sciences expérimentales. En sociolinguistique, la prévisibilité est comprise de la même manière que dans les autres sciences humaines. Dans les grammaires formelles, selon le modèle de CHOMSKY (1957) qui cherche à traiter la théorie de la grammaire comme une science exacte telle la physique ou l'astronomie, une prédiction signifie habituellement que la grammaire d'une langue prédit quelles sont les phrases à classer comme correctes et incorrectes parmi les phrases réalisées ou possibles.

Les prédictions probabilistes sont plus faibles. Prenons par exemple la distribution des composés et de leurs

[1] Traduction de *Vorhersehbarkeit in der Sprachentwicklung und die Bedeutung elektronischer Corpora*, effectuée par Marianne Kilani-Schoch.

constituants respectifs dans les textes. Nous avons soutenu que ceux-ci se répartissent de façon inverse suivant qu'il s'agit du titre ou du corps du texte : les composés apparaissent plutôt dans les titres (ex. *Ausfuhrzölle* 'droits à l'exportation') et les constituants ou membres des composés plutôt dans le corps des textes (*Ausfuhr* 'exportation', *Zölle* 'droits') (DRESSLER & MÖRTH 2012a; b). Une telle prédiction est rendue possible par la convergence de deux facteurs : la pression actuelle en faveur de la brièveté des titres et le fait que les composés sont plus courts que la combinaison syntaxique de leurs parties, comparez *Meeresfreiheit* 'liberté des mers' et *Freiheit der Meere* 'liberté des mers' (sur la convergence des motivations en linguistique, voir DRESSLER *et al.* 2014).

La vérification empirique de cette prédiction probabiliste est difficilement réalisable à partir de la lecture de textes imprimés, à la différence d'une analyse linguistique de corpus basée sur des textes électroniques annotés. C'est ainsi que dans l'étude de DRESSLER & MÖRTH (2012b), le nombre des textes à examiner a considérablement diminué et est passé de 13 000 à 2087 grâce la segmentation automatique des composés en leurs constituants (au moyen du programme *Noun Splitter* de l'*Institut für Corpuslinguistik und Texttechnologie* ou ICLTT 'Institut de linguistique de corpus et de technologie du texte' de l'Académie des sciences autrichienne). Le temps de lecture en a été réduit de plus de 80%. Cette diminution a été d'autant plus marquée que les composés identifiés dans les textes et leurs constituants potentiels ont pu être extraits automatiquement. Les résultats ont dû néanmoins être vérifiés manuellement dans chaque texte.

Notons que les composés représentent un défi considérable pour tout ce qui est traitement automatique des langues naturelles (au sens de *Natural Language Processing*), en particulier pour les nombreuses procédures de traduction

automatique ou de récupération de l'information. On compte de nombreux essais d'automatisation de la segmentation des composés – des techniques d'apprentissage automatique sont notamment utilisées – mais nos travaux avec le *Noun Splitter* ont montré que dans bien des cas le problème ne peut être résolu que par le recours à de grands dictionnaires numériques, malheureusement encore en nombre insuffisant aujourd'hui.

Dans la présente contribution, nous allons concentrer notre attention sur la prévisibilité et la prédictibilité probabiliste dans deux domaines du développement linguistique, d'une part l'acquisition par l'enfant dans sa phase précoce (particulièrement l'acquisition de la grammaire), d'autre part le changement diachronique, au cours de l'histoire, où nous tenterons aussi une comparaison avec les sciences historiques classiques.

2. Acquisition de la langue première

Considérons d'abord l'acquisition de la langue première par les enfants ainsi que l'acquisition successive d'une langue seconde. Ici nous pouvons sans grands risques prévoir des différences interindividuelles dans le rythme d'acquisition des enfants en fonction de la typologie linguistique, c'est-à-dire par exemple prévoir que la vitesse d'acquisition de la morphologie, en l'occurrence celle de la flexion, dépend de de la richesse relative, de la transparence et de l'univocité de la morphologie de la langue-cible (DRESSLER 2010 ; SLOBIN 1985a ; 1985b ; 1992 ; 1997a ; 1997b ; XANTHOS *et al.* 2011). Ainsi, dans les langues qui font l'objet des recherches que nous venons de mentionner, le pluriel est-il d'abord acquis en turc, parce que la formation biunivoque du pluriel turc peut être prévue, c'est-à-dire que la terminaison du pluriel est toujours *-ler* ou *-lar*, selon la voyelle qui précède, par ex. *ev-ler* 'maisons', *oda-lar* 'chambres'. En allemand, par contre, le pluriel, *-(e)n, -s, -e, -er* est largement imprévisible – mais

presque toujours conventionnel, c'est-à-dire fixé pour chaque mot (sur les tendances en matière de prévisibilité partielle des formes de pluriel allemand, voir KÖPCKE 1993 ; WEGENER 2004).

