**Cahiers de praxématique**

**54-55 | 2010**
**Corpus, données, modèles**

# Le corpus PhoDiFLE : un corpus commun de français langue étrangère pour une étude phonétique des productions de locuteurs de langues maternelles plurielles

*The PhoDiFLE Corpus : a Joint Corpus of French Speech, for a Phonetic Study of L1-L2 Contrasts*

**Simon Landron, Nikola Paillereau, Ahmad Nawafleh, Christelle Exare, Hirofumi Ando et Jiayin Gao**

**Édition électronique**
URL : http://journals.openedition.org/praxematique/1119
DOI : 10.4000/praxematique.1119
ISSN : 2111-5044

**Éditeur**
Presses universitaires de la Méditerranée

**Édition imprimée**
Date de publication : 1 janvier 2010
Pagination : 73-86
ISBN : 978-2-36781-012-6
ISSN : 0765-4944

**Référence électronique**
Simon Landron, Nikola Paillereau, Ahmad Nawafleh, Christelle Exare, Hirofumi Ando et Jiayin Gao, « Le corpus PhoDiFLE : un corpus commun de français langue étrangère pour une étude phonétique des productions de locuteurs de langues maternelles plurielles », *Cahiers de praxématique* [En ligne], 54-55 | 2010, document 4, mis en ligne le 01 janvier 2013, consulté le 08 septembre 2020. URL : http://journals.openedition.org/praxematique/1119 ; DOI : https://doi.org/10.4000/praxematique.1119

*Cahiers de praxématique 54-55, 2010, 73-86*

Simon Landron, Nikola Paillereau, Ahmad Nawafleh, Christelle Exare, Hirofumi Ando & Jiayin Gao
Laboratoire de phonétique et de phonologie (L.P.P.), université de Paris 3, Sorbonne nouvelle

# Le corpus PhoDiFLE [1] : un corpus commun de français langue étrangère pour une étude phonétique des productions de locuteurs de langues maternelles plurielles

## 1. Le Projet

Nous nous intéressons à la création d'un corpus commun pour l'enseignement des sons du français (les voyelles et les consonnes). Les cibles acoustiques des phonèmes et les effets de la coarticulation sont évalués à partir de la production de natifs du français et les résultats des mesures acoustiques sont comparés avec ceux des apprenants de diverses langues. Trois types de tâches sont proposées : 1) prononciation des voyelles isolées (cibles) et dans des logatomes (effet de la coarticulation), 2) de la lecture, et 3) de la parole spontanée. Ces trois tâches sont analysées acoustiquement et perceptivement.

Après avoir défini l'objet de notre travail, nos objectifs et nos besoins, nous présenterons le corpus PhoDiFLE et ses particularités : les objectifs phonétique et didactique et la démarche envisagée pour ce corpus avec une approche mixte, guidée et fondée sur corpus. Nous décrirons finalement la méthodologie utilisée, pour terminer avec les perspectives offertes par notre travail.

### 1.1. Genèse du corpus : objet, objectifs et besoins

Le terme « corpus » étant polysémique, nous l'utilisons ici pour désigner, d'une part, l'ensemble des tâches soumises à des locuteurs dont les réalisations sont enregistrées et, d'autre part, les données recueillies,

1. Phonétique didactique du français langue étrangère.

autrement appelées « base de données ». En phonétique, cet ensemble de données acoustiques peut être couplé avec d'autres informations, comme des données aérodynamiques, physiologiques, articulatoires, visuelles et des questionnaires.

Jeunes chercheurs en phonétique didactique et enseignants de français langue étrangère (F.L.E.), nous nous sommes associés en 2009 au sein du Laboratoire de phonétique et phonologie (université Paris 3). Le groupe compte aujourd'hui 11 membres et travaille sur des apprenants d'une dizaine de langues d'origine. Au départ réunis par des préoccupations similaires concernant nos recherches respectives, nous nous sommes donné :

a) comme objet d'étude : les réalisations phonétiques du français oral (F.L.E.) produites par des apprenants de langues maternelles différentes,

b) comme objectif primaire : l'inventaire des écarts phonétiques produits par des apprenants du F.L.E. par rapport à la norme dans le cadre d'une étude à visée essentiellement didactique, qui analyse et quantifie les difficultés des apprenants de langues maternelles différentes,

c) comme objectif secondaire : une contribution tangible à la recherche en didactique des langues, via la création d'une base de données phonétiques avec toutes les consonnes et les voyelles du français, étiquetée et annotée, en libre accès et qui donnera lieu à des propositions d'exercices.

