Dictionnaires vs. Données orales : Le cas de la gémination de /r/ en anglais

Journée data SHS - 8 décembre 2022

Quentin Dabouis

Le projet a été mené en collaboration avec **Olivier Glain** (Université Jean Monnet de Saint-Etienne) et **Sylvain Navarro** (Université de Paris)

Plan

- La gémination en anglais
- Le cas de /r/
- Objectifs
- Méthodologie
- Résultats

La gémination

On parle de consonne géminée pour désigner des consonnes « doubles ».

Par exemple, en italien, on peut opposer *fato* et *fatto* sur la seule base de l'opposition [t] ~ [tː].

On parle alors de géminées « phonologiques ».

La géminination en anglais

L'anglais n'a pas de géminées phonologiques mais peut avoir des "géminées morphologiques" au niveau de frontières de morphèmes ou de mot (e.g. *right time, unnecessary*)

Dans un mot comme *illegal*, on a

il-	+	*legal*	→	*illegal*
préfixe		base		dérivé

C'est parce que la dernière consonne du préfixe et la première consonne de la base sont identiques qu'on a la possibilité de géminer

La gémination en anglais

La gémination peut être traitée comme un processus de doublement phonétique ou, pour l'anglais, de longueur phonétique (Kaye 2005)

Les études existantes se sont concentrées sur les mots préfixés et ont traité les consonnes [l, n, m] (e.g. *illegal, immoral, unnamed*)

La gémination en anglais

Ces études ont montré que l'on pouvait trouver de la gémination consonantique en anglais pour les mots préfixés dont le sens est transparent, bien que cela varie en fonction de

- la vitesse d'élocution
- la capacité du préfixe à être utilisé pour former des nouveaux mots (« productivité »)
- l'environnement phonologique
  - présence d'un accent tonique sur la syllabe suivante
  - nature des segments suivants (consonne ou voyelle)

(Bauer 2003 ; Cruttenden 2014 : 248 ; Ben Hedia & Plag 2017 ; Kaye 2005 ; Oh & Redford 2012 ; Videau 2013)

Le cas de /r/

Aucune des études existantes n'a traité de la possible gémination de /r/

Dabouis (2016) a trouvé que les dictionnaires de prononciation font état d'une différence entre anglais britannique (RP) et anglais américain (GA) sur la possibilité de géminer /r/:

→ il serait possible de géminer /r/ en GA mais pas en RP dans des mots comme *irrational* ou *irremovable*.

Le cas de /r/

On trouve cette différence dans plusieurs dictionnaires.
(Kenyon & Knott 1953; Merriam-Webster online; Upton & Kretzschmar 2017; Wells 2008).

Dabouis (2016) propose que cette différence soit attribuée à la **rhoticité** :

Les variétés d'anglais se distinguent notamment par la possibilité d'avoir un /r/ qui n'est pas suivi d'une voyelle :

- les variétés qui le peuvent, comme le GA, sont dites **rhotiques** (e.g. *car* [ˈkʰɑɹ])
- celles qui ne le peuvent pas, comme le RP, sont dites **non-rhotiques** (e.g. *car* [ˈkʰɑː])

L'hypothèse de la segmentabilité

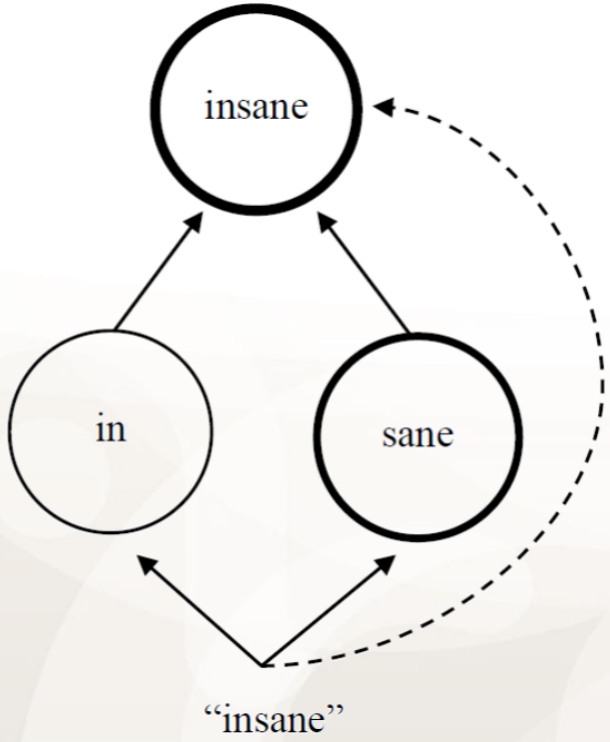
Une étude précédente sur la gémination, Ben Hedia & Plag (2017), a testé l'hypothèse de la segmentabilité, et nous allons chercher à la tester également

Cette hypothèse a été proposée par Hay (2001, 2003), et elle dit que :

- Les mots complexes les plus segmentables sont plus susceptibles de préserver les propriétés phonologiques de leur base
- Les mots complexes les moins segmentables sont plus susceptibles de s'éloigner phonologiquement de leur base

L’hypothèse de la segmentabilité

Cette hypothèse se fonde sur un modèle d’accès lexical à deux voies :



Voie directe

Voie décomposée

Il y aurait une “course” entre ces deux voies. La plus rapide est celle empruntée la plus souvent → la fréquence d’utilisation du dérivé et de la base détermine quelle voie l’emporte

L’hypothèse de la segmentabilité

Quel lien avec la gémination ?