L'étude longitudinale de la langue spontanée requiert une préparation informatique précise des données, par exemple à l'aide de l'ensemble des programmes CLAN (MACWHINNEY 2000) qui permettent une annotation automatique et l'exploitation de données massives (voir également KORECKY-KRÖLL soumis ; LAAHA & KORECKY-KRÖLL à paraitre). Dans un premier temps les enregistrements sont saisis dans un format défini, le format CHAT (MACWHINNEY 2000), puis ils sont vérifiés selon trois procédures différentes : d'abord la transcription est contrôlée par un expert pour minimiser les erreurs d'écoute, ensuite l'alignement des niveaux de CHAT est vérifié à l'aide du programme CHECK, enfin les lacunes dans la standardisation de mots (par ex. *gehma* pour *gehen wir* 'allons') sont repérées. Une fois ces étapes franchies avec succès, le programme MOR ainsi qu'un lexique électronique recherchent quelles entrées de la nouvelle transcription à coder ne sont pas encore enregistrées dans le lexique. L'annotation de ces entrées doit être effectuée manuellement. Selon les objectifs de la recherche, il peut être décidé d'ajouter des annotations supplémentaires et de réaliser une annotation plus détaillée (par exemple distinguant entre verbes faibles et forts et les divisant en sous-classes).

Dans un deuxième temps, on procède à l'annotation morphologique automatique des mots de chaque ligne de la transcription. Comme toutes les formes ambigües ont été relevées, un travail supplémentaire de désambigüisation est nécessaire. Il consiste à décider laquelle des annotations possibles s'impose en contexte, par exemple si le mot *der* dans *da ist der Hund* 'le chien est là' est seulement un article et non pas un pronom démonstratif ou un pronom relatif.

Grâce aux divers programmes de contrôle, ces données codées peuvent rapidement se prêter à une analyse quantitative.

L'acquisition du pluriel en turc est facilitée par le fait que la formation du pluriel est transparente : dans *ev-ler* 'maison-pluriel', *oda-lar* 'salle-pluriel', etc. il est très facile de séparer le marqueur du pluriel du radical du mot. L'allemand en revanche a deux marqueurs, la désinence et, dans beaucoup de mots, l'umlaut, dont l'occurrence n'est généralement pas prévisible. Dans les pluriels allemands *Häus-er* 'maisons', *Mütter* 'mères', le marqueur de pluriel ne peut être isolé aussi aisément des bases *Haus* 'maison', *Mutter* 'mère' qu'il ne l'est en turc. À ces pluriels s'ajoutent des pluriels sans marqueur au niveau du radical, comme *der – die Lehrer* 'le, les professeurs'. Ils constituent un troisième exemple de pluriel qui n'est pas non plus transparent et univoque en général. Car dans le pluriel *die* du singulier *der/das*, il n'y a pas de désinence plurielle transparente segmentable ; en outre *die* est aussi le féminin singulier, et l'indéfini se caractérise au pluriel par l'absence d'article. Ceci montre que la formation du pluriel en allemand n'est ni biunivoque ni univoque, mais ambigüe.

Enfin, intervient le fait que le turc est morphologiquement bien plus riche que l'allemand (spécialement dans la flexion), c'est-à-dire qu'il exprime un plus grand nombre de catégories grammaticales, et par exemple recourt souvent à des désinences au lieu de subordonnées. Un exemple de langue très pauvre en flexion est l'anglais. Or on constate que les enfants turcs prêtent beaucoup plus d'attention à l'acquisition de la flexion (par ex. le pluriel ou les formes verbales) que les enfants qui acquièrent l'anglais comme langue première. La flexion anglaise n'est en outre ni biunivoque (cf. *cow-s* 'vaches' et *ox-en* 'bœufs'), ni toujours transparente (par ex. dans la formation du pluriel *wife* 'femme' – *wiv-es*, *mouse* 'souris' – *mice*). La prédiction,

avérée, est donc que les petits anglophones acquièrent la morphologie de l'anglais plus tard et plus lentement que les enfants turcs. Et, comme prédit, les enfants germanophones acquièrent la flexion, par ex. le pluriel, plus tôt et plus vite que les enfants anglophones, mais plus lentement et plus tard que les enfants turcs. Des langues slaves comme le russe ont aussi une flexion peu transparente et le plus souvent ambigüe (par ex. dans la formation du pluriel), mais plus riche que l'allemand et un peu moins que le turc. À partir de l'interaction des trois facteurs de richesse morphologique, biunivocité et transparence, l'échelle suivante d'acquisition de la morphologie, basée sur des données empiriques, est prévisible : turc – russe – allemand – anglais.

Cette échelle est valable non seulement pour l'acquisition typique, sans problème spécifique, mais aussi pour l'acquisition non-typique perturbée ou retardée. Néanmoins, dans le cas de l'acquisition non-typique, la nature et le degré de sévérité du trouble sont des facteurs plus importants que ceux que nous venons de mentionner (cf. BARTKE & SIEGMÜLLER 2004 ; BAVIN 2009). La prévisibilité du développement langagier y est comparable à la prévisibilité de l'évolution de la maladie dans le pronostic médical, à ceci près que dans la recherche sur l'acquisition du langage, le degré d'incertitude du pronostic ne peut pas (pas encore ?) être mesuré quantitativement.