Tout corpus répond au départ à un besoin. Notre recherche concernant la phonétique et la didactique des langues, nous avons d'abord regardé ce qui existait en la matière. Parmi les corpus récents, les objectifs les plus proches des nôtres sont ceux du corpus *I.P.F.C.* (Detey *et al.* 2008, Racine *et al.* 2010) car il vise l'étude des caractéristiques phonétiques de la parole en français L2. Mais le corpus *I.P.F.C.* a été construit en vue d'analyser des unités supérieures telles que la liaison et l'intonation. Cette approche ne prévoit pas d'analyses fines à partir de valeurs formantiques (mais celles-ci pourront être faites par la suite, et les résultats pourront être comparés aux nôtres), pas d'exemplaires de voyelles prononcées isolément, ni des données suffisantes pour caractériser les habitudes de coarticulation des locuteurs par exemple.

Nous avons ainsi quatre objectifs principaux :

- un corpus de parole, propre à une étude phonétique systématique pour comparer (qualitativement et quantitativement) des réalisations de locuteurs,
- un corpus qui soit « commun » : pour une seule langue étrangère (F.L.E.), des locuteurs de langues maternelles différentes,
- un corpus qui aboutirait à une base de données accessible en ligne,
- un corpus qui permettrait de construire des exercices (remédiation et évaluation).

Le corpus PhoDiFLE a été initié afin de répondre à ces besoins.

## 2. Particularités du corpus PhoDiFLE

Le corpus PhoDiFLE concerne les domaines de la Phonétique et de la Didactique du Français Langue Étrangère. Sa construction s'est organisée autour de ces deux domaines et permet d'explorer la réalisation de l'ensemble des phonèmes du français dans différents contextes (approche guidée par corpus) ou de tester des hypothèses (approche fondée sur corpus).

### 2.1. Un corpus pour la phonétique

La phonétique s'intéresse aux sons de la parole. Elle est la branche qui se préoccupe de la face sonore du langage. Elle appuie ses conclusions à la fois sur une analyse détaillée de chaque production et sur l'analyse statistique de vastes bases de données. Pour notre étude, la tâche demandée aux locuteurs doit cibler en priorité les sons de la langue.

Si la réalisation d'un phonème est sujette à une grande variabilité (intra-locuteur, inter-locuteur, intra-langue), les sources de variabilité sont à ce jour pour la plupart bien analysées (Gendrot, 2006, entre autres) et il est possible de définir des contextes qui contrôlent les facteurs de variation. Ceci peut être contrôlé dans la création des tâches. Notre corpus vise à l'étude de deux sources de variabilité : l'influence du contexte phonétique immédiat (coarticulation) et de la position du phonème dans le mot.

L'informatique permet l'étude systématique de très grandes quantités de réalisations d'un même phonème. Elle permet d'avoir une

représentation visuelle des segments via le spectrogramme, d'effectuer des mesures semi-automatiques (avec vérification manuelle) et reproductibles, et offre la possibilité de dégager des tendances générales. L'étude des réalisations des phonèmes ne se place donc plus dans une approche par transcription en API [1], globalisante et subjective. Au contraire, les productions des apprenants sont situées par rapport à des données chiffrées, et des variations visualisables sur des graphiques.

### 2.2. Un corpus pour la didactique

La vérification et la quantification des écarts observés entre la norme et la production réelle des apprenants doivent permettre de rendre compte des difficultés spécifiques de locuteurs d'une langue d'origine donnée. En effet, la langue d'origine (L1) des apprenants introduit un crible phonologique (Troubestkoy, 1949) conduisant à des productions de sons dans la langue cible influencés par cette L1, et en conséquence potentiellement différents selon l'origine des apprenants. Nous envisageons ainsi de définir un modèle des écarts typiques en fonction de la L1 des locuteurs. Il existe déjà des typologies d'écarts de prononciation d'apprenants par nationalité (Lauret, 2007), mais nous envisageons une description plus précise et quantifiée.

