

ENGLISH
UNE BASE DE DONNÉES COMPARATIVES
DE L'ANGLAIS LU, RÉPÉTÉ ET PARLÉ EN L1 & L2

Anne Tortel

Résumé

La littérature et les recherches menées dans le cadre de l'apprentissage de l'anglais L2 ne cessent de se développer tant au niveau segmental qu'au niveau suprasegmental. Pourtant, lorsque l'on se penche sur le domaine de l'acquisition des langues secondes et plus particulièrement l'acquisition de l'anglais L2 par des francophones, la constitution et le travail sur de grandes bases de données orales semblent inexistantes, ou difficilement accessibles. De ce constat est née la constitution d'une ressource dénommée ANGLISH.

ANGLISH, base de données d'anglais britannique L1 et L2, constitue une ressource importante pour une analyse comparative de l'anglais produit par des francophones et des anglophones. À l'heure actuelle, ANGLISH comporte plus de 5h30 de parole en anglais L1 et L2. Déposé sur le site du CRDO, ANGLISH sera prochainement mis en ligne de manière à ce que la communauté scientifique et enseignante puisse accéder à cette ressource à des fins scientifiques et pédagogiques.

Mots-clés : base de données, anglais britannique, prosodie, langue maternelle, langue seconde, apprenants français.

Abstract

Comparative database of read, repeated and spoken English in L1 and L2.

Literature and studies on English second language acquisition have been recently developed either on a segmental level or a suprasegmental level. Despite this fact, when looking closer at the field of second language acquisition and in particular the acquisition of English as second language by French speakers, the constitution and work on large oral databases seem to be either non-existent or uneasy to access. From this noting was created a database called ANGLISH.

ANGLISH, database of British English in L1 & L2, constitutes an important resource for a comparative analysis of English produced by French and English native speakers. ANGLISH is currently made up of more than 5h30 of oral English in L1 & L2. Registered on the CRDO website, ANGLISH will be shortly available online in order to be accessible to the scientific and teaching community for scientific and pedagogical purposes.

Keywords: database, British English, prosody, first and second language, French learners, English native speakers.

TORTEL, Anne (2008) ANGLISH. Une base de données comparatives de l'anglais lu, répété et parlé en L1 & L2, *Travaux Interdisciplinaires du Laboratoire Parole et Langage*, vol. 27, p. 111-122.

Introduction

L'anglais est la langue officielle de plus de 50 pays et est parlé en tant que langue maternelle par plus de 330 millions de terriens (Weber, 1997). Estimée comme l'actuelle « langue internationale », elle est assurément la seconde langue la plus apprise et étudiée au monde (Crystal, 2003) avec un nombre toujours accru d'apprenants (estimé à 150 millions en 2008 par Weber d'après une interview sur son ouvrage en 2008 : *“The number of speakers of all the top ten languages have gone up in the last quarter century but relative to each other, the situation among the top ten remains unchanged”*).

Parallèlement, la littérature sur l'acquisition des langues est de plus en plus abondante, les travaux sur l'acquisition de l'anglais L2 se font de plus en plus nombreux. Lorsque l'on se penche sur le domaine de l'acquisition des langues secondes et notamment l'acquisition de l'anglais L2 par des francophones, la constitution et le travail sur des corpus oraux sont plus rares. Pourtant, les travaux menés dans le cadre de l'apprentissage de l'anglais L2 ne cessent d'augmenter tant au niveau segmental qu'au niveau suprasegmental. Actuellement de nombreuses études sont menées, par exemple, sur « l'accent étranger » d'un français parlant anglais (Horgues, 2005) ou de locuteurs d'autres nationalités parlant anglais (Cheong, 2007 ; Jilka, 2000).

