Chapitre 3

Rythme & fluence

Nous avons vu dans le chapitre précédent que de nombreux facteurs, côté locuteur comme auditeur, impactent le degré d'effort requis pour comprendre le message. Parmi les facteurs côté locuteur, la fluence et le rythme de la parole sont deux éléments qui reviennent régulièrement dans les grilles d'évaluation de la production orale. Dans ce chapitre, nous proposons d'approfondir ces deux notions, et présenter en détails deux phénomènes linguistiques qui y sont étroitement liés : les pauses et l'accent lexicale.

3.1 Définitions

Avant de rentrer dans le vif du sujet, il nous paraît important de faire un point terminologique sur ces deux termes qui reviennent souvent dans la littérature, mais pour lesquels les définitions varient parfois selon les auteurs.

Commençons par le terme le plus fréquent – la fluence. Si le terme *fluency* en anglais fait souvent référence au niveau global de compétence en langue étrangère, nous nous intéressons ici à sa définition restreinte, plus commune dans le domaine de l'enseignement/apprentissage des langues étrangères, et qui concerne plus spécifiquement la production de parole (*speech fluency* ou *oral fluency*). Cette fluence est souvent interprétée comme le niveau d'automatisation et de contrôle du locuteur sur les processus cognitifs impliqués dans la planification et la production de la parole (Thomson, 2015). Segalowitz (2010) distingue 3 types de fluences : la fluence cognitive *(cognitive fluency)*, la fluence de phrase *(utterance fluency)* et la fluence perçue *(perceived fluency)*. La première correspond à la fluidité des processus cognitifs en amont de la production, la seconde à la fluidité de la parole produite, la troisième enfin à la perception de flui-

dité par l'auditeur. Lickley (2015) reprend les mêmes catégories mais les appelle les deux premières fluence de planification *(planning fluency)* et fluence de surface *(surface fluency)*. Les trois catégories sont étroitement liées, mais une disfluence dans l'une n'entraîne pas nécessairement une disfluence dans les autres. La plus importante pour la réussite de la communication est la troisième, mais l'évaluer de manière systématique n'est possible que sur la deuxième, tandis que remédier au problème n'est envisageable qu'en agissant sur la première. C'est le plus souvent la fluence perçue qui est évaluée, de manière intuitive et holistique, mais certains auteurs tentent de mesurer directement la fluence de surface, à partir de l'analyse du signal de parole. Les critères les plus souvent utilisés dans ce cas sont le débit de parole ou d'articulation (nombre de syllabes par seconde ou minute, avec ou sans pauses), le ratio de phonation (temps de parole sans pause divisé par temps de parole total), le nombre moyen de mots ou de syllabes par segments entre pauses, le nombre de pauses par seconde ou par minute ou encore leur durée moyenne (Thomson, 2015). On constate que la présence ou l'absence de pauses et leur durée semblent être fortement liées à la notion de fluence. En effet, selon Derwing et Munro (2015), la fluence se caractérise principalement par la présence de pauses ou d'autres marqueurs de disfluence tels que les faux départs ou les répétitions. Ainsi on a souvent tendance à considérer les pauses et autres interruptions du flux de parole comme des disfluences – ou « dysfluences » en contexte pathologique (Kernou, 2022) – mais nous allons voir que les pauses sont loin d'être nécessairement problématiques.

Le deuxième terme qui nous intéresse est le rythme. Là encore, de nombreuses définitions co-existent, mais la plupart des auteurs semblent s'accorder sur le fait qu'il fait référence à la façon dont se succèdent des éléments forts et des éléments faibles le long d'un axe temporel. Gibbon et Gut (2001) le définissent par exemple comme la récurrence de patterns temporels perceptibles de valeurs plus ou moins marquées d'un paramètre à travers le temps[1]. Di Cristo et Hirst (1997) le définissent comme l'organisation temporelle des proéminences. On parle aussi parfois de tempo (Frost & Picavet, 2014). Maintenant, il est légitime de discuter l'applicabilité de ce concept à une langue humaine naturelle, a priori beaucoup moins prédictible et régulière que ne peut l'être la musique par exemple (encore que). Toujours est-il que ce terme est largement utilisé pour qualifier un on-ne-sait-trop-quoi qui rend la parole plus « naturelle », mais surtout et comme nous allons le voir, plus facile à comprendre. Si la notion de fluence de la parole renvoie souvent aux patterns de pauses, celle du rythme renvoie principalement à l'accentuation des syllabes, et en particulier au phénomène d'accentuation lexicale.

[1] "*Rhythm is the recurrence of a perceivable temporal patterning of strongly marked (focal) values and weakly marked (non-focal) values of some parameter as constituents of a tendentially constant temporal domain (environment).*" (Gibbon & Gut, 2001, p. 95)

3.2 Les pauses

On appelle communément « pauses » les interruptions ponctuelles du flux de parole du locuteur. Ces interruptions sont le résultat complexe d'un compromis entre des contraintes physiologiques, linguistiques et culturelles, et n'ont pas toutes le même impact sur l'auditeur. Contrairement à la ponctuation dans un texte écrit, les pauses n'interviennent pas nécessairement pour structurer l'énoncé ; et leur position – bien que contrainte – est plus variable et semble dépendre de plus nombreux facteurs. Les interruptions du flux de parole peuvent être acoustiques (silences), ou linguistiques (allongements, interjections, mots de remplissage etc.), elles peuvent être physiquement présentes (pauses objectives) ou parfois seulement perçues par l'auditeur (pauses subjectives), et peuvent plus ou moins l'aider ou le perturber dans la compréhension du message.

Nous tenterons dans un premier temps de lister les différents types de pauses recensés, ainsi que leurs rôles. Nous décrirons ensuite les caractéristiques physiques de ces pauses ainsi que les contraintes syntaxiques auxquelles elles sont soumises, avant de nous intéresser à leur impact sur la perception de fluence et de compréhensibilité.

3.2.1 Types et rôles des pauses

Il existe probablement autant de typologies de pauses que d'auteurs ayant écrit à leur sujet. Certains les catégorisent selon leurs fonctions, d'autres selon leurs caractéristiques physiques, d'autres encore selon leur impact sur l'auditeur. Di Cristo (2013) identifie 6 types de pauses : les pauses respiratoires, les pauses structurales, les pauses pragmatiques, les pauses d'hésitation, les pauses aléatoires et les pauses phonostylistiques. Si les pauses respiratoires sont a priori issues de contraintes de bas-niveau, elles ont toutefois tendance à éviter de perturber la cohérence grammaticale et sémantique du discours – une pause respiratoire ne peut donc pas survenir n'importe où dans l'énoncé et sera contrainte par sa structure syntaxique. On peut regrouper les autres types de pauses en deux catégories : les pauses volontaires et les pauses involontaires. Les pauses structurales, qui ont pour objectif de délimiter les groupes syntaxiques de l'énoncé, et les pauses pragmatiques, qui ont un rôle plutôt rhétorique, sont en principe plutôt volontaires et planifiées par le locuteur. Les pauses d'hésitation, engendrées par la recherche lexicale ou la planification du discours, et les pauses aléatoires, causées par des troubles du langage, sont a priori plutôt involontaires et viendront potentiellement perturber la compréhension du discours. Enfin, les pauses phonostylistiques caractérisent le style de parole ou celui du locuteur, elles sont plus ou moins volontaires, et peuvent être plus ou moins perturbantes. Candea (2000) propose une classification binaire plutôt tournée vers l'impact sur auditeur : elle oppose les pauses

structurantes, à fonction de segmentation de la parole, aux pauses non-structurantes, à fonction d'hésitation. Dodane et Hirsch (2018) considèrent quant à eux les pauses en contexte de conversation, et distinguent d'abord les pauses inter-tours, pour la gestion du dialogue, et les pauses intra-tours, comprenant des pauses tantôt dues à des mécanismes physiologiques (déglutition, respiration), tantôt à l'organisation structurelle du discours (délimitation des unités de sens, mise en relief d'informations), ou enfin à la planification de l'énoncé (recherche lexicale, élaboration mentale).