Si un mot complexe est plus segmentable, on s’attend à ce que la frontière entre le préfixe et la base soit plus marquée

Les prédictions de cette hypothèse sont donc :

- Fréquence du dérivé + élevée → [ɹ] + court
- Fréquence de la base + élevée → [ɹ] + long
- Relation sémantique base-dérivé transparente → [ɹ] + long

Objectifs

Vérifier si la différence entre RP et GA peut être confirmée avec des **données orales**

Si elle est confirmée, évaluer si la gémination de /r/ dépend des **mêmes variables** que celles trouvées pour les autres consonnes

Quelles données utiliser ?

En linguistique, les types de données utilisées peuvent être classes en deux grandes catégories :

- Les corpus (qui peuvent être oraux ou écrits) de données (quasi-)spontanées

- Les données expérimentales :
  - Jugements de grammaticalité
  - Elicitation

Quelles données utiliser ?

Tous les types de données ne sont pas adaptés à chaque thématique étudiée

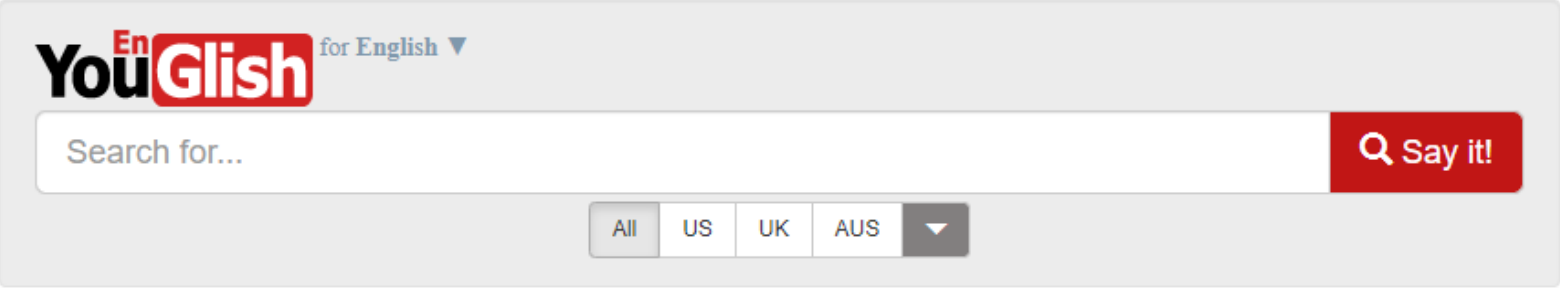
Quand plusieurs méthodologies sont possibles, elles n'ont pas les mêmes avantages et inconvénients

Pour notre étude, nous avons choisi de constituer un corpus de données (quasi-)spontanées car nous souhaitions comparer des données dictionnairiques à des données orales

Méthodologie

Source des données

Youglish est un site qui permet d'effectuer des requêtes dans les sous-titres de Youtube et qui permet d'accéder à des réalisations orales de mots ou expressions



Il est possible de filtrer par variété d'anglais

Méthodologie

Mots pertinents

Nous avons sélectionné 16 mots en <irr-> qui varient en fréquence, en transparence sémantique, selon qu'ils ont ou non une base attestée (e.g. *irregular* ← *ir- + regular*) et selon la présence ou absence d'un accent sur la deuxième syllabe

Les mots sont les suivants : *irradiate, irradiated, irradiation, irrational, irrefutable, irregular, irrelevance, irrelevant, irreplaceable, irresponsibility, irresponsible, irreverence, irrevocably, irrigation, irritate, irritation*

Mots pertinents

Mots pertinents

Parmi les 16 mots, 9 sont des préfixés transparents (ci-après "mots dérivés") : *irrational, irrefutable, irregular, irrelevance, irrelevant, irreplaceable, irresponsible, irreverence, irrevocably*

Les mots restants sont :

- 3 mots avec un préfixe locatif assez opaque : *irradiate, irradiation, irradiated*
- 3 préfixés opaques : *irrigation, irritate, irritation*
- *Irresponsibility*, formé comme *irresponsible + -ity*

Méthodologie

Les 16 mots étudiés ont été extraits automatiquement de Youglish

Un maximum de 25 occurrences par genre (homme/femme) et par variété d'anglais ont été sélectionnées

→ maximum théorique de 100 occurrences par mot

→ jamais atteint car on trouve moins d'occurrences britanniques que d'occurrences américaines, et moins d'occurrences produites par des femmes que d'occurrences produites par des hommes

Méthodologie

	UK			US			Total
	F	M	Total	F	M	Total	général
irradiate	1	3	*4*	3	25	*28*	***32***
irradiated	2	5	*7*	8	25	*33*	***40***
irradiation	1	4	*5*	12	25	*37*	***42***
irrational	16	25	*41*	25	25	*50*	***91***
irrefutable	0	8	*8*	18	25	*43*	***51***
irregular	21	25	*46*	25	25	*50*	***96***
irrelevance	1	7	*8*	10	22	*32*	***40***
irrelevant	24	22	*46*	24	24	*48*	***94***
irreplaceable	6	13	*19*	23	26	*49*	***68***
irresponsibility	0	4	*4*	7	25	*32*	***36***
irresponsible	15	24	*39*	22	25	*47*	***86***
irreverence	2	1	*3*	7	16	*23*	***26***
irrevocably	1	9	*10*	13	25	*38*	***48***
irrigation	1	12	*13*	25	24	*49*	***62***
irritate	4	18	*22*	25	25	*50*	***72***
irritation	7	9	*16*	25	25	*50*	***66***
Total général	**102**	**189**	***291***	**272**	**387**	***659***	***950***