La recherche en linguistique, quel que soit son domaine, semble se tourner vers la constitution et l'utilisation de production de parole naturelle. De ce point de vue-là, le travail sur corpus représente la base fondamentale des travaux entrepris. Les études portant sur l'oral et l'acquisition des langues L2 ont alors grand intérêt à se fonder sur de grands corpus de parole dont le style répond aux exigences et objectifs visés. L'utilisation de corpus dans l'apprentissage des langues n'est d'ailleurs pas récente et c'est dans les années 1990 que le travail sur de larges corpus est apparu notamment avec la 'naissance' du TaLC (Teaching and Language Corpora). Cependant, comme le remarque Mauranen (2004), si la production et les échanges oraux en L2 font l'objet de nombreux travaux et occupent une place largement notable, le travail sur corpus oraux reste particulièrement mis de côté, les corpus utilisés étant majoritairement des corpus écrits. Le corpus ICLE³ (International Corpus of Learner English) en est une parfaite illustration. En effet, l'ICLE est un important projet qui est le résultat d'une collaboration de plus de 10 ans d'activités ; sa conception est très riche et originale, mais ce corpus ne comporte qu'une base écrite reposant sur deux types d'exercices : commentaire composé argumentatif et dissertation littéraire (environ 500 à 1000 mots pour chacun des exercices). Nous ne trouvons aucun équivalent du même type pour des corpus oraux. Boulton (2008) fait état des différents corpus existants en Didactique des Langues ; sur 39 recensés, nous constatons que seulement 11 portent sur l'oral des apprenants et semblent de surcroît être difficilement accessibles.

Parallèlement à cela, nombreux sont les corpus d'anglais L1, dont les utilisations sont spécifiques à certains types de travaux ou répondent à des exigences en fonction des diverses approches choisies. En effet, de nombreux travaux sur l'anglais, quel que soit leur cadre théorique, semblent

3. <http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm>

s'appuyer sur la volonté d'utiliser des stimuli naturels. Ainsi plusieurs grandes bases de données, orientées selon le type de travaux, sont exploitables et disponibles (études inter-dialectales avec IVIE⁴ (Kochanski *et al.*, 2004), étude d'un phénomène particulier sur de l'anglais spontané avec la base de données Aix-Marsec (Auran *et al.*, 2004), ou encore une base de données comparatives sur les langues européennes telle que EUROM1 (Chan *et al.*, 1995), *etc.*). On retrouve également plus d'une vingtaine de corpus d'anglais L1 sur le site de l'UCREL⁵ (University Centre for Computer Corpus Research on Language) tels que le BNC (90% de données écrites), le Lancaster/IBM Spoken English Corpus devenu SEC puis Aix-Marsec ou encore ITU (International Telecommunication Union) pour exemples de corpus oraux.

De ce constat général, s'est imposée la nécessité de la constitution d'une ressource telle que ANGLISH.

1. ANGLISH : nouvelle base de données de parole d'anglais L2

“Do you speak English?” Si cette phrase devait être prononcée par un Français qui n'a jamais entendu d'anglais au cours de sa vie, nous entendrions très probablement quelque chose comme ‘douillouspe-akanglich’... ANGLISH, corpus d'anglais L1 & L2 constitue une ressource importante pour une analyse comparative de l'anglais produit par des francophones et des anglophones. À l'heure actuelle, il n'existe, à notre connaissance, aucun corpus de ce type distribué en accès libre. Les corpus existants précédemment cités sont soit des corpus écrits, soit des corpus oraux anglais L1. ANGLISH a la particularité de proposer des productions en anglais sur une base de données comparatives natifs anglais/apprenants français et se répertorie comme corpus oral. Nous entendons par corpus oral, un corpus de données orales disponibles sous un format audio.

1.1. Lignes directrices à la création du corpus ANGLISH

La conception retenue pour ce corpus s'imposait par différents souhaits et exigences que nous exposons par la suite dans les différents thèmes développés ci-dessous. Les quatre axes majeurs interdépendants qui ont guidé la conception du corpus reposaient sur (i) la volonté de mener à bien des travaux sur l'évaluation de la prosodie d'anglophones et de francophones apprenants en anglais L2, permettant de créer une évaluation objective intégrable à des logiciels de langues ; (ii) la volonté de combler le manque de larges corpus oraux en anglais L2 ; (iii) la volonté de créer une ressource qui serait utilisable, outre par ses propriétaires, par la communauté scientifique mais aussi enseignante ; (iv) une libre diffusion qui serait destinée à un public d'utilisateurs secondaires avec des buts différents allant du simple apport de support sonore de natifs anglais (anglais L1) à un contenu théorique sur un fait de langue particulier à l'analyse d'erreurs segmentales ou prosodiques de productions de francophones en anglais L2 par exemple.