La notion de « norme » du français en didactique du F.L.E. est problématique. Face à la grande variabilité du français qui existe de fait entre les locuteurs, mais aussi entre les communautés francophones, les questions d'orthoépie deviennent relatives. L'intercompréhension entre les locuteurs pourrait être la seule vraie garante de ce que l'on peut appeler « les bonnes prononciations ». Pourquoi, dès lors, nous intéresser à la question fine des écarts par rapport à une norme en F.L.E. ?

Si la réponse était que l'enseignant de prononciation de F.L.E. ne doit pas s'intéresser à la langue telle qu'elle est parlée, mais telle qu'elle est la mieux comprise (Lauret, 2007), il nous faudrait cependant toujours définir cette norme de compréhension puis en enseigner la prononciation. Notre travail de phonéticien doit s'intéresser à un moment donné à la prononciation des Français et en dégager ses normes.

1. Alphabet phonétique international.

Tout d'abord, comme l'analyse de productions réelles en français est nécessaire pour identifier les particularités de la langue et pour pouvoir comparer avec une référence, nous avons choisi d'enregistrer des locuteurs dont la façon de parler ne peut pas indiquer qu'ils viennent d'une autre région que de la région parisienne. Le critère déterminant ce choix a d'abord été choisi pour des raisons pratiques : notre laboratoire se situe à Paris.

Ensuite, en tant que didacticiens, nous devons viser la production du français natif pour nos apprenants. Même si, avec l'enjeu identitaire, certains apprenants ne veulent pas perdre leur accent d'origine — et dans un tel cas, ils ne modifieront sans doute pas leur façon de parler — d'autres veulent acquérir une façon de parler non marquée (Pillot-Loiseau *et al.*, 2010) et le devoir de l'enseignant de F.L.E. est de les y aider (Lauret, 2007).

Au plan pédagogique, ce corpus permet, par une comparaison des données des deux groupes de locuteurs, les natifs et les apprenants, de représenter visuellement le décalage entre ces réalisations. De plus, grâce à l'analyse des réalisations des apprenants pour les voyelles isolées, en contexte et en parole spontanée, nous sommes en mesure de vérifier la capacité des apprenants à reproduire les phonèmes comme les natifs, que ce soit avec un effort conscient sur des sons isolés ou en parole continue.

## 2.3. Contenu des tâches

À partir des mesures acoustiques (des formants vocaliques, de la durée, du VOT [1] des occlusives, des transitions formantiques, de la limite inférieure du bruit de friction des fricatives, etc.) nous décrivons les voyelles et les consonnes réalisées de façon isolée et/ou dans des contextes phonétiques différents (coarticulation) au travers des tâches de parole contrôlée (productions de sons de façon isolée ou dans des logatomes) et des tâches de parole lue et de parole spontanée.

Ainsi, les différentes tâches se répartissent entre de la lecture 1) des voyelles isolées, 2) de logatomes, 3) d'un texte et de phrases contenant tous les phonèmes du français ainsi que 4) de la production orale spontanée.

1. *Voice Onset Time*, le temps écoulé entre le relâchement d'une occlusive et le début de voisement.

Il est parfois difficile pour les locuteurs d'associer correctement les sons à leur graphème, notamment pour les voyelles (Delattre, 1945). Ainsi, les voyelles moyennes présentées aux locuteurs peuvent être prononcées avec un timbre fermé ou ouvert. Nous avons par conséquent inséré des images présentant des mots contenant la voyelle en question accompagnées de leur graphie sur les diapositives présentées aux sujets, d'abord dans une phase d'entraînement, puis sur ces mêmes diapositives pendant l'enregistrement. Ce protocole facilite la tâche d'identification du son et permet de prévenir des confusions dans la réalisation des phonèmes.

Les cibles acoustiques des voyelles sont étudiées lorsque les voyelles sont prononcées isolément, et ne sont pas influencées par des phonèmes environnants. Les voyelles isolées sont les treize voyelles du français (Vaissière, 2006), c'est-à-dire les dix voyelles orales /a, i, u, y, e, ɛ, o, ɔ, ø, œ/ et les trois voyelles nasales /ã, ɛ̃, õ/ placées dans une phrase cadre du type : *« bébé, il a dit < é > comme dans bébé »*. Le choix de placer le mot contenant la voyelle au début de la phrase cadre permet au locuteur d'identifier la voyelle cible avant de la produire isolément (entre deux courtes pauses).