Les pauses inter-tours interviennent comme leur nom l'indique entre les tours de parole des locuteurs. Elles sont souvent rapidement écartées des analyses, soit parce que les corpus analysés sont des monologues, soit parce qu'on ne les considère pas comme relevant de la fluence du locuteur. Or, il arrive que des pauses intra-tours soient utilisées par l'interlocuteur comme une opportunité de prendre la parole, et il s'avère que leur utilisation est contrainte par de nombreux facteurs linguistiques et culturels. Selon Fox et al. (1996), en anglais, les auditeurs sont capables de prédire avec précision quand un énoncé en construction va se terminer. Ils peuvent ainsi planifier leur énoncer et prendre leur tour de parole précisément à un moment de fin possible *(possible completion point)* sans laisser de pause entre les deux tours de parole. D'après Fox et al. (1996) et Sacks (1992), du point de vue d'un anglophone natif, les pauses sont souvent considérées comme un moment de malaise qui perturbe la conversation, et qui incite l'interlocuteur à prendre la parole. De ce fait, de nombreux outils existent pour éviter de se faire prendre la parole, généralement en remplissant tout moment de silence possible ("eh", "yeah", "well", "you know" etc., Sacks, 1992). Suivant ce raisonnement, une pause vide en anglais peut avoir tendance à être considérée comme un manque de contrôle sur la conversation par le locuteur.

Il en va autrement en japonais. Selon Shigemitsu (2007), la syntaxe du japonais permet difficilement de prédire la fin de l'énoncé ; c'est une pause à la fin de celui-ci qui indique à l'interlocuteur qu'il peut prendre la parole. En outre, les locuteurs japonophones ont tendance à séparer par des pauses des segments de mots assez courts entre lesquels les interlocuteurs ont la possibilité de réagir, sans pour autant prendre le tour de parole *(backchannel*[2]*)*. Maynard (1989) appelle ces segments courts entre pauses des *Pause-bounded Phrasal Units* (PPU), caractéristiques par leur courte durée (2,36 mots en moyenne en japonais d'après son étude). Les pauses séparant les PPU servent au locuteur à s'assurer que l'interlocuteur comprend le message, car celui-ci aura tendance à ne pas demander explicitement de clarification lorsqu'il ne comprend pas ; au contraire, il va plutôt avoir tendance à attendre que le locuteur apporte des informations complémentaires par lui-même. Ce type de pauses en japonais a donc, à

[2]Le backchannel, ou 相槌 *aizuchi* en japonais, correspond à l'utilisation fréquente d'interjections dans une conversation pour indiquer que le locuteur est écouté. C'est un phénomène particulièrement courant en japonais, mais qui existe aussi dans une moindre mesure en anglais et en français (White, 1989).

l'opposé de l'anglais, un rôle d'encouragement du locuteur à poursuivre son discours.

Par ailleurs, la dynamique des tours de parole en japonais est fortement influencée par la relation sociale entre les locuteurs : l'un des participants de la conversation détient généralement le contrôle de la dynamique de conversation – il a le *speakership* – et la prise de parole inattendue de l'un des autres participants peut provoquer un malaise. On parle aussi de *pauses de politesse*, qui sont attendues de la part de certains locuteurs vis-à-vis de certains autres, et interprétées de manière différente en fonction du statut conversationnel du locuteur (Shigemitsu, 2007). Une prise de parole en japonais est donc à la fois régie par la syntaxe, mais également par les contraintes sociales entre les locuteurs.

Shigemitsu (2007) s'intéresse à l'effet que peut avoir l'utilisation de stratégies pausales culturellement différentes dans une conversation entre des locuteurs de langue maternelle différente. Elle analyse 4 conversations spontanées d'une trentaine de minutes en anglais et en japonais, entre 2 ou 4 locuteurs qui ne se connaissent pas. Dans chacune d'elles, la moitié des participants sont de langue maternelle japonaise, l'autre moitié de langue maternelle anglaise. Chaque conversation est suivie d'un entretien individuel avec les locuteurs, pour leur demander ce qu'ils ont ressenti pendant la conversation et s'ils se sont sentis à l'aise ou non. Seules les pauses silencieuses (interruption de phonation) sont considérées dans cette étude. Shigemitsu observe que l'utilisation de stratégies pausales japonaises en anglais, ou anglaises en japonaise, peut considérablement impacter la réussite de la conversation. Dans les conversations en anglais qualifiées de moins réussies par les participants, elle observe que les pauses sont rares et très courtes, empêchant les participants japonophones d'y placer une réaction, ou trop courtes pour qu'ils la considèrent comme un moment de prise de parole potentielle. Les participants anglophones ont eu tendance à remplir chaque moment de silence, jusqu'à ceux des locuteurs japonophones, qui l'ont souvent interprété comme une coupure de parole. Par ailleurs, si les anglophones considéraient important que tout le monde parle autant, certains locuteurs japonais se satisfaisaient de participer sans pour autant prendre la parole, et sans sentir de gêne vis-à-vis de cela. À l'inverse, les locuteurs anglophones ont perçu les participants japonais comme peu coopératifs et parfois impolis par leur manque de conversation et d'initiative de prise de parole, résultant pour certains en un sentiment de culpabilité de ne pas leur laisser le temps de parler. L'utilisation adéquate des pauses est donc clé pour mener à bien une conversation.

Les pauses peuvent ainsi avoir des causes et des objectifs variés. En outre, Grosjean et Deschamps (1975) suggèrent qu'une même pause peut porter plusieurs fonctions différentes en même temps, comme profiter d'une frontière syntaxique ou d'une hésitation pour respirer ou pour reformuler, il est donc important de ne pas lui attribuer un type exclusif. Par ailleurs, une pause peut avoir un objectif précis souhaité par

le locuteur, comme vérifier que l'interlocuteur comprend, mais être interprétée différemment par ce-dernier, comme un manque d'intérêt dans la conversation ou une invitation à prendre la parole.

3.2.2 Caractéristiques physiques

Plusieurs phénomènes dans la parole du locuteur peuvent être interprétés par l'auditeur comme des pauses. Le premier et le plus évident est l'interruption de la phonation, ou l'arrêt temporaire de production de parole. On parle dans ce cas de « pause silencieuse », et ce sont elles qui sont le plus largement analysées dans la littérature. La plupart des études s'accordent à fixer un seuil minimum de durée à partir duquel considérer une interruption de phonation comme une pause silencieuse, mais la valeur de ce seuil est très variable d'une étude à l'autre, comme le suggère le tableau 3.2.2. La revue d'une quarantaine d'études analysant les phénomènes de pauses dans la parole non pathologique nous montre qu'il varie entre 0 ms et 3 s, avec un grand nombre d'entre elles le fixant entre 100 ms et 300 ms. L'ensemble des études présentées ici traitent des phénomènes d'hésitation, d'organisation des pauses ou d'évaluation de la fluence en parole native ou L2, dans différentes langues.

De Jong et Bosker (2013) constatent qu'un seuil de 250 ms à 300 ms obtient la meilleure corrélation avec le niveau de compétence en langue des locuteurs non-natifs en néerlandais (déterminé par un test de vocabulaire), amenant de nombreuses études à fixer un seuil à 250 ms par la suite. Une autre justification souvent donnée pour ne pas considérer les silences inférieurs à 200 ms est le fait que les pauses plus courtes sont moins à même de refléter les difficultés linguistiques de construction du discours, mais semblent plutôt liées à des contraintes coarticulatoires de bas-niveau, qui ne sont généralement pas le sujet d'intérêt de ces études. En effet, les études qui considèrent des silences très courts ajoutent souvent un délai supplémentaire devant les consonnes occlusives (50 ms pour Fauth et Trouvain, 2018 ; Smiljanić et Bradlow, 2005), ou complètent la détection automatique des silences par une annotation manuelle (Matzinger et al., 2020).

Campione et Véronis (2002) mettent en garde sur le fait que le choix du seuil minimal de durée peut largement impacter les conclusions des analyses qui suivent. Ils observent notamment que la durée moyenne des pauses est plus courte en parole spontanée qu'en parole lue si on ne définit aucun seuil, mais qu'elle est plus longue si on ne considère que les pauses supérieures à 200 ms, et qu'elle est égale si on ajoute un seuil maximum à 2 s. Il devient ainsi pratiquement impossible de comparer les résultats obtenus par différentes études, si celles-ci choisissent des seuils différents.