4. http://www.phon.ox.ac.uk/old_IViE

5. <http://ucrel.lancs.ac.uk/corpora.html>

1.1.2. ANGLISH : corpus, base de données ?

« Corpus » est le premier terme qui émerge de manière dominante lorsque l'on nomme un matériel linguistique de ce type. Si nous prenons en compte la définition de Habert *et al.*, (1997), nous pouvons alors considérer qu'ANGLISH a toutes les caractéristiques d'un corpus traditionnel :

« Une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon de langage ». Cela nous conduit à une définition classique du corpus, comme un rassemblement d'enregistrements audio caractérisant ainsi ANGLISH en tant que tel. Cependant cette définition semble assez limitative. Delais-Roussarie & Durand (2003) proposent une définition un peu plus complète issue de Gibbon *et al.* (1997) qui correspondrait à notre propre collection d'enregistrements : *“A corpus is any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use of the data-in-house, or by people in other organisations”.*

Ainsi nous complétons notre première définition par l'idée que l'ensemble des enregistrements audio est alors accompagné par différents niveaux (ou niveau unique) de transcriptions. C'est d'ailleurs dans ce sens-là que nous relevons la tâche considérable et la difficulté de travail qui nous sont apparues tant au niveau du traitement des données audio (environ 5h30) qu'au niveau des enregistrements et des différents traitements orthographiques, extractions des données, annotations *etc.*

Cependant, par contraste avec la représentation classique du corpus comme élément achevé ou abouti, un second concept consiste à examiner ANGLISH également en tant que base de données. Selon le jargon français⁶, une base de données se définit comme telle : *« Une base de données doit être conçue pour permettre une consultation et une modification aisée de son contenu, si possible par plusieurs utilisateurs en même temps ».*

ANGLISH répond donc également aux exigences d'une base de données comportant une structuration particulière des données (enregistrements structurés en catégorie *cf.* § Annotation et transcription, par exemple) et une possibilité de requête (Auran, 2003) dans la mesure où les informations classées de type TextGrid dans le logiciel Praat (Boersma & Weenick, 2000) sont isolées sur des niveaux distincts nommés « tiers » et permettent commodément la formulation de diverses requêtes. Ainsi l'annotation des données peut se faire sur plusieurs années.

ANGLISH en tant que base de données s'inscrit donc dans le cadre d'une conception dynamique qui est appelée à une constante évolution.

1.1.3. Parole de laboratoire *vs* parole spontanée vers parole naturelle

Outre les corpus écrits, on distingue dans les travaux de phonétique et phonologie, deux types de corpus, à savoir les corpus de parole de laboratoire et les corpus de parole spontanée. La parole de laboratoire correspond à des enregistrements de parole effectués dans des conditions

6. <http://www.linux-france.org/prj/jargonf>

expérimentales, très souvent en chambre sourde de manière à obtenir une qualité de son. Les données se composent de lecture de logatomes, de mots, de phrases isolées, de textes ou de phrases préparées en amont. Les exercices proposés étant plutôt de la parole contrainte, cela peut parfois aboutir à des résultats peu naturels voire artificiels. En revanche, cela permet à l'expérimentateur d'étudier un phénomène précis en utilisant des stimuli contrôlés. Nous retiendrons comme atouts pour le corpus ANGLISH que ce type de méthodologie permettra d'obtenir une qualité du signal sonore optimale (*cf.* qualité technique du corpus) mais aussi d'en retirer une base de données comparables pour l'ensemble des locuteurs, ce qui nous fera pencher pour une tâche de lecture de textes. Le deuxième type de parole, qui est celui de la parole spontanée ou authentique, vient s'opposer à la parole de laboratoire dans le sens où aucune contrainte n'est imposée aux locuteurs, ces derniers étant enregistrés dans des conditions naturelles de production (échanges, débats, interviews, *etc.*) qui ne nécessitent pas à l'expérimentateur d'avoir préalablement organisé des stimuli. L'avantage de ce type de corpus est qu'il peut être recueilli en chambre sourde. De plus en plus d'études se tournent vers de la parole naturelle de manière à obtenir des caractéristiques authentiques des productions des locuteurs. C'est dans ce sens-là que nous avons souhaité allier à la fois qualité optimale du son (enregistrements effectués en chambre sourde), données comparables par les tâches de lecture et répétitions (*cf.* variétés des genres) et enfin parole naturelle (il a été demandé aux locuteurs de parler spontanément d'un sujet de leur choix, racontant par exemple une anecdote sur leurs dernières vacances). Ainsi ANGLISH combine trois types de données orales tout en gardant une qualité acoustique idéale pour des analyses phonétiques.