La coarticulation et l'influence de la position du phonème dans le mot sont étudiées dans des logatomes. Les logatomes permettent d'analyser et de décrire l'effet du contexte phonétique immédiat (coarticulation) et des positions prosodiques, sur les propriétés acoustiques des voyelles et des consonnes. Ils sont de type CVCVCVC, placés dans la phrase cadre *« le mot CVCVCVC peut bien coller »*, pour obtenir la configuration des segments dans trois positions différentes (début, milieu et fin de mot). Les treize voyelles du français sont placées dans cinq contextes consonantiques (/p, t, k, r, m/) représentant quatre lieux d'articulation et le mode nasal. Les dix-sept consonnes du français sont présentées adjacentes aux trois voyelles extrêmes (/i, a, u/).

Le texte, un conte pour enfants, et les phrases choisis se composent de tous les phonèmes du français, ce qui permet de les analyser dans la parole lue.

Le dernier volet de notre corpus est une série d'enregistrements de parole spontanée. Chaque locuteur est invité à parler pendant dix minutes de son apprentissage d'une langue étrangère, de ses études, de ses loisirs ou de son travail.

### 2.4. Une approche mixte : guidée par corpus (*corpus-driven*) et fondée sur corpus (*corpus-based*)

Notre travail vise à identifier les écarts de prononciation entre un groupe d'apprenants et des natifs. Notre première approche consiste donc à explorer des données sans *a priori*, sans hypothèse précise à vérifier. Il s'agit donc d'une approche de type « guidée par corpus » (Williams, 2005). Ces premières observations permettent de formuler des hypothèses plus précises mais uniquement parmi les éléments ciblés lors de la construction du corpus. Le corpus a été créé de façon à ce que les hypothèses formulées à partir des écarts observés puissent être précisément validées ou non (et les écarts, quantifiés). Cette approche est ainsi également de type « basée sur corpus ».

Une démarche similaire est utilisée sur les données récoltées de type « parole spontanée ». Si les résultats sont moins ciblés, nous y cherchons les mêmes éléments que dans la parole contrôlée. Ainsi, ces données peuvent servir soit à identifier des écarts entre les réalisations d'apprenants et de natifs soit à tester ces écarts qui ont été constatés sur de la parole spontanée (avec le risque de ne pas avoir de données suffisantes) ou sur de la parole contrôlée pour comparaison.

Ce corpus volumineux présente donc la possibilité, sinon l'obligation, de travailler sur des sous-corpus, définis par la langue d'origine des apprenants étudiés, et définis par l'élément précis de la langue étudiée. Des comparaisons entre des groupes d'apprenants sont envisageables.

## 3. Méthodologie : pour un corpus homogène et représentatif

Face à une problématique de travail en commun, il nous a fallu négocier les éléments de ce corpus, définir les critères de comparabilité, ce qui se joue finalement autour des concepts d'homogénéité et de représentativité. En effet, l'homogénéité concerne la définition des paramètres qui doivent être les mêmes pour chaque locuteur ou groupes de locuteurs (par langue d'origine) et qui permet, tous paramètres égaux par ailleurs, de définir une variabilité entre les locuteurs ou groupes de locuteurs. La représentativité concerne les éléments comparés : en effet l'objectif est de comparer des communautés, et non simplement quelques locuteurs entre eux.

Ainsi, en premier lieu, nous avons tenté de définir l'homogénéité de notre corpus à deux niveaux. Tout d'abord l'homogénéité se joue au niveau des conditions d'enregistrement et des prises de données. Il s'agit ici des conditions de collecte des données. Ensuite l'homogénéité se joue au niveau même de l'élaboration des tâches qui doivent avoir leur cohérence. Il ne s'agit cependant plus là seulement de comparabilité, mais aussi de définition même de l'objet d'étude. Il doit être le même pour tous.

Enfin, la construction du corpus doit également prendre en compte les paramètres de représentativité afin que les productions des locuteurs reflètent l'ensemble de leurs productions et qu'à partir de quelques locuteurs puissent être définies des règles pour un ensemble plus vaste.