Dans les prévisions typologiques que nous venons de présenter, nous avons procédé à une simplification importante. Car ce n'est naturellement pas la structure de la langue à acquérir qui agit directement sur le processus de développement mais la réalisation de ce système linguistique dans le parler adressé à l'enfant, ce qu'on appelle l'*input*, dont dépend la production de l'enfant telle que nous l'analysons. Et cet *input* que la personne responsable (le plus souvent la mère) adresse au jeune enfant peut se différencier fortement de la langue des adultes (CAMERON-FAULKNER *et al.* 2003 ;

RAVID *et al.* 2008 ; XANTHOS *et al.* 2011). Par exemple, la morphologie turque comporte un certain nombre de difficultés, tout particulièrement les longues formes fléchies, c'est-à-dire les chaines de désinences successives, auxquelles s'ajoutent les éventuels changements de position de ces suffixes, sans que la signification morphologique n'en soit modifiée. Aucune mère turque, cependant, ne parle à des enfants en bas âge en utilisant des formes aussi complexes. C'est la raison pour laquelle, dans notre « Crosslinguistic Project on Pre- and Protomorphology in Language Acquisition » (DRESSLER 2010 ; XANTHOS *et al.* 2011) qui porte sur la production d'enfants provenant de 18 pays différents, nous enregistrons et étudions aussi bien le développement longitudinal de l'input que celui de l'output.

3. Input, saisie, intégration, production

Il n'y a pas non plus de relation directe entre l'input et l'output de l'enfant, dans la mesure où les êtres humains ne sont pas des perroquets. La chaine causale entre production de l'enfant (*output*) et input est la suivante :

> input → saisie (*intake*) → intégration (*uptake*) → production (*output*).

La saisie (*intake*) correspond à cette partie de l'input que l'enfant capte, notamment pour des raisons de perception. Les syllabes accentuées sont ainsi mieux perçues que les syllabes atones. Les enfants (en tout cas les plus jeunes) enregistrent plus facilement la partie finale d'unités structurales comme les syntagmes nominaux *des klein-en Kind-es* 'du petit enfant', *die klein-en Kind-er* 'les petits enfants' que la partie initiale. Bien que l'occurrence de l'article soit beaucoup plus fréquente dans l'input que celle des désinences flexionnelles (comme dans l'exemple *-en, -es, -er* ci-dessus), les désinences sont acquises plus tôt par les jeunes enfants que l'article (KORECKY-KRÖLL 2011: 190).

Les différences interindividuelles mentionnées plus haut dépendent d'autres facteurs encore, par exemple de l'étape intermédiaire d'intégration (*uptake*) qui consiste dans la forme que les enfants donnent à leur grammaire en la construisant à partir des éléments saisis (*intake*). Dans la relation entre intégration et production enfantine interviennent des facteurs relatifs à l'extraction des données : ainsi les enfants produisent-ils moins "d'erreurs" (dans la perspective de la langue adulte) dans leur usage spontané de la langue que dans des tests formels, comme nous l'avons montré pour l'acquisition des pluriels et des cas en allemand (KORECKY-KRÖLL 2011 ; KORECKY-KRÖLL & DRESSLER 2015). C'est la raison pour laquelle la recherche ne peut se limiter à des tests formels, comme les tests psychologiques et pédagogiques réalisés à des fins de diagnostic et de thérapie (qui ne prennent ni l'input, ni le développement langagier individuel en considération).

En d'autres termes, la prévisibilité en ce qui concerne l'acquisition de la langue première d'un enfant est tributaire d'un nombre important de facteurs. Le principal problème réside donc dans la manière de les évaluer et de les hiérarchiser. L'ordre de présentation suivi ici reflète notre proposition de hiérarchisation.

4. Variable sociolinguistique

En ce qui concerne le recueil des données, une variable supplémentaire est pertinente, à savoir une variable sociolinguistique. Les recherches sur le langage des enfants, en particulier sur la langue spontanée, sont dans la majeure partie des cas conduites auprès de familles dont le niveau de formation est élevé, parce que l'accès est à bien des égards beaucoup plus facile (c'est également le cas dans notre « Crosslinguistic Project on Pre- and Protomorphology in Language Acquisition »). Mais la question se pose de savoir

comment le développement du langage s'effectue dans des familles dont le niveau de formation est peu élevé.