Méthodologie

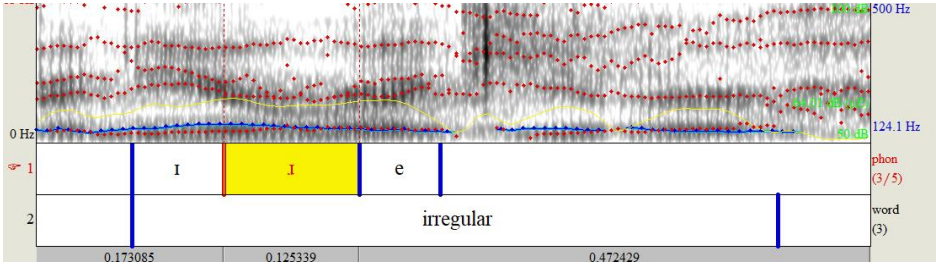
En condensé, la répartion se fait comme suit :

	UK	US	
Hommes	189	387	*576*
Femmes	102	272	*374*
	291	*659*	*950*

Méthodologie

Les données ont été analysées spectrographiquement sur Praat (Boersma & Weenik 2014) de façon à opérer deux mesures :

- la durée de [ɹ]
- celle du mot entier

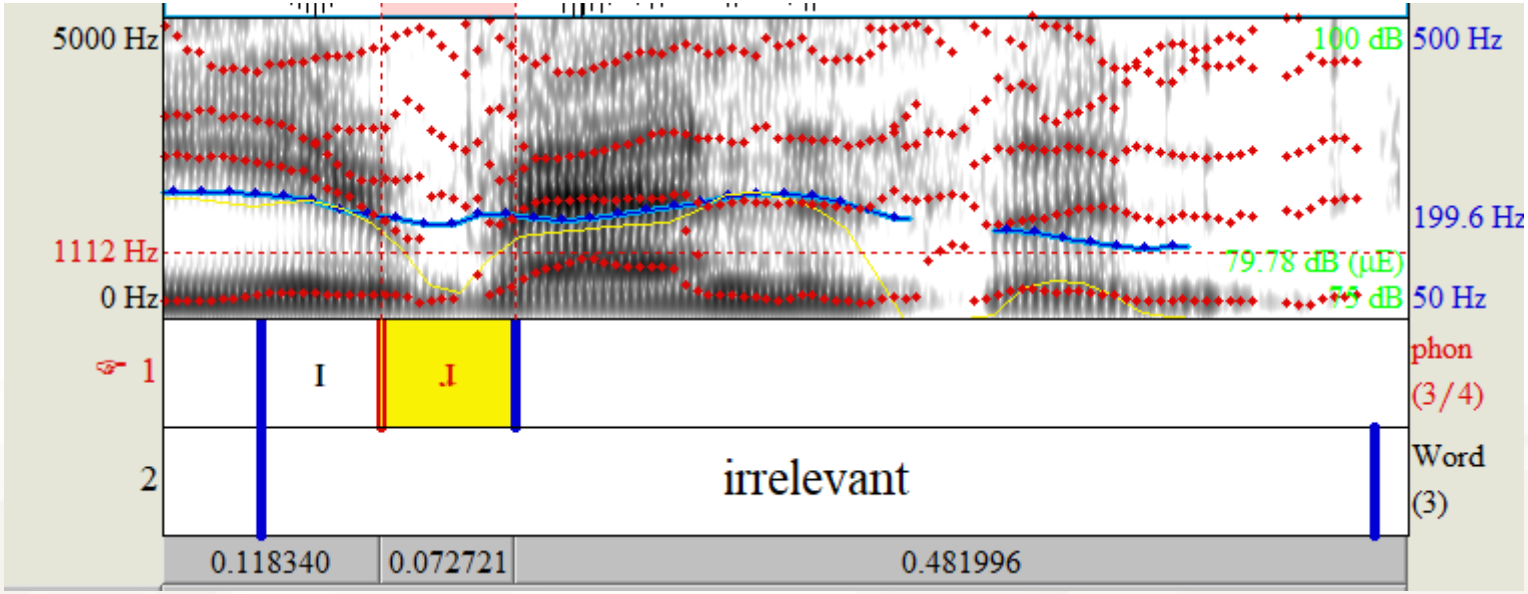


L'analyse a été faite manuellement par un des auteurs de l'étude, puis vérifiée par un des co-auteurs

La segmentation est basée sur une analyse auditive des extraits (l'analyse visuelle n'étant pas possible pour [ɹ], ce son étant trop similaires aux voyelles qui l'entourent)