Seuil minimum	**Sources**
Pas de seuil	Fauth et Trouvain, 2018 ; Maclay et Osgood, 1959 ; Wilkes et Kennedy, 1969
1 ms	Matzinger et al., 2020
5 ms	Owoicho et al., 2024 ; Smiljanić et Bradlow, 2005
20 ms	Cucchiarini et al., 2000 ; Kirsner et al., 2005
60 ms	Campione et Véronis, 2002
80 ms	Levin et al., 1967
100 ms	Butcher, 1981 ; Kang et Johnson, 2018 ; Lounsbury, 1954 ; Trouvain, 2004
200 ms	Candea, 2000 ; Cucchiarini et al., 2002 ; Fletcher, 1987 ; Goldman-Eisler, 1968 ; Grosjean, 1980 ; Kahng, 2014 ; Lennon, 1990 ; Zellner, 1994
250 ms	De Jong et Bosker, 2013 ; de Jong, 2016 ; Grosjean et Deschamps, 1975 ; Kahng, 2018 ; Kallio et al., 2022 ; Shea et Leonard, 2019 ; Suzuki et al., 2021 ; Witton-Davies, 2018
300 ms	Grosjean et Deschamps, 1972 ; Lacheret-Dujour et Victorri, 2002
400 ms	Tavakoli, 2010
1 s.	Lay et Paivio, 1969 ; Levin et Silverman, 1965
2 s.	Siegman et Feldstein, 1979
3 s.	Siegman et Pope, 1966

TAB. 3.1 : Seuils de durée minimum de pause utilisés dans la littérature

À travers une analyse de corpus écrit multilingue[3], ils observent que la distribution des durées de pauses suit une distribution logarithmique multimodale et non une loi arithmétique et normale comme il est couramment admis jusqu'alors. Ils identifient deux gaussiennes autour de 150 ms et 500 ms, quelque soit la langue. Ces deux gaussiennes sont observées dans des études ultérieures et semblent relativement stables. Kirsner et al. (2005) vont jusqu'à faire l'hypothèse que la première catégorie (pauses courtes, 50 ms à 70 ms) est due aux processus d'articulation, tandis que la deuxième (pauses longues, 500 ms à 700 ms) l'est plutôt à la structuration du discours. Demol et al. (2007) identifie également ces deux gaussiennes et constatent qu'elles ne sont liées ni à la langue ni au débit de parole[4]. Enfin, Goldman et al. (2010) analysent un corpus de 40 min de français de différentes situations de communication[5], et constatent que le nombre de gaussiennes fluctue entre 1 et 3 en fonction des situations, mais étant majoritairement bimodal.

Du côté de la parole spontanée, Campione et Véronis (2002) observent toujours deux gaussiennes (autour de 80 et 430 ms), accompagnée d'une troisième autour de 1500 ms. Leur corpus est constitué d'entretiens d'une quinzaine de minutes avec 10 locuteurs, issus du Corpus Français Oral de Référence. Les auteurs en viennent à proposer la catégorisation des durées de pauses suivante : pauses brèves ($<$200 ms), pauses moyennes (entre 200 ms et 1 s) et pauses longues ($>$1 s). Cette catégorisation sera souvent citée par la suite, mais semble peu utilisée dans les faits – la plupart des auteurs préférant fixer un seuil fixe et unique.

Si l'on part du principe qu'une pause est un phénomène perceptif, il semble peu pertinent de déterminer un seuil absolu de durée qui ne tienne pas compte du contexte (présence d'hésitations, longueur des segments) ou du débit de parole du locuteur. Certaines études choisissent ainsi de ne pas fixer de seuil mais plutôt de se fier à la perception d'annotateurs humains, amenant parfois tout de même à des pauses inférieures à 100 ms (Fauth & Trouvain, 2018). De rares études font état d'un seuil relatif au débit de parole du locuteur, variant entre 180 ms et 250 ms chez Duez (1982, 1991) (calculé à partir de la durée moyenne des occlusives intervocaliques), de 98 ms à 490 ms chez Kirsner et al. (2003) (calculé à partir de la distribution des durées de pauses par locuteur), ou de 138 ms à 384 ms chez De Jong et Bosker (2013) (calculé à partir du débit d'articulation de chaque enregistrement). De son côté, Zellner (1994) montre que le seuil de perception des pauses varie en fonction du segment précédent, et Duez (1993) constate que certaines pauses sont perçues même sans interruption de phonation – elle les appelle "pause subjectives".

[3] Campione et Véronis (2002) analysent l'anglais, le français, l'allemand, l'italien et l'espagnol dans le corpus *Eurom*.

[4] Demol et al. (2007) analysent l'anglais, le français, l'italien, l'espagnol, le roumain et le néerlandais.

[5] Lecture, narration conversationnelle, journal télévisé et conférence scientifique.

Grosman et al. (2018) identifient également 2 gaussiennes dans la distribution des durées de pauses du corpus LOCAS-F[6], toutefois, ils remarquent que cette bimodalité ne se retrouve pas nécessairement dans toutes les situations de parole et pour tous les locuteurs : le journal radiophonique et les conférences scientifiques semblent relativement standardisés avec une distribution bimodale similaire pour tous les locuteurs ; celles-ci sont plus hétérogènes en discours politique et présentent une bimodalité pour 3 locuteurs sur 5, tandis que les récits conversationnels et les homélies sont plutôt unimodaux. Quant à la durée médiane des pauses, elle varie de 289 ms à 518 ms selon les situations mais également largement à l'intérieur de celles-ci. Les auteurs conseillent de considérer la distribution des durées de pause en fonction des situations de parole, voire en fonction des locuteurs. Ils ne préconisent pas de définir de seuil de durée fixe pour exclure certaines données, et exclure seulement les mesures aberrantes (ils n'ont ainsi supprimé que 4 % des pauses de leur corpus).

- durée arithmétique ou logarithmique ? GrosmanAl2018 utilisent $\log_{10}$ ms. Kahng2018 log-transforme les durée moyennes de pauses, mais aussi les fréquences de pauses intra-proposition et inter-proposition pour approximer une distribution normale. SheaLeonard2019 transforment toutes leurs distributions qui ne sont pas normales avec log et square root transformations. Mais ChristodoulidesAl2017 critique l'utilisation des transformations logarithmiques → est-ce que ça change vraiment quelque chose ?

Les pauses ne se limitent toutefois pas aux phénomènes d'interruption de phonation. On parle de « pauses pleines » lorsque qu'il n'y a pas d'interruption de phonation (allongements, « heu » et autres interjections. Certains auteurs, comme Fauth et Trouvain (2018), considèrent même les faux-départs, les répétions ou les reformulations comme pauses pleines : tout ce qui, en somme, interrompt le flux du discours.

3.2.3 Pauses et localisation syntaxique

De nombreuses études montrent que la fréquence et la durée des pauses sont corrélées avec la position de celles-ci dans l'énoncé, et en particulier avec le type de frontière syntaxique où elles se trouvent. Lorsque la position d'une pause est inattendue, on parle souvent de pause non-structurante (Candea, 2000), de pause agrammaticale ou encore disfluente (Fauth & Trouvain, 2018).

[6] *LOuvain Corpus of Annotated Speech-French* (L. Martin et al., 2014). Durée : 3 h38 min, 76 locuteurs belges, français et suisses, en situation de monologue, dialogue, ou multilogues. 14 situations de communication différentes comprenant conférences scientifique, débats et discours politiques et académiques, interactions formelles et informelles, interviews, journaux radiophoniques, lectures radiophonique.

Tauberer (2008) utilise les informations de catégories grammaticales des mots et la structure syntaxique de l'énoncé pour prédire la position et la durée des pauses en anglais spontané dans le corpus de conversations téléphoniques Switchboard. Il observe que les pauses ont tendance à apparaître autours des conjonctions, des compléments, ou avant les pronoms ou les sujets. En revanche, elles sont beaucoup plus rares après les sujets, entre les verbes et les syntagmes prépositionnels, ou entre les prépositions et les syntagmes nominaux. Tauberer teste différentes combinaisons entre 12 paramètres[7] pour obtenir la meilleure prédiction. D'après ses résultats, l'analyse structurale par constituants a un plus grand pouvoir prédictif que l'analyse lexicale seule, mais la simple information de durée du constituant précédent combinée au nombre de mots du constituant suivant prédit à peu près aussi bien que l'ensemble des paramètres combinés (F-score de 78,2% contre 78,5% avec tous les paramètres).

Cao et Chen (2019) s'intéressent quant à eux aux caractéristiques de la parole préparée de ceux qu'ils appellent des "*successful speakers*" : 15 locuteurs anglophones natifs et non-natifs enregistrés lors de discours politiques, de Ted Talks, ou dans des vidéos à succès sur les réseaux sociaux. Ils constatent que les pauses sont souvent placées avant les conjonctions de subordination (exemple : *« we must never forget // **that** those heroes // **who** fought against evil // also fought for // the nations // **that** they loved »*, p. 2050), et plus généralement à la frontière syntaxique entre deux proposition (*« if it is not available in your area // you can also use ham instead »*, p. 2050), et ce sans différence perceptible entre les locuteurs natifs et non natifs.