Cet aspect est étudié de manière systématique dans notre programme de recherche INPUT[2] (KORECKY-KRÖLL *et al.* 2015), dans le cadre d'une approche linguistique basée sur corpus, s'appuyant sur des travaux américains et israéliens (HART & RISLEY 1995 ; WEISLEDER & FERNALD 2013). La variable socioculturelle du niveau socioéconomique[3] et du niveau de formation déterminent le degré de richesse de l'input enfantin, parce qu'en règle générale, les mères (plus exactement les personnes en charge des enfants) dont le niveau de formation est plus élevé développent plus d'interactions verbales avec leurs enfants que celles dont le niveau de formation est moins élevé. La chaine causale est donc la suivante :

> niveau socioéconomique → richesse de l'input → saisie → intégration → production

Elle permet de formuler deux prévisions : premièrement l'output des enfants provenant de familles dont le niveau de formation est limité est moins riche que celui d'enfants provenant de familles dont le niveau de formation est plus élevé ; deuxièmement leur développement langagier est plus tardif que celui d'enfants de familles dont le niveau de formation est plus élevé. Les observations dans ce sens se sont multipliées à partir du 20e siècle (cf. OEVERMANN 1972). Mais il n'y a pas pour autant de rapport direct entre niveau socioéconomique et output enfantin, car entre les deux intervient de manière déterminante le style communicatif des parents avec les enfants.

La richesse de l'input et de l'output peut être évaluée par diverses mesures. Dans notre projet nous recourons à la

---

[2] Le projet est financé par l'Académie des sciences de Vienne et le fonds Technologie.

[3] En anglais, *SES* pour *socioeconomic status*.

longueur moyenne des phrases, la diversité du vocabulaire, la quantité de mots utilisés, des mesures de complexité des différentes catégories grammaticales, l'élaboration textuelle d'un récit rapporté, ainsi que la manière dont les adultes s'adressent à leurs enfants (par exemple comment ils réagissent aux erreurs des enfants, cf. KILANI-SCHOCH *et al.* 2009). Sur tous ces points on observe des différences dans l'input et dans l'output en fonction du niveau socioéconomique. À ce stade de la recherche, ces différences correspondent au minimum à des tendances, mais nous pensons parvenir, au terme de l'analyse, à des résultats statistiquement significatifs. À ces différences s'ajoute, comme prévu, un retard du développement linguistique des enfants dont les familles ont un niveau de formation limité. On constate chez ces enfants un retard dans l'acquisition des groupes consonantiques, par exemple à la finale des mots *Obst* 'fruit', *du lobst* 'tu loues' (KORECKY-KRÖLL & DRESSLER à paraitre). La question de savoir si ces enfants, devenus adultes, n'atteindront pas non plus le niveau linguistique des enfants de familles dont le niveau de formation est plus élevé se pose. On s'attend à la persistance d'une différence en ce qui concerne le degré de complexité des phrases et des mots (voir déjà OEVERMANN 1972).

Notre recherche porte aussi sur l'effet de l'input adressé aux enfants dans les garderies. Cependant il est encore plus difficile de prévoir si les éducatrices compensent dans leurs interactions avec les enfants les déficits linguistiques familiaux. En effet, des facteurs relatifs à la formation des éducatrices à côté de variables pédagogiques et personnelles, de la taille et de la composition du groupe d'enfants, ainsi que des facteurs spécifiques aux garderies interviennent.

Considérons maintenant le deuxième domaine de prévisibilité du développement linguistique, le changement des langues au cours de l'histoire, c'est-à-dire le changement

diachronique. Ce domaine est en relation étroite avec celui que nous venons d'examiner puisque le produit de l'acquisition du langage ne coïncide pas avec tous les aspects de la langue des parents, c'est-à-dire avec la langue adulte, et qu'il implique donc le changement linguistique. Le rôle précis de l'acquisition du langage par les enfants dans le changement diachronique est toutefois fortement débattu.

Commençons par un type de prévision qui implique une langue dans sa totalité. Il s'agit de la prédiction très commune selon laquelle une langue minoritaire menacée dans un État où une langue est dominante disparaitra rapidement. Cette prédiction souvent n'est pas réalisée. Elle s'apparente le plus souvent à une projection démographique très vague concernant la diminution du nombre des locuteurs qui ont cette langue minoritaire comme langue maternelle. Car la réduction du système linguistique, au sens d'érosion ou de déclin linguistique, est bien plus importante pour le devenir d'une langue que la réduction du nombre de locuteurs. Dans ce domaine du déclin linguistique, DRESSLER (2011) a proposé de considérer le tarissement de la créativité linguistique comme un facteur prédictif prometteur. Il se manifeste dans l'incapacité à exprimer de nouveaux concepts à travers de nouveaux mots (néologismes) acceptés par la communauté. Par exemple en breton, au 19e siècle, non seulement les néologismes français comme *moissonneuse-batteuse* étaient traduits (breton *dorn-erezh*), mais de nouvelles formations comme *marc'h-houarn* 'bicyclette' (littéralement 'cheval de fer', cf. allemand *Drahtesel*) étaient créées. Après la première guerre mondiale, cependant, ce mouvement a cessé et les nouvelles formations bretonnes proposées n'ont plus été acceptées. Il faut souligner que le déclin linguistique tel qu'il est illustré par le breton ne peut être empêché que par des tentatives énergiques et réussies de revitalisation linguistique comme dans le cas de l'hébreu moderne.