1.1.4. Focus sur l'oral

Comme nous l'avons expliqué précédemment, alors qu'il existe un nombre assez important de corpus écrits disponibles et exploitables pour travailler sur l'anglais et sur le français, la situation est un peu moins fructueuse concernant les corpus oraux et tout particulièrement concernant l'anglais L2 ; en effet, il existe une réelle paucité de corpus oraux d'anglais L2 produit par des francophones. Par ailleurs, parallèlement à cela, un travail considérable reste à entreprendre dans le domaine de l'apprentissage de l'oral de l'anglais. À plusieurs reprises, nous avons pu constater le peu de place laissée à l'oralisation des langues proportionnellement à celle qui est faite à l'écrit. Les épreuves de langues du baccalauréat, par exemple, sont d'abord menées sur un support écrit, l'oral n'arrive que de manière optionnelle ; il a d'ailleurs été constaté que plus de 70% des lycéens au terme de leur scolarisation de second cycle et ayant étudié l'anglais pendant plus de 7 ans, atteignent un niveau de compétence médiocre selon les rapports européens⁷. Par ailleurs, lorsque l'on se penche sur le premier système majeur d'évaluation fondé sur le cadre commun européen du Conseil de l'Europe, DIALANG, aucune place pour l'oral n'a été faite lors de la conception du

7. European network of Policy Makers for the evaluation of Education Systems (2004) The Assessment of Pupils' Skills in English in Eight European Countries 2002. A European Project. <http://cisad.adc.education.fr/revu/pdf/lire16.pdf>

‘test-diagnostique’. Ainsi, de nouveau, ce sont des compétences écrites de la langue qui sont évaluées. Il nous semblait donc important qu’une ressource de productions orales soit créée afin (i) de pallier un manque de corpus anglais L2 et (ii) d’obtenir une base de données qui pourrait donner lieu à de nombreux travaux sur l’évaluation de l’oral et qui permettrait d’inclure un système d’évaluation automatique de l’oral à ces tests déjà existants et incomplets.

1.1.5. Variétés des genres

Pour cette base de données, l’intention première était de réaliser à la fois (i) des tâches représentatives du contenu d’un logiciel de langue destiné à l’apprentissage de la prosodie de l’anglais ; (ii) des exercices que l’on retrouve en contexte d’apprentissage en parcours universitaire, et (iii) d’obtenir des données identiques en termes de contenu qui soient comparables pour l’ensemble des locuteurs ; notre choix s’est donc porté sur une tâche de lecture, de répétition de phrases, et de parole naturelle (*cf.* § contenu de la base de données pour une description détaillée).

1.1.6. Variétés des niveaux d’apprentissage des locuteurs

Afin de constituer nos différents groupes de locuteurs, la question de la représentativité a été de nouveau abordée. La constitution d’un précédent corpus (Tortel, 2004) a permis de mettre en avant la problématique des groupes de niveaux. Quatre groupes avaient été constitués : français non spécialistes de la langue, étudiants de 1^{ère} année, étudiants de 4/5^{ème} année, et natifs anglais. Il était apparu, au vu des résultats obtenus, que les groupes de non spécialistes et étudiants de 1^{ère} année avaient des résultats quasi équivalents, même chose pour les étudiants de 4/5^{ème} année avec le groupe des natifs. Il était donc souhaitable de constituer des groupes de niveau plus distincts par rapport à leur niveau de langue (niveau académique). Trois groupes ont donc été constitués : le groupe témoin qui est celui des natifs anglais, le groupe d’étudiants de 2/3^{ème} année, spécialistes de la langue, et un groupe de « faux débutants » (Hadley, 2002) avec un niveau Bac en anglais, n’ayant pas poursuivi d’études supérieures spécialisées dans l’anglais, qui pourraient être dans l’optique de se ‘remettre’ à l’anglais dans un but spécifique ou personnel et qui pourraient ainsi être des utilisateurs potentiels d’un logiciel de langues. Concernant l’homogénéité des groupes, afin de valider la répartition des locuteurs non seulement selon des critères de niveaux ‘académiques’ annoncés mais aussi en fonction de variabilités intra-locuteurs possibles au sein d’un même groupe, une évaluation subjective est en cours de réalisation par le biais d’un test de perception effectué par un jury de natifs et d’experts. Cette question n’étant pas l’objet de cet article, nous ne rentrerons pas plus dans les détails.