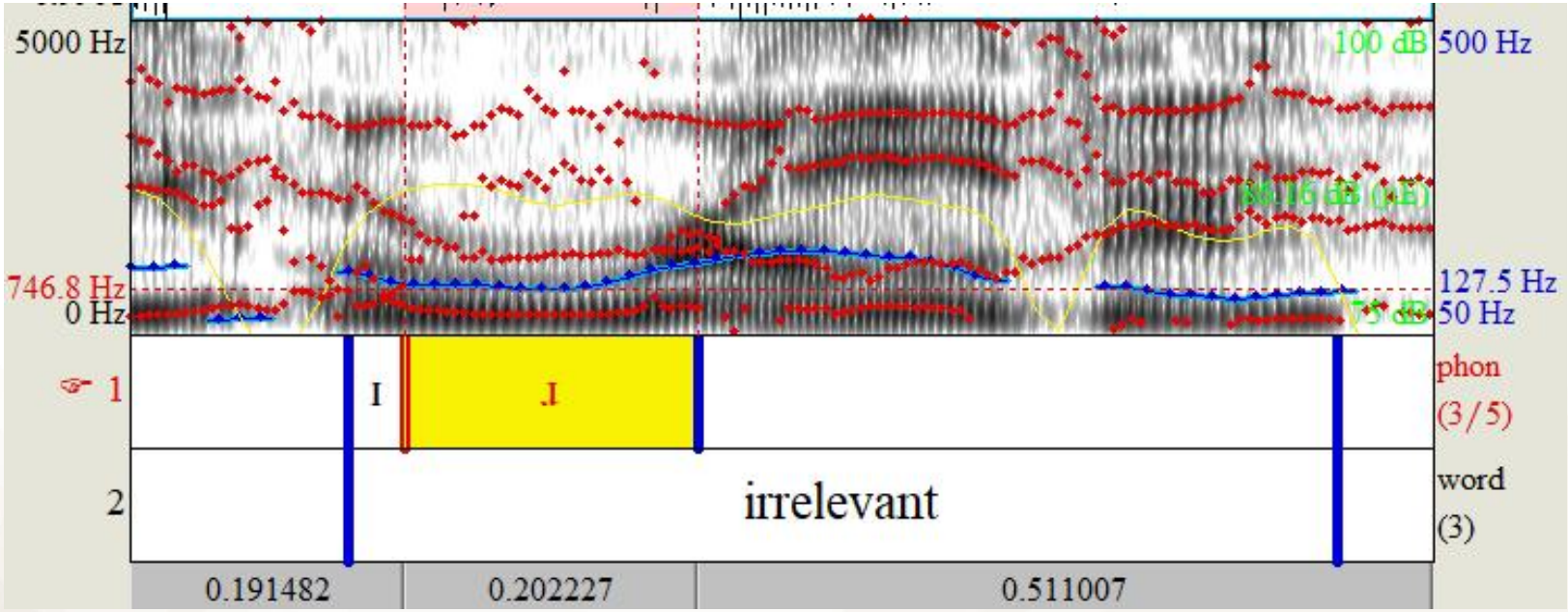
Exemples : [ɻ] non-géminé

irrelevant, UKF4



Exemples : [ɻ] géminé

irrelevant, USM1



Le codage des données

Nous avons codé les variables suivantes :

- **R-LENGTH** : Durée du [ɹ] en secondes
- **SPEECHRATE** : Ratio du nombre de segment dans le mot et de sa durée en secondes. Par exemple :
  - Pour l'occurrence de USM18 de *irradiate*, on a une durée de 0,959589s
  - Ce mot a 7 segments :

ɪ	ɹ	eɪ	d	i	eɪ	t
1	2	3	4	5	6	7

  - La vitesse d'élocution est donc 7/0,959589 = 7,294789749 segments/s

Le codage des données

- **LogFrequency** : fréquences d'utilisation de la base et du dérivé
  - Relevées dans les corpus SUBTLEX-UK et US et du COCAE
  - log-transformées ($\log_e(x+1)$) pour ressembler à la manière dont le cerveau humain traite les informations de fréquence
- **RelativeFrequency** : ratio de la fréquence de la base et de celle du dérivé
- **SecondSyllable** : Stressed ou Unstressed
- **Gender** : Male ou Female
- **SemanticTransparency** : les mots ont été codés comme Transparent ou Opaque