5. Changement diachronique

En ce qui concerne la prévisibilité du changement linguistique dans une langue dont la vitalité est entière et n'est pas menacée de disparition ou de déclin, nous prendrons comme exemple de prévisibilité partielle le pluriel allemand des mots étrangers. WEGENER (2004) a constaté que la terminaison du pluriel des mots étrangers en allemand dans un premier temps est *-s*, puis est remplacée par des désinences plus courantes, par exemple *Ballon-s* 'ballons' devient *Ballon-e*. Cette observation permet de prévoir que les pluriels en *-s* des mots étrangers actuels seront dans le futur substitués par d'autres désinences. Mais, naturellement, elle ne prédit pas quand le changement se produira. Il est aussi possible d'appliquer cette prédiction aux mots étrangers qui ont été empruntés antérieurement. Ce faisant on ne réalise pas une prédiction au sens propre mais une rétrodiction (BARRETT & STANFORD 2006). Une telle rétrodiction s'applique à des cas comme en allemand *General* dont le pluriel a d'abord été *General-s*, puis est devenu *General-e* et enfin *Generäl-e*. Toutefois, dans notre corpus électronique (DRESSLER & MÖRTH 2012a ; MÖRTH & DRESSLER 2014) nous avons trouvé beaucoup d'exemples de pluriels en *-s* utilisés concurremment à d'autres pluriels dans un même mot dès la première attestation, par exemple au début du 19^{e} siècle *Pizza-s* et *Pizz-en*, *Scheich-s* 'cheikhs' et *Scheich-e.* On ne peut donc, en l'occurrence, prévoir qu'une tendance dont la probabilité dépend aussi d'autres facteurs.

De telles rétrodictions ne sont habituellement que probabilistes, et souvent aussi très faibles. Une rare exception est constituée par notre étude sur le développement de la 1ère personne du pluriel du présent des dialectes italoromans depuis leur formation jusqu'à aujourd'hui (SPINA & DRESSLER 2011), c'est-à-dire ceux des dialectes romans (le standard y compris) qui n'appartiennent pas à une autre langue romane d'Italie. En proto-italien, c'est-

à-dire les étapes intermédiaires entre le latin vulgaire et l'attestation la plus ancienne de l'italien, étapes reconstruites avec un haut degré de certitude, on peut partir des formes de 1ère personne du pluriel du présent des trois classes flexionnelles telles qu'elles sont illustrées par les verbes 'aimer', 'craindre', 'terminer' :

> indicatif : *-amo* (par ex. *amamo* 'nous aimons'), *-emo* (par ex. *tememo* 'nous craignons'), *-imo* (par ex. *finimo* 'nous terminons') ;
> subjonctif : *-emo* (*amemo*), *-iamo* (*temiamo*), *-iamo* (*finiamo*).

Dans plusieurs dialectes la distribution de ces six formes est restée la même jusqu'à aujourd'hui. Nous ne cherchons évidemment pas à prédire rétrodictivement quel changement s'est produit dans quel dialecte, et quand (le "problème d'actualisation" de LABOV 2001 : 466 ; 2014). Nous ne traitons pas non plus la question de savoir si la désinence flexionnelle *-mo* a changé et comment. Seul nous intéresse ici le destin des voyelles *-a-*, *-e-*, *-i-* et *-ia-* du radical. Nous ne pouvons donc pas prédire que dans certains dialectes *-e-* accentué s'est développé en *-i-* selon les lois phonétiques générales. Nous limitons le champ de recherche de notre rétrodiction à la distribution morphologique des voyelles du radical dans les classes flexionnelles, distribution qui est largement indépendante des changements des autres formes personnelles (même de la 2e personne du pluriel). Nous prédisons deux résultats pour le domaine de la 1ère personne du pluriel du présent : premièrement, les systèmes susceptibles de succéder au système proto-italien décrit ci-dessus, tels qu'ils sont possibles déductivement et les systèmes impossibles déductivement ; deuxièmement, la probabilité d'occurrence de ces systèmes possibles. Les deux prédictions ont été comparées au développement des dialectes italoromans actuels – lorsque les données disponibles le permettaient. Le champ de recherche est donc

suffisamment vaste pour que la réfutation des rétrodictions soit facilitée. En ce qui concerne l'examen de textes italiens plus anciens, le *Corpus testuale del Tesoro della lingua italiana delle origini* de l'Accademia della Crusca (Florence) a constitué un instrument très utile.