1.1.7. Qualité technique du corpus

Pour rendre possible l’analyse expérimentale, la constitution d’une base de données sonores impose une précaution particulière et des caractéristiques techniques précises quant à l’enregistrement effectué ; en effet, le choix du lieu d’enregistrement et du matériel utilisé (entre autres) sont des questions primordiales pour mener à bien une analyse phonétique du signal

(conditions identiques pour chaque enregistrement, absence de bruits de fond, matériel de qualité, bon rapport signal/bruit, *etc.*). Ainsi, pour une qualité sonore optimale, tous les enregistrements ont eu lieu en chambre anéchoïque (dans les locaux du LPL) à une fréquence de 20KHz et une résolution de 16 bits et ont été effectués à l'aide d'un micro-casque. Les données ont été recueillies directement sur disque dur sous format wav.

1.1.8. Dispositif actuel

Actuellement, ANGLISH comporte environ 5h30 de parole en anglais (L1) par des natifs anglais britanniques et en L2 par des francophones de deux niveaux différents en anglais. Les productions enregistrées se répartissent sur trois tâches différentes (lecture, répétition, monologue) et comptent 63 locuteurs. Une description plus détaillée des tâches effectuées et des locuteurs est proposée dans le paragraphe suivant.

2. Description de la base de données ANGLISH

2.1. Contenu de la base de données

Le corpus ANGLISH se compose de trois phases d'exercices différents en anglais, enregistrées, comme dit précédemment, en chambre sourde, selon l'ordre suivant :

- Première tâche proposée :
Il s'agit d'une tâche de lecture de 4 passages issus du corpus EUROM 1 (P9, Q0, Q9, R3), lus par 63 locuteurs, ce qui représente un total de 252 passages (soit 1260 phrases) pour une durée approximative d'une heure et demie de lecture (soit 4765,3138 secondes). Chacun des passages se compose de 5 phrases liées sémantiquement racontant un événement, une histoire de la vie quotidienne. La consigne donnée à la fois oralement et à l'écrit était de :
« Lire les passages comme si l'histoire vous était arrivée et que vous la racontiez à un ami ». Les passages étaient donnés à l'avance de manière à ce que les locuteurs se familiarisent avec les textes et puissent éviter les bafouillages lors de l'enregistrement. Ci-dessous, un exemple de passage proposé : *"I've always found it difficult to sleep on long train journeys in Britain. For one thing, I can never make myself comfortable in the seat. Then the other passengers usually talk so loudly or worse still they snore. In addition, there's the constant clickety-click of the wheels on the track. If I do manage to doze off the ticket inspector comes along and wakes me"*.
- Deuxième tâche proposée :
Il s'agit d'une répétition de 25 phrases (par 63 locuteurs), selon un modèle natif (féminin ou masculin) enregistré dans le cadre du corpus EUROM 1, issues des 4 passages utilisés pour la tâche de lecture ainsi qu'un passage (R3) ajouté, toujours extrait du corpus

EUROM 1. La consigne donnée était de « répéter les phrases suivantes en essayant d'imiter la façon dont elles sont produites ». Cette tâche représente un total de 1575 phrases, soit une durée totale approximative d'une heure et demie (soit 4811, 2246 secondes) pour l'ensemble des phrases répétées. En cas de bafouillage, le sujet pouvait se reprendre ou recommencer la phrase, le pilotage de la bande-son des phrases étant mené par l'examinatrice elle-même (sachant qu'à la base, une durée de 10 secondes avait été fixée entre chaque phrase). Ci-dessous, un exemple de quelques phrases proposées sur les 25 données :