### 3.1. Conditions d'enregistrement et traitement des données

Tout d'abord, le corpus est homogène parce que nous utilisons le même protocole : nous enregistrons toujours les mêmes tâches (moins directives pour la parole spontanée) avec la même consigne, le même matériel, les mêmes paramètres d'enregistrement, dans des conditions similaires pour tous les locuteurs et avec les mêmes procédures de traitement et d'analyse.

Les enregistrements sont effectués dans une pièce tranquille pour assurer une bonne qualité sonore, à l'aide d'un microphone serre-tête AKG C 520L. Nous utilisons un microphone serre-tête pour assurer que la distance entre le microphone et la source sonore reste constante. Nous utilisons finalement des paramètres d'enregistrement identiques (une fréquence d'échantillonnage à 44 100 Hz et un taux d'échantillonnage à 16 bits). Le son est enregistré au format numérique (*.wav*). La qualité de l'enregistrement est vérifiée par spectrogramme.

Après la transcription — le codage orthographique des formes entendues — nous procédons à l'alignement semi-automatique des données, c'est-à-dire à la segmentation en mots et en segments. Ceci est effectué sous *Praat* (Boersma P. et Weenink D., 2011) avec des scripts écrits par les membres du laboratoire L.P.P. [1], et EasyAlign, outil d'alignement automatique (Goldman, 2011). En raison de problèmes de

1. Laboratoire de phonétique et phonologie de l'université de Paris 3, Sorbonne nouvelle.

paramétrages du logiciel liés aux écarts de prononciation, les données des apprenants nécessitent un traitement manuel important.

Nous intéressant à la production de phonèmes du français, la transcription des phones est effectuée à partir du phonème cible, et non de ce qui est perçu. Les variations de production sont ensuite étudiées dans nos analyses. Une autre ligne d'annotation est ajoutée avec les mots tels qu'ils devraient être produits, notamment afin d'identifier d'éventuels ajouts ou absences de phonèmes.

### 3.2. Les tâches et les locuteurs

Il faut distinguer un second niveau d'homogénéité. L'élaboration des tâches et la définition des locuteurs sont également essentielles. En parallèle, il faut prendre en compte les éléments qui permettent de rendre ce corpus représentatif de son objet d'étude. Cela détermine le contenu, et la taille de celui-ci. Il a donc fallu réfléchir aux éléments suivants :

- la quantité d'items pour la taille des tâches de lecture et leur qualité,
- la quantité de répétitions de chaque item,
- la quantité de locuteurs et leur qualité,
- la quantité d'origines différentes des locuteurs et leur qualité.

### 3.3. Viser l'homogénéité : quantité/qualité des items et quantité/qualité d'origines linguistiques étrangères

Le contenu des tâches et la définition des locuteurs constituent des repères de comparabilité. Les éléments qui intéressent notre étude doivent être les seuls à diverger afin de pouvoir être comparés.

Couvrir la prononciation de l'ensemble des phonèmes du français suppose une variété d'éléments étudiés, facteurs d'hétérogénéité, analysables au travers de sous-corpus. Ils constituent les éléments à étudier. Même si chaque chercheur n'est pas amené à étudier précisément les mêmes éléments que les autres, l'étude commune reste une possibilité offerte car tous les locuteurs réalisent les mêmes tâches.

La durée moyenne d'enregistrement de ce corpus est de l'ordre d'une heure par locuteur. Il a été nécessaire de limiter les différents éléments car un long enregistrement génère des phénomènes de fatigue, ou de rejet.

La quantité de nationalités différentes n'a pas été définie *a priori*, car notre volonté est de rassembler des données sur la plus grande quantité de langues possibles. Les langues constituent ainsi un aspect de l'hétérogénéité du corpus, mais cette variété est également ce que nous voulons étudier. Ce corpus est en fait constitué d'un corpus de référence : les productions des locuteurs francophones, et d'une multitude de corpus d'étude : les productions des apprenants par langue et l'élément de la langue étudié. À l'heure actuelle, nous avons essentiellement enregistré des apprenants d'origine tchèque, japonaise, taïwanaise et bosnienne.