Dans une analyse de la position des pauses et des marqueurs d'hésitation dans des récits produits par des élèves de 4ème en classe de français, Candea (2000) catégorise les pauses en « structurantes » (lorsqu'elles sont non immédiatement précédées par un marqueur d'hésitation) et « non-structurantes » (lorsqu'elles sont immédiatement précédées par un marqueur d'hésitation). Selon sa définition, elle note que 82,5% des pauses sont structurantes. Parmi elles, 78% sont placées en fin d'énoncé ou de proposition syntaxique, tandis que 19% seulement se trouvent en fin de syntagme (qu'elle appelle constituant syntaxique), et 3% à l'intérieur d'un syntagme. Dans un corpus plus long et diversifié en situations de parole (LOCAS-F), Grosman et al. (2018) font des observations similaires : 78% des pauses sont structurantes (selon la même définition que Candea). Toutes pauses confondues, 36% d'entre elles sont en fin de proposition (qu'ils appellent unité de rection, constituée d'un verbe accompagné de ses dépendants), 11% entre ce qu'ils appellent séquences syntaxiques, ou unités syntaxiques intermédiaires, et 9% à l'intérieur des groupes accentuables, leurs unités syntaxiques minimales, qui correspondent à la combinaison d'un mot lexical et des mots gram-

[7] Catégorie du mot précédent, du mot suivant, et combinaison des deux ; catégorie du constituant le plus grand se terminant, se commençant, et combinaison des deux ; nombre de mots et durée du constituant le plus grand se terminant, et commençant ; profondeur syntaxique ; et temps de fin du mot précédent calculé depuis le début de l'énoncé et relatif sa longueur totale.

maticaux qui en dépendent (Mertens, 2008), soit une unité légèrement plus petite que le syntagme. Les 44% restants se situent entre des groupes accentuables. D'après leurs observations, la majorité des pauses surviennent entre les unités syntaxiques, et rarement à l'intérieur des groupes accentuables (ci-après ga). Par ailleurs, plus la frontière syntaxique est grande, plus la pause est longue. Les auteurs observent également que la parole spontanée est caractérisée par plus de pauses intra-ga, mais aussi plus de pauses entre les unités syntaxiques maximales[8]. Les pauses inter-séquence syntaxique semblent quant à elles plus fréquentes en parole préparée.

En ce qui concerne la durée des pauses, Candea (2000) observe que les pauses sont significativement plus longues en fin d'énoncé, qu'en fin de proposition, et qu'en fin de syntagme. C'est également ce que constatent Grosman et al. (2018) : plus la frontière syntaxique est grande, plus la pause est longue (intra-ga $<$ inter-ga $<$ inter-séquence syntaxique $<$ inter-unité de rection), quelque soit la situation de parole. Ajoutons que, bien que significativement corrélée, la durée de la pause ne dépend pas que de sa position, mais peut être également influencée par la longueur des constituants la précédant ou la suivant par exemple (Krivokapić, 2007).

Grosjean et Deschamps (1975) comparent quant à eux la distribution des pauses en français et en anglais dans des interviews radiophoniques. Ils fixent un seuil de durée minimale de pause à 250 ms, et considèrent 7 positions de pauses possibles : soit en fin de proposition (qu'ils appellent phrases, combinant un syntagme nominal (SN) et un syntagme verbal (SV), éventuellement accompagné de compléments), soit à l'intérieur d'une proposition, entre ou à l'intérieur des syntagmes. D'après leurs observations, les locuteurs français ont tendance à faire plus de pauses en fin de proposition (60%) que les locuteurs anglais (55 %, $p < 0,05$), mais surtout moins de pauses à l'intérieur d'un syntagme SN ou SV (16 % contre 26 %, $p < 0,001$). La différence se joue surtout au niveau du SV, où les anglophones font 14 % plus de pauses que les francophones, tandis qu'ils en font 5 % moins à l'intérieur du SN. De plus les anglophones semblent répartir les pauses plus librement à l'intérieur du SV, avec une préférence devant le complément prépositionnel (45%), alors que les francophones les placent majoritairement entre le verbe et son objet (70%). Il semble donc y avoir des différences de tendance dans la distribution des pauses en français et en anglais, du moins en parole radiophonique, dans les années 70.

Qu'en est-il pour les locuteurs non-natifs ? Dans une analyse de la distribution des pauses en parole lue en français, Fauth et Trouvain (2018) observent que les lecteurs non-natifs font plus de pauses à l'intérieur des énoncés que les lecteurs natifs, et les débutants plus que les avancés. Le premier groupe est constitué de 20 lecteurs germanophones lisant à haute voix un texte en français des Trois Petits Cochons issu du

[8] Les auteurs expliquent cette observation par le fait que les propositions sont plus courtes en parole spontanée.

corpus IFCASL (Trouvain et al., 2016). Dix d'entre eux ont un niveau A2-B1, les dix autres un niveau B2-C1. Dix autres locuteurs francophones natifs sont également enregistrés pour comparaison. Ils constatent par ailleurs que les lecteurs non-natifs font plus de pauses et des pauses plus longues en général, et plus encore pour les lecteurs débutants, mais sans toutefois observer de différence significative entre les niveaux.

de Jong (2016) observe également que les locuteurs non-natifs (dans son cas, anglophones et turcophones) ont tendance à faire plus de pauses à l'intérieur des énoncés[9] que les locuteurs natifs en néerlandais. Elle observe aussi une corrélation avec le niveau du locuteur : plus celui-ci a un niveau élevé, moins il fait de pauses intra-énoncé. En outre, la fréquence des pauses entre les énoncés semble indépendante de la langue maternelle du locuteur et de son niveau de compétence. Des résultats similaires sont observés par Kahng (2014) et Shea et Leonard (2019).

3.2.4 Perception des pauses

Les pauses sont perçues différemment selon leur position et leur nature. Duez (1985) montre par exemple que les pauses en français sont mieux perçues lorsqu'elles sont situées entre deux propositions, qu'à l'intérieur de l'une d'elles. Cette observation est également confirmée par Collard (2009) et Lickley (1995). Candea (2000) et Duez (1995) remarquent par ailleurs que les pauses qu'ils catégorisent comme « non-structurantes » (immédiatement précédées d'une hésitation) n'occasionnent presque jamais un changement de tour de parole : elles ne sont pas perçues comme des indices de coupe par les auditeurs. Mieux encore, lorsque J. Martin et Strange (1968) demandent à 129 étudiants anglophones natifs de répéter un énoncé spontané avec ses hésitations, ou de le transcrire avec ses hésitations, ils constatent que les hésitations intra-constituants sont systématiquement déplacées en frontière de constituant. Simon et Christodoulides (2016) proposent une expérience intéressante où ils demandent à des auditeurs naïfs d'annoter en temps réel des échantillons de parole francophone de genres variés, en signalant chaque fois qu'ils perçoivent la fin d'un groupe de mots. Les résultats montrent que la simple complétude syntaxique provoque la perception d'une frontière même sans autre indice acoustique. La syntaxe semble donc jouer un rôle important sur la perception et la tolérance des pauses.

Par ailleurs, si Bard et Lickley (1997) observent que les auditeurs peinent à se souvenir des éléments disfluents dans la parole au profit du contenu du message, ils peuvent aussi avoir tendance à mieux retenir les informations lorsqu'elles sont précédées d'une hésitation (Fox Tree, 2001). Corley et al. (2007) et MacGregor (2008)

[9] de Jong (2016) les appelle des « unités de paroles » *(speech units)*, constituée d'une proposition indépendante et de ses subordonnées éventuelles.

constatent par exemple que la présence d'une pause (pleine ou silencieuse) à l'intérieur d'un énoncé augmente la probabilité que le locuteur se souvienne du mot qui suit. Lundholm Fors (2015) fait le même constat, et ajoute que les pauses inférieures à 500 ms semblent avoir un meilleur impacte que les pauses plus longues.

Les pauses semblent donc mieux perçues et acceptées aux frontières syntaxiques qu'à l'intérieur des constituants. Même pleines, elles peuvent avoir un effet positif sur l'auditeur, en augmentant ponctuellement son niveau d'attention et en facilitant la mémorisation du message.