« [...] 04. In addition, there's the constant clickety-click of the wheels on the track.
05. If I do manage to doze off the ticket inspector comes along and wakes me.
06. What can I have for dinner tonight?
07. I do have some fresh pasta in the fridge.
08. The trouble is, I eat that two or three times a week. [...] »

- Troisième tâche proposée :

Elle consiste en un monologue de 2 minutes environ (produit par 63 locuteurs), sur un sujet non imposé (le thème des vacances avait été suggéré à chacun des sujets mais n'était pas imposé afin que le locuteur soit le plus possible à l'aise et confiant dans cet exercice). Ainsi, les locuteurs avaient 'libre antenne' pendant environ deux minutes pour parler d'un sujet de leur choix, non préparé, si ce n'est quelques minutes, le temps de réfléchir à un thème approprié. Cette tâche représente une totalité de 63 monologues, soit une durée totale approximative de parole de trois heures (soit 10799,9619 secondes). Ci-dessous, un extrait d'un monologue produit par une locutrice (F03GB) :

“so for the last three or perhaps since for four weeks I've been learning this new Chopin ballad it's the first one in C minor it's terribly difficult and [erm] although I've been working at it quite a lot there's still loads more to do what I have to do is break it down into small parts and try and work on the technical aspects of some of it [...]”.

2.2. Locuteurs

Les 63 locuteurs, volontaires pour enregistrer les trois exercices proposés forment 3 groupes distincts, comme expliqué précédemment (*cf.* § Variétés des niveaux d'apprentissage) :

- GB : 23 locuteurs, 13 femmes et 10 hommes, britanniques anglais ;
- FR2: 20 locuteurs, 10 étudiantes et 10 étudiants anglicistes, fin de 2^{ème}/début 3^{ème} année ;
- FR1: 20 locuteurs, 10 femmes et 10 hommes en activité ayant niveau Bac anglais L2.

Concernant les locuteurs du groupe GB, la moyenne d'âge est de 31 ans. Tous viennent d'Angleterre, et plus précisément, 6 des Midlands, 4 du Nord et 13 du Sud de l'Angleterre. Pour le groupe FR2, tous sont français de langue maternelle française, étudiants en anglais, spécialistes de la langue cible, ayant suivi le cursus universitaire de phonétique à Aix-en-Provence avec un apprentissage de la prosodie et des caractéristiques segmentales ; ils sont soit en fin de 2^{ème} année,

soit en début de 3^{ème} année. Aucun n'a effectué de séjour dans un pays anglophone de plus de quatre semaines consécutives, et aucun ne fait partie d'une famille dont l'un des parents ou membres serait anglophone. Tous ont suivi en moyenne 9 à 10 ans d'études d'anglais en milieu scolaire. La moyenne d'âge du groupe se situe entre 19 et 22 ans.

Enfin, le groupe FR1 se compose d'adultes français de langue maternelle française, d'un niveau d'études post-baccalauréat et exerçant une activité professionnelle. La moyenne d'âge de ce groupe est de 37,5 ans. Aucun ne possède de connaissances approfondies en phonétique anglaise ; tous ont un niveau bac d'anglais et ont suivi un cursus scolaire d'anglais en tant que L2. Aucun n'a effectué de séjour récent de plus d'un mois dans un pays anglophone.

Toutes ces informations ont été recueillies par le biais d'un questionnaire présenté aux sujets en amont des exercices à réaliser. Un consentement de participation et de droit de diffusion des enregistrements a également été présenté et signé par chacun des participants.

2.3. Codage du corpus

Afin de pouvoir repérer plus facilement les locuteurs pour la suite de ce travail, nous avons codé leur prestation en fonction des tâches effectuées et de leur groupe d'appartenance :

- Pour la tâche de lecture (exemple : F02GBP2)
 1. La première lettre indique le sexe du locuteur : **F** pour une femme et **H** pour un homme.
 2. En deuxième position, le nombre indique le numéro attribué au locuteur (ici, 02)
 3. Les deux lettres suivantes indiquent le groupe d'appartenance du locuteur : **GB**, **FR1** ou **FR2** (ici, GB), le numéro indique le niveau (1 = faux débutants ; 2 = étudiants anglicistes)
 4. Enfin le dernier groupe de caractères du codage indique l'exercice effectué : **P** pour passage et **1, 2, 3** ou **4** pour le numéro du passage.