Néanmoins, il est nécessaire que les locuteurs d'une langue donnée gardent une certaine homogénéité. Ainsi, pour les locuteurs natifs du français, afin d'avoir une référence du français contemporain tout en gardant certains phonèmes que les jeunes sont en train de perdre (par exemple l'opposition entre /e/ et /ɛ/ dans certaines paires minimales) nous avons décidé d'enregistrer des locuteurs francophones appartenant à la tranche d'âge 18-40 ans, et avec un accent de type « parisien ». Au plan socioprofessionnel, le niveau du baccalauréat est requis.

Pour les langues d'origine des apprenants, selon sa situation, le chercheur doit systématiquement déterminer l'homogénéité de son groupe (langue maternelle, langue de communication...). Afin de dissocier le facteur du niveau de français sur les difficultés de prononciation, nous avons choisi d'enregistrer des apprenants de niveau intermédiaire-avancé qui sont capables de lire le corpus proposé et de produire de la parole spontanée en français. L'analyse de leurs productions permet du reste de déterminer leur niveau de maîtrise des sons du français.

Ces critères sont complétés par deux questionnaires, renseignés respectivement par les locuteurs natifs et par les apprenants, afin de mieux connaître leur profil individuel, de pouvoir contrôler les facteurs impliqués et pour faciliter l'interprétation de la variabilité acoustique observée. Quant aux francophones natifs, nous leur avons notamment demandé leur biographie linguistique pour vérifier d'éventuelles influences sur leur accent actuel.

### 3.4. Viser la représentativité : quantité de répétitions et quantité de locuteurs

La quantité de répétitions d'items relève de la représentativité. En effet, pour avoir une idée de la façon dont un locuteur prononce réellement un son, il faut en avoir une certaine quantité de réalisations. Quatre répétitions permettent de représenter assez fidèlement la prononciation habituelle d'un sujet. Malheureusement, cela augmente la durée totale de l'enregistrement et constitue pour ce corpus la limite de répétitions que l'on peut exiger d'un locuteur.

La quantité de locuteurs définit également la représentativité du corpus. Ainsi, un minimum de trente locuteurs par sexe seront enregistrés (quarante femmes et une dizaine d'hommes ont été enregistrés à ce jour). Les locuteurs natifs servent de référence commune pour le corpus. En ce qui concerne les apprenants (décrits précédemment), chaque chercheur doit enregistrer dix locuteurs (hommes ou femmes séparément) de la langue étudiée. Jusqu'à présent, nous avons enregistré quarante-quatre apprenants (Tchèques, Japonais, Taïwanais et Bosniens essentiellement). Ces locuteurs sont ainsi représentatifs de la langue définie par les critères d'homogénéité.

## 4. Perspectives

Nous proposons quatre pistes pour la poursuite de cette recherche.

### 4.1. Formalisation : une analyse contrastive et empirique pour chaque langue d'origine

Pour formaliser les difficultés phonétiques que peuvent rencontrer les apprenants du français, nous fournissons une analyse contrastive, non pas *a priori*, comme une comparaison des inventaires phonologiques des langues en présence, mais *a posteriori*, basée sur des données phonétiques et sur leur analyse.

### 4.2. Représentation des écarts : en production et en perception

L'étude doit aboutir à une représentation des écarts non seulement pour la production, mais également pour la perception. En effet, si

la première étape du travail consiste en des tâches de production, des tests de perception informent sur 1) la perception par des apprenants des sons articulés par des natifs et 2) la perception par des natifs des sons produits par des non-natifs (évaluation de l'intelligibilité et de l'exactitude). Ainsi, ce travail doit permettre de corréler les écarts constatés à une réalité perceptive des natifs français.

### 4.3. Remédiation : des exercices à pratiquer

La conséquence directe sera la réalisation d'exercices ciblés sur les difficultés avérées. Notons cependant que les exercices ne viseront pas nécessairement la réalisation moyenne des locuteurs de référence, mais plutôt une réalisation qui soit la mieux comprise par ces locuteurs de référence.

### 4.4. Évaluation : une liste de mots clés pour tester un niveau

Le C.E.C.R. [1] (Division des Politiques Linguistiques, 2001) propose une grille d'évaluation commune pour attester du niveau des apprenants dans une langue étrangère. On notera que les critères restent flous en ce qui concerne la compétence phonétique. Nous proposerons une liste de mots clés, avec pour chacun, une cible précise à évaluer.