3.2.5 Pauses et évaluation de la fluence

Shea et Leonard (2019) font une revue approfondie des mesures relatives aux pauses utilisées pour l'évaluation de la parole L2. La plupart des études mesurent des fréquences générales de pauses : nombre de pauses par minute, par mot, par syllabe, par proposition ou par énoncé (généralement défini comme une proposition principale avec ses relatives), ou encore par tour de parole. La durée des pauses est généralement considérée à travers des ratios de durée totale de pause par rapport au temps de parole, ou à l'inverse, la durée de phonation par rapport au temps de parole. Les mesures qui prennent en compte la position des pauses sont plus rares, et considèrent généralement celles-ci vis-à-vis de la frontière des propositions *mid-clause vs. end-of-clause)*, ou plus largement de l'énoncé (proposition principale avec ses relatives). Il peut s'agir de fréquence de pause par type (par exemple, inter- ou intra-proposition), ou une durée moyenne, ou encore le nombre d'énoncés suivis d'une pause par exemple.

On peut donc constater que la majorité des études recourent à une fréquence globale ou une durée moyenne de pauses en général (Kahng, 2018 ; Saito et al., 2022). Ces deux paramètres apparaissent effectivement très corrélés avec le niveau global d'un apprenant, ces-derniers ayant tendance à faire plus de pauses et des pauses plus longues quand leur niveau est moins élevé. Toutefois, comme nous l'avons vu dans les sections précédentes, les pauses ne sont pas nécessairement un problème ; au contraire, lorsqu'elles sont bien placées, les pauses permettent une meilleure compréhensibilité (Cao & Chen, 2019 ; Isaacs et al., 2018). La question reste de savoir quelles pauses sont susceptibles d'être problématiques, et lesquelles le sont moins.

De récentes études se sont penchées sur la relation entre la distribution syntaxique des pauses et la perception de fluence ou de compréhensibilité. Kahng (2018), par exemple, recrute une cohorte de 46 évaluateurs et leur fait évaluer 80 extraits de paroles au moyen d'une échelle de Likert à neuf points (1=très disfluent, 9=très fluent). Les évaluateurs sont tous de langue maternelle anglaise et étudiants dans une université aux États-Unis ; les locuteurs sont de langue maternelle coréenne ($n = 37$, 74

extraits) et anglaise ($n = 3$, 6 extraits). Les extraits font environ 20 s et sont issus d'un enregistrement plus long dans lequel le locuteur répond à deux questions, sur sa spécialité à l'université et ses loisirs. En parallèle, tous les silences de plus de 250 ms ont été annotés et catégorisés en fonction de leur position dans l'énoncé : entre ou à l'intérieur des propositions ; le ratio pauses/minute, leur durée moyenne et le débit d'articulation par extrait sont également calculés. La durée moyenne, la fréquence des pauses inter- et intra-proposition sont log-transformées pour approximer une distribution normale. Au moyen d'une régression multiple par étapes, Kahng constate que la fréquence des pauses intra-proposition est le paramètre le plus corrélé avec le jugement de fluence, expliquant à lui seul plus de 54 % de sa variance. Combiné avec la fréquence des pauses inter-proposition, seuls 6 % supplémentaires de la variance sont expliqués, et ni la fréquence, ni la durée moyenne des pauses en général ne sont capable d'améliorer significativement ce modèle[10]. La distribution syntaxique des pauses semble donc jouer un rôle important dans la perception de la fluence.

Dans une seconde expérimentation, Kahng (2018) tente de vérifier l'impact des pauses sur le jugement de fluence en modifiant artificiellement une sélection de 24 extraits (L1=anglais) et 24 extraits (L1=coréen). Il propose 3 conditions : condition 1) les pauses sont supprimées (réduites à 150 ms) ; condition 2) à partir de ces extraits sans pauses, 5 pauses inter-proposition de 600 ms sont insérées ; condition 3), à partir de ces extraits sans pauses, 5 pauses intra-proposition de 600 ms sont insérées. Kahng fait alors évaluer les extraits ainsi modifiés selon le même protocole, à 92 locuteurs natifs de l'anglais, en veillant à ce qu'ils n'écoutent pas deux fois le même enregistrement original. En comparant les jugements par condition, il observe que les extraits avec pauses ajoutées sont jugés significativement moins fluents que les extraits sans pauses ($p < 0,001$), et que les extraits avec pauses intra-proposition sont jugés significativement moins fluents que les extraits avec pauses inter-proposition ($p = 0,048$).

Ces observations sont confirmées par d'autres études par la suite. Suzuki et Kormos (2020) font évaluer par 10 locuteurs anglophones natifs des enregistrements produits par 40 locuteurs japonophones de niveau A2 à C1. Il s'agit cette fois de parole argumentative, les locuteurs doivent donner leur avis sur un sujet d'actualité. L'évaluation est faite sur deux dimensions : la compréhensibilité et la fluence du locuteur, toujours via une échelle de 9 points. Parmi de nombreux paramètres couvrants la complexité et la précision de la réponse, la fluence, la prononciation ou la cohérence du discours, les auteurs observent que le débit de parole est le plus corrélé avec le jugement de compréhensibilité, tandis que le jugement de fluence est en particulier influencé par la fréquence de pauses inter-proposition. Dans une autre étude, Kallio et al. (2022) vont

[10] Notons que cela ne signifie pas que la fréquence et la durée moyenne des pauses n'expliquent pas une partie de la variance des jugements de fluence. Kahng note que la fréquence seule explique 31 % de la variance, et la fréquence et la durée moyenne en expliquent 43 %.

plus loin en étudiant l'impact de la position des pauses vis-à-vis du syntagme dans des extraits de 200 locuteurs non-natifs du finnois. Ils classent les pauses en 5 catégories : inter- et intra-proposition, inter- et intra-syntagme, et intra-mot (pauses intervenant à l'intérieur d'un mot non terminé). Une évaluation de la perception de fluence sur une échelle de 4 points et une évaluation du niveau global sur une échelle de 7 points est effectuée par deux évaluateurs parmi une cohorte de 16 évaluateurs certifiés par l'Agence Nationale de l'Éducation Finnoise. Comme dans les études précédentes, des modèles de régression multiples sont utilisés pour déterminer quels paramètres influencent le plus l'évaluation parmi. Les auteurs observent que la fréquence des pauses intra- et inter-syntagme sont les indicateurs les plus corrélés avec le jugement de niveau global ($R^2 = -4,96$ et $R^2 = -4,33$, $p < 0,001$) et la perception de fluence ($R^2 = -6,93$ et $R^2 = -5,33$, $p < 0,001$). La fréquence des pauses intra-mot est également un indicateur fort, suivi par celle des pauses inter-proposition ($R^2 = 2,23$, $p < 0,05$ pour la fluence, mais non significatif pour le niveau global).

3.3 L'accent lexical

L'accent tonique *(stress)* fait référence au degré de force utilisé pour produire une syllabe (Crystal, 2008). C'est un phénomène relatif, c'est à dire qu'une syllabe pourra être plus ou moins accentuée qu'une autre, mais sa valeur absolue n'a pas réellement d'intérêt. On distingue généralement trois catégories fonctionnelles : l'accent de mot, l'accent de phrase et l'accent contrastif (Frost, 2023). Nous nous concentrerons ici sur la première catégorie. Toutes les langues n'ont pas d'accent de mot *(word stress, word-level stress)* ; parmi celles qui en ont un, certaines ont un accent à position fixe (*fixed stress languages*, comme le finnois, le polonais ou le français, où il se place systématiquement sur la première, la pénultième ou la dernière syllabe, respectivement), d'autres ont un accent à position variable comme l'anglais, l'allemand ou l'espagnol (Cutler & Jesse, 2021). Lorsque la position de l'accent de mot est variable, on parle d'accent lexical *(lexical stress)* car il joue un rôle pour l'accès lexical : certains mots ne se distinguent que par sa position, comme ***differ*** et ***defer*** en anglais, ou ***aun*** et ***aún*** en espagnol.

Au delà de son intérêt sémantique, l'accent lexical – et plus généralement l'accent de mot – joue un rôle important pour aider l'auditeur à segmenter le flux de parole (Cutler, 2015).