Le codage des données

1	Word	Speaker	R-length	Wd-length	Nb-sgmts	SpeechRat	Gender	EngVa	FqD-US-SL	logFq-US-S	FqB-US-SL	RelFq-US-S	FqD-US-CC	logFq-US-C	FqB-US-CC	RelFq-US-C
2	irradiate	USM18	0,23086	0,959589	7	7,2947897	M	US	6	1,9459101			236	5,4680601		
3	irrevocably	USM23	0,227184	0,985942	10	10,142584	M	US	17	2,8903718	0	#DIV/0!	597	6,3935908	1	597
4	irrevocably	USM15	0,222138	1,074108	10	9,3100508	M	US	17	2,8903718	0	#DIV/0!	597	6,3935908	1	597
5	irrevocably	USF8	0,215218	1,007829	10	9,9223182	F	US	17	2,8903718	0	#DIV/0!	597	6,3935908	1	597
6	irradiate	UKM2	0,214162	0,731691	7	9,56688	M	UK	6	1,9459101			236	5,4680601		
7	irrational	USF13	0,212341	0,802564	6	7,4760393	F	US	149	5,0106353	256	0,5820313	2817	7,9437827	7924	0,3555023
8	irrefutable	USM8	0,208105	0,904632	10	11,054219	M	US	13	2,6390573	0	#DIV/0!	394	5,9788858	13	30,307692
9	irrelevance	USM19	0,207946	0,73494	8	10,885242	M	US	0	0	82	0	428	6,0614569	4985	0,0858576
10	irrational	USM14	0,207767	0,876374	6	6,8463921	M	US	149	5,0106353	256	0,5820313	2817	7,9437827	7924	0,3555023
11	irrelevant	USM1	0,202227	0,687058	8	11,64385	M	US	262	5,572154	286	0,9160839	5122	8,5414955	19870	0,2577755
12	irrelevant	USF2	0,2005	0,7996	8	10,005003	F	US	262	5,572154	286	0,9160839	5122	8,5414955	19870	0,2577755
13	irradiated	USF8	0,199268	0,794691	9	11,325157	F	US	6	1,9459101			190	5,2522734		
14	irreverence	USM10	0,197045	0,816721	8	9,7952667	M	US	9	2,3025851	68	0,1323529	237	5,4722707	1759	0,1347356
15	irrelevance	USM10	0,195866	0,846899	8	9,4462268	M	US	0	0	82	0	428	6,0614569	4985	0,0858576
16	irrelevant	USM11	0,1956	0,7215	8	11,088011	M	US	262	5,572154	286	0,9160839	5122	8,5414955	19870	0,2577755
17	irreverence	USF5	0,194988	0,784668	8	10,195395	F	US	9	2,3025851	68	0,1323529	237	5,4722707	1759	0,1347356
18	irrevocably	USM17	0,19474	0,814516	10	12,27723	M	US	17	2,8903718	0	#DIV/0!	597	6,3935908	1	597
19	irrelevant	USF9	0,1942	0,9372	8	8,5360649	F	US	262	5,572154	286	0,9160839	5122	8,5414955	19870	0,2577755
20	irrational	USM17	0,193642	0,686147	6	8,7444819	M	US	149	5,0106353	256	0,5820313	2817	7,9437827	7924	0,3555023
21	irrelevant	USF8	0,19	0,592	8	13,513514	F	US	262	5,572154	286	0,9160839	5122	8,5414955	19870	0,2577755
22	irrational	USF24	0,189932	0,792912	6	7,567044	F	US	149	5,0106353	256	0,5820313	2817	7,9437827	7924	0,3555023
23	irrelevance	USM13	0,189507	0,885542	8	9,0340153	M	US	0	0	82	0	428	6,0614569	4985	0,0858576
24	irradiation	UKM2	0,186817	0,835088	8	9,5798287	M	UK	0	0			524	6,2633983		
25	irradiated	USM24	0,186456	0,865835	9	10,39459	M	US	6	1,9459101			190	5,2522734		
26	irradiate	USM15	0,18549	0,819237	7	8,544536	M	US	6	1,9459101			236	5,4680601		
27	irreverence	USM4	0,1849	0,747015	8	10,70929	M	US	9	2,3025851	68	0,1323529	237	5,4722707	1759	0,1347356
28	irrevocably	USM19	0,18412	0,800557	10	12,491303	M	US	17	2,8903718	0	#DIV/0!	597	6,3935908	1	597
29	irrelevance	USM21	0,183821	0,945732	8	8,4590561	M	US	0	0	82	0	428	6,0614569	4985	0,0858576
30	irrelevant	USF12	0,1816	0,726	8	11,019284	F	US	262	5,572154	286	0,9160839	5122	8,5414955	19870	0,2577755
31	irreverence	USM7	0,17713	0,648587	8	12,334506	M	US	9	2,3025851	68	0,1323529	237	5,4722707	1759	0,1347356
32	irrefutable	USM4	0,175828	0,718231	10	13,923097	M	US	13	2,6390573	0	#DIV/0!	394	5,9788858	13	30,307692
33	irregular	USM14	0,17493	0,677536	8	11,807491	M	US	129	4,8675345	1727	0,074696	2972	7,9973268	38143	0,0779173

UNIVERSITÉ Clermont Auvergne

Résultats

Analyse statistique

Une analyse de régression linéaire a été conduite pour chaque variété

Toutes les variables ont été testées et seules celles qui se sont révélées significatives ont été préservées

Par exemple, formule utilisée pour l'analyse des données US :

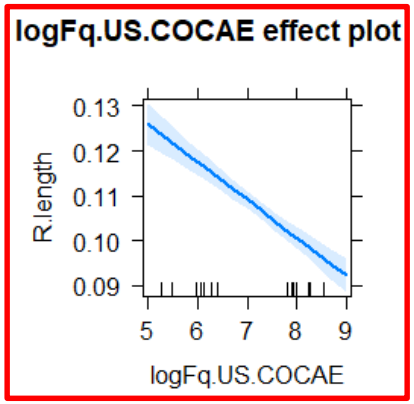
```
TestUS <- lm(R.length ~ logFq.US.COCAE + SpeechRate + Gender + Syll2 + SemTrans, data = donneesUS)
```