- Pour la tâche de répétition (exemple : H05FR1phrase01_Britain)
 1. Les points 1, 2 et 3 précédents restent inchangés
 2. Le mot 'phrase' indique que c'est la tâche de répétition de phrases qui est codée
 3. Le nombre suivant est le numéro attribué à la phrase (ici, 01)
 4. Le mot suivant est un mot-clé de la phrase, il ne correspond pas nécessairement au noyau de la phrase (ici, Britain).

- Pour la tâche du monologue (exemple : F10FR2_monologue)
 1. Les points 1, 2 et 3 précédents restent inchangés
 2. La tâche du monologue est indiquée par le mot « monologue » comme dans l'exemple cité).

2.4. Annotation, transcription

Le travail de transcription et d'annotation a été effectué avec le logiciel PRAAT (Boersma & Weenick, 2009) par l'expérimentatrice elle-même. La transcription de ENGLISH est essentiellement orthographique, non ponctuée, mais indique les phénomènes typiques de l'oral tels que les faux-départs, les pauses pleines, les répétitions, les mots tronqués. En voici quelques exemples :

- Les mots tronqués sont suivis d'un tiret, par exemple lors d'une hésitation : « bene- » pour « benefited »
- Les aspects non linguistiques tels que les rires sont marqués entre crochets, exemple : [laugh]
- Les interjections de type 'hésitation' où le locuteur cherche un mot particulier ou cherche à gagner du temps pour réfléchir à la suite de son discours sont marquées 'er' ou 'erm' pour les locuteurs britanniques et 'heu' pour les locuteurs français.
- Les formes contractées ont été retranscrites telles quelles sans modification, c'est-à-dire sans avoir été ramenées à une forme non contractée, comme le mot 'cos', forme contractée du mot « because » dans la phrase « I think just 'cos I met erm a friend ».

À partir de la transcription orthographique, une segmentation manuelle en unités intonatives a été effectuée. Cette unité est codée « UI » et inclut les pauses, notées par le symbole « # ». La deuxième étape a été une segmentation manuelle en mots dans une nouvelle tire nommée « mot ». Actuellement un découpage en unités rythmiques selon le modèle de Jassem (1952) est en cours de réalisation. La prochaine étape sera une segmentation en phonèmes, codée « CVC » pour les besoins de notre propre travail de thèse, portant sur la définition de critères prosodiques évaluatifs du rythme de l'anglais d'apprenants français.

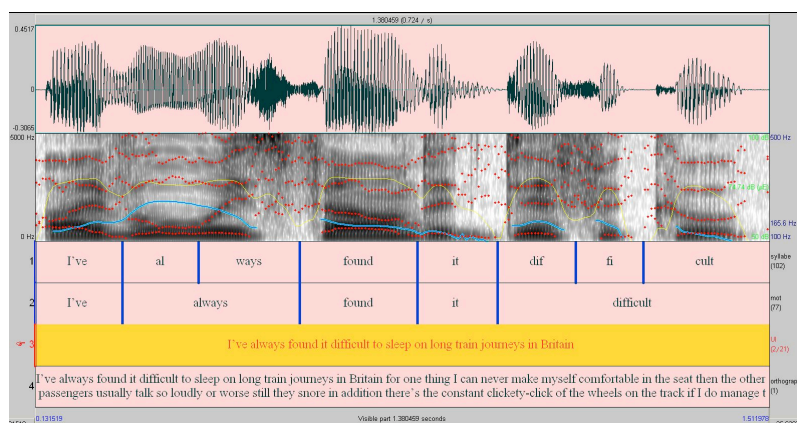


Figure 1
Exemple d'alignement du locuteur F09FR2 effectué dans Praat

2.5. Accessibilité

Le corpus ANGLISH est en libre accès sur le Centre de Ressources pour la Description de l'Oral (CRDO) <<http://crdo.fr>> Le CRDO (Bel & Blache, 2006) est le CRN (Centre de Ressources Numériques) centré sur les ressources orales, initiative conjointe de la Direction de l'Information Scientifique et du Département scientifique « Homme et Société » du CNRS sous la responsabilité des laboratoires LACITO et LPL. La mise en ligne de ANGLISH sur le CRDO constitue un accès ouvert à la communauté scientifique et enseignante dans le but que chacun puisse accéder à cette ressource à des fins scientifiques et pédagogiques.