Au niveau acoustico-phonétique, l'accentuation peut se caractériser par des variations au niveau suprasegmental (fréquence fondamentale (F_0), intensité et durée), mais également au niveau segmental (qualité vocalique). Toutefois ces 4 niveaux ne sont pas nécessairement exploités dans toutes les langues, et leur poids respectif peut varier. Ainsi, en espagnol, seuls les 3 niveaux suprasegmentaux sont en jeu, là où en

thaï, la F_0 est réservée pour le ton et ne participe pas à l'accent de mot (Cutler & Jesse, 2021). En anglais ou en allemand, en revanche, les 4 niveaux de variation peuvent être exploités pour marquer la syllabe accentuée. Les 3 niveaux suprasegmentaux apparaissent très corrélés entre eux, et si de nombreuses études ont analysé un seul niveau à la fois, ou encore tenté de hiérarchiser leur importance respective, il semble important de ne pas les dissocier complètement et adopter une approche intégrée de l'accentuation (Vaissière, 1983).

Comme pour le phénomène de pause que nous avons analysé dans la section précédente, il est important de différencier l'accent phonétique (physiquement présent et mesurable) de l'accent perçu par l'auditeur, qui sera à la fois influencé par l'accent phonétique et l'accent linguistique (théorique, plus ou moins conscientisé par l'auditeur).

Nous proposons de présenter en détails les caractéristiques fonctionnelles et acoustiques de l'accent lexical en anglais, puis celles de l'accent en français et en japonais. Nous nous intéresserons ensuite à l'impact de l'accent lexical sur la compréhensibilité du locuteur, aux difficultés de perception et de production de l'accent en anglais L2, et enfin aux différents moyens existants pour mesurer automatiquement l'accent lexical.

3.3.1 L'accent lexical en anglais

En anglais, l'accent lexical se manifeste par des modifications à la fois prosodiques et segmentales des voyelles. Les syllabes accentuées sont généralement plus longues, plus fortes, plus hautes, et présentent un mouvement de la F_0 plus important, avec une qualité vocalique dite « pleine », comparativement aux syllabes non accentuées qui auront tendance à être « réduites » (Cutler, 2015). Ainsi, l'accentuation d'une syllabe affecte les syllabes non accentuées environnantes, les rendant plus courtes, moins fortes, moins hautes, centralisées et relâchées (Tortel, 2021). La voyelle réduite par excellence en anglais est le *schwa*, noté /ə/, mais le phénomène d'accentuation-réduction doit plutôt être considéré comme un continuum, allant de très accentué (quand se superposent l'accent lexical, l'accent de phrase et l'accent contrastif) à complètement réduit, voire supprimé. On distingue ainsi jusqu'à 7 niveaux d'accentuation, mais la plupart des auteurs s'accordent à dire que 4 niveaux sont phonologiquement pertinents : l'accent primaire, l'accent secondaire, la syllabe pleine non-accentuée (ou accent tertiaire), et la syllabe réduite (Frost, 2023).

Le rôle principal de l'accent lexical est la segmentation du flux de parole et la désambiguïsation lexicale. En anglais, les mots pleins (noms, verbes, adjectifs, adverbes etc.) sont généralement accentués, tandis que les mots grammaticaux (prépositions, déterminants, particules etc.) sont généralement réduits (Tortel, 2021). L'accent

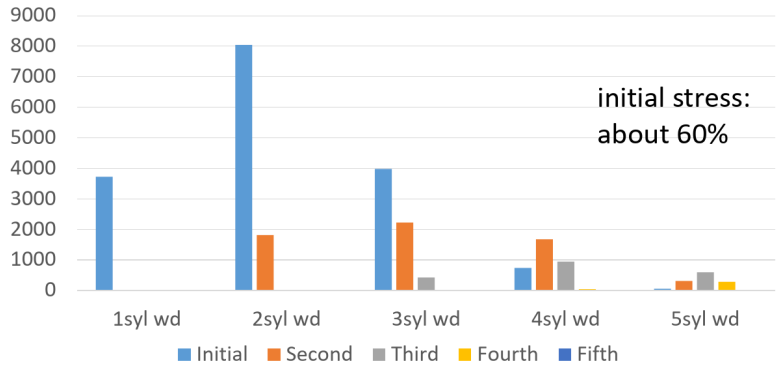

FIG. 3.1 : Distribution de la position de l'accent lexical en anglais dans les mots de 1 à 5 syllabes, à partir de la base de données CELEX (Sugahara, 2020, p. 8)

joue également un rôle important dans la morphologie dérivationnelle, car il change fréquemment selon la catégorie du mot (**per**son vs. per**so**nifier) et aide à distinguer des mots au sein de la même catégorie (**pho**tograph vs. pho**to**grapher). Ainsi, les noms et adjectifs ont tendance à porter l'accent sur la première syllabe, tandis que les verbes sont plus souvent accentués sur la deuxième syllabe.

Les paires minimales se distinguant seulement par la position de l'accent sont rares en anglais, étant donné que celui-ci s'accompagne souvent de variations segmentales des voyelles (Cutler & Jesse, 2021).

Si la position de l'accent lexical est variable en anglais, il est toutefois important de noter que la majorité des mots se trouveront accentués sur la première syllabe. Selon Sugahara (2020), 60 % des lemmes de la base de données CELEX (Baayen et al., 1995) sont accentués sur la première syllabe (cf. graphique 3.1). En analysant un corpus de 190 000 mots issus de conversations spontanées en anglais britannique, Cutler et Carter (1987) constatent que 90 % des mots lexicaux commencent par une syllabe accentuée. Ils en concluent que l'auditeur anglophone s'appuie certainement sur les syllabes accentuées pour repérer les frontières de mot dans le flux de parole.

3.3.2 L'accent lexical en français et en japonais

Le français n'a pas d'accent lexical (Vaissière & Michaud, 2006), mais il n'est pas pour autant dénué d'accentuation. On distingue en général deux types d'accents : l'accent emphatique, qui permet d'attirer l'attention de l'auditeur de manière ponctuelle sur un mot de l'énoncé, et l'accent non-emphatique qui, contrairement au premier, est systématiquement placé en fin de groupe rythmique. Nous nous intéresserons ici seulement à l'accent non-emphatique.

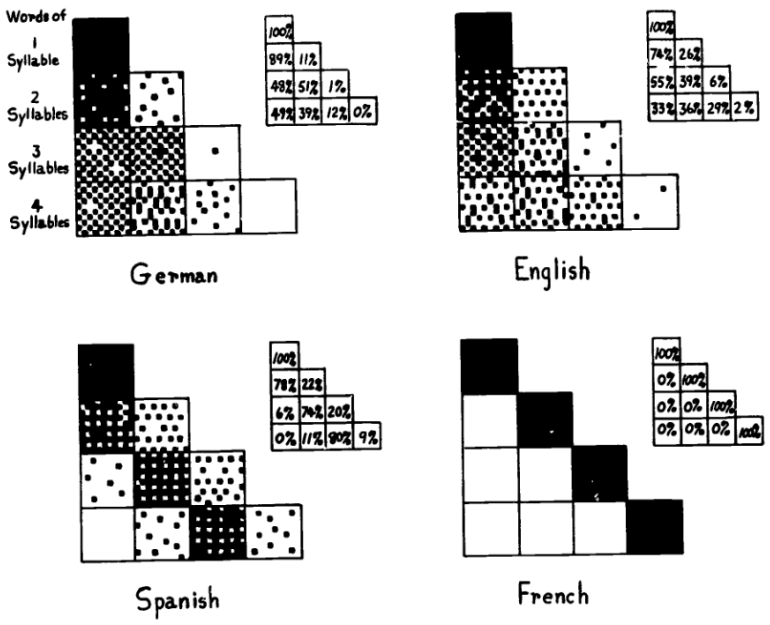

FIG. 3.2 : Comparaison de la position de l'accent primaire dans quatre langues (Delattre, 1963, p. 200)

Le français est traditionnellement décrit comme une langue à accentuation finale, aussi appelée oxytonique, où l'accent (non-emphatique, donc) tombe sur la dernière syllabe d'un groupe de mots (Astesano, 2001). On distingue généralement deux niveaux de groupes rythmiques en français. Le plus grand est appelé groupe de souffle ou unité intonative ; il peut contenir plusieurs groupes plus petits appelés groupes accentuables, composés d'un mot lexical (accentuable) et éventuellement de mots grammaticaux qui en dépendent (généralement non accentués). Selon Di Cristo (1998), la dernière syllabe des groupes accentuables est systématiquement accentuée. Il est parfois fait mention d'un accent rythmique secondaire placé sur la syllabe initiale et permettant de délimiter les unités intonatives (Di Cristo & Hirst, 1993 ; Fónagy, 1980).