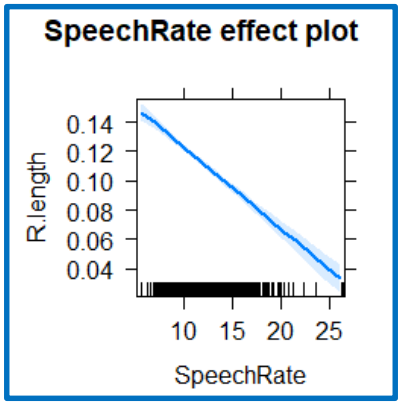
Les résultats sont présentés avec des graphiques représentant les effets des différentes variables

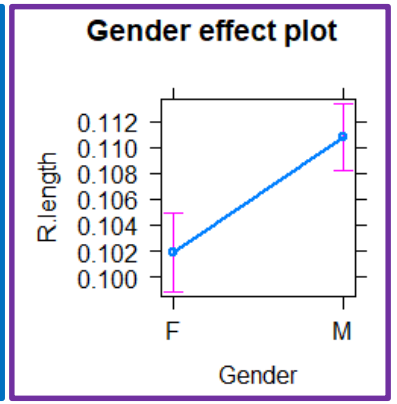
Résultats

Anglais américain - corpus complet

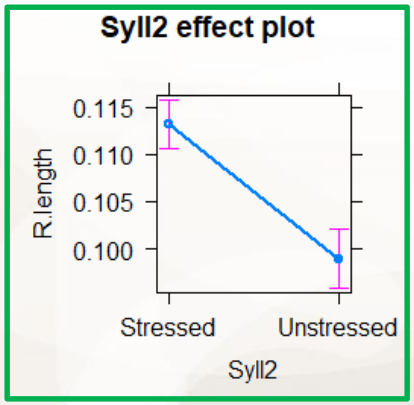
[ɹ] + long si la fréquence est basse

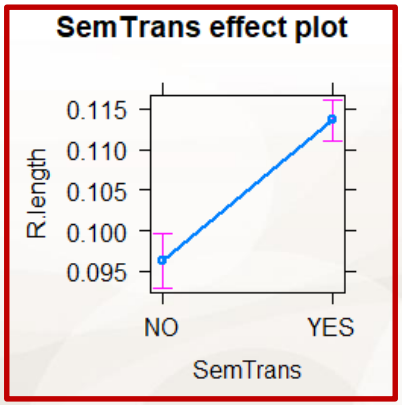






[ɹ] + long si Syll2 est accentuée





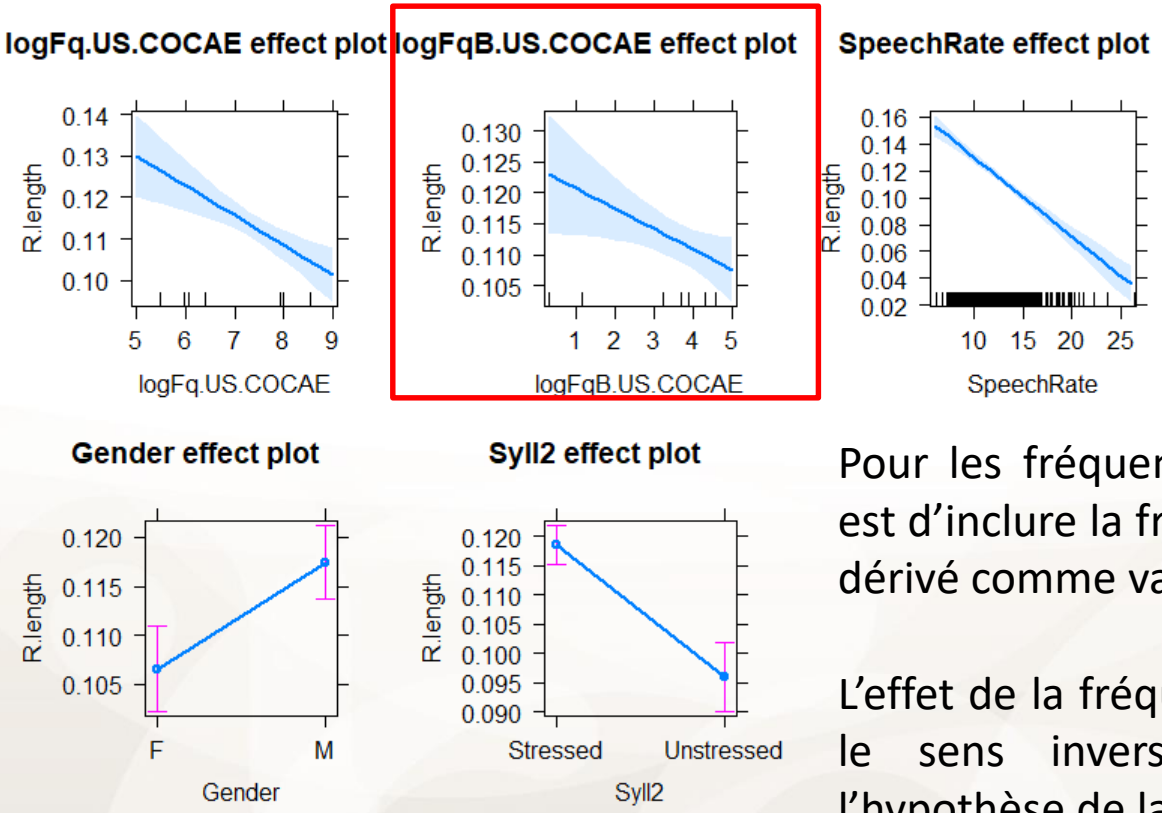
[ɹ] + long si la vitesse d'élocution est basse