D'abord nous pouvons déduire du modèle de la morphologie naturelle, et spécialement de son application à l'histoire de la langue (DRESSLER 1997 ; 2002 ; KILANI-SCHOCH & DRESSLER 2005), quels changements concevables dans la distribution des formes du radical du proto-italien sont permis par la théorie et quels changements sont exclus. Ce champ de prédictions est ensuite restreint aux changements internes à la morphologie ; en d'autres termes, les changements qui trouvent leur origine dans la phonologie ou la syntaxe ainsi que les dialectes dans lesquels un changement syntaxique analogue à la substitution actuelle en français de *nous parlons* par *on parle* s'est produit, sont exclus de la recherche. Enfin, une prémisse supplémentaire pose que dans le développement typologique du latin à l'italien ainsi qu'à la plupart des langues romanes, la morphologie flexionnelle n'a connu qu'une réduction ou des échanges de formes, et aucun développement (voir, dans les langues romanes, premièrement la disparition des cas latins, du participe futur, du supin, du gérondif, du passif, deuxièmement de l'impératif, de l'infinitif passé, de l'imparfait et du parfait du subjonctif ainsi que la réduction d'autres catégories). Autrement dit il s'est produit une perte de complexité morphologique. Il s'agit donc seulement d'établir quelles formes parmi les six formes flexionnelles mentionnées plus haut se sont substituées aux autres et lesquelles ont disparu. La disparition la plus marquée est survenue en italien standard et dans les dialectes toscans de sa base qui ont remplacé toutes les autres formes par *-iamo*.

À partir du modèle de développement morphologique diachronique de la morphologie naturelle et de son

application aux dialectes italiens (sur une période de plus de mille ans), on peut prédire que 64 changements sont concevables. Parmi ceux-ci la théorie en exclut 52 comme étant impossibles et en retient 12. À notre connaissance les changements exclus par la théorie n'ont effectivement pas eu lieu. Parmi les 12 changements admissibles, seuls deux ne se sont pas produits ; ce sont des changements consécutifs à d'autres changements très rares, de nature vraisemblablement accidentelle parce que des aires dialectales de petite dimension tendent à ne pas se subdiviser en dialectes encore plus limités.

Comme la morphologie naturelle est une théorie des préférences, on peut prévoir quels changements parmi les changements admissibles sont les plus vraisemblables. Ces rétrodictions doivent être compatibles avec le nombre relatif de dialectes différents qui ont connu le même changement. C'est-à-dire que plus un changement déterminé est préféré (dans la dérivation déductive à partir de la théorie), plus son occurrence dans les dialectes italoromans doit être fréquente. De fait, six des changements documentés correspondent à des aires plus étendues et souvent discontinues, tandis que quatre d'entre eux sont limités à des aires très restreintes. Cette contradiction apparente est compatible avec les hypothèses théoriques, dans la mesure où le nombre exact de dialectes qui manifestent un changement particulier/déterminé ne peut être prédit.

Un tel exemple de forte prévisibilité du changement linguistique historique est exceptionnel. Il est le résultat des contraintes très restrictives imposées aux prémisses et aux conditions des rétrodictions. Généralement la prévisibilité des changements diachroniques est plus faible et seulement partielle.

Néanmoins, la prévisibilité en linguistique reste supérieure à la prévisibilité dans les sciences historiques, car ni le développement de l'enfant dans le cours de l'acquisition du

langage, ni le changement diachronique des langues ne connaissent des latitudes de variation comparables à celles des changements historiques. Et dans les deux domaines linguistiques, les intentions des personnes et des groupes jouent un rôle beaucoup plus limité.

De même que les autres sciences cognitives, sociales et culturelles, la linguistique a donc affaire à des phénomènes complexes, comme on l'a vu également avec le développement du langage. Ces phénomènes ne sont que partiellement saisissables du point de vue quantitatif et, en raison du nombre des facteurs, leur prévisibilité ne peut correspondre qu'à des degrés variables de probabilité.

L'ample travail de vérification des phénomènes langagiers qui ont fait l'objet de prévisions est opéré au moyen de l'analyse informatique de grands corpus électroniques. À nos yeux cette entreprise constitue le fondement de la linguistique de corpus. En ce qui concerne les méthodes, celles qui se basent sur des modèles probabilistes se sont largement imposées en linguistique informatique au cours des années. Elles s'appliquent aussi bien aux exigences de base de la linguistique de corpus, telle la lemmatisation automatique (l'attribution/assignation de formes de mots comme *Haus* 'maison', *Hauses, Häuser* au même lemme *Haus*) qu'à l'attribution automatique de classes de mots à l'ensemble des lemmes (par ex. nom à *Haus*, adjectif à *häuslich*, verbe à *hausen*). Dans toutes les procédures d'analyse qui concernent les textes, celles-ci représentent le niveau de base. Leur large diffusion tient principalement au fait que l'application de méthodes statistiques à de nouvelles données apparait comme plus robuste et plus efficace que d'autres méthodes. Ainsi, par exemple l'adaptation d'outils de la linguistique de corpus à de nouvelles langues au moyen d'une approche stochastique est-elle dans de nombreux cas réalisée beaucoup plus rapidement et avec des résultats aussi bons sinon meilleurs que ceux qui étaient obtenus avec une

approche basée sur des règles. Les modèles de Markov cachés font partie des algorithmes le plus souvent utilisés en linguistique informatique aujourd'hui (CARSTENSEN *et al.* 2009).