## Conclusion

La construction d'un corpus est une tâche complexe. Lorsqu'il s'agit de construire un corpus pour un groupe de chercheurs, ces questions prennent un enjeu supplémentaire relatif à la défense des intérêts de recherche de chacun. Sachant qu'il est déjà difficile de définir l'homogénéité d'un corpus pour la recherche d'un chercheur, lorsque cela concerne la recherche de plusieurs chercheurs, les problèmes se multiplient. Néanmoins, les gains en sont proportionnels : il s'agit tout d'abord d'un travail formateur qui permet à la fois de se poser des questions sur ce que l'on fait et qui permet aussi d'avoir une perspective plus large sur son travail. Il s'agit ensuite d'obtenir des résultats comparables avec les résultats d'autres chercheurs sur des sujets proches. Il s'agit enfin d'un véritable travail d'équipe où chacun peut bénéficier

1. Cadre européen commun de référence pour les langues.

des conseils des autres, de leur différence de point de vue, et finalement cela bénéficie également à la qualité de la recherche de chacun.

Ainsi, afin d'identifier des écarts typiques de prononciation entre des apprenants et des locuteurs natifs du français, nous avons créé un corpus que nous allons explorer pour identifier des spécificités de langage (approche *corpus-driven*) que nous allons ensuite tester (approche *corpus-based*). Ce travail permettra à terme de proposer des exercices, avec nous l'espérons, des idées nouvelles. Dans tous les cas, il s'agira de descriptions détaillées des écarts de production entre des apprenants non-natifs et des natifs, qui pourront aider les enseignants de F.L.E., et qui nous serviront dans notre propre activité d'enseignement.

## Références bibliographiques

Biber D., Conrad S. & Reppen R.,
1998, *Corpus linguistics investigating language structure and use*, Cambridge, Cambridge University Press.

Bilger M. (dir.),
2000, *Corpus : Méthodologie et applications linguistiques*, Paris, Champion.

Boersma P. & Weenink D.,
2011, *Praat : Doing phonetics by computer* [Computer program]. Version 5.2.19. Récupéré le 16 mars 2011 de `www.praat.org/`.

Delattre P., 1945, « Prononciation graphique et prononciation phonétique : II. Les voyelles », *The French Review*, 18, 5.

Detey S. & Kawaguchi Y.,
2008, « Interphonologie du Français contemporain (I.P.F.C.) : Récolte automatisée des données et apprenants japonais », Communication présentée aux Journées P.F.C. : *Phonologie du français contemporain : variation, interfaces, cognition*, Paris.

Division des politiques linguistiques, conseil de l'Europe,
2001, *Cadre européen commun de référence pour les langues* (C.E.C.R.), Paris, Didier.

Gendrot C. & Adda-Decker M.,
2006, « Analyses formantiques automatiques en français : périphéralité des voyelles orales en fonction de la position prosodique », *Actes des XXVI^es Journées d'étude de la parole*, 12-16 juin 2006, 205-208.

GOLDMAN J. P., 2011, *EasyAlign : an automatic phonetic alignment tool under Praat Proceedings of InterSpeech*, septembre 2011, Firenze.

LAURET B., 2007, *Enseigner la prononciation du français : Question et outils*, Paris, Hachette.

PILLOT-LOISEAU C., FRÉDET F. & AMELOT A., 2010, « Apports de la phonétique expérimentale à la didactique de la prononciation du Français Langue Étrangère : Étape 1 : Réflexion autour de l'établissement d'un corpus », *Cahiers de l'APLIUT* 29(2), 75-88.

RACINE I., DETEY S., ZAY F. & KAWAGUCHI Y., 2012, « Des atouts d'un corpus multitâches pour l'étude de la phonologie en L2 : L'exemple du projet "Interphonologie du français contemporain" (I.P.F.C.) », in KAMBER A. & SKUPIENS C. (dir.), *Recherches récentes en F.L.E.*, Berne, Peter Lang, 1-19.

RASTIER F., 2001, *Arts et sciences du texte*, Paris, Presses universitaires de France.

TROUBETZKOY N. S., 2005 (1949, 1re éd.), *Principes de phonologie*, trad. J. Cantineau, Paris, Klincksieck.

VAISSIÈRE J., 2006, *La phonétique*, Paris, Presses universitaires de France.

WILLIAMS G. (dir.), 2005, *La linguistique de corpus*, Rennes, Presses universitaires de Rennes.