Bibliographie

- AURAN, C. (2004) *Prosodie et anaphore dans le discours en anglais et en français: cohésion et attribution référentielle*, Thèse de doctorat, Université de Provence.
- AURAN, C. ; BOUZON, C. ; HIRST, D. (2004) The Aix-Marsec Project: an evolutive database of spoken British English, in *Proceedings of 2nd International Conference on Speech Prosody*, Nara, Japan, March 2004, p.561-564.
- BEL, B. ; BLACHE, P. (2006) Le Centre de Ressources pour la Description de l'Oral (CRDO), *Travaux Interdisciplinaires du Laboratoire Parole et Langage (TIPAL)*, vol. 25, p. 13-18.
- BOERSMA, P. ; WEENICK, D. (2009) *Praat, a system for doing phonetics by computer, version 5.1.04*. Téléchargeable à partir de <www.praat.org>.
- BOULTON, A. (2008) Esprit de corpus : promouvoir l'exploitation de corpus en apprentissage des langues, revue électronique *Texte et Corpus*, n°3, Actes des Journées de la linguistique de Corpus 2007, p. 37-46.
- CHAN, D. ; FOURCIN, A. ; GIBBON, D. ; GRANDSTRÖM, B. ; HUCKVALE, M. ; KOKKINAKIS, G. ; KVALE, K. ; LAMEL, L. ; LINDBERG, B. ; MORENO, A. ; MOUROPOULOS, J. ; SENIA, F. ; TRANSCOSO, I. ; VELT, C. & ZEILIGER, J. (1995) EUROM – A Spoken Language Ressource for the EU, in *Proceedings of Eurospeech'95*, Madrid, 1995.
- CHEONG, H.S (2007) *The role of listener affiliated socio-cultural factors in perceiving native accented versus foreign accented speech*, Doctoral Dissertation, The Ohio State University.
- CRYSTAL, D. (2003) *English as a Global Language*, Cambridge: Cambridge University Press, 2nd ed.
- DELAIS-ROUSSARIE, E. & DURAND, J. (2003) *Corpus et variation en phonologie du français*, PUM.
- GIBBON *et al.* (1997) *Handbook of standards and resources for spoken language systems*, vol.4: *Spoken language reference materials*, XVI, Berlin: Mouton de Gruyter.
- GOLDMAN, J.-P. *et al.* (2007) Phonostylographe : un outil de description prosodique. Comparaison du style radiophonique et lu, *Nouveaux cahiers de linguistique française*, n° 28, p. 219-237.
- HABERT, B. *et al.* (1997) *Les linguistiques de corpus*, Paris : Armand Colin.

- HADLEY, G. (2002) Sensing the winds of change: an introduction to data-driven learning, *RELC Journal*, vol.33,2, p. 99-124.
- HORGUES, C. (2005) Contribution à l'étude de l'accent français en anglais. Quelques caractéristiques prosodiques de l'anglais parlé par des apprenants francophones et leur évaluation perceptive par des juges natifs, *Actes des VIIIes RJC Langage et Langues*.
- JASSEM, W. (1952) *Intonation in conversational English*, Warsaw: Polish Academy of Science.
- JILKA, M. (2000) *The contribution of intonation to the perception of foreign accent*, Doctoral Dissertation, Arbeiten des Instituts für Maschinelle Sprachverarbeitung (AIMS), vol. 6(3), University of Stuttgart.
- KOCHANSKI, G., GRABE, E. & COLEMAN, J. (2004) The difference between a question and a statement: a survey of English dialects, *The Journal of the Acoustical Society of America*, 115(5), p. 2398.
- MAURANEN, A. (2004) Speech Corpora in the Classroom, in G. Aston *et al.* (eds), *Corpora and Language Learners*, Amsterdam, John Benjamins, p. 195-211.
- TORTEL, A. (2004) *Evaluation of Intonation and Rhythm of French and Native Speakers*, mémoire de recherche de master I, Université de Provence.
- WEBER, G. (1997) The World's 10 most influential Languages, *Language Today*, 3, p. 12-18.