[ɹ] + long si le locuteur est un homme

[ɹ] + long si le mot a un sémantisme transparent

Résultats

Anglais américain - mots dérivés

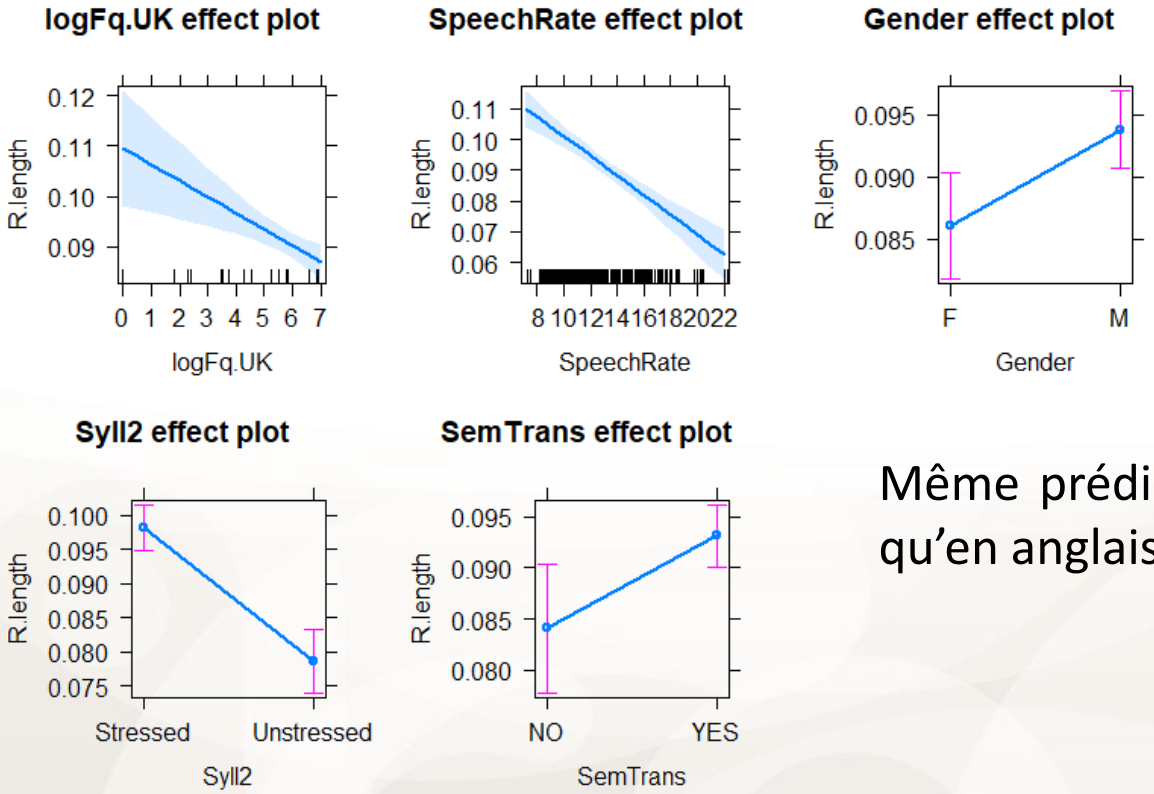


Pour les fréquences, le meilleur modèle est d'inclure la fréquence de la base et du dérivé comme variables indépendantes

L'effet de la fréquence de la base va dans le sens inverse de ce que prédit l'hypothèse de la segmentabilité !

Résultats

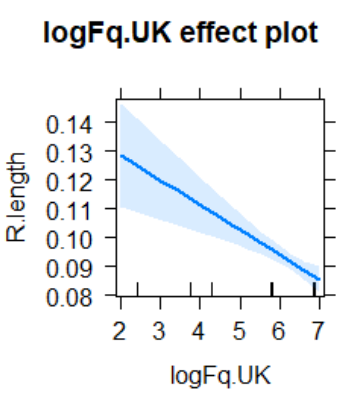
Anglais britannique - corpus complet

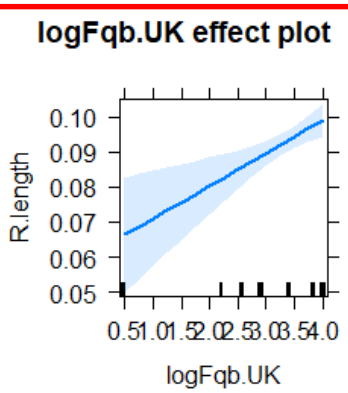


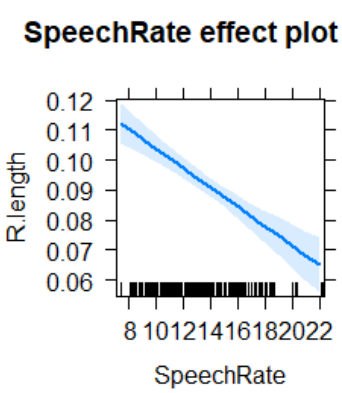
Même prédicteurs significatifs qu'en anglais américain