Références

BARRETT Jeff & STANFORD P. Kyle. (2006). Prediction. In PFEIFER Jessica & SARKAR Sahotra (Eds), *The Philosophy of Science: An Encyclopedia*, New York: Routledge, 585-599.

BARTKE Susanne & SIEGMÜLLER Julia (Eds), (2004). *Williams Syndrome across Languages*. Amsterdam/Philadelphia: John Benjamins.

BAVIN EDITH LAURA (Ed.) (2009). *The Cambridge Handbook of Child Language*. New York: Cambridge University Press.

CAMERON-FAULKNER Thea, LIEVEN Elena & TOMASELLO Michael. (2003). A Construction Based Analysis of Child Directed Speech. *Cognitive Science 27-6*, 843-873.

CARSTENSEN Kai-Uwe, EBERT Christian, EBERT Cornelia, JEKAT Susanne, KLABUNDE Ralf & LANGER Hagen (Eds), (2009). *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Dordrecht: Springer Verlag.

CHOMSKY Noam. (1957). *Syntactic Structures*. The Hague: Mouton.

DRESSLER Wolfgang U. (1997). "Scenario" as a Concept for the Functional Explanation of Language Change. In GVOZDANOVIĆ Jadranka (Ed.), *Language Change and Functional Explanations*, Berlin: Mouton de Gruyter, 109-142.

DRESSLER Wolfgang U. (2002). Naturalness and Morphological Change. In JOSEPH Brian D. & JANDA Richard D. (Eds), *The Handbook of Historical Linguistics*, Oxford: Blackwell, 461-471.

DRESSLER Wolfgang U. (2010). A Typological Approach to First Language Acquisition. In KAIL Michèle & HICKMANN Maya (Eds), *Language Acquisition Across Linguistic and Cognitive Systems*, Amsterdam: Benjamins, 109-124.

DRESSLER Wolfgang U. & MÖRTH Karlheinz. (2012a). Vom Einfluss der Pragmatik auf die Grammatik, insbesondere in der Entwicklung der Pluralbildung. Eine corpusbasierte Untersuchung. *Historische Pragmatik. Jahrbuch für Germanistische Sprachgeschichte 3-1*, 75-93.

DRESSLER Wolfgang U. & MÖRTH Karlheinz. (2012b). Produktive und weniger produktive Komposition in ihrer Rolle im Text an Hand der Beziehungen zwischen Titel und Text. In GAETA Livio & SCHLÜCKER Barbara. (Eds), *Das Deutsche als kompositionsfreudige Sprache*, Berlin: de Gruyter, 219-233.

DRESSLER Wolfgang U., LIBBEN Gary & KORECKY-KRÖLL Katharina. (2014). Conflicting vs. Convergent vs. Interdependent Motivations in Morphology. In MACWHINNEY Brian, MALCHUKOV Andrej & MORAVCSIK Edith (Eds), *Competing Motivations in Grammar and Usage*, Oxford: Oxford University Press, 181-196.

HART Betty & RISLEY Todd R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore: Paul H. Brookes.

KILANI-SCHOCH Marianne & DRESSLER Wolfgang U. (2005). *Morphologie Naturelle et Flexion du Verbe Français*. Tübingen: Gunter Narr.

KILANI-SCHOCH Marianne, BALČIUNIENE Ingrida, KORECKY-KRÖLL Katharina, LAAHA Sabine & DRESSLER Wolfgang U. (2009). On the Role of Pragmatics in Child-Directed Speech for the Acquisition of Verb Morphology. *Journal of Pragmatics 41*, 219-239.

KÖPCKE Klaus-Michael. (1993). *Schemata bei der Pluralbildung im Deutschen: Versuch einer kognitiven Morphologie*. Tübingen: Narr.

KORECKY-KRÖLL Katharina (2011). *Der Erwerb der Nominalmorphologie bei zwei Wiener Kindern: Eine Untersuchung im Rahmen der Natürlichkeitstheorie*. Wien: Universität Wien. Dissertation.

KORECKY-KRÖLL Katharina & DRESSLER Wolfgang U. (2015). The Acquisition of Case in German: A Longitudinal Study of Two Viennese Children. *Studi e Saggi Linguistici 53-1*, 9-36.

KORECKY-KRÖLL Katharina, UZUNKAYA-SHARMA Kumru, CZINGLAR Christine & DRESSLER Wolfgang U. (2015). Das INPUT-Projekt: Herausforderungen auf dem Weg zum Bildungserfolg von ein- und zweisprachigen Wiener Kindergartenkindern. In ANREITER Peter, MAIRHOFER Elisabeth & POSCH Claudia (Eds), *ARGUMENTA. Festschrift für Manfred Kienpointner zum 60. Geburtstag*, Wien: Praesens, 201-213.