Dans une étude comparative, Delattre (1963) analyse la position de l'accent primaire (théorique) dans un corpus de textes de 1500 mots en français et en espagnol, 2400 mots en allemand et 5800 mots en anglais. Il montre que l'anglais et le français se comportent de manière diamétralement opposée, le premier exhibant une grande variabilité, l'autre étant étonnamment stable. La figure 3.2 présente le pourcentage d'accentuation de chaque syllabe pour les mots d'une à quatre syllabes, dans les quatre langues analysées.

L'accentuation du français est avant tout caractérisée par une variation de durée de syllabe (Astesano, 2001 ; Di Cristo, 1998). Cette variation est d'autant plus grande

qu'elle se cumule avec l'allongement naturel de fin de groupe, ainsi la dernière syllabe d'un groupe a tendance à paraître notoirement plus longue que les précédentes (Nord et al., 1990), sans qu'il soit clairement établi quelle proportion de l'allongement final est due à la position en fin de groupe, et quelle proportion est due à l'accentuation (Astesano, 2001).

D'après Vaissière (1991), la fréquence fondamentale sert en français principalement à marquer la frontière des mots, et non pas à accentuer une syllabe. La F_0 a tendance à monter en fin de mot (de manière étalée sur plusieurs syllabes), pour reprendre plus bas au début du mot suivant ; ou bien à tomber si le mot se situe en fin de groupe de souffle.

L'intensité quant à elle ne semble pas être un paramètre déterminant de l'accentuation du français. Il apparaît même que la voyelle d'une syllabe finale (donc accentuée) est en moyenne moins intense que les autres syllabes (-0.5 dB en français, contre 4.4 dB en anglais pour la syllabe accentuée, Delattre, 1966). Comme pour les autres dimensions, il n'est pas évident de distinguer la variation due à la position de la syllabe (on observe souvent une baisse de l'intensité en fin de groupe de souffle, Di Cristo, 1998), de celle qui serait due spécifiquement à la position de l'accent.

La fonction première de l'accent non emphatique en français est de structurer l'énoncé en groupes de sens (Astesano, 2001). Il permet à l'auditeur de segmenter le flux de parole et de focaliser son attention sur les informations importantes ou nouvelles. Il peut avoir également des fonctions secondaires expressives, contrastives ou rythmiques.

Selon Sugahara (2020), les deux dialectes principaux du japonais, à savoir le dialecte de Tōkyō et celui du Kansai, ont tous les deux une accent lexical caractérisé seulement par une variation de la F_0. De ce fait, on parle souvent d'accent de hauteur ou *pitch accent*, mais il s'agit bien d'un accent lexicalement contrastif. D'après Shibata et Shibata (1990), 13,6 % des homophones japonais sont distingués exclusivement par la position de l'accent, par exemple ***hasi*** (baguettes de table), ***hasi*** (pont) ou ***hasi*** (limite). Toutefois la position de l'accent varie en fonction du dialecte. On pourra ainsi avoir ***kokoro*** à Tōkyō et ***kokoro*** dans le Kansai (cœur, esprit), ou encore ***kamakiri*** à Tōkyō et ***kamakiri*** dans le Kansai (mante religieuse).

Précisons que l'accent est communément rattaché à la more, et non à la voyelle ou à la syllabe. La more est une unité légèrement plus petite que la syllabe, composée dans le cas du japonais d'un noyau vocalique éventuellement précédé d'une consonne et d'un glide, ou elle peut être aussi une consonne nasale seule, un coup de glotte ou un allongement vocalique.

La position de l'accent est régie par un système complexe qui varie selon la région et l'origine du mot, mais sa position la plus courante semble être la more antépénul-

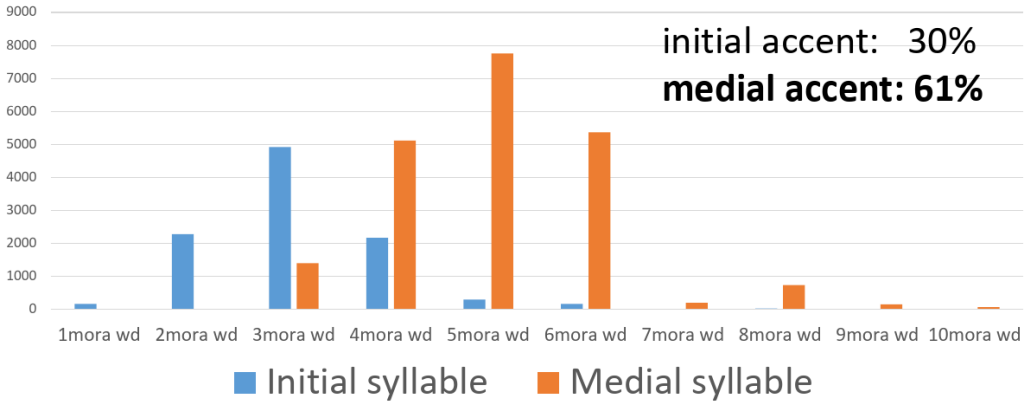

FIG. 3.3 : Distribution de la position de l'accent lexical dans les mots de 1 à 10 mores du Ōsaka-Tōkyō Accent Dictionary (Sugahara, 2020, p. 15)

tième ou la pénultième Kubozono (2006). La figure 3.3 présente la distribution de l'accent lexical dans les mots de 1 à 10 mores dans le Ōsaka-Tōkyō Accent Dictionary REFERENCE. On peut voir que la majorité des mots sont accentués sur une more en position médiale (61%). Si la more initiale porte généralement l'accent sur les mots de 1 à 3 mores, Sugahara (2020) indique que le japonais a tendance à avoir des mots de plus de trois mores (comme le montre la distribution), de part sa morphologie agglutinante. Par ailleurs, l'accent a tendance à se rapprocher de la frontière du morphème (et donc en position médiale) quand de nouveaux éléments viennent s'y ajouter. Ainsi ***kyooto*** (Kyōto) devient *kyooto-si* (la ville de Kyōto), mais *kyooto-**dai**gaku* (l'université de Kyōto). Il serait intéressant de connaître la distribution de le position de l'accent sur un corpus de mots courants, mais nous n'avons malheureusement pas pu trouver cette information.

Ainsi il est généralement admis que le japonais est une langue à accent lexical, majoritairement en position médiale, et caractérisé par une variation de la F_0 sans modification significative des autres dimensions prosodiques ou segmentales (Sugahara, 2020).

3.3.3 L'accent lexical en anglais L2

Dans les contextes d'apprentissage d'une langue seconde, les locuteurs/auditeurs non-natifs sont souvent influencés par les règles prosodiques de leur langue maternelle, et cela peut poser plus ou moins de problème selon que ces règles ou tendances diffèrent de la langue cible (Cutler, 2015). Par exemple, le locuteur francophone, habitué à un accent fixe sur le syllabe finale et une qualité et une quantité vocalique

constante dans les voyelles non accentuées, aura tendance à accentuer la dernière syllabe des mots en anglais et à ne pas réduire les syllabes précédentes (Tortel & Hirst, 2010). On peut s'attendre par ailleurs à ce que cette accentuation soit plus prononcée en termes de durée de syllabe, que de variation de F_0 ou d'intensité, comme ces deux derniers paramètres ne semblent pas particulièrement exploités pour accentuer les syllabes. De plus, puisque l'accentuation en français ne joue pas de rôle de désambiguïsation lexicale comme en anglais, les locuteurs francophones ont souvent des difficulté à conscientiser les patterns accentuelles de l'anglais, et peuvent avoir du mal à reconnaître leur propre tendance à accentuer les syllabes finales. Dupoux et al. (1997) proposent le terme de « surdité accentuelle » *(stress deafness)* pour décrire cette capacité limitée à percevoir et à être conscient de l'accent, notant que les locuteurs de langues à accent fixe rencontrent plus de difficultés comparativement à ceux des langues à accent lexical. De plus, adopter un rythme différent de celui de sa langue maternelle peut être psychologiquement éprouvant, car celui-ci est ancré depuis l'enfance et fortement associé avec sa personnalité et sa culture (Calbris & Montredon, 1975). En conséquence, un accent mal placé et l'absence de réduction syllabique peuvent significativement entraver la segmentation et la reconnaissance des mots pour les auditeurs (Cutler, 2015). Tortel (2021) souligne que les apprenants francophones de l'anglais devraient prioriser l'amélioration de la position de l'accent lexical, le contraste entre les syllabes accentuées et réduites, éviter l'allongement des syllabes finales non accentuées, et réduire les mots fonctionnels.