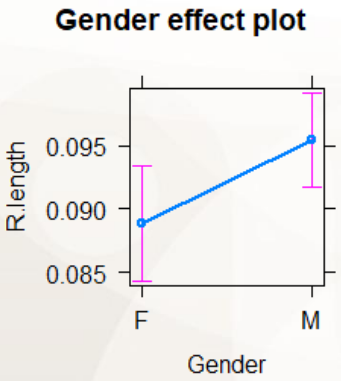
Résultats

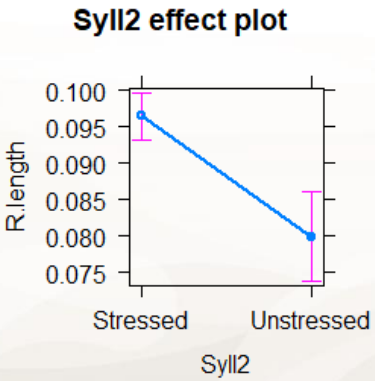
Anglais britannique - mots dérivés











Ici, l'effet de la fréquence de la base va dans le "bon" sens, celui conforme à l'hypothèse de la segmentabilité

Discussion

- Les mêmes variables sont significatives dans les deux variétés
- Une différence majeure : la direction de l'effet de la fréquence de la base
- Nos résultats vont dans le même sens que les études existantes portant sur d'autres consonnes que [ɹ]:
  - Durées plus courtes pour les mots fréquents
  - Durées plus longues si la consonne est suivie d'une syllable accentuée

Discussion

- Résultats contrastés sur l'hypothèse de la segmentabilité :
  - Effet significatif de la transparence sémantique
  - Résultats contradictoires sur la fréquence de la base entre les deux variétés
    - Ajouter de nouveaux mots pour augmenter le nombre de valeurs possible pour les fréquences (9 mots en cours d'ajout)

- Nouveau résultat: [ɻ] plus long pour les hommes que pour les femmes

Discussion

- Comparaison entre anglais britannique et anglais américain
  - Différence significative, [ɹ] plus long en anglais américain qu'en anglais britannique
  - On ne peut pas dire que la gémination est possible en anglais américain mais pas en anglais britannique
  - Qualitativement, on trouve des réalisations clairement géminées en anglais britannique

irrelevant, UKM23

Conclusion

L’analyse des données demande à être affinée mais, pour le moment, les résultats confirment l’existence d’une différence entre anglais britannique et américain

Dans les deux variétés, on trouve des effets de la fréquence absolue, de l’accentuation de la deuxième syllabe, du genre et de la transparence sémantique

Les résultats sont à ce stade inconclusifs sur les effets de la fréquence de la base

Merci de votre attention !

Références

Bauer, L. (2003) 'The Phonotactics of Some English Morphology', in Jacobsen, H. G. et al. (eds) *Take Danish for Instance. Linguistic studies in honour of Hans Basbøll presented on the occasion of his 60th birthday 12 July 2003.* Odense: University Press of Southern Denmark, pp. 1–8.

Ben Hedia, S. and Plag, I. (2017) 'Gemination and degemination in English prefixation: Phonetic evidence for morphological organization', *Journal of Phonetics*, 62, pp. 34–49.

Boersma, P. and Weenink, D. . (2018) 'Praat: doing phonetics by computer [Computer program]. Version 6.1.10'.

Cruttenden, A. (2014) *Gimson's Pronunciation of English*. 8th editio. Oxon & New York: Routledge.

Corpus of Contemporary American English [online]. URL: http://corpus.byu.edu/coca/

Dabouis, Q. (2016) *L'accent secondaire en anglais britannique contemporain*. Ph.D. dissertation. University of Tours.

Dabouis, Q. (to appear) 'English Phonology and the Literate Speaker: Some Implications for Lexical Stress', in Ballier, N. et al. (eds) *English Word Stress: Theories, Data and Variation*.

Hay, J. (2001) 'Lexical Frequency in Morphology: Is Everything Relative?', *Linguistics*, 28(6), pp. 1041–70.

Hay, J. (2003) *Causes and Consequences of Word Structure*. London: Routledge.

Van Heuven, W. V. J. *et al.* (2014) 'Subtlex-UK: A new and improved word frequency database for British English', *Quarterly Journal of Experimental Psychology*, (67), pp. 1176–1190.

Kaye, A. S. (2005) 'Gemination in English', *English Today*, 21(2), pp. 43–55.

Kenyon, J. S. and Knott, T. A. (1953) *A Pronouncing Dictionary of American English*. Springfield, MA: Merriam.

Oh, G. E. and Redford, M. A. (2012) 'The production and phonetic representation of fake geminates in English', *Journal of Phonetics*, 40(1), pp. 82–91.

Merriam-Webster [online]. URL: https://www.merriam-webster.com/

Navarro, S. (2016). *Le /r/ en anglais : histoire, phonologie et variation*. Dijon : Editions Universitaires de Dijon.

Upton, C. and Kretzschmar, W. A. (2017) *The Routledge Dictionary of Pronunciation for Current English*, *The Routledge Dictionary of Pronunciation for Current English*. Abingdon & New York: Routledge.

Videau, N. (2013) *Préfixation et phonologie de l'anglais : Analyse lexicographique, phonétique et acoustique*. Ph.D. dissertation. Université de Poitiers.

Wells, J. C. (2008) *Longman Pronunciation Dictionary*. 3rd ed. London: Longman.