KORECKY-KRÖLL Katharina & DRESSLER Wolfgang U. (à paraitre). (Mor)phonotactics in High vs. Low SES Children. In DZIUBALSKA-KOŁACZYK Katarzyna & WECKWERTH Jaroslaw (Eds), *Volume in Memoriam of Rajendra Singh*, Poznań: Adam Mickiewicz University Press.

KORECKY-KRÖLL Katharina. (soumis). Kodierung und Analyse mit CHILDES: Erfahrungen mit kindersprachlichen Spontansprachkorpora und erste Arbeiten zu einem rein erwachsenensprachlichen Spontansprachkorpus. In RESCH Claudia & DRESSLER Wolfgang U. (Eds), *Korpusbasierte Linguistik*, Wien: Verlag der Österreichischen Akademie der Wissenschaften.

LAAHA Sabine & KORECKY-KRÖLL Katharina. (à paraitre). Verschriftung, Kodierung und Analyse von Kindersprache mit CHILDES. In WANDL-VOGT Eveline & KORECKY-KRÖLL Katharina (Eds), *Transkriptionssysteme im Vergleich: Sprache - Ton - Bild. Kodierung gesprochener Sprache*, Wien: Praesens.

LABOV William. (2001). *Principles of Linguistic Change: Social Factors*. Oxford: Blackwell.

LABOV William. (2014). The Sociophonetic Orientation of the Language Learner. In CELATA Chiara & CALAMAI Silvia. (Eds), *Advances in Sociophonetics*, Amsterdam: John Benjamins, 17-29.

MACWHINNEY Brian. (2000). *The CHILDES Project: Tools for Analyzing Talk. Volume 1: Transcription Format and Programs*. Mahwah: Lawrence Erlbaum.

MÖRTH Karlheinz & DRESSLER Wolfgang U. (2014). German Plural Doublets with and without Meaning Differentiation. In RAINER Franz, DRESSLER Wolfgang U., GARDANI Francesco & LUSCHÜTZKY, Hans C. (Eds), *Morphology and Meaning*, Amsterdam: John Benjamins, 249-258.

OEVERMANN Ulrich. (1972). *Sprache und soziale Herkunft. Ein Beitrag zur Analyse schichtenspezifischer Sozialisationsprozesse und ihrer Bedeutung für den Schulerfolg*. Frankfurt am Main: Suhrkamp.

RAVID Dorit, DRESSLER Wolfgang U., NIR-SAGIV Bracha, KORECKY-KRÖLL Katharina, SOUMAN Agnita, REHFELDT Katja, LAAHA Sabine, BERTL Johannes, BASBØLL Hans & GILLIS Steven. (2008). Core morphology in child directed speech: Crosslinguistic corpus analyses of noun plurals. In BEHRENS Heike (Ed.), *Corpora in Language Acquisition Research: History, Methods, Perspectives*, Amsterdam/Philadelphia: John Benjamins, 25-60.

SLOBIN Dan Isaac (Ed.) (1985a). *The Crosslinguistic Study of Language Acquisition. Volume 1.* Hillsdale: Lawrence Erlbaum.

SLOBIN Dan Isaac (Ed.) (1985b). *The Crosslinguistic Study of Language Acquisition. Volume 2.* Hillsdale: Lawrence Erlbaum.

SLOBIN Dan Isaac (Ed.) (1992). *The Crosslinguistic Study of Language Acquisition. Volume 3*. Hillsdale: Lawrence Erlbaum.

SLOBIN Dan Isaac (Ed.) (1997a). *The Crosslinguistic Study of Language Acquisition. Volume 4*. Mahwah: Lawrence Erlbaum.

SLOBIN Dan Isaac (Ed.) (1997b). *The Crosslinguistic Study of Language Acquisition. Volume 5*. Mahwah: Lawrence Erlbaum.

SPINA Rossella & DRESSLER Wolfgang U. (2011). How far Can Diachronic Change be Predicted: the Case of Italo-Romance First Person Plural Present Indicative. *Diachronica 28*, 499-544.

WEGENER Heide. (2004). Pizzas und Pizzen, die Pluralformen (un)assimilierter Fremdwörter im Deutschen. *Zeitschrift für Sprachwissenschaft 23 - 1*, 47-112.

WEISLEDER Adriana & FERNALD Anne. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science 24-11*, 2143 –2152.

XANTHOS Aris, LAAHA Sabine, GILLIS Steven, STEPHANY Ursula, AKSU-KOÇ Ayhan, CHRISTOFIDOU Anastasia, GAGARINA Natalia, HRZICA Gordana, KETREZ F. Nihan, KILANI-SCHOCH Marianne, KORECKY-KRÖLL Katharina, KOVAČEVIĆ Melita, LAALO Klaus, PALMOVIC Marijan, PFEILER Barbara, VOEIKOVA Maria D. & DRESSLER Wolfgang U. (2011). On the Role of Morphological Richness in the Early Development of Noun and Verb Inflection. *First Language 31-4*, 461-479.