Si le japonais a quant à lui un accent lexical, et que les locuteurs japonophones semblent avoir moins de difficultés à percevoir et produire l'accent en anglais, ils restent toutefois influencés par la distribution de l'accent du japonais – plus souvent en position médiale –, et ont tendance à ne pas réduire les voyelles non-accentuées (Sugahara, 2011, 2016).

3.3.4 Accent lexical et évaluation de la compréhensibilité

Isaacs et Trofimovich (2012) constatent que l'accentuation lexicale est le troisième paramètre le plus corrélé avec la compréhensibilité, parmi les 19 paramètres qu'ils analysent (cf. section 2.2.1). Ils calculent un *word stress error ratio* à partir du nombre de mots polysyllabiques dont l'accent primaire est mal placé ou absent, divisé par le nombre de mots polysyllabiques. La corrélation entre la proportion d'erreur d'accentuation et le jugement de compréhensibilité est de $-0,76$ ($p < 0,01$), suivie immédiatement de la proportion de réduction vocalique ($0,74$, $p < 0,01$). Contrairement aux paramètres de fluence qui apparaissent plus discriminant dans les petits niveaux, l'accentuation lexicale est discriminante pour tous les niveaux de locuteurs.

Saito et al. (2015) reprennent les données de Isaacs et Trofimovich (2012), et pro-

posent à une autre cohorte d'évaluateurs experts et non-experts d'évaluer chaque locuteur en termes de compréhensibilité, puis selon 11 critères linguistiques[11]. Il s'agit cette fois de voir sur quelles dimensions les évaluateurs s'appuient explicitement pour juger les locuteurs, et observer quels scores obtenus par dimension sont les plus corrélés avec le jugement global de compréhensibilité. Deux éléments dans leurs résultats sont intéressants à mentionner ici. Tout d'abord, les évaluations subjectives du rythme et de l'accentuation lexicale apparaissent fortement corrélées avec les annotations effectuées par Isaacs et Trofimovich (2012) ($r = 0,76$, $p < 0,01$ entre le critère « rythme » et la proportion de réduction vocalique – c'est la deuxième corrélation la plus haute parmi les 11 critères, après celle calculée entre le débit de parole et la longueur moyenne des énoncés, $0,78$ $p < 0,01$ – et $r = 0,70$, $p < 0,01$ entre le critère « accentuation lexicale » et le *word stress error ratio*). Ensuite, le critère « rythme » apparaît le plus corrélé avec le jugement de compréhensibilité parmi les 5 critères de prononciation $(0,79)$, alors que le critère « accentuation lexicale » n'est qu'en quatrième position $(0,62)$ après les erreurs segmentales $(0,75)$, le débit de parole $(0,66)$, et avant l'intonation $(0,54)$.

Une autre étude intéressante est celle de Field (2005). Il s'intéresse à l'impact du déplacement de l'accent lexical sur la reconnaissance de mots isolés. Il constate qu'un déplacement de l'accent vers la droite impacte plus l'intelligibilité du mot qu'un déplacement vers la gauche, ce qui semble cohérent avec la tendance majoritaire en anglais à l'accent en initiale. Par ailleurs, le même déficit d'intelligibilité est observé quelque soit la langue maternelle de l'auditeur[12].

3.3.5 Mesures automatiques de l'accent lexical

Plusieurs études ont proposé des systèmes de classification automatique de l'accent lexical depuis le début des années 2000. La plupart de ces systèmes s'appuient sur des mesures de F0, d'intensité et de durée de syllabe ou de segments vocaliques (J.-Y. Chen & Wang, 2010 ; L.-Y. Chen & Jang, 2012 ; Deshmukh & Verma, 2009 ; Johnson & Kang, 2015 ; K. Li et al., 2018 ; Tepperman & Narayanan, 2005). Certains systèmes intègrent également des informations segmentales, telles que les coefficients cepstraux (Ferrer et al., 2015 ; C. Li et al., 2007). Toutes ces études ont entraîné leurs systèmes sur des corpus de parole lue, voire de mots isolés, parfois générés artificiellement à l'aide de systèmes de synthèse de parole. C'est le cas par exemple de Korzekwa

[11] 5 critères d'évaluation de l'enregistrement audio (erreurs segmentales, accent lexical, intonation, rythme, débit de parole) et 6 critères d'évaluation de la transcription des enregistrements (précision et richesse du lexique, précision et complexité grammaticale, richesse et cohésion du discours).

[12] Field a mené son expérimentation sur 82 auditeurs natifs et 76 locuteurs non-natifs, dont les langues maternelles sont le coréen (16), le japonais (15), le mandarin (10), l'espagnol (9), le portugais (6) et l'italien (6), ainsi que d'autres langues avec moins de 5 locuteurs respectifs.

et al. (2021), qui ont entraîné un réseau de neurones à attention pour détecter les mots dont l'accent primaire n'est pas placé correctement. À partir des paramètres acoustiques classiques (F_0, intensité et durée) extraits sur chaque phonème d'un corpus de mots isolés partiellement générés par synthèse vocale, ils mettent au point un modèle capable de déterminer si un mot est bien accentué avec une précision de 94,8 %. Toutefois, la moitié des erreurs présentes dans le corpus ne sont pas détectée (rappel de 49,2 %). Les auteurs ajoutent que leur modèle n'est pas adapté à l'analyse de mots en contexte, les résultats étant biaisés par les phénomènes de liaison et de coarticulation. Des performances similaires étaient obtenues quelques années avant en combinant coefficients cepstraux et paramètres prosodiques dans un modèle à mélange de gaussiennes (Ferrer et al., 2015).

On constate que les études qui tentent de mesurer les patterns accentuels des mots, malgré leurs infrastructures complexes et leur grand nombre de paramètres (jusqu'à 39 lorsque les coefficients cepstraux sont utilisés par Ferrer et al., 2015), sont limitées et semblent souvent déconnectées de la réalité de la parole et des enjeux pédagogiques. Saito et al. (2022) se rendent à l'évidence : on ne sait pas encore analyser automatiquement la précision de l'accentuation lexicale. Par ailleurs, nous n'avons trouvé à ce jour aucun système proposant de mesurer le degré de contraste prosodique entre les syllabes.

3.4 Conclusion

Nous avons vu dans ce chapitre que les patterns de pause et d'accentuation lexicale participent à rendre la parole du locuteur plus ou moins compréhensible. Les pauses, ou plus largement les interruptions du discours – qu'il s'agisse d'hésitations, de faux départ, de répétitions etc. –, participent à structurer le message et aident l'auditeur à traiter l'information. Leur présence peut aussi perturber la compréhension, en particulier lorsqu'elles interviennent à l'intérieur d'un constituant syntaxique. Le lien entre les pauses et la syntaxe est clair, et il semble difficilement concevable d'effectuer des mesures sur les pauses sans considérer leur position dans l'énoncé. En outre, nous retenons que la notion de pause n'est pas absolue. On peut percevoir une pause purement par cohérence syntaxique (sans aucun indice acoustique), ou au contraire ne pas la percevoir malgré la présence d'un silence ou d'une hésitation prolongée. Toujours est-il que la présence d'une pause n'est pas anodine : perçue ou non, elle semble avoir un impact sur la qualité de la transmission du message et sur la perception de fluence qu'en aura l'auditeur.

La qualité de réalisation de l'accent lexical semble aussi, du moins en anglais, impacter la perception de compréhensibilité du locuteur. Le fait que les patterns ac-

centuels de la L1 transparaissent souvent dans la L2 peut rendre la segmentation ou la reconnaissance du mot difficile, surtout quand les tendances sont opposées, comme en français (accent en finale) et en anglais (tendance à l'initiale). Quatre dimensions sont en jeu en anglais pour réaliser cette accentuation : une variation de hauteur, d'intensité, de durée et de qualité vocalique. Ces 4 dimensions sont intriquées et peuvent être plus ou moins exploitées selon les contextes et les locuteurs. Selon les langues, l'accent n'est pas toujours réalisé à l'aide de ces 4 paramètres : le français aura tendance à privilégier la variation de durée, là où le japonais s'appuiera sur la hauteur ; et l'utilisation des autres dimensions peut s'avérer difficile. Ce qui ressort clairement est la tendance des syllabes de l'anglais à se réduire ou s'accentuer de manière marquée, tandis que ce phénomène n'est pas présent, ou beaucoup moins marqué, en français et en